Problems with the Differential Receiver Operating Characteristic (DROC) Method

Dev P. Chakraborty
University of Pittsburgh, Department of Radiology,
3520 5th Avenue, Pittsburgh, PA 15213

ABSTRACT

Most papers in these proceedings present ideas that work. This is the story of an idea that did not work as intended. The differential receiver operating characteristic (DROC) method was proposed about 8 years ago. It was intended to measure the difference in performance between two imaging modalities. It was expected that the DROC method could outperform the ROC method in statistical power. This expectation has not been borne out and the author no longer recommends the DROC method. The purpose of this paper is to present a critical look at this method, why the author initially believed it should work, the assumptions involved and the fallacies. The author believes there is value to this frank account as it has yielded, at least for the author, new insights into ROC analysis. The author concludes with a few personal reflections on his experience with this project and advice on how to deal with negative results.

Keywords: Observer performance, ROC, DROC, differential ROC, modality evaluation

1. INTRODUCTION

A significant goal is finding alternatives to ROC studies that are more clinically more relevant and statistically more powerful **[1]**. Achieving this goal would allow imaging system optimization to be conducted more rapidly and with less chance of missing the beneficial effects of the optimal system. Others are working on non-ROC methods **[2], [3]** and the author has introduced a few non-ROC methods of his own. About eight years ago the author became interested in a method of evaluating imaging modalities that appeared to have more statistical power than the standard receiver operating characteristic (ROC) method. The latter **[4-6]** involves acquiring data with many cases and several readers, which can be time consuming and expensive.

The particular approach the author undertook was motivated by image processing algorithm evaluations performed prior to the widespread usage of the ROC method. In these studies radiologists are asked to give their assessment of clinical image quality, which is subjective in the sense that the reader does not pay a price for blindly preferring a modality, in other words there is no control. Studies like these are still occasionally submitted to journals (and sometimes they get accepted) but when the author is asked to review them, he usually states that they amount to a "beauty-contest", totally subjective and hence not scientifically valid. The concern is that the radiologist may have a bias for a particular aesthetic look of the image, and will *prefer* the image-processing algorithm that delivers that look. To the extent that the preference is unrelated to actual diagnostic performance, the results of such preference studied might be misleading. [Note that in the clinic the result of a blind preference for a modality can have significant negative repercussions for the business practice of the radiologists, so there is, in fact, a control, which is difficult to implement in the laboratory.]

Another way of stating the problem with blind preference is that the preferred image processing algorithm may improve the detectability of a particular structure that is actually present (true positive) but, if it also improves the detectability of spurious structures (false positives) that is not good. This suggested a possible way of controlling for the subjectivity – why not make two measurements, one to evaluate the effect of the image processing algorithm on signal enhancement, and one to evaluate its effect on noise enhancement and take the difference. This was the basis of the differential receiver operating characteristic (DROC) method.

Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment,
edited by Dev P. Chakraborty, Miguel P. Eckstein, Proceedings of SPIE Vol. 5372
(SPIE, Bellingham, WA, 2004) · 1605-7422/04/$15 · doi: 10.1117/12.533300

2. EXPERIMENTAL METHODS

The DROC method
In the DROC method the observer is shown a pair of images of the same patient as rendered by the two modalities (A and B). The questions to the observer are is the patient normal or abnormal and which image of the image pair yields higher confidence in the normal-abnormal diagnosis?

The geometrical picture
A geometrical picture was used to guide the analysis of this data. The ROC model **[7]** uses the concept of a decision variable (DV) to analyze ratings data. It is assumed that there exist two distributions, one corresponding to normal cases and one to abnormal cases, referred to as the noise (N) and signal (S) distributions, respectively. The geometrical picture for the DROC study (misleading in retrospect) that was adopted was that corresponding to each modality one had a pair of such distributions: i.e., for modality-A one had N_A and S_A and for modality-B one had N_B and S_B and the two DV axes coincided. See Fig. 1.

There are two ways of showing that, as depicted in Fig. 1, the net separation of the abnormal and normal distributions is larger for modality-B than for modality-A. The standard way is to measure the separations separately for each modality and then take the difference. This requires two ROC studies, corresponding to the two modalities, conversion of the area under the curve (AUC) to a detectability index d **[8]**, and finally one takes the difference of the two detectability indices. Note that the final result [$\Delta d = d(B) - d(A)$] does not require the two DV-axes to be being collinear. The two axes could in fact be perpendicular, and the result would still be true.

