Chapter 12: The FROC paradigm

Table of contents

Online Supplementary Material

A. Online Appendix 12.A: Code used to generate the FROC plots
B. Online Appendix 12.B: CAMPI and cross correlation
C. Online Appendix 12.C: The Bunch transforms

12.1: Introduction

Until now focus has been on the receiver operating characteristic (ROC) paradigm. For diffuse interstitial lung disease[a], and diseases like it, where disease location is implicit (by definition *diffuse interstitial lung disease* is spread through and confined to lung tissues) this is an appropriate paradigm in the sense that possibly essential information is not being lost by limiting the radiologist's response in the ROC study to a single rating. The *extent* of the disease, i.e., how far it has spread, is an example of essential information that is still lost[1]. Anytime essential information is not accounted for in the analysis, as a physicist, the author sees a red flag. There is room for improvement in basic ROC methodology by modifying it to account for extent of disease. However, this is not the direction taken in this book. Instead, the direction taken is accounting for *location* of disease.

In clinical practice it is not only important to identify if the patient is diseased, but also to offer further guidance to subsequent care-givers regarding other characteristics (such as location, size, extent) of the disease. In most clinical tasks if the radiologist believes the patient may be diseased, there is a location (or more than one

[a] Diffuse interstitial lung disease refers to disease within both lungs that affects the interstitium or connective tissue that forms the support structure of the lungs' air sacs or alveoli. When one inhales, the alveoli fill with air and pass oxygen to the blood stream. When one exhales, carbon dioxide passes from the blood into the alveoli and is expelled from the body. When interstitial disease is present, the interstitium becomes inflamed and stiff, preventing the alveoli from fully expanding. This limits both the delivery of oxygen to the blood stream and the removal of carbon dioxide from the body. As the disease progresses, the interstitium scars with thickening of the walls of the alveoli, which further hampers lung function.

location) associated with the manifestation of the suspected disease. Physicians have a term for this: "focal disease", defined as "a disease located at a specific and distinct area".

For focal disease, the ROC paradigm restricts the collected information to a single rating representing the confidence level that there is disease *somewhere* in the patient's imaged anatomy. The emphasis on "*somewhere*" is because it begs the question: if the radiologist believes the disease is *somewhere*, why not have them to point to it? In fact they do "point to it" in the sense that they record the location(s) of suspect regions in their clinical report, but the ROC paradigm cannot use this information. *Neglect of location information leads to loss of statistical power as compared to paradigms that account for location information.* One way of compensating for reduced statistical power is to increase the sample size, which increases the cost of the study and is also considered unethical, because one is subjecting more patients to imaging procedures[2] and is not using the optimal paradigm/analysis. This is the *practical* reason for accounting for location information in the analysis. The *scientific* reason is that including location information yields a wealth of insight into processes limiting performance; these are discussed in **Chapter 16** and **Chapter 18**. This knowledge could have significant implications – currently widely unrecognized and unrealized - for how radiologists and algorithmic observers are designed, trained and evaluated. There are other scientific reasons for accounting for location, namely it accounts for unexplained features of ROC curves. Clinicians have long recognized problems with ignoring location[1,3] but, with one exception[4], much of the observer performance research community has yet to grasp this.

This part of the book, the subject of which has been the author's prime research interest over the past three decades, starts with an overview of the FROC paradigm introduced briefly in **Chapter 01**. Practical details regarding how to conduct and analyze an FROC study are deferred to **Chapter 18**. Here is an outline of this chapter. Four observer performance paradigms are compared with a visual schematic as to the kinds of information collected. An essential characteristic of the FROC paradigm, namely search, is introduced. Terminology to describe the FROC paradigm and its historical context is described. A pioneering FROC study using phantom images is described. Key differences between FROC ratings and ROC data are noted. The FROC plot is introduced and illustrate with **R** examples. The dependence of population and empirical FROC plots on perceptual signal-to-noise-ratio *(pSNR*) is shown. The expected dependence of the FROC curve on *pSNR* is illustrated with a "solar" analogy – understanding this is key to obtaining a good intuitive feel for this paradigm. The finite extent of the FROC curve, characterized by an end-point, is emphasized. Two sources of expertise are identified in a search task: search and lesion-classification performances, and it is shown that there is an expected inverse correlation between them.

The starting point is a comparison of four current observer performance paradigms.

12.2: Location specific paradigms

Location-specific paradigms take into account, to varying degrees, information regarding the locations of perceived lesions, so they are sometimes referred to as *lesion-specific* (or lesion-level[5]) paradigms. Usage of this term is discouraged. In this book *the term "lesion" is reserved for true malignant*[b] *lesions*[c] (as distinct from "*perceived lesions*" or "*suspicious regions*" that may not be true lesions).

All observer performance methods involve detecting the presence of true lesions; so ROC methodology is, in this sense, also lesion-specific. On the other hand location is a *characteristic* of true and perceived focal lesions, and methods that account for location are better termed *location-specific* than lesion-specific.

There are three location-specific paradigms: the free-response ROC (FROC) [6-11], the location ROC (LROC)[12-16] and the region of interest (ROI)[17,18].

Fig. 12.1 shows a mammogram as it might be interpreted according to current paradigms – these are not actual interpretations, just schematics to illustrate essential differences between the paradigms. The arrows point to two real lesions (as determined by subsequent follow-up of the patient) and the three lightly shaded crosses indicate perceived lesions or *suspicious regions*. From now on, for brevity, the author will use the term "*suspicious region*".

The numbers and locations of suspicious regions depend on the case and the observer' skill level. Some images are obviously non-diseased that the radiologist sees nothing suspicious in them, or they are obviously diseased that the suspicious regions are conspicuous. Then there is the gray area where one radiologist's suspicious region may not correspond to another observer's suspicious region. This difference is particularly apparent when one of the observers is an algorithmic observer.

In Fig. 12.1, evidently the radiologist found one of the lesions (the lightly shaded cross near the left most arrow), missed the other one (pointed to by the second arrow) and mistook two normal structures for lesions (the two lightly shaded crosses that are relatively far from any true lesion). To repeat, the term *lesion* is always a true or real lesion. *The prefix "true" or "real" is implicit*. The term *suspicious region* is reserved for any region that, as far as the observer is concerned, has "lesion-like" characteristics, but may not be a true lesion.

[b] Benign lesions are simply normal tissue variants that resemble a malignancy, but are not malignant.
[c] Lesion: a region in an organ or tissue that has suffered damage through injury or disease, such as a wound, ulcer, abscess, tumor, etc.

