Robustness and Generalizability of Smallest Detectable Difference in Radiological Progression

MARISSA N.D. LASSERE, DÉSIRÉE van der HEIJDE, KENT JOHNSON, KARIN BRUYNESTEYN, ESMERALDA MOLENAAR, ANNELIES BOONEN, ARCO VERHOEVEN, PAUL EMERY, and MAARTEN BOERS

ABSTRACT. The smallest detectable difference (SDD) reflects that component of a measure statistically attributable to error from the measurement process itself. As such it is an irreducible component of the inherent variability in measurements in clinical trials and will affect their design, whether randomized or observational. Even though the application of the SDD concept to assaying radiographs in rheumatoid arthritis is relatively new and not well understood, systematic work on the influences of radiographic SDD can be done. This report describes the effects of a number of clinical aspects of the disease and operational aspects of trials on the values of the SDD of radiographic progression data. We show that if conditions affecting SDD are known and kept constant across datasets, the SDD of radiological progression from one study may be generalizable to other studies. However, if any one condition varies, the SDD is distinctly unrobust and cannot be generalized to other studies. (J Rheumatol 2001;28:911–3)

Key Indexing Terms:
MINIMAL CLINICALLY IMPORTANT DIFFERENCE RELIABILITY
MEASUREMENT ERROR RADIOGRAPHIC SCORING RHEUMATOID ARTHRITIS

INTRODUCTION

The initial work on minimum clinically important differences (MCID) in radiological measures in rheumatoid arthritis (RA) was begun at OMERACT 4[1]. Here, as elsewhere in this conference, it was agreed that, prior to the existence of data to justify an MCID determination demonstrated by predictive data ("What change in the radiographic progression translates into a future change of an important clinical outcome?"), proxy MCID could be formulated with a so-called distribution based method[2,3]. The most intuitively attractive distribution based concept is the smallest detectable difference (SDD), defined as that amount of difference for which anything smaller cannot be reliably distinguished from random error in measurement. The SDD can be determined using the 95% confidence interval of the standard deviation of the error of 2 (or more) measurements (limits of agreement method of Bland and Altman[4]).

Measurement error in radiographic progress in RA depends on multiple factors. There are aspects intrinsic to the measurement process itself, such as the particular scoring method used, and whether the reader is experienced or naïve. If the intent is to generalize to a population of readers, additional uncertainty is injected. Further, clinical features external to measurement also affect radiographic scores and so will affect measurement error. These include the baseline damage seen on the radiograph, and the numerous covariates that affect the rate of radiographic progression. Therefore, the SDD for radiographic progression is condition- and context-specific.

Given the above, the question arises as to how robust the radiological SDD is to changes in these factors? Which influence the SDD most, and, most importantly, is an SDD from one study generalizable to other studies? This article presents the results of several small studies conducted to examine the effect of conditions external to measurement — baseline damage, rate of progression, and disease activity, and a condition intrinsic to measurement process — different readers — on the robustness, and thus potential generalizability, of the SDD of radiographic progression.

From the Department of Rheumatology, St. George Hospital, Sydney, Australia; Department of Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands; Limburg University Center, Diepenbeek, Belgium; Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, USA; Department of Rheumatology and Department of Clinical Epidemiology, VU University Hospital, Amsterdam, The Netherlands; and Rheumatology Research Unit, University of Leeds, Leeds, UK.

M. Lassere, MB BS, PhD, FAFPHM, FRACP, Staff Specialist in Rheumatology, Department of Rheumatology, St. George Hospital; D. van der Heijde, MD, PhD, Professor of Rheumatology, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and Limburg University Center; K. Johnson, MD, Medical Officer, Center for Drug Evaluation and Research, Food and Drug Administration (views expressed are not necessarily those of the FDA); K. Bruynesteyn, MD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht; E. Molenaar, MD, Department of Rheumatology, VU University Hospital; A. Boonen, MD, Rheumatologist, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht; A. Verhoeven, MD, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht; P. Emery, MA, MD, FRCP, ARC Professor in Rheumatology, Rheumatology Research Unit, University of Leeds; M. Boers, MSc, MD, PhD, Rheumatologist, Professor of Clinical Epidemiology, Department of Clinical Epidemiology, VU University Hospital.

