Published in final edited form as:

Acad Radiol. 2009 April ; 16(4): 507–510. doi:10.1016/j.acra.2008.12.011.

Counterpoint to “Performance assessment of diagnostic systems under the FROC paradigm” by Gur and Rockette

Dev P. Chakraborty, Ph.D.

University of Pittsburgh, Department of Radiology, 3520 Forbes Ave, Suite 109, Pittsburgh, PA 15261

This is a response to a recent thought provoking paper [1] by Gur and Rockette (henceforth referred to as “the authors”) that raises issues regarding the applicability of free-response receiver operating characteristic (FROC) methodology [2] to imaging systems evaluations. Unlike many tests, diagnostic imaging provides information about the location(s) of disease, and other information, in addition to its presence or absence. However, the receiver operating characteristic (ROC) method only considers the disease presence or absence information and disregards location. For some clinical tasks the ROC method is more relevant. For example, a task like detecting diffuse pulmonary fibrosis that does not involve focal lesions is appropriately analyzed by the ROC method. However, tasks such as detecting lung nodules in chest radiography, or microcalcifications in screening mammography, that involve detecting localized lesions, especially multiple lesions, are more appropriately handled by FROC analysis.

By way of disclosure, having worked in this area since ca. 1984, I am vested in FROC methodology. There are two other location-specific paradigms not mentioned in the authors' paper, the location ROC (LROC) paradigm [3] and the region of interest (ROI) paradigm [4]. In the LROC paradigm the radiologist provides an overall rating and marks the most suspicious region. In the ROI paradigm the investigator segments the image into ROIs and the radiologist rates each ROI for presence of disease. Like FROC these paradigms were developed to address the localization and multiple lesions limitations of ROC methodology, and most of the issues attributed to FROC apply to these methods as well.

Neglect of location information implies suboptimal measurement precision, ie, low statistical power, which diminishes the ability to detect differences between modalities, the most common application of observer performance studies. Early analyses tools developed by me drew fair criticism because they ignored correlations. This issue was resolved in 2004 by the jackknife alternative free-response operating characteristic (JAFROC) method [5] which was demonstrated to have substantially higher power than the ROC method and passed rigorous statistical validation. As evidenced by recent publications in journals and proceeding papers at a major international conference on medical imaging, JAFROC usage is gaining acceptance. However resistance to it is also increasing which is to be expected as part of normal scientific discourse. Gur and Rockette have done a service to the imaging community by expressing their concerns publicly and I am grateful to the Editor for giving me the opportunity to respond.

E-mail: dpc10@pitt.edu, 412-605-1553 (phone), 412-605-1554 (fax).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

I agree with the authors on some of the issues: ambiguity of the acceptance target (how close a mark has to be to a lesion in order to be counted as a true positive); suitability of the figure of merit for multiple lesions with different clinical significances; handling multiple views per case, simulator related issues such as distributional assumptions and lack of accounting for satisfaction of search; etc. However, I choose to regard these as research opportunities and thank the authors for laying out a detailed research road-map. This research, quite apart from its obvious application to modality assessment, could substantially extend our understanding of medical decision making. I am making progress in some of these areas but others need to get involved. Unfortunately, if one accepts the authors' premise that ROC is more clinically relevant than FROC in location-specific tasks such as screening mammography, then few will be motivated to do research in FROC analysis.

When a screening radiologist refers a patient for further investigation to a colleague, the location of the lesion and which breast is involved, is crucial. Screening programs require documentation of lesion characteristics, including location, in addition to the overall recommendation to “recall” or “return to screening”. The location(s) identified at screening guide the subsequent diagnostic workup and the decision to biopsy the lesion(s). Just knowing that the woman has a malignancy somewhere in her breasts is obviously less helpful to the mammographer doing the diagnostic workup than knowing the locations and types of abnormalities detected at screening by a colleague. Neglecting location information can lead to the scenario where in the ROC paradigm the radiologist is credited for detecting an abnormal condition when in fact a lesion was missed and a normal structure was mistaken for a lesion (“right for the wrong reason”). It is obvious that the clinical consequences of the two mistakes are serious: the undetected cancer is allowed to grow and a biopsy is made at the wrong location (see #1 below).

