Tree-Structured Reliability Analysis for Magnetic Resonance Imaging (MRI) Data

Ruji Yao[1], Tulin Shekar[1], Hanzhe Zheng[1], Norman Y. Yao[2]

[1]Merck Research Laboratory
2015 Galloping Hill Road, K15-2-2435, Kenilworth, NJ 07033
[2]Department of Physics, Harvard University, Cambridge, MA 02138

ruji.yao@merck.com; tulin.shekar@merck.com;hanzhe.zheng@merck.com;
nyao@fas.harvard.edu

Abstract

Inter-rater reliability refers to a comparison of scores assigned to the same target by two or more raters. The simplicity of the two-way fixed effect model has rendered it a popular method for reliability estimation. On the other hand, "trees" – a class of non-parametric methods used to break partition-able data into several pieces (nodes), allows each node to then be fit with most suitable method. Here, we present an intuitive tree-based approach for reliability estimation. In a recent clinical trial, we used MRI data to evaluate the treatment effect on subjects with active axial spondyloarthritis. Each MRI slide was scored by two independent raters and the average of the two scores was used as an endpoint. In a routine reliability check, we noticed several intuitively incorrect reliability results as compared with the raw data. Motivated by several tree-based methods, we partition the response data to create a simple 2-node tree; we then combine the results with a modified reliability formula. With this new formula, the reported reliability scores follow intuitively from the raw data and also provide additional insight into the source of the variance of the MRI data itself.

Key words: Reliability, Regression Tree, Magnetic Resonance Imaging

1. Introduction

Statistically gauging the agreement achieved when multiple raters describe an object of analysis is an important cornerstone in a wide-range of subjects ranging from psychometrics and survey analysis to image recognition and clinical trials [1-5]. In the case of clinical trials, the data is often a sequence of repeated measurements and inter-rater disagreements can be caused by variations in procedure, interpretation of results and data presentation [6-9]. The character and distribution of such differences are often themselves quite different and may further depend on interested endpoints, disease, treatment, subjects selected and stage. Common procedure is to utilize an average of multiple measures as the endpoint for analysis and generically, a routine quality control procedure is set up to monitor inter-rater reliability.

To this end, there exist a number of statistics, which are often used to characterize reliability; the richness of this variety, which includes Krippendorff's Alpha, Cohen's kappa, Fleiss' kappa, joint-probability agreement, concordance correlation coefficients, and many others, is a testament to the diverse array of situations that can arise when attempting to quantify rater agreement [1,2,5,10,11]. Here, we describe and analyze a novel inter-rater reliability test based upon non-parametric trees [12-15]. Specifically, we consider a recent clinical trial in patients with early active axial spondyloarthritis, a disease characterized by chronic low back pain and which precedes the development of radiographic sacroiliitis [16].

The effects of treatment are assessed by analyzing the improvement in active inflammatory lesions as seen on magnetic resonance imaging (MRI) scans of the spine and sacroiliac joints.
The duration of the trial was 52 weeks and consisted of a 28-week treatment phase (156 randomized subjects) and a second follow-up phase from week 28 to week 52; this second phase included only patients (80 randomized subjects) who met the ASAS partial remission criteria at week 28. The MRI data are collected at baseline, week 28 and week 52. Subjects must have a non-zero inflammatory lesion score at baseline to be eligible for enrollment. Table 1 below depicts the data collected for each subject at baseline, week 28 and 52 (indexed by location and vertebral unit).

