Published in final edited form as:
Acad Radiol. 2013 July ; 20(7): 915–919. doi:10.1016/j.acra.2013.03.001.

A brief history of FROC paradigm data analysis

Dev P. Chakraborty, Ph.D.
Department of Radiology, University of Pittsburgh, Presbyterian South Tower, Room 4771, 200 Lothrop Street, Pittsburgh, PA 15213, 412-605-1553 (p), 412-605-1554 (f), dpc10@pitt.edu, 412-605-1553 (phone), 412-605-1554 (fax)

Abstract

In the receiver operating characteristic (ROC) paradigm the observer assigns a single rating to each image, and the location of the perceived abnormality, if any, is ignored. In the free-response receiver operating characteristic (FROC) paradigm the observer is free to mark and rate as many suspicious regions as are considered clinically reportable. Credit for a correct localization is given only if a mark is sufficiently close to an actual lesion; otherwise the observer's mark is scored as a location-level false positive. Until fairly recently there existed no accepted method for analyzing the resulting relatively unstructured data containing random numbers of mark-rating pairs per image. This paper reviews the history of work in this field which has now spanned more than 5 decades. It introduces terminology used to describe the paradigm, proposed measures of performance (figures of merit), ways of visualizing the data (operating characteristics) and software for analyzing FROC studies.

Keywords

ROC; free-response; FROC; localization tasks; observer performance; JAFROC; software

INTRODUCTION

The term "free-response" was coined by Egan in 1961 in connection with studies involving the detection of brief audio tone(s) against a white-noise background [1]. The tone(s) could occur at any instant within an active listening interval (e.g., while an indicator light was on) and the listener's task was to respond by pressing a button at any instant(s) when a tone(s) was perceived. The listener was uncertain how many true tones, if any, could occur in the active interval and when they might occur. Therefore the number of responses per active interval could be ≥ 0 and was a-priori unpredictable. With two-dimensional space replacing time the acoustic study is analogous to a common task in medical imaging, namely, prior to interpreting an image for possible breast cancer the mammographer does not know a-priori how many lesions (i.e., cancers) are present, if any, and where they are located. Consequently the image must be searched for regions that appear suspicious for cancer. If the level of suspicion of a particular suspicious region exceeds the minimum clinical reporting threshold the mammographer reports it (at our institution they digitally outline and annotate the suspicious region). Conceptually a screening report consists of the locations of regions that exceed the threshold and the corresponding levels of suspicion (reported as a

BIRADS rating [2]). This type of information defines the free-response paradigm as it applies to medical imaging. At its essence free-response is a search paradigm.

The free-response receiver operating characteristic (FROC) curve was introduced, also in the auditory domain, by Miller as a way of visualizing performance in the free-response task [3]. The importance of the free-response paradigm for radiology applications was first recognized by Bunch et al in [4]. Their paper describes several ambiguities that arise when the receiver operating characteristic (ROC) method is applied to a localization task (the interested reader is referred to Table I in their paper). A well-known one is the ambiguity when a location-level false positive and a location-level false negative occur on the same image. The two mistakes effectively "cancel" each other in ROC analysis and the image is scored as a "perfect" image-level true positive. In other words the radiologist was right – a cancer containing image was diagnosed abnormal – but for the wrong reason –an incorrect lesion location was reported.

Bunch et al [4] conducted the first imaging FROC experiment. Under certain assumptions, appropriate to their data, they showed that it was possible to derive ROC operating points from FROC operating points and they also anticipated the alternative FROC (AFROC) curve. The author and colleagues at the University of Alabama at Birmingham were the first to apply the free-response paradigm to the clinical problem of comparing a prototype digital chest imaging device to a conventional analog device in a lesion localization task [5]. The method was soon applied in a second study [6] to evaluate a prototype dual-energy chest imaging system by the same manufacturer.

In an FROC study the number of marks on an image can be 0 or more, and must be regarded as a modality, reader and image dependent random variable. The randomness in the number of marks, in addition to the usual sources of randomness of the ratings due to image sampling and reader sampling, is the main reason why analysis of FROC data has been a challenge. Work in this area has now spanned over five decades. This paper traces the history of developments in free-response analysis.

