STABILITY OF HEALTH STATUS MEASUREMENT IN RHEUMATOID ARTHRITIS

R. GEENEN,*† J. W. G. JACOBS,* G. GODAERT;‡ F. W. KRAAIMAAT,‡ M. R. BRONS,*
A. van der HEIDE,*§ and J. W. J. BIJLSMA*

*Department of Rheumatology, University Hospital Utrecht, †Department of Clinical and Health Psychology, University of Utrecht, ‡Department of Medical Psychology, University of Nijmegen and §Arthritis Research Foundation Utrecht (SRU), The Netherlands

SUMMARY

Health status measures in rheumatoid arthritis that have been extensively validated for use in clinical trials are generally used also in correlative studies, e.g. to predict future health status. This application requires stability (repeatability of measurements). The purpose of our study was to determine the stability of commonly used health status measures. Two measurements at an interval of 6 months were taken in 99 patients. High stability (a = 0.78 to 0.94) was observed for five biomedical measures (grip strength, walking time, platelet count, haemoglobin and erythrocyte sedimentation rate) and five self-report measures (mobility, self care, impact daily activities, anxiety and cheerful mood). Moderate stability (a = 0.65 to 0.72) was observed for joint scores, pain, C-reactive protein and depressive mood. The highly stable measures most adequately reflect individual differences, may be applied most reliably in correlative studies and appear to have the largest clinical utility with regard to long-term prediction of health status.

KEY WORDS: Rheumatoid arthritis, Stability, Reliability, Validity, Outcome measures, Clinimetrics.

In rheumatoid arthritis (RA) health status measurement for the purpose of therapy evaluation has been the subject of extensive research and has been validated in several ways [1, 2]. Recently, recommendations on the use of process and outcome measures [3], definitions of clinically important changes in patients and groups [4] and a core set of disease activity measures in clinical trials of RA patients have been published [5].

The same measures that have been validated for the purpose of evaluation are generally used also in correlative studies, examining whether health status variables predict future health status or vary concomitantly with other variables (e.g. demographic, biological or social variables). However, in correlative studies other considerations should guide the choice of variables than those used in evaluation studies. In evaluation studies the ability of a measure to detect intra-individual change in response to treatment is important. Variance due to change, the responsiveness coefficient, reflects this property [6]. In correlative studies the ability of a measure to discriminate between individuals is important. Discrimination refers to variation between individuals. This property is reflected in the stability coefficient and other reliability coefficients [6]. Stability refers to the extent to which measurements are the same when retested after some period of time. Stability is determined by: (a) repeatability (or changeability) of measurements and (b) the range of inter-individual differences. When variation between individuals is small then even minor intra-individual changes in health status may reduce stability. But when individual differences are large then even clinically relevant intra-individual changes may only marginally affect the relative position of patients towards each other, and stability will be high.

Stability has hardly been investigated systematically [7] or only for very short intervals between measurements [8–10]. A single measurement of health status reflects first of all enduring differences between individuals, i.e. the irreversible outcome of the disease and other stable aspects of individuals such as constitutional differences. But a single assessment also reflects current disease activity and other temporary fluctuations such as seasonal variations and inaccuracy of measurements [7]. Whatever the cause of low stability, fluctuations of the disease or measurement error, if a health status measure predominantly reflects a transient state, little can be expected when this measure is correlated with enduring aspects of persons or when it is used to predict future health status. The first and most important purpose of stability analysis is to quantify the extent to which health status measures in RA reflect stable individual differences.

In correlative studies conclusions are often based on single measurements, based on the assumption that these assessments reflect to a large extent stable individual differences between patients. Except for pain assessment [11], validation of this current practice to infer from periodic single assessments has not gained much attention. The second purpose of stability analysis is to determine whether it is meaningful to infer from a single assessment, e.g. to predict future health status.

If a single assessment of a variable proves to be unsuitable for prediction purposes, one could use: (a) composite measures of multiple variables or (b) the mean of repeated measurements of the same variable.
[11–13]. The mean of repeated measurements of a variable more accurately reflects relatively enduring characteristics of patients than does a single measurement. This mean is consequently more appropriate to predict future health status. Stability analysis provides the tools to estimate the number of measurements needed to arrive at a mean value that adequately reflects individual differences. This is the third purpose of stability analysis.

