The RJafroc Froc Book

Dev P. Chakraborty, PhD

2021-12-10

Contents

Preface

- This book is currently (as of November 2021) in preparation.
- It is intended as an online update to my "physical" book (Chakraborty, 2017). Since its publication in 2017 the `RJafroc` package, on which the `R` code examples in the book depend, has evolved considerably, causing many of the examples to "break". This also gives me the opportunity to improve on the book and include additional material.
- The physical book chapters are referred to as *book-chapters*, to distinguish them from the chapters in this online book.

TBA How much finished

10%

The pdf file of the book

Go here and then click on `Download` to get the `RJafrocFrocBook.pdf` file.

The html version of the book

Go here to view the `html` version of the book.

A note on the online distribution mechanism of the book

- In the hard-copy version of my book (Chakraborty, 2017) the online distribution mechanism was `BitBucket`.

- `BitBucket` allows code sharing within a *closed* group of a few users (e.g., myself and a grad student).
- Since the purpose of open-source code is to encourage collaborations, this was, in hindsight, an unfortunate choice. Moreover, as my experience with R-packages grew, it became apparent that the vast majority of R-packages are shared on `GitHub`, not `BitBucket`.
- For these reasons I have switched to `GitHub`. All previous instructions pertaining to `BitBucket` are obsolete.
- In order to access `GitHub` material one needs to create a (free) `GitHub` account.
- Go to this link and click on `Sign Up`.

Structure of the book

The book is divided into parts as follows:

- Part I: Quick Start: intended for existing Windows `JAFROC` users who are seeking a quick-and-easy transition from Windows `JAFROC` to `RJafroc`.
- Part II: ROC paradigm: this covers the basics of the ROC paradigm
- Part III: Significance Testing: The general procedure used to determine the significance level, and associated statistics, of the observed difference in figure of merit between pairs of treatments or readers
- Part IV: FROC paradigm: TBA

Contributing to this book

I appreciate constructive feedback on this document. To do this raise an `Issue` on the `GitHub` interface. Click on the `Issues` tab under `dpc10ster/RJafrocFrocBook`, then click on `New issue`. When done this way, contributions from users automatically become part of the `GitHub` documentation/history of the book.

Is this book relevant to you and what are the alternatives?

- Diagnostic imaging system evaluation
- Detection
- Detection combined with localization
- Detection combined with localization and classification
- Optimization of Artificial Intelligence (AI) algorithms

- CV
- Alternatives

ToDos TBA

- Check Bamber theorem derivation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.
- Change third person to first person in references to myself.

Chapters needing heavy edits

- 12-froc.
- 13-froc-empirical.
- 13-froc-empirical-examples.

Shelved vs. removed vs. parked folders needing heavy edits

- replace functions with ; eg. erf and exp in all of document
- Also for TPF, FPF etc.
- Temporarily shelved 17c-rsm-evidence.Rmd in removed folder
- Now 17-b is breaking; possibly related to changes in RJafroc: had to do with recent changes to RJafroc code - RSM_xFROC etc requiring intrinsic parameters; fixed 17-b
- parked has dependence of ROC/FROC performance on threshold

Coding aids

- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:
- kbl(dfA, caption = "...", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")

- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

FROC paradigm

Chapter 1

The FROC paradigm

1.1 TBA How much finished

70%

1.2 Introduction

Until now the focus has been on the receiver operating characteristic (ROC) paradigm. For diagnostic tasks such as detecting diffuse interstitial lung disease[1], or diseases similar to it, where *disease location is implicit*, this is an appropriate paradigm in that essential information is not being lost by limiting the radiologist's response to a single rating categorizing the likelihood of presence 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, type, size, extent) of the disease. In most clinical tasks if the radiologist believes the patient is diseased there is a location (or locations) associated with the suspected disease. Physicians term this *focal disease*, i.e., disease located at specific region(s) of the image.

[1] 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. By definition, diffuse interstitial lung disease is spread through, and confined to, lung tissues.

For focal disease the ROC paradigm constrains 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. Clinicians have long recognized problems with ignoring location (Black and Dwyer, 1990; Black, 2000). From the observer performance measurement point of view the most important consideration is that neglect of location information leads to loss of statistical power. The basic reason for this is that additional noise is introduced in the measurement due to crediting the reader for correctly detecting the diseased condition but pointing to the wrong location - i.e., *being right for the wrong reason.* One can compensate for reduced statistical power by increasing the numbers of readers and cases, which increases the cost of the study and is also unethical because, by not using the optimal paradigm and analysis, one is subjecting more patients to imaging procedures (Halpern et al., 2002).

1.2.1 Chapter outline

Four observer performance paradigms are compared as to the kinds of information collected and ignored. An essential characteristic of the FROC paradigm, namely *visual search*, is introduced. 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. A "solar" analogy is introuduced – understanding this is key to obtaining a good intuitive feel for this paradigm.

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

1.3 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) paradigms: usage of these terms is discouraged. For example, 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.

The term *lesion* always refers to 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. *A lesion is a real while a suspicious region is perceived.*

There are three location-specific paradigms:

- the free-response ROC (FROC) (Bunch et al., 1977; Chakraborty, 1989);
- the location ROC (LROC) (Starr et al., 1977; Swensson, 1996);
- the region of interest (ROI) (Obuchowski et al., 2000).