FROC DATA: MARK-RATING PAIRS

The mark is the location of the suspicious region and the rating is the confidence level that the region contains a lesion. The data analyst decides whether a mark is close enough to a real lesion to qualify as lesion localization (LL) – a location-level "true positive" - and otherwise the mark is classified as non-lesion localization (NL) – a location-level "false positive". The quotes are intended to emphasize the confusion that can arise if one uses terminology developed for image-level ROC studies to location-level paradigms. What constitutes "close enough" (i.e., the proximity criterion or "acceptance radius") is a clinical decision which should be made based on the application [7–9]. Two physicians do not need to agree on the exact center of a lesion in order to appropriately assess and treat it. The proximity criterion should be similar for all modalities under comparison as otherwise there would be a bias favoring the modality with the more lenient criterion [8].

DATA ANALYSIS

Operating characteristics and figures-of-merit

Data analysis starts with the selection of a figure-of-merit (FOM) and a procedure for estimating it from the observed collection of NLs and LLs, each with an associated rating (the rating does not have to be a discrete integer). A valid figure of merit rewards the observer for correct decisions and penalizes for incorrect decisions. Finding a suitable figure of merit usually starts with a way of visualizing the data. For example, the ROC curve

suggests the area under the curve as a suitable figure of merit for ROC data. Bunch et al [4] suggested two ways of visualizing FROC data.

The FROC curve and associated figures of merit—The FROC curve was defined [4] as the plot of lesion localization fraction (LLF) vs. non-lesion localization fraction (NLF), where LLF is defined as the total number of lesion localizations at a given threshold divided by the total number of lesions, and NLF is defined as the total number of non-lesion localizations at that threshold divided by the total number of images (as discussed later, the mixing of normal and abnormal images in the definition of NLF has an undesirable effect). Pairs of (NLF, LLF) values can be plotted corresponding to the different cumulated ratings. For example, in a 5-rating FROC study, with increasing numbers representing increasing confidence in presence of lesion, one gets 5 FROC operating points as follows: the point corresponding to the 5's, the 5's and 4's, the 5's, 4's and 3's, the 5's, 4's, 3's and 2's and finally, the 5's, 4's, 3's, 2's and 1's. If continuous ratings are used then the procedure is to start with a high threshold such that none of the ratings exceed it, and slowly lower the threshold and count the total numbers of LLs and NLs exceeding the threshold and divide by the appropriate denominators – this yields the so-called "raw" FROC curve. For example, when a LL rating just exceeds the threshold, the operating point jumps upwards by 1/(total number of lesions) and when a NL rating exceeds the threshold, the operating point jumps to the right by 1/(total number of images).

The FROC curve is not contained within the unit square. While the y-axis is the probability *P(l)* that a lesion localization occurs, estimated by LLF, the x-axis is the mean number of non-lesion localizations per image, estimated by NLF (for notational symmetry the author terms this a "fraction" when in fact it is an improper fraction). The x-axis can potentially tend to large values, especially if the total lesion area is much smaller than the total image area, as was the case in the Bunch et al experiment. Partial area measures, such as the area under the FROC curve to the left of a predefined abscissa value or the value of the ordinate at the predefined abscissa have been used as figures of merit. The latter figure of merit was used by the author in the first clinical application of the FROC method [5].

The AFROC curve and associated figure of merit—Bunch et al [4] also introduced the plot of LLF vs. false positive fraction (FPF) which was subsequently termed the alternative FROC (AFROC) by the author [10]. Since the AFROC curve is completely contained within the unit square, since both axes are probabilities, the author suggested that, analogous to the area under the ROC curve, the area under the AFROC be used as a figure-of-merit for FROC performance [10,11].

In the author's experience the question of the end-point of the AFROC curve (reached when the decision threshold is infinitely low) often creates confusion. If every region in the image generates a finite decision variable sample, no matter how small, then when the observer's threshold is lowered to negative infinity all regions, including all lesions, will be marked and the end-point (1,1) will be reached trivially. The continuous approach to (1,1) is implicit in early models [10–12]. In newer models [13,14] not all regions generate decision variable samples and the observer generally cannot reach (1,1). Nevertheless, the area under the total AFROC curve, including that under the straight line extension from the uppermost reached point to (1,1), has to be included to properly credit perfect decisions such as normal images with no marks and to penalize unmarked lesions [15].