Since claims are being made that ROC is clinically more relevant than FROC in some scenarios, a definition of clinical relevance is needed. Evaluation methods form a six-level hierarchy [6] of efficacies: technical, diagnostic, diagnostic-thinking, therapeutic, patient outcome and societal. I will interpret “clinical relevance” of a measurement as its hierarchy-level. The difficulty/cost of measurement increases as one moves up the hierarchy. At the lowest level, technical efficacy (e.g., spatial resolution) is easiest to measure. The level-2 ROC measurement has a reputation of being time consuming and costly. Level-3 measurements such as positive predictive values are even more laborious [7]. One way of showing clinical relevance is to perform measurements at the higher level and show that they confirm the lower level measurements regarding which modality is superior. If the performance difference is small, demonstrating clinical relevance can be very difficult. As an example, the initial optimistic expectations of mammography CAD, which were based on ROC studies, have not been confirmed in some large-scale clinical trials [8,9]. Since it is difficult to prove the clinical relevance of ROC one can hardly claim it is *more* relevant than FROC. Black and Dwyer [10] studied the issue of global vs. local measures of accuracy and their effects on post test probability of disease, which is a level-3 measure. They considered mediastinal lymph node metastasis (LNM) which is more likely to be present in the right lower paratracheal region (4R) than in the left lower paratracheal region (4L). As expected, post test probability is higher if the radiologist knows that LNM was found in 4R rather than 4L, and this knowledge will influence the subsequent action (e.g., biopsy or surgery). But if the location information is ignored, the posttest probability is equal for the two cases. Black and Dwyer conclude “*the local versus global distinction supports the commonsense notion that information pertaining to the anatomic distribution of disease is crucial for test interpretation*”.

One cannot fault the authors' recommendation that end-users become aware of issues with FROC and address them in the study design, but the authors fail to provide guidance on how one is to address them, apart from, by implication, *not conducting an FROC study*. The one

example given of an appropriate application of the FROC method actually follows the ROI paradigm (see #8 below). An end-user may reasonably conclude from the authors' paper: "reconsider using FROC; consider using binary or ROC methods instead".

What follows are specific response to some of the more salient issues.

1. "... location is not as primary a factor in the screening arena ... additional abnormalities often are found ... as a result of this recommendation during diagnostic workup..."

Suppose that during screening the radiologist reports a false positive region on an abnormal image but misses the lesion. The patient would be referred for a diagnostic workup during which the lesion(s) may be found and appropriate clinical action taken. For this patient the two incorrect decisions at screening (missed lesion and false positive) did not affect the final clinical outcome. The problem is that the radiologist/modality combination that "encourages" this type of error will also tend to generate more false positives on normal images, and in screening this has serious consequences. To pose a rhetorical question, given a choice between two modalities, on one the radiologist is "right for the right reason", and on the other the radiologist is "right for the wrong reason", which modality would insurers, radiologists and their patients trust?

2. "...FROC data can be viewed as clustered data ... analysis of such data is complicated by the necessity to account for the correlations..."

While JAFROC is not specifically mentioned, since it is currently the most easily accessible and widely used method for analyzing FROC data, and since presumably Gur and Rockette do not mean to imply that their own methods [11,12] neglect data clustering, I am assuming that their issue is with JAFROC. JAFROC does not ignore correlations. It has been acknowledged by experts as a valid method of analyzing FROC data. Ref. [13,14] states "... *their paradigm successfully passes a rigorous statistical validation test.*" Ref. [15] states "... *to accommodate the correlations within an image... have suggested a jackknife approach.... Extensive simulations have been conducted... (JAFROC) preserves the power advantage ... while maintaining the appropriate rejection rate....*"

3. "... (the) assumption that the search process can be adequately described by the existing 'search models' ... is yet to be proven in the actual experimental domain."

The search model goes back at least four decades [16] and is widely accepted in the non-medical and medical imaging literature. In medical imaging it has been validated by eye-tracking measurements on radiologists while they perform clinical tasks. The authors should express their concerns in journals that specialize in visual perception so that appropriate experts can respond (I am not such an expert). My search model [17] is a signal-detection theory based mathematical parameterization of the Kundel-Nodine perceptual model of search in diagnostic imaging. The distributional assumptions it makes, e.g., Poisson distributed false positives, normally distributed ratings, etc., can be modified to more realistic distributions.

4. "... (FROC) loses statistical power for actually clinically relevant questions..."

As noted previously clinical relevance is difficult to prove, but in the hope that the authors will counter with a *testable scenario* (e.g., a specific simulator) demonstrating their counterintuitive claim that *ignoring* available information results in *more* power, I will explain why FROC has greater statistical power. In the "right for the wrong reason" scenario noted previously the locations of the two marks carry *additional independent* information which allows one to *penalize* the observer for *two* mistakes

unlike the ROC method which *rewards* the observer. Under the FROC paradigm a modality that yields more “right for the wrong reason” outcomes will be judged substantially *inferior* to a modality that yields “right for the right reason” outcomes. Since the ROC paradigm cannot distinguish between these outcomes, the modalities will be closer in performance, i.e., less statistical power.

