Aspects of methodology in assessing inflammation and damage in rheumatoid arthritis and axial spondyloarthritis

Navarro Compan, Maria Victoria

5

Measurement error in the assessment of radiographic progression in rheumatoid arthritis clinical trials: the smallest detectable change revisited

V. Navarro-Compán, D. van der Heijde, H.A. Ahmad, C.G. Miller,
R. Wolterbeek, R. Landewé

Ann Rheum Dis 2014;73:1067-70

ABSTRACT

Objectives

To evaluate if the mean smallest detectable change (SDC) of multiple time intervals using the Bland & Altman (B&A) levels of agreement (LoA) method is an appropriate surrogate for the generalizability analysis method for estimating the overall SDC of radiological progression in rheumatoid arthritis (RA) trials. Secondly, to compare the SDC based on 95% LoA with the SDC based on 80% LoA, and to investigate the association between SDC and baseline damage and progression.

Methods

Fifteen datasets from randomized controlled trials in RA were scored by 13 experienced readers as pairs according to the modified Sharp-van der Heijde method. The SDC using the 95% and 80% LoA and the generalizability methods were calculated.

Results

21,295 radiographic time points from 7,643 patients were included. The mean (range) SDC for the LoA and the generalizability methods was 3.1 (2.3-4.3) and 3.2 (2.3-4.6) units, respectively. The mean ± SD difference between the two methods was -0.13 ± 0.28. The mean SDC including all intervals (n=31) was 3.0 ± 0.7 for 95% LoA and 2.0 ± 0.4 for 80% LoA. No relationship was observed between baseline damage and the SDC, whereas the SDC increased with increasing radiological progression.

Conclusions

The mean of the interval SDCs obtained by the simple LoA method is a valid surrogate for the SDC obtained by complex generalizability methods. The SDC depends on the level of radiographic progression rather than on the level of absolute damage. In addition, the use of an SDC based on 80% rather than on 95% LoA is proposed.

INTRODUCTION

Prevention of structural damage has been included as one of the claims in the US Food and Drug Administration [1] and the European Medicines Agency (EMA) Guidelines on clinical investigation of new drugs for treatment of rheumatoid arthritis (RA) [2]. The gold standard imaging technique for assessing the degree of structural damage is conventional radiography [3], and the Sharp-van der Heijde (SHS) method recommended by the EMA for assessing radiological progression [2].

Reliability of the scoring method is essential to be able to detect differences in radiological progression between treatment arms, in order to assess the efficacy of therapeutic interventions in randomized controlled trials (RCTs). Reliability can be reported in relative terms using statistics such as the intraclass correlation coefficients; however, descriptions of other statistics such as the smallest detectable difference (SDD) and the smallest detectable change (SDC) is recommended because they provide an estimation of absolute rather than relative reliability, and they may give clinical guidance for assessing real changes at the individual patient level [4]. While the SDD has been recommended as one of the measures in the guidelines on reporting radiographic data of RCTs in RA, the SDC is nowadays recognized as the preferred measure for absolute agreement [5,6]. The SDD is appropriate to determine if progression in patient A is different from progression in patient B. In order to determine if progression in an individual patient is beyond measurement error, however, the SDC is the most appropriate statistic [5,6].

At least two analytical methods for estimating the SDC are available: one 'simple' method is based on the standard deviation (SD) of the difference between change scores obtained by two readers resulting in 95% levels of agreement (LoA) (also referred to as the Bland & Altman (B&A) method); the other method is more complex and based on generalizability analysis (the analysis of variance (ANOVA) method).

Two arguments challenge the methodology of obtaining SDC cut-off levels as appropriate surrogates for inter-reader reliability:

1) The simple LoA method is only applicable if two scores (twice scored by the same observer or two observers) are obtained. In the case of multiple time points or multiple readers (complex databases), which is common in RA trials, only a generalizability analysis is appropriate. However, estimating the SDC in complex databases requires more statistical expertise and is more laborious, and a simpler method is warranted.

