Published in final edited form as:
Acad Radiol. 2008 October ; 15(10): 1312–1315. doi:10.1016/j.acra.2008.05.006.

Performance Assessments of Diagnostic Systems under the FROC Paradigm: Experimental, Analytical, and Results Interpretation Issues

David Gur, ScD[1] and **Howard E. Rockette, PhD**[2]

[1]*Department of Radiology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15261*

[2]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261*

Abstract

As use of Free Response Receiver Operating Characteristic (FROC) curves gains more acceptance for quantitatively assessing the performance of diagnostic systems it is important that the experimentalist understands the possible role of this approach as one of the experimental design paradigms that are available to him/her amongst all other approaches as well as some of the issues associated with FROC type studies. In a number of experimental scenarios, the FROC paradigm and associated analytical tools have theoretical and practical advantages over both the binary and the ROC approaches to performance assessments of diagnostic systems but it also has some limitations related to experimental design, data analyses, clinical relevance, and complexity in the interpretation of the results. These issues are rarely discussed and are the focus of this paper.

In medicine in general and radiology in particular when an observer becomes an integral part of a diagnostic system there are often difficulties assessing the performance of systems in an objective generalizeable manner. We often become enamored by the development of new and more complex methodologies that reflect more subtle details of the evaluation performed by the human observers and for very good reasons. However, these studies are difficult to design and perform, are extremely costly, and may lead to an analysis that losses statistical power for actually clinically relevant questions because of the attempt to distinguish the effects of different aspects of the performance assessments. As important, generalizeability of results depends, among others, on the assumptions that are built into the study design. We should not forget that observer performance studies are designed in the hope of appropriately addressing a relevant clinical question in a manner that will withstand the test of time, hence enable important clinical practice decisions to be made. A specific question to be answered should lead to a less burdensome, practical (doable) and hopefully an optimal specific study design. The specific analytical tool to be used for analyses is but one of several available to the investigator in his/her "tool box" and the one selected should be the one most likely to provide reliable but clinically relevant conclusions [1,2]. Obviously the experimentalist always desires to minimize the required sample size in terms of both readers and cases that are needed for

Correspondence: David Gur, University of Pittsburgh, Department of Radiology, Radiology Imaging Research, 3362 Fifth Avenue, Pittsburgh, PA 15213-3180, Phone : 412-641-2513, Fax : 412-641-2582, Email : gurd@upmc.edu.

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.

inference generation; hence, statistical power is always an important consideration in selecting a specific study design [3,4].

Recent developments in the field of observer performance experiments resulted in significant improvements of a very valuable tool, namely the Free Response Receiver Operating Characteristic (FROC) paradigm and the possibility of performing and analyzing studies under this approach. FROC methodology takes into consideration location of the suspected abnormality and allows for more than one location to be identified as suspicious [5,6]. As a result, the FROC approach enables detecting the differences within-subject (location-based) diagnostic performances which are ignored in the subject-based ROC analysis [7]. Both parametric and nonparametric approaches can be used to analyze FROC data [8–11]. Some of the parametric approaches [8] attempt to describe latent (unobserved) characteristics of a search process conducted specifically by human observers. As investigators learn more about the FROC approach and begin to understand it we should remember that, like all other approaches that yield tools for analyses of observer performance studies, FROC has both advantages and disadvantages that need to be understood. As others emphasize the advantages, and there are many, the experimentalist should be aware of some practical issues associated with the FROC approach and address these issues in his/ her study design.

First, unlike ROC where an overall rating is provided for an image / case, under the FROC approach observers are free to mark as many suspicious regions as they wish. Although some alternatives have been proposed [12], it is currently a standard practice that to analyze the results of these studies an “acceptance target” has to be determined [12]. This target defines the distance from (or specific locations on) the center of the abnormalities in question that if marked inside this distance the observer gets the credit for “detecting” the suspicious region. Obviously the size and/or shape of the acceptance target affect the results of the study as described by Chakraborty et al. [12]. We should remember that, in standard FROC studies, at one extreme when the full image represents an acceptance target all marks on actually positive images (cases) are considered positive findings. At the other extreme, where acceptance target is but one pixel (voxel) all marks (whether actually positive or not) are likely to be assigned as negative findings; hence, by definition the results will be affected by the acceptance target and the number of actual abnormalities present.