Estimating FPF from FROC data—The y-axis of the AFROC is identical to that of the FROC curve. So the problem is how to estimate FPF, which is an image-level (i.e., ROC) quantity from FROC data. In the ROC context FPF is an estimate of the probability *P(FP)* of observing a FP. It is estimated by counting the number of normal images declared abnormal

and dividing by the total number of normal images. When one has FROC data it is customary to take the rating of the highest rated NL on an image, and assume that is the ROC rating of the image (often termed the highest rating assumption).

In the Bunch et al study only simulated abnormal images were used (each image contained from 10 to 20 simulated lesions) so they needed a way of inferring the probability *P(FP)* of an image-level FP, from the non-lesion localization marks made on abnormal images only. They assumed that the probability $P(n)$ of observing an image with n non-lesion localizations is given by the Poisson distribution: i.e., $P(n) = e^{-\lambda} \lambda^n / n!$, where λ is the mean of the distribution, estimated by NLF, defined previously in connection with the FROC curve. The probability of observing an image with a false positive *P(FP)* is the complement of the probability of observing an image with zero non-lesion localizations and therefore $P(FP) = P(0) = 1 - e^{-\lambda}$. Since λ is estimated by NLF, this yields $FPF = 1 - e^{-NLF}$. This method can obviously be applied to a mixture of abnormal and normal cases, as is typical with most datasets [10].

Estimating the figure-of-merit: parametric methods

A parametrically fitted curve allows estimation of the figure-of-merit. Much previous work focused on parametric fitting of FROC [4,11,12] or AFROC curves [10]. All of these made untenable independence assumptions which drew valid criticisms [12,16]. For example, the models assume that the probability of occurrence of non-lesion localizations on an image is independent of the number of true lesions present in the image. In reality the probability of non-lesion localizations is typically larger on normal images than on abnormal images. As another example, they assumed that the lesion localization mark-rating pairs on an image are independent, when in fact satisfaction of search effects have been reported in which the observer, having marked one lesion, is less likely to mark other lesions present in the image [17].

Estimating the figure-of-merit: non-parametric methods

In recent years the focus has shifted towards non-parametric estimates of the figure of merit. In non-parametric ROC analysis one uses the trapezoidal area under the ROC curve as the figure of merit. The calculation is simple and no curve fitting is necessary: one compares all possible pairs of normal and abnormal images; if the abnormal image rating exceeds the normal image rating one cumulates unity in a zero-initialized counter variable. If the two ratings are identical one cumulates 0.5. This is done for all possible pairings of normal and abnormal images and the final result is divided by the total number of comparisons. This yields the empirical probability that an abnormal image rating exceeds that of a normal image and can be shown to be identical to the area under the trapezoidal ROC curve. Non-parametric FROC analysis is possible [18,19] along similar lines. One approach is to use the area under the raw FROC curve to the left of a specified value of the abscissa as the figure of merit. Image-level bootstrapping is used to estimate the 95% confidence interval for the FOM, or if two modalities are being compared, to estimate a 95% confidence interval for the difference in FOMs. [While it is true that the non-parametric FOM underestimates that derived by curve-fitting, since interest is generally in the difference between two FOMs the underestimates tend to cancel out, and this is generally not an issue.]

Other FROC figures of merit

The Λ figure of merit—A FROC curve-based figure of merit has been introduced [20] as the "area under the empirical FROC curve penalized for the number of erroneous marks, rewarded for the fraction of detected abnormalities, and adjusted for the effect of the acceptance radius". This index is defined by $\Lambda = A_0 - NLF_0 + LLF_0 / \varphi$. Here A_0 is the non-parametric area under the complete FROC curve, the zero subscript denotes the end-point of

the FROC curve (reached when all marks are cumulated), and the parameter φ, is defined as the ratio of the total area per image occupied by the acceptance regions surrounding each target to the rest of the image. The method has been applied in one clinical study [21]. The cited publications do not describe how φ is estimated in clinically realistic situations. The breast is not a homogeneous structure and not all regions within the skin-line are equally likely to have lesions so exactly what is meant by "rest of the image" in the definition of φ is not clear. This could lead to arbitrariness in the usage of this figure of merit.