2) The SDC is calculated using 95% LoA, basically assuming that 95% of the inter-reader differences of paired observations in a scenario with two readers is captured within the area delineated by the upper and lower 95% LoA, and not more than 5% of the differences

are more extreme. It can be argued that this requirement is rather strict. For example, it has been shown that the SDD (SDC multiplied by SQRT2) is a conservative estimate, as rheumatologists have rated progression at or below this level as clinically significant [7]. Further, there is no scientific basis for choosing a 95% limit over a less strict limit, and one may argue that the use of 80% LoA is not only sufficiently strict to select a cut-off to determine if a patient shows progression beyond measurement error, but is also closer to reality in terms of what clinicians consider relevant.

The principal aim of this study is twofold: first, to evaluate if the mean SDC of multiple time intervals in complex databases using the 'simple' LoA method per interval is an appropriate surrogate for the generalizability analysis for estimating the overall SDC of radiological progression; second, to compare the SDC based on 95% LoA with the SDC based on 80% LoA, and to investigate the association between baseline radiological damage/radiological progression and the magnitude of the SDC.

METHODS

Data were extracted from 15 databases of RCTs testing biological treatments in patients with RA. All these trials were performed according to good clinical practice and all studies received ethical approval. We selected these studies because they had been used for registration purposes using similar methodologies and all scored by members of our group according to the SHS method [8]. Thirteen experienced readers, who had all received the same training, scored all digitized films in pairs on a 21 CFR Part 11 compliant read system deployed by BioClinica, Code of Federal Regulations. The readers were blinded to patient identification, treatment and chronology of the time points. Initially, the total SHS for all patients was calculated per visit and per reader for all visits. Next, the SDC was calculated using the simple LoA [9] and a generalizability analysis as follows.

LoA (B&A method)

First, the change score per reader was calculated on a per time-point basis [baseline-first follow-up, first follow-up-second follow-up and second follow-up-third follow-up (if applicable)], and subsequently the difference in change scores between the two readers was calculated. Second, the SD of that difference was calculated. The SDC for all intervals of each trial was estimated using the formula $(\pm 1.96*SD)/(\sqrt{2}*\sqrt{k})$ for 95% LoA and $(\pm 1.28*SD)/(\sqrt{2}*\sqrt{k})$ for 80% LoA, in which k represents the number of readers within the same reading session (equals 2 in this study). Finally, we estimated a mean 95% LoA SDC per study by calculating the average of the 95% LoA SDCs of all intervals of the study $(SDC^{1st\ interval} + SDC^{2nd\ interval} + SDC^{3rd\ interval} \ldots\ldots + SDC^{n\ interval})/n$.

Generalizability analysis (ANOVA method)

For the generalizability analysis, we performed an ANOVA as proposed by Bruynesteyn et al [6]. Random variation in change scores (the residual error) per trial was determined, taking into account all the time points from the same trial, using a full-factorial univariate linear model, as detailed in the statistical analysis below. The standard error of the mean (SEM) was calculated by taking the square root of this residual error and the SDC for all intervals of each trial was estimated using the formula (±1.96*SEM)/√k for 95% LoA and (±1.28*SEM)/√k for 80% LoA, where k represents the number of readers (equals 2 in this study).

To compare the B&A method with the ANOVA method, we excluded studies with only two time points (n=2) from this analysis.

Statistical analysis

For descriptive purpose, the values including the characteristics of all RCTs are presented as median (interquartile range -IQR-). All treatment arms were considered as one per trial. The variance components (including residual error) were estimated by three-way ANOVA, with change score between two time points per reader as the dependent variable, patient and reader as random factors, and time interval as fixed factor and all possible interactions (patient*reader, patient*time interval and patient*reader*time interval) were also included in the ANOVA to obtain the residual error components. Statistical analysis was performed using SPSS software version 18.0.

