REPORT

Imaging: do erosions heal?

D van der Heijde, R Landewé

Ann Rheum Dis 2003;**62**(Suppl II):ii10–ii12

It was shown previously that reparative changes in erosions can be seen in individual joints and individual patients. Whether repair may occur at a group level, and can be induced by treatment, is not known. This manuscript describes a means of visualising data obtained in a clinical trial by the use of probability plots to better understand the results. These probability plots give a good insight into the coherence of the data. They can also be used to make the interpretation of repair at a group level easier. Probability plots also explain the hazard of using binomial cut off points to compare treatment effects. The interference of true repair with measurement error is demonstrated. Repair at a group level is suggested if the mean progression score is statistically significantly different from zero, which can be visualised by a 95% confidence interval of this mean change score below zero. Application of this technique may give us better information on the effects of new drugs on the induction of repair.

Why is it important to answer the question, do erosions heal? Firstly, it would be valuable in helping us to better understand the pathophysiological processes in rheumatoid arthritis (RA). Secondly, it is suggested that repair can occur only when disease activity is absent (over the long term). Therefore, signs of repair may represent absence of inflammation. However, it would only be important from a patient's perspective if there was a correlation with final outcome: patients who show repair having a better course than those patients in whom the progression of structural damage has stopped. It would be particularly important if the ability to induce repair could differentiate between the efficacy of different drugs. Currently, it is unclear how long a follow up should be to be able to show repair. It is assumed that there should be complete absence of inflammation for a reasonably long time before this can occur. On the other hand, if negative scores in clinical trials are really an expression of repair, this can already be seen after six months of follow up. Moreover, different lengths of follow up may be needed in various phases of disease. In early disease it may be easier to see repair than in joints with longstanding inflammation and damage.

There is indeed evidence that bony erosions in RA occasionally show some degree of repair. The support for this lies in individual case reports and small series, and in supportive studies.[1–6] One of these studies was undertaken by the OMERACT study group on imaging and included a group of experts in scoring of radiographs in RA.[6] The study showed that the experts agreed on which films showed the smallest erosion, which was seen in about half of the second films taken, but they were unable to put the films in the correct time sequence. Moreover, there was no agreement on the features which are thought to represent repair, such as cortication, sclerosis, filling in, remodelling, and restoration. Explanations for this may be that these features were not present in the films included in the study; that the experts were not appropriately trained to observe these features; or, indeed, that these features are not as specific as previously assumed.

All the present evidence is based on individual joints in individual patients. Little information exists on the occurrence of repair on entire films of hands and feet (for example, does repair and progression in joints of the same patient occur simultaneously?). Negative scores in clinical trials may indicate that repair may at least dominate progression in individual patients, although this conclusion cannot be drawn without taking into account measurement error. Measurement error is a well known phenomenon in scoring radiographs, and therefore it is accepted that, for clinical trials, films should be scored by two observers in order to get a better impression of the true score.[7]

Moreover, when a randomised clinical trial is analysed the null hypothesis is that there is no difference in change of structural damage between the two treatment arms. The primary analysis is a between-group analysis. Treatment effect is the only subject of interest, and measurement error is assumed to be similar in both groups and therefore irrelevant. Change in structural damage may represent both progression and repair, but to examine the question whether repair really occurs in a group of patients, a within-group analysis is appropriate. In this case, measurement error has a significant role.

It is well known that in clinical trials, in groups of patients in general, no change in structural damage occurs in a large proportion of patients and only a small fraction of patients show substantial progression. Therefore, the appropriate way of presenting the data is as box and whisker plots or as medians with various centiles (usually 25 and 75). The disadvantage is that a lot of information on the data is lost, as these values only relate to the localisation in the distribution. The presentation of radiographic data as means, with a measure to address the variation (SD, SE or confidence interval (CI)), includes all data, but is strongly influenced by subtle changes at the upper extremes.[8] Another way is the presentation of the percentage with a change above or below a certain cut off point. However, the result is largely dependent on the chosen cut off value, and can therefore be influenced by the analysts, as we will show below.