Fig. 1.1 shows a schematic mammogram interpreted according to current observer performance paradigms. The arrows point to 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 rating that the image contains at least one lesion, somewhere in the image.
- 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 the the image contains at least one lesion and marks the most suspicious region.
- ROI (bottom-right): the image is divided – by the researcher – into a number of regions-of-interest (ROIs) and the radiologist rates each ROI for presence of at least one lesion somewhere within the ROI.

The numbers and locations of suspicious regions depend on the image and the radiologists' expertise. In general the numbers of missed lesions and incorrect localizations increase as lesion contrast and/or reader expertise decreases.

In Fig. 1.1, 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).

- In the ROC paradigm, Fig. 1.1 (top-left), the radiologist assigns a single rating indicating the confidence level that there is at least one lesion somewhere in the image. 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).
- In the free-response (FROC) paradigm, Fig. 1.1 (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 now 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

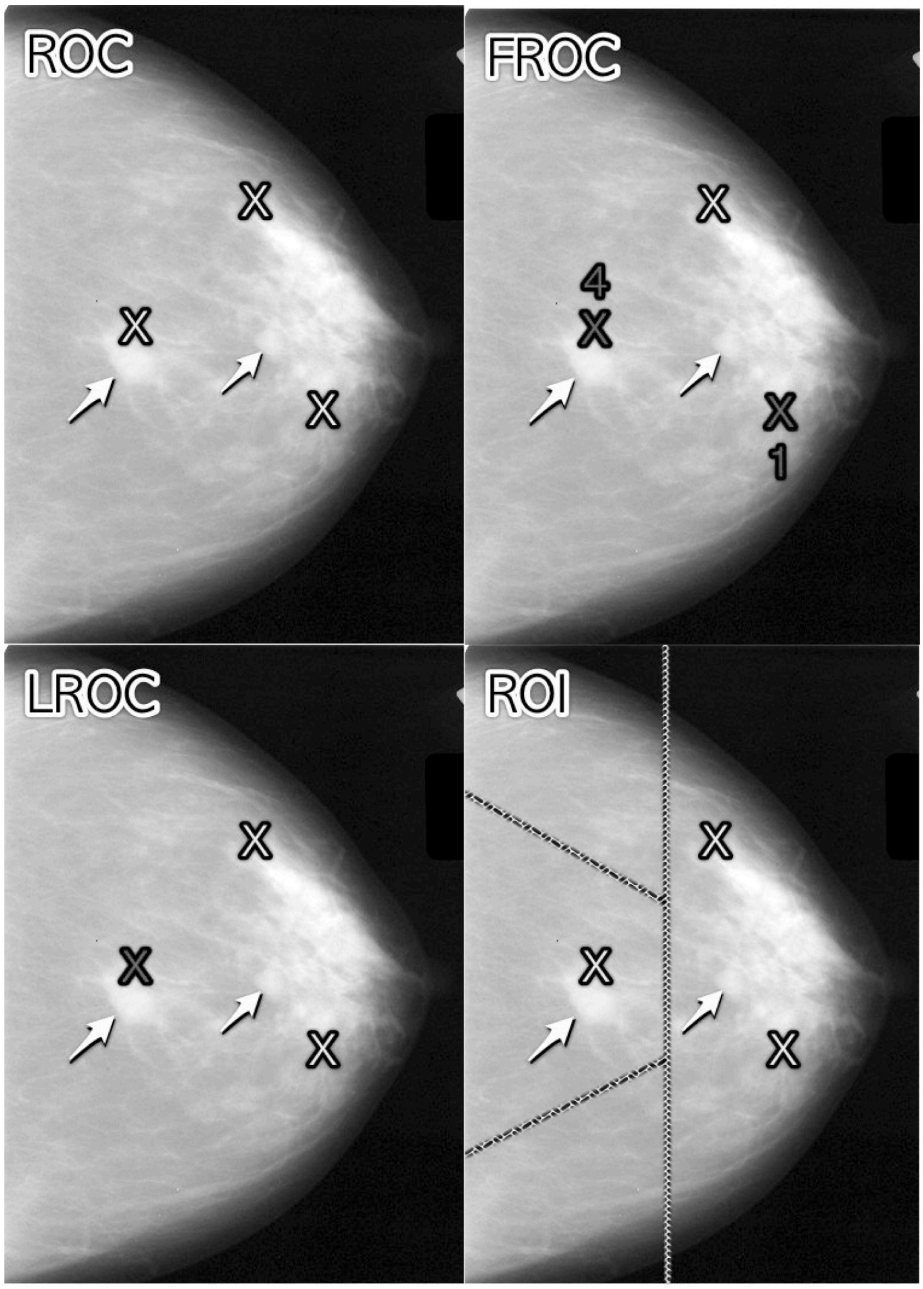


Figure 1.1: Upper Left: ROC, Upper Right: FROC, Lower Left: LROC, Lower Right: ROI

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.

- In the LROC paradigm, Fig. 1.1 (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 regions to report. The late Prof. Swensson has been the prime contributor to this paradigm.

- 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. 1.1 (bottom-right) there are four ROIs. 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. In the example shown in Fig. 1.1 (bottom-right), each case yields 4 ratings. Prof. Obuchowski has been the principal contributor to this paradigm. [2]

The rest of this book part focuses on the FROC paradigm.

1.4 Visual search

The FROC paradigm in medical imaging is equivalent to a visual search task. Any search task has two components: (i) finding something and (ii) acting on

[2]When different views of the same patient anatomy (perhaps in different 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). 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.

it. Examples of a search tasks are 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 expertise associated with any search task. Husbands are notoriously bad at finding the milk carton in the refrigerator (analogy due to Dr. Elizabeth Krupinski at an SPIE course taught jointly with the author). Like anything else, search expertise is honed by experience, i.e., lots of practice.