RESULTS

A total of 21,295 time points from 7,643 patients were included in the analysis. From all RCTs, two studies had two time points, 10 studies had three time points and three studies had four time points. The median (range) sample size of the studies was 517 (103-921) patients, and the number of time points within one reading session was two for 1,172 patients, three for 5,296 patients and four for 1,175 patients. The median (IQR) disease duration of patients included in the studies was 6 (3-7) years, and the median (IQR) baseline radiological damage and progression in SHS to last follow-up across all studies was 32 (18-48) and 1.1 (0.5-2.0), respectively.

Since the principal aim of this study was to propose a surrogate for the ANOVA method for calculating the SDC when more than two time points are scored within the same reading session, we evaluated the agreement between the two different methods (LoA method and ANOVA method) employing a B&A plot; this means plotting the difference between the methods against their mean as shown in Figure 1. The mean (range) SDC over the included studies based on the 95% LoA and ANOVA methods was 3.1 (2.3-4.3) and 3.2 (2.3-4.6) units,

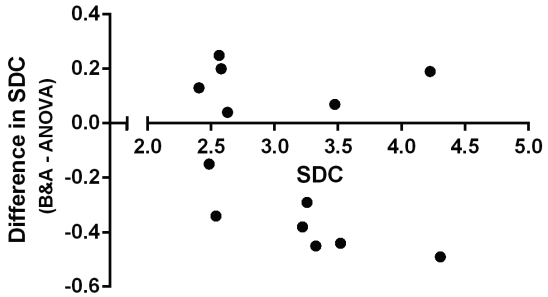

Figure 1: Difference in smallest detectable change (SDC) between the Bland & Altman (B&A) levels of agreement (LoA) method (mean of all intervals) and the analysis of variance (ANOVA) method (taking into account all intervals), for trials with two or more intervals.

respectively. The mean ± SD difference between the two methods was -0.13 ± 0.28, range (-0.48, 0.25) units. The mean of the SDC for all studies was somewhat higher for the ANOVA method (not statistically significant). No particular trend was observed, and therefore the difference between the two methods did not tend to get larger (or smaller) as the average discrepancy increased. The variability was also consistent along the range of observations (homoscedasticity of the scatter). Moreover, median values for the difference between the 95% LoA and ANOVA methods were higher in studies with less radiographic damage at baseline and less radiographic progression compared with studies with more radiographic damage and progression (-0.22 vs 0.07 and -0.22 vs 0.04, respectively), but no differences in the range was observed (supplementary Table S1).

Second, we compared the SDC based on the 95% LoA with the SDC based on the 80% LoA using the LoA method. Figure 2 shows the SDC values for all intervals and studies based on both LoAs. The median (range) difference between the 95% and 80% LoA SDCs was 1.1 (0.8-1.6) for the first interval, 0.9 (0.7-1.5) for the second interval and 1.3 (1.1-1.4) for the third interval. The mean ± SD SDC including all the SDCs calculated for all intermediate intervals (n=31) was 3.0 ± 0.7 for 95% LoA and 2.0 ± 0.4 for 80% LoA.

Finally, we also investigated if there was an association between baseline radiological damage and radiographic progression to last follow-up with the SDC. We did not observe any relationship between the degree of damage at baseline (SHS) and the SDC (r^2= 0.01, p=0.8) (Figure 3a), while an association between radiological progression and the SDC was obvious (r^2= 0.64, p<0.001) (Figure 3b), indicating that the SDC is higher in trials with more progression, although this relationship is strongly influenced by two trials with the highest progression rate.

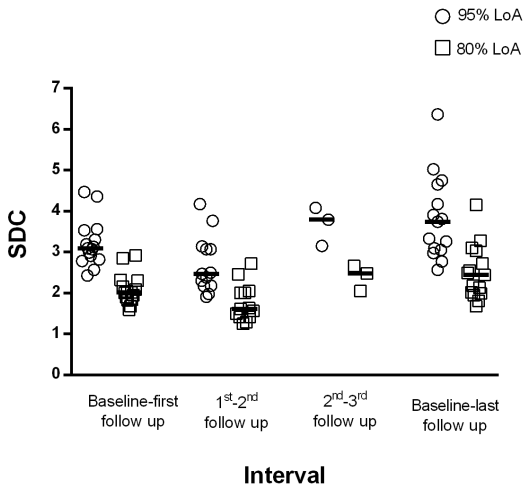

Figure 2: Smallest detectable change (SDC) for 95% and 80% levels of agreement (LoA) based on the LoA method.