In many situations the acceptance target of different abnormalities could overlap and one has to a priori decide how to address the marks that belong to the intersection. Some of the possible approaches can involve setting smaller targets or having a rule based approach to address this issue. This 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). Incorrect handling of the marks that “hit” both lesions could have a significant impact on the actual relevance of the results of these studies. Investigators should be aware of this and other similar issues when performing FROC studies with multiple abnormalities, some of which are in close proximity.

Correct handling and assignments of ratings becomes an even more difficult issue when more than one image per examination is provided to the observer (e.g. 2 view mammography or 2 view chest) and “paired” markings are ascertained between the two presentations of the same abnormality (on each of the images). When analyzing a “case based” (e.g. breast based) performance using both ratings of an abnormality by averaging the two ratings, or using the maximum rating between the two, correct assignments of the ratings to the appropriate abnormality is of utmost importance. The effect of “acceptance target” aggravates the problem even further. Therefore, decision rules regarding how these situations are considered and accounted for should be addressed (and stated) apriori.

Second, the current FROC approaches ignore the possible difference in the distributions of numbers and ratings of false positive marks in the positive and negative cases. If the distributions were similar in reality one could perform studies with only positive cases. We know that similarity of the distributions is theoretically a very convenient assumption but it is also a restrictive one if indeed there is a “satisfaction of search effect” in observer performance studies and as important perhaps there is a different type of satisfaction of search in positive versus negative images (or cases).

Third, often in diagnostic clinical medicine in general and during screening in particular the decision about a patient is not whether or not he has cancer but rather whether or not the procedure in question leads to suspicious findings that warrant additional diagnostic workup. The FROC paradigm is not designed for this purpose and it is not easy to generate a summary index (in particular when more than one abnormality is present in a case or the abnormality in question is depicted on more than one view) that would mimic in some way the subject-based diagnostic decision process. In this respect the question about the applicability of FROC summary indices to Computer Aided Detection (CAD) performance is an important one, namely the use of FROC may enable one to identify small differences between schemes at the individual mark level but not at the image or the examination level. To obtain the latter a case-based approach may be needed and a straightforward, albeit somewhat arbitrary, reduction of the FROC to a ROC analysis is possible [13,14]. However, when doing so, one loses some of the advantages of the FROC approach while complicating the study methodology.

We also need to remember that performance in the diagnostic arena is frequently measured differently than that in the screening one. While the tools we discuss here may be quite relevant and appropriate in both areas, in many instances location is not as primary a factor in the screening arena; hence, observer performance studies in this context may be best served by binary or ROC type ratings and analyses. At the same time, if one takes the more “global” view that the primary goal of screening is to segment a population with low prevalence of disease into two groups, namely “negative” (or “come back for your regularly scheduled screening examination”) and “recommended for diagnostic work-up” (or “it should be investigated further”), the reality is that often additional abnormalities are found and investigated as a result of this recommendation during diagnostic workup and the exact location of the originally suspected abnormality is important but not sufficient to fully provide a summary index of performance of the screening practice. Hence, the ultimate validity of the FROC approach is not always optimal in this regard.

There is no doubt that in all systems where the output ratings are generated and the “search” for suspicious points (or regions) is complete, namely all pixels or voxels are evaluated in the same manner during the initial stage (e.g. CAD system), the FROC paradigm provides a clear advantage of taking into account a correct (or incorrect) localization which is ignored in subject-based ROC approach. Moreover, the underlying approach used in an FROC study, namely identify (mark) suspicious locations and then provide a likelihood rating of the suspected region being positive at each location is a natural progressive approach to address this problem. Much has been written about the use of the FROC approach for this purpose [15,16]. Whether or not the initial stage of identifying all suspicious locations, as typically performed by CAD systems, can be efficiently described by a single process, as in FROC paradigm [8], is an issue that needs to be addressed, because pruning of initially suspected false positive regions in CAD systems is often done in several stages, but the FROC approach seems to be satisfactory for this purpose in many of the instances investigated to date. In the case of human observers the pruning mechanism can also have certain latent components, in the sense that more locations could have been actually noted than finally reported [6].

