HYPOTHESIS

# Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change

K Bruynesteyn, M Boers, P Kostense, S van der Linden, D van der Heijde

*Ann Rheum Dis* 2005;**64**:179–182. doi: 10.1136/ard.2003.018457

Progression of radiological joint damage is usually based on the simultaneous assessment of a series of films from an individual patient (''paired'', with or without known sequence). In this setting the degree of progression that can be reliably detected above the measurement error is best determined by the smallest detectable *change*, and overestimated by the traditionally calculated smallest detectable *difference*. This knowledge is important for calculation of the proportion of patients showing radiographic progression in clinical trials.

See end of article for authors' affiliations

Correspondence to:
Professor D van der Heijde, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, PO Box 5800, 6202 AZ Maastricht, The Netherlands;
dhe@sint.azm.nl

Accepted 16 July 2004
**Published Online First 29 July 2004**

In therapeutic trials the main interest is usually the analysis based on the mean or median change at a group level of the outcome measure(s) over time. The percentage of patients responding to a particular intervention is, however, included increasingly in trial analyses because this can add valuable information.[1] Presenting the percentage of (non) responders helps the practising rheumatologist to individualise group data found in trials, the ultimate goal of evidence based medicine. It further enables the practitioner to calculate the so-called number needed to treat (NNT).[2]

For the interpretation of radiological progression of joint damage due to rheumatoid arthritis (RA) in the hands and feet it can be also very useful to present the number patients responding in addition to the means and medians.[3] Particularly, because radiographic data show a highly skewed distribution pattern, the majority of patients show only mild or no progression in the observation period, and only a subset of patients show substantial progression. Note that for radiological joint damage due to RA, ''response'' can be translated as no progression of damage (and theoretically, also improvement or repair).

To determine the percentage of patients who showed a relevant change over time continuous data have to be dichotomised, and thus a valid and clinically relevant cut off level should be chosen (called the minimal clinically important difference). It seems logical that such a cut off value should at least be greater than the measurement error of the instrument used to quantify the response. As a starting point the smallest detectable difference (SDD) has therefore been suggested as the cut off level.[3–5] The SDD expresses the smallest difference between two *independently* obtained measures that can be interpreted as ''real''—that is, a difference greater than the measurement error.

Radiological joint damage due to RA is usually assessed with the films of one particular patient side by side—that is, simultaneously. The purpose of the simultaneous reading of the films is that the raters can compensate for variation in positioning of the hands and feet and in film quality, minimising the measurement error.[6 7] But when films are read side by side, the measures are not obtained independently and therefore the SDD is not the appropriate way to define a cut off level.

This paper aims at presenting the way in which the smallest change in scores that can be deemed as a ''real'' change can be assessed correctly, from here on called the smallest detectable change (SDC) for the setting in which films are read simultaneously. In subsequent paragraphs, we will first explain further why the regular SDD should not be used. Then, we will describe two methods of estimating the measurement error of change and accompanying SDC. Finally, we will give an example of both calculations based on the reliability data of 10 random subjects of the COBRA trial[8] and demonstrate that if films are scored simultaneously, the SDD results in a cut off level which is too high to detect change in an individual patient.

## READING RADIOGRAPHS AS SINGLE FILMS COMPARED WITH READING AS PAIRS

Whether the SDD or the SDC should be used to determine if a patient's score has changed more than the measurement error depends on whether the change is based on two *independently* obtained scores or not.

Radiographs can be presented to a rater completely at random—that is, a single film at a time, or they can be grouped for each patient so that all films of one particular patient are read simultaneously, with or without data on the sequence of the films. As mentioned briefly in the introduction, reading all films of one patient simultaneously has the major advantage that the rater will be able to correct for variation in positioning of the hands and feet or variation of the film quality. When films are grouped for each patient, a rater compares all the films of one

**Abbreviations:** ANOVA, analysis of variance; SDC, smallest detectable change; SDD, smallest detectable difference patient and judges whether a change in joint damage has occurred. Paired (without information on the chronology of the films) or sequenced (with information on the chronology of the films) reading thereby aims at reducing the measurement error of the object of interest in trials: the change in joint damage. However, the measurement errors of the scores are thereby correlated and are no longer independent. Furthermore, the change-score (the score obtained by subtracting the status-score at time 1 from the status-score at time 2) can actually be interpreted as a single score. Because one actually tests the hypothesis whether the difference between two *independent* status-scores differs from 0[1 5 9 10] when determining a regular SDD, it is inappropriate to use an SDD based on *dependent* status-scores.

