Assessment of the utility of patient outcome measures in rheumatoid arthritis

Charles Thurston, Medical Student, Lancaster Medical School, Lancaster University
Dr Marwan Bukhari PhD, FRCP, Consultant Rheumatologist, RLI

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory condition that leads to joint destruction and ultimately disability. This is often seen in combination with pain, fatigue, and extra-articular manifestations. It affects 0.5-1% of the population,[1] and therefore represents a challenge for NHS services and a large case load for rheumatologists. In recent years there has been a revolution in new drug therapies with the development of biologics. These drugs are effective at treating RA with the aim of total disease remission, but are more expensive than the traditional disease-modifying anti-rheumatic drugs (DMARDs). The cost of an annual course of adalimumab (a tumour necrosis factor-α monoclonal antibody)[3] is £9295.00[4] compared to the cost of methotrexate (DMARD) which is £10.00 (prices vary, also dependent on dose).[5] The National Institute of Health and Care Excellence (NICE) has developed guidance to allow for more efficient prescribing and ensure proper use of the NHS budget, but during this process NICE used the Health Assessment Questionnaire (HAQ) to evaluate the effectiveness of the treatment, rather than more traditional measures like the Disease Activity Score-28 (DAS-28).[4] This review will examine the literature behind NICE's choice to use the HAQ over the DAS-28, including the validity of the measures and their components, and the relationship between the two measures, with the ultimate goal of determining the appropriateness of this decision.

DAS-28

The DAS-28 score is made up of 4 parts: a swollen and tender joint count, the patient's erythrocyte sedimentation rate, and the patient's global assessment.[6] The joints counted in the score are the metacarpophalangeal and proximal interphalangeal joints of the hand (totalling 20), the wrists, elbows, shoulders, and knees. The scoring can be seen in Table 1. Each component of the score has advantages and disadvantages to its inclusion, and the role of each component will be further discussed.

Table 1: The scoring boundaries of the DAS-28.

Scoring range	Level of disease activity
>5.1	High
5.1 – 3.2	Moderate
3.2 – 2.6	Low
<2.6	Remission

Swollen joint counts (SJC)

Swollen joints are a key identifying symptom of an inflammatory arthritis, it forms part of the prerequisites of using the American College of Rheumatologists and European League against Rheumatism's (ACR/EULAR) 2010 classification of RA, which are the patient must have at least 1 swollen joint with definite clinical synovitis and that it cannot be explained by any other disease[7] (see Table 2).

Table 2: ACR/EULAR classification of rheumatoid arthritis. A score of ≥6 indicates rheumatoid arthritis.

	Score
A. Joint involvement	
1 large joint	0
2-10 large joints	1
1-3 small joints (with or without large joint involvement)	2
4-10 small joints	3
>10 joints (with at least 1 small joint)	5
B. Serology	
Negative rheumatoid factor (RF) or anti-CPA	0
Low-positive RF or anti-CPA	2
High-positive RF or anti-CPA	3
C. Acute-phase reactants	
Normal CRP or ESR	0
Abnormal CRP or ESR	1
D. Duration of symptoms	
<6 weeks	0
≥6 weeks	1

However, SJCs have a poor reproducibility and standardisation, as there is a large variance both within and between clinicians. Intra-observer reproducibility (using Cohen's kappa values) indicates that there is a large range in reproducibility, from good to poor (0.31-0.77). This remained the trend in inter-observer reproducibility (with 50% reference standard) kappa values ranged from 0.40-0.62. The DAS-28 was calculated using the score derived by the clinicians and compared to data obtained by ultrasound, and the mean variation due to a difference in SJC was 0.59, and the maximum variation seen was 0.92. These variations are large enough to have an effect on decisions to initiation or withhold treatment.[8] However, other studies have shown that training increases the reproducibility and standardisation of SJC,[9] which is helpful if undertaken on a departmental