Likewise, a medical imaging search task has two components (i) finding suspicious regions and (ii) acting on each finding ("finding" 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 may be 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[3]. Consequently the radiologist must search each image 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.

1.4.1 Proximity criterion and scoring the data

In the first two clinical applications of the FROC paradigm (Chakraborty et al., 1986; Niklason et al., 1986) 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[4], was given only if a mark was sufficiently close (as per proximity criterion, see below) to an actual diseased region; otherwise, the observer's mark-rating pair was scored as a non-lesion localization or NL-event.

The use of ROC terminology, such as true positives or false positives to describe FROC data is not conducive to clarity, and is strongly discouraged.

Definitions:

- 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

[3]The probability of benign suspicious regions is much higher (Ernster, 1981), about 13% for women aged 40-45.

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

application. 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 (i.e., not one used in the observer study) how to score ambiguous marks. A rigid definition of the proximity criterion should not be used.

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 (Dobbins III et al., 2016) using FROC analysis gives more details on appropriate proximity criteria in the clinical context. [5]

1.4.2 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. [6] However, algorithmic readers, such as computer aided detection (CAD) algorithms, tend to find multiple regions in the same area. Algorithm designers generally incorporate a clustering step 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. [7]

1.4.3 Historical context

The term "free-response" was coined by (Egan et al., 1961) to describe a task involving the detection of brief audio tone(s) against a background of white-noise

[5] 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.

[6] 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 mark individual specks with great precision.

[7] 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.

(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 two-dimensional images replacing the listening intervals, the acoustic signal detection study is similar to medical imaging search tasks, e.g., screening mammography. At my former institution (University of Pittsburgh) 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. 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 (David Gur, private communication, ca. 2015). Conceptually, a mammography 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 (the BIRADS rating is actually assigned after the diagnostic workup following a 0-screening rating; the screening rating itself is binary: 0 for recall or 1 for normal).

1.5 The free-response receiver operating characteristic (FROC) plot

The free-response receiver operating characteristic (FROC) plot was introduced (Miller, 1969) as a way of visualizing performance in the free-response auditory tone detection task.

In the medical imaging context, assuming the mark rating pairs have been classified as NLs (non-lesion localizations) or LLs (lesion localizations):

- Non-lesion localization fraction (NLF) is defined as the total number of NLs rated 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 rated at or above the same threshold rating divided by the total number of lesions.

- The FROC plot is defined as that of LLF (ordinate) vs. NLF as the threshold is varied.
- The upper-right most operating point is termed the *observed end-point* and its coordinated are denoted $(\mathrm{NLF}_{\max}, \mathrm{LLF}_{\max})$.
- Unlike the ROC plot which is completely contained in the unit square, the FROC plot is not.

If integer ratings are used for each recorded mark, then (for example) in a four-rating FROC study, 4 FROC operating points will result: one corresponding to marks rated 4s; another corresponding to marks rated 4s or 3s; another to the 4s, 3s, or 2s; and finally the 4s, 3s, 2s, or 1s. An R-rating study yields at most R operating points.

If continuous ratings are used, the procedure is to start with a very high threshold so that none of the ratings exceed the threshold and then to 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. 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.

1.6 The "solar" analogy

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, assume the observer knows his 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., 1961) sense, 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, $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$.

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. Consequently, even in this example $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$.

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 $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$. Why are LLF and NLF subscripted "max"? By randomly choosing to not mark 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 $NLF = 0$. The reason for allowing the observer to "walk down" the vertical is simply to demonstrate that a continuous FROC curve from the origin to (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 $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$. 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 a fundamentally different mechanism from that considered in conventional observer performance methodology, namely *search*, is at work. Search describes the process of *finding* the lesion while *not finding* non-lesions.

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 magnification "spotter scope".

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 $\mathrm{LLF}_{\max} = 0$ and $\mathrm{NLF}_{\max} = 0$, 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.

The analogy is not restricted to the sun, which one might argue is an almost infinite SNR object and therefore atypical. 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 , that is much further than even the newly demoted Pluto.

1.7 Discussion and suggestions

This chapter has introduced the FROC paradigm, the terminology used to describe it and a common operating characteristic associated with it, namely the FROC. There are several areas of possible confusion to avoid which consider the following suggestions:

- Avoid using the term "lesion-specific" to describe location-specific paradigms.
- Avoid using the term "lesion" when one means a "suspicious region" that may or 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.
- 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 clinically important. Interactions with clinicians will allow selection of an appropriate proximity criterion for the task at hand and the multiple mark problem only occurs with algorithmic observers and is readily fixed.

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

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

1.8 References

Chapter 2

Visual Search

2.1 TBA How much finished

10%

2.2 Introduction

To understand free-response data, specifically how radiologists interpret images, one must come to grips with visual search. Casual usage of everyday terms like "search", "recognition" and "detection" in specific scientific contexts can lead to confusion. *Visual search is defined in a broad sense as grouping and labeling parts of an image.*

A schema of how radiologists find perform the search task, termed the Kundel-Nodine search model is described. The importance of this major conceptual model is not widely appreciated by researchers. It is the basis of the radiological search model (RSM) described in a later chapter TBA.

The following sections draw heavily on work by Nodine and Kundel (Nodine and Kundel, 1987; Kundel et al., 2007; Kundel and Nodine, 2004, 1983; Kundel et al., 1978). The author acknowledges critical insights gained through conversations with Dr. Claudia Mello-Thoms.