1) In the ROC paradigm, Fig. 12.1 (top-left), the radiologist assigns a single rating indicating the confidence level that there is at least one lesion *somewhere* in the image[d]. Assuming a 1 – 5 positive directed integer rating scale, if the left-most lightly shaded cross is a highly suspicious region then the ROC rating might be 5 (highest confidence for presence of disease).
2) In the free-response (FROC) paradigm, Fig. 12.1 (top-right), the dark shaded crosses indicate *suspicious regions that were marked or reported in the clinical report,* and the adjacent numbers are the corresponding ratings, which apply to specific regions in the image, unlike ROC, where the rating applies to the whole image. Assuming the allowed FROC ratings are 1 through 4, two marks are shown, one rated FROC-4, which is close to a true lesion, and the other rated FROC-1, which is not close to any true lesion. The third suspicious region, indicated by the lightly shaded cross, was not marked, implying its confidence level did not exceed the lowest reporting threshold. The marked region rated FROC-4 (highest FROC confidence) is likely what caused the radiologist to assign the ROC-5 rating to this image in the top-left ROC paradigm figure. [To avoid confusion, the rating is specified alongside the applicable paradigm.]
3) In the LROC paradigm, Fig. 12.1 (bottom-left), the radiologist provides a rating summarizing confidence that there is at least one lesion somewhere in the image (as in the ROC paradigm) *and* marks the *most suspicious region* in the image. In this example the rating might be LROC-5, the five rating being the same as in the ROC paradigm, and the mark may be the suspicious region rated FROC-4 in the FROC paradigm, and, since it is close to a true lesion, in LROC terminology it would be recorded as a *correct localization*. If the mark were not near a lesion it would be recorded as an *incorrect localization*. Only one mark is *allowed* in this paradigm, and in fact one mark is *required* on every image, even if the observer does not find any suspicious region to report. The forced mark, even when there is nothing to report, has cause confusion in the interpretation of this paradigm and its usage. The late Prof. Swensson has been the prime contributor to this paradigm.
4) In the region of interest (ROI) paradigm, the researcher segments the image into a number of regions-of-interest (ROIs) and the radiologist rates each ROI for presence of at least one suspicious region somewhere within the ROI. The rating is similar to the ROC rating, except it applies to the segmented ROI, not the whole image. Assuming a 1 – 5 positive directed integer rating scale in Fig. 12.1 (bottom-right) there are four ROIs and the ROI at ~9 o'clock might be rated ROI-5 as it contains the most suspicious light cross, the one at ~11 o'clock might be rated ROI-1 as it does not contain any light crosses, the one at ~3 o'clock might be rated LROC-2 or 3 (the unmarked light cross would tend to increase the confidence level) and the one at ~7 o'clock might be rated ROI-1. When different views of the same patient anatomy (perhaps in different

[d] The author's imaging physics mentor, Prof. Gary T. Barnes, had a unique way of emphasizing the word "somewhere" when he spoke about the neglect of localization in ROC methodology, as in "what do you mean the lesion is *somewhere* in the image? If you can see it you should point to it". Some of his grant applications were turned down because they did not include ROC studies, yet he was deeply suspicious of the ROC method because it neglected localization information. Around 1983 he guided the author towards a publication by Bunch et al. to be discussed below, and that started the author's career in this field. The author has since sensed that Gary was not totally happy with the author's obsession with this field, because it presumably distracted from other projects.

modalities) are available, it is assumed that all images are segmented consistently, and the rating for each ROI takes into account all views of that ROI in the different views (or modalities). In the example shown in Fig. 12.1 (bottom-right), each case yields 4 ratings. The segmentation shown in the figure is a schematic. In fact the ROIs could be clinically driven descriptors of location, such as "apex of lung" or "mediastinum", and the image does not have to have lines showing the ROIs (which would be distracting to the radiologist). The number of ROIs per image can be at the researcher's discretion and there is no requirement that every case have a fixed number of ROIs. Prof. Obuchowski has been the principal contributor to this paradigm.

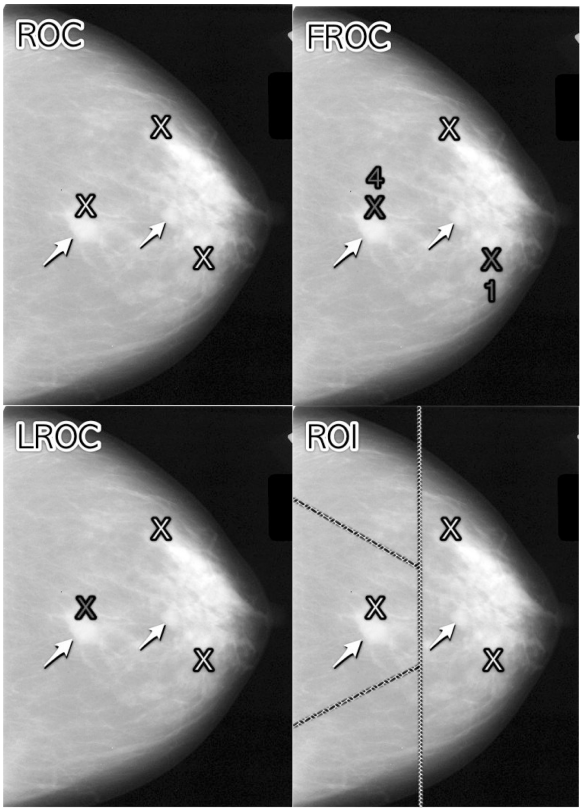

Fig. 12.1: A mammogram interpreted according to current observer performance paradigms. The arrows indicate two real lesions and the three light crosses indicate suspicious regions. Evidently the radiologist saw one of the lesions, missed the other lesion and mistook two normal structures for lesions. ROC (top-left): the radiologist assigns a single confidence level that *somewhere* in the image there is at least one lesion. FROC (top-right): the dark crosses indicate suspicious regions that are *marked* and the accompanying numerals are the FROC ratings. LROC (bottom-left): the radiologist provides a single rating that somewhere in the image there is at least one lesion and marks the most suspicious region. ROI (bottom-right): the image is divided into a number of regions-of-interest (by the researcher) and the radiologist rates each ROI for presence of at least one lesion somewhere within the ROI.

The rest of the book focuses on the FROC paradigm. It is the most general paradigm, special cases of which accommodate other paradigms. As an example, for diffuse interstitial lung disease the radiologist will implicitly "point to" the lung when disease is seen and the remaining cases are "definite normals".

12.3 The FROC paradigm as a search task

The FROC paradigm is equivalent to a *search* task. Any search task has two components: (i) *finding* something and (ii) *acting* on it. An example of a search task is looking for lost car-keys or a milk carton in the refrigerator. Success in a search task is finding the searched for object. Acting on it could be driving to work or drinking milk from the carton. There is *search-expertise* associated with any search task. Husbands are notoriously bad at finding the milk carton in the refrigerator (the author owes this analogy to Dr. Elizabeth Krupinski). Like anything else, *search expertise is honed by experience*, i.e., lots of practice. While the author is not good at finding the milk carton in the refrigerator he is good at finding files in his computer.

Likewise, a medical imaging search task has two components (i) finding suspicious regions and (ii) acting on each finding ("*finding*", used as a noun, is the actual term used by clinicians in their reports), i.e., determining the relevance of each finding to the health of the patient, and whether to report it. A general feature of a medical imaging search task is that the radiologist does not know a-priori if the patient is diseased and, if diseased, how many lesions are present. In the breast-screening context, it is known a-priori that about 5 out of 1000 cases have cancers, so 99.5% of the time odds are that the case has no malignant lesions (the probability of benign suspicious regions is much higher[19], about 13% for women aged 40-45). The radiologist searches the images for lesions. If a suspicious region is found, and provided it is sufficiently suspicious, the relevant location is *marked* and *rated* for confidence in being a lesion. The process is repeated for each suspicious region found in the case: a radiology report consists of a listing of search related actions specific to each patient. To summarize:

Free-response data = variable number (≥0) of mark-rating pairs per case. It is a record of the search process involved in finding disease and acting on each finding.

12.3.1 Proximity criterion and scoring the data

In the first two clinical applications of the FROC paradigm[8,20] the marks and ratings were indicated by a grease pencil on an acrylic overlay aligned, in a reproducible way, to the CRT displayed chest image. Credit for a correct detection and localization, termed a lesion-localization or *LL*-event[e], was given only if a mark was sufficiently close to an actual diseased region; otherwise, the observer's mark-rating pair was scored as a non-lesion localization or *NL*-event.

[e] The proper terminology for this paradigm has evolved. Older publications and some newer ones refer to these as true positive (TP) event, thereby confusing a ROC related term that does not involve search with one that does.

T*he use of ROC terminology, such as true positives or false positives to describe FROC data, seen in the literature on this subject, including the author's earlier papers*[9]*, is not conducive to clarity, and is strongly discouraged*.

The classification of each mark as either a LL or a NL is referred to as *scoring* the marks.