The EFROC figure of merit—Popescu et al [22] proposed an exponential transformation, $1-e^{-NLF}$, of the abscissa of the FROC curve, and the resulting plot termed EFROC is fully contained within the unit square. This transformation is identical to that suggested by Bunch et al, but the latter needed to invoke the Poisson and highest rating assumptions to derive it, as opposed to simply postulating the transformation as Popescu et al did.

Since the uppermost operating point (reached by cumulating all the points) generally lies below (1,1), Popescu et al added a linear extension from the uppermost point to the (1,1) to obtain the total area under the EFROC (as opposed to a partial area measure). In the author's opinion the EFROC is a reasonable figure of merit provided one uses only normal images to estimate the abscissa. It has the advantage of using the multiple NLs on a normal image, unlike the AFROC approach which uses only the highest rated one (this will give the EFROC a statistical power advantage). The current limitation of the Popescu et al approach to statistically independent marks (they needed this assumption to derive non parametric estimates of standard error of the area under the EFROC) can be circumvented by resampling significance testing techniques, as described below.

Should one count NLs on both normal and abnormal images?

The FROC curve abscissa is traditionally defined over all images: NLs on abnormals and on normal images both contribute to NLF. If the observer's tendency to generate NLs is independent of the presence or absence of true lesions, this would be perfectly legitimate, but, as noted earlier, this is often not the case. At the same confidence level normal cases usually generate more NLs than abnormal cases. It follows that the (asymptotic) FROC curve will depend on the case mix, i.e., the ratio of abnormal to the total number of cases. Two investigators using different case mixes would get different (asymptotic) FROC curves even if all other conditions were identical. This is undesirable and the only way to remedy it is to define NLF over normal images only. Similar comments apply to the AFROC: FPF should be calculated over normal cases only. [By "asymptotic" we mean the infinite sample size limit, so as to eliminate sampling variability.] Good figures of merit should measure the observer's ability to correctly discriminate LLs from NLs as happens when LLs are rated higher than NLs. However, discriminating LLs from NLs on an abnormal image (these can be from the same image) is less clinically meaningful than discriminating LLs from NLs on a normal image (these have to be different images). Defining the curves over all images has the undesirable consequence of mixing the two types of discrimination abilities[1].

[1]Since this issue is dedicated to the memory of Prof. Charles Metz, a historical note is appropriate. Around May 2005 the author gave a talk on FROC analysis at the University of Chicago. When he stated that in JAFROC analysis one ignores the non-lesion localizations on abnormal images, Dr. Metz in the audience stated "That is a good idea". At that time it was believed that including the non-lesion localizations on abnormal images led to incorrect null hypothesis behavior, which was subsequently proven to be incorrect. So the right decision (at least in Dr. Metz's view) was made for the wrong reason.

JAFROC SOFTWARE

Figures of merit

The AFROC plot (LLF vs. FPF) is amenable to defining a non-parametric FOM. One compares all pairing of LLs and highest rated NLs on normal images. If the LL rating is greater, one cumulates unity; if they are equal one cumulates 0.5, and at the end of the process one divides by the total number of comparisons. Except for some nuances, described in a document available on the website www.devchakraborty.com, this is the figure of merit used in jackknife alternative FROC (JAFROC) analysis [23], currently the most widely used method for analyzing FROC data acquired using the multiple reader multiple case (MRMC) protocol where each reader interprets each case in all modalities. The software includes tools for sample size estimation and empirical and parametric plots of operating characteristics. Since its introduction in 2004, 46 publications have appeared that have used JAFROC software. The software supports various non-parametric figures of merit: the JAFROC figure of merit, a weighted JAFROC figure of merit, which corrects for the tendency of the JAFROC figure of merit to be dominated by abnormal images with relatively large numbers of lesions, the highest rating inferred ROC trapezoidal area under the curve, an average rating-based figure of merit [24] and a stochastic dominance based figure-of-merit [24].

