Diagnostic performance of the ACR/EULAR 2010 criteria for rheumatoid arthritis and two diagnostic algorithms in an early arthritis clinic (REACH)

Celina Alves,[1] Jolanda Jacoba Luime,[1] Derkjen van Zeben,[2] Anne-Margriet Huisman,[2] Angelique Elisabeth Adriana Maria Weel,[3] Pieternella Johanna Barendregt,[1,3] Johanna Maria Wilhelmina Hazes[1]

► Additional data (supplementary figures) are published online only. To view these files please visit the journal online at (http://ard.bmj.com)

[1]Department of Rheumatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
[2]Department of Rheumatology, Sint Franciscus Gasthuis, Rotterdam, The Netherlands
[3]Department of Rheumatology, Maasstad Ziekenhuis, Rotterdam, The Netherlands

Correspondence to
J. Luime, Department of Rheumatology, Erasmus MC, University Medical Center Rotterdam, PO Box 2040, 3000 CA Rotterdam, The Netherlands; j.luime@erasmusmc.nl

Accepted 24 April 2011
Published Online First
27 May 2011

ABSTRACT

Introduction An ACR/EULAR task force released new criteria to classify rheumatoid arthritis at an early stage. This study evaluates the diagnostic performance of these criteria and algorithms by van der Helm and Visser in REACH.

Methods Patients with symptoms ≤12 months from REACH were used. Algorithms were tested on discrimination, calibration and diagnostic accuracy of proposed cut-points. Two patient sets were defined to test robustness; undifferentiated arthritis (UA) (n=231) and all patients including those without synovitis (n=513). The outcomes evaluated were methotrexate use and persistent disease at 12 months.

Results In UA patients all algorithms had good areas under the curve 0.79, 95% CI 0.73 to 0.83 for the ACR/EULAR criteria, 0.80, 95% CI 0.74 to 0.87 for van der Helm and 0.83, 95% CI 0.77 to 0.88 for Visser. All calibrated well. Sensitivity and specificity were 0.74 and 0.66 for the ACR/EULAR criteria, 0.1 and 1.0 for van der Helm and 0.59 and 0.93 for Visser. Similar results were found in all patients indicating robustness.

Conclusion The ACR/EULAR 2010 criteria showed good diagnostic properties in an early arthritis cohort reflecting daily practice, as did the van der Helm and Visser algorithms. All were robust. To promote uniformity and comparability the ACR/EULAR 2010 criteria should be used in future diagnostic studies.

Recently an American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) task force released new classification criteria for rheumatoid arthritis (RA) at an early stage.[1] These criteria might also have diagnostic value early in the disease process although this has not yet been evaluated. Early diagnosis is important to improve patient outcome by early treatment to prevent joint damage and functional impairment.[2]

The previous classification criteria for RA (the 1987 ACR criteria) proved inadequate in the early stages of disease.[3 4] This led to the development of other diagnostic algorithms.[5 6] These algorithms showed good diagnostic performance and identified patients at an early stage of the disease.[7–9]

Diagnostic algorithms tend to be overoptimistic in their capabilities when only tested in the population they were derived from.[10] For instance, if a high erythrocyte sedimentation rate (ESR) is an important predictor for RA but in the derivation cohort by chance only a few patients had a high ESR, the data-driven way in which these algorithms are build will not identify this predictor. Therefore before use in practice the discriminative abilities of such algorithms should be tested in another cohorts with similar patients (similar incidence rate). In addition, the robustness of algorithms to variation of incidence rates can be tested in cohorts with different previous disease probabilities.[11–13]

We aim to evaluate the diagnostic performance of the ACR/EULAR 2010 criteria and two diagnostic algorithms simultaneously to predict methotrexate use or persistent disease in the Rotterdam Early Arthritis Cohort (REACH). In addition, we will test robustness after defining two patient sets in the same cohort resulting in different previous probabilities of developing RA.