wide basis, however this aspect of clinical examination is fundamental to the working life of a rheumatologist and may be better implemented during specialist training. It has been suggested that this could also be counteracted by having an assigned rheumatologist that the patient consistently sees.[10] However this remains unlikely due to the changeability of clinics, and could increase the difficulty of shared care between the rheumatologist and the GP.[10] It has also been observed that synovitis does not necessarily lead to swollen joints, 185 clinically synovitic joints were detected by clinicians from 644 painful joints from 80 patients. Ultrasound confirmed synovitis in 79% of these joints, but found synovitis in 33% of the remaining 459 non-clinical joints, meaning overall the ultrasound detected 64% more synovitis than clinical examination. This further highlights the inaccuracies in examination, and that simple joint counts are not reflective of the scope of disease.[11] However this study was carried out on patients with recent (<12 months) oligoarthritis (≤5 joints), they were not diagnosed with RA, but the study assumes that the basic inflammatory process is similar and that by extension the oligoarthritis will become RA. However, due to the small symptomatic window, there may not have been time for sub-clinical cases to develop into clinically detectable cases. This may link more to a lack of understanding of the natural history of oligoarthritis and RA rather than the clinician's inaccuracy.

Tender joint counts (TJC)

Tender joints are another key symptom of RA, and in the DAS-28 the TJC carries twice the weight of the SJC.[12] TJCs have been shown to be one of the components most sensitive to change. The TJC showed an adjusted t-statistic of 3.58, compared to those of the ESR at 3.30, SJC at 2.18, and the patient global assessment 1.89.[13] This highlights that the DAS-28 is a dynamic score, and as all components are sensitive to change will be reflective of changing disease activity. However this was a small review, only encompassing 3 trials. The DAS-28 components were compared to other factors like swelling and tenderness scores, platelet count, pain scores, and PIP joint circumference, but not each parameter was reported in all the trials meaning there is a lack of consistency in the result. TJCs are not as good at predicting structural joint damage as swelling or confirmed synovitis on ultrasound, however when the two are measured simultaneously there is a high predictive value of disease progression.[14] This highlights that the composite nature of the DAS-28 increases its face validity. However TJCs can easily be affected by RA co-morbidities like fibromyalgia (FM); estimates are that 13-17% of the RA population would also fulfil the criteria for fibromyalgia.[15] In a comparison between RA and RA co-morbid with FM, the co-morbid group scored 1.33 more on the DAS-28, and the main factors were TJC and ESR. When FM is not diagnosed, this will lead to a misclassification of active RA disease by the DAS-28 score and may lead to unnecessary changes in therapy that will have no therapeutic effect on the FM.[15]

Erythrocyte sedimentation rate (ESR)

ESR contributes 15% of the overall DAS-28 score[16] and is the laboratory measure of inflammation. ESR is linked to the blood concentration of fibrinogen, an acute phase reactant that has a half-life of 4 days.[17] The ESR is also an indirect measure of the presence of immunoglobulins17, of which rheumatoid factor is predominantly an IgM, and therefore has a link to the pathophysiology of RA. ESR is regarded as a longer term measure of inflammation because of the relatively long half-lives of fibrinogen and immunoglobulins,[17] making it useful in RA investigation due to the chronic nature of the inflammation. ESR is consistently raised in RA patients, especially with an associated rise in RF and anti-CCP, it was shown that the average ESR of an RA patient was 56.6mm/h compared to control of 11.2mm/h (18). However ESR has a weak correlation with SJC (r=0.315, p value = 0.686) and TJC (r=0.1, p value = 0.173), and overall has a weak correlation with function (r=0.315, p value 0.0001) and pain (r=0.212, p value = 0.0034).[19] This leads to a scenario where there is little evidence of synovitic joints but a high DAS-28 because of a high ESR that has no correlation to the clinical examination, and by extension, a low DAS-28 due to underestimated ESR but a large number of clinically significant joints.[16] The ESR is also vulnerable as it is not a specific measure of inflammation associated with RA, it can be affected by other co-morbidities. Reduced haemocrit increases ESR,[20] which is a known extra-articular manifestation.[21] However, it could be argued that this further represents the activity of the disease, but this is not clearly defined within the DAS-28 literature. However in an inflammatory condition such as RA it is important that we measure the level of inflammation to know the extent of its activity and to what extent we can reduce the inflammation. ESR shows much greater sensitivity to change than other inflammatory markers. There exists a variant of DAS-28 where the C-reactive protein (CRP) is measured instead of the ESR. However ESR is 15-20% more sensitive to change than CRP, with ESR having a greater effect size in 58% of trials compared to CRP.[22] This indicates that the pathophysiology of RA is linked with ESR, and therefore it is a good measure of disease activity.