The aim of the present study was to quantify the stability of commonly used measures of health status in RA in order to validate their use in correlative studies.

PATIENTS AND METHODS

Data were collected from 99 out-patients who regularly visited one of four hospitals in the Utrecht area of The Netherlands. Selection criteria were: RA according to the 1987 classification criteria [14], a minimal age of 20 and a minimal disease duration of 1 yr. Patients participated in a study on behavioural and occupational interventions or were on a waiting list [15]. In the current analysis data comprised two measurements collected after completion of the intervention. The interval between these measurements was 6 months. During these months on-going antirheumatic drug therapies were continued. Biomedical measurements at the two time points were carried out by the same rheumatology research nurse; self-report questionnaires were collected on the same day. The characteristics of the patients are shown in Table I.

Two types of biomedical measures were taken:

<table>
<thead>
<tr>
<th>Table I</th>
<th>Characteristics of the 99 patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (number of patients)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>68</td>
</tr>
<tr>
<td>Male</td>
<td>31</td>
</tr>
<tr>
<td>Age (yr)</td>
<td></td>
</tr>
<tr>
<td>Mean (s.d.)</td>
<td>58 (12.7)</td>
</tr>
<tr>
<td>Range</td>
<td>25 82</td>
</tr>
<tr>
<td>Disease duration (yr)</td>
<td></td>
</tr>
<tr>
<td>Mean (s.d.)</td>
<td>14.8 (11.5)</td>
</tr>
<tr>
<td>Range</td>
<td>2 59</td>
</tr>
<tr>
<td>ARA functional class (number of patients)</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>13</td>
</tr>
<tr>
<td>II</td>
<td>76</td>
</tr>
<tr>
<td>III</td>
<td>10</td>
</tr>
<tr>
<td>IV</td>
<td>0</td>
</tr>
<tr>
<td>Medication* (number of patients)</td>
<td></td>
</tr>
<tr>
<td>NSAID or analgesic, no DMARD</td>
<td>29</td>
</tr>
<tr>
<td>DMARD (HCQ/gold/pen/SASP/esp)</td>
<td>39</td>
</tr>
<tr>
<td>DMARD (methotrexate/azathioprine)</td>
<td>25</td>
</tr>
<tr>
<td>Prednisone alone</td>
<td>0</td>
</tr>
<tr>
<td>Complementary treatment alone</td>
<td>2</td>
</tr>
</tbody>
</table>

*HCQ, hydroxychloroquine; gold, aurothioglucose or auranofin; pen, p-penicillamine; SASP, sulphasalazine; esp, cyclosporin. Patients using disease-modifying anti-rheumatic drugs (DMARDs) could also be using non-steroidal anti-inflammatory drugs (NSAIDs). Nine of the patients using DMARDs were also on prednisone (2.5 15 mg/day) treatment. The medication of four patients was unknown.

Clinical assessments and laboratory measures. Clinical assessments consisted of grip strength (patients were asked to squeeze the cuff of a partially inflated sphygmomanometer as tightly as possible; the mean value of the best grip strength of three attempts in the left and right hand was taken), Thompson joint score (clinical assessment of tender and swollen joints) [16] and 30-m walking time. Blood samples were analysed for erythrocyte sedimentation rate (ESR) (Westergren method), C-reactive protein (CRP: nephelometric), haemoglobin and platelet count (both by Coulter counter).

To assess physical and psychosocial health status, the Impact of Rheumatic Diseases on General Health and Lifestyle (IRGL) self-report questionnaire was used [17]. This validated questionnaire is partly derived from the Arthritis Impact Measurement Scales 1 [18]. Of the seven scales used in the current analyses, three assess the physical dimension of health (mobility, self care and pain), three address psychological discomfort (anxiety, depressive mood and cheerful mood) and one scale assesses the impact of RA upon daily activities such as work, household activities and sleeping (impact daily activities).