Lastly, for addressing clinically relevant issues the performances under Binary, ROC, and FROC paradigm can be summarized with True Positive Fraction and False Positive Fraction (or False Positive Rate in FROC) and combined with the utilities of different types of classifications (i.e. True Positive, True Negative, False Positive, and False Negative) [17]. However, in contrast to the Binary and ROC approaches, the FROC curve specifically characterizes the performance at the location level and hence is more difficult to determine utilities of correct localization. For addressing the overall performances under the Binary and ROC paradigm there are some commonly accepted, reasonably intuitive, easily estimable and interpretable indices (e.g. Youden's index, area under the ROC curve). However, there is still no widely accepted index of the overall performance under the FROC paradigm. Although there are several appealing indices in this regard based on the area under the FROC curve [8, 10] they do not have as simple a probability interpretation as does the area under the ROC curve. The development and validation of easily computable, intuitively interpretable, clinically relevant, summary indices that take into account all experimentally ascertained data remains an important area for future investigations. Finally, the FROC data can be viewed as clustered data with random cluster sizes and the analysis of such data is complicated by the necessity to account for the correlations and additional sources of variability. Hence, while more elaborate, FROC is not always the method of choice for data analyses.

We believe that the clinical question being posed should be the overriding factor in determining which paradigm one should use to assess and compare systems and practice performance levels. The tool to be used for analyses is just that namely a tool, and its properties or elegance cannot override the practical objective of a specific study. The assumption that the search process can be adequately described by the existing "search models" and FROC tools is yet to be proven in the actual experimental domain. If indeed we are convinced that this approach better represents the actual observer than by all means we should try to use it whenever we can.

We attempted to address here practical issues related to the use of a FROC approach to data collection, analysis, and interpretation of results. There are many other practical issues related to case selection, training of observers, and generalizeability of results from the laboratory experiment to the clinical environment that are applicable to all methodologies for performance assessment studies in general and in observer performance studies that include the observer as an integral part of the diagnostic system or the clinical practice of interest, in particular. These have been discussed elsewhere and are beyond the scope of this article [1,18]. FROC is an extremely important methodology that should be available for consideration in the "tool box" of every experimentalist and it should be employed when appropriate. However, it is ultimately the responsibility of the investigator to understand that like all other methodologies FROC should be the preferred approach in specific scenarios and that like all other approaches it has both advantages as well as limitations.

Acknowledgements

This work is supported in part by Grants EB006388 and EB002106 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

References

1. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol 2007;14(6):723–748. [PubMed: 17502262]Review
2. Krupinski EA, Jiang Y. Anniversary Paper: Evaluation of medical imaging systems. Med Physics 2008;35(2):645–659.
3. Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Acad Radiol 2004;11(11):1260–1273. [PubMed: 15561573]

4. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Academic Radiology 1995;2(8):709–716. [PubMed: 9419629]

5. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 2004;31(8):2313–2330. [PubMed: 15377098]

6. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol 2006;51(14):3449–3462. [PubMed: 16825742]

7. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur DA. comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. Med Phys 2005;32(4):1031–1034. [PubMed: 15895587]

8. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys 2002;29(12):2861–2870. [PubMed: 12512721]

9. Yoon HJ, Zheng B, Sahiner B, Chakraborty DP. Evaluating computer-aided detection algorithms. Med Phys 2007;34(6):2024–2038. [PubMed: 17654906]

10. Bandos AI, Rockette HE, Song T, Gur D. Area under the free-response ROC curve (FROC) and a related summary index. Biometrics. 2008(in press)

11. Samuelson, FW.; Petrick, N. Comparing image detection algorithms using resampling. Biomedical Imaging: Macro to Nano; 3rd IEEE International Symposium Arlington; April 6–9; Virginia. 2006. p. 1312-1315.

12. Chakraborty D, Yoon HJ, Mello-Thoms C. Spatial localization accuracy of radiologists in free-response studies: Inferring perceptual FROC curves from mark-rating data. Acad Radiol 2007;14(1): 4–18. [PubMed: 17178361]

13. Chakraborty DP. ROC curves predicted by a model of visual search. Phys Med Biol 2006;51(14): 3463–3482. [PubMed: 16825743]

14. Song T, Bandos AI, Rockette HE, Gur D. On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data. Med Phys. 2008(in press)

15. van Engeland S, Karssemeijer N. Combining two mammographic projections in a computer aided mass detection method. Med Phys 2007;34(3):898–905. [PubMed: 17441235]

16. Wei J, Chan HP, Sahiner B. Dual system approach to computer-aided detection of breast masses on mammograms. Med Phys 2006;33(11):4157–4168. [PubMed: 17153394]

17. Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population of readers and cases. Med Decis Making 2004;24(6):561–572. [PubMed: 15534338]

18. Gur D, Bandos AI, Cohen CS, et al. The “Laboratory” Effect: Comparing Radiologists’ Performance and Variability during Clinical Prospective and Laboratory Mammography Interpretations. Radiology. in press