A Review of Evidence on the Discriminant Validity of Outcome Measures in Rheumatoid Arthritis

DAVID T. FELSON and JENNIFER J. ANDERSON

ABSTRACT. We have assessed the discriminant validity of functional status measures and measures that are part of the core set for rheumatoid arthritis. Papers were identified by a systematic literature search in MEDLINE and hand searching of references. (J Rheumatol 2001;28:422–6)

Key Indexing Terms:
RHEUMATOID ARTHRITIS
DISCRIMINATION
MINIMAL CLINICALLY IMPORTANT DIFFERENCES
VALIDITY

Sensitivity to change is present when a measure captures the change that occurs in a patient or a group of patients. In rheumatoid arthritis (RA), the past 15 years has seen an explosion in our knowledge of sensitivity to change, validity, and reliability of different ways of measuring the status of patients with RA. Discriminant validity is a broader concept that includes sensitivity to change as well as the ability of a measure to detect differences between groups in the amount of change experienced. It is a critical type of validity because without it we cannot evaluate change over time; we cannot test treatments to see whether they cause change that other treatments don't, or evaluate whether and why certain patients are improving or deteriorating.

There are many elements to discriminant validity and, ideally, the set of measures we use to evaluate patients with RA over time should have proven validity in most, if not all, facets of discriminant validity. These facets include, but are not limited to, measuring change within a group followed over time, discriminating between one treatment that induces improvement versus another treatment that does not, and evaluating whether improvements reach a level that would be considered clinically important. Some of these elements of discriminant validity have been studied and others have not. It is the goal of this review paper to critically assess the current state of knowledge with respect to the discriminant validity of outcome measures used in RA. We will focus especially on 2 categories of outcome measures that have been most comprehensively tested for discriminant validity, functional status measures (often also called health status measures) and measures that are members of the RA core set.

From the Boston University Arthritis Center, Boston, Massachusetts, USA.
Supported by NIH Grant AR20613.
D.T. Felson, MD, MPH; J.J. Anderson, PhD.
Address reprint requests to Dr. D. Felson, Boston University School of Medicine, A203, 715 Albany Street, Boston, MA 02118.

MATERIALS AND METHODS

Papers reporting on the discriminant validity of outcome measures were collected from a MEDLINE and bibliographic search. The MEDLINE search used search terms shown in Table 1. This was supplemented by bibliographic references.

We focused on reviewing studies that evaluated longitudinal change in patients with RA and ways to measure that change. To evaluate which types of validation studies have been performed on outcome measures in RA that related to the discriminant validity of these measures, we used the taxonomy of the discrimination cube, altering it slightly for the purposes of the evaluation of RA outcome measures. Our modifications were as

Table 1. Search strategy for finding articles on rheumatoid arthritis outcome measures.

Terms	No. Retrieved
arthritis, rheumatoid/	4783
rheumat$ arthritis.tw.	5286
15 or 16	6310
improvement.tw.	25518
important.tw.	102830
18 or 19	126142
exp "outcome assessment (health care)"/	52575
treatment outcome/	45884
21 or 22	52575
20 and 23	8095
(mcid or mcids).tw.	6
minimal$ clinical$ important difference$.tw	9
(improvement adj2 criteria).tw.	67
(improvement adj2 definition$).tw.	18
or/24–28	8147
clinical judgement analys#s.tw.	2
judgement.tw.	703
clinical$ important change.tw.	9
decision making.tw.sh.	6364
assessment of change.tw.	31
disease activity.tw.	1948
change in activity.tw.	170
difference criteria.tw.	3
change criteria.tw.	8
sensitivity to change.tw.	139
outcome assessment.tw.	173
responsiveness improvement.tw.	1
or/29–41	17360

follows: (1) We eliminated the Differences Between row, which focuses on cross sectional comparisons, as we were interested in longitudinal ones. (2) We deleted the Minimum Potentially Detectable row, which deals solely with a delineation of the scales that are used. We also operationalized each of the remaining cells of the cube. The clinical trials in which changes in 2 groups were compared with respect to a particular outcome measure were classified as studies focusing on both differences between groups and changes within groups, as both are required to show significant differences between treatments. Changes within group cells were filled in when studies dealt only with a single group followed over time.

A second dimension of the cube consists in types of changes or differences that are reported in a study. We characterized the study as having reporting minimum change actually detectable beyond error when a significance level based on the error rate in the measurement of the measure was utilized. For example, if the group change in a particular outcome measure reached statistical significance, we characterized that significance as demonstrating that the outcome measure was valid to detect minimal change beyond error. Any study reporting change within or between groups was characterized as showing change that was actually "observed in the population" (means or standardized response means or effect sizes were all acceptable methods of demonstrating observed change). To characterize a study as having evaluated an outcome measure with respect to its validity in those estimated to improve, we required an independent measure of improvement, often a patient self-report measure of improvement, and change in the outcome measure must be compared in those improved versus those who were not improved.

