Estimating the parameters of a model of visual search from ROC data: an alternate method for fitting proper ROC curves

D. P. Chakraborty*[a] and Tony Svahn[b]
[a]Department of Radiology, Univ. of Pittsburgh, 200 Lothrop Street, Pittsburgh, PA 15213
[b]Medical Radiation Physics, Department of Clinical Sciences, Lund University,
Malmö University Hospital, SE 205 02, Malmö, Sweden

ABSTRACT

The binormal receiver operating characteristic (ROC) model often predicts an unphysical "hook" near the upper-right corner (1,1) of the ROC plot. Several models for fitting proper ROC curves avoid this problem. The purpose of this work is to describe another method that involves a model of visual search that models free-response data, and to compare the search-model predicted ROC curves with those predicted by PROPROC (proper ROC) software. The highest rating rule was used to infer ROC data from FROC data. An expression for the search-model ROC likelihood function is derived, maximizing which yielded estimates of the parameters and the fitted ROC curve. The method was applied to a dual-modality 5-reader FROC data set. The relative difference between the average AUCs for the two methods was less than 1%. A linear regression of the AUCs yielded an adjusted R-squared of 0.95 indicative of strong linear correlation between the search model AUC and PROPROC AUC, although the shapes of the predicted ROC curves were qualitatively different. This study shows the feasibility of estimating parameters characterizing visual search from data acquired in a non-search paradigm.

Keywords: Observer performance, ROC, FROC, proper ROC, search model, PROPROC, maximum likelihood

1. INTRODUCTION

The RSCORE [1, 2] and ROCFIT [3] software are widely used for fitting receiver operating characteristic (ROC) ratings data. They are based on the unequal variance binormal model according to which the decision variables - or z-samples - for disease-free cases are drawn from a unit variance normal distribution with mean zero, and for disease-present cases are drawn from a normal distribution with variance σ^2 and mean μ. These parameters are related to the traditional *a* and *b* parameters of the binormal model by $a = \mu/\sigma$ and $b = 1/\sigma$. Most clinical datasets are consistent with $\sigma > 1$, in which case the fitted curve has a "hook" near the upper-right corner (1,1) [4] where it falls under the chance diagonal and hooks upward to (1,1). Fig. 1 shows the predicted ROC curve for a = 1 and b = 0.3. While less commonly observed, if $\sigma < 1$ the binormal model predicts an unphysical "hook" near the lower-left corner (0,0) of the ROC plot.

Although the hook may not be readily evident for most datasets, the fact that it exists is theoretically unsatisfactory because it implies worse than chance level performance in a region of the curve. There are at least three models for fitting proper ROC curves that avoid this problem: (i) the contaminated binormal model [5], (ii) the bi-gamma model [6], and (iii) the proper binormal ROC model [7, 8]. The proper ROC (PROPROC) model is based on the likelihood ratio corresponding to a binormal model, and decisions are made based on the likelihood ratio exceeding a threshold, rather than z exceeding a threshold. The model has two parameters d_a and *c* which are related to the *a* and *b* parameters of the conventional binormal model, and the area under the binormal ROC curve, as described in Reference [7].

Free-response data consists of mark-rating pairs classified as lesion or non-lesion localizations, according to their proximity, near or far, respectively, from real lesions. The FROC curve is defined as the plot of lesion localization fraction vs. non-lesion localization fraction. Lesion localization fraction is the probability that a lesion is correctly localized to within the proximity criterion, and non-lesion localization fraction is the average number of non-lesion localizations per image. Recently a search-model for free-response receiver operating characteristic (FROC) data was introduced by the first author [9, 10]. Unlike the 2-parameter binormal model, the search model is characterized by three parameters (not counting the cutoffs). A method has been described for estimating the parameters of the search-model by maximizing the likelihood function of the observed FROC data [11].

If one adopts a rule for reducing the multiple mark-rating pairs on an image to a single ROC-like rating, then the search model also predicts ROC data. A common rule is the highest rating assumption, according to which the inferred ROC rating is the rating of the highest rated mark on the image, or negative infinity if the image has no marks. It has been shown that the search-model predicted ROC curve is "proper". This paper describes fitting proper ROC curves using the search-model and compares them to PROPROC fits.

