# Online Appendix 18.A: Instructive correspondence with a user (identifiers suppressed)

Following is an email from a researcher to the author.

Dear Dev,

We are conducting a reader study on xx where readers are freely asked to mark and score xx lesions. We are using your JAFROC software (v4.2.1) to analyze the data, but we are encountering some doubts that we hope you can kindly help us with.

Attached you can find the original data, the .xls sheet, as well as the results of using JAFROC-2, and JAFROC-2-WEIGHTED (following the definition from your 2004 Med Phys paper). We have xx cases (xx abnormal), xx modalities and xx readers. We treat everything as random.

As you can see, WEIGHTED analysis results into non-significant differences between modalities, as opposed to the non-weighted analysis. Is there any particular reason for that? Only three abnormal cases have 2 lesions, for the rest is 1. On your readme file you write this:

Weighted vs. non-weighted: Weighting assures that all abnormal cases get equal importance, regardless of the number of lesions on them, a desirable statistical characteristic. The author has encountered a situation in nuclear medicine bone scans where the number of lesions can vary over a very large range (1 – 100) and in this situation only weighted analysis should be reported. Based on our experience, there is little difference between the two analyses when the number of lesions varies from 1-3. There is some loss of statistical power in using weighted over non-weighted figures-of-merit, but the benefits of localization (vs. ROC analysis) are largely retained. Unless there are clinical reasons for doing otherwise, equal weighting is recommended (assigning zero weights to all images achieves this).

A second doubt is that, when I use 3 modalities (instead of 4) and 6 readers (5 from above, plus one who only read 3/4 modalities), get significant differences. I cannot really understand why, so if you have some quick thoughts we would gladly hear them.

Finally, JAFROC calculates a figure-of-merit defined as the probability that a lesion is rated higher than any mark on a normal image. It is estimated using the non-parametric Mann-Whitney-Wilcoxon U-statistic applied to the lesion ratings and the highest rating on normal images, where unmarked lesions and unmarked normal images are assigned the negative infinity rating. However, on the result files that I attach, which are generated by the software, these FOMs are still labeled as "Area". Is this corrected, am I missing something or it is a small mistake of the software?

Thank you very much for your time!

This is the author's response.

Dear xx,

I have examined your dataset and there is nothing unusual about your results. The p-value is a realization of a random variable and it is extremely sensitive to small differences in the observed F-statistic; one expects wAFROC to give slightly greater p-value, which you observe, but the values of wAFROC and AFROC are actually quite close; nothing to get excited about.

If you apply two sequential tests to get the result that you prefer, you should, in all honesty, be using an alpha of 0.025 (Bonferroni correction), according to which neither method rejects the NH. Or you could simply not report that you tried wAFROC first, which I would not advise, especially when you are beginning your research career.

There is nothing wrong with reporting a negative result, as long as you quote a CI for the differences and variance components, output by the software. There is a paper by Metz on the importance of confidence intervals when failing to reject the NH. These results in the archival literature are helpful to others, especially for sample size estimation for future studies.

Results for all three methods as excel files are attached.

There is no mistake in the software; the “area” that you refer to is standard format used in all ROC software, including U of Iowa and U of Chicago software, it is generic for the area measure; when organizing a tabular output it is helpful to have short names. A line in the output file gives the precise definitions:

Analysis method: WEIGHTED JAFROC FOM DBM-MRMC SIGNIFICANCE TESTING

Analysis method: JAFROC FOM DBM-MRMC SIGNIFICANCE TESTING

Dev

The correspondence brings out several instructive points.

1. Over-reliance on small differences between p-values can be misleading. The p-value is a realization of a random variable. Even if one method has higher statistical power than another, one expects instances where the opposite result is observed. One expects wAFROC to have slightly less power than AFROC, which the correspondent observed. The reason is that in wAFROC there is *one* wLL value per diseased case while with AFROC there are *multiple* LL values per diseased case. The single vs. multiple values would tend to favor AFROC, but counteracting this is the fact that the single value is more stable than each of the multiple values.
2. It is necessary to apply a Bonferroni type correction when applying multiple tests to the same data. The latter amounts to "fishing the data": one can usually tease out a desired result by applying enough different tests; the Bonferroni correction is there to keep one honest.

Not explicitly stated, but relevant:

1. One needs to be aware of the equivalence1 between a figure of merit (a number calculated from the data) and an area measure (obtained by numerical integration of an empirical operating characteristic).
2. The JAFROC FOM statistic is not equivalent to a Wilcoxon statistic. The latter applies to ROC data, where localization is not involved, while localization is definitely involved in calculating the JAFROC (i.e., AFROC) FOM. The author prefers the term quasi-Wilcoxon statistic when referring to the AFROC FOM.

# References

1. Chakraborty DP, Zhai X. On the meaning of the weighted alternative free-response operating characteristic figure of merit. *Medical physics.* 2016;43(5):2548-2557.