Address reprint requests to Dr. M. Lassere, Staff Specialist in Rheumatology, Department of Rheumatology, St. George Hospital, Kogarah, NSW 2217, Australia. E-mail: lasserem@sesahs.nsw.gov.au

METHODS

Selection of patients. The design and characteristics of 5 studies are shown in Table 1. The first 4 (datasets 1–4) were one year, randomized trials of patients with early, active RA[1,5,6]. Two experienced readers were used in each, reading films in paired, chronological order. Five experienced readers (A to E) were used in total, distributed 2 per study as shown in Table 1. Dataset 4 used a 2 by 2 factor design, wherein patients were selected to reflect the spectrum of baseline damage score (factor 1) and the spectrum of radiographic progression over 12 months (factor 2)[1]. The last study (dataset 5) comprised patients with longer duration RA, recruited from an observational cohort of patients in remission by American College of Rheumatology (ACR) criteria (unpublished data). The radiographs for dataset 5 were also read by 2 experienced readers (D and E), in paired, but now random order. Radiographs in all studies were scored using the Sharp-van der Heijde method (0–448 scale)[7].

Statistical analysis. Bland and Altman's 95% limits of agreement method was used to estimate the SDD of radiological progression[4,8]. The mean ($mean_{diff}$) and standard deviation (SD_{diff}) of the paired differences of radiological progression between readers were calculated. Because radiological progression was calculated in all studies using the mean score of 2 readers, one must first divide the SD of the paired differences by the square root of 2. The confidence interval (CI) for the difference between measurements on the same subject is estimated by $mean_{diff} - t_{0.05,\ n-1} * (SD_{diff}/\sqrt{2})$ and $mean_{diff} + t_{0.05,\ n-1} * (SD_{diff}/\sqrt{2})$, where $t_{0.05,\ n-1}$ is the critical point of the t distribution. This CI is the 95% limits of agreement as defined by Bland and Altman, and is the SDD of radiological progression for the mean score if the same 2 readers are used in all studies. However, if different readers are used across studies, then the 95% upper and lower limits of agreement in turn have their own 95% CI, and these wider limits must be used to determine the SDD of radiological progression for the mean score of *any* 2 readers. The 95% CI around the upper and lower limits were determined first by calculating the standard error (SE) of the limits of agreement, SE_{limits}, which is about $\sqrt{[(\{3*(SD_{diff}/\sqrt{2})\}^2)/n]}$. Therefore the 95% CI for the lower limit of agreement is: $mean_{diff} - t_{0.05,\ n-1} * (SD_{diff}/\sqrt{2})$ $(t_{0.05,\ n-1\ (df)}\ (SE_{limits})$, and the 95% CI for the upper limit of agreement is: $mean_{diff} + t_{0.05,\ n-1} * (SD_{diff}/\sqrt{2})$ $(t_{0.05,\ n-1\ (df)}\ (SE_{limits})$.

RESULTS

Table 2 gives the SDD of radiological progression for the Sharp-van der Heijde method for each of the studies. In dataset 1 and 2 the following conditions were constant: distribution of disease progression, distribution of disease duration (< 12 mo), and distribution of disease activity (active RA, by ACR criteria). Further, the same 2 readers, both experienced, read all radiographs in paired, chronological order. Given these constants between datasets 1 and 2, one would surmise the SDD for radiological progression to be very similar. Indeed, this proved to be the case: 4.9 and 4.7 for the "same 2 readers" SDD, and 7.1 and 6.6 for "any 2 readers" SDD (see Table 2).

Dataset 3 differs from datasets 1 and 2 with the replacement of reader B with reader C; otherwise, conditions were held constant. This maneuver alone was associated with an increase of the SDD to 8.8 and 11.0, reflecting the greater lack of agreement of readers A and C in dataset 3 compared to readers A and B in datasets 1 or 2.

Dataset 4 used the same readers as dataset 3, but now features external to measurement, the underlying distributions of baseline damage and of radiological progression, have changed. Unlike all other studies using patients without regard to baseline damage and progression, patients in study 4 were selected to ensure the distributions of these factors were representative. The resulting SDD for the mean score are 11 (for the same 2 readers) and 16 (for any 2 readers); both results are more than twice the respective SDD for dataset 1 and 2 and about 25% larger than SDD for dataset 3.