2. METHODS

2.1 The relation of a perceptual model to FROC data

According to a perceptual model [12, 13] image interpretation involves a brief (~300 ms per image) global-response stage that identifies potential lesion candidates and (ii) a decision-making stage (~1 s per candidate) at which decisions are made whether to report the lesion candidates. In search-model terminology, the lesion candidates are referred to as *decision sites*. Sites that are in fact normal anatomic variants, but which the observer may mistake for cancer, are termed *noise sites* and sites that represent true lesions are termed *signal sites*. It is assumed that based on image information and the nature of the task the reader calculates a scalar z-sample at each site, which represents the confidence level that the site is a true lesion. The site is marked if the z-sample exceeds the observer's lowest reporting threshold. The rating assigned to a mark is the value of the z-sample, or the bin index 1, 2, ….R, if a discrete R-rating reporting format is used. It is assumed that higher ratings correspond to greater confidence level. A marked noise site is recorded as non-lesion localization and a marked signal site as lesion localization.

2.2 Notation and search model parameterization

Modalities are indexed by i (i = 1, 2, ..., I, for I modalities), readers by j (j = 1, 2, ..., J, for J readers) and cases by k_t, where N_N disease-free cases are indexed by k_1 = 1, 2, ..., N_N; and N_A disease-present cases are indexed by k_2 = 1, 2, ..., N_A. The case-truth index t refers to the patient as a whole: disease-free cases are denoted by t = 1 and disease-present cases by t = 2. The site-truth index s distinguishes noise sites from signal sites: s = 1 represents a noise site and s = 2 represents a signal site. On disease-free cases (t = 1) only noise-sites (s = 1) are possible, but on disease-present cases (t = 2) both noise and signal sites are possible (s = 1 or 2). The shorthand ijk_t denotes the combination of modality i, reader j and case k_t. Let n_{ijk_ts} denote the number of sites of type s on ijk_t. The number of noise sites n_{ijk_t1} is modeled [10] as an integer random variable sampled from a Poisson distribution with mean λ_{ijt}, i.e., $n_{ijk_t1} \sim Poi(\lambda_{ijt})$. The number of signal-sites n_{ijk_t2} is modeled as an integer random variable sampled from a binomial distribution with trial size L_{k_2} and probability of success ν_{ij}, i.e., $n_{ijk_t2} \sim binomial(L_{k_2}, \nu_{ij})$, where L_{k_2} is the number of lesions in disease-present case k_2. The maximum number of lesions per case is $L^{MAX} \equiv \max_{k_2}\left(L_{k_2}\right)$. The sites are indexed by l_s where l_1 indexes noise-sites with possible values $l_1 = 1, 2, ..., n_{ijk_t1}$, provided the image has at least one noise-site, and l_2 indexes signal sites with possible values $l_2 = 1, 2, ..., n_{ijk_t2}$, provided the image has at least one signal-site. The z-samples are modeled by $z_{ijk_tl_s} \sim N(\mu_{ijs}, 1)$ where $\mu_{ij1} = 0$ and $\mu_{ij2} \equiv \mu_{ij}$. In other words, noise sites are sampled from a unit variance normal distribution with mean zero and signal sites from a unit variance normal distribution with mean μ_{ij}.

2.3 **Search model predicted ROC curves**

Using the highest rating assumption, ROC data can be inferred from FROC data. Generalizing the results derived in Ref. [9] for constant number of lesions per disease-present case, the ROC curve predicted by the search model for variable numbers of lesions per disease-present case is given by:

$$\left.\begin{aligned} FPF_{ij}(\zeta) &= 1-\exp\left(-\frac{\lambda_{ij1}}{2}+\frac{\lambda_{ij1}}{2}erf\left(\frac{\zeta}{\sqrt{2}}\right)\right) \\ TPF_{ij}(\zeta) &= \sum_{L=1}^{L^{MAX}} f_L\left[1-\left(1-\frac{\nu_{ij}}{2}+\frac{\nu_{ij}}{2}erf\left(\frac{\zeta-\mu_{ij}}{\sqrt{2}}\right)\right)^L \exp\left(-\frac{\lambda_{ij2}}{2}+\frac{\lambda_{ij2}}{2}erf\left(\frac{\zeta}{\sqrt{2}}\right)\right)\right] \end{aligned}\right\}, \quad \text{Eqn. 1}$$

