ROC or FROC? It depends on the research question

Stephen L. Hillis, Ph.D.
Departments of Radiology and Biostatistics, The University of Iowa, Iowa City, Iowa 52242-1077, USA
(Tel: 319-338-0581; E-mail: steve-hillis@uiowa.edu)

Dev P. Chakraborty, Ph.D.
ExpertCAD Analytics, LLC, 2103 Noble Court, Murrysville, Pennsylvania 15668, USA
(Tel: 412-349-5305; E-mail: dpc10ster@gmail.com)

Colin G. Orton, Ph.D., Moderator
(Received 27 January 2017; accepted for publication 2 February 2017; published 17 March 2017)

[https://doi.org/10.1002/mp.12151]

OVERVIEW

Receiver Operating Characteristic (ROC) and Free-Response Operating Characteristic (FROC) methods are used to assess the accuracy of radiological imaging systems. ROC methods analyze an observer’s confidence that an abnormality is or is not present, whereas FROC methods additionally require the observer to locate abnormalities. Typically, ROC and FROC methods are applied to answer different research questions: sometimes ROC is the most appropriate and sometimes FROC. However, some believe that ROC is usually either equivalent or inferior and FROC is preferred over ROC for all research questions. This is the topic debated in this month’s Point/Counterpoint.

FOR THE PROPOSITION: Stephen L. Hillis, Ph.D.
Opening Statement

When comparing imaging modalities in a diagnostic radiologic observer study, what type of data should a researcher collect? Receiver operating characteristic data consist of likelihood-of-disease ratings, one for each case (i.e., patient); free-response ROC data consist of localization (i.e., specification of location) of suspected diseased areas (e.g., malignant tumors), referred to as targets, and target-specific likelihood-of-disease ratings; and localization ROC (LROC) consists of both types of data. I argue that the appropriate data type and corresponding analysis combination is the one that best answers the research question.
In many medical centers, screening mammography recalled cases will undergo extensive further evaluation, making target location of minor importance. Thus, for these centers a researcher might ask, “Which modality is best for classification of cases as diseased versus non-diseased?” Here, an ROC approach is appropriate. An ROC curve answers the question, “If a reader incorrectly classifies X% (e.g., X = 10) of nondiseased cases, what percent of diseased patients does the reader classify correctly?” The ROC area-under the curve (AUC) estimates the probability that a reader will correctly classify a randomly chosen pair of diseased and nondiseased cases.

In contrast, for diagnostic mammography for patients with suspicious screening mammograms, accurate localization of all actual targets is necessary to ensure appropriate treatment. Thus, a researcher might ask, “Which modality is best for classification when accurate localization of targets is needed?” Here, a modified LROC approach that requires a reader to accurately localize all of a patient’s actual targets is appropriate. The resulting LROC curve answers the question, “If a reader incorrectly classifies X% of non-diseased cases, for what percent of diseased patients does this reader jointly provide correct classification and accurate target localization?” The LROC AUC estimates the probability that a reader will correctly classify a randomly chosen diseased/nondiseased pair of cases and provide accurate target localization.

Now consider the frequently used adjusted FROC (AFROC) analysis. An estimated AFROC curve answers the question, “If a reader incorrectly classifies X% of non-diseased cases, what percent of actual targets (across patients) will the reader accurately localize?” This statement does not answer either of the two previous research questions. The corresponding jackknife AFROC (JAFROC) summary statistic estimates the probability that a randomly selected actual target will be rated higher than the maximum rating given to a randomly selected normal image, which is not clinically relevant to the previous two research questions. On the other hand, AFROC seems more suitable than ROC or LROC for assessing performance of a computer algorithm used in computer-aided diagnosis (CAD) to suggest sites for a human reader to examine.

In conclusion, different approaches estimate different aspects of reader performance and hence answer different research questions. Furthermore, a researcher may want to estimate two or more aspects of reader performance. In the words of Charles E. Metz: “How effective is a particular diagnostic imaging procedure? . . . To address the question in a meaningful way, we must decide exactly what information is sought, and in answering we must state precisely what information we are giving.”

**AGAINST THE PROPOSITION: Dev P. Chakraborty, Ph.D.**

**Opening Statement**

The FROC-paradigm radiologist marks and rates suspicious regions. Based on a proximity criterion, marks close to lesions are credited as correct localizations. ROC is a subset of FROC: if the proximity criteria are large enough, and the radiologist knows it, the two paradigms are indistinguishable: specifically, the radiologist will make at most one mark/case and unmarked cases are “definite normals.” A FROC model predicts ROC curves, but one cannot go the other way. For interstitial lung disease, where location is implicit, ROC is appropriate, but then so is FROC. However, in clinical tasks which involve finding focal-disease, for example, screening mammography or lung nodules, if the radiologist suspects the patient is diseased, there is at least one associated suspicious location. For these tasks the ROC paradigm obtains a rating that there is disease “somewhere,” which begs the question: if disease is “somewhere”, why not point to it (Prof. Gary Barnes, private communication, ca. 1985)? In fact they do—radiologists mark and annotate suspicious regions, but the ROC paradigm ignores this information, leading to loss of statistical power relative to FROC. It is unethical to use a method with lower statistical power when one with greater power is available. Over 104 publications, mostly non-US, have used JAFROC to analyze FROC studies. Clinicians have long recognized the importance of accounting for localization, and a leading statistician has recognized it. Yet there is opposition to JAFROC within the US. Here, is a recent reviewer comment: “…the JAFROC statistic…does not yield a meaningful clinical interpretation.” I ask medical physicists: if the probability that lesions are rated higher than nondiseased cases (the JAFROC statistic) equals unity, is this a good thing? I hope your answer is a resounding “yes” because a unit value means all diseased patients are correctly recalled and no nondiseased patients are incorrectly recalled. (A zero value reverses correct/incorrect in the preceding sentence). This is the clinical interpretation the reviewer finds so elusive. The subject of this debate is not “rocket science”, but it does require one to be open-minded and unbiased. Dirac addressed an analogous then-existing criticism against quantum mechanics, namely, it did not provide a “satisfying picture” (translate “satisfying picture” to the reviewer’s “meaningful clinical interpretation”) as did classical mechanics. To paraphrase Dirac, the purpose of science is not to provide satisfying “pictures” but to explain phenomena. Since it allows zero or more mark-rating pairs per image, FROC is inherently more complex than ROC, no doubt about it. More importantly, it mirrors clinical practice, which is also more complex than the “somewhere” that the ROC paradigm accommodates. It is about time to stop using “I don’t understand” as an excuse for impeding scientific progress and patient care. ROC methods should not be used to analyze localization tasks.

