Search, lesion-classification and area under ROC curve – results from 236 fits using 3 proper ROC methods and implications for optimizing observer performance

ABSTRACT

This document presents an overview of the most recent results of three proper ROC-curve fitting methods. It is the subject of Chapter 18 of an upcoming book. It corrects erroneous negative statements regarding PROPROC published in an earlier document. The area under the ROC curve measures the ability of the observer to discriminate diseased from non-diseased cases. Perfect discrimination yields ROC-AUC = 1 while guessing performance yields ROC-AUC = 0.5. Search performance is the ability of the observer to find true lesions while avoiding false lesions. It is measured by a quantity denoted S. Whether a found lesion is marked depends on the reporting threshold. Having found a suspicious lesion, lesion classification performance is the ability to correctly classify it is malignant or benign. It ranges from 0.5 (chance level ability) to 1 (perfect). The primary conclusion of this document is that search performance is the "bottleneck", averaging about 17%, while lesion-classification performance is about 89%. Unless researchers recognize this fact and continue to focus on ROC performance, their efforts will not be productive. A fundamental rule of science is to first focus on the weak-link. There exists a strong inverse correlation between lesion-classification and search performance. Observers tend to compensate for deficiencies in search performance by greater performance in lesion-classification, and vice-versa.

OVERVIEW

This document presents an overview of the most recent results of three proper[1] ROC-curve fitting methods. It is the subject of Chapter 18 of an upcoming book[1].

It corrects erroneous negative statements regarding PROPROC published in an earlier document[2].

The three methods were applied to 14 datasets, indexed by *d* (*d* = 1, 2, ..., 14) With **I** modalities, **J** readers, each dataset yielded **3IJ** ROC plots. Table 1 summarizes the datasets and acknowledges the sources.

[1] As the operating point moves up the curve, a proper ROC curve has monotonically decreasing slope; it will not cross the chance diagonal nor will it show a "hook" near the top-right corner, as usually observed with the binormal model.

[2] https://www.researchgate.net/publication/317087463_Quantifying_search_lesion_classification_and_case_classification_performances; the erroneous conclusions were due to programming errors.

Table 1: This table lists the summary characteristics of the datasets used in this book. The dataset type is ROC, FROC or LROC. The total number of individual modality-reader combinations is 236, i.e., $\sum_{d=1}^{14} I_d J_d$. [I = # modalities, J = # readers, K_1 = number of non-diseased cases, K_2 = number of diseased cases; K = total # cases.]

Dataset # d	Dataset Name	Data Type	I	J	K_1	K_2	K	Description
1	TONY[2]	FROC	2	5	96	89	185	Digital breast tomosynthesis vs. mammography
2	VD[3]	ROC	2	5	69	45	114	Cine vs. SE MRI for aortic dissection
3	FR[4]	ROC	2	4	33	67	100	Digital vs. analog pediatric chest
4	FED[5]	FROC	5	4	100	100	200	Image processing in mammography: FROC
5	JT[6]	FROC	2	9	45	47	92	Nodule detection in an thorax CT phantom
6	MAG[7]	FROC	2	4	47	42	89	Tomosynthesis Vs. Radiography Pulmonary Nodules
7	OPT[8]	FROC	5	7	81	81	162	Calcification detection in digital mammography
8	PEN[9]	FROC	5	5	48	64	112	Image compression in mammography
9	NICO[10]	LROC	1	10	120	80	200	Standalone CAD (modality 1) vs. 9 radiologists
10	RUS[11]	FROC	3	8	50	40	90	Lesion detection in digital mammography
11	DOB1[12]	FROC	4	5	43	115	158	Tomosynthesis, Dual-Energy & Conventional Chest
12	DOB2[12]	ROC	4	5	64	88	152	do:
13	DOB3[12]	FROC	4	5	52	106	158	do:
14	FZR[13]	ROC	2	4	100	100	200	Image processing in mammography: ROC

THREE BASIC CONCEPTS

1. The **area under the ROC** curve measures the ability of the observer to **discriminate diseased from non-diseased cases**. Perfect discrimination yields ROC-AUC = 1 while guessing performance yields ROC-AUC = 0.5.
2. **Search performance** is the ability of the observer to find true lesions while avoiding false lesions. It is measured by a quantity denoted **S**. Whether a found lesion is marked depends on the reporting threshold.
3. Having found a suspicious lesion, **lesion classification** performance is the ability to **correctly classify it is malignant or benign**. It is denoted by A_C, ranging from 0.5 (chance level ability) to 1 (perfect).

The primary conclusion of this document is that search performance is the "bottleneck", averaging about 17%, while lesion-classification performance is about 89%. Unless researchers recognize this fact, and continue to focus on ROC performance their efforts will not be productive. A fundamental rule of science is to first focus on the weak-link. The radiological search model is the only existing method of estimating search performance. It is implemented in RJafroc Version 1.0.0 (currently undergoing QC checks prior to upload to CRAN).