## CHANGE IN AN INDIVIDUAL PATIENT COMPARED WITH CHANGE BETWEEN PATIENTS

To determine if the scores of a patient changed over time in the case of dependent scores, the hypothesis to be tested should be whether the change-score in an individual patient differs from 0. In the past, researchers have calculated SDDs based on the change-score. However, with SDDs based on change-scores, one actually tests the hypothesis whether the difference between two independent change-scores differs from 0. So the SDD based on change-scores is the correct measure if we want to test whether the change-score from one patient is different from the change-score of another patient. However, when we want to assess if an individual patient shows progression, this is not appropriate and an alternative, the SDC, should be used.

Note that with independent scores we do not refer to the condition that scores of one person are independent statistically. Scores of one patient are always paired observations and thereby statistically dependent. With independent scores we refer to the fact that the measurement error of the scores of the films from time 2 is not related to the measurement error of the scores of the films from time 1. When scoring films of times 1 and 2 simultaneously this is not true.

## APPROPRIATE METHODS TO ASSESS THE SDC FOR SIMULTANEOUS READING OF FILMS

In the field of radiological joint damage, it is customary to assess the measurement error of the scoring methods by rescoring the radiographs by the same rater (intrarater reliability) or by a different rater (interrater reliability). The option of obtaining a second radiograph is never applied on a large scale in trials because this would require additional exposure of the patients to radiation. Because the same radiographs are re-examined, differences between the two observations on each individual subject are caused by measurement error.

Here we shall describe two correct methods of estimating the measurement error of the change-score of two simultaneously scored films and the accompanying SDC. The first method quantifies measurement error by the standard error of measurement of the change-score ($\mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}$) derived from a two way analysis of variance (ANOVA). The second method estimates the measurement error by calculating the standard deviation of the difference between change-scores of two reading sessions ($\mathrm{SD}_{\Delta(\mathrm{CHANGE\text{-}SCORES})}$). The latter resembles the estimation of the measurement error as described by Bland and Altman,[4 9] in which the standard deviation of the difference between status scores of two reading sessions ($\mathrm{SD}_{\Delta(\mathrm{STATUS\text{-}SCORES})}$) estimates the measurement error. This second method is only applicable for two raters or one repetition; the first method, using an ANOVA analysis, can be applied also if measurements of three or more raters or two or more repetitions are available.

### Measurement error estimated with the $\mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}$

For this method one runs a two way ANOVA to estimate measurement error expressed by the $\mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}$, with the change-scores of repeated measurements or the change-scores of two or more raters. The two way ANOVA will result in mean squares for the different sources of variation in the change-scores. These are the between-patient variation (variance caused by the variation in change between the patients) and the within-patient variance. The within-patient variance is composed of the variance caused by the systemic variation between the first and the second reading session or the first and the second rater (between-measurement variation) and the random variation in change-scores (the residual error). The SEM is calculated by taking the square root of the error variance. The error variance can constitute the total within-patient variance (including the between-measurement variance) or the residual variance only. If one does not want to generalise to other raters not included in the reliability study, the between-measurement variance should not be included in the error variance. In this paper we will base the $\mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}$ on residual error, only, to be able to compare directly the SDC calculated with an $\mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}$ with the SDC calculated with the $\mathrm{SD}_{\Delta(\mathrm{CHANGE\text{-}SCORES})}$.

Because an SEM is a variable that expresses the amount of measurement error in the original metric unit of the measurement it can be used to calculate an interval of error around scores, assuming that the measurement error is distributed normally. At the 95% confidence level, the interval around a change-score is calculated according to the formula:

$$\text{change-score} \pm (1.96 \times \mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}/\sqrt{k})$$

where k represents the number of readings or raters used for the actual analyses of a trial. For a trial in which the results are based on the mean scores of k raters/readings, the measurement error diminishes by a factor $\sqrt{k}$.

If a calculated 95% interval around a change-score contains the value 0, the null hypothesis that the change-score is 0 cannot be rejected. So, values of the change-score lying in this interval other than 0 might be induced by error alone. On the other hand, if the interval does not include 0 we reject the null hypothesis and conclude that the change-score really differs from 0 and state that there is a ''real'' change in joint damage.

From the formula of the 95% interval around the change-score we can see that a change-score larger than $\pm 1.96 \times \mathrm{SEM}_{\mathrm{CHANGE\text{-}SCORE}}/\sqrt{k}$ can be regarded as larger than the measurement error. This formula is thus the formula used to calculate the SDC based on change-scores obtained with simultaneous reading of films.