Significance testing

The software performs case-level jackknifing and analyzes the resulting pseudovalue matrix using the Dorfman Berbaum and Metz (DBM) analysis of variance (ANOVA) algorithm developed for MRMC ROC data [25,26]. Because the DBM pseudovalue model is applicable to any scalar figure-of-merit [27,28], not just the area under the ROC curve, it is permissible to apply this method to a pseduovalue matrix derived from non-ROC paradigm data. In spite of the zero or more mark-rating pairs that could occur on a case, in the pseudovalue matrix *each case is represented by a single pseudovalue*. If case is treated as a random factor the analysis generalizes to the population of cases with lesion prevalence similar to that in the analyzed dataset. Since lesions do not exist without reference to images, it is incorrect to regard the analysis as "generalizing to the population of lesions". It is more accurate to state that the analysis generalizes to the population of images with lesion distribution similar to that in the sampled images.

Validation studies

JAFROC has been validated using simulations to generate FROC data in two statistically identical null hypothesis (NH) modalities. The nominal α (probability of a Type I error) of the test is set at 5%, and the expected NH behavior (the test should reject the NH in 5% of the simulations) has been confirmed [15,23,29,30].

DISCUSSION

This paper has summarized the history of research in free-response data analysis. The history of research in this field is essentially that of finding a good figure of merit and a method for testing the significance of the difference between two figures of merit. The significance testing methodology has benefited immensely from work by DBM [25] and subsequent refinements by Hillis et al [26,31–35].

Not discussed in this paper are the modeling advances that have taken place in connection with FROC research [13–15,36–39]. Besides allowing one to simulate realistic FROC data in order to validate proposed methods of analysis, they have yielded insight into the effect of search on performance.

There are other approaches besides free-response to accounting for localization information. In the localization ROC (LROC) approach the observer gives a single rating to the image and marks the most suspicious region in the image [12,40–42]. In the region of interest (ROI) approach [43,44] the investigator divides the image into a number of ROIs and the observer's task is to rate each region for presence of disease. All of these approaches, and indeed the ROC approach, have roles in imaging system assessment. Based on a good understanding of the clinical task one should select the paradigm that most closely resembles it. For example, in diagnostic mammography one is interested in determining whether or not a lesion already identified at screening is in fact a cancer or benign. The ROC paradigm should be used in this binary discrimination task.

Acknowledgments

The author is grateful to Ms. Kun-Wan Chen for proofing the manuscript. 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. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. J Acoust Soc Am. 1961; 33:993–1007.
2. Horsch K, Giger ML, Metz CE. Potential Effect of Different Radiologist Reporting Methods on Studies Showing Benefit of CAD. Academic Radiology. 2008; 15:139–152. [PubMed: 18206613]
3. Miller H. The FROC curve: a representation of the observer's performance for the method of free response. The Journal of the Acoustical Society of America. 1969; 46:1473–1476. [PubMed: 5361517]
4. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. J of Appl Photogr Eng. 1978; 4:166–171.
5. Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes GT, et al. Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules. Radiology. 1986; 158:35–39. [PubMed: 3940394]
6. Niklason LT, Hickey NM, Chakraborty DP, Sabbagh EA, Yester MV, et al. Simulated Pulmonary Nodules: detection with Dual-Energy Digital versus Conventional Radiography. Radiology. 1986; 160:589–593. [PubMed: 3526398]
7. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. Medical Physics. 1999; 26:267–275. [PubMed: 10076985]
8. Chakraborty DP, 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; 14:4–18. [PubMed: 17178361]
9. Haygood TM, Ryan J, Brennan PC, Li S, Marom EM, et al. On the choice of acceptance radius in free-response observer performance studies. The British journal of radiology. 2012
10. Chakraborty DP, Winter LHL. Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment. Radiology. 1990; 174:873–881. [PubMed: 2305073]
11. Chakraborty DP. Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys. 1989; 16:561–568. [PubMed: 2770630]
12. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys. 1996; 23:1709–1725. [PubMed: 8946368]
13. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys Med Biol. 2006; 51:3463–3482. [PubMed: 16825743]
14. 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]
15. Chakraborty DP. New Developments in Observer Performance Methodology in Medical Imaging. Semin Nucl Med. 2011; 41:401–418. [PubMed: 21978444]