In the second method one measures the separation of the two signal distributions (labeled A+, B+) and similarly the separation of the two noise distributions (labeled A-, B-) and takes *their* difference. How does one measure the separation of the two signal distributions? According to Fig. 1 one shows the same signal (abnormal) case in each modality, and asks the reader which image is more "positive for abnormality". From the proportion of times that the reader picks modality-B, one can determine the corresponding d-index, which is the separation of the two abnormal distributions in Fig. 1. This involves an additional assumption – that the decision variable for the modality "positive for abnormality" task is related to that for the original detection task and that modality bias (preference for a particular modality that is independent of the detection task DV) has been corrected for. Similarly, showing pairs of normal cases from the two modalities, one obtains the separation of the two normal distributions in Fig. 1. Finally one takes the differences of the two measured separations. If the two DV axes are collinear, as shown in Fig. 1, this procedure should yield the same result as the 2-ROC study method just described above. However, the second method fails if the two DV-axes are not colinear. Since the DV is not directly measurable it is dangerous to assume that the two DV-axes are collinear.

Reasons why the method should work
The method had the appearance of objectivity. The DROC method involved measuring the shift (of modality-B relative to modality-A) of the distributions for positive and negative cases separately and taking their difference. A uniform shift to both distributions would produce no net effect. The author could show that if a random (non-diagnostic) observer always picked modality-B then such an observer would yield zero net incremental separation – in other words, the DROC method was resistant to machinations by the non-diagnostic-reader.

Assume that in the ROC study context the rating standard deviation is about 0.5 units (on a 5-point scale). An image that is rated 4 might on successive and independent viewings (i.e., memory effects are minimized) be rated 3 or 5. Assume that the case is actually abnormal and that the second modality raises the average rating for this case by 0.5. The within-reader variability will tend to mask the modality-effect. However, if the images from the two modalities are shown side-by-side, the observer may be more consistent at detecting the difference between the modalities, i.e., the observer may consistently give modality-B a higher rating than modality-A, when they are shown simultaneously. This argument was persuasive on a number of people that the author tried it on. The weakness to this argument is that within-reader variability is a relatively small component in ROC studies, and inter-observer variability is much larger. In other words, the DROC method is attempting to minimize a quantity that is not limiting the measurement anyway.

The high point
Initial results of ROC and DROC experiments conducted on the same cases and readers were quite encouraging as the DROC method predicted the correct direction of the inter-modality AUC change (as measured by independent ROC experiments), but at a much higher significance level (smaller p-value). Based on these ideas and data the author made a proposal to the National Institutes of Health (NIH) titled "Measurement of Differential Image Quality". The reviewers liked the idea (the project was rated at the 6th percentile level on the second round) and the project was funded. The author presented the idea at three meetings **[9], [10], [11]**. The first presentation at the 1997 Far West Image Perception Meeting created excitement and debate (he remembers a particular exchange with Art Burgess that the author is not proud of). At the 1999-SPIE meeting a senior researcher came up to the author after his talk and told him that this was a "fantastic" idea. [A student at the same meeting was not quite as impressed. His objection was that a modality preference could not be construed as proof that the preferred modality was actually providing more diagnostic information, see below.] Based on his new-found fame, the author was asked to contribute a chapter on methodologies for an SPIE special issue **[12]**.

Comparison to traditional 2AFC
More insight into the nature of the problem was gained by appealing to the correspondence between the ROC and two-alternative-forced-choice (2AFC) methods. The DROC method seemed to be giving information about the difference in AUC between two modalities with far fewer comparisons. Consider two 2AFC studies, one with modality-A, and the other with modality-B, performed with 10-normal patients labeled N1 through N10, and 10 signal-containing (i.e., abnormal) patients labeled S1 through S10. All comparisons of the type [Ni, Sj] are permitted in the 2AFC study, where i and j are both integers ranging from 1 through 10. If one repeats this study with the second modality, then the entire set of comparisons can be denoted by [Ni, Sj]-A and [Ni, Sj]-B. In other words there are 100 independent comparisons involved for each modality. In the DROC mode one has comparisons of the type [Ni-A, Ni-B] and [Si-A, Si-B], since one is comparing only A and B-modality images of the same case. This yields a total of only 20 DROC comparisons, much less than the 200 that are possible in the 2AFC study. [This argument is still valid even if one performs an ROC experiment, which does not involve directly performing normal- abnormal case comparisons. For, given the ROC ratings one can *infer* the result of the 2AFC comparisons for all the case pairings. For example, if N1 received a rating of 5 (highly abnormal) and S9 received a rating of 2 (possibly normal), then for the 2AFC study (which was not actually performed) one would predict that for the pairing [N1, S9] the observer would pick N1 as the abnormal case (incidentally, an incorrect choice). The point is that the trapezoidal area under the ROC curve and the Wilcoxon statistic calculated from the 2AFC study (which was not actually performed) would be identical.] The DROC method has insufficient data – only 20 independent numbers in this example compared to the ROC method, which yields 200 numbers.