Please note that the absolute change-score (that is, disregarding the sign) needs only to be half the size of the full interval around the change-score. In other words: in the case of two readings, the first can be either larger or smaller than the second purely by measurement error, and the full range of possible changes (with positive and negative sign) is described by the interval in the formula. However, once we find that the first reading is larger than the second we will reject the null hypothesis once the change falls outside the range of the interval, in this case on the positive side. This will occur when the change is half the size of the full interval.

**Table 1** Scores of two different raters of 10 patients

| | Baseline score | | Difference status-scores | Change score | | Difference change-scores |
|---|---|---|---|---|---|---|
| Subject | 1st Rater | 2nd Rater | between raters | 1st Rater | 2nd Rater | between raters |
| 1 | 12 | 15 | 3 | 2 | 3 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 23 | 20 | −3 | 9 | 10 | 1 |
| 4 | 5 | 2 | −3 | 4 | 1 | −3 |
| 5 | 7 | 11 | 4 | 11 | 10 | −1 |
| 6 | 3 | 0 | −3 | 1 | 0 | −1 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1 | 1 | 0 | 5 | 8 | 3 |
| 9 | 9 | 6 | −3 | 17 | 13 | −4 |
| 10 | 19 | 25 | 6 | 10 | 11 | 1 |
| Mean | | | 0.1 | | | −0.3 |
| SD | | | 3.28 | | | 2.06 |

### Measurement error estimated with the SD of the differences between change-scores of two reading sessions ($SD_{\Delta\ (CHANGE\text{-}SCORES)}$)

If repeated measurements or scores of two raters are used to assess the SDC, the measurement error can also be estimated with the standard deviation of the differences between change-scores of two reading sessions ($SD_{\Delta(CHANGE\text{-}SCORES)}$). For this method one firstly calculates the differences between the change-scores obtained in the repeated reading session. Secondly, the SD of these differences is calculated. This $SD_{\Delta(CHANGE\text{-}SCORES)}$ reflects the measurement error of the difference between two change-scores—that is, the measurement error when discriminating between two change-scores. However, for evaluating a clinical trial one is interested in whether a change-score in an individual patient really can be distinguished from 0 (in other words, if progression of damage is seen in this patient) and not whether two change-scores differ from each other (in other words, if the progression in one patient is significantly different from the progression in another patient). The measurement error of a single change-score is obtained by dividing the $SD_{\Delta(CHANGE\text{-}SCORES)}$ by $\surd 2$.[10] The error interval around a change-score, at the 95% confidence level, is further calculated by:

$$\text{change-score} \pm (1.96 \times SD_{\Delta(CHANGE\text{-}SCORES)})/(\surd 2 \times \surd k)$$

and the SDC by:

$$\pm\ 1.96 \times SD_{\Delta(CHANGE\text{-}SCORES)}/(\surd 2 \times \surd k)$$

in which k again represents the number of readings over which one wants to average the analyses of the trial. These formulae show that:

$$SD_{\Delta(CHANGE\text{-}SCORES)}/(\surd 2 \times \surd k)\ =\ SEM_{CHANGE\text{-}SCORE}/\surd k$$

if the data of one repetition or two raters is used.[11]

**Table 2** Two way analysis of variance for the change-score data of table 1

| Source of variation | Degrees of freedom | Sum of squares | Mean square |
|---|---|---|---|
| Between-patient* | 9 | 520.25 | 57.81 |
| Within-patient† | 10 | 19.50 | 1.95 |
| Between-measurement‡ | 1 | 0.45 | 0.45 |
| Residual§ | 9 | 19.05 | 2.12 |
| Total | 19 | 539.75 | |

* Variance in scores due to differences between patients; †variance in scores due to differences within a patient; ‡variance in scores due to the differences between the raters; §variance in scores due to unknown sources.

## EXAMPLE OF THE TWO METHODS TO ASSESS THE SDC FOR SCORES BASED ON SIMULTANEOUS READING OF FILMS

In this section we will show both calculations for the reliability data of 10 subjects from the COBRA trial.[8] These 10 subjects were randomly selected from the subgroup of patients showing progression of <25 Sharp/van der Heijde (90th centile) to ensure homoscedasticity. As we know that radiological data are often highly skewed and because it is known that measurement error tends to be larger in patients with more baseline damage and radiological progression,[4] the assumption of homoscedasticity can be violated.

Table 1 shows the status and change-scores of two different raters of the 10 subjects, and table 2 shows the two way ANOVA for the change-scores. The $SD_{\Delta(CHANGE\text{-}SCORES)}$ in this example is 2.06 and the $SEM_{CHANGE\text{-}SCORE}$, calculated by extracting the root of the residual mean square ($\surd 2.12$), is 1.46. Putting these figures into the formulae presented in the section above results in both cases in an SDC of 2.85 scoring units, if not using average scores ($k = 1$).