2.3 Grouping and labeling ROIs

Looking at and understanding an image involves grouping and assigning labels to different regions of interest (ROIs) in the image, where the labels correspond to entities that exist (or have existed in the examples to follow) in the real world.

As an example, if one looks at Fig. 2.1, one would label them (from left to right and top to bottom, in raster fashion): Franklin Roosevelt, Harry Truman, Lyndon Johnson, Richard Nixon, Jimmy Carter, Ronald Reagan, George H. W. Bush, and the presidential seal. The accuracy of the labeling depends on prior-knowledge, i.e., expertise, of the observer. If one were ignorant about US presidents one would be unable to correctly label them.

Figure 2.1: This image consists of 8 sub-images or ROIs. Understanding an image involves grouping and assigning labels to different ROIs, where the labels correspond to entities that exist in the real world. One familiar with US history would label them, from left to right and top to bottom, in raster fashion, Franklin Roosevelt, Harry Truman, Lyndon Johnson, Richard Nixon, Jimmy Carter, Ronald Reagan, George H. Bush and the presidential seal. Labeling accuracy depends on expertise of the observer. The row and column index of each ROI identifies its location.

Image interpretation in radiology is not fundamentally different. It involves assigning labels to an image by grouping and recognizing areas of the image that have correspondences to the radiologist's knowledge of the underlying anatomy, and, most importantly, deviations from the underlying anatomy. Most doctors, who need not be radiologists, can look at a chest x-ray and say, "this is the heart", "this is a rib", "this is a clavicle", "this is the aortic arch", etc., the top image in Fig. 2.2. This is because they know the underlying anatomy, the bottom image in Fig. 2.2 and have a basic understanding of the x-ray image formation physics that relates the anatomy to the image.

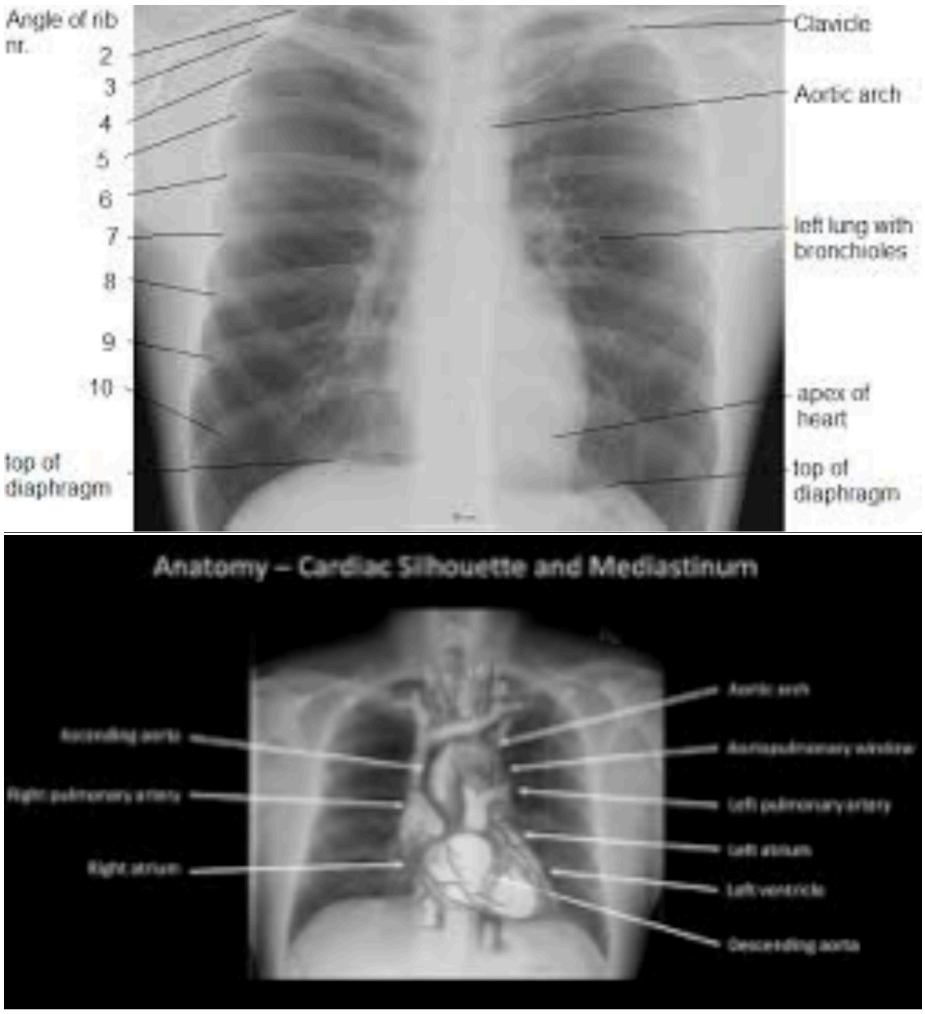


Figure 2.2: Image interpretation in radiology involves assigning labels to an image by grouping and recognizing areas of the image that have correspondences to the radiologist's knowledge of the underlying anatomy. (TOP) Most doctors can look at a chest x-ray and say, "this is the heart", "this is a rib", "this is the clavicle", "this is the aortic arch", etc. (BOTTOM) This is because they know the underlying anatomy and have a basic understanding of x-ray image formation physics that relates anatomy to the image.