and

$$\sum_{L=1}^{L^{MAX}} f_L = 1. \quad \text{Eqn. 2}$$

Here ζ is the cutoff parameter defining an operating point on the ROC, L is the number of lesions in a disease-present case, f_L is the fraction of disease-present cases with L lesions, and *erf* is the error function. It is seen that $TPF_{ij}(\zeta)$ is a weighted average over all values of L = 1, 2, ..., L^{MAX}, of true positive fraction for modality i, reader j, defined for a dataset where every disease-present case has L lesions. In this work, it is assumed that $\lambda_{ij1} = \lambda_{ij2} \equiv \lambda_{ij}$. In this case it can be shown [9] that the ROC curve predicted by the above equation is "proper", i.e., it does not cross the chance diagonal and has monotonically decreasing slope, but the uppermost operating point, i.e., that reached when ζ = -∞, is not (1,1) rather it is:

$$\left.\begin{aligned} FPF_{ij,\max} &= 1-\exp\left(-\lambda_{ij}\right) \\ TPF_{ij,\max} &= \sum_{L=1}^{L^{MAX}} f_L\left[1-\left(1-\nu_{ij}\right)^L \exp\left(-\lambda_{ij}\right)\right] \end{aligned}\right\}. \quad \text{Eqn. 3}$$

Since the term multiplying the exponential in the second equation in Eqn. 3 is less than one, the uppermost operating point satisfies TPF > FPF, i.e., it lies above the chance diagonal. Differentiating TPF and FPF with respect to ζ and taking the ratio, the limiting slope of the ROC curve at the end point can be shown to be (assuming μ > 0)

$$\left.\frac{TPF'_{ij}}{FPF'_{ij}}\right|_{\zeta=-\infty} = \sum_{L=1}^{L^{MAX}} f_L(1-\nu)^L \quad . \quad \text{Eqn. 4}$$

Here the prime denotes the derivative with respect to ζ. The area under the ROC curve equals that under the continuous section plus the area under the straight-line segment connecting the uppermost point to (1,1). The area under the continuous section can be calculated by numerical integration.

2.4 **Likelihood function**

Suppressing for simplicity the modality and reader subscripts, let (F_b, T_b) denote the number of false positive and true positives, respectively, in ratings bin *b* defined by neighboring cutoffs (ζ_b, ζ_{b+1}). Here *b* = 0, 1, …, R and R is the number of FROC bins, one less than the number of ROC bins, and $\zeta_0 = -\infty$ and $\zeta_{R+1} = +\infty$. The cutoffs are collectively denoted by the cutoff-vector $\vec{\zeta}$. F_0 and T_0 represent the number of normal images with no marks and the number of disease-present cases with no marks, respectively. Ignoring terms that do not depend on search model parameters, the contribution to the likelihood function from bin *b* is

$$\mathcal{L}_b = \left[FPF(\zeta_b) - FPF(\zeta_{b+1})\right]^{F_b} \left[TPF(\zeta_b) - TPF(\zeta_{b+1})\right]^{T_b}. \quad \text{Eqn. 5}$$

The net likelihood $\mathcal{L}$ is the product of (R+1) terms like the one shown above, i.e.,

$$\mathcal{L} = \prod_{b=0}^{b=R} \mathcal{L}_b \ . \quad \text{Eqn. 6}$$

To estimate the parameters one maximizes the logarithm of the likelihood with respect to the 3+R parameters. It was found that allowing all parameters to vary independently could lead to unrealistic parameter estimates (typically λ is driven to large values and ν to unity). This is due to a near degeneracy of the likelihood function whereby the effects on L of a positive change in λ or ν can be almost cancelled by a positive change in ζ_1. The following algorithm was more successful. (i) For given λ, ν and cutoffs $\vec{\zeta}$ the μ parameter was determined by minimizing the chi-square goodness of fit statistic. (ii) The log likelihood $\mathcal{LL}$ was calculated. (iii) The parameters $\lambda, \nu, \vec{\zeta}$ were varied and the preceding steps repeated until $\mathcal{LL}$ was maximized. The maximization routine involved simulated annealing combined with a Simplex search algorithm. PROPROC analysis was also conducted using DBM-MRMC software [14].