Because we could not separate out studies that evaluated minimally important differences or changes from clearcut changes, we did not feel that we could accurately characterize studies that had evaluated change in those "observed to have changed" (one column from the far right) and therefore left this column as a question (see Figure 2, for example). Studies evaluating definite change were placed in the right hand column, and this was the criterion we used to evaluate whether studies had evaluated those estimated to have experienced important improvement. It was our opinion that current measures of response such as ACR/EULAR definitions focused on minimally *important* differences and therefore would fit into the right hand column. We created a subcategory of this column, depicted under a diagonal slash in this column, that evaluated whether studies had evaluated definitions of major improvement (see Figure 2, for example).

For the horizontal dimension of the cube, on the type of change reported, there is a nested hierarchy, such that a study that falls in a box further away from the origin usually also falls in boxes closer to the origin. (For example, if a study presented data on the percentage of patients who experienced important improvement in an RA trial and reported whether the difference in the number of patients was significantly different in the 2 treatment groups, the study would have been classified as including information on change beyond detectable (cell 1), on observed change (cell 2), and on important improvement (cell 4).

The last dimension in which we evaluated studies was whether they were studies of individual patients or groups of patients. Studies in which means or group tendencies were reported were characterized as studies of groups of patients. Studies of individual patients were difficult to characterize because some studies differentiated improved versus not-improved patients and then evaluated the mean response of an outcome measure within each group. We characterized these types of studies as studies of groups of patients. If, on the other hand, the study focused on the number of patients who had actually experienced improvement in the outcome measure of interest, we characterized this study as one of individual patients.

We also subcharacterized RA publications into 2 groups: (1) a group of studies that focused on the broad definitions of response (e.g., core set or index measure or evaluations of multiple outcome measures, which are elements of the core set), and (2) a group that studied functional status instruments and their validity.

RESULTS

Our search yielded over 300 articles. After excluding reviews, studies on related diseases, and reports that did not contain data on discriminant validity of RA measures (most of these were cross sectional studies in which measures were used to separate one subgroup of RA patients from another), we were left with 36 articles for review.

Studies of outcome measures in the core set. First, we turn to the subset of 19 studies whose focus was on validating core set measures over time (Figures 1 and 2). Our results showed a concentration of study validation types in which treatment groups in trials were compared or patients evaluated over time. These focused mostly on whether changes improved more than chance and whether one treatment group experienced more change than another. Recently, because of the availability of definitions of individual patients' response such as the ACR criteria and the EULAR response criteria, studies have increasingly dealt with changes in individuals over time. Few studies have evaluated whether particular outcome measures identify subjects who improved based on independent criteria, and no studies have done this in the context of trial data. In addition, if we assume that the ACR improvement criteria and the EULAR criteria define minimal clinically important differences (MCID) (we characterized these as estimated to have an important difference/change), there is a remarkable paucity of data on whether the core set measures or thresholds of response in the core set measures identify those with major improvement (something more than ACR improvement, for example).

Studies on functional status instruments. When we evaluated the 17 functional status instruments studies (Figures 3 and 4), we found the preponderance of studies focused on changes within groups of patients followed over time and characterized their functional status change as to whether it was significant beyond chance (thus satisfying our criteria for whether the study focused on minimum change actually detectable beyond error). Because most of these studies reported effect sizes or standardized response means, we also characterized them as having studied changes observed in the population. A lesser number of studies characterized those estimated to have improved. Almost all the studies that looked at the discriminant validity of functional status measures did so in groups of patients (Figure 4) rather than individual patients (Figure 3), with one study focusing on individual patients. A number of studies looked at functional status change in clinical trials, and therefore were characterized as looking at group change both between groups and within groups, although none of these focused on a threshold for individual improvement of functional status change. Further, no functional status validity studies evaluated or attempted to validate the notion of important improvement in functional status change.