Finally, dataset 5 differs from the first 4 datasets on at least 4 conditions — the source population, patient disease activity (disease remission at baseline), 2 new readers, and a new radiograph order (random, not chronological, see Table 1). Therefore one would expect cross-study comparisons to be more problematic. Disease remission should reduce radiological progression, and therefore decrease the SDD. However, the films being read in paired but random order could cancel this effect on the SDD, since random reading is more likely to add to measurement error. The similarity of the SDD for dataset 5 and those for datasets 1 and 2 (see Table 2) is likely fortuitous and would not be predictable.

DISCUSSION

In this brief report we show that conditions external to the

Table 1. Description of study populations and methods.

Dataset	Source of Patients	RA Status	Distribution: Baseline Damage and Progression	Readers	Median Disease Duration	Order of Readings	No. of Pairs of Films	Interval
1	RCT	Active	Unselected	A and B	12 mo	Paired, chronological	46	1 year
2	RCT	Active	Unselected	A and B	8 mo	Paired, chronological	61	1 year
3	RCT	Active	Unselected	A and C	4 mo	Paired, chronological	135	1 year
4	RCT	Active	Selected to maximize	A and C	6 mo	Paired, chronological	52	1 year
5	Cohort study	Remission	Unselected	D and E	7 yrs	Paired, random	112	1 year

RCT: randomized controlled trial.

Table 2. Smallest detectable difference of radiological progression: van der Heijde-Sharp method.

Dataset	SDD, mean score of same 2 readers	SDD, mean score of any 2 readers
Dataset 1	± 4.9	± 7.1
Dataset 2	± 4.7	± 6.6
Dataset 3	± 8.8	± 11.0
Dataset 4	± 11.0	± 15.5
Dataset 5	± 5.0	ND

ND: not determined.

measurement process as well as intrinsic factors influence the magnitude and direction, and thereby the variability, of the smallest detectable difference of radiological progression. The SDD differs across datasets, sometimes by more than 2-fold, and these differences are probably important. Further, other differences may be missed because they cancel each other out. Finally, there may be other factors such as the total number of readings, the division between repeated readings by one observer, and use of multiple readers that would also be expected to influence the variability of the SDD. These are topics of future research.

In this report we were not able to rank the conditions that influence the magnitude of the SDD, nor even predict their direction. Generally, one would expect that experience and calibration between readers would reduce the SDD, as would reading the radiographs in paired chronological order[5]. However, because in most systems measurement error usually increases monotonically with the underlying quantity measured[9,10], patients with longer disease duration, more active disease, more baseline damage, and more radiological progression per year would tend to show larger SDD.

If all conditions affecting SDD are known and kept constant across datasets, then the SDD of radiological progression from one study may be generalizable to other studies. However, if any one condition varies, the SDD is distinctly unrobust and cannot be generalized to other studies. We recommend that the SDD be determined as a part of the analysis for all studies. Although the SDD is study-specific, it will be a useful tool in assisting in the interpretation of the results.

REFERENCES

1. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
2. Lassere M, van der Heijde D, Johnson K. Foundations of the minimal clinically important difference for imaging. J Rheumatol 2001;28:890-1.
3. van der Heijde D, Lassere M, Edmonds J, Kirwan J, Strand V, Boers M, for the OMERACT Imaging Task Force. Minimal clinically important difference in plain films in rheumatoid arthritis: group discussions, conclusions and recommendations. J Rheumatol 2001;28:914-7.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
5. 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 1999;38:1213-20.
6. Proudman SM, Conaghan PG, Richardson G, et al. Treatment of poor-prognosis rheumatoid arthritis: A randomized study of treatment with methotrexate, cyclosporin A and intraarticular corticosteroids compared with sulfasalazine alone. Arthritis Rheum 2000;43:1809-19.
7. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 1999;26:743-5.
8. Bland JM, Altman DG. Measurement error. BMJ 1996;313:744.
9. Lassere M. Pooled metaanalysis of radiographic progression: comparison of Sharp and Larsen methods. J Rheumatol 2000;27:269-75.
10. Healy MJR. Measuring measuring errors. Stat Med 1989; 8:893-906.