2.5 **Application to a clinical data set**

The method was applied to human observer data from a dual-modality FROC study in which 5 readers interpreted 96 normal and 89 disease-present cases. The total number of lesions was 95 and there were at most two lesions per disease-present case (L^{MAX} = 2). A 5-point rating scale was used (R = 5). The FROC data was reduced to ROC data by assigning the rating of the highest-rated mark on an image as its ROC-equivalent rating. If the image had no marks, the default 0 rating was assigned to it (any number less than 1 would have sufficed). This resulted in ROC data on a 6-point scale. If necessary, the data was re-binned by combining neighboring bins to achieve $F_b \geq 5, T_b \geq 5$ for all bins. This was necessary because a meaningful chi-square statistic requires at least 5 counts in each bin.

3. RESULTS

Table 1 lists the estimated search-model and PROPROC parameters; the search model area (AUC) under the ROC curve and the PROPROC estimated AUC. For all readers except reader 4, both PROPROC and search model AUCs were greater for modality 1 than for modality 2. For reader 4 the reverse was true for both PROPROC and search-model AUCs. In other words, both fitting methods agreed on the ordering of the modalities, even for the "outlier" reader. Table 1 also lists for each modality the average AUCs over the five readers and the last two rows list the relative difference between the two reader-averaged AUC estimates (search-model and PROPROC) for each modality. The relative

difference was less than 1%, indicating that the two methods of calculating AUC tracked each other. Fig. 1 is a plot of SM AUC vs. PROPROC AUC. The linear regression line is also shown in Fig. 1. The adjusted R^2 was 0.9502 showing that the search-model and PROPROC estimates tracked each other.

The ROC curves using both methods are shown in Fig. 2. The left panels are search-model fitted curves and the right panels are PROPROC-fitted curves. The shapes of the curves are qualitatively different. A distinguishing feature of search-model fitted ROC curves, implied by Eqn. 3 is that they never extend continuously to (1,1). This has been termed the finite end-point property and is shared by all search-model fitted operating characteristics [15]. The alternative free-response receiver operating characteristic (AFROC) curve does not continuously extend to (1,1); the location receiver operating characteristic (LROC) curve abscissa does not extend to unity; and the FROC curve does not extend to (∞, 1). The reason is that the search-model allows for the possibility that some images have no decision sites and on them there is literally nothing to report. We illustrate this idea for the ROC curve. According to the search-model the probability that a disease-absent case has zero noise sites is $e^{-\lambda}$ and the probability that a disease-present case with L lesions has zero noise sites *and* zero signal sites is $e^{-\lambda}(1-\nu)^{L}$. For λ = 1, ν = 0.7 and L^{MAX} = 1, these translate to 0.368 and 0.110, respectively. In other words, in this example, about 37% of disease-free cases and 11% of disease-present cases have nothing to report. Suppose the observer is using a 101-point ROC scale, where zero represents certainty that there is no disease and 100 represents certainty that there is disease. If there is nothing to report, then the observer has to give the case the zero rating. On the remaining cases, there is something to report, and the observer is free to bin the cases using the full 100-point scale ranging from one to 100. When constructing operating points, cumulating the cases rated one or greater will yield the operating point FPF = 1-0.368 = 0.632 and TPF = 1-0.11 = 0.89. Note that the observer cannot operate infinitesimally close to (1,1) by lowering the cutoff. When the cutoff is lowered to its lowest level the operating point approaches (0.632, 0.89), which is well short of (1,1). At the highest achievable operating point, the observer has exhausted all cases with something to report. The point (1,1) is not reached by the observer, rather the experimenter upon cumulating all cases, including the ones rated 0, trivially reaches it. The inaccessible section of the ROC curve is depicted Fig. 3 by the dotted lines. All search-model predicted ROC curves (left panels in Fig. 3) have an inaccessible section. With appropriate choices of parameters the uppermost operating point can be practically indistinguishable from (1,1) but it can never exactly reach the upper right corner. For example, for λ = 10, ν = 1 and L^{MAX} = 1, the uppermost operating point is (0.99996, 1). The large value of λ ensures that practically every disease-free case has a noise site and ν = 1 ensures that every lesion is a signal site, i.e., every disease-present case has a decision site.