## OVERESTIMATION OF THE MEASUREMENT ERROR BY THE SDD

To complete this report, we will show that if one uses the regular SDD—for example determined with the baseline status scores of the two raters—to determine whether a patient's scores have really changed, this will result in an overestimation of the measurement error. The SDD according to Bland and Altman[4, 9] is estimated by calculating a 95% interval around the difference between two single status scores with the formula:

**Table 3** Two way analysis of variance for the baseline status-score data of table 1

| Source of variation | Degrees of freedom | Sum of squares | Mean square |
|---|---|---|---|
| Between-patient* | 9 | 1298.45 | 144.27 |
| Within- patient† | 10 | 48.5 | 4.85 |
| Between-measurement‡ | 1 | 0.05 | 0.05 |
| Residual§ | 9 | 48.45 | 5.38 |
| Total | 19 | 1346.95 | |

$$\pm 1.96\times\sqrt{2}\times(\mathrm{SEM}_{\mathrm{STATUS\text{-}SCORE}}/\sqrt{k}$$

or by

$$\pm 1.96\times \mathrm{SD}_{\Delta(\mathrm{STATUS\text{-}SCORES})}/\sqrt{k}$$

Calculating the SDD with the $\mathrm{SD}_{\Delta(\mathrm{STATUS\text{-}SCORES})}$ (see table 1; $\pm 1.96\times 3.28/\sqrt{k}$) or with the $\mathrm{SEM}_{\mathrm{STATUS\text{-}SCORE}}$ (see table 3; $\pm 1.96\times\sqrt{2}\times\sqrt{5.38}/\sqrt{k}$) results in an SDD of 6.4 units if not using average scores ($k = 1$). So, the measurement error of detecting a change within patients is clearly smaller than the measurement error of detecting a difference between two single baseline scores, as expected.

## Conclusions

In summary, progression of radiological joint damage is usually based on simultaneous assessment of a series of films from an individual patient (''paired'', with or without known sequence). In this setting, the amount of progression that is reliably detectable above the measurement error is best determined by the smallest detectable *change*, and overestimated by the traditionally calculated smallest detectable *difference*. This is important knowledge for the calculation of the proportion of the patients showing radiographic progression in clinical trials.

## Authors' affiliations

**K Bruynesteyn, S van der Linden, D van der Heijde,** Department of Internal Medicine, Division of Rheumatology, University Hospital of Maastricht and CAPHRI Research Institute University of Maastricht, Maastricht, The Netherlands

**M Boers, P Kostense,** Department of Clinical Epidemiology and Biostatistics, VU University Medical Centre, Amsterdam, The Netherlands

## REFERENCES

1 **Streiner DL**, Norman GR. *Measuring change. Health measurements scales. A practical guide to their development and use.* Oxford: Oxford University Press, 1995:163–80.

2 **Guyatt GH**, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;**316**:690–3.

3 **van der Heijde D**, Simon L, Smolen J, Strand V, Sharp J, Boers M, *et al.* How to report radiographic data in randomized clinical trials in rheumatoid arthritis: guidelines from a roundtable discussion. *Arthritis Rheum* 2002;**47**:215–18.

4 **Lassere M**, Boers M, van der Heijde D, Boonen A, Edmonds J, Saudan A, *et al.* Smallest detectable difference in radiological progression. *J Rheumatol* 1999;**26**:731–9.

5 **Ravaud P**, Giraudeau B, Auleley GR, Edouard-Noel R, Dougados M, Chastang C. Assessing smallest detectable change over time in continuous structural outcome measures: application to radiological change in knee osteoarthritis. *J Clin Epidemiol* 1999;**52**:1225–30.

6 **van der Heijde D**, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. *Rheumatology (Oxford)* 1999;**38**:1213–20.

7 **Bruynesteyn K**, van der Heijde D, Boers M, Saudan A, Peloso P, Paulus P, *et al.* Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. *J Rheumatol* 2002;**29**:2306–12.

8 **Boers M**, Verhoeven AC, Markusse HM, van de Laar MAFJ, Westhovens R, van Denderen JC, *et al.* Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet* 1997;**350**:309–18.

9 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.

10 **Beckerman H**, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res* 2001;**10**:571–8.

11 **Hopkins WG**. Measures of reliability in sports medicine and science. *Sports Med* 2000;**30**:1–15.