Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

METHODOLOGY

# Methodological issues in the design of a rheumatoid arthritis activity score and its cut-offs

This article was published in the following Dove Press journal:
Clinical Epidemiology
10 July 2014
Number of times this article has been viewed

Olivier Collignon

Centre de Recherche Public de la Santé (CRP-Santé), Competences Centre for Methodology and Statistics (CCMS), Strassen, Luxembourg

**Abstract:** Activity of rheumatoid arthritis (RA) can be evaluated using several scoring scales based on clinical features. The most widely used one is the Disease Activity Score involving 28 joint counts (DAS28) for which cut-offs were proposed to help physicians classify patients. However, inaccurate scoring can lead to inappropriate medical decisions. In this article some methodological issues in the design of such a score and its cut-offs are highlighted in order to further propose a strategy to overcome them. As long as the issues reviewed in this article are not addressed, results of studies based on standard disease activity scores such as DAS28 should be considered with caution.

**Keywords:** DAS28, disease activity score, penalized logistic regression, clinical prediction, modeling

Video abstract

Point your SmartPhone at the code above. If you have a QR code reader the video abstract will appear. Or use: http://dvpr.es/1m25ZAa

## Introduction

Rheumatoid arthritis (RA) is a systemic disease which occurs in about 1% of the world population and triggers joint inflammations that may worsen patients' quality of life. Nowadays, efficient disease-modifying antirheumatic drugs (DMARDS), such as Methotrexate, as well as targeted immunomodulating agents, are available to relieve patients.[1–4] In order to define treatment strategy and to evaluate response to therapy, disease activity may be measured via several scoring schemes, including the Disease Activity Score, involving 28 joint counts (DAS28), the Simplified Disease Activity Score (SDAI), and the Clinical Disease Activity Score (CDAI),[5,6] among others. All these measures consist of a weighted sum of bioclinical features, or a transform (eg, logarithm) of these, such as the number of tender joints in the hands, the number of swollen joints in the hands, the erythrocyte sedimentation rate (ESR), the C-reactive protein (CRP) concentration, the patient global assessment, and the physician global assessment. Several authors proposed cut-offs for these scores to help physicians classify patients in a particular disease activity state, ranging from remission to high activity.[7,8]

Among all disease activity scores, DAS28 is the most widely used. In clinical practice, it may be used as a monitoring tool to define treatment strategy and to further adapt it during patients' follow up.[9,10] For example, Van der Cruyssen et al[11] investigated the potential of DAS28 to help decide on a dose increase of infliximab to improve response to treatment in RA patients. On the other hand, DAS28 can be used in clinical trials to evaluate disease improvement from baseline using the European League Against Rheumatism (EULAR) response criteria.[12–14]

A lot of detail on RA scores can be found in two very helpful reviews from Anderson et al published in 2011 and 2012.[5,6] Some of them have been compared several times.[15,16]

Correspondence: Olivier Collignon
Centre de Recherche Public de la Santé
1A rue Thomas Edison,
L-1445 Strassen, Luxembourg
Tel +352 26 970 775
Fax +352 26 970 719
Email olivier.collignon@crp-sante.lu

An interesting introduction to the use of DAS28 can also be found online at http://www.das-score.nl.

However, one should be very careful about the design of such RA activity scores and cut-offs. Indeed, inaccurate evaluation of the RA activity using these scores may lead to inappropriate treatment administration or unreliable conclusions in clinical trials. For example, a DAS28 score lower than 3.6 is often considered as evidence of low disease activity and thus a target to reach for the physicians.[17] However, if the initial design of DAS28 or corresponding cut-off does not allow accurate evaluation of RA activity, such a patient's disease could actually be classified as being in the moderate or high activity phase and would thus necessitate initiation of treatment. In clinical trials, change in DAS28 from baseline could then be an inappropriate measure of drug efficacy.