PROBABILITY PLOTS

To overcome the problems discussed above, cumulative probability plots can be used as a means to better visualise all data and consequently make the results easier to interpret correctly (Landewé R and van der Heijde D, unpublished data). A cumulative probability plot is a frequency distribution that plots the observed cumulative proportion (scores ranked from the lowest to the highest values, and presented as a cumulative proportion of all scores) on the x axis against the variable's actual values. In that case, the cumulative probability indicates the proportion of observations with a value less

Abbreviations: AUC, area under the curve; CI, confidence interval; RA, rheumatoid arthritis; SDD, smallest detectable difference

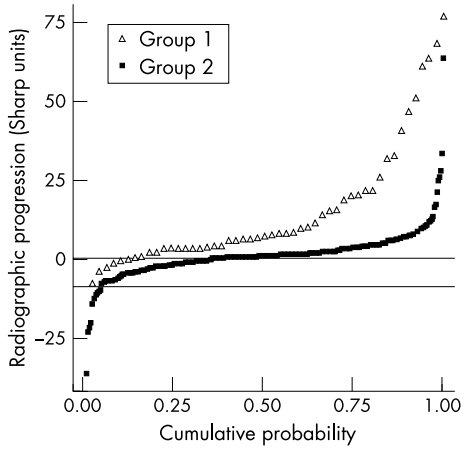

Figure 1 Probability plot of change scores of radiographic damage of two drugs. The x axis shows the cumulative probability and the y axis the actual values. Every single observation is plotted in the curve. The horizontal lines represent a possible binomial cut off point (change of 0 and change of –7).

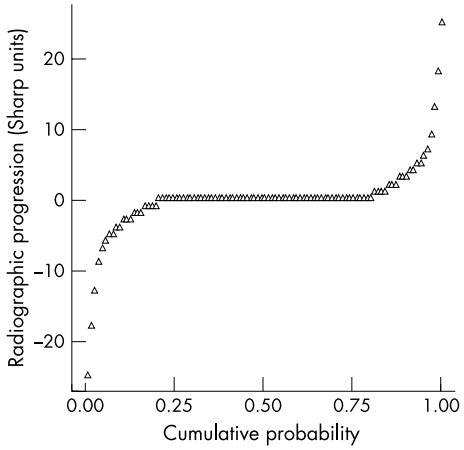

Figure 2 The entire group in reality shows no progression, but there is random measurement error (films scored with the time sequence concealed). The mean change score is 0 (95% CI –1.06 to 1.06).

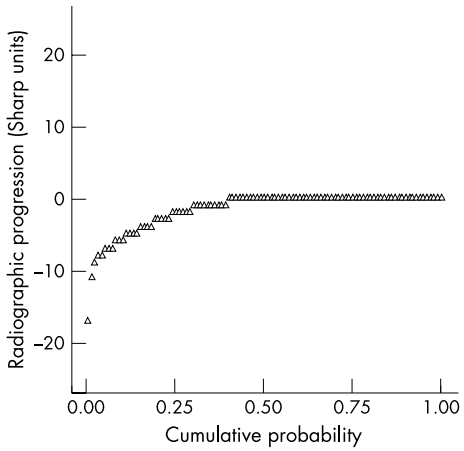

Figure 3 Forty per cent of the patients show true repair and no measurement error is present. The mean change score is –1.62 (95% CI –2.30 to –1.04).

than the value of the corresponding variable. Figure 1 presents an example of a probability plot of two drugs. The values of both drugs belonging to the various centiles such as median (50 centile) and 25 and 75 centiles can easily be deduced from the plot by drawing a vertical line at the centile and connecting this to the corresponding value on the y axis. It is obvious that the cumulative probability curves are not entirely "smooth", and the space between both curves, which is an indication for the treatment contrast, varies along the axis of cumulative probability. As mentioned earlier, this irregularity is important in choosing a binomial cut off level for radiographic progression to describe the magnitude of the treatment effect. The probability curves show us that the choice of the cut off level is relevant for the magnitude of the treatment contrast. As a consequence, an optimal cut off level (that is, that which provides the highest contrast) can easily be constructed by the investigator, as we already have shown previously.[8] Figure 1 gives an example of two cut off levels (change of ≤ 0 and change of ≤ -7) which give a completely different treatment contrast.