**Rebuttal: Stephen L. Hillis, Ph.D.**

An ROC analysis performed using FROC data by treating the highest target rating per case as the decision variable has been called an inferred ROC analysis. This is different from a conventional ROC analysis of ROC data where the decision variable is a case-specific overall likelihood-of-disease rating.
Although the two analyses are equivalent under the assumption that the two decision variables provide the same case rankings, this assumption has never been conclusively demonstrated empirically.\textsuperscript{15} Furthermore, the assumption is intuitively not reasonable; for example, it implies that two cases are considered by a reader to have equal likelihood of disease if one case receives one mark and the other receives multiple marks, with each marked site rated as having 80% disease probability. Thus, inferred ROC answers a different question than conventional ROC. Although I disagree with Dr. Chakraborty’s statement that “\textit{ROC is a subset of FROC}.” I have no problem agreeing that \textit{inferred ROC} is a subset of FROC.”

Dr. Chakraborty has demonstrated through simulations that FROC (specifically, JAFROC) is more powerful than inferred ROC (conventional ROC is not included in these simulations). The problem with this kind of comparison is that JAFROC and inferred ROC have different hypotheses, and hence answer different research questions.

In conclusion, I agree that FROC data are required for assessing reader performance with respect to localization. However, this does not mean that one FROC analysis method (e.g., JAFROC) is suitable for all or most research questions. JAFROC is a statistically viable approach, but it should only be used if it answers the research question. A future area for research is development of FROC-ROC analysis methods designed to answer specific research questions. Before new analysis methods can be developed, however, there is a need for identifying important research questions and providing precise statements of them.

\textbf{Rebuttal: Dev P. Chakraborty, Ph.D.}

Let me address some of the statements made by my distinguished colleague, which I dispute:

- Location-specific methodologies (LROC/FROC/ROI) were developed not to address some abstract research question, but to account for clinical reality.
- In my opinion, Dr. Hillis has the roles of screening and diagnostic mammography reversed. While screening mammography results in a binary decision (recall: yes/no?), radiologists also report locations of suspicious regions. Diagnostic mammograms are used to further investigate these suspicious regions.\textsuperscript{16} There is an analogous difference between CADe (screening) and CADx (diagnostic).\textsuperscript{16} The starting point for diagnostics is not that there are suspicious regions “somewhere in the breast”, but that there are specific regions found at screening. Yet Dr. Hillis supports using ROC methodology, where location is ignored, to analyze screening mammography. As justification, he repeats an incorrect argument\textsuperscript{8} that, because recalled cases will undergo “extensive evaluation”, location is of “minor importance”. The fallacy is that at the end of the extensive evaluation one is down to a few localized regions, whose truth is established by needle-biopsy, that is, \textit{one is down to FROC data}.

- The issue is clinical, not statistical. Lesions are location-level manifestations of patient-level disease. A malignant lesion means the patient (not the lesion) has breast cancer. The lesion is not recalled—the patient is. If a lesion is rated higher than a nondiseased case, the patient who is attached to it is also being effectively rated higher, as accounted for in the weighted AFROC (wAFROC) figure-of-merit, which is a case-level figure-of-merit, where every diseased case effectively contributes exactly one lesion.\textsuperscript{17} If the case is regarded as a random factor, results extrapolate to the population of cases. A “population of lesions”, as Dr. Hillis’ second-last paragraph seems to imply, is a contradiction in terms, as lesions have no independent existence.
- If an incorrect location is identified in a diseased case, then the recall is technically “correct” but for the wrong reason\textsuperscript{18}—two canceling errors actually occurred, a missed lesion and a location-level false positive. A modality that minimizes “right for wrong reason” outcomes would have an advantage when analyzed by wAFROC-AUC figure-of-merit\textsuperscript{5} but not when analyzed by ROC-AUC, because in ROC the two canceling errors count as a perfect decision.

\textbf{ACKNOWLEDGMENTS}

Dr. Hillis thanks Craig Abbey (U.C. Santa Barbara), Kevin Berbaum (U. Iowa), Patrick Brennan (U. Sydney), Leonel Vasquez (U. Iowa), and Tamara Miner Haygood (U. Texas) for their helpful discussions. This research was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB019967.

\textbf{CONFLICTS OF INTEREST}

The authors have no relevant conflicts of interest to disclose.

\textbf{REFERENCES}