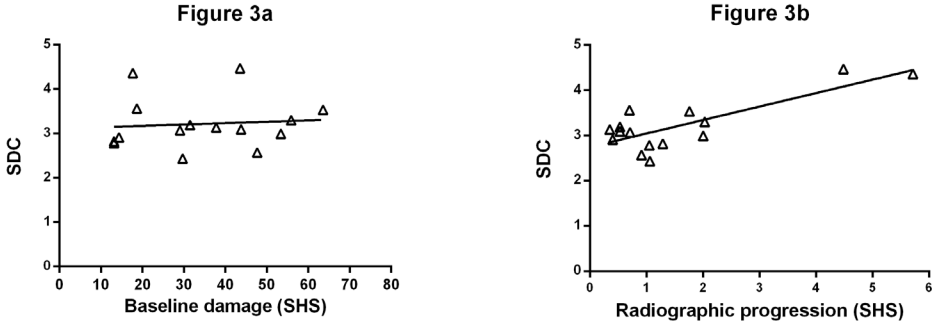

Figure 3: Association between baseline radiographic damage and radiographic progression with smallest detectable change (SDC) for 95% level of agreement. Regression lines summarize the association between radiographic damage (Figure 3a) and radiographic progression to last follow-up (Figure 3b) with the SDC. SHS: Sharp-van der Heijde score.

DISCUSSION

The results of this analysis suggest that, in complex databases with multiple time points and time intervals, the mean of the interval SDCs obtained by the simple LoA method is a valid surrogate for the 'umbrella SDC' obtained by complex methods based on the generalizability theory. Further, we have found arguments that the SDC is dependent on the level of radiographic progression in a trial rather than on the level of absolute damage. In addition, we here propose to consider the use of an SDC based on 80% rather than 95% LoA, for reasons explained below.

The maximum discrepancy in SDCs of 0.48 units when calculated by the two methods, and the systematic difference of only 0.13 units is negligible, in the light of the minimal clinically important difference (MCID) of 3-4.5 units for radiographic progression [10]. On the other hand, we should take into account the fact that this MCID cut-off was selected based on results of a study performed when biological agents were just entering clinical use and therefore when tolerance for progressive joint damage was less strict. However, we consider it very unlikely that an updated MCID would even approach 0.48 units.

Further, there was consistency in variability, and no particular trend was observed when the two methods were compared, which adds to the validity of the mean LoA method. Obviously, the mean LoA method has important advantages in that it is simpler, less time consuming, and more familiar to researchers.

With respect to the proposal to base SDCs on the 80% LoA, it can be argued that there is no solid scientific basis to choose a 95% instead of a lower LoA. The 95% cut-off level has its basis in distribution theory, where it is a boundary for including 95% of observations of a distribution with standard normal ('bell-shaped') properties (the mean ± 2 SDs), and was therefore probably chosen because it resembles the 95% confidence interval (CI) used in statistical hypothesis testing. Conceptually, though, CI and LoA are not related: whereas 95% CI statistically tests the null hypothesis that the mean difference in change scores obtained by two readers is zero, the 95% LoA quantifies the boundaries that include 95% of all paired observations and has nothing in common with hypothesis testing [11]. The justification for choosing boundaries other than 95% as LoA depends on the relevance of avoiding potential misclassifications. In radiographic analysis, the SDC concept is used to determine whether a patient is a 'true progressor' (i.e. progression beyond reasonable measurement error) or not (i.e. progression still compatible with measurement error, and therefore classified as zero progression). If an 80% LoA is accepted as the basis for the SDC, and the SDC is accepted as the level that distinguishes 'progression beyond measurement error' from 'progression still compatible with measurement error', more patients will be accepted as 'true progressors'. Obviously, there will also be some more 'progressors' for whom progression is due to measurement error, but this misclassification will affect both arms of an RCT in an unbiased