Table 1. Berlin MRI Scoring assessment

Spine on 23 vertebral units each score = 0, 1, 2 or 3 0 = no lesion 3 = very severe lesion	Location	# of units	Score Range
	Cervical	7	0 to 21
	Thoracic	11	0 to 33
	Lumbar	5	0 to 15
	Overall Spine	23	0 to 69
Sacroiliac (**SI**) Joints 4 locations on each side each score = 0, 1, 2 or 3	Left SI	4	0 to 12
	Right SI	4	0 to 12
	Overall SI	8	0 to 24

Each MRI slide is scored by two independent raters who are given the same rating guidelines and criteria; moreover, the raters must first pass an initial consistency test with sample slides. Then, each rater is given 2 slides, either baseline / week 28 or week 28 / week 52, from the same patient. The time point of the slides is blinded so the rater does not know which particular slide precedes treatment. Finally, since weak 28 is included in both studies, such slides may be scored twice by the same rater, in principle enabling an internal consistency check. Each time, a rater will report 14 scores on 2 slides, 7 endpoints on each slide. Thus, the total numbers of scores from a rater will be 156*14 + 80*14 = 3304 for 156 subjects in the treatment phase and for 80 subjects in follow-up phase. Combining these yields the total number of scores from 2 raters as 6608. We note that the actual number used in our study is 6184 (3092 pairs) since some slides are deemed unreadable.

In this study, we focus on the agreement issue between 2 raters scoring the same MRI slide. Figure 1 below shows the histogram of the raw MRI data and evinces the fact that the data is strongly skewed to the right. Similarly, Figure 2 depicts the histogram of differences between the MRI score (rater 1 minus rater 2)

Figure 1. Histogram of MRI scores in log10 scale

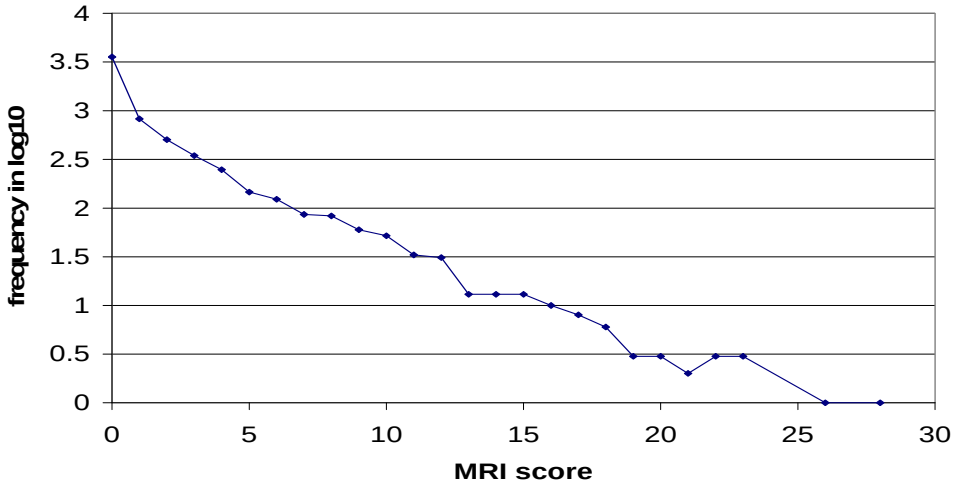
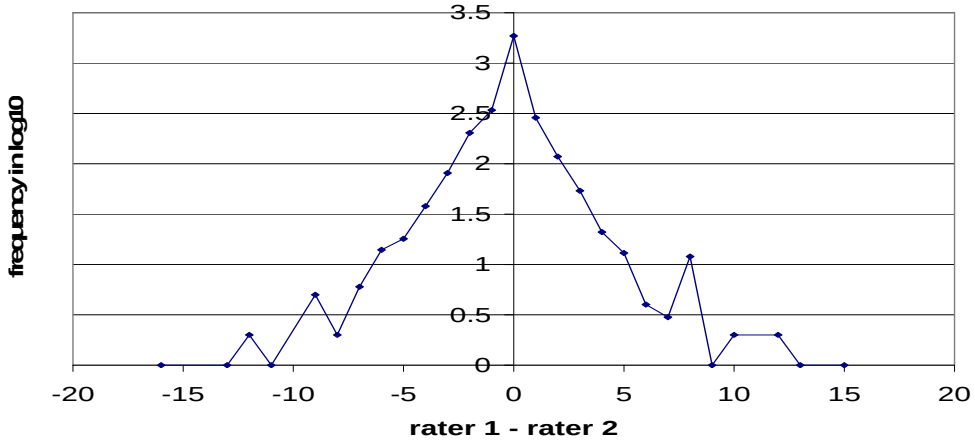