4. DISCUSSION

This work shows that the search model can be used to fit ROC curves. The PROPROC and search model fitted curves had different shapes but the areas under the ROC curves were close. The PROPROC fitted curve ends at (1,1) while the search model fitted curve can end anywhere above the chance diagonal inside the ROC plot. This flexibility allows the search model to more easily fit an empirical ROC curve whose shape is inconsistent with a curve whose slope is monotonically decreasing from the origin to (1,1). An example of such a curve would be one that rises sharply from the origin and ends relatively far from (1,1), for example reader 5 in Fig. 2. When there is only one operating point, both methods yield exact fits, but the chi-square statistic is not meaningful since it has zero degrees of freedom.

The search model parameters have physical correspondences. For example, the λ and ν parameters characterize search performance – defined as the ability to find lesions (ν) without finding non-lesion candidates (λ). Perfect search performance is characterized by ν approaching unity and λ approaching zero. In these limits, see Eqn. 3, the ROC curve would consist of a continuous section hugging the y-axis and ending at (0,1). The inaccessible section would be the horizontal line connecting (0,1) to (1,1) and AUC = 1. The μ parameter characterizes the ability of the observer to discriminate between noise sites and signal sites. Higher values of μ would correspond to increasing ability to correctly classify noise and signal sites. Since the fitted parameter values characterize visual search performance, the method in essence estimates search parameters from a non-search task. The reason this is possible is that there is information in the location of the uppermost operating point that is related to search performance. When the uppermost operating point is close to the upper left corner of the ROC plot, search performance is high (small λ and ν almost unity). When it is close to the upper right corner, λ and ν have more typical values. The shape of the curve between the origin and the upper

most point is affected by μ. If μ is zero, the curve is linear, and the steepness of the initial section of the curve increases with μ. If search performance is ideal (λ = 0 and ν = 1) then good discrimination between signal and noise sites (high μ) is not necessary for good overall performance. In this limit even if μ = 0 the ROC curve will be a vertical line from the origin to (0,1) followed by a horizontal line from (0,1) to (1,1) yielding AUC = 1. This is because the search has already screened out any noise sites, and any z-sample is necessarily from a signal site, and there is no need to further discriminate between noise and signal sites.

The fitting algorithm is implemented in JAFROC 4.0 software available on www.devchakraborty.com.

Acknowledgments: DPC was supported in part by grants from the Department of Health and Human Services, National Institutes of Health, R01-EB005243 and R01-EB008688. TS was supported by the Cancer 450 Research Foundation at the Department of Oncology, Franke and Margareta Bergqvist Foundation. DPC is grateful to Mr. Hong-Jun Yoon, MSEE, for assistance with the programming.

REFERENCES

[1] D. D. Dorfman, RSCORE II. In J. A. Swets and R. M. Pickett, Evaluation of diagnostic systems: Methods from signal detection theory, Academic Press, New York (1982).
[2] D. D. Dorfman, and E. Alf, "Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating-Method Data.," Journal of Mathematical Psychology, 6, 487-496 (1969).
[3] C. E. Metz, Current problems in ROC analysis. In W. W. Peppler and A. A. Alter, Proceedings of the Chest Imaging Conference 1987, Department of Medical Physics, University of Wisconsin-Madison., Madison, WI, (1988).
[4] C. E. Metz, "Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC studies," Investigative Radiology, 24, 234-245 (1989).
[5] D. D. Dorfman, and K. S. Berbaum, "A contaminated binormal model for ROC data: Part II. A formal model," Acad Radiol., 7(6), 427-37 (2000).
[6] D. D. Dorfman, K. S. Berbaum, C. E. Metz *et al.*, "Proper Receiving Operating Characteristic Analysis: The Bigamma model," Acad. Radiol., 4(2), 138-149 (1997).
[7] C. E. Metz, and X. Pan, "Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation," J Math Psychol, 43(1), 1-33 (1999).
[8] L. L. Pesce, and C. E. Metz, "Reliable and Computationally Efficient Maximum-Likelihood Estimation of Proper Binormal ROC Curves," Acad Radiol, 14(7), 814-829 (2007).
[9] D. P. Chakraborty, "ROC Curves predicted by a model of visual search," Phys. Med. Biol., 51, 3463–3482 (2006).
[10] D. P. Chakraborty, "A search model and figure of merit for observer data acquired according to the free-response paradigm," Phys. Med. Biol., 51, 3449–3462 (2006).
[11] H. J. Yoon, B. Zheng, B. Sahiner *et al.*, "Evaluating computer-aided detection algorithms," Medical Physics, 34(6), 2024-2038 (2007).
[12] H. L. Kundel, and C. F. Nodine, "A visual concept shapes image perception," Radiology, 146, 363-368 (1983).
[13] H. L. Kundel, C. F. Nodine, E. F. Conant *et al.*, "Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study," Radiology, 242(2), 396-402 (2007).
[14] K. S. Berbaum, C. E. Metz, L. L. Pesce *et al.*, [DBM MRMC User's Guide, DBM-MRMC 2.1 Beta Version 2, available from www-radiology.uchicago.edu/cgi-bin/roc_software.cgi and from perception.radiology.uiowa.edu, last accessed Dec 28, 2009].
[15] D. P. Chakraborty, and H. J. Yoon, "Operating characteristics predicted by models for diagnostic tasks involving lesion localization," Med. Phys., 35(2), 435-445 (2008).