REPAIR OF EROSIONS

Repair at a group level scientifically refers to the question of whether the null hypothesis of no true change over time can be rejected by finding a proportion of patients with negative progression scores that cannot solely be attributed to measurement error and/or occasional repair phenomena. We will show here by using simulated data how cumulative probability plots can help in answering this question.

Suppose an observational study includes 100 patients with RA in whom there is no true radiographic progression after one year of follow up. Radiograms, at baseline and after one year, are scored by two readers, with concealed time order, and a progression score is calculated. Under ideal circumstances (no measurement error, no variability in readings) a progression score of zero will be assigned to all patients in such an imaginary situation, resulting in a flat probability plot: all dots are on the y=0 line.

In reality, measurement error and other sources of variation will be operative, and will be two sided because of concealment of the reading sequence. Figure 2 shows a cumulative probability curve under the null hypothesis of no progression; a proportion of change scores will deviate from zero owing to error. Under the null hypothesis, the effect of measurement error (and other sources of variation) is best reflected by the area under the probability curve (AUC): a sum of the positive AUC on the right side of the curve and the negative AUC on the left side of the curve. Note that the AUC is zero here, because of the symmetry of the curve. Note also that the area under the probability curve is mathematically identical to the mean value, calculated as the quotient of the sum score of all observations and the number of observations.

Now suppose that, under the assumptions of two sided symmetrical measurement error and independence of error and true signal, there is true repair in 40% of the patients. If there were no measurement error operative, the probability curve might be as in fig 3: 60% of patients with values equal to zero, many patients with small negative values, and a few patients with highly negative values.

We simulated the aggregated effects of "true repair" and all sources of measurement error by randomly matching the imaginary repair scores against the imaginary error scores obtained under the null hypothesis, and plotting the resultant scores in a probability plot again (fig 4). The resulting curve is, as expected, translated to the right, with a greater negative AUC on the left part of the curve (and a greater proportion of negative scores), and a smaller positive AUC on the right part of the curve (and a smaller proportion of positive scores). The extreme values, however, are similar to the values obtained by measurement error alone. As a result of this translation, the mean progression (the AUC of the entire curve) becomes

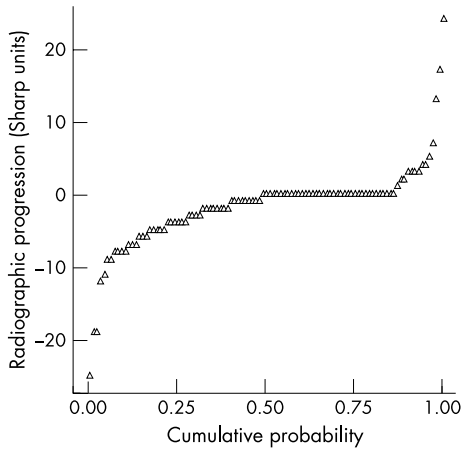

Figure 4. Forty per cent of the cases showing true repair combined with measurement error, by randomly matching the putative true repair scores with putative error scores. The mean change score is –1.7 (95% CI –2.9 to –0.52).

negative, and the statistical confirmation for repair at a group level simply follows a statistical test of the null hypothesis (*t* test for paired observations or Wilcoxon's signed rank test depending on the skewness of the data), which is that the mean change is zero. In this particular example, the mean change from baseline is –1.7 Sharp units, with a 95% CI from –2.9 to –0.52.

Note that presentation of centiles does not give justice to the phenomenon of repair here: the median score and 75 centile are (still) zero Sharp units; only the 25 centile indicates a negative change. The presentation of means and standard deviations show a mean change of –1.7 Sharp units, which may easily be considered as clinically unimportant, and may lead to a neglect of information about the repair in 40% of the patients. The probability plots, however, better visualise the negative scores, and immediately make clear that these negative scores occur in 40% of the patients, thus explaining why the median score is still zero. They also show the symmetry, and thus the relation between the impact of the negative scores and that of the positive scores, giving information on the coherence of the data.