Definition:
NL = non-lesion localization, i.e., a mark that is not close to any lesion
LL = lesion localization, i.e., a mark that is close to a lesion

What is meant by *sufficiently close*? One adopts an *acceptance radius* (for spherical lesions) or *proximity criterion* (the more general case). What constitutes "close enough" is a clinical decision the answer to which depends on the application[21-23]. This source of arbitrariness in the FROC paradigm, which has been used to question its usage[24], is more in the mind of some researchers than in the clinic. It is not necessary for two radiologists to point to the same pixel in order for them to agree that they are seeing the same suspicious region. Likewise, two physicians (e.g., the radiologist finding the lesion on an x-ray and the surgeon responsible for resecting it) do not have to agree on the exact center of a lesion in order to appropriately assess and treat it. More often than not, "clinical common sense" can be used to determine if a mark actually localized the real lesion. When in doubt, the researcher should ask an independent radiologist how to score ambiguous marks.

For roughly spherical nodules a simple rule can be used. If a circular lesion is 10 mm in diameter, one can use the "touching-coins" analogy to determine the criterion for a mark to be classified as lesion localization. Each coin is 10 mm in diameter, so if they touch, their centers are separated by 10 mm, and the rule is to classify any mark within 10 mm of an actual lesion center as a LL mark, and if the separation is greater, the mark is classified as a NL mark. A recent paper[25] using FROC analysis gives more details on appropriate proximity criteria in the clinical context. Generally the proximity criterion is more stringent for smaller lesions than for larger one. However, for very small lesions allowance is made so that the criterion does not penalize the radiologist for normal marking "jitter". For 3D images the proximity criteria is different in the x-y plane vs. the slice thickness axis.

For clinical datasets, a rigid definition of the proximity criterion should *not* be used.

12.3.2 Multiple marks in the same vicinity

Multiple marks near the same vicinity are rarely encountered with radiologists, especially if the perceived lesion is mass-like (the exception would be if the perceived lesions were speck-like objects in a mammogram, and

even here radiologists tend to broadly outline the region containing perceived specks – they do not spend their clinical time marking individual specks with great precision). However, algorithmic readers, such as a CAD algorithm, are not radiologists and do tend to find multiple regions in the same area. Therefore, algorithm designers generally incorporate a clustering step[26] to reduce overlapping regions to a single region and assign the highest rating to it (i.e., *the rating of the highest rated mark, not the rating of the closest mark*). The reason for using the highest rating is that this gives full and deserved credit for the localization. Other marks in the same vicinity with lower ratings need to be discarded from the analysis; specifically, they should not be classified as NLs, because each mark has successfully located the true lesion to within the clinically acceptable criterion, i.e., any one of them is a good decision because it would result in a patient recall and point further diagnostics to the true location.

12.3.3: Historical context

The term "free-response" was coined in 1961 by Egan et al[6] to describe a task involving the detection of brief audio tone(s) against a background of white-noise (white-noise is what one hears if an FM tuner is set to an unused frequency). The tone(s) could occur at any instant within an active listening interval, defined by an indicator light bulb that is turned on. The listener's task was to respond by pressing a button at the specific instant(s) when a tone(s) was perceived (heard). The listener was uncertain how many true tones could occur in an active listening interval and when they might occur. Therefore, the number of responses (button presses) per active interval was a priori unpredictable: it could be zero, one or more. The Egan et al study did not require the listener to rate each button press, but apart from this difference and with a two-dimensional image replacing the listening interval, the acoustic signal detection study is similar to a common task in medical imaging, namely, prior to interpreting a screening case for possible breast cancer, the radiologist does not know how many diseased regions are actually present and, if present, where they are located. Consequently the case (all 4 views and possibly prior images) is *searched* for regions that appear to be suspicious for cancer. If one or more suspicious regions are found, and the level of suspicion of at least one of them exceeds the radiologists' minimum reporting threshold, the radiologist reports the region(s). At the author's former institution (University of Pittsburgh, Department of Radiology) the radiologists digitally outline and annotate (describe) suspicious region(s) that are found. As one would expect from the low prevalence of breast cancer, in the screening context about 5 per 1000 cases in the US, and assuming expert-level radiologist interpretations, about 90% of breast cases do not generate any marks, implying case-level specificity of 90%. About 10% of cases generate one or more marks and are recalled for further comprehensive imaging (termed *diagnostic workup*). Of marked cases about 90% generate one mark, about 10% generate 2 marks, and a rare case generates 3 or more marks. Conceptually, a mammography screening report consists of the locations of regions that exceed the threshold and the corresponding levels of suspicion, reported as a Breast Imaging Reporting and Data System (BIRADS) rating[27,28]. This type of information defines the free-response paradigm as it applies to medical imaging. Free-

response is a clinical paradigm: *it is a misconception that the paradigm forces the observer to keep marking and rating many suspicious regions per case* – as the mammography example shows, this is not the case. The very name of the paradigm, namely "free-response", implies, in plain English, "no forcing".

Described next is the first medical imaging application of this paradigm.

12.4: A pioneering FROC study in medical imaging

This section details an FROC paradigm phantom study with x-ray images conducted in 1978 that is often overlooked. With the obvious substitution of clinical images for the phantom images, this study is a template for how an FROC experiment should ideally be conducted. A detailed description of it is provided to set up the paradigm, the terminology used to describe it, and concludes with the FROC plot, which is still widely (and *incorrectly*, see **Chapter 17**) used as the basis for summarizing performance in this paradigm.

12.4.1 Image preparation

Bunch et al. [4] conducted the first radiological free-response paradigm study using simulated lesions. They drilled 10-20 small holes (the simulated lesions) at random locations in ten 5 cm x 5 cm x 1.6 mm Teflon™ sheets. A Lucite™ plastic block 5 cm thick was placed on top of each Teflon™ sheet to decrease contrast and increase scatter, thereby appropriately reducing visibility of the holes (otherwise the hole detection task would be too easy; as in ROC it is important that the task not be too easy or too difficult). Imaging conditions (kVp, mAs) were chosen such that, in preliminary studies, approximately 50% of the simulated lesions were correctly located at the observer's lowest confidence level. To minimize memory effects, the sheets were rotated, flipped or replaced between exposures. Six radiographs of 4 adjacent Teflon sheets, arranged in a 10 cm x 10 cm square, were obtained. Of these six radiographs one was used for training purposes, and the remaining five for actual data collection. Contact radiographs (i.e., with high visibility of the simulated lesions, similar in concept to the insert images of computerized analysis of mammography phantom images (CAMPI) described in Section 11.2 and Online Appendix 12.B) of the sheets were obtained to establish the true lesion locations. Observers were told that each sheet contained from 0 to 30 simulated lesions. A mark had to be within about 1 mm to count as a correct localization; *a rigid definition was deemed unnecessary* (the emphasis is because of this simple and practical advice is ignored, not by the user community, but by "experts"). Once images had been prepared, observers interpreted them. This is how Bunch et al conducted the image interpretation part of their experiment.

12.4.2 Image Interpretation and the 1-rating

Observers viewed each film and *marked* and *rated* any visible holes with a felt-tip pen on a transparent overlay taped to the film at one edge (this allowed the observer to view the film directly without the distracting effect of

previously made marks – in digital interfaces it is important to implement a show/hide feature in the user interface). The record of mark-rating pairs generated by the observer constitutes free-response data.

The observers used a 4-point ordered rating scale with 4 representing "most likely a simulated lesion" to 1 representing "least likely a simulated lesion". Note the meaning of the 1 rating: least likely a simulated lesion. There is confusion with some using the FROC-1 rating to mean "definitely not a lesion". If that were the observer's understanding, then logically the observer would "fill up" the entire image, especially parts outside the patient anatomy, with 1's, as each of these regions is "definitely not a lesion". Since the observer did not behave in this unreasonable way, the meaning of the FROC-1 rating, as they interpreted it, or were told, must have been "I am done with this image, show me the next one".