In the current rheumatology literature, no article seems to have pointed out incorrect design of such scores as potential sources of misevaluation of RA activity yet. For this reason, methodological problems related to the development of RA activity scores and cut-offs are emphasized in this article. This study is devoted especially to DAS28, since it is the most widely used scoring system in clinical practice and research. The seminal papers of Prevoo et al in 1995[18] and Aletaha et al in 2005,[7] where DAS28 and its cut-offs for disease activity were constructed, are therefore discussed. A strategy to address these issues is then proposed in order to further improve RA activity scoring and thus patient care.

## Review of methodological issues

### Definition and evaluation of rheumatoid arthritis activity

Before developing a score, clear guidelines designed by expert physicians are needed to evaluate disease activity in order to limit inter-physician variability.[19] Such gold standards can be related, for example, to the level of physical impairment, the type of treatment needed, etc. However, it seems that no gold standard has yet been reached for describing disease activity in RA. For example, no consensus on the definition of remission has yet been reached. Remission can currently be confirmed when the DAS28 score is lower than 2.4, or confirmed independently using the 2010 American College of Rheumatology (ACR)/EULAR criteria.[7,17,20] Although both rules share common items, discrepancies exist and the need for guidelines has been pointed out.[21] Shaver et al already reported some inconsistencies using published cut-offs for remission of DAS28 and CDAI and therefore recommend cautious use of these with patients.[22]

The lack of guidelines was also illustrated in the study of Aletaha et al in 2005.[7] In this article, 35 experts had to judge the disease activity state of 32 RA patients. No reference explaining how patients were rated by the experts was given, although objective criteria had been clearly established when setting up DAS28 earlier in Prevoo et al's paper.[18] Only two of these were unanimously classified in the same disease activity category by every expert: those of lowest and highest disease activity. Over the whole sample, the mean percentage of judges classifying a given patient into a group other than the majority reached 28.42%. Even if the proposed statistical analysis tried to smooth the inter-expert variability by averaging the expert specific cut-offs, it seems then somewhat illusory to hope that precise cut-offs will help to classify patients when experts in the field experience some difficulty reconciling their judgments.

Besides this, a scoring scale has to replace a gold standard that is difficult or expensive to measure directly. It has to rely on other features than the ones used to define the guidelines and to offer a comparable efficiency.[19] For example, if the number of tender joints was used by the 35 experts as the reference test to assess disease activity, then including it in a score such as DAS28 becomes redundant.

### Sampling from database

Inclusion criteria defining target patients should be clearly defined in order to build a database. For example, in the study of Aletaha et al,[7] it was not specified if patients met the inclusion criteria demanded when developing DAS28 in the original study of Prevoo et al.[18] Although the ACR criteria were respected, it was not specified if patients had not been treated previously with DMARDS and if the disease duration did not exceed one year, as required in Prevoo et al's article.[18] This leads to a lack of comparability between studies.

Moreover, DAS28 was initially designed with 227 RA patients sampled from a longitudinal, hospital-based database to distinguish between high and low disease activity phases using Canonical Discriminant Analysis (CDA).[18] In this study, patients went through several disease activity periods, among which only two were randomly selected. As a result, a given patient could have contributed to both high and low disease activity groups. This leads to a violation of the assumption of independence in the statistical analysis, which in turn leads to erroneous coefficients estimates of the features in CDA and thus to an incorrect score. This constraint could have been accounted for using mixed models.[23] Furthermore, the longitudinal aspect of the disease activity phases could

also be very informative in predicting a novel activity phase. One may then wonder why the complete database was not used, and why only two phases per patient were kept in the analysis.