Is the selected modality actually providing more information?
Suppose that one has a lesion with 6 features, and at least 4 have to be detected for the observer to recognize and detect the lesion. Assume that modality-A allows the observer to visualize 3 features and modality-B allows the observer to visualize 3 different features. Viewed individually neither modality would have detected the lesion, but viewed as a pair the observer would detect the lesion with high confidence, and pick A or B as superior, according to his preference for the features. In either case a spurious difference between the modalities would be revealed by the DROC method.

3. RESULTS

The low point
The first hint of trouble came when the author found that the DROC results for different readers were highly variable and the p-values for differences between modalities were overly optimistic **[13]**. The corresponding ROC analyses yielded more consistent results and less optimistic p-values. This led the author to think about the idea in more depth and develop a simulation model for it. It soon became apparent that while the DROC method was resistant to machinations by the *non-diagnostic* reader, a *typical* reader would be able to affect the results by bias for a modality. For example, if a reader had an AUC = 0.9, then if this reader always picked modality-B, the analysis would show that modality-B was better (the quantity termed the area under the DROC curve would also equal 0.9). While one can argue that real observers do not try to 'beat the study', this presented a basic problem to the method. An evaluation method should be able to deal with this kind of challenge, especially when one is trying to improve on the ROC method, which is resistant

to problems like this. Needless to say, this was a low point for this investigator. A manuscript titled "A Paired Image Method for Measuring Small Differences in Image Quality", which had been accepted by Medical Physics, was withdrawn.

Bias correction

One way to solve the bias issue was to exploit the correlation introduced between the two responses when the observer was not giving an independent response in the second task. In other words, if the observer always picked modality-B, *that* was a predictable event, and in the way the analysis had been formulated it would introduce a spurious correlation between the ROC-performance measured using the first response, and the DROC performance measured using the second response. One can use CORROC **[14]** to determine this correlation. The method of bias removal that suggested itself was to insert a compensatory bias into the analysis program, and adjust its magnitude until cancellation occurred, i.e., when the correlation was minimized. At this point the observed DROC-AUC would be bias free **[12]**. However, the author found that the bias estimation error resulted in no net improvement over the ROC method (by this point the author had developed great respect for the standard ROC method).

4. DISCUSSION

For all the reasons stated the DROC method is on shaky scientific grounds and the author no longer recommends it. So what are the lessons that can be drawn from this study?

Getting funding is seen as a measure of success in this business. However, before one can sell an idea to a funding agency, one has to believe in it. Ideas can be seductive and there is a real danger of believing ones own "propaganda". Should one be more suspicious of a "simple" idea that appears to have been overlooked by the experts? If the author had exercised more caution he would not have sought funding for the project and the subsequent work that it led to may not have happened. The simulation work that began in this project actually led to more new ideas **[15]** and methodologies **[16], [17]**.

Ideas are not good or bad - they represent a dynamic process, an evolution of thinking. Almost two decades ago the author developed AFROC-scoring of FROC data and a reviewer of a prestigious journal used terms like "seminal work" to describe it **[18]**. In a companion paper **[19]** in these proceedings the author is now recommending *against* further usage of AFROC methodology. Two decades from now the author hopes that JAFROC will get supplanted by an even better method. That is the essence of progress in science.

Unfortunately there is no "Journal of Negative Results". Perhaps there needs to be, since it is when ideas do not work that one gets the most insight into how they are supposed to work. Sometimes, even the best laid plans go wrong. A funded project can turn out to have fundamental problems. So what is the investigator to do? Once the mistake is recognized, one cannot go forward with the original plan. Instead, one must develop alternate approaches that address the same overall aim. One needs to confront the fact of life that since the original plans have been severely disrupted, it will be difficult to appear productive, at least in the short term. One needs to follow other ideas, ones that were believed in but did not get funded, and grasp the opportunity to work on these ideas.