Figure 2. Histogram of differences of MRI scores in log10 scale (rater 1 minus rater 2)



The strong skewness of Figure 1 is common for subjects with moderate-to-severe active axial spondyloarthritis. Furthermore, the distribution of differences between the two raters is qualitatively symmetric, implying that two raters are scoring consistently over all MRI slides.

Reliability measure:

We now turn to a description of the simple two-way fixed effect model, which is commonly employed in such clinical trials [6]. For n MRI slides, we let $R1_i \ and \ R2_i \quad i=1,..,n$ be 2 n scores from 2 raters. Thus,

$M_i = (R1_i + R2_i)/2, \ \ i=1,..,n$ are the endpoints used for our analysis.

$SSE = \sum SSE_i = \sum \{(R1_i - M_i)^2 + (R2_i - M_i)^2\}$ is then the measure of disagreement between 2 raters on n slides, while

$SSM = 2\sum_i (M_i - \overline{M})^2$ is the measure of the mean variance over n slides.

With an ANOVA model, the reliability is defined [6] as: $reliability = 1 - \frac{SSE}{SSM}$.

Here, it behooves us to briefly consider what we ought to expect from the reliability analysis on the 28 endpoints listed in Table 1. In particular, we expect three key factors to determine the reliability measure:

1. Raters have the same training for scoring and score all slides consistently
2. The quality of MRI slides are consistent over time
3. Proper reliability statistics are chosen to summarized the reliability/agreement.

If all 3 factors are satisfied, then reliability results should merely fluctuate around a constant, as expected for a set of normal quality control data (taken over time). In Table 2, we present the reliability as obtained from the naïve two-way fixed effect model.

Table 2. Reliability based on above ANOVA model (1 – SSE/SSM)

	Reliability (treatment phase)			Reliability(follow-up phase)	
LOCATION	Baseline (1)	Week 28 (2)		Week 28 (3)	Week 52 (4)
Spine Overall (SO)	92%	75%		62%	90%
Spine Cervical (SC)	75%	45%		0%	56%
Spine Thoracis (ST)	92%	75%		79%	88%
Spine Lumbar (SL)	86%	76%		58%	85%
Overall SI (OS)	92%	85%		86%	92%
Left SI (LS)	92%	85%		84%	90%
Right SI (RS)	93%	82%		83%	90%

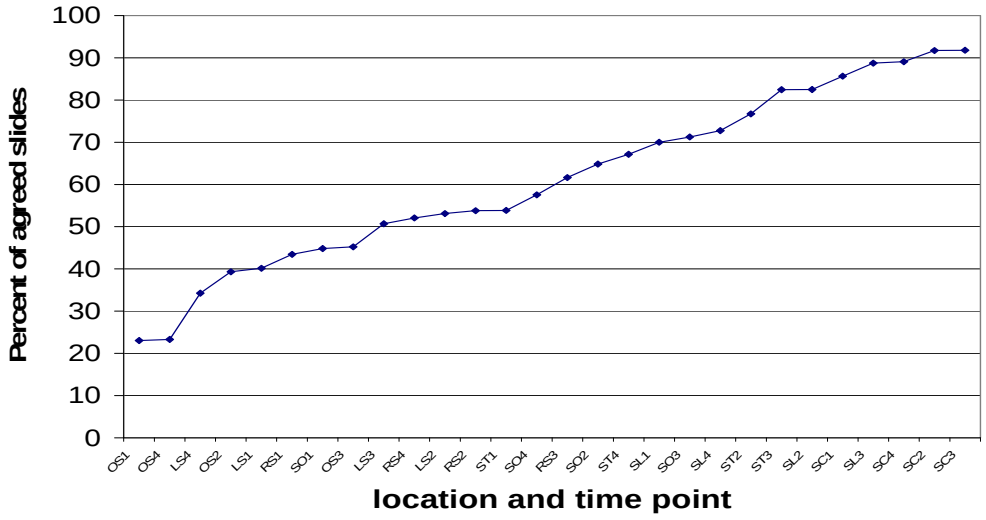
We notice that for the (follow-up) spine-cervical endpoint at week 28, the reliability is 0%, an indication of extremely poor inter-rater agreement. Interestingly however, after carefully examining the raw data, one finds that 67 out of a total of 73 slides agree. Thus, intuitively, it seems that our naive statistic does not correctly summarize the inter-rater reliability [8].