## Design of the score

The scoring objective (diagnosis, prognosis, choice of treatment, etc) should help the practitioner define which variables to integrate in the model. For example, some authors evocated the possibility of adding extra variables as imaging to DAS28 in order to predict remission.[24] Others raised the issue that DAS28 does not take into account age and sex,[25] nor the number of swollen and tender joints in the feet.[26]

Furthermore, sample size restrains the number of predictors to be used. Indeed, according to Whitehead,[27] when regressing an ordinal outcome with $q$ categories with a dataset of size $n$, the maximal number of predictors should be lower than $m/10$ to avoid overfitting, where

$$m = n - \frac{1}{n^2}\Sigma_{k=1}^{q} n_k^3 \quad (1)$$

($n_1$, …, $n_q$ being the frequencies of each category). For example, using Aletaha's sample[7] (remission: $n_1$=6, low activity: $n_2$=13, moderate activity: $n_3$=9, high activity: $n_4$=4), no more than two to three predictors should have been included to build the score.

## Definition of disease activity categories and design of the cut-offs

Predictive ability, ie, the propensity of the score to recover the disease activity state of patients, is the primary objective of defining classification rules. For example, three methods were proposed by Aletaha et al in 2005[7] to define the cut-offs of DAS28 and SDAI and to classify patients in "remission" versus "low to high activity", "remission to low activity" versus "moderate to high activity", and "remission to moderate activity" versus "high activity". In the first method, upper quartiles were used as optimal cut-offs, although they does not maximize any classification performance. The second technique relied on maximizing the $\kappa$ statistic, which is more a measure of agreement between two rankings than a measure of performance. Conversely, the third approach, relying on ROC (receiver operating characteristic) curves, did not suffer from these drawbacks. Besides, DAS28 was originally developed as a discriminative tool for a two-level disease activity state, which was precisely defined by the frequency of DMARDS administrations. Then, disease activity was defined as an ordinal outcome with four levels: remission; low activity; moderate activity; and high activity. Defining four states and cut-offs from a score designed primarily as a two-level disease activity measurement[18] seems suboptimal. New scores and cut-offs should be built to have the same number of categories as the gold standard they are replacing.

## Validation

Developing complex scores and cut-offs, especially when the features are numerous and the sample size is low, can lead to overfitting problems. This means that model performances may decrease when applied to other datasets. In Aletaha's study in 2005,[7] no actual statistical validation step had been proposed to address this issue. The cut-offs were validated by applying them to other datasets and by showing a significant increase of surrogates of disease activity (like the Health Assessment Questionnaire [HAQ] functional index, which also combines information on damage and comorbidity) across the categories defined by those cut-offs. Proper statistical validation procedures are needed to avoid developing scores with over-optimistic predictive ability.

# Possible guidelines to define a disease activity score and cut-offs

In this section a relevant strategy is proposed to define a tool to be used in assessing disease activity in RA patients, and to address previous issues.

## Patients

Patient inclusion criteria should be clearly defined (eg, ACR criteria, DMARDS administration, disease duration, etc). This would clarify which patients are on target and ensure comparability between studies.

## Disease activity

A gold standard to assess disease activity should be defined by expert physicians. They need to define the number of disease activity groups in which patients can be classified and how to do it. Guidelines were recently published about panel diagnosis.[28]

## Surrogate variables

If this gold standard is considered too complex or too expensive to manage, physicians should review possible surrogate variables to replace the ones used in the gold standard. In that case, constructing a disease activity score using these variables is relevant.

## Study design

For example, cross-sectional studies may be considered for diagnosis purposes, whereas cohort studies are preferred when prognosis is the aim of the score.[19]

## Statistical analysis

A complete approach of clinical prediction models can be found in Steyerberg's 2009 publication.[19]

### Estimation of disease activity state probabilities

Any classification procedure can be used to predict disease activity states using patient's features $x_1, \ldots, x_p$, such as forward continuation ratio ordinal logistic regression.[29] The probabilities of disease activity states $DA$ can be modeled as follows:

$$P(DA = k \mid DA \geq k, x_1, \ldots, x_p) = \frac{e^{\alpha_k + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\alpha_k + \beta_1 x_1 + \cdots + \beta_p x_p}} \quad (2)$$

where $\alpha_k$ represents the baseline disease activity state of category $k$, varying from 1 to $q$. The parameters $\beta_1, \ldots, \beta_p$ are the coefficients of the features and measure their contribution to disease activity. Parameters are estimated by maximizing the likelihood of the model. Forward continuation ratio ordinal logistic regression is often seen as a discrete Cox survival model.