No one who has studied the probability curves of radiographic progression scores will dispute any more the view that it is tricky to adopt a cut off level of zero (or 0.5 if the average of two readers is used) (fig 1) for differentiating between patients with and without progression, an issue that we recently encountered in a published meta-analysis on the efficacy of disease modifying antirheumatic drugs in slowing radiographic progression.[9]

Earlier, we advocated the concept of the smallest detectable difference (SDD) beyond measurement error as a minimum cut off level for distinguishing between patients with and without radiographic progression.[7 10] The SDD can be easily plotted in the probability curve, and the consequence of doing so is obvious at first glance. It is easy to see whether the SDD cut off point is a conservative cut off point with respect to the treatment difference, and one can find out the implications of different cut off levels immediately. The disadvantage of the SDD indeed is that it chooses only one cut off point. Often it is a conservative cut off point indicating only those patients with high scores as showing progression or repair. Probability plots do not replace Bland and Altman plots.[11] The latter are useful in determining an important source of measurement error: interreader variability. Probability plots of change scores aggregated from two or more readers do not provide an insight into this factor. Rather, they visualise the entire level of measurement error, as shown.

Cumulative probability plots are an aid in the explorative analysis. They certainly do not replace statistical testing, and should only be used as an adjunct to formal hypothesis testing. However, they may give useful information if in a comparative clinical trial a between-group difference appears not to be statistically significant. They can help in interpreting type II error as the cause when a trend is not found to be statistically significant.

It would be intriguing to see the results of the trials published in recent years on the effects of new treatments, including biological agents, presented as probability plots. It would give us more insight into the meaning of negative scores present in these studies. A mean negative progression with the entire 95% CI below zero would suggest repair at a group level. It is important to realise that a 95% CI including zero does not rule out individual cases of repair.

Authors' affiliations

D van der Heijde, R Landewé, University Hospital Maastricht, Department of Internal Medicine/Rheumatology, and Care and Public Health Research Institute (CAPHRI) University Maastricht, Maastricht, The Netherlands

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

REFERENCES

1 **Moeser PJ**, Baer AN. Healing of joint erosions in rheumatoid arthritis [letter]. Arthritis Rheum 1990;33:151–2.
2 **Menninger H**, Meixner C, Sondgen W. Progression and repair in radiographs of hands and forefeet in early rheumatoid arthritis. J Rheumatol 1995;22:1048–54.
3 **Rau R**, Wassenberg S, Herborn G, Perschel WT, Freitag G. Identification of radiologic healing phenomena in patients with rheumatoid arthritis. J Rheumatol 2001;28:2608–15.
4 **Sokka T**, Hannonen P. Healing of erosions in rheumatoid arthritis. Ann Rheum Dis 2000;59:647–9.
5 **van der Heijde D**, Sharp JT, Rau R, Strand V. OMERACT Workshop: repair of structural damage in rheumatoid arthritis. J Rheumatol 2003;30:1108–9.
6 **Sharp JT**, van der Heijde D, Boers M, Boonen A, Bruynesteyn K, Emery P, *et al*. Repair of erosions in rheumatoid arthritis does occur. Results from 2 studies by the OMERACT Subcommittee on Healing of Erosions. J Rheumatol 2003;30:1102–7.
7 **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.
8 **Landewé RB**, Boers M, van der Heijde DM. How to interpret radiological progression in randomized clinical trials? Rheumatology (Oxford) 2003;42:2–5.
9 **Jones G**, Halbert J, Crotty M, Shanahan EM, Batterham M, Ahern M. The effect of treatment on radiological progression in rheumatoid arthritis: a systematic review of randomized, placebo-controlled trials. Rheumatology (Oxford) 2002;41:6–13
10 **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.
11 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.