Since the overall aim was developing methodology for more efficient observer studies, and since the author had a prior interest in the free-response ROC (FROC) method, which he had not been able to get funded, he made the decision to switch the focus to solving the FROC problem.

In hindsight, the author made two mistakes that he is aware of. One mistake was to submit the new FROC ideas as a "competitive renewal" to the original DROC idea. This resulted in a classic case of the 'bad' tainting the 'good'. The reviewers were understandably concerned about the lack of productivity on the prior grant, and the appearance that the Principal Investigator had "lost interest in the DROC method". This brings one to the point of this paragraph, and the author owes this gem to someone else. If a grant-idea does not work, simply *close out* the project – do not submit a competitive renewal. In the final report state that "we tried the idea and for these reasons it did not work". Submit the new idea as a fresh RO1 untainted by its association with the older grant. The other mistake was to prematurely recommend the DROC method to another investigator in 1997.

5. ACKNOWLEDGMENTS

This work was supported by grants from the Department of Health and Human Services, National Institutes of Health, National Cancer Institute, RO-1 CA75145 and 8 RO1-EB002120. The author is grateful to a grant reviewer for comments that led the author to summon enough courage to write the final chapter to the DROC story, and to his wife, Beatrice, for sustaining him through a difficult period.

6. REFERENCES

1. Krupinski, E.A. and H.L. Kundel, *Update on Long-Term Goals for Medical Image Perception Research.* Acad Radiol, 1998. **5.**: p. 629-633.

2. Good, W.F., D. Gur, J.H. Feist, F.L. Thaete, C.R. Fuhrman, C.A. Britton, and B.S. Slasky, *Subjective and Objective Assessment of Image Quality--a Comparison.* J Digit Imaging, 1994. **7**(2): p. 77-8.

3. Good, W., J. Sumkin, N. Dash, C. Johns, M. Zuley, H. Rockette, and D. Gur, *Observer Sensitivity to Small Differences: A Multipoint Rank-Order Experiment.* Am. J. Roentgenol., 1999. **173**(2): p. 275-278.

4. Metz, C.E., *Roc Methodology in Radiologic Imaging.* Investigative Radiology, 1986. **21**(9): p. 720-733.

5. Metz, C.E., *Some Practical Issues of Experimental Design and Data Analysis in Radiological Roc Studies.* Investigative Radiology, 1989. **24**: p. 234-245.

6. Metz, C.E., *Basic Principles of Roc Analysis.* Seminars in Nuclear Medicine, 1978. **VIII**(4): p. 283-298.

7. Dorfman, D. and E.J. Alf, *Maximum Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating Method Data.* Journal of Mathematical Psychology, 1969. **6**: p. 487-496.

8. Burgess, A.E., *Comparison of Receiver Operating Characteristic and Forced Choice Observer Performance Measurement Methods.* Medical Physics, 1995. **22**(5): p. 643-655.

9. Chakraborty, D.P. *The Differential Receiver Operating Characteritics (Droc) Method*. in *Far West Image Perception Conference*. 1997. Tucson, AZ.

10. Chakraborty, D.P. *Bias Correction in the Droc Experiment*. in *Far West Image Perception Conference*. 1999. Calgary.

11. Chakraborty, D.P., H.L. Kundel, C.F. Nodine, T.K. Narayan, and V. Devaraju, *The Differential Receiver Operating Characteristic (Droc) Method.* SPIE Proc, Medical Imaging, 1998. **3338**: p. 234-240.

12. Chakraborty, D.P., *The Froc, Afroc, and Droc Variants of the Roc Analysis*, in *Handbook of Medical Imaging*. 2000, SPIE: Bellingham, Washington. p. 771-796.

13. Chakraborty, D.P., N.S. Howard, and H.L. Kundel, *The Differential Receiver Operating Characteristic (Droc) Method: Rationale and Results of Recent Experiments.* SPIE, 1999. **3663**: p. 82-90.

14. Metz, C.E., P.-L. Wang, and H.B. Kronman, eds. *A New Approach for Testing the Significance of Differences between Roc Curves Measured from Correlated Data*. Information Processing in Medical Imaging, ed. F. Deconinck. 1984, Nijhoff: The Hague. 432-445.

15. Chakraborty, D.P., *Proposed Solution to the Froc Problem and an Invitation to Collaborate.* Proc. SPIE, Medical Imaging 2003: Image Perception, Observer Performance and Technology Assessment, 2003. **5034**: p. 204-212.