Table 1: This table lists the estimated search-model and PROPROC parameters, the search model area AUC and the PROPROC estimated AUC. For all readers both fitting methods agreed on the ordering of the modalities, even for the "outlier" reader #4 who ordered the modalities differently from the rest. The last two rows list the relative difference (Rel. Diff.) between the two reader-averaged AUC estimates (search-model and PROPROC) for each modality, which was less than 1%, indicating that the two methods of estimating AUC tracked each other.

Modality i	Reader j	SM			PROPROC		AUC_{SM}	$AUC_{PROPROC}$
		μ_{ij}	λ_{ij}	ν_{ij}	c_{ij}	$(d_a)_{ij}$		
1	1	2.077	1.057	0.705	-0.132	1.197	0.819	0.801
	2	1.681	0.682	0.866	-0.087	1.771	0.882	0.895
	3	2.799	18.63	0.938	-0.144	1.482	0.849	0.853
	4	0.900	0.797	1.000	0.080	1.514	0.856	0.858
	5	0.961	0.245	0.790	0.223	1.740	0.858	0.891
2	1	1.183	1.051	0.486	-0.082	0.628	0.680	0.672
	2	0.860	0.978	0.747	0.050	0.974	0.750	0.754
	3	2.297	4.589	0.751	-0.133	1.156	0.794	0.793
	4	0.516	0.304	0.875	0.118	1.620	0.862	0.874
	5	0.562	0.776	0.673	0.078	0.893	0.720	0.736
1	AVG	1.684	4.282	0.860	-0.012	1.541	0.853	0.860
2		1.084	1.540	0.706	0.0063	1.054	0.761	0.766
1	Rel. Diff. $\frac{AUC_{PROPROC} - AUC_{SM}}{AUC_{PROPROC}}$							0.0077
2								0.0063

5. FIGURES

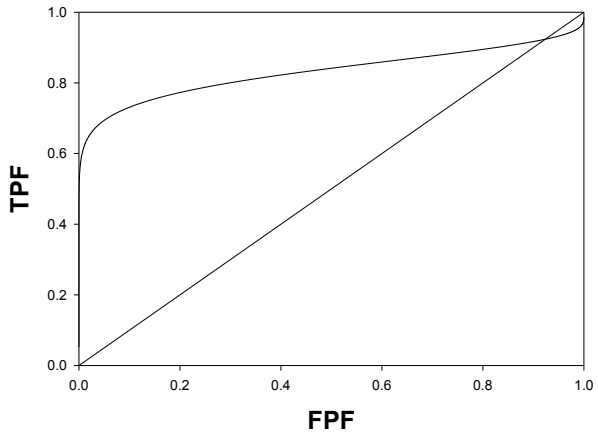


Fig. 1: An exaggerated example of an improper ROC curve, generated by the binormal model with model parameters a = 1 and b = 0.3. The "hook" near (1,1) where the curve falls below the chance diagonal, is readily apparent.