Patient global assessment (PGA)

At first the role of the PGA may seem similar to the HAQ, they both are looking at the patient's experience of RA in a more holistic way. However, the PGA is a much less structured way of obtaining this information. This is exemplified by the different ways in which the PGA can be assessed, there is no standardised question used to obtain the score. One common way is to ask specifically about the effect of the arthritis, which focuses more directly on the level of disease activity. Another commonly used question is asking about their health overall, which alludes more to their global health. This difference allows for different patient and clinician interpretations of their disease activity, some patients may factor in the effect of other co-morbidities, whilst the others may focus purely on the level of joint pain they are experiencing, as this is the only symptom they associated with their disease. Moreover, patients with longer term disease may score themselves as having lower disease activity, as they have come to a stage of acceptance with their symptoms, or have reached a period of stability after a longer period of flares.[23] However,

studies show the PGA has the same level of efficiency as other traditional disease measures. In a trial comparing abatacept and placebo, the PGA showed a treatment different of 1.04% compared to TJC.[24] This number does not indicate a large difference, but it highlights that it has similar results to already ratified disease measures. However in the trial there is standardisation of how PGA was measured, so does not fully address the concerns about reproducibility.

HEALTH ASSESSMENT QUESTIONNAIRE (HAQ)

The HAQ was first developed in the 1980s by the rheumatology community in response to a growing need to base medical decisions on the patient's experience of their disease rather than purely biomedical measures.[25] It has been copyrighted since its development so that it remains unmodified, which strengthens its consistency of results and ensuring standardisation across testing.[25]

In a breakdown of the HAQ score, regression analysis showed that the pain score was the primary explanatory variable (41.4%) for the HAQ score, the Larsen score (a score given to x-rays on the level of joint destruction)[26] was 7.3%, and the Beck Depression Inventory was 5.5%.[27] Other factors measured such as age, sex, duration of disease, and serological features were not statistically significant; this leaves roughly half of the model unexplained by unmeasured factors. One possibility is that the rest is explained by motivation and psychological factors that are measured in the HAQ but were not independently analysed in the study,[27] which suggests that the HAQ is able to measure disease activity much more comprehensively and holistically than multiple measures combined. However, age has been consistently shown to be a negative predictive factor in HAQ score.[28] It seems unclear whether this is related to the natural progression of RA and the result of lifelong inflammation or the effect of co-morbidities. This highlights one of the difficulties in using the HAQ, as the questions do not actively discriminate between the disability associated with RA and that of other co-morbidities.

It has also been suggested that the HAQ lacks face validity as there is no inclusion of a joint examination or count, a measure that would bring the measure into focus on the "organ" that is affected by RA. This is reinforced by a series of recommendations made by an international expert committee on the goals of RA therapy, where the committee recommends that regular joint counts are the main tool of assessing disease activity; 93.4% of the committee made up of 60 members, 55 rheumatologists and 5 RA patients, agreed with this statement.[29] This is important as this suggests that the majority of the patients on the committee agreed with a more biomedical approach to the measure of their disease, contradicting the theory behind the development of the HAQ. The HAQ also faces cultural issues, it was developed in English largely for the American and Canadian market, and therefore some questions show cultural bias and make them unanswerable in other cultures.[25] For example, the questionnaire asks about the ability to open milk cartoons, which are very rarely used outside of North America. Secondly, the level of disability attached to the scoring in the HAQ is open to change across the cultures, as there are varying ideas in the appropriate level of function across the ages. The fact that this international committee has steered away from a solely patient-centred measure suggests that these cultural differences can be superseded by a measure like the DAS-28.