2.4 Recognition vs. detection

The process of grouping and labeling parts of an image is termed recognition. This was illustrated with the pictures of the US presidents, Fig. 2.1. Recognition is distinct from detection, which is deciding about the presence of something that is unexpected or the absence of something that is expected, in other words, a deviation, in either direction, from what is expected. An example of detecting the presence of something that is unexpected would be a lung nodule and an example of detecting the absence of something that is expected would be an image of a patient with a missing rib (yes, it does occur, even excluding the biblical Adam).

The terms "expected" and "unexpected" are important: they imply expertise dependent expectations regarding the true structure of the non-diseased image, which I term a non-diseased template, and therefore an ability to recognize clinically relevant deviations or perturbations, in either direction, from this template; e.g., a lung nodule that could be cancer. By "clinically relevant" I mean perturbations related to the patient's health outcome – recognizing scratches, dead pixels, artifacts of know origin, and lead patient ID markers, do not count. There is a location associated with recognition, but not with detection. Detection is the presence or absence of something, i.e., the perturbation, which could be anywhere. For example, in Fig. 2.1, recognizing a face is equivalent to assigning a row and column index in the image. Specifically, recognizing of George H.W. Bush implies pointing to row = two and column = three. Detecting George H.W. Bush implies stating that George H.W. Bush is present in the image, but the location could be in any of the eight locations. Recognition is an FROC paradigm task, while detection is an ROC paradigm task. Instead of recognition, I prefer the more clinical term "finding", as in "finding" a lesion.

2.5 TBA Search vs. classification

Since template perturbations can occur at different locations in the images, the ability to selectively recognize them is related to search expertise. The term "selectively" is important: a non-expert can trivially recognize all perturbations by claiming all regions in the image are perturbed. Search expertise is the selective ability to find clinically relevant perturbations that are actually present while minimizing finding what appear to be clinically relevant perturbations but which are actually not present. In FROC terminology, search expertise is the ability to find latent LLs while minimizing the numbers of found latent NLs. Lesion-classification expertise is the ability to correctly classify a found suspicious region as malignant or benign.

The skills required to recognize a nodule in a chest x-ray are different from that required to recognize a low-contrast circular or Gaussian shaped artificial nodule against a background of random noise. In the former instance the skills

of the radiologist are relevant: e.g., the skilled radiologist knows not to confuse a blood vessel viewed "end on" for a nodule, especially since the radiologist knows where to expect these vessels, e.g., the aorta. In the latter instance, (i.e., viewing artificial nodules superposed on random noise) there are no expected anatomic structures, so the skills possessed by the radiologist are nullified. This is the reason why having radiologists interpret random noise images and pretending that this somehow makes it "clinically relevant" is a waste of reader resources and represent bad science. One might as well used undergraduates with good eyesight, motivation and training. To quote (Nodine and Kundel, 1987)

> Detecting an object that is hidden in a natural scene is not the same as detecting an object displayed against a background of random noise.

This paragraph also argues against usage of phantoms as stand-ins for clinical images for "clinical" performance assessment. Phantoms are fine in the quality control context, but they do not allow radiologists the opportunity to exercise their professional skills.

2.6 The Kundel - Nodine search model

The Kundel-Nodine model (Kundel et al., 2007; Kundel and Nodine, 2004) is a schema of events that occur from the radiologist's first glance to the decision about the image.

Assuming the task has been defined prior to viewing, based on eye-tracking recordings obtained on radiologists while they interpreted images, Kundel and Nodine proposed the following schema for the diagnostic interpretation process, consisting of two major components: (1) glancing or global impression and (2) scanning or feature analysis:

2.6.1 Glancing / Global impression

The colloquial term "glancing" is meant literally. The glance is brief, typically lasting about 100 - 300 ms, too short for detailed foveal examination and interpretation. Instead, during this brief interval peripheral vision and reader expertise are the primary mechanisms responsible for the identification of the perturbations. The glance results in a global impression, or gestalt, that identifies perturbations from the template. Object recognition occurs at a holistic level, i.e., in the context of the whole image, as there is insufficient time for detailed viewing and all of this is going on using peripheral vision. It is remarkable that radiologists can make reasonably accurate interpretations from information obtained in a brief glance, see Fig. 6 in (Nodine and Kundel, 1987).