When correctly used, the 1-rating means there is some finite, perhaps small, probability that the marked region is a lesion. In this sense the free-response rating scale is *asymmetric*. Compare the 5 rating ROC scale, where ROC-1 = "patient is definitely not diseased" and ROC-5 = "patient definitely diseased". This is a *symmetric* confidence level scale. In contrast the free-response confidence level scale labels different degrees of *positivity* in presence of disease. Table 12.1 compares the ROC 5-rating study to a FROC 4-rating study.

Table 12.1: comparison of ROC and FROC rating scales: note that the FROC rating is one less than the corresponding ROC rating and that there is no rating corresponding to ROC-1 The observer's way of indicating definitely non-diseased images is by simply not marking them. NA = not available.

ROC paradigm		FROC paradigm	
Rating	Observer's categorization	Rating	Observer's categorization
1	Definitely not-diseased	NA	Image is not marked
2	…	1	Just possible it is a lesion
3		2	…
4		3	
5	Definitely diseased	4	Definitely a lesion

The FROC rating is one less than the corresponding ROC rating because the ROC-1 rating is not used by the observer; the observer indicates such images by the simple expedient of *not* marking them.

12.4.3 Scoring the data

Scoring the data was defined (§12.3) as the process of classifying each mark-rating pair as NL or LL, i.e., as an incorrect or a correct decision, respectively. In the Bunch et al study, after each case was read the person running the study (i.e., Phil Bunch) compared the marks on the overlay to the true lesion locations on the contact radiographs and scored the marks as lesion localizations (LLs: lesions correctly localized to within about 1 mm radius) or non-lesion localizations (NLs: all other marks). Bunch et al actually used the terms "true

positive" and "false positive" to describe these events. This practice, still used in publications in this field, is confusing because there is ambiguity about whether these terms, commonly used in the ROC paradigm, are being applied to the case as a whole or to specific regions in the case.

12.4.4: The free-response receiver operating characteristic (FROC) plot

The free-response receiver operating characteristic (FROC) plot was introduced, also in an auditory detection task, by Miller[29] as a way of visualizing performance in the free-response auditory tone detection task. In the medical imaging context, assume the marks have been classified as NLs (non-lesion localizations) or LLs (lesion localizations), along with their associated ratings. Non-lesion localization fraction (*NLF*) is defined as the total number of NLs at or above a threshold rating divided by the total number of cases. Lesion localization fraction (*LLF*) is defined as the total number of LLs at or above the same threshold rating divided by the total number of lesions in the case set. The FROC plot is defined as that of LLF (ordinate) vs. NLF as the threshold is varied. While the ordinate *LLF* is a proper fraction, e.g., 30/40 assuming 30 LLs and 40 true lesions, the abscissa is an *improper* fraction that can exceed unity, like 35/21 assuming 35 NLs on 21 cases). The NLF notation is not ideal: it is used for notational symmetry and compactness.

Definitions:

- NLF = cumulated NL counts at or above threshold rating divided by total number of cases.
- LLF = cumulated LL counts at or above threshold rating divided by total number of lesions.
- The FROC curve is the plot of LLF (ordinate) vs. NLF.
- The upper-right most operating point is termed the *end-point* and its coordinates are denoted $\left(NLF_{\max}, LLF_{\max}\right)$.

Following Miller's suggestion, Bunch et al[7,30] plotted lesion localization fraction (*LLF*) along the ordinate vs. non-lesion localization fraction (*NLF*) along the abscissa. Corresponding to the different threshold ratings, pairs of (*NLF*, *LLF*) values, or operating points on the FROC, were plotted. For example, in a positive directed four-rating FROC study, such as employed by Bunch et al, 4 FROC operating points resulted: that corresponding to marks rated 4s; that corresponding to marks rated 4s or 3s; the 4s, 3s, or 2s; and finally the 4s, 3s, 2s, or 1s. An R-rating (integer R > 0) FROC study yields at most R operating points. So Bunch et al were able to plot only 4 operating points per reader, Fig. 6 *ibid*.[f] Lacking a method of fitting a continuous FROC curve to the operating points, they did the best they could, and manually "French-curved" fitted curves. In 1986, the author followed the same practice in his first paper on this topic[8]. In 1989 the author described[9] a method for fitting such

[f] Fig. 7 *ibid* has about 12 operating points, as it includes three separate interpretations by the same observer; moreover the area scaling implicit in the paper assumes a homogenous and isotropic image, i.e., the probability of a NL is proportional to the image area over which it is calculated, which is valid for a uniform background phantom. Clinical images are not homogenous and isotropic and therefore not scalable in the Bunch et al. sense.

operating points, and developed software called FROCFIT, but the fitting method is obsolete, as the underlying statistical model has been superseded, see **Chapter 18**, and moreover, it is shown that the FROC plot is a poor visual descriptor of performance.

If continuous ratings are used, the procedure is to start with a high threshold so none of the ratings exceed the threshold, and gradually lower the threshold. Every time the threshold crosses the rating of a mark, or possibly multiple marks, the total count of LLs and NLs exceeding the threshold is divided by the appropriate denominators yielding the ‘‘raw’’ FROC plot widely (and inappropriately, see **Chapter 17**) used in current research. For example, when an LL rating just exceeds the threshold, the operating point jumps up by 1/(total number of lesions), and if two LLs simultaneously just exceed the threshold, the operating point jumps up by 2/(total number of lesions). If an NL rating just exceeds the threshold, the operating point jumps to the right by 1/(total number of cases). If an LL rating and a NL rating simultaneously just exceed the threshold, the operating point moves diagonally, up by 1/(total number of lesions) and to the right by 1/(total number of cases). The reader should get the general idea by now and recognize that the cumulating procedure is very similar to the manner in which ROC operating points were calculated, the only differences being in the quantities being cumulated and the relevant denominators.

Having seen how a binned data FROC study is conducted and scored, and the results “French-curved” as an FROC plot, typical simulated plots, generated under controlled conditions, are shown next, both for continuous ratings data and for binned rating data. Such demonstrations, that illustrate trends, are impossible using real datasets. The reader should take the author's word for it (for now) that the simulator used is the simplest one possible that incorporates key elements of the search process. Details of the simulator are given in **Chapter 16**, but for now the following summary should suffice.

The simulator is characterized by three parameters μ, λ and ν. The ν parameter characterizes the *ability of the observer to find lesions*, the λ parameter characterizes the *ability of the observer to avoid finding non-lesions* and μ parameter characterizes the *ability of the observer to correctly classify a found suspicious region as a true lesion or a non-lesion*. The reader should think of μ as a perceptual signal-to-noise ratio (*pSNR*) or conspicuity of the lesion, similar to the separation parameter of the binormal model, that separates two normal distributions describing the sampling of ratings of NLs and LLs. Finally, there is a threshold parameter ζ_1 that determines if a found suspicious region is actually marked. If ζ_1 is negative infinity, then all found suspicious regions are marked and conversely, as ζ_1 increases, only those suspicious regions whose confidence level exceeds ζ_1 are marked. The concept of *pSNR* is clarified in §12.5.2.

12.5: Population and binned FROC plots

Fig. 12.2 (A - C) shows simulated population FROC plots when the ratings are not binned, generated by file **mainFrocCurvePop.R** described in Online Appendix 12.A. FROC data from 20,000 cases, half of them non-diseased are generated (the code takes a while to finish). The very large number of cases minimizes sampling variability; hence the term "*population*" curves. Additionally, the reporting threshold ζ_1 was set to negative infinity to ensure that all suspicious regions are marked; with higher thresholds, suspicious regions with confidence levels below the threshold would not be marked and the rightward and upward traverses of the shown curves would be truncated. Plots (A) – (C) correspond to μ equal to 0.5, 1 and 2, respectively. Plots (D) – (F) correspond to 5-ratings binned data for 50 non-diseased and 50 diseased cases, and the same values of μ; the relevant file is **mainFrocCurveBinned.R**. [Binning 20,000 cases requires much more time and is not useful.]