Correlation between DAS-28 and HAQ

Limited studies have directly compared the two as disease outcome measures, they are often both included and the difference individually reviewed, but they are not normally the main focus of the study. However one study does report their correlation, finding that after a 24 month period with DMARD or steroid treatment there is a correlation between the scores.[2] Correlation was

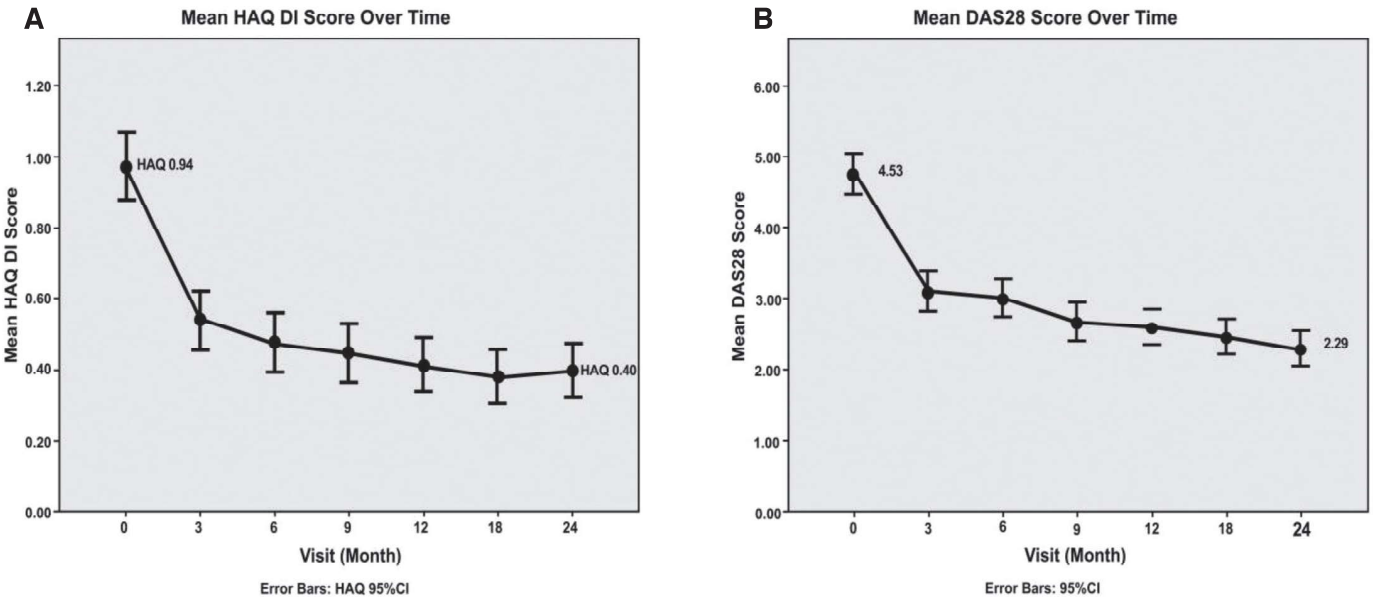

Figure 1: Graphs showing mean HAQ DI score over time (A) and mean DAS28 score over time (B).[2]

moderate at baseline (r=0.53), 18 months (r=0.57), and 24 months (r=0.59). At 6, 9, and 12 months the correlation was weaker, but still mainly moderate where r=0.41, 0.30, and 0.40, respectively (see figure 1).[2] Patients that were RF positive had consistently higher HAQ and DAS-28 scores and also showed a stronger correlation than RF negative (r=0.63 in RF positive and r=0.47 in RF negative).[2] These results do suggest a link between the two measures, however the study again is measured in early inflammatory arthritis, so therefore assumes that there is pathological overlap between the early process and established RA. This does seem likely that they are similar processes but it still remains an assumption.