Furthermore, looking at each endpoint for percent of agreement, we notice that, as shown in Table 3 below, the average agreement is approximately 60% (ranging from 23% up to 92%). Generally speaking, these MRI data do not fit the ANOVA model well, since it uses the ratio of the variance of disagreement (SSE) to the mean variance over all slides (SSM); however, the slides with agreeing scores have no contribution to the SSE. Thus, in contrast to our intuition, when there are a many of slides with perfectly agreeing MRI scores, the reliability statistic is heavily dependent on the few slides which do not agree. Figure 3 below is the schematic plot of the data in Table 3 and it confirms that unstable reliability results are closely tied with a higher percentage of agreeing slides [8].

Table 3. Percent of agreement over all 28 endpoints

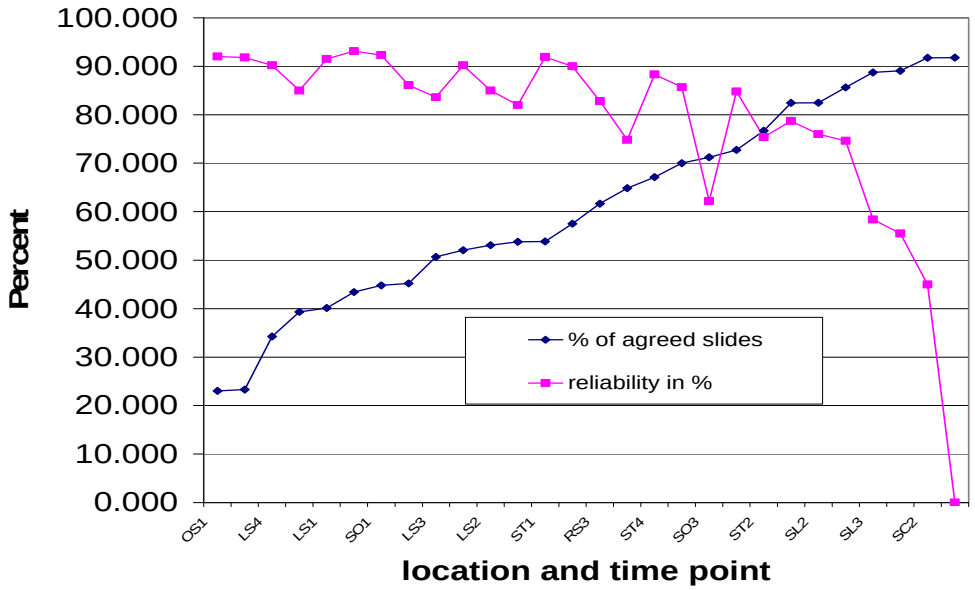
	Agreement			Agreement
LOCATION	Baseline (1)	Week 28 (2)		Week 28 (3)	Week 52 (4)
Spine Overall (SO)	45%	65%		71%	58%
Spine Cervical (SC)	86%	92%		92%	89%
Spine Thoracis (ST)	54%	77%		82%	67%
Spine Lumbar (SL)	70%	82%		89%	73%
Overall SI (OS)	23%	39%		45%	23%
Left SI (LS)	40%	53%		51%	34%
Right SI (RS)	43%	54%		62%	52%

Figure 3. Data in Table 3. Location and time point are sorted by percent of agreed slides for each endpoint from lowest to highest.