### Variable selection

The most relevant predictors should be chosen using a variable selection scheme to avoid over-estimating their effects on disease activity state. Measuring only a limited number of predictors will also make the model more robust, easier to use, and cheaper.

In order to remove irrelevant variables from the predictive model, the penalization technique can be performed using the $L^1$ criterion. This widely-used method has several advantages over stepwise selection methods.[29] Only the relevant features receive a non-zero coefficient $\beta_i$ in the logistic regression formula and are further integrated in the score. To do so the following quantity has to be maximized:

$$L(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p) - \lambda \sum_{i=1}^{p} |\beta_i| \quad (3)$$

where $L(\alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)$ is the likelihood of the logistic regression and $\lambda$ is a parameter that controls the amount of penalization. The optimal $\lambda$ value is searched within a prespecified grid to optimize the Akaike information criterion (AIC) by crossvalidation.[29] As an example, Hirata et al used penalization to define a disease activity score using serum biomarkers.[30]

### Development of a score and associated cut-offs

The term $S = \beta_1 x_1, + \cdots + \beta_p x_p$ in logistic regression can be used as a score itself, although computation of the probabilities of the RA activity categories allows classifying patients directly. Note that some variables might not appear in the score if they have been eliminated during the variable selection step.

### Development of associated cut-offs

If for example two categories "low activity" and "high activity" are desired ($q$=2), the cut-off can be selected using criteria based on sensitivity and specificity.[19] However, more categories can be defined. For example, if four categories are desired, cut-offs can be defined by analyzing every possible triplet of score values ($c_1$, $c_2$, $c_3$), such as the disease activity state is predicted as "remission" if $S \leq c_1$, "low" if $c_1 < S \leq c_2$, "moderate" if $c_2 < S \leq c_3$ and "high" if $S > c_3$. The set of triplets has to optimize a performance measure like the correct classification rate, as defined in the following paragraph.

### Predictive ability

The performance of the classification technique may be evaluated with the correct classification rate, which is the percentage of patients whose disease activity state is correctly predicted by the logistic regression formula or using the cut-offs. The $C$ classification index,[29] which is an equivalent of the area under the ROC curve, is another widely used discrimination index.

### Model validation

External validity has to be assessed by collecting a blinded independent dataset meeting the criteria enounced in 1) of "Possible guidelines to define a disease activity score and cut-offs". This test set is classified according to the predictive model built previously. It is then unblinded to evaluate the performance of the model. If new data are unavailable, internal validation techniques such as bootstrap may be used to correct original results for optimism.[29] Bootstrap consists in drawing about 100 to 150 random samples of the same size from the actual dataset, with replacement. This means that these samples have the same number of observations that the original dataset had, and one observation can be selected several times in the same bootstrap sample. All modeling steps

are processed again on each bootstrap sample to compute correct classification rates and $C$ indexes. The classification techniques are then applied to the original dataset as if it were an independent test set. Optimism is calculated as the difference between the classification indexes obtained with the bootstrap sample and with the test sample. Optimism is finally averaged across all bootstrap samples and subsequently subtracted to the original classification indexes computed with the actual dataset. To sum up, if we denote by $Q$ any performance measure of the model built on the original data, $b$ the number of bootstrap samples, $Q_{boot,i}$ the performance obtained when re-regenerating the model with the $i$-th bootstrap sample, $Q_{test,i}$ the performance obtained when applying it to the original data, the optimism corrected performance $Q_{corr}$ can be computed using the following formula:

$$Q_{corr} = Q - \frac{1}{b}\sum_{i=1}^{b}(Q_{boot,i} - Q_{test,i}) \quad (4)$$

## Conclusion

In this article, some methodological issues in developing an RA activity score and its cut-offs are reviewed and addressed. Particular attention is devoted to DAS28, although most of the comments could be applied to SDAI and CDAI since they are direct modifications of DAS28. Despite its limits, DAS28 is widely used in clinical trials and for treatment monitoring. However, developing a new score following the guidelines proposed in this article could offer an alternative tool to accurately measure RA activity and could thus improve patients' health care. Moreover, it is now very important to define a gold standard to evaluate RA activity, to collect reliable data, and to apply a relevant methodology in order to develop a valid bioclinical score to assess RA disease activity states. Indeed, inappropriate medical decisions such as treatment administration could be the result of an inaccurate score. Meanwhile, results of studies based on classic disease activity scores should be considered with caution.

## Acknowledgments

The author would like to thank Professor Stephen Senn and Dr Daniel Witte from CRP Santé, and Dr Sophie Norman from LIMIDRA, for their fruitful review of the manuscript. The comments of the reviewers were also very useful and helped to improve the quality of the paper.

## Disclosure

The author has no conflict of interest to disclose.

## References

1. Gartlehner G, Hansen RA, Jonas BL, Thieda P, Lohr KN. The comparative efficacy and safety of biologics for the treatment of rheumatoid arthritis: a systematic review and metaanalysis. *J Rheumatol*. 2006;33(12):2398–2408.
2. Lee SJ, Chang H, Yazici Y, Greenberg JD, Kremer JM, Kavanaugh A. Utilization trends of tumor necrosis factor inhibitors among patients with rheumatoid arthritis in a United States observational cohort study. *J Rheumatol*. 2009;36(8):1611–1617.
3. Lipsky PE, van der Heijde DM, St Clair EW, et al; Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Infliximab and methotrexate in the treatment of rheumatoid arthritis. Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. *N Engl J Med*. 2000;343(22):1594–1602.
4. Smolen JS, Aletaha D, Koeller M, Weisman MH, Emery P. New therapies for treatment of rheumatoid arthritis. *Lancet*. 2007;370(9602):1861–1874.
5. Anderson J, Caplan L, Yazdany J, et al. Rheumatoid arthritis disease activity measures: American College of Rheumatology recommendations for use in clinical practice. *Arthritis Care Res (Hoboken)*. 2012;64(5):640–647.
6. Anderson JK, Zimmerman L, Caplan L, Michaud K. Measures of rheumatoid arthritis disease activity: Patient (PtGA) and Provider (PrGA) Global Assessment of Disease Activity, Disease Activity Score (DAS) and Disease Activity Score with 28-Joint Counts (DAS28), Simplified Disease Activity Index (SDAI), Clinical Disease Activity Index (CDAI), Patient Activity Score (PAS) and Patient Activity Score-II (PASII), Routine Assessment of Patient Index Data (RAPID), Rheumatoid Arthritis Disease Activity Index (RADAI) and Rheumatoid Arthritis Disease Activity Index-5 (RADAI-5), Chronic Arthritis Systemic Index (CASI), Patient-Based Disease Activity Score With ESR (PDAS1) and Patient-Based Disease Activity Score without ESR (PDAS2), and Mean Overall Index for Rheumatoid Arthritis (MOI-RA). *Arthritis Care Res (Hoboken)*. 2011;63 Suppl 11:S14–S36.
7. Aletaha D, Ward MM, Machold KP, Nell VP, Stamm T, Smolen JS. Remission and active disease in rheumatoid arthritis: defining criteria for disease activity states. *Arthritis Rheum*. Sep 2005;52(9):2625–2636.
8. Aletaha D, Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clin Exp Rheumatol*. 2005;23(5 Suppl 39):S100–S108.
9. van Riel PL, Schumacher HR Jr. How does one assess early rheumatoid arthritis in daily clinical practice? *Best Pract Res Clin Rheumatol*. 2001;15(1):67–76.
10. van Riel PL, Fransen J. DAS28: a useful instrument to monitor infliximab treatment in patients with rheumatoid arthritis. *Arthritis Res Ther*. 2005;7(5):189–190.
11. Vander Cruyssen B, Van Looy S, Wyns B, et al. DAS28 best reflects the physician's clinical judgment of response to infliximab therapy in rheumatoid arthritis patients: validation of the DAS28 score in patients under infliximab treatment. *Arthritis Res Ther*. 2005;7(5):R1063–R1071.
12. Fransen J, van Riel PL. The Disease Activity Score and the EULAR response criteria. *Clin Exp Rheumatol*. 2005;23(5 Suppl 39):S93–S99.
13. van Riel PL. Provisional guidelines for measuring disease activity in clinical trials on rheumatoid arthritis. *Br J Rheumatol*. 1992;31(12):793–794.
14. Behrens F, Tony HP, Alten R, et al. Development and validation of a new disease activity score in 28 joints-based treatment response criterion for rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2013;65(10):1608–1616.
15. Gaujoux-Viala C, Mouterde G, Baillet A, et al. Evaluating disease activity in rheumatoid arthritis: which composite index is best? A systematic literature analysis of studies comparing the psychometric properties of the DAS, DAS28, SDAI and CDAI. *Joint Bone Spine*. 2012;79(2):149–155.