1) Plots (A) – (C) show quasi-continuous plots, while (D) – (F) show operating points, five per plot, connected by straight line segments, so they are termed *empirical FROC curves*, analogous to the empirical ROC curves encountered in previous chapters. At a "microscopic level" plots (A) – (C) are also discrete, but one would need to "zoom in" to see the discrete behavior (upward and rightward jumps) as each rating crosses a sliding threshold.
2) The empirical plots in the bottom row (D - F) of Fig. 12.2 are subject to sampling variability and will not, in general, match the population plots. The reader should try different values of the **seed** variable in the code.
3) *In general FROC plots do not extend indefinitely to the right*. Fig. 5 in the Bunch et al paper is incorrect in implying, with the arrows, that the plots extend indefinitely to the right. [Notation differences: In Bunch et al *P(TP)* or v is equivalent to *LLF*. To avoid confusion with the λ - parameter of the radiological search model, the variable Bunch et al call λ is equivalent to *NLF* in this book.]
4) Like an ROC plot, the population FROC curve rises monotonically from the origin, initially with infinite slope (this may not be evident for Fig. 12.5. (A), but it is true, see code snippet below). If all suspicious regions are marked, i.e., $\zeta_1 = -\infty$, the plot reaches its upper-right most limit, termed the *end-point*, with zero slope (again, this may not be evident for (A), but it is true, see code snippet below; here *x* and *y* are arrays containing NLF and LLF, respectively). In general these characteristics, i.e., initial infinite slope and zero final slope, are not true for empirical plots Fig. 12.2 (D – F).

12.5.1: Code Snippet

```
> mu
[1] 0.5
```

```
> (y[2]-y[1])/(x[2]-x[1]) # slope at origin
[1] Inf
> (y[10000]-y[10000-1])/(x[10000]-x[10000-1]) # slope at end-point
[1] 0
```

5) *Assuming all suspicious regions are marked, the end-point* $\left(NLF_{\max}, LLF_{\max}\right)$ *represents a literal end of the extent of the population FROC curve. This will become clearer in following chapters, but for now it should suffice to note that the region of the population FROC plot to the upper-right of the end-point is inaccessible to the observer. If sampling variability is involved it is possible for the observed end-point to extend into this inaccessible space.*

6) *There is an inverse correlation between* $LLF_{\max}$ *and* $NLF_{\max}$ *analogous to that between sensitivity and specificity in ROC analysis. The end-point* $\left(NLF_{\max}, LLF_{\max}\right)$ *of the FROC tends to approach the point (0,1) as the perceptual SNR of the lesions approaches infinity.* As μ decreases the FROC curve approaches the x-axis and extends to large values along the abscissa, as in Fig. 12.2 (B). This is the "chance-level" FROC, where the reader detects few lesions, and makes many NL marks.

7) The slope of the population FROC decreases monotonically as the operating point moves up the curve, always staying non-negative and it approaches zero, flattening out at an ordinate *less than unity*. Some publications[31] (Fig. 3 *ibid.*) and Ref. [32] (Fig. 1 *ibid.*) incorrectly show *LLF* reaching unity. This is generally not the case unless the lesions are particularly conspicuous. This is well known to CAD researchers and to anyone who has conducted FROC studies with radiologists. *LLF* reaches unity for large μ, which can be confirmed by setting μ to a large value, e.g., 10, Fig. 12.3 (A). On the unit variance normal distribution scale, a value of 10, equivalent to 10 standard deviations, is effectively infinite]

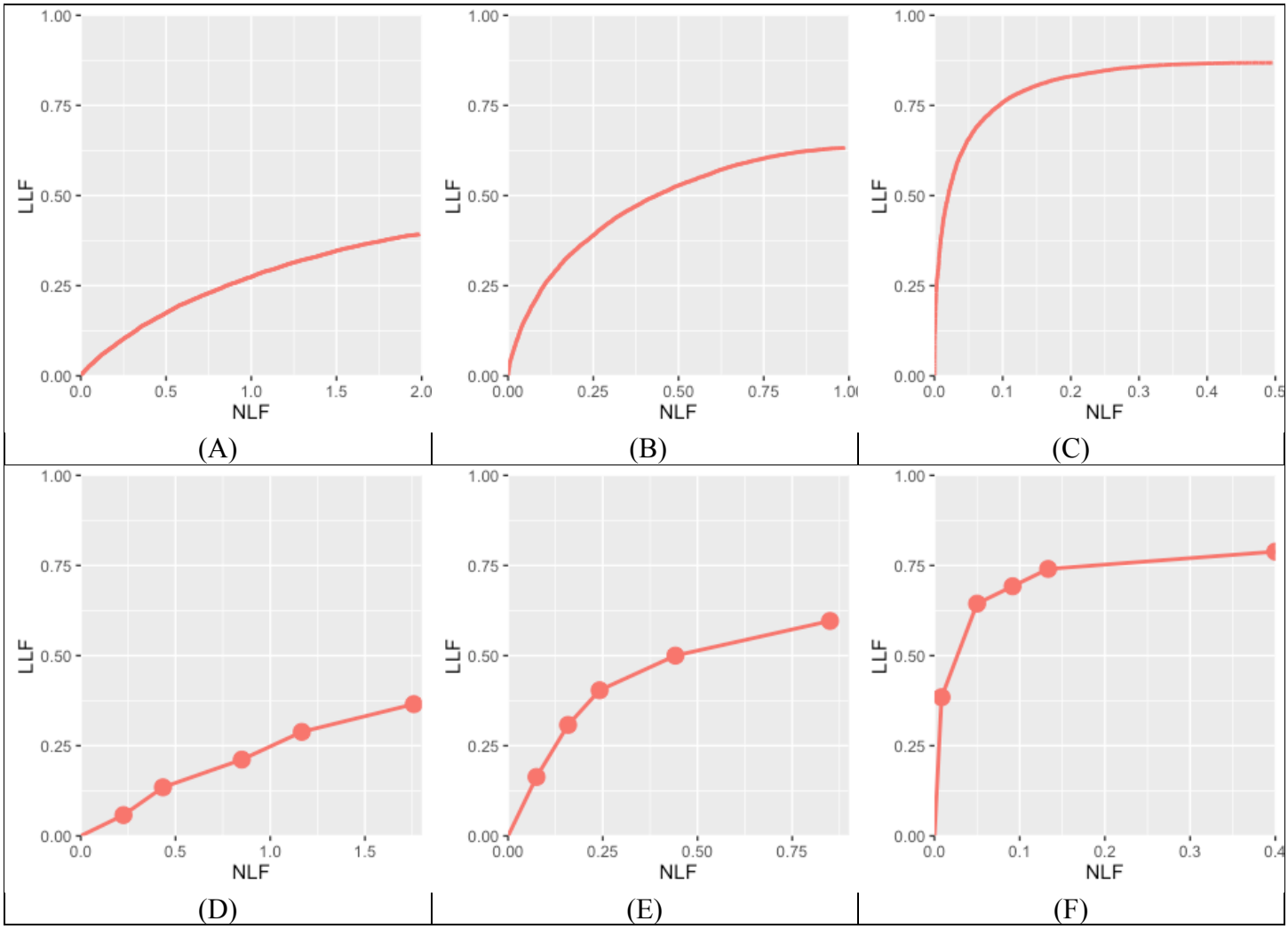

Fig. 12.2 (A – F): Top row, (A) through (C): Population FROC plots for μ = 0.5, 1, 2; the other parameters are $\lambda = 1$, $\nu = 1$, $\zeta_1 = -\infty$ and $L_{max} = 2$ is the maximum number of lesions per case in the dataset. The plots in the bottom row (D - F) correspond to 50 non-diseased and 70 diseased cases, where the data was binned into 5 bins, and other parameters are unchanged. As μ increases, the uppermost point moves upwards and to the left, approaching the top-left corner in the limit μ = infinity, Fig. 12.3 (A). The top row of images was produced by **`MainFrocCurvePop.R`** while the bottom row by **`mainFrocCurveBinned.R`**.