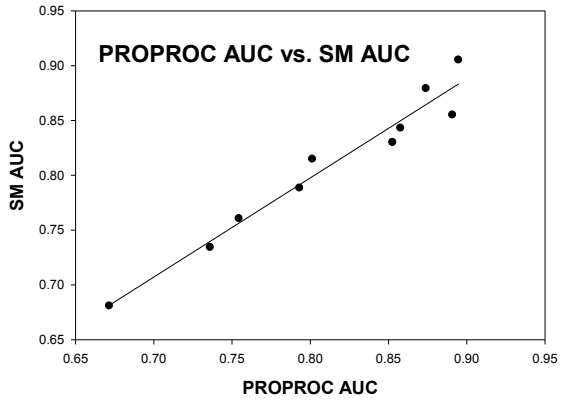
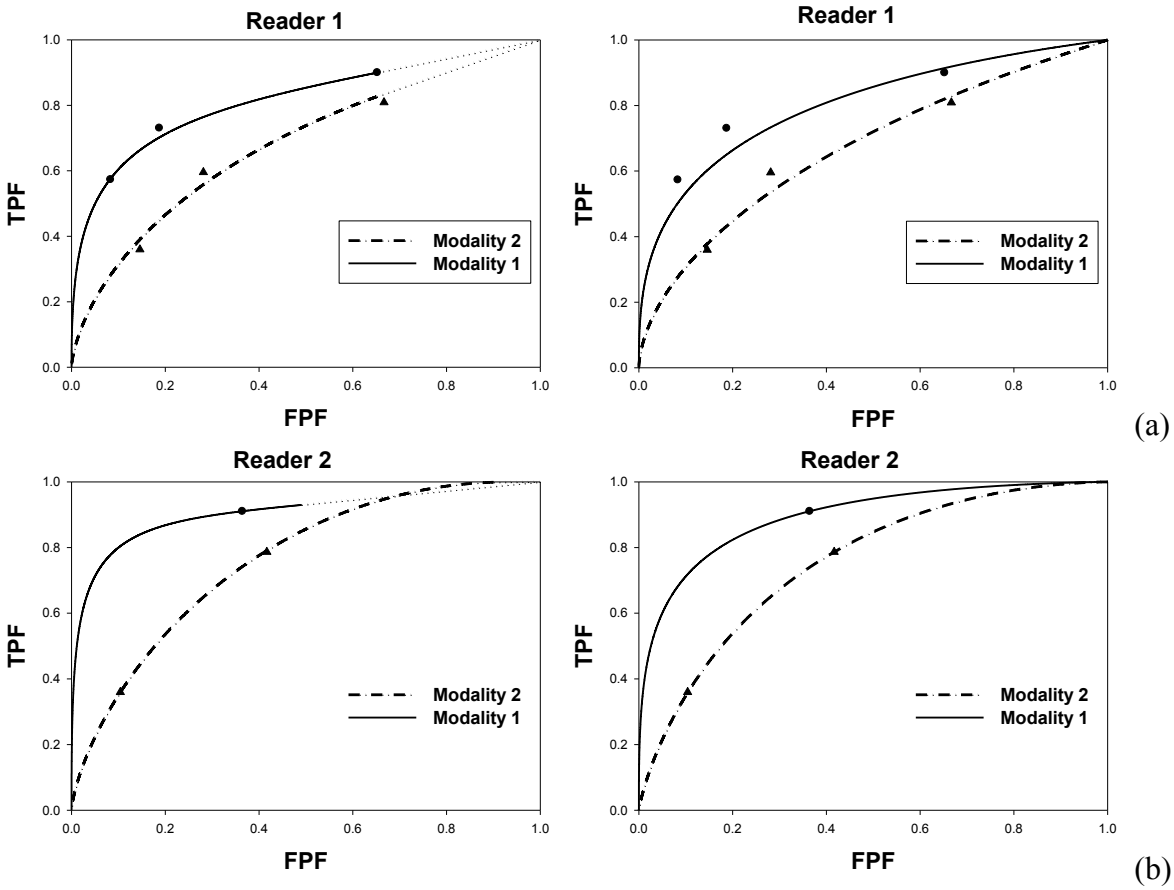


Fig. 2: Plot of search model fitted area AUC under the ROC curves vs. PROPROC fitted AUC for the five readers and two modalities (10 data points). The straight line is a linear regression fit, y = 0.9047 * x + 0.0738, adjusted R^2 = 0.9502. The search model AUCs tracked the PROPROC AUCs.