Statistical analyses

To estimate test–retest stability of measures Cronbach’s α was computed using the SPSS procedure reliabilities [19]. This coefficient reflects the proportion of variance of the mean score of two measurements that is explained by stable individual differences. It can take any value between 0 and 1: a value approaching 1 reflects a stable position of patients towards each other with two repeated measurements. The more repeated measurements are taken of a variable, the better the mean score of a patient reflects the usual position of that patient relative to other patients. The Spearman–Brown formula was used to estimate stability coefficients of health status measures as a function of one to six theoretical measurements [8, 20, 21], in which the estimate of a single measurement reflects the proportion of variance of a single measurement (in the 6-month interval) that must be ascribed to individual differences. These estimated stability coefficients show how many measurements are needed to arrive at a specific level of stability. As the required level of stability of a measure depends on the specific use of the measure, no single level can be considered adequate in all circumstances. For studies into groups of patients coefficients of 0.70 are considered sufficient; for clinical outcome studies with relatively few subjects or examining individual responses to treatment 0.90 may be a minimum [7, 13].

Lack of change in the test–retest interval as well as the range of inter-individual differences determine stability estimates. To be able to interpret stability estimates the percentage of patients changing more than 36% relative to initial measurements was computed for each variable. This 36% criterion has been proposed to define the cut-off point of clinically important change in individual patients [4].
In all analyses missing values were deleted pair-wise. To meet requirements for parametric statistics, transformations were applied where appropriate.

RESULTS

Stability

Descriptive statistics of the first of the two assessments are shown in Table II. Estimates of stability coefficients as a function of one to six measurements are shown in Table III. Alpha coefficients are shown in column 2 of Table III. Overall, 65% ($\alpha = 0.65$) to 94% ($\alpha = 0.94$) of the observed score variances reflect stable inter-individual differences. The proportion of variance in the group explained by individual differences is high ($\alpha = 0.78$ to 0.94) in five biomedical measures (grip strength, walking time, platelet count, haemoglobin and ESR) and five self-report measures (mobility, self care, anxiety, cheerful mood and impact daily activities). Moderate stability ($\alpha = 0.65$ to 0.72) is observed for Thompson joint score, CRP, pain and depressive mood. Estimates of stability coefficients as a function of one to six measurements (columns 1–6 in Table III) show that with a single assessment the stability coefficients of more than half of the measures, including all clinical observations and self-report measures of functional ability, are greater than 0.70; with three measurements all coefficients are in excess of 0.70.

Changeability

Taking 36% intra-individual change from the first to the second measurement as a cut-off criterion for clinically relevant change, it was proved that 70% of the patients changed at Thompson joint score, whereas the percentages for depressive mood and ESR were 56 and 54, respectively. Also, considerable proportions of the sample changed at CRP (43%), grip strength (34%), cheerful mood (29%) and pain (24%).

The estimated stability of health status measures as a function of the number of measurements:

\begin{table}
\centering
\caption{The estimated stability of health status measures as a function of the number of measurements.}
\begin{tabular}{l|c|c|c|c|c|c}
\hline
& \multicolumn{6}{c}{Number of measurements} \\
\hline
& 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
Grip strength & 0.89 & 0.94 & 0.96 & 0.97 & 0.98 & 0.98 \\
Mobility & 0.83 & 0.91 & 0.94 & 0.95 & 0.96 & 0.96 \\
Self care & 0.80 & 0.89 & 0.92 & 0.94 & 0.95 & 0.96 \\
Anxiety & 0.79 & 0.88 & 0.92 & 0.94 & 0.95 & 0.96 \\
Walking time & 0.77 & 0.87 & 0.91 & 0.93 & 0.94 & 0.95 \\
Platelet count & 0.74 & 0.85 & 0.89 & 0.92 & 0.93 & 0.94 \\
Haemoglobin & 0.72 & 0.84 & 0.89 & 0.91 & 0.93 & 0.94 \\
Cheerful mood & 0.72 & 0.84 & 0.89 & 0.91 & 0.93 & 0.94 \\
Impact daily activities & 0.71 & 0.83 & 0.88 & 0.91 & 0.92 & 0.94 \\
ESR & 0.64 & 0.78 & 0.84 & 0.88 & 0.90 & 0.91 \\
Thompson joint score & 0.56 & 0.72 & 0.79 & 0.84 & 0.87 & 0.89 \\
Pain & 0.56 & 0.72 & 0.79 & 0.84 & 0.87 & 0.88 \\
C-reactive protein & 0.52 & 0.68 & 0.76 & 0.81 & 0.84 & 0.86 \\
Depressive mood & 0.48 & 0.65 & 0.74 & 0.79 & 0.82 & 0.85 \\
\hline
\end{tabular}
\end{table}