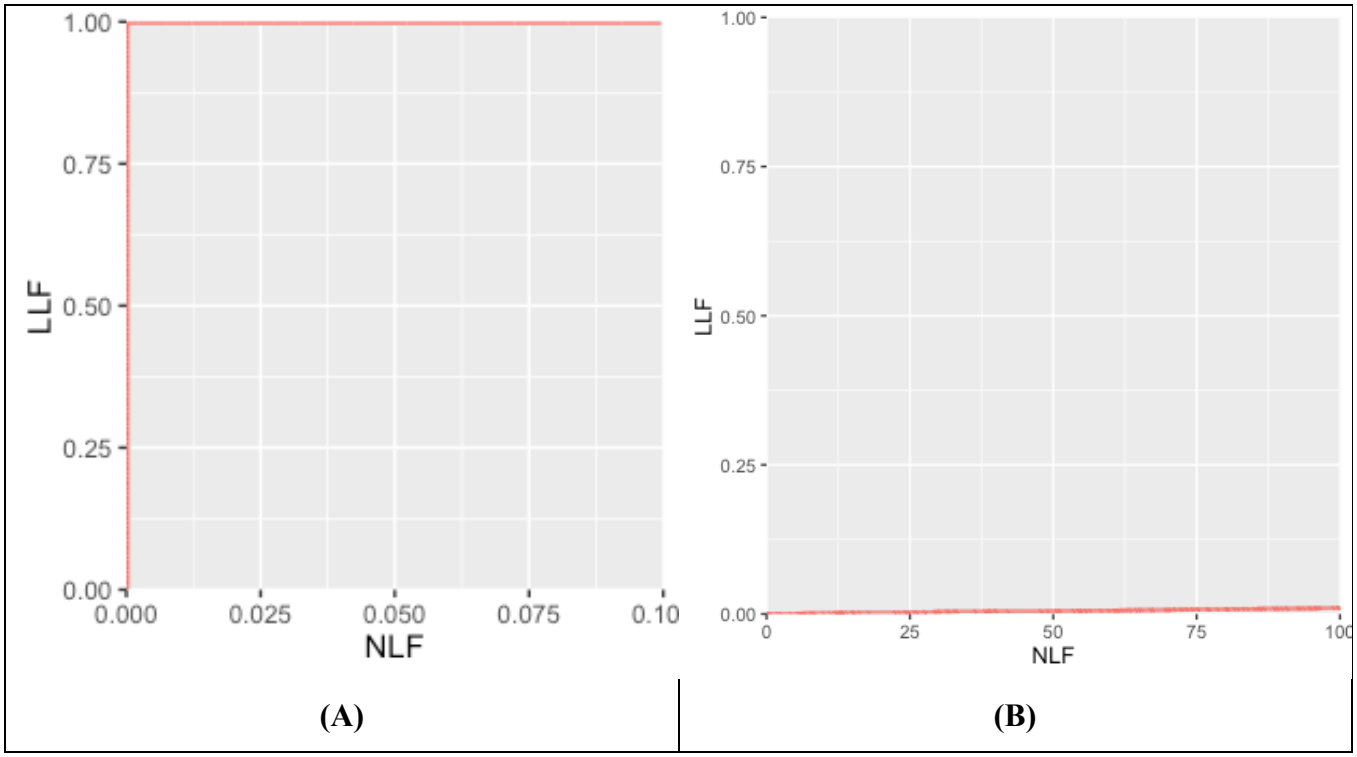

Fig. 12.3 (A - B): (A) FROC plot for μ = 10 in code file **`mainFrocCurvePop.R`**. Note the small range of the NLF axis (it extends to 0.1). In this limit the ordinate reaches unity, but the abscissa is limited to a small value; see "solar analogy" §12.6 for explanation. (B) This plot corresponds to μ = 0.01, depicting near *chance-level* performance. Note the greatly increased traverse in the x-directions and the slight upturn in the plot near NLF = 100.

12.5.2: Perceptual SNR

The shape and extent of the FROC plot is to a large extent determined by the *perceptual*[g] SNR of the lesions, *pSNR*, modeled by a parameter μ. Perceptual SNR is the ratio of perceptual signal to perceptual noise. To get to perceptual variables one needs a model of the eye-brain system that transforms physical image brightness variations to corresponding perceived brightness variations, and such models exist[33-35]. For uniform background images, like the phantom images used by Bunch et al, physical signal can be measured by a *template* function that has the same attenuation profile as the true lesion. Assuming the template is aligned with the lesion the *cross-correlation* between the template function and the image pixel values is related to the numerator of SNR. The cross correlation is defined as the summed product of template function pixel values times the corresponding pixel values in the actual image. Next, one calculates the cross-correlation between the template function and the pixel values in the image when the template is centered over regions known to be *lesion free*. Subtracting the mean of these values (over several lesion free regions) from the centered value gives the numerator of SNR. The denominator is the standard deviation of the cross correlation values in the lesion free areas. Appendix 12.C has further details on calculating *physical* SNR, which derives from the author's CAMPI (computer analysis of mammography phantom images) work[36-40]. To calculate *perceptual* SNR one repeats

[g] Since humans make the decisions, it would be incorrect to label these as *physical* signal-to-noise-ratios; that is the reason for qualifying them as *perceptual* SNRs.

these measurements but the visual process, or some model of it (e.g., the Sarnoff JNDMetrix visual discrimination model[35,41,42]), is used to filter the image prior to calculation of the cross-correlations.

An analogy may be helpful at this point. *Finding the sun in the sky is a search task, so it can be used to illustrate important concepts.*

12.6: The "solar" analogy: search vs. classification performance

Consider the sun, regarded as a "lesion" to be detected, with two daily observations spaced 12 hours apart, so that at least one observation period is bound to have the sun "somewhere up there". Furthermore, the observer is assumed to know their GPS coordinates and have a watch that gives accurate local time, from which an accurate location of the sun can be deduced. Assuming clear skies and no obstructions to the view, the sun will always be correctly located and no reasonable observer will ever generate a non-lesion localization or *NL,* i.e., no region of the sky will be erroneously "marked".

FROC curve implications of this analogy are:

- Each 24-hour day corresponds to two "trials" in the Egan et al sense[1], or two cases – one diseased and one non-diseased - in the medical imaging context.
- The denominator for calculating *LLF* is the total number of AM days, and the denominator for calculating *NLF* is twice the total number of 24-hour days.
- *Most important,* $LLF_{max} = 1$ and $NLF_{max} = 0$.

In fact, even when the sun is not directly visible due to heavy cloud cover, since the actual location of the sun can be deduced from the local time and GPS coordinates, the rational observer will still "mark" the correct location of the sun and not make any false sun localizations or "non-lesion localizations", NLs. Consequently even in this example $LLF_{max} = 1$ and $NLF_{max} = 0$.