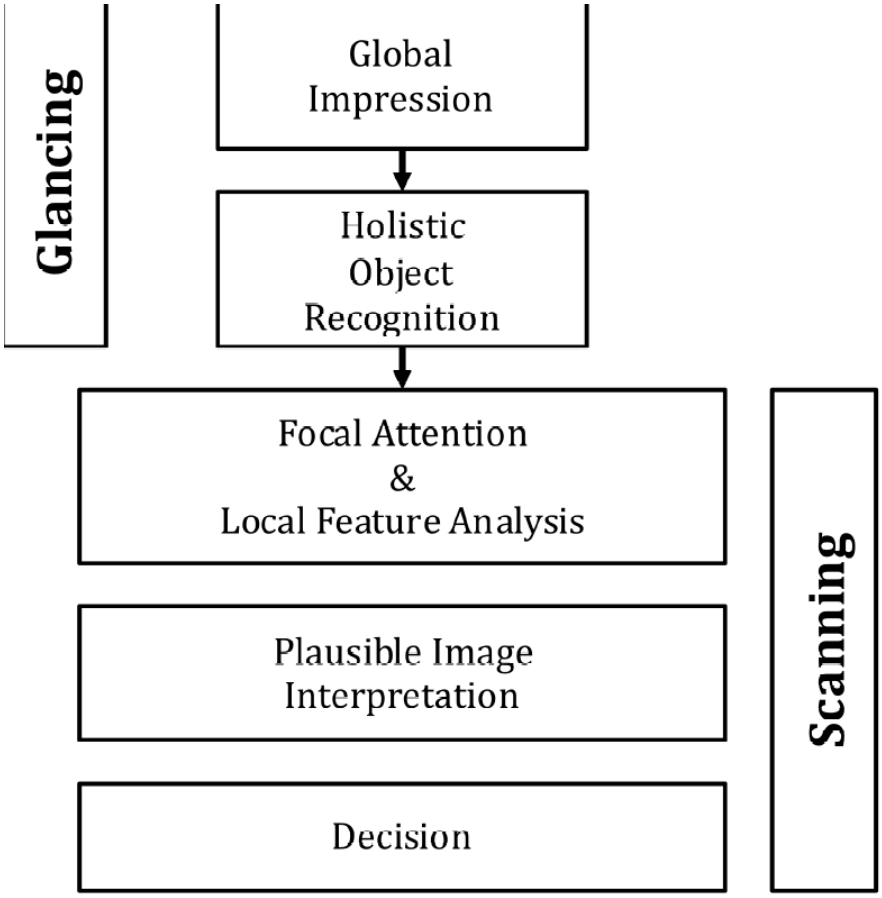


Figure 2.3: The Kundel-Nodine model of radiological search. The glancing/global stage identifies perturbations from the template of a generic non-diseased case. The scanning stage analyzes the each perturbation and calculates the probability it is a true lesion. Only perturbations with sufficiently high probability are marked.

Perturbations are flagged for subsequent detailed viewing, i.e., the initial glance tells the visual system where to look more closely.

2.6.2 Scanning / Local feature analysis

The global impression identifies perturbations for detailed foveal viewing by the central vision. During this process - termed scanning or feature analysis - the observer scrutinizes and analyzes the suspicious regions for evidence of possible disease. In principle, they calculate the probability of malignancy. For those readers of this book familiar with how CAD works, this corresponds to the feature analysis stage of CAD where regions found by the global search, termed *initial detections* in the CAD literature, are analyzed for probability of malignancy.

The essential point that emerges is that decisions are made at a finite, relatively small, number of regions. Attention units are not uniformly distributed through the image in raster-scan fashion; rather the global impression identifies a smaller set of regions that require detailed scanning.

Eye-tracker recordings for a two-view digital mammogram for two observers are shown in Fig. 2.4, for an inexperienced observer (upper two panels) and an expert mammographer (lower two panels). The small circles indicate individual fixations (dwell time ~ 100 ms). The larger high-contrast circles indicate clustered fixations (cumulative dwell time ~ 1 s). The larger low-contrast circles indicate a mass visible on both views. The inexperienced observer finds many more suspicious regions than does the expert mammographer but misses the lesion in the MLO view. In other words, the inexperienced observer generates many latent NLs but only one latent LL. The mammographer finds the lesion in the MLO view, which qualifies as a latent LL, without finding suspicious regions in the non-diseased parenchyma, i.e., the expert generated zero latent NLs on this case and one latent LL. It is possible the observer was so confident in the malignancy found in the MLO view that there was no need to fixate the visible lesion in the other view - the decision had already been made to recall the patient for further imaging.

Details: Eye-tracking recordings for a two-view digital mammogram display for two observers, an inexperienced observer (upper two panels) and an expert mammographer (lower two panels). The small circles indicate individual fixations (dwell time ~ 100 ms). The larger high-contrast circles indicate clustered fixations (cumulative dwell time ~ 1 sec). The latter correspond to the latent marks in the search-model. The larger low-contrast circles indicate a mass visible on both views. The inexperienced observer finds many more suspicious regions than does the expert mammographer but misses the lesion in the MLO view. In other words the inexperienced observer generates many latent NLs but only one latent LL. The mammographer finds the lesion in the MLO view, which qualifies as a latent LL, without finding suspicious regions in the non-diseased parenchyma, i.e., the expert generated zero latent NLs on this case