16. Metz, CE. Evaluation of digital mammography by ROC analysis. In: Doi, K., editor. Digital Mammography '96. Amsterdam, the Netherlands: Elsevier Science; 1996. p. 61-68.
17. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Kathol MH, et al. Satisfaction of Search in Diagnostic Radiology. Invest Radiol. 1990; 25:133–140. [PubMed: 2312249]
18. Samuelson, FW.; Petrick, N. Comparing image detection algorithms using resampling; IEEE International Symposium on Biomedical Imaging: From Nano to Micro; 2006. p. 1312-1315.
19. Samuelson, FW.; Petrick, N.; Paquerault, S. Advantages and examples of resampling for CAD evaluation; 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro; 2007. p. 492-495.
20. Bandos AI, Rockette HE, Song T, Gur D. Area under the Free-Response ROC Curve (FROC) and a Related Summary Index. Biometrics. 2008; 65:247–256. [PubMed: 18479482]
21. Gur D, Bandos AI, Rockette HE, Zuley ML, Sumkin JH, et al. Localized Detection and Classification of Abnormalities on FFDM and Tomosynthesis Examinations Rated Under an FROC Paradigm. American Journal of Roentgenology. 2011; 196:737–741. [PubMed: 21343521]
22. Popescu LM. Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve. Medical Physics. 2011; 38:5690–5702. [PubMed: 21992384]
23. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. Med Phys. 2004; 31:2313–2330. [PubMed: 15377098]
24. Song T, Bandos AI, Rockette HE, Gur D. On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data. Medical Physics. 2008; 35:1547–1558. [PubMed: 18491549]
25. Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. Invest Radiol. 1992; 27:723–731. [PubMed: 1399456]
26. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol. 2008; 15:647–661. [PubMed: 18423323]
27. Beiden SV, Wagner RF, Campbell G. Components-of Variance Models and Multiple-Bootstrap Experiments: An Alternative Method for Random-Effects, Receiver Operating Characteristic Analysis. Academic Radiology. 2000; 7:341–349. [PubMed: 10803614]
28. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. Statistics in Medicine. 2005; 24:1579–1607. [PubMed: 15685718]
29. Chakraborty DP. Validation and Statistical Power Comparison of Methods for Analyzing Free-response Observer Performance Studies. Acad Radiol. 2008; 15:1554–1566. [PubMed: 19000872]
30. Chakraborty, DP. Recent developments in free-response methodology. In: Samei, E.; Krupinski, E., editors. The Handbook of Medical Image Perception and Techniques. Cambridge: Cambridge University Press; 2010. p. 216-239.
31. Hillis SL, Berbaum KS. Power Estimation for the Dorfman-Berbaum-Metz Method. Acad Radiol. 2004; 11:1260–1273. [PubMed: 15561573]
32. Hillis, SL.; Berbaum, KS. Recent developments in the Dorfman-Berbaum-Metz (DBM) procedure for multi-reader ROC study analysis; Medical Image Perception Society (MIPS) Conference XI; 2005.
33. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudovalues and less data-based model simplification. Acad Radiol. 2005; 12:1534–1541. [PubMed: 16321742]
34. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC studies. Statistics in Medicine. 2007; 26:596–619. [PubMed: 16538699]
35. Hillis, S. Multireader ROC analysis. In: Samei, E.; Krupinski, E., editors. The Handbook of Medical Image Perception and Techniques. Cambridge: Cambridge University Press; 2010. p. 204-215.
36. 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; 29:2861–2870. [PubMed: 12512721]

37. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 2011; 648:S297–S301.

38. Chakraborty, DP. Recent developments in imaging system assessment methodology. Stockholm, Sweden: 2010.

39. Chakraborty DP, Yoon H-J, Mello-Thoms C. Inverse dependence of search and classification performances in lesion localization tasks. Proc SPIE. 2012; 8318:83180H.

40. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology. 1975; 116:533–538. [PubMed: 1153755]

41. Starr SJ, Metz CE, Lusted LB. Comments on generalization of Receiver Operating Characteristic analysis to detection and localization tasks. Phys Med Biol. 1977; 22:376–379. [PubMed: 854532]

42. Swensson RG, Judy PF. Detection of noisy visual targets: Models for the effects of spatial uncertainty and signal-to-noise ratio. Perception & Psychophyics. 1981; 29:521–534.

43. Obuchowski NA, Lieber ML, Powell KA. Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography. Acad Radiol. 2000; 7:516–525. [PubMed: 10902960]

44. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient-clustered data. Acad Radiol. 2000; 7:413–419. [PubMed: 10845400]