Next, in Figure 4, we combine the results from Table 2 and Table 3, showing that the higher the percentage of agreed slides at an endpoint, the more unstable our reliability from the ANOVA is.

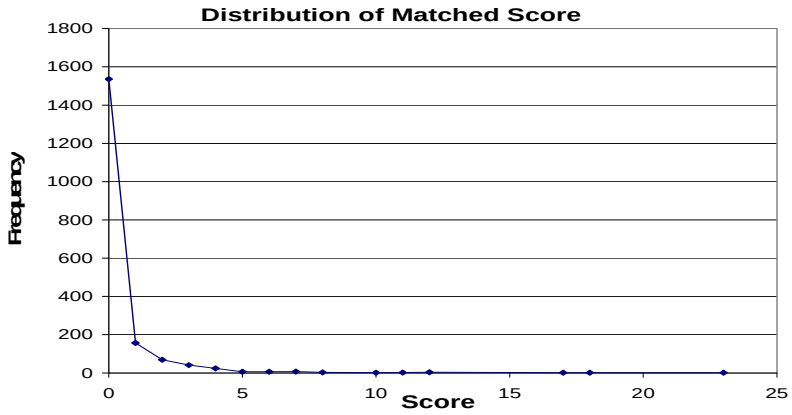
Figure 4. Reliability plot based on data in Table 2. Overlay of data in Figure 3



.

Finally, figure 5 below summarizes another important character of the MRI data, depicting that the agreed scores are heavily concentrated in several low MRI regions. It is then possible, for a specific endpoint (if we are considering slides with agreed scores between) that their averages might be a constant (e.g. zero on all averages). This leads to zeros on both SSE and SSM if the ANOVA model is used to fit these subgroup MRI data.

Figure 5. Frequency plot of agreement over each score



Tree-structured reliability analysis:

Trees are a class of non-parametric methods which can be used to fit many types of data [12-15]. Usually, one starts by growing a tree with many branches, before using n-fold cross-validation to prune it (controlled by a loss function); this then yields the final simplified tree. The data sets that benefit most from such tree-based methods are those where the final tree contains only few nodes with the data in each node generically showing different character. In such cases, the summary statistic on these final nodes are simple and easy to understand.

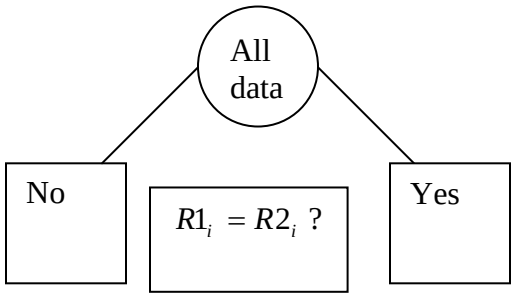
The unique feature of the scored MRI data is that it contains 20% to 90% slides with agreed scores between 2 raters over all 28 endpoints. Thus, the idea of our Tree-Structured Reliability analysis is to partition the data into 2 nodes based on their contribution to the SSE term in the ANOVA based reliability analysis. In particular, we propose the following procedure:

1. For a node with $R1_i = R2_i$ assign reliability: 100%

2. For a node with $R1_i \neq R2_i$ assign reliability: $1 - \frac{SSE}{SSM}$

3. Then combine the reliability estimation on 2 nodes where we define the modified reliability statistic as:

$$\max(0, 1 - \frac{SSE}{SSM}) * (1 - p) + p, \textit{where } p = \textit{percent of data on } R1_i = R2_i \textit{ node}$$