The conclusion is that in a task where a target is known to be present in the field of view and its location is known, the observer will always reach $LLF_{max} = 1$ and $NLF_{max} = 0$. Why are *LLF* and NLF subscripted *max*? By randomly not marking the position of the sun even though it is visible, for example, using a coin toss to decide whether or not to mark the sun, the observer can "walk down" the *y*-axis of the FROC plot, reaching $LLF = 0$ and $NLF = 0$.[h] Alternatively, the observer uses a very large threshold for reporting the sun, and as this threshold is lowered the operating point "walks down" the curve. The reason for allowing the observer to

[h] The logic is very similar to that used in §3.9.1 to describe how the ROC observer can "walk along" along the chance diagonal of the ROC curve.

"walk down" the vertical is simply to demonstrate that a continuous FROC curve from the origin to the highest point (0,1) can in fact be realized.

Now consider a fictitious otherwise earth-like planet where the sun can be at *random* positions, rendering GPS coordinates and the local time useless. All one knows is that the sun is somewhere, in the upper or lower hemispheres subtended by the sky. If there are no clouds and consequently one can see the sun clearly during daytime, a reasonable observer would still correctly located the sun while not marking the sky with any incorrect sightings, so $LLF_{max} = 1$ and $NLF_{max} = 0$. This is because, in spite of the fact that the expected location is unknown, the high contrast sun is enough the trigger the peripheral vision system, so that even if the observer did not start out looking in the correct direction, peripheral vision will drag the observer's gaze to the correct location for foveal viewing.

The implication of this is that fundamentally different mechanisms from that considered in conventional observer performance methodology, namely *search* and *lesion-classification*, are involved. Search describes the process of *finding* the "lesion" while not finding "non-lesions". Classification describes the process, once a possible sun location has been found, of recognizing that it is indeed the sun and marking it. Recall that search involves two steps: finding the object of the search and acting on it. Search and lesion-classification performances describe the abilities of an observer to efficiently perform these steps.

Think of the eye as two cameras: a low-resolution camera (peripheral vision) with a wide field-of-view plus a high-resolution camera (foveal vision) with a narrow field-of-view If one were limited to viewing with the high-resolution camera one would spend so much time steering the high-resolution narrow field-of-view camera from spot-to-spot that one would have a hard time finding the desired stellar object. Having a single high-resolution narrow field of view vision would also have negative evolutionary consequences as one would spend so much time scanning and processing the surroundings with the narrow field of view vision that one would miss dangers or opportunities. Nature has equipped us with essentially two cameras; the first low-resolution camera is able to "digest" large areas of the surround and process it rapidly so that if danger (or opportunity) is sensed, then the eye-brain system rapidly steers the second high-resolution camera to the location of the danger (or opportunity). This is Nature's way of optimally using the eye-brain system. For a similar reason astronomical telescopes come with a wide field of view lower resolution "spotter scope".

Since the large field-of-view low-resolution peripheral vision system has complementary properties to the small field-of-view high-resolution foveal vision system, one expects an inverse correlation between search and

lesion-classification performances. Stated generally, search involves two complementary processes: *finding* the suspicious regions and *deciding* if the found region is actually a lesion, and that there should be an inverse correlation between performance in the two tasks, see **Chapter 19**.

When cloud cover completely blocks the fictitious random-position sun there is no stimulus to trigger the peripheral vision system to guide the fovea to the correct location. Lacking any stimulus, the observer is reduced to guessing and is led to different conclusions depending upon the benefits and costs involved. If, for example, the guessing observer earns a dollar for each LL and is fined a dollar for each NL, then the observer will likely not make any marks as the chance of winning a dollar is much smaller than losing many dollars. For this observer $LLF_{\max} = 0$ and $NLF_{\max} = 0$, and the operating point is "stuck" at the origin. If, on the other hand, the observer is told every LL is worth a dollar and there is no penalty to NLs, then with no risk of losing the observer will "fill up" the sky with marks. In either situation the locations of the marks will lie on a grid determined by the ratio of the 4π solid angle (subtended by the spherical sky) and the solid angle Ω subtended by the sun. By marking every possible grid location the observer is trivially guaranteed to "detect" the sun and earn a dollar irrespective of its random location and reach $LLF = 1$, but now the observer will generate lots of non-lesion localizations, so maximum *NLF* will be large:

$$NLF_{\max} = \frac{4\pi}{\Omega} \quad . \qquad \textbf{(12.6.1)}$$

The FROC plot for this guessing observer is the straight line joining (0,0) to $\left(NLF_{\max}, 1\right)$. For example, if the observer fills up half the sky then the operating point, averaged over many trials, is

$$\left(0.5\, NLF_{\max}, 0.5\right) \quad . \qquad \textbf{(12.6.2)}$$

Radiologists do not guess – there is much riding on their decisions to allow them that luxury – so in the clinical situation, if the lesion is not seen, the radiologist will not mark the image at random.

The analogy is not restricted to the sun, which one might argue is an almost infinite SNR object and therefore atypical. As another example, consider finding stars or planets. In clear skies, if one knows the constellations, one can still locate bright stars and planets like Venus or Jupiter. With less bright stars and / or obscuring clouds, there will be false-sightings and the FROC plot could approach a flat horizontal line at ordinate equal to zero, but the observer will not fill up the sky with false sightings of a desired star.

False sightings of objects in astronomy do occur. Finding a new astronomical object is a search task, where as always one can have two outcomes, correct localization (LL) or incorrect localizations (NLs). At the time of writing there is a hunt for a new planet, possibly a gas giant[i], that is much further away than even the newly demoted Pluto. There is an astronomer in Australia[j] who is particularly good at finding super novae (an exploding star; one has to be looking in the right region of the sky at the right time to see the relatively brief explosion). His equipment is primitive by comparison to the huge telescope at Mt. Palomar, but his advantage is that he can rapidly point his 15" telescope at a new region of the sky and thereby cover a lot more sky, in a given unit of time, than is possible with the 200" Mt. Palomar telescope. His *search expertise* is particularly good. Once correctly pointed, the Mt. Palomar telescope will reveal a lot more detail about the object than is possible with the smaller telescope, i.e., it has high lesion-classification accuracy. In the medical imaging context this detail (the shape of the lesion, its edge characteristics, presence of other abnormal features, etc.) allows the radiologist to diagnose whether the lesion is malignant or benign. Once again one sees that there should be an inverse correlation between search and lesion-classification performances.

Prof. Jeremy Wolfe of Harvard University is an expert in visual search, and the interested reader is referred to his many publications on search[43,44]. As noted by him, rare items are often missed: to paraphrase him, "*things that are not seen often are often not seen*" [45]. So the problem faced by an astronomer looking for supernova events, a terminal security agency baggage inspector looking for explosives, and the radiologist interpreting a screening mammogram for rare cancers, are similar at a fundamental level. All of these tasks are low prevalence search tasks.

12.7: Discussion / Summary

This chapter has introduced the FROC paradigm, the terminology used to describe it and a common operating characteristic associated with it, namely the FROC. In the author's experience this paradigm is widely misunderstood. The following suggested rules might reduce the confusion:

- Avoid using the term "lesion-specific" to describe location-specific paradigms.
- Avoid using the term "lesion" when one means a "suspicious region" that may not be a true lesion.
- Avoid using ROC-specific terms, such as true positive and false positive, that apply to the whole case, to describe location-specific terms such as lesion and non-lesion localization, that apply to localized regions of the image. This issue will come up in later chapters.

[i] https://en.wikipedia.org/wiki/Tyche_(hypothetical_planet)
[j] https://en.wikipedia.org/wiki/Robert_Evans_(astronomer)

- Avoid using the FROC-1 rating to mean in effect "*I see no signs of disease in this image* ", when in fact it should be used as the lowest level of a reportable suspicious region. The former usage amounts to wasting a confidence level.
- Do not show FROC curves as reaching the unit ordinate, as this is the exception rather than the rule.
- Do not conceptualize FROC curves as extending to large values to the right.
- Arbitrariness of the proximity criterion and multiple marks in the same region are not clinical constraints - they are problems only in the mind of the data analyst unfamiliar with clinical practice. Interactions with clinicians, preferably using a medical physicist, will allow selection of an appropriate proximity criterion for the task at hand and the latter problem only occurs with algorithmic observers and is readily fixed.