The stability of two measurements over 6 months was computed using Cronbach's $\alpha$. This coefficient reflects the proportion of variance of the mean score of two measurements that is explained by stable individual differences. Coefficients of one measurement (reflecting the proportion of variance of scores of a single measurement that must be ascribed to individual differences) and of three to six measurements were estimated with the Spearman–Brown formula. From the Table it can be seen how many measurements are needed to arrive at a coefficient of, e.g., 0.70 or 0.90 (dotted lines; see text for the meaning of these levels). For instance 90% of the mean values of five repeated measurements of ESR reflect stable individual differences; the remaining 10% reflects 'true' changes in time (e.g. due to changes in disease activity) or measurement error.

Although some variables perform better than others, most of the commonly used health status variables in RA largely reflect stable individual differences. Table III may be used to guide the choice of variables in correlative studies and to decide whether a single measurement or the mean of repeated measurements is to be used.

In the current study the stability of measures of physical functioning (mobility, self care, walking time) reflects a lack of intra-individual change during the course of the study as well as inter-individual differences. This contrasts with measures that have been proven to be sensitive to change in clinical trials, e.g. ESR and joint scores [3, 5, 22]. In our study less than 8% of the patients changed more than 36% at physical functioning, while more than half of the patients changed more than 36% at ESR and joint scores. These meaningful changes in a large proportion of the sample did not, however, affect stability to a large extent. This

\begin{table}
\centering
\caption{Descriptive statistics of health status measures of the 99 patients.}
\begin{tabular}{l|ccc}
\hline
\textbf{Biomedical measures} & \textbf{Mean} & \textbf{s.d.} & \textbf{Range} \\
\hline
Grip strength (mmHg) & 35 & 27 & 0–144 \\
Walking time (s) & 31 & 16 & 18–116 \\
Platelet count (10$^9$/l) & 287 & 91 & 55–569 \\
Haemoglobin (mmol/l) & 8.3 & 0.9 & 5.5–10.5 \\
ESR (mm 1$^4$h) & 34 & 28 & 3–140 \\
Thompson joint score* & 103 & 98 & 0–467 \\
C-reactive protein (mg/l) & 11 & 24 & 0–204 \\
\hline
\textbf{Self-report measures} & \textbf{Mean} & \textbf{s.d.} & \textbf{Range} \\
\hline
Mobility & 19.0 & 6.3 & 7–28 \\
Self care & 24.4 & 6.0 & 10–32 \\
Anxiety & 18.8 & 6.0 & 10–32 \\
Cheerful mood & 10.8 & 4.8 & 0–24 \\
Impact daily activities & 22.0 & 5.1 & 10–35 \\
Pain & 15.6 & 4.5 & 6–25 \\
Depressive mood & 3.2 & 4.0 & 0–17 \\
\hline
\end{tabular}
\end{table}

*Theoretical range 0–534.

†Scales of the RQL. Theoretical ranges: mobility 7–28, self care 8–32, pain 6–25, anxiety 10–40, depressive mood 0–24, cheerful mood 0–24, impact daily activities 10–40. A large impact of the disease is reflected in low scores for the scales mobility, self care and cheerful mood, and in high scores for the other scales.
can only mean that the stability of these variables must be explained in terms of individual differences. Due to this broad range of individual scores the relative position of patients towards each other is maintained, even if a large part of the sample of patients shows clinical amelioration or deterioration while another part does not change. The observation of large heterogeneity in health status is in agreement with other studies [5, 23, 24].