METHODS

Diagnostic algorithms

Three diagnostic algorithms were evaluated. The first is the new ACR/EULAR 2010 criteria set.[1] The other two, the algorithms by van der Helm and the one by Visser, are existing, well-known models.[5 6]

Validation cohort

Clinical data used were from REACH. This ongoing, prospective, inception cohort study was set up in the greater Rotterdam area in July 2004. Patients were recruited either via their general practitioner, or via the outpatient rheumatology clinic of three hospitals at first consultation. Patients were included in case of one or more swollen joint or, in the absence of joint swelling, if they had two or more joints with pain or loss of movement with two or more of the following criteria: morning stiffness for more than 1 h; unable to clench a fist in the morning; pain when shaking someone's hand; pins and needles in the fingers; difficulties wearing rings or shoes; a family history of RA; unexplained fatigue for less than 1 year. Patients were excluded if their symptoms resulted from trauma or overexertion, were for over 12 months, or if they were younger than 16 years.

A trained research nurse at the REACH clinic took a standardised history and conducted a physical examination at baseline, 6 and 12 months, including blood and urine samples. For the current analysis data from baseline and 1 year were used. Physical examination included the measurement of tender and swollen joints, using a 44 joint count. Laboratory variables included IgM-rheumatoid factor (ELISA), anti-cyclic citrullinated peptide (Elia CCP on immunoCAP 250; Phadia Freiburg, Germany),C-reactive protein (local standards) and ESR (local standards). x-Rays of hands and feet were assessed for bony erosions at baseline. For a detailed description of REACH, see Geuskens et al.[14]

Statistical analyses

To asses overall performance the prediction algorithms were tested on discrimination and calibration.[15] Discrimination is the ability of an algorithm to differentiate correctly between patients with and without the disease. Calibration reveals the ability to estimate the probability of the diagnosis for individuals correctly by comparing the probability predicted by the algorithm and the observed probability. To evaluate discrimination receiver operating characteristic curves, including corresponding areas under the curve (AUC), were calculated. Calibration was evaluated using calibration plots and the Hosmer–Lemeshow test.[15] The latter indicates good calibration if a non-significant result appears. To assess diagnostic performance of the algorithms sensitivity, specificity, positive predictive values (PPV) and negative predictive values (NPV) were estimated at cut-points proposed for treatment initiation among patients at risk of RA. For the ACR/EULAR 2010 criteria and Visser algorithm a score of 6 or more[1 16] was used and for van der Helm a score of 8 or more[5] was used. To test robustness this analysis was repeated among all patients included in REACH. This group had a lower previous disease probability by a case-mix of synovitis and inflammatory joint complaints without synovitis. Synovitis was defined as joint swelling.

As a classifier for correct diagnosis two outcomes were evaluated at 1 year: the use of methotrexate and persistent disease, defined as synovitis present at physical examination after 1 year, or the use of disease-modifying antirheumatic drugs (DMARD) including biological agents. Patients with a definite alternative diagnosis such as gout were not classified as persistent disease. A complete case analysis was done.

RESULTS

Validation cohort

Up to 31 October 2008, 875 patients were referred to REACH and had 1-year follow-up. One hundred and 13 patients did not fulfil the inclusion criteria and 31 patients were lost to follow-up at baseline (see supplementary figure S1, available online only). Patients used in the development of the ACR/EULAR 2010 criteria were excluded (n=216).[1] Table 1 reports baseline characteristics of all patients (n=513). Patients had a mean age of 50 years, 73% were women and the median symptom duration was 106 days (range 1–366 days). At baseline 48% (n=246) presented with synovitis. After 1 year, 148 of 513 used methotrexate, of whom 22 did not have synovitis at baseline, and 231 of 513 patients had persistent disease, of whom 59 did not have synovitis at baseline.