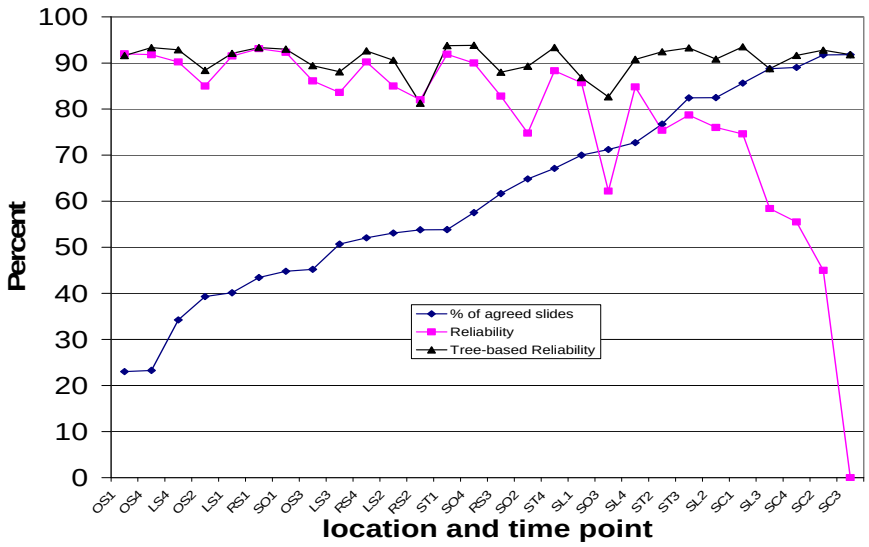


The crucial intuition is that the impact of disagreeing slides is now weighted by their proportion. This allows us to bypass the previous instability in our ANOVA model, which was dominated by the variance of disagreement. Table 4 below shows the reliability statistics based upon our Tree-Structured model, while Figure 6 is the plot of the data in Table 4.

Table 4. Results with Modified reliability statistics

	Reliability			Reliability	
LOCATION	Baseline (1)	Week 28 (2)		Week 28 (3)	Week 52 (4)
Spine Overall (SO)	93%	89%		83%	94%
Spine Cervical (SC)	94%	93%		92%	92%
Spine Thoracis (ST)	94%	92%		93%	93%
Spine Lumbar (SL)	87%	91%		89%	91%
Overall SI (OS)	92%	88%		89%	93%
Left SI (LS)	92%	91%		88%	93%
Right SI (RS)	93%	81%		88%	93%

Figure 6. Tree-structured reliability results as in Table 4. Overlay of the Figure 4 of percent of agreed slides and the reliability from ANOVA model



We provide two simple observations on Figure 6:

1. The Tree-Structured reliability statistic produces a more consistent estimation of agreement from 2 raters across all endpoints.

2. By partitioning the MRI slides using their residuals (e.g. the contribution to the SSE term in ANOVA model), the Tree-Structured reliability statistic improves upon the original reliability statistic in terms of how to utilize agreeing data points between 2 raters.

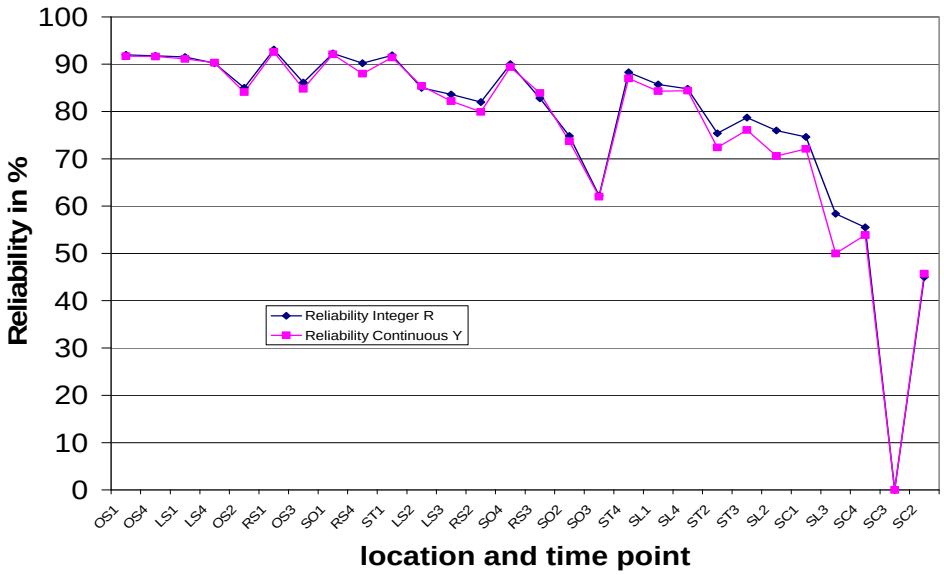
Extending the modified reliability statistics to continuous data.

When collecting noisy MRI data people often use a VAS (visual-analog-scale) score, instead of integer valued categories [17]. Instead of asking a rater to score the image of a spine vertebrae from 0 to 3 for severity of disease, a continuous line of 3 inches long may be associated with each vertebrae image; then the rater will be asked to mark a position on the line to indicate the severity of the disease, with e.g. the left end indicating no lesions and the right end indicating very severe lesions. In such cases, the MRI data is now a continuous variable from 0 to 3 for each vertebrae.
The question is thus whether our the tree-structured reliability analysis has a natural continuous analog?

For illustration purposes, we can make our MRI data be a continuous variable, as if it came from a VAS scoring procedure (e.g. by adding a uniform random variable on each score). More specifically, let the original MRI score be variable **R**, and **Y** = |R + U| be a continuous version of such a score, where U be is a uniform random variable between (-

0.5, 0.5). As expected, the plot of this continuous data shows that the set is again skewed (Figure 7).

Figure 7. Comparing reliability of original MRI data **R** with that from above continuous variable **Y**, based on ANOVA model



For continuous scoring, the pair of scores from 2 raters on an MRI slide cannot in general be equal; however, in terms of contributions to the SSE term of the ANOVA model, they are essentially the same. With this idea in mind, we define a modified tree below:

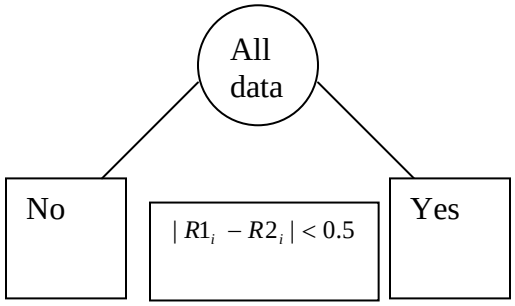
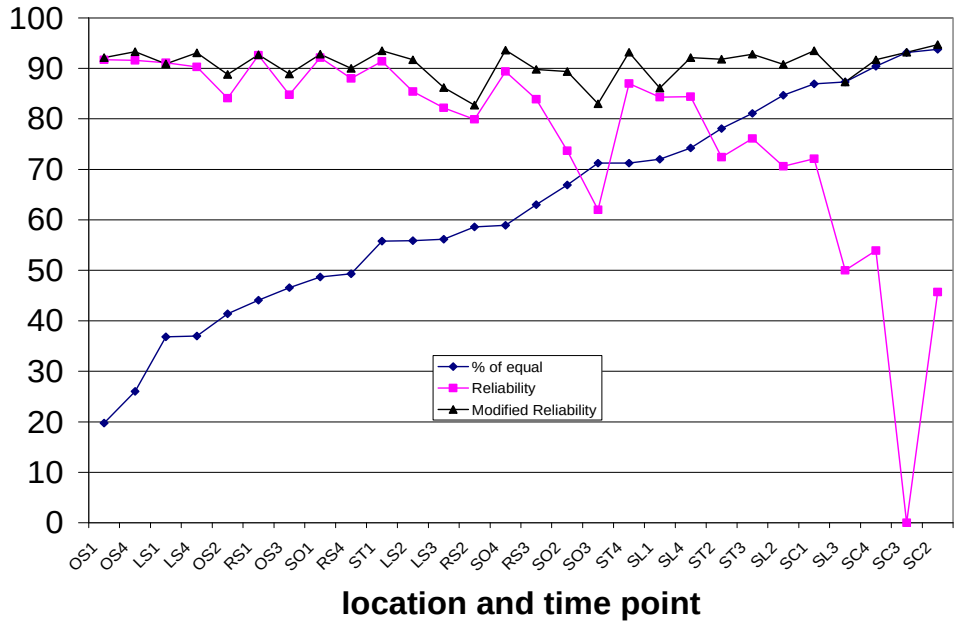