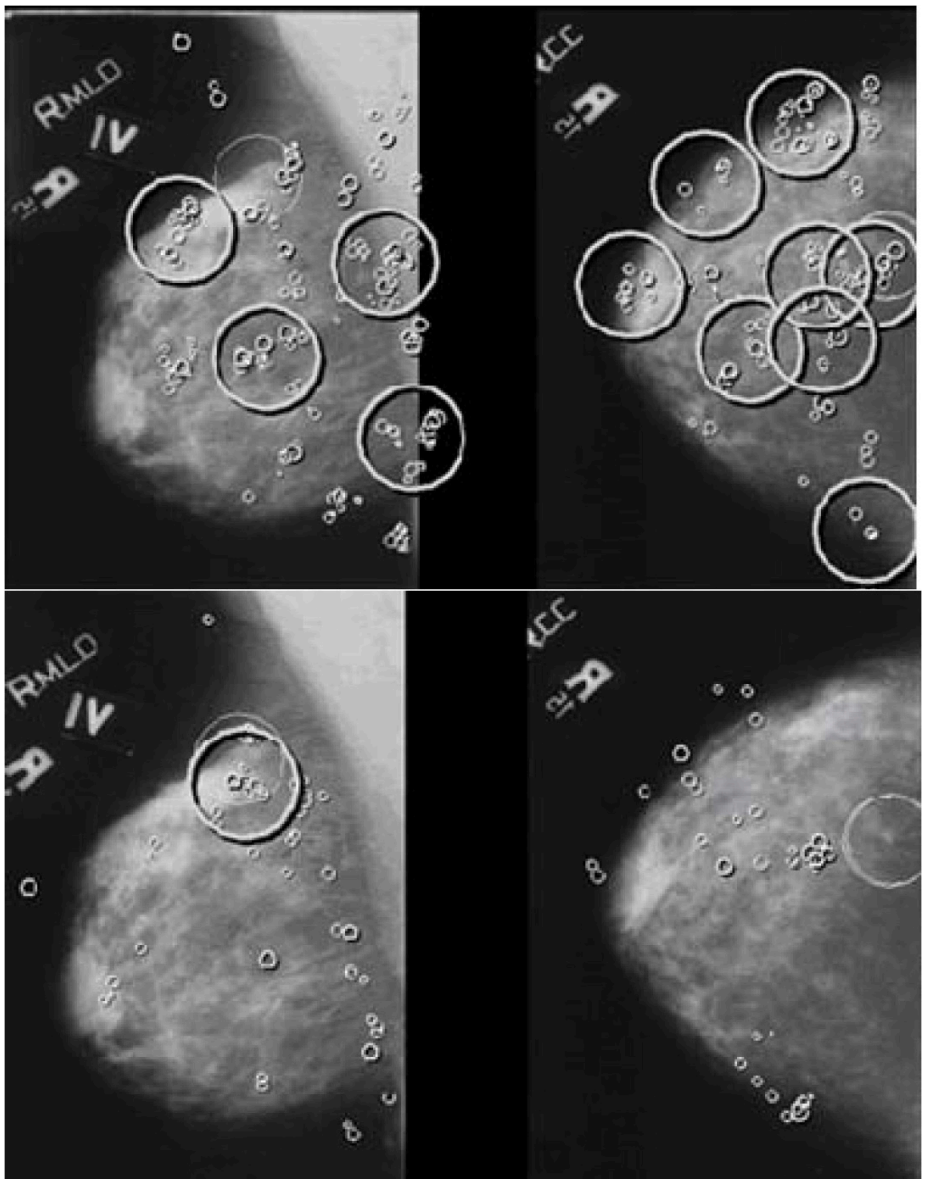


Figure 2.4: Eye-tracking recordings for a two-view digital mammogram: see details.

and one latent LL. It is possible the observer was so confident in the malignancy found in the MLO view that there was no need to fixate the visible lesion in the other view - the decision had already been made to recall the patient for further imaging, which confirmed the finding.

2.7 Kundel-Nodine model and CAD algorithms

It turns out that the designers of CAD algorithms independently arrived at a two-stage process remarkably similar to that described by Kundel-Nodine for radiologist observers. CAD algorithms are designed to emulate expert radiologists, and while this goal is not yet met, these algorithms are reasonable approximations to radiologists, and include the critical elements of search and localization that are central to clinical tasks. CAD algorithms involve two steps analogous to the holistic and cognitive stages of the Kundel-Nodine visual search model (Nodine and Kundel, 1987; Kundel and Nodine, 2004, 1983). In other words, CAD has a perceptual correspondence to human observers that to my knowledge is not shared by other method of predicting what radiologists will call on clinical images.

In the first stage of CAD, termed initial detections (Edwards et al., 2002), the algorithm finds "all reasonable" regions that could possibly be a malignancy. The term "all reasonable" is used because an irrational observer could trivially "find" every malignancy by marking all regions of the image. Most of these regions would be unreasonable to a rational observer, who would prefentially marks lesions while minimizing marking other regions. Therefore, the idea of CAD's initial detection stage is to find as many of the malignancies as possible while not finding too many non-diseased regions. This corresponds to the search stage of the Kundel-Nodine model. Unfortunately, CAD is rather poor at this task compared to expert radiologists. Progress in this area has been stymied by lack of understanding of search and how to measure performance in the FROC task. Indeed a widely held misconception is that CAD is perfect (!) at search, because it "looks at" everything (Dr. Ron Summers, NIH, private communication, Dublin, ca. 2010). In giving equal attention units to all parts of the image, CAD will trivially find all cancers, but it will also find a large number of NLs.

CAD researchers are, in my opinion, at the forefront of those presuming to understand how radiologists interpret cases. They work with real images and real lesions and the manufacturer's reputation is on the line, just like a radiologist's, and Medicare even reimburses CAD interpretations. While their current track record is not that good for breast masses compared to expert radiologists, with proper understanding of what is limiting CAD, namely the search process, there is no doubt in my opinion, that future generations CAD algorithms will approach and even surpass expert radiologists.

2.8 TBA Discussion / Summary

This chapter has introduced the terminology associated with a search task: recognition/finding, classification, and detection. Search involves finding lesions and correctly classifying them, so two types of expertise are relevant: search expertise is the ability to find (true) lesions without finding non-lesions, while classification accuracy is concerned with correct classification (benign vs. malignant) of a suspicious region that has already been found. Quantification of these abilities is described in the next chapter. Two paradigms are used to measure search, one in the non-medical context and the other, the focus of this book, in the medical context. The second method is based on the eye tracking measurements performed while radiologists perform quasi-clinical tasks (performing eye-tracking measurements in a true clinical setting is difficult). A method for analyzing eye-tracking data using methods developed for FROC analysis has been described. It has the advantage of taking into account information present in eye-tracking data, such as dwell time and approach rate, in a quantitative manner, essentially by treating them as eye-tracking ratings to which modern FROC methods can be applied. The Kundel-Nodine model of visual search in diagnostic imaging was described. The next chapter describes a statistical parameterization of this model, termed the radiological search model (RSM).