Discrimination

Table 2 shows AUC of each diagnostic algorithm for both outcomes. In undifferentiated arthritis (UA) patients (n=231) the AUC for methotrexate use were comparable, with overlapping 95% CI, 0.79 (95% CI 0.73 to 0.83) for the ACR/EULAR 2010 criteria, 0.80 (95% CI 0.74 to 0.87) for the van der Helm algorithm and 0.83 (95% CI 0.77 to 0.88) for the Visser algorithm. For persistent disease the AUC were 0.77 (95% CI 0.71 to 0.85) for the ACR/EULAR 2010 criteria, 0.78 (95% CI 0.71 to 0.85) for the van der Helm algorithm and 0.77 (95% CI 0.71 to 0.83) for the Visser algorithm. In all patients (n=513) the AUC were comparable for both outcomes, with slightly better performance of the van der Helm algorithm; 0.88 (95% CI 0.84 to 0.91) and 0.83 (95% CI 0.79 to 0.87).

Calibration

Calibration plots of all diagnostic algorithms are shown in figure S2 (see supplementary figure S2, available online only). In UA patients (n=513) calibration was worse than in all patients, although the Hosmer–Lemeshow test was not significant for any of the calibration plots. All algorithms calibrated well in all patients (n=513), confirmed by the Hosmer–Lemeshow test.

Evaluating diagnostic performance using proposed cut-points

To identify patients in need of treatment proposed cut-points were tested in UA patients. The ACR/EULAR criteria showed a sensitivity of 0.74 (95% CI 0.65 to 0.82) and a specificity of 0.66 (95% CI 0.54 to 0.76), with the cut-point of 6 or higher using methotrexate as a classifier for correct diagnosis (table 3). The Visser algorithm and the van der Helm algorithm had a lower sensitivity, 0.47 and 0.59 for the Visser algorithm for both outcomes and 0.08 and 0.10 for the van der Helm algorithm. Specificity was higher: 0.93 for the Visser algorithm and 1.0 for the van der Helm algorithm.

The PPV is the probability that a patient has the disease if the test is positive. The van der Helm algorithm had the highest PPV; 1.0. The NPV is the opposite probability and was highest for the ACR/EULAR criteria with 0.63 for methotrexate use and 0.46 for persistent disease, slightly higher than the Visser algorithm.

DISCUSSION

The results of our study show that the new ACR/EULAR 2010 criteria could aid diagnostics in early arthritis patients. They had good overall performance, with a sufficiently high AUC and good performance of the proposed cut-point of 6 for persistent disease, which could be considered RA. The other algorithms performed well when tested for discriminatory properties

Table 1 Patient characteristics for each patient set

	UA (n=231)	All patients (n=513)
Women (%)	68	73
Age, years (mean, SD)	53 (14)	50 (14)
SJC (median, range)	4 (1–38)	0 (0–38)
TJC (median, range)	7 (0–42)	6 (0–42) mv=2
RF positive (%)	35%	26%
Anti-CCP positive (%)	28% mv=6	19% mv=10
ESR, mm/h (median, range)	18 (1–103) mv=7	14 (0–103) mv=15
CRP, mg/l (median, range)	6 (1–180) mv=16	5 (1–180) mv=40
Erosions (%)	9% mv=4	4% mv=9
RA, according to 1987 ACR criteria	29% mv=3	14% mv=5
RA, according to 2010 ACR/EULAR criteria	45% mv=12	58% mv=6
Persistent arthritis at 1 year	45% mv=9	71% mv=3

ACR, American College of Rheumatology; CCP, cyclic citrullinated protein; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate; EULAR, European League Against Rheumatism; mv, missing values; RA, rheumatoid arthritis; RF, rheumatoid factor; SJC, swollen joint count; TJC, tender joint count; UA, undifferentiated arthritis.