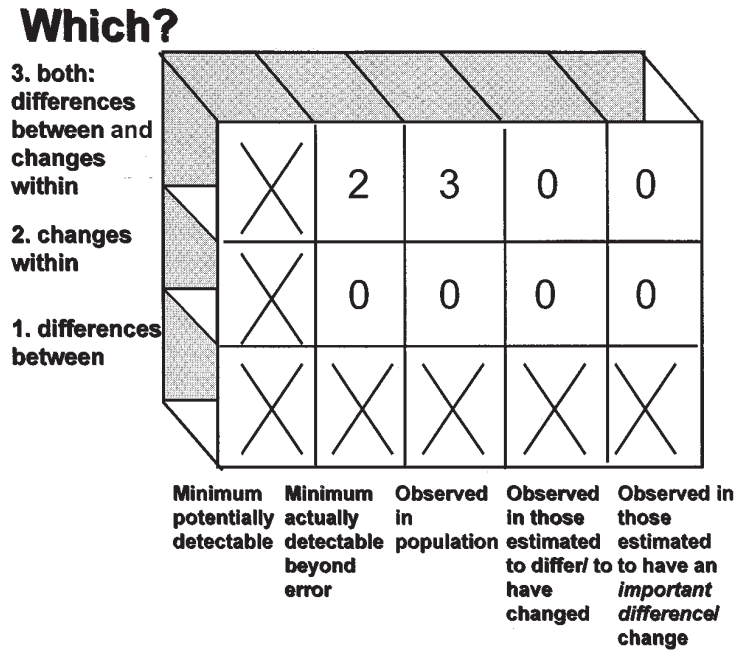

Figure 1. A number of studies providing data on the validity of RA core set measures — studies on individual patient changes/differences.

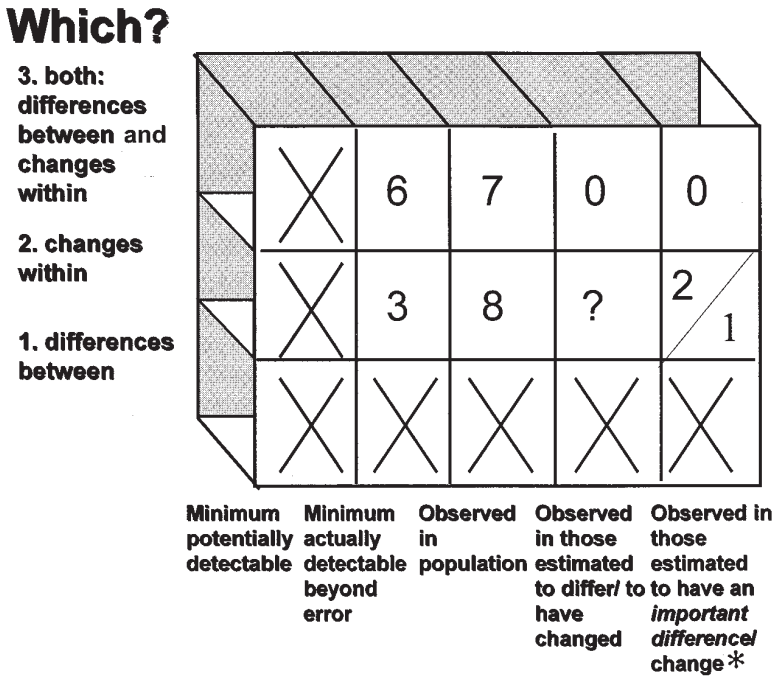

Figure 2. A number of studies providing data on the validity of RA core set measures — studies on groups of patients. *In second row, numbers above the diagonal correspond to minimal important improvement (e.g., ACR improvement) and numbers below the diagonal to major improvement.

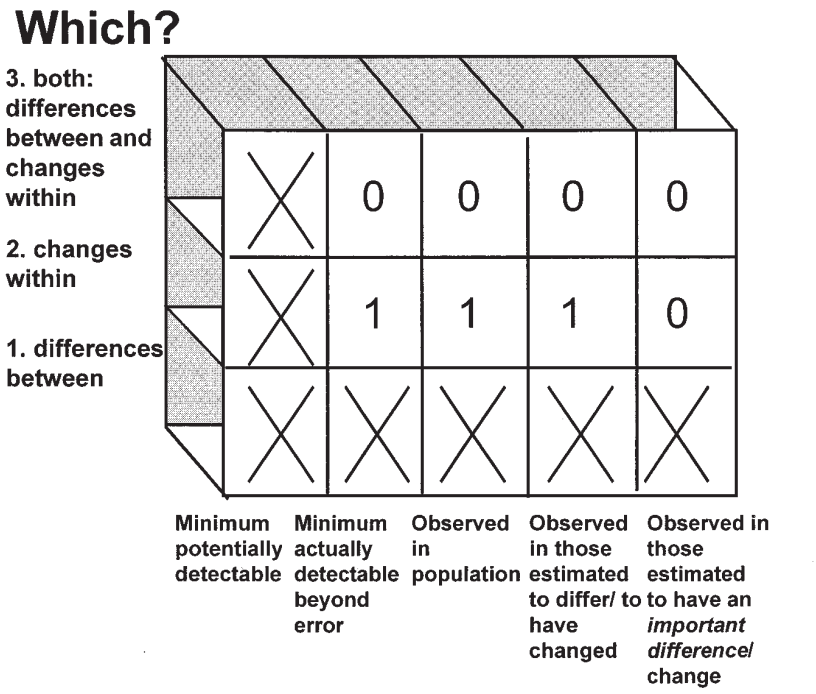

Figure 3. A number of studies providing data on the validity of functional status measures in RA — studies of individual patient changes/differences.