16. Hamdi W, Néji O, Ghannouchi MM, Kaffel D, Kchir MM. Comparative study of indices of activity evaluation in rheumatoid arthritis. *Ann Phys Rehab Med*. 2011;54(7):421–428.
17. Fransen J, Moens HB, Speyer I, van Riel PL. Effectiveness of systematic monitoring of rheumatoid arthritis disease activity in daily practice: a multicentre, cluster randomised controlled trial. *Ann Rheum Dis*. 2005;64(9):1294–1298.
18. Prevoo ML, van't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum*. 1995;38(1):44–48.
19. Steyerberg EW. *Clinical Prediction Models: A Practical Approach To Development, Validation, And Updating*. New York: Springer; 2009.
20. Felson DT, Smolen JS, Wells G, et al; American College of Rheumatology; European League Against Rheumatism. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. *Arthritis Rheum*. 2011;63(3):573–586.
21. Sheehy C, Evans V, Hasthorpe H, Mukhtyar C. Revising DAS28 scores for remission in rheumatoid arthritis. *Clin Rheumatol*. 2014;33(2):269–272.
22. Shaver TS, Anderson JD, Weidensaul DN, et al. The problem of rheumatoid arthritis disease activity and remission in clinical practice. *J Rrheumatol*. 2008;35(6):1015–1022.
23. Brown H, Prescott R. *Applied Mixed Models in Medicine*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2006.
24. Haavardsholm EA, Lie E, Lillegraven S. Should modern imaging be part of remission criteria in rheumatoid arthritis? *Best Pract Res Clin Rheumatol*. 2012;26(6):767–785.
25. Radovits B, Fransen J, van Riel PL, Laan RF. Influence of age and gender on the 28-joint Disease Activity Score (DAS28) in rheumatoid arthritis. *Ann Rheum Dis*. 2008;67(8):1127–1131.
26. Bakker MF, Jacobs JW, Kruize AA, et al. Misclassification of disease activity when assessing individual patients with early rheumatoid arthritis using disease activity indices that do not include joints of feet. *Ann Rheum Dis*. 2012;71(6):830–835.
27. Whitehead J. Sample size calculations for ordered categorical data. *Stat Med.* 1993;12(24):2257–2271.
28. Bertens LC, Broekhuizen BD, Naaktgeboren CA, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med.* 2013;10(10):e1001531.
29. Harrell FE. *Regression Modeling Strategies: With Applications To Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer; 2001.
30. Hirata S, Dirven L, Shen Y, et al. A multi-biomarker score measures rheumatoid arthritis disease activity in the BeSt study. *Rheumatology (Oxford)*. 2013;52(7):1202–1207.