Table 2 Area under the receiver operating characteristic curves with 95% CI for each algorithm and each patient set

	ACR/EULAR 2010	Van der Helm	Visser
Outcome methotrexate use			
UA patients	0.79 (0.73 to 0.85)	0.80 (0.74 to 0.87)	0.83 (0.77 to 0.88)
All patients	0.79 (0.75 to 0.83)	0.88 (0.84 to 0.91)	0.80 (0.76 to 0.85)
Outcome persistent disease			
UA patients	0.77 (0.71 to 0.85)	0.78 (0.71 to 0.85)	0.77 (0.71 to 0.83)
All patients	0.74 (0.70 to 0.78)	0.83 (0.79 to 0.87)	0.74 (0.70 to 0.79)

UA, undifferentiated arthritis.

Table 3 Sensitivity, specificity, and PPV and NPV at the proposed cut-points in UA patients

	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
methotrexate use				
ACR/EULAR 2010 criteria	0.74 (0.65 to 0.82)	0.66 (0.54 to 0.76)	0.76 (0.67 to 0.83)	0.63 (0.52 to 0.73)
van der Helm algorithm	0.10 (0.05 to 0.17)	1.0 (0.95 to 1.0)	1.0 (0.74 to 1.0)	0.43 (0.35 to 0.50)
Visser algorithm	0.59 (0.50 to 0.68)	0.93 (0.85 to 0.97)	0.92 (0.84 to 0.97)	0.62 (0.53 to 0.71)
Persistent disease				
ACR/EULAR 2010 criteria	0.69 (0.61 to 0.76)	0.72 (0.59 to 0.83)	0.87 (0.80 to 0.92)	0.46 (0.35 to 0.56)
van der Helm algorithm	0.08 (0.05 to 0.14)	1.0 (0.93 to 1.0)	1.0 (0.75 to 1.0)	0.27 (0.21 to 0.34)
Visser algorithm	0.47 (0.39 to 0.55)	0.93 (0.84 to 0.98)	0.95 (0.87 to 0.99)	0.40 (0.31 to 0.48)

ACR, American College of Rheumatology; EULAR, European League Against Rheumatism; NPV, negative predictive value; PPV, positive predictive value, UA, undifferentiated arthritis.

(AUC and calibration), although the van der Helm algorithm failed to detect cases at the proposed cut-point. To promote uniformity and comparability of studies we would suggest using the ACR/EULAR 2010 criteria in future diagnostic studies.

The cut-point of 6 in the ACR/EULAR 2010 criteria was well chosen and showed good diagnostic performance, even though it was not intended for diagnostic purposes.[1] Choosing a cut-point is a trade-off between harm of treatment in non-cases (overtreatment) and harm of no treatment in cases (undertreatment).[13 17] Ideally a cut-point has a high sensitivity to prevent undertreatment and a high specificity to prevent overtreatment. However, a high specificity is often accompanied by a low to moderate sensitivity and vice versa. For the ACR/EULAR 2010 criteria both sensitivity and specificity were approximately 70%. Using this cut-point of 6 to start treatment, in this study 30% of persistent patients would not be treated, whereas 30% of the non-persistent patients would have been. Lowering the cut-point to 4 increases sensitivity to 0.92 at the cost of specificity (0.33). Increasing it to 7 had a sensitivity of 0.53 and a specificity of 0.85. Perhaps creating a low, intermediate and high-risk group for disease using dual cut-points would enable treatment with different intensities.

The cut-point of 6 was chosen using the AUC of three cohorts, including our own. In this study the AUC for methotrexate use, 0.79, was similar to that in the derivation article (0.66–0.82), indicating consistency. It was also similar (0.77) for persistent disease. It could be argued that this is a direct result of the use of our data in the derivation cohort. However, the decision to use 6 as cut-point was based on expert opinion and two other cohorts. Furthermore, patients included in the derivation of the ACR/EULAR 2010 criteria were removed from analyses.

