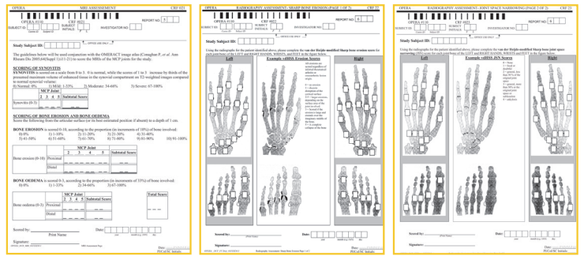
The Effect of Evaluation Variability at the Unit of Measurement on the Reliability of OMERACT RAMRIS and van der Heijde-Modified Sharp Score

Ruben Tavares[1], Naveen Parasu[2], Karen Finlay[2], Erik Jurriaans[2], Hao Wu[3], Karen A. Beattie[1], Maggie Larche[1,4], Lawrence E. Hart[1,4], William G. Bensen [1,4], Raja S. Bobba[1,4], Alfred A. Cividino[1,4], Colin E. Webber[1,3], Jean-Eric Tarride [1,4,5], and Jonathan D. Adachi[1,4]

[1]McMaster University, [2]Hamilton Health Sciences, [3]Adachi Medicine Professional Corporation, [4]St. Joseph's Healthcare Hamilton, [5]Programs for Assessment of Technology in Health Research Institute, Hamilton, ON, Canada

Background

- 2 major rheumatoid arthritis (RA) diagnostic imaging disease progression measures include the following (**Fig.1**):
  - MRI: OMERACT RA MRI Score (RAMRIS) (1)
  - X-ray: van der Heijde-modified Sharp score (vdHSS) (2)



a b c

Fig. 1. Evaluation forms for RAMRIS (a) and vdHSS erosions (b) and JSN (c).

- The smallest detectable difference (SDD) is used to quantify status score reliability in scale units (3)
- To date, the RAMRIS and vdHSS SDD have been evaluated at the overall or component feature score levels of analysis ignoring variability at the unit of measurement (4-11)

Objectives

- To determine the reliability of RAMRIS and vdHSS at the unit of measurement across four radiologists and to compare it to the conventional approach

Methods

- A cross-sectional study of RA patients conducted
- 19 MR image sets of metacarpophalangeal joints (MCP) 2-5
  - 1.0T ONI OrthOne small-bore extremity MRI
  - Patient RA symptom duration: mean (SD) 6.8 (6.4) years
- 9 X-ray image sets of both hands, wrists, and forefeet
  - Patient RA symptom duration: 7.6 (7.3) years
- Imaging sets independently evaluated by 4 radiologists
- Shrout and Fleiss fixed and random effects intra-class correlation coefficients (fICC & rICC) calculated (12)
- Overall, feature & unit of measurement score SDD calculated

Results

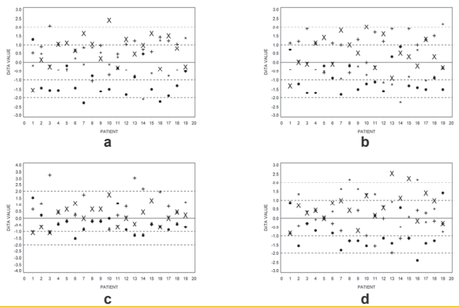
- Overall and component feature reliability measures were dependent on the anatomy compared (**Table 1**)

Table 1. Summary of reliability results.

Measure	n	fICC	rICC	SDD	% SDD / Max Score	% SDD / Scale Max
OMERACT RAMRIS						
Overall	19	0.66	0.54	11	37	10
Component Subscore						
Erosion	*19*	*0.55*	*0.42*	*7*	*39*	*9*
Edema	*19*	*0.60*	*0.54*	*4*	*50*	*17*
Synovitis	*19*	*0.39*	*0.29*	*5*	*45*	*42*
Unit of Measurement						
Erosion	*152*	*0.71*	*0.66*	*2*	*22*	*20*
Edema	*152*	*0.56*	*0.54*	*1*	*33*	*33*
Synovitis	*76*	*0.41*	*0.35*	*2*	*66*	*66*
vdHSS (hands and feet)						
Overall	7	0.69	0.57	39	45	9
Component Subscore						
Erosion	*7*	*0.60*	*0.52*	*34*	*56*	*12*
JSN	*7*	*0.85*	*0.75*	*11*	*31*	*7*
vdHSS (hands only)						
Overall	9	0.65	0.56	34	41	13
Component Subscore						
Erosion	*9*	*0.42*	*0.34*	*26*	*53*	*16*
JSN	*9*	*0.83*	*0.78*	*11*	*31*	*9*
vdHSS (Unit of Measurement)						
Erosion	*372*	*0.61*	*0.59*	*2*	*22*	*20*
JSN	*354*	*0.69*	*0.67*	*2*	*50*	*50*

Intraclass Correlation Coefficient Plots (13)

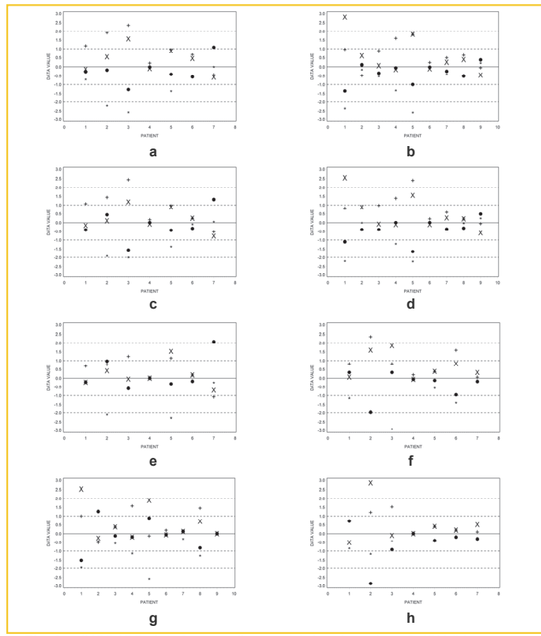
- ICC plots of RAMRIS MCP 2-5 evaluations (**Fig 2**).



a b c d

Fig. 2. RAMRIS ICC plots for overall score (a), erosion (b), edema (c), and synovitis (d). Each symbol (+, *, x, o) represents a specific radiologist.

- From **Table 1** (left)
  - rICC more conservative than fICC
  - At aggregate score levels, SDD are large
  - At unit of measurement, only edema SDD ≤ scale unit
  - Synovitis had the lowest reliability of the MRI features
  - Similar '% SDD / Max Score' & '% SDD / Scale Max' between scoring systems
- From ICC plots of RAMRIS evaluations **Fig. 2** (left, below)
  - Tendency towards bias between raters
- ICC plots of vdHSS evaluations (**Fig 3**).
  - Tendency towards bias between raters
  - Bias may be less pronounced than RAMRIS evaluations



a b c d e f g h

Fig. 3. vdHSS ICC plots for overall score (a), hand erosions & JSN (b), hands/feet erosions (c), hand erosions (d), feet erosions (e), hands/feet JSN (f), hands JSN (g), feet JSN (h). Each symbol (+, *, x, o) represented a specific radiologist.

Conclusions

- Mean ICC ranged from moderate to good (4,14)
  - rICC is more conservative than fICC & is generalizable
- Inter-rater variability of evaluations may not be random
  - Results in small ICC & large SDD for aggregate scores
    - Decreases sensitivity of abnormality detection
    - Negatively impacts sensitivity to change
- Assessment of reliability for overall or component feature scores ignores variability at the unit of measurement
  - A potentially false assumption is made that the SDD at the unit of measurement is less than the scale unit
  - Unit-of-measurement-SDD defined changes may be used to more accurately score RAMRIS and vdHSS
- Aggregate score reliability evidence is suspect if it does not account for variability at the unit of measurement

References

1. Østergaard M, et al. J Rheumatol 2003;30:1385-6.
2. van der Heijde D. J Rheumatol 2000;27:261-3.
3. Bruynesteyn K, et al. J Clin Epidemiol 2004;57:502-12.
4. Lassere M, et al. J Rheumatol 2003;30:1366-75.
5. Conaghan PG, et al. J Rheumatol 2007;34:857-8.
6. Boers M, et al. Lancet 1999;350:309-18.
7. Klareskog L, et al. Lancet 2004;363:675-81.
8. St Clair EW, et al. Arthritis Rheum 2004;50:3432-43.
9. Lassere MN, et al. J Rheumatol. 2001;28:911-3.
10. Bruynesteyn K, et al. Ann Rheum Dis 2005;64:179-82.
11. Lassere M, et al. J Rheumatol 1999;26:731-9.
12. Shrout PE & Fleiss JL.. Psychol Bull 1979;86:420-8.
13. Bland JM & Altman DG. Lancet 1986;1:307-10.
14. Landis R & Koch GG. Biometrics 1977;33:159-74.

Acknowledgements

Ruben Tavares was the recipient of Graduate studentship from Canadian Arthritis Network/The Arthritis Society. The project was supported by a Canadian Initiative for Outcomes in Rheumatology (CIORA) and Ministry of Health and Long Term Care grant via collaboration with the PATH Research Institute. Image acquisition & data collection & management support from the following individuals is acknowledged: Erika Arseneau, Christine Fyfe, Craig MacDougald, Erica Nunes, Caitlin Steven, Mary Strain, and Steven Tytus.