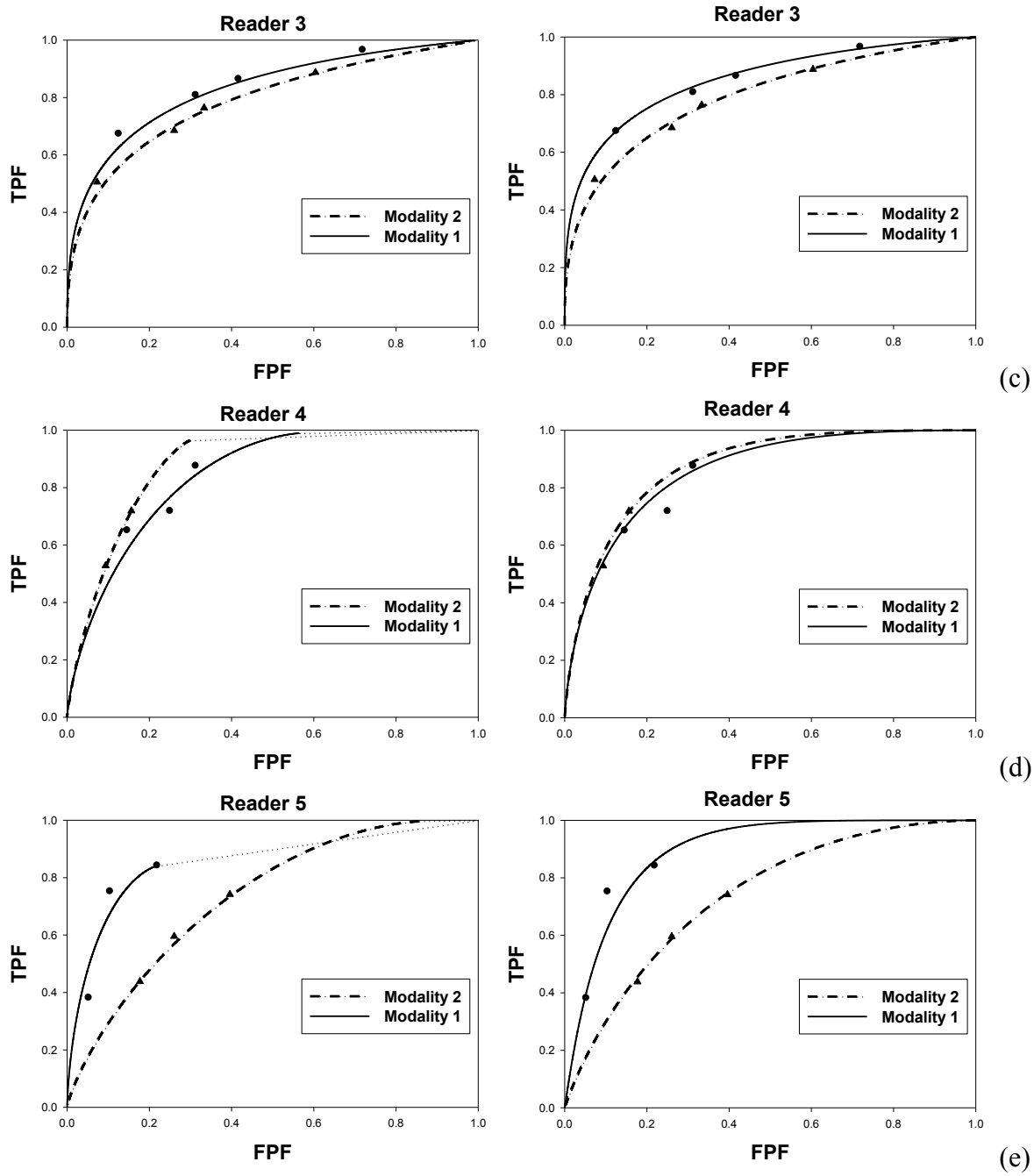


Fig. 3: (a) – (e) ROC curves for readers 1 thru 5 for the two modalities. The left panels are search model fits and the right panels are PROPROC fits. Although the shapes of the ROC curves are qualitatively different, the AUC's are, with one exception, in good agreement (Table 1). The exception is reader 5, modality 1 where $AUC_{SM} = 0.891$ and $AUC_{PROPROC} = 0.858$.