The strengths of our study include the heterogeneity of patients' subsets to test robustness of the algorithms and simultaneous evaluation of three diagnostic algorithms in one study sample. We showed that the ACR/EULAR criteria and both algorithms were robust in a case-mix of synovitis and non-synovitis patients. Calibration was good for all algorithms, but not perfect. Calibration and robustness have not been evaluated before by others, but discrimination was. The van der Helm algorithm showed AUC of 0.82–0.88 and the Visser algorithm an AUC of 0.70, both similar to the AUC in the present study.[7–9]

This study should be interpreted in the light of current developments in diagnostic research in RA. Current diagnostic studies within RA are faced with defining a suitable outcome. We defined two outcomes; methotrexate use similar to the definition of the ACR/EULAR 2010 and persistent disease (either synovitis or DMARD use at 12 months).[1] This may have led to misclassification in two ways. Patients could be classified as true positive because they were still using methotrexate or other DMARD at 12 months, whereas in fact some patients may not need treatment. Likewise, patients may have had episodes of arthritis with no episode or DMARD use at 12 months, while later on they developed persistent arthritis.

In conclusion, the new ACR/EULAR 2010 criteria showed good diagnostic properties in an early arthritis cohort reflecting daily clinical practice, as did the van der Helm and Visser algorithms. All were robust. To promote uniformity and comparability we would suggest using the ACR/EULAR 2010 criteria in future diagnostic studies.

Acknowledgements The authors would like to thank all patients and general practitioners who participated as well as the research nurses, trainee rheumatologists and rheumatologists (other than the authors) who contributed to the REACH cohort.

Funding Data management was partly funded by the Dutch Arthritis Association.

Competing interests None.

Ethics approval This study was conducted with the approval of the medical ethics committee at Erasmus MC.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Aletaha D,** Neogi T, Silman AJ, *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis* 2010;**69**:1580–8.
2. **Finckh A,** Liang MH, van Herckenrode CM, *et al.* Long-term impact of early treatment on radiographic progression in rheumatoid arthritis: a meta-analysis. *Arthritis Rheum* 2006;**55**:864–72.
3. **Arnett FC,** Edworthy SM, Bloch DA, *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988;**31**:315–24.
4. **Banal F,** Dougados M, Combescure C, *et al.* Sensitivity and specificity of the American College of Rheumatology 1987 criteria for the diagnosis of rheumatoid arthritis according to disease duration: a systematic literature review and meta-analysis. *Ann Rheum Dis* 2009;**68**:1184–91.
5. **van der Helm-van Mil AH,** le Cessie S, van Dongen H, *et al.* A prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: how to guide individual treatment decisions. *Arthritis Rheum* 2007;**56**:433–40.
6. **Visser H,** le Cessie S, Vos K, *et al.* How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. *Arthritis Rheum* 2002;**46**:357–65.
7. **van der Helm-van Mil AH,** Detert J, le Cessie S, *et al.* Validation of a prediction rule for disease outcome in patients with recent-onset undifferentiated arthritis: moving toward individualized treatment decision-making. *Arthritis Rheum* 2008;**58**:2241–7.
8. **Kuriya B,** Cheng CK, Chen HM, *et al.* Validation of a prediction rule for development of rheumatoid arthritis in patients with early undifferentiated arthritis. *Ann Rheum Dis* 2009;**68**:1482–5.
9. **Visser H.** *Diagnosis and prognosis in early arthritis.* Leiden: Leiden University Medical Center, 2003.
10. **Harrell FE,** Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361–87.
11. **Altman DG,** Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.
12. **Justice AC,** Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515–24.
13. **Steyerberg EW.** *Clinical prediction models: a practical approach to development, validation and updating.* New York: Springer Science, Business Media, 2009.
14. **Geuskens GA,** Hazes JM, Barendregt PJ, *et al.* Work and sick leave among patients with early inflammatory joint conditions. *Arthritis Rheum* 2008;**59**:1458–66.
15. **Steyerberg EW,** Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–38.
16. **Claessen SJ,** Hazes JM, Huisman MA, *et al.* 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.
17. **Hunink MG,** Glasziou P, Siegel JE, *et al.* *Decision making in health and medicine: integrating evidence and values.* Cambridge, UK: Cambridge University Press, 2001.