Overall however, correlation does not always lead to causation, and it is not possible to say with certainty that because the HAQ and DAS-28 are correlated that they do have a relationship that means they can both reflect the same changes in disease activity. Bradford-Hill developed a list of criteria that can be used to determine the likelihood of the causation being related to causation rather than chance.[30] The first criterion is strength, the two values in this study are only moderately correlated, but this judgement is limited because it is the only study found that directly compares the two, this suggests that further research is necessary. Second is consistency, the study also highlights other studies that have attempted to correlate the two scores. However most of them use a variation of the DAS-28 defined earlier, either using CRP over ESR or using a different joint count. This also highlights a problem with the use of DAS-28; different formats may lead to differently defined results and clinical states. Third is specificity, the HAQ is used in multiple settings, as is not a measure designed for sole use in RA, therefore it seems unlikely that it would be as efficient at measuring disease activity as one that was developed solely for use in RA. However, it does measure many of the functional ways in which RA can affect a patient, and asks some questions that are commonly employed by rheumatologists during a history. Fourth is temporality, which is not necessary to consider in this scenario as both measures were taken concurrently and at the same time periods. Fifth is biological gradient, and having a moderate correlation coefficient does suggest that there is a dose-response relationship. Sixth is plausibility, although the DAS-28 uses different measures to the HAQ there is some overlap with their methods. When the HAQ asks "are you able to dress yourself, including tying shoelaces and doing buttons?" this can be seen as a reflection of the TJC and SJC, because as the counts rise then it will become more difficult to carry out dextrous activities. Therefore the causation does seem plausible as the two measures relate the interface between disease severity and disability. Seventh is coherence, it is known that RA is a disease that causes disability, and it is agreed that this disability is caused by the underlying inflammatory process associated with RA, therefore the association would be coherent. Eighth is experiment, and this study shows that the two do respond to treatment, so reducing the inflammation and disability associated with RA reduces the scores. Ninth is analogy, the HAQ has been validated in other inflammatory conditions like systemic lupus erythematosus, psoriatic arthritis, and systemic sclerosis since its development[31] meaning it is well placed to measure disease activity in another inflammatory conditions such as RA. Overall analysis using the Bradford-Hill criteria suggests that there is a possibility of a causative relationship between disease activity, the HAQ, and DAS-28. However it would need further research, particularly with a predetermined and standardised definition of DAS-28.

CONCLUSION

In conclusion, there are both strengths and weaknesses to NICE's decision to use HAQ over the DAS-28 in evaluating the use of biologics in RA therapy. The DAS-28 is a composite score that encompasses traditional biological measures of disease activity along with a patient derived one. Although there are issues with the reproducibility of aspects like the SJC, TJC, and the PGA, and that the ESR reflects more than inflammation associated with RA, overall the score is well established in the rheumatology community with significant effect. The HAQ may not have been specifically developed for use in RA, but it does come with a rising need to measure disease more holistically, and does ask questions that replicate the categories of the DAS-28. Although studies have shown that there is a possibility of a meaningful correlation and causation, further research is still needed. Overall the use of the HAQ does highlight a move to more patient-centred disease control, which is a positive in theory, but the literature suggests that the DAS-28 covers disease activity more comprehensively than the HAQ, particularly as the DAS-28 is specifically designed for RA. However, it seems short sighted to limit the conclusion to the use of just one tool, the use of both for the evaluation of biologics would provide a way to link the interface between disease activity and disability so commonly seen in RA patients.

Correspondence to:
c.thurston@lancaster.ac.uk

REFERENCES
(A full list is available on request)

1. Scott DL, Wolfe F, Huizinga TW. Rheumatoid arthritis. Lancet 2010;376(9746):1094-108.
2. Boyd TA, Bonner A, Thorne C, Boire G, Hitchon C, Haraoui BP, *et al.* The relationship between function and disease activity as measured by the HAQ and DAS28 varies over time and by rheumatoid factor status in early inflammatory arthritis (EIA). Results from the CATCH Cohort. The Open Rheumatology Journal 2013;7:58-63.
3. Gabay C, Hasler P, Kyburz D, So A, Villiger P, von Kempis J, *et al.* Biological agents in monotherapy for the treatment of rheumatoid arthritis. Swiss Medical Weekly 2014;144:w13950.
4. National Institute for Health and Care Excellence. Adalimumab, etanercept, infliximab, certolizumab pegol, golimumab, tocilizumab and abatacept for rheumatoid arthritis not previously treated with DMARDs or after conventional DMARDs only have failed. Technology appraisal guidance [TA375] 2016 Available from: https://www.nice.org.uk/guidance/ta375 (accessed 12.06.2020).
5. British Medical Association, Royal Pharmaceutical Society of Great Britain, National Institute for Health and Care Excellence. British National Formulary [electronic resource]. BNF.