2.9 References

Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. RadioGraphics. 1987;7(2):1241-1250. 2. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. Radiology. 2007;242(2):396-402. 3. Kundel HL, Nodine CF. Modeling visual search during mammogram viewing. Proc SPIE. 2004;5372:110-115. 4. Kundel HL, Nodine CF. A visual concept shapes image perception. Radiology. 1983;146:363-368. 5. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol. 1978;13:175-181. 6. Horowitz TS, Wolfe JM. Visual search has no memory. Nature. 1998;394(6693):575-577. 7. Wolfe JM. Guided Search 2.0: A revised model of visual search. Psychonomic Bulletin & Review. 1994;1(2):202-238. 8. Wolfe JM, Cave KR, Franzel SL. Guided search: an alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human perception and performance. 1989;15(3):419. 9. Carmody DP, Kundel HL, Nodine CF. Performance of a computer system for recording eye fixations using limbus reflection. Behavior Research Methods & Instrumentation. 1980;12(1):63-66. 10. Duchowski AT. Eye Tracking Methodology: Theory and Practice. Clemson, SC: Clemson University; 2002. 11. Nodine C, Mello-Thoms C, Kundel H, Weinstein S. Time course of perception and decision making during mammographic interpretation. AJR. 2002;179:917-923. 12. Nodine CF, Kundel HL, Mello-

Thoms C, et al. How experience and training influence mammography expertise. Acad Radiol. 1999;6(10):575-585. 13. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic radiology. 2012;19(12):1474-1483. 14. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. Med Phys. 1995;22(5):643-655. 15. 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. 16. Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. Investigative Radiology. 1989;24:472-478. 17. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic Radiology. 2012;In press. 18. Wolfe JM. Visual Search. In: Pashler H, ed. Attention. London, UK: University College London Press; 1998. 19. Larson AM, Loschky LC. The contributions of central versus peripheral vision to scene gist recognition. Journal of Vision. 2009;9(10):6-6. 20. Pritchard RM, Heron W, Hebb DO. Visual perception approached by the method of stabilized images. Canadian Journal of Psychology/Revue canadienne de psychologie. 1960;14(2):67. 21. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys. 2002;29(12):2861-2870. 22. Kundel HL, Nodine CF, Krupinski EA, Mello-Thoms C. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. Academic Radiology. 2008;15(7):881-886. 23. Mello-Thoms C, Hardesty LA, Sumkin JH, et al. Effects of lesion conspicuity on visual search in mammogram reading. Acad Radiol. 2005;12:830-840.

Chapter 3

The radiological search model

3.1 TBA How much finished

70%

3.2 Introduction

Brief accounts of the radiological search model (RSM) were presented earlier in connection with the simulator used to generate FROC data. This chapter describes the model in more detail.

All models of ROC data *not incorporating search* involve two fundamental parameters (i.e, not including binning-related threshold parameters). For example, the unequal variance binormal model in Chapter TBA (binormal-model) requires the a, b parameters. Alternative ROC models described in TBA Chapter 20 also require two fundamental parameters.

It turns out that all that is needed to model as seemingly complex a process as visual search, at least to first order, is one additional fundamental parameter. The RSM contains three fundamental parameters: μ, λ and ν. However, it is easier to introduce the RSM via μ and intermediate primed parameters, λ' and ν'. The model is then re-parameterized to take into account that λ' and ν' must depend on μ via un-primed parameters λ and ν *which are intrinsic parameters*, i.e., independent of μ.

The RSM is a model of the FROC paradigm. It accounts for all features characterizing the FROC paradigm, including localization and the random non-negative numbers of NLs and LLs per image.

3.3 The radiological search model

The radiological search model (RSM) for the free-response paradigm is a statistical parameterization of the Nodine-Kundel model. It consists of:

- A *search stage* corresponding to the initial glance in the Nodine-Kundel model, in which suspicious regions, i.e., the latent marks, are flagged for subsequent foveal scanning. The total number of latent marks on a case is ≥ 0; some cases may have zero latent marks, a fact that will turn out to have important consequences for the shapes of all RSM predicted operating characteristics.
- A *decision stage* during which each latent mark is closely examined (via foveal scanning), relevant features are extracted and analyzed and the observer calculates a decision variable or z-sample for each latent mark. The number of z-samples equals the number of latent marks.
- If the z-sample exceeds a pre-selected minimum reporting threshold the location is marked, i.e., the latent mark is recorded as an actual mark.
- Latent marks can be either latent NLs (corresponding to non-diseased regions) or latent LLs (corresponding to diseased regions). The number of latent NLs on a case is denoted l_1. The number of latent LLs on a diseased case is denoted l_2. Latent NLs can occur on non-diseased and diseased cases, but latent LLs can only occur on diseased cases. We will initially assume that every diseased case has L actual lesions. Later this will be extended to arbitrary number of lesions per diseased case. Since the number of latent LLs cannot exceed the number of lesions, $0 \leq l_2 \leq L$. The symbol l_s denotes a location with site-level truth state s, where $s = 1$ for a NL and $s = 2$ for a LL. [1]

3.4 RSM assumptions

Assumption 1: The number of latent NLs, $l_1 \geq 0$, is sampled from the Poisson distribution Poi with mean λ':

$$l_1 \sim \text{Poi}\,(\lambda') \tag{3.1}$$

The probability mass function (pmf) of the Poisson distribution is defined by:

[1] In this chapter distributional assumptions are made for the numbers of latent NLs and LLs and the associated z-samples. Since the RSM is a parametