Psoriasis severity assessment with a computational similarity-clustering programme reduces intra- and inter-observer variation

Arman Garakani (1), Ionela Manole (1,2), Wouter Walmink (1), Adrian Young-San Rössler (1), John R Zibert (1)

1) LEO Innovation Lab, Silkegade, 1113 Copenhagen, Denmark 2) Dermatology Research Unit, Colentina Clinical Hospital, Bucharest, Romania

john@leoilab.com

INTRODUCTION

Clinical studies of psoriasis severity using photos, a.k.a image data, where multiple raters repeatedly rate the image data according to PASI are widely used to assess treatment efficacy and progression in trials and clinical practice. PASI is widely considered subjective and of poor reproducibility. Difficulty of calculating inter-rater and intra-rater reproducibility remains an obstacle in using PASI . Current best practices include PASI assessment training, only using one evaluator per case and use of reference baseline images.

HYPOTHESIS

Similarity Clustering of all pairwise comparisons in a set of psoriasis photos reduces inter- and intra-rater variability in the context of PASI components of redness, thickness and scaliness.

METHODS

Images were provided from patients with psoriasis via a digital mobile application (Imagine™, CPH, DK) after they had accepted our terms and conditions to use the pictures for research purposes.
Five dermatologists evaluated the severity of psoriasis photo sets by modifid (mPASI)PASI absolute scoring and a relative pairwise PASI scoring using similarity-clustering and conducted using a web-program (Pairoscope™, CPH, DK)

Standard Method

Distribution of visual gradeing of redness of photos in the set by 5 dermatologists for both standard scoring and Similarity Clustering scoring is shown in Fig 2. Every plaque is graded for existance of disease related symptoms: scaliness, redness, and thickness from 0 to 4 for most sever and most commonly in integer increments. mPASI for a photo is computed as an average of the 3 scores. Hence, a progression evaluation for a time-ordered set of photos is the corresponding mPASI scores.

Similarity Clustering Method

We redefine scoring to ranking of images in from all pairwise comparisons[6,9]. A pairwise comparison captures how similar / dissimilar two photos are in the context of a mPASI component. All unique pairings of photos in are presented to the dermatologist in random order. Time is hours between photo capture.

RESULTS

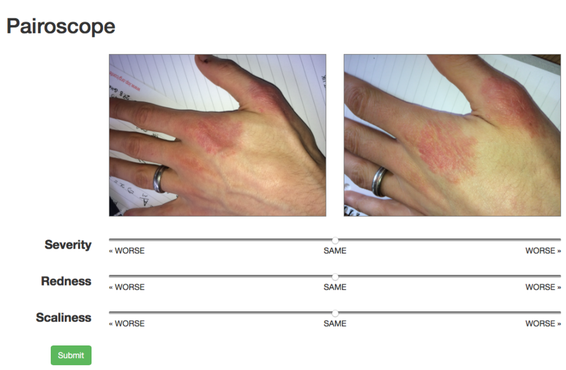

Figure 1. Pairoscope

Web Interface for relative assessment of pair of Photos. Displaying two images at a time, it allows zoom and pan of each image independably, 3 sliders with range of -0.5 to +0.5 set at default position of 0 and a button to submit the comparison.