A FEW DETAILS

- Search performance is low (0.17) compared to lesion-classification performance (0.89). **Search performance is by far the weak link in observer performance**. It is not measurable under any conventional ROC model, i.e., those that generate decision variable samples on every image. It is measurable using RSM fitting model.
- There exists a **strong inverse correlation** between **C** and **S** and weaker positive correlations between **A** vs. **S** and **A** vs. **C**. Observers tend to compensate for deficiencies in search by greater performance in lesion-classification, and vice-versa.
- **All three methods yielded almost identical AUCs. On the average, PROPROC AUC was 1% larger than RSM AUC, while CBM AUC was 1% smaller than RSM AUC**.[3]
- These findings are consistent with each proper-ROC method being a realization of an ideal observer[14,15].
- RSM-μ and CBM-μ parameters are correlated, consistent with their physical meanings.
- RSM-ν' and CBM-α parameters are correlated, consistent with their physical meanings.
- For degenerate datasets[4] PROPROC yields gross overestimates of performance [16]. This problem is readily fixed but unfortunately, this important software due to Metz and colleagues is no longer being supported.

Due to its size, the detailed analysis, including plots, will be published separately.

[3] This is the correction to the main conclusion of the earlier document mentioned above.

[4] Defined as not having any interior points.

REFERENCES

1. Chakraborty DP. *OBSERVER PERFORMANCE METHODS FOR DIAGNOSTIC IMAGING - Foundations, Modeling, and Applications with R-Based Examples.* Taylor-Francis LLC; 2017 (under production, expected to be available by Dec 15, 2017).
2. Svahn T, Andersson I, Chakraborty D, et al. The Diagnostic Accuracy of Dual-View Digital Mammography, Single-View Breast Tomosynthesis and a Dual-View Combination of Breast Tomosynthesis and Digital Mammography in a Free-response Observer Performance Study. *Radiat Prot Dosimetry.* 2010;139:113–117.
3. Van Dyke CW, White RD, Obuchowski NA, Geisinger MA, Lorig RJ, Meziane MA. Cine MRI in the diagnosis of thoracic aortic dissection. *79th RSNA Meetings.* 1993.
4. Franken EA, Jr., Berbaum KS, Marley SM, et al. Evaluation of a Digital Workstation for Interpreting Neonatal Examinations: A Receiver Operating Characteristic Study. *Investigative Radiology.* 1992;27(9):732-737.
5. Zanca F, Jacobs J, Van Ongeval C, et al. Evaluation of clinical image processing algorithms used in digital mammography. *Medical Physics.* 2009;36(3):765-775.
6. Thompson JD, Chakraborty DP, Szczepura K, et al. Effect of reconstruction methods and x-ray tube current-time product on nodule detection in an anthropomorphic thorax phantom: a crossed-modality JAFROC observer study. *Medical Physics.* 2016;43(3):1265-1274.
7. Vikgren J, Zachrisson S, Svalkvist A, et al. Comparison of Chest Tomosynthesis and Chest Radiography for Detection of Pulmonary Nodules: Human Observer Study of Clinical Cases. *Radiology.* 2008;249(3):1034-1041.
8. Warren LM, Mackenzie A, Cooke J, et al. Effect of image quality on calcification detection in digital mammography. *Medical Physics.* 2012;39(6):3202-3213.
9. Penedo M, Souto M, Tahoces PG, et al. Free-Response Receiver Operating Characteristic Evaluation of Lossy JPEG2000 and Object-based Set Partitioning in Hierarchical Trees Compression of Digitized Mammograms. *Radiology.* 2005;237(2):450-457.

10. Hupse R, Samulski M, Lobbes M, et al. Standalone computer-aided detection compared to radiologists’ performance for the detection of mammographic masses. *Eur Radiol.* 2013;23(1):93-100.

11. Ruschin M, Timberg P, Bath M, et al. Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies. *Med Phys.* 2007;34:400 - 407.

12. Dobbins III JT, McAdams HP, Sabol JM, et al. Multi-Institutional Evaluation of Digital Tomosynthesis, Dual-Energy Radiography, and Conventional Chest Radiography for the Detection and Management of Pulmonary Nodules. *Radiology.* 2016;282(1):236-250.

13. Zanca F, Hillis SL, Claus F, et al. Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: Results from independently conducted FROC/ROC studies in mammography. *Med Phys.* 2012;39(10):5917-5929.

14. Macmillan NA, Creelman CD. *Detection Theory: A User's Guide.* New York: Cambridge University Press; 1991.

15. Barrett HH, Myers K. *Foundations of Image Science.* Hoboken, N.J.: John Wiley and Sons; 2003.

16. Zhai X, Chakraborty DP. A bivariate contaminated binormal model for robust fitting of proper ROC curves to a pair of correlated, possibly degenerate, ROC datasets. *Med Phys.* 2017;44(3):in press.