The stability estimates in the first column of Table III may be used to decide whether a single assessment of the health status measures is sufficiently reliable to predict future health status or to correlate it with more enduring characteristics of patients. The single assessment stability estimates of five health status measures do not exceed 0.70. Defining a value of 0.70 as a relevant boundary, these five measures do not provide a reliable characterization of health status during the 6-month interval of investigation. Whatever the cause of this lower stability, whether it be real fluctuations in disease activity or inaccurate measurements, a single measurement of these less stable measures reflects a more temporary and transient state of affairs. Therefore, correlations with enduring characteristics of patients or with future health status will very likely be low. Taking into account transient variation, the mean score of repeated measurements of a less stable measure should be used to predict endpoints rather than a single assessment. The mean of three to five measurements of joint scores, pain and acute-phase reactants reflects individual differences as adequately as does a single assessment of functional ability. To be able to remove from the score both the variations that are due to inaccurate measurements and fluctuations of the disease, these measurements should be taken at several time points during the 6-month interval. In the current study only grip strength was repeatedly measured at the same moment. This may have inflated the stability coefficient a little, because only true changes during the 6-month interval did reduce the stability, whereas the reduction of stability in other measurements was due to inaccurate measurement as well.

The sample studied seems rather representative for the population of non-hospitalized patients with RA. In comparison with other samples from the same area, the patients did not differ with respect to gender, age or duration of the disease [17, 23, 24]. Moreover, the means and ranges of the measures evaluated do not differ widely from those found in other samples of out-patients [7, 12, 16, 24–26]. Conclusions of the current study apply to health status scores of patients in a relatively short time frame of 6 months. With shorter time intervals stability obviously will be higher [9, 10]. Over a longer time course patients will show differential clinical progression of the disease, resulting in smaller stability [27]. On a much longer time scale a joint score that varies between fixed limits may be more stable than a measure of functional ability that may steadily progress in some patients but not in others. The results of this study may not apply to early RA. It is conceivable that general health and well-being are more affected by inflammatory flares and remissions in patients with a recent diagnosis, in whom there are less irreversible consequences of the disease, than in later phases of the disease [28]. The stability of measurements within these groups (and therefore, for example, the prediction of future health status) may vary accordingly.

Our study was designed before the American College of Rheumatology (ACR) and EULAR core sets of disease activity measures for RA clinical trials were established [1, 5, 29]. Nevertheless, most disease activity measures from these core sets were included in the current study, viz., a joint score, assessments of pain and physical function and the acute-phase reactants ESR and CRP. The choice of the ACR core set was guided by several validity considerations, especially sensitivity to change in clinical trials and the ability of measures to predict important long-term outcomes in RA. The extent to which the measures of the ACR core set comply with these two types of validity relates strongly to the stability of measures observed in this study. The moderately stable measures (joint scores, ESR and pain) are most sensitive to change in clinical trials, while the highly stable measures of physical function are the stronger predictors of long-term outcome in RA. The fact that certain variables were chosen in the core sets because of their superior characteristics regarding therapy evaluation does not mean that these variables are also superior in predicting disease course or in correlative studies. Values of grip strength, walking time, button test and questionnaire scores of activities of daily living proved to have a substantial clinical utility regarding long-term prediction of health status, predicting morbidity and mortality 9 yr later [30]. By inference, it is worthwhile to study the long-term prediction of the other variables that proved to be stable in this study. Stability is an important feature of health status variables that should be considered next to (not opposed to) other generally recognized features such as sensitivity to change. Stability should guide the choice of variables in correlative studies and sensitivity to change in evaluation studies. The results of the current study therefore do not give rise to change the ACR preliminary core set of disease activity measures. Rather, it enhances our insight into the stability of distinct measures of the set, and consequently into the applicability of these measures in correlative investigations.

CONCLUSION

Stability should guide the choice of variables in correlative studies as sensitivity to change guides the choice of variables in evaluation studies. The stability of commonly used health status measures in RA varies from moderate to high. The highly stable measures most adequately reflect individual differences. These highly stable measures may reliably be applied in correlative studies and appear to be relevant with regard to the long-term prediction of health status.
ACKNOWLEDGEMENTS

For help with data acquisition the authors are indebted to the rheumatology research nurse (C. Cornelis), researchers of the University Hospital Utrecht (L. Gijis, C. J. A. E. Huiskes and M. Richardson) and rheumatologists participating in the Arthritis Research Foundation Utrecht (SRU) (G. A. Albada-Kuipers, H. G. J. Haanen, Y. Schenk and M. J. van der Veen). For useful suggestions following an earlier draft we express our gratitude to F. W. M. M. Touw-Otten (University Hospital Utrecht) and J. R. Kirwan (University of Bristol). This study was supported by a grant from the Dutch League Against Rheumatism (Nationaal Reumafonds).

REFERENCES