5. “... to analyze the results of (FROC) studies an “acceptance target” has to be (arbitrarily defined which) ... affect(s) the results of the study...”

While different FROC curves are realized with different choices of acceptance target it does not follow that the arbitrariness affects the conclusions of a modality comparison study. My simulations show that the *direction* of the performance change between two modalities is not sensitively dependent on reasonable values for acceptance targets, but statistical power is affected.

6. “(need for acceptance target) may become a significant problem when an easily detectable benign finding (with relatively low importance) is located near a subtle malignant finding (with high importance).”

It is fair to ask how the ROC paradigm would handle the same issue. An expert clinician might base his rating on the subtle finding, whereas a non-expert might miss the subtle finding and base his rating on the easily detectable finding. On this image the expert would give a lower rating than the non-expert and with enough such instances, *the area under the ROC curve would be higher for the non-expert than for the expert. Information, however imperfect, is better than no information.* Over time methods for dealing with such imperfect information will undoubtedly emerge. Similar comments apply to the point about multiple views per patient.

7. “...there is still no widely accepted index of the overall performance under the FROC paradigm... they do not have as simple a probability interpretation as does the area under the ROC curve”

FROC is a new and evolving paradigm and over time consensus on an index will emerge. For one lesion per image the JAFROC figure of merit does have a simple probabilistic interpretation, but I agree that the JAFROC definition needs to be improved to better account for multiple lesions with different importance and the method needs to be validated with more sophisticated simulators that relax some of the distributional assumptions. As noted previously, the area under the ROC curve is not immune to these problems, but since ROC is ignorant of the numbers and importance of lesions, it is impossible to correct for these effects.

8. “(if)... all pixels ... are evaluated in the same manner... the FROC paradigm provides a clear advantage ...”

If ratings are available for all pixels, the data collection is in fact following the ROI paradigm [4] with the ROIs identified as individual pixels. If the number of pixels is the same for all images, the method described in Ref. [4] can be used, or if they differ the bootstrap variant can be used [18].

Acknowledgements

This work was supported by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688.

References

1. Gur D, Rockette HE. Performance Assessment of Diagnostic Systems under the FROC paradigm: Experimental, Analytical, and Results Interpretation Issues. *Acad Radiol* 2008;15:1312–1315. [PubMed: 18790403]
2. Bunch PC, et al. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr Eng* 1978;4(4):166–171.
3. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys* 1996;23(10):1709–1725. [PubMed: 8946368]
4. Obuchowski NA, Lieber ML, Powell KA. Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography. *Acad Radiol* 2000;7(7):516–525. [PubMed: 10902960]
5. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. *Medical Physics* 2004;31(8):2313–2330. [PubMed: 15377098]
6. Fryback DG, Thornbury JR. The Efficacy of Diagnostic Imaging. *Med Decis Making* 1991;11(2):88–94. [PubMed: 1907710]
7. Journal of the ICRU, Receiver Operating Characteristic Analysis in Medical Imaging, ICRU Report 79. Vol. 8. O.U. Press; 2008.
8. Fenton JJ, et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N Engl J Med* 2007;356(14):1399–1409. [PubMed: 17409321]
9. Astley SM, Gilbert FJ. Computer-aided detection in mammography. *Clinical Radiology* 2004;59(5):390–399. [PubMed: 15081844]
10. Black WC, Dwyer AJ. Local versus Global Measures of Accuracy: An Important Distinction for Diagnostic Imaging. *Med Decis Making* 1990;10(4):266–273. [PubMed: 2172690]
11. Song T, et al. On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data. *Medical Physics* 2008;35(4):1547–1558. [PubMed: 18491549]
12. Bandos AI, et al. Area under the Free-Response ROC Curve (FROC) and a Related Summary Index. *Biometrics*. 2008
13. Wagner RF, Metz CE, Campbell G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Academic Radiology* 2007;14(6):723–748. [PubMed: 17502262]
14. Chakraborty DP. Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies. *Academic Radiology* 2008;15(12):1554–1566. [PubMed: 19000872]
15. Dodd LE, et al. Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: contemporary research topics relevant to the lung image database consortium. *Acad Radiol* 2004;11(4):462–475. [PubMed: 15109018]
16. Neisser, U. *Cognitive psychology*. New York: Appleton-Century-Crofts; 1967.
17. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol* 2006;51:3449–3462. [PubMed: 16825742]
18. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Acad Radiol* 2000;7(6):413–9. [PubMed: 10845400]