Standard and Pairwise Scores by All Dermatogists

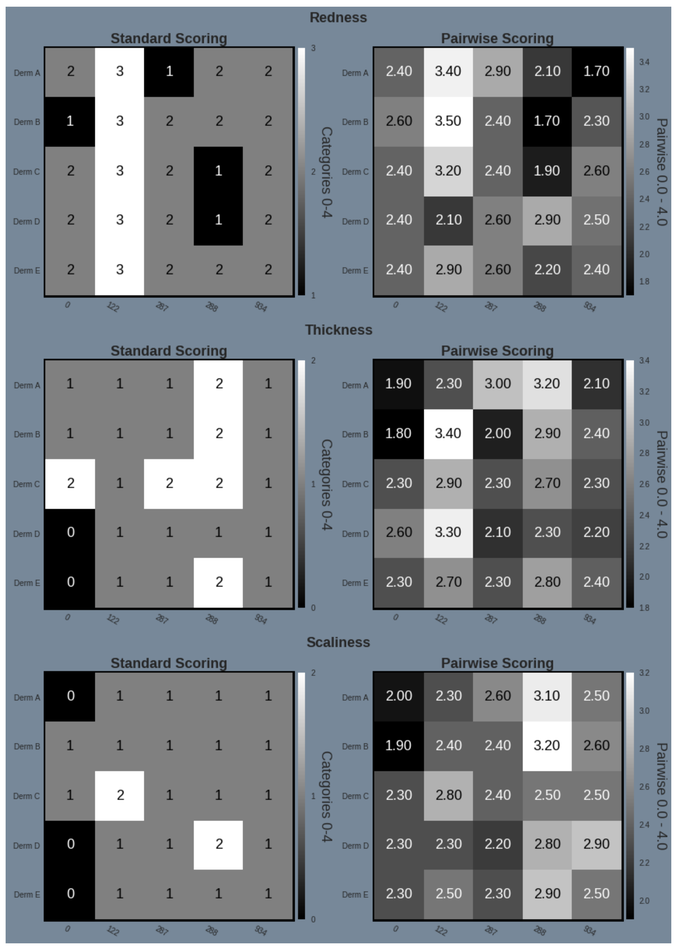

Figure 2. All scores for redness, thickness and scaling over 5 time points

Agreement Among Raters Intra/inter Variations in Scoring

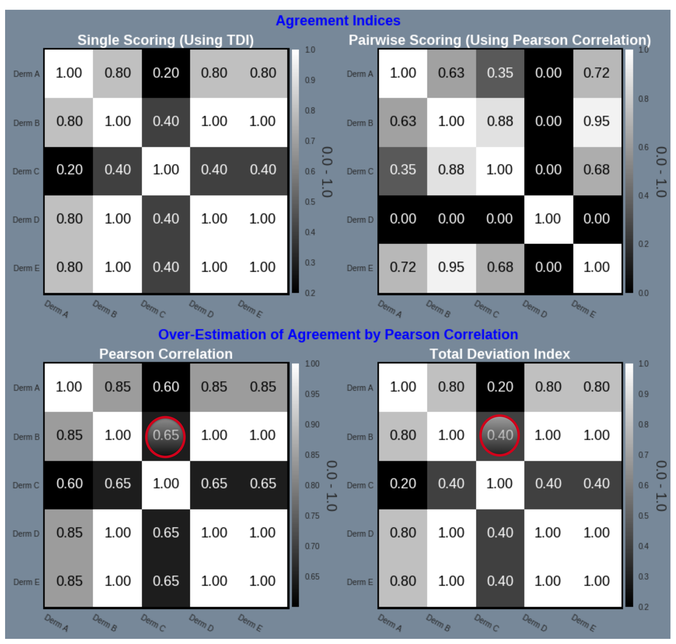

Figure 3A. Inter-rater Reproducability
Figure 3B. Agreement Indices For Standard Scoring

Progression

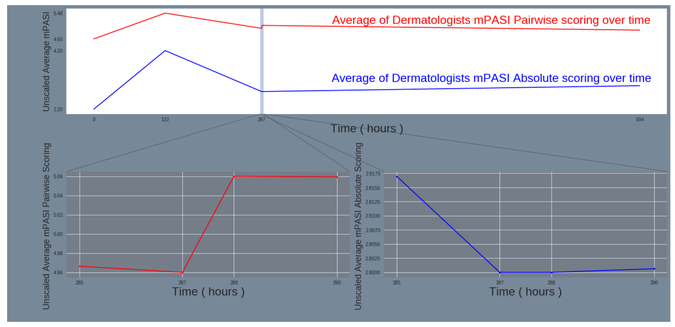

Figure 4. Disease Progression Over Time

Repeated mPASI scoring of single photos by the same or different dermatologist showed consistent mPASI ratings of 50% to 80%, respectively. Repeated mPASI comparison using the similarity clustering programme showed consistent mPASI ratings of >95%.
Pearson correlation between absolute scoring and pairwise scoring progression was 0.72

DISCUSSION

Categorical Variables
In a categorical variable say one to represent race:
  with 5 values: 1=White, 2=Black, 3=Asian, 4=Mixed, and 5=Other
Values do not represent anything and the order is arbitrary.
In analyzing a collection of race values,
it is better to bin the data and calculate rank rather than to caclulate an average of the numeric values.

At first glance, values in a variable representing each component do represent presence or absence of disease
relatedness and the order is not arbitrary,
  with 0 = no disease related, 1 = mild, 2 = moderate, 3 = severe and 4 = very severe.
Wider close examination has revealed skewed distribution[3,4,5] and inherent subjectivity:
  variability in redness and scaling is due to their “relative” nature,
  variability inarea is due to regional variations and its subjective definition,
  variability in thickness is due to inconsistency in training environment and many others.
Pearson Correlation
Pearson corrleation is invariant to linear transformations:
  correlation of two sets of PASI scores [1,2,0,0,1] and [2,4,0,0,2] is 1.0 !

CONCLUSIONS

A • Pairwise Scoring reduces inter/intra variability.
B • Pairwise Scoring produces accurate estimate of disease progression.
C • Use of Pearson correlation in Calculation of InterClass Correlation can lead to over-estimation of inter-agreement.
D • In larger image sets the true pairwise comparisons may not be actively measured, however, algorithms do exist for their estimation.
E . A similarity-clustering programme significantly reduces the intra- and inter-observer variation assessing severity in dermatoses, indicating an application for clinical trials and practice. Further studies are warranted to generate a new standard for severity assessments.

REFERENCES

[1] A rator here may be both a human rater or a machine rater.
[2] PASI = psoriasis severity index
[3] Ribas, Jonas, Cunha, Maria da Graça Souza, Schettini, Antônio Pedro Mendes, & Ribas, Carla Barros da Rocha. (2010). Agreement between dermatological diagnoses made by live examination compared to analysis of digital images. Anais Brasileiros de Dermatologia, 85(4), 441-447.

[4] Youn, S. W., Choi, C. W., Kim, B. R., & Chae, J. B. (2015). Reduction of Inter-Rater and Intra-Rater Variability in Psoriasis Area and Severity Index Assessment by Photographic Training. Annals of Dermatology, 27(5), 557–562.
[5] Pierre-Antoine Gourraud, Caroline Le Gall, Eve Puzenat, Francois Aubin, Jean-Paul Ortonne and Carle F. Pau, Why statistics matter: limited inter-rater agreement prevents using the psoriasis area and severity index as a unique determinant of therapeutic decision in psoriasis, journal of Investigative Dermatology (2012) 132, 2171–2175; published online 17 May 2012

[6] Balcan MF, Blum A, Vempala S. Clustering via similarity functions: theoretical foundations and algorithms. In: 40th ACM Symposium on Theory of Computing Conference; 17–20 May 2008; Victoria, Canada. New York, NY, USA: ACM. pp. 1-42.
[7] Huiman X. Barnhart, Michael J. Haber & Lawrence I. Lin (2007) An Overview on Assessing Agreement with Continuous Measurements, Journal of Biopharmaceutical Statistics, 17:4, 529-569,
[8] Lin. L. I. (2000), “Total Deviation Index for Measuring Individual Agreement: With Application in Lab Performance and Bioequivalence,” Statistics in Medicine, 19, 255–270.

[9] E.E.M. van Berkum. "Bradley-Terry model". Encyclopedia of Mathematics. Retrieved 18 November 2014.
[10] XLSTAT 2018.2.50384 - Reliability analysis - One-sample ICC test / Two-tailed test
[11] Wauthier, F., Jordan, M. and Jojic, N., 2013, February. Efficient ranking from pairwise comparisons. In International Conference on Machine Learning (pp. 109-117)