manner. Given the context of the RCT, in which a treatment is tested against a comparator for its potential to *avoid* radiographic progression, and the current mean progression scores observed in such trials, it is unlikely that a cut-off level based on an 80% SDC will spuriously influence the trial results. In fact, a trial with higher percentages of patients with progression per trial arm may provide increased conservatism, which is advantageous from the perspective of internal validity of a trial. In the light of the well-recognized phenomenon of deflating radiographic progression rates over time in clinical trials [12], increased rates of 'progressors' per trial arm using more lenient SDC cut-offs is advantageous for the statistical power of a trial. Increased misclassification is unlikely to be relevant here, since one may expect that these misclassifications will be evenly distributed among trial arms. However, the ultimate effect will depend on the analyses and the degree of misclassification.

It is therefore proposed to use 80% LoA-based SDCs instead of 95% LoA-based SDCs, so that measurement error is substantially lower than the change in radiographic damage that rheumatologists consider clinically relevant: approximately 3 units [10].

Another observation of note was that the degree of joint damage at baseline did not influence the SDC, whereas the level of radiographic progression did have a slight influence on the SDC, that is, the SDC tended to increase with increasing radiographic progression rates. The first observation is somewhat unexpected, since readers usually recognize unaffected joints relatively easily and in general achieve a high level of agreement, while they have to make far more decisions in case of multiple affected joints with different states of joint involvement. An explanation could be that the studies included in this analysis covered a relatively small range of potential involvement at baseline (10 to 65 units). SDCs may therefore still be relatively low if baseline joint damage is low to moderate, but increase if baseline damage exceeds 65 units. Moreover, trials with even lower baseline damage were not included and therefore baseline damage below 10 could not be tested. This study does not provide resolution for this. The second observation of increasing SDCs with increasing progression rates is in compliance with what has been found in detailed analyses. In a recent analysis of the TEMPO trial [13], with four independent reads of the same patient, a very high level of agreement was reached for the great majority of individual joints that showed zero progression in SHS. In contrast, agreement on a per joint basis was poor in those joints that were scored as 'progressive' by at least one of the four readers. This lack of agreement is lost when total per-patient scores are calculated, as is standard practice or in evaluation in RCTs, explaining increasing SDCs with increased progression rates. A limitation of this observation is that this positive correlation was largely determined by two trials.

A limitation of our study is that the maximum number of time points within the same reading session in RCTs included was four, so we do not know if these results would be applicable if five or more time points are present. However, not many clinical trials include more than four time points in one read campaign, and the question is therefore rather theoretical.

Importantly, all images were scored in unknown chronological order by experienced readers, and these results cannot be extrapolated to reads with known time order and to reads by inexperienced readers or those that have not been trained similarly. Although a recent study suggests that chronological reading is more precise than random reading [14], regulatory agencies still require radiographs to be scored randomly. Random scoring is therefore still considered the reference setting [15]. Moreover, although not tested, it may be assumed that the issues addressed in this paper are equally applicable to studies scored in chronological order, as the topics under investigation in this manuscript are not directly influenced by the (un)blinding of the time order.

In conclusion, for reasons of convenience, we propose to report the mean of all interval SDCs as an appropriate surrogate for the ANOVA-based SDC in trials with multiple time points. In addition, we consider an SDC based on an 80% LoA to be an acceptable alternative to an SDC based on a 95% LoA. For the SHS method, based on these large datasets involving many different readers, we propose a cut-off level of 3.0 units for a 95% LoA SDC and of 2.0 units for an 80% LoA SDC as the threshold for deciding if the RA of an individual patient shows radiographic progression.