Figure 8. The same as Figure 6, but using the above Modified Tree with continuous MRI data **Y**



It is unsurprising that Figure 8 demonstrates that our Tree-structured reliability method can be employed even when the rating scores are continuous variables.

Conclusion:

We propose the Tree-Structured reliability method as a modification to the standard two-way fixed effect reliability statistic. We provide a specific example which demonstrates that our method characterizes an improvement over the original method when the distribution of differences from 2 raters is strongly skewed.

Reference

[1] Cohen, J. "A coefficient for agreement for nominal scales", **Education and Psychological Measurement.** Vol. 20, pp. 37–46 (1960).

[2] Fleiss, J. L. "Measuring nominal scale agreement among many raters", **Psychological Bulletin.** Vol. 76, No. 5, pp. 378–382 (1971).

[3] Saal, F.E., Downey, R.G. and Lahey, M.A "Rating the Ratings: Assessing the Psychometric Quality of Rating Data", **Psychological Bulletin.** Vol. 88, No. 2, pp. 413–428 (1980).

[4] Saal, F.E., Downey, R.G. and Lahey, M.A "Rating the Ratings: Assessing the Psychometric Quality of Rating Data", **Psychological Bulletin.** Vol. 88, No. 2, pp. 413–428 (1980).

[5] Bland, J. M., and Altman, D. G. "Statistical methods for assessing agreement between two methods of clinical measurement", **Lancet** 307–310 (1986).

[6] Fayers, P. and Hays, R. Assessing Quality of Life in Clinical Trials: Methods and Practice, Oxford University Press, (2005)

[7] Gwet, Kilem L. Handbook of Inter-Rater Reliability, Gaithersburg, Advanced Analytics (2012).

[8] Gwet, K. L. "Computing inter-rater reliability and its variance in the presence of high agreement." **British Journal of Mathematical and Statistical Psychology**, 61, 29-48 (2008).

[9] Shoukri, M. M. Measures of Interobserver Agreement and Reliability. Boca Raton, Chapman & Hall/CRC Press, (2010)

[10] Hayes, A. F. and Krippendorff, K. "Answering the call for a standard reliability measure for coding data", **Communication Methods and Measures,** 1, 77-89 (2007).

[11] Shrout, P. and Fleiss, J. L. "Intraclass correlation: uses in assessing rater reliability", **Psychological Bulletin.** Vol. 86, No. 2, pp. 420–428 (1979).

[12] Breiman, L., Friedman, J., Stone , C. J., Olshen, R. A., Classification and Regression Trees Boca Raton, Chapman & Hall/CRC Press, (1984).

[13] Wasserman, L. All of Nonparametric Statistics. Springer Texts in Statistics, (2005).

[14] Schmoor C, Ulm K, Schumacher M. "Comparison of the Cox model and the regression tree procedure in analysing a randomized clinical trial", **Stat Med.** 24 2351-2366 (1993).

[15] Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R. "Piecewise-polynomial regression trees" **Statistica Sinica** 4 143-167 (1994).

[16] Rudwaleit, M., van der Heijde, D., Khan, M., Braun, J., and Sieper, J. " How to diagnose axial spondyloarthritis early", **Ann Rheum Dis.** 63 535–543 (2004).

[17] S. Grant, T. Aitchison, E. Henderson, J. Christie, S. Zare, J. McMurray, and H. Dargie., "A comparison of the reproducibility and the sensitivity to change of visual analogue scales" **Chest**, 116 1208-1217 (1999).