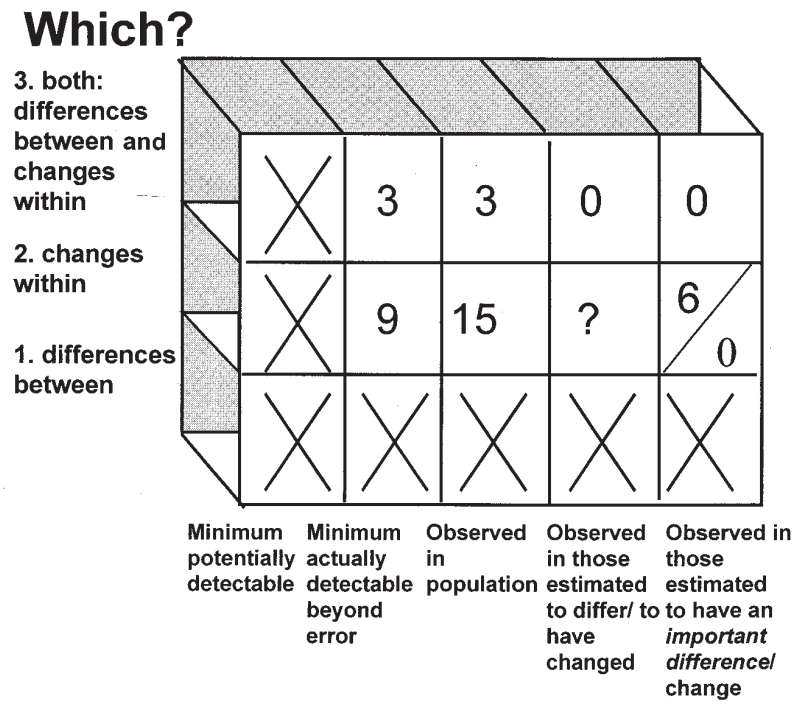

Figure 4. A number of studies providing data on the validity of functional status measures in RA — studies on groups of patients.

DISCUSSION

In summary, our review of RA literature suggests that most outcome measures used in practice and clinical trials settings have been validated with an eye toward demonstrating statistically significant changes over time and demonstrating significant differences in change between treatment groups in a trial. Until recently, there has been little, if any, attention to defining thresholds for improvement, and despite ACR and EULAR criteria, which focus on the number of patients improved or having responded to treatment, there is still a dearth of studies that have validated their measures with an independent evaluation of patient improvement. Further, there is almost no literature on major improvement and its definitions. Studies focusing on the response of individuals as opposed to group changes (such as means or medians) are also lacking, notwithstanding the recent focus on defining response in individuals.

There are a number of important caveats that should be considered in evaluating the results presented here. First, we may have had a unique interpretation of how certain studies fit within cells of the discriminant validity cube. Other reviewers might have classified such studies differently and therefore left certain cells filled while others were vacant. Also, it is not clear that all cells in this cube need to be filled with multiple studies, nor that a cell currently filled with 2 or 3 studies ought not to be a focus of yet other studies. Finally, we focused on a variety of subcategories of RA outcome measures, but combined data on studies of a group of outcome measures when one or more of the outcome measures in that group might not have been the topic of any of the studies. For example, one cell within the cube could be filled with 4 or 5 studies on core set measures, yet it is possible no studies considered such a measure as physician global assessment (although in fact this was not the case). Further, our summary of the literature on functional status measures may be too crude, missing the fact that one important functional status measure that might be widely used had not undergone the considerable validation that others had. Lastly, our literature review might have been incomplete, despite a comprehensive search. It is likely that articles that contain validations of outcome measures in RA have another primary focus (e.g., clinical trials in RA) and therefore were not indexed in such a way as to be captured by our search.

In summary, this review of the validity of outcome in RA suggests that most published studies have focused on observable change, especially in groups of patients. There has been a relative lack of studies that treat individuals and that study important or major improvement.