16. Chakraborty, D.P. and K.S. Berbaum, *Methodologies for Observer Studies Involving Detection and Localization: Modeling, Analysis and Validation.* Submitted to Medical Physics, 2004(#04-041.).

17. Chakraborty, D.P. and E.A. Krupinski, *Statistical Methods in Medical Imaging and Bioengineering with Applications to Observer Performance Evaluation.* 2004, SPIE: San Diego.

18. Chakraborty, D.P. and L. Winter, *Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment.* Radiology, 1990. **174**: p. 873-881.

19. Chakraborty, D.P. and K.S. Berbaum, *Jackknife Free-Response Roc Methodology.* SPIE Proc, Medical Imaging, 2004. **5372**(This issue).

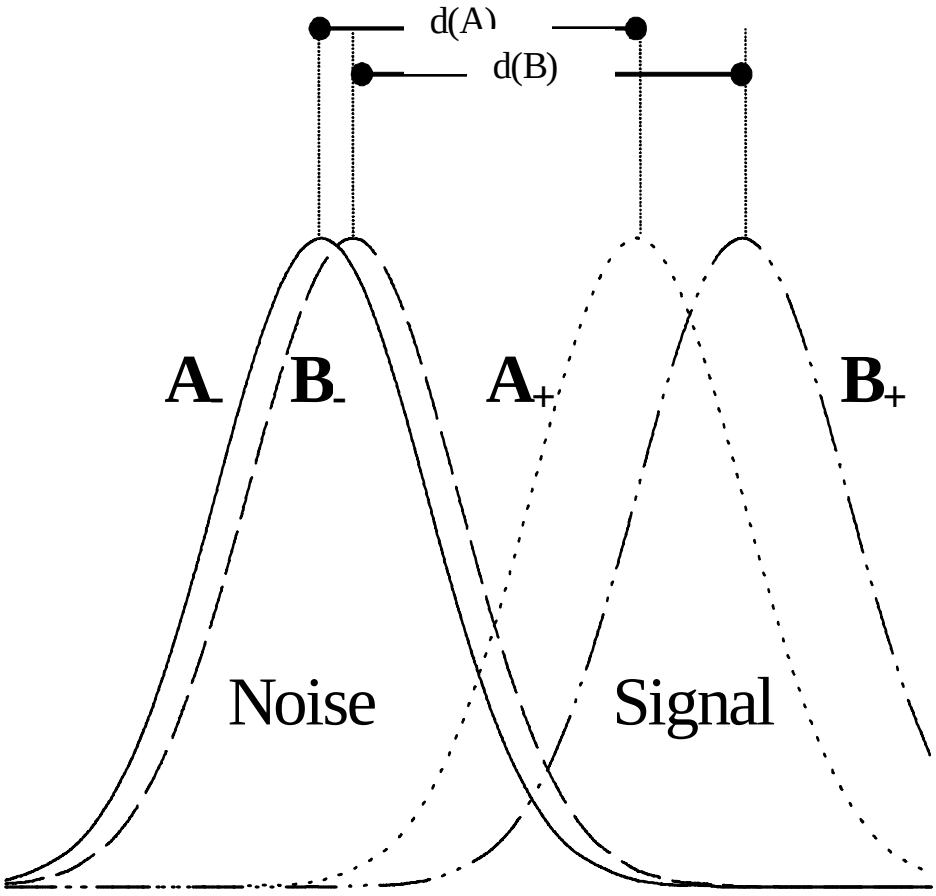

Figure 1: The geometrical picture, misleading in hindsight, which led to the DROC project. Corresponding to each modality one has a pair of distributions (signal and noise) and the two DV axes coincide. The modalities are A and B and the subscripts – and + refer to normal (noise) and abnormal (signal) cases, respectively. The standard ROC method of determining the inter-modality difference consists of sampling from the – and + distributions for a modality, calculating the detection index, repeating the process for the second modality. The final quantity of interest is Δd (ROC) = d(B) – d(A). In the DROC method one samples from the two + distributions corresponding to modalities A and B, calculates $\Delta d(+)$ and similarly for the two – distributions and finally takes the difference: $\Delta d(DROC) = \Delta d(+) - \Delta d(-)$. The two methods yield identical results, $\Delta d(DROC) = \Delta d(ROC)$, *only if the two DV axes, corresponding to A and B, are collinear*. This assumption is incorrect.