REFERENCES

1. US Food and Drug Administration. Guidance for industry: clinical development programs for drugs, devices and biological products for the treatment of rheumatoid arthritis. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071579.pdf.)
2. Guideline on clinical investigation of medicinal products other than NSAIDs for treatment of rheumatoid arthritis. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/12/WC500119785.pdf.)
3. van der Heijde DM. Radiographic imaging: the 'gold standard' for assessment of disease progression in rheumatoid arthritis. Rheumatology (Oxford) 2000;39 Suppl 1:9-16.
4. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033-9.
5. van der Heijde D, Simon L, Smolen J, et al. How to report radiographic data in randomized clinical trials in rheumatoid arthritis: guidelines from a roundtable discussion. Arthritis Rheum 2002;47:215-8.
6. Bruynesteyn K, Boers M, Kostense P, et al. Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. Ann Rheum Dis 2005;64:179-82.
7. Bruynesteyn K, van der Heijde D, Boers M, et al. Minimal clinically important difference in radiological progression of joint damage over 1 year in rheumatoid arthritis: preliminary results of a validation study with clinical experts. J Rheumatol 2001;28:904-10.
8. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261-3.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
10. Bruynesteyn K, van der Linden S, Landewe R, et al. Progression of rheumatoid arthritis on plain radiographs judged differently by expert radiologists and rheumatologists. J Rheumatol 2004;31:1088-94.
11. Landewe RB, van der Heijde D. Principles of assessment from a clinical perspective. Best Pract Res Clin Rheumatol 2003;17:365-79.
12. Rahman MU, Buchanan J, Doyle MK, et al. Changes in patient characteristics in anti-tumour necrosis factor clinical trials for rheumatoid arthritis: results of an analysis of the literature over the past 16 years. Ann Rheum Dis 2011;70:1631-40.
13. Lukas C, van der Heijde D, Fatenajad S, et al. Repair of erosions occurs almost exclusively in damaged joints without swelling. Ann Rheum Dis 2010;69:851-5.
14. van Tuyl LH, van der Heijde D, Knol DL, et al. Chronological reading of radiographs in rheumatoid arthritis increases efficiency and does not lead to bias. Ann Rheum Dis 2014;73:391-5.
15. van der Heijde D, Boonen A, Boers M, et al. 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.

Supplementary Table S1: Characteristics of studies and difference in smallest detectable change between the Bland & Altman levels of agreement method (average of all intervals) and the analysis of variance (ANOVA) method (taking into account all intervals).

Study ID	Mean disease duration (years)	Study duration (weeks)	Baseline SHS (mean ± SD)	SHS progression to last follow up (mean ± SD)	Bland&Altman method	ANOVA method	Difference between two methods
2 Time points							
1	6	24	47.7 ± 55.0	0.91 ± 3.03	-	2.57	-
14	6	24	43.8 ± 54.2	0.53 ± 6.83	-	3.09	-
3 Time points							
2	6	24	53.4 ± 53.1	2.00 ± 4.91	2.69	2.44	0.25
3	7	48	31.5 ± 50.2	0.53 ± 3.59	2.68	2.48	0.20
5	12	48	63.6 ± 71.2	1.76 ± 5.35	3.30	3.74	-0.44
7	1	52	13.2 ± 26.8	1.05 ± 5.79	2.47	2.34	0.13
9	2	52	17.7 ± 20.1	5.71 ± 9.30	4.06	4.55	-0.48
10	6	52	18.7 ± 32.8	0.70 ± 4.03	3.03	3.41	-0.38
11	6	104	14.4 ± 27.0	0.40 ± 4.15	2.41	2.56	-0.15
12	8	104	29.7 ± 43.6	1.06 ± 4.88	2.37	2.71	-0.35
13	6	52	37.8 ± 46.4	0.35 ± 3.39	3.10	3.55	-0.45
15	3	52	43.6 ± 43.3	4.48 ± 10.27	4.32	4.13	0.19
4 Time points							
4	7	96	29.1 ± 48.0	0.71 ± 5.12	3.11	3.40	-0.29
6	12	96	55.9 ± 60.5	2.03 ± 7.00	3.51	3.44	0.07
8	1	108	13.2 ± 28.1	1.29 ± 7.05	2.65	2.61	0.04