Additional points made in this chapter are: There is an inverse correlation between $LLF_{\max}$ and $NLF_{\max}$ analogous to that been sensitivity and specificity in ROC analysis. The end-point $\left(NLF_{\max}, LLF_{\max}\right)$ of the FROC curve tends to approach the point (0,1) as the perceptual SNR of the lesions approaches infinity. The solar analogy is highly relevant to understanding the search task. In search tasks two types of expertise are at work: search and lesion-classification performances, and there is an expected inverse correlation between them.

Online Appendix 12.A describes, and explains in detail, the code used to generate the population FROC curves shown in Fig. 12.2 (A - C). Online Appendix 12.B details how one calculates *physical* signal to noise ratio (SNR) for an object on a uniform noise background. This is useful in understanding the concept of *perceptual* signal to noise ratio denoted μ. Online Appendix 12.C is for those who wish to understand the Bunch et al paper[7] in more depth. This paper has certain transformations, sometimes referred to as the *Bunch transforms*, which relate an ROC plot to an FROC plot and vise-versa. *It is not a model of FROC data*. The reason for including it is that this important paper is much overlooked, and if the author does not write it, no one else will.

The FROC plot is the first proposed way of visually summarizing FROC data. The next chapter deals with different empirical operating characteristics that can be defined from an FROC dataset.

12.8 References

1. Black WC. Anatomic Extent of Disease: A Critical Variable in Reports of Diagnostic Accuracy. *Radiology.* 2000;217(2):319-320.
2. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *Jama.* 2002;288(3):358-362.
3. Black WC, Dwyer AJ. Local versus Global Measures of Accuracy: An Important Distinction for Diagnostic Imaging. *Med Decis Making.* 1990;10(4):266-273.

4. Obuchowski NA, Mazzone PJ, Dachman AH. Bias, underestimation of risk, and loss of statistical power in patient-level analyses of lesion detection. *Eur Radiol.* 2010;20:584–594.
5. Alberdi E, Povyakalo AA, Strigini L, Ayton P, Given-Wilson R. CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer Assisted Radiology and Surgery.* 2008;3(1):115-122.
6. Egan JP, Greenburg GZ, Schulman AI. Operating characteristics, signal detectability and the method of free response. *J Acoust Soc Am.* 1961;33:993-1007.
7. 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.
8. Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes GT, Fraser RG. Digital and Conventional Chest Imaging: A Modified ROC Study of Observer Performance Using Simulated Nodules. *Radiology.* 1986;158:35-39.
9. Chakraborty DP. Maximum Likelihood analysis of free-response receiver operating characteristic (FROC) data. *Med Phys.* 1989;16(4):561-568.
10. Chakraborty DP, Winter LHL. Free-Response Methodology: Alternate Analysis and a New Observer-Performance Experiment. *Radiology.* 1990;174:873-881.
11. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis and validation. *Med Phys.* 2004;31(8):2313-2330.
12. Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. *Radiology.* 1975;116:533-538.
13. 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.
14. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys.* 1996;23(10):1709 -1725.
15. Judy PF, Swensson RG. Lesion detection and signal-to-noise ratio in CT images. *Medical Physics.* 1981;8(1):13-23.
16. Swensson RG, Judy PF. Detection of noisy visual targets: Models for the effects of spatial uncertainty and signal-to-noise ratio. *Perception & Psychophysics.* 1981;29(6):521-534.
17. Obuchowski NA, Lieber ML, Powell KA. Data Analysis for Detection and Localization of Multiple Abnormalities with Application to Mammography. *Acad Radiol.* 2000;7(7):516-525.
18. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient-clustered data. *Acad Radiol.* 2000;7(6):413-419.
19. ERNSTER VL. The epidemiology of benign breast disease. *Epidemiologic reviews.* 1981;3(1):184-202.
20. Niklason LT, Hickey NM, Chakraborty DP, et al. Simulated Pulmonary Nodules: detection with Dual-Energy Digital versus Conventional Radiography. *Radiology.* 1986;160:589-593.
21. Haygood TM, Ryan J, Brennan PC, et al. On the choice of acceptance radius in free-response observer performance studies. *BJR.* 2012;Published online before print May 9, 2012.
22. 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.
23. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Medical Physics.* 1999;26(2):267-275.
24. Gur D, Rockette HE. Performance Assessment of Diagnostic Systems under the FROC paradigm: Experimental, Analytical, and Results Interpretation Issues. *Acad Radiol.* 2008;15:1312-1315.
25. 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.
26. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics).* 1979;28(1):100-108.
27. D'Orsi CJ, Bassett LW, Feig SA, et al. *Illustrated Breast Imaging Reporting and Data System.* Reston, Va: American College of Radiology; 1998.

28. D'Orsi CJ, Bassett LW, Berg WA. *ACR BI-RADS-Mammography.* 4th ed. Reston, Va: American College of Radiology; 2003.
29. 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(6(2)):1473-1476.
30. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *Proc SPIE.* 1977;127:124-135.
31. Popescu LM. Model for the detection of signals in images with multiple suspicious locations. *Medical Physics.* 2008;35(12):5565-5574.
32. Popescu LM. Nonparametric ROC and LROC analysis. *Medical Physics.* 2007;35(5):1556-1564.
33. Van den Branden Lambrecht CJ, Verscheure O. Perceptual quality measure using a spatiotemporal model of the human visual system. Paper presented at: Electronic Imaging: Science & Technology1996.
34. Daly SJ. Visible differences predictor: an algorithm for the assessment of image fidelity. Paper presented at: SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology1992.
35. Lubin J. A visual discrimination model for imaging system design and evaluation. *Vision models for target detection and recognition.* 1995;2:245-357.
36. Chakraborty DP, Sivarudrappa M, Roehrig H. Computerized measurement of mammographic display image quality. Paper presented at: Proc SPIE Medical Imaging 1999: Physics of Medical Imaging1999; San Diego, CA.
37. Chakraborty DP, Fatouros PP. Application of computer analyis of mammography phantom images (CAMPI) methodology to the comparison of two digital biopsy machines. Paper presented at: Proc SPIE Medical Imaging 1998: Physics of Medical Imaging; 24 July 1998, 1998.
38. Chakraborty DP. Comparison of computer analysis of mammography phantom images (CAMPI) with perceived image quality of phantom targets in the ACR phantom. Paper presented at: Proc. SPIE Medical Imaging 1997: Image Perception; 26-27 February 1997, 1997; Newport Beach, CA.
39. Chakraborty DP. Computer analysis of mammography phantom images (CAMPI). *Proc SPIE Medical Imaging 1997: Physics of Medical Imaging.* 1997;3032:292-299.
40. Chakraborty DP. Computer analysis of mammography phantom images (CAMPI): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics.* 1997;24(8):1269-1277.
41. Siddiqui KM, Johnson JP, Reiner BI, Siegel EL. Discrete cosine transform JPEG compression vs. 2D JPEG2000 compression: JNDmetrix visual discrimination model image quality analysis. Paper presented at: Medical Imaging2005.
42. Chakraborty DP. An alternate method for using a visual discrimination model (VDM) to optimize softcopy display image quality. *Journal of the Society for Information Display.* 2006;14(10):921-926.
43. Wolfe JM. Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review.* 1994;1(2):202-238.
44. Wolfe JM. Visual Search. In: Pashler H, ed. *Attention*. London, UK: University College London Press; 1998.
45. Wolfe JM, Horowitz TS, Kenner NM. Rare items often missed in visual searches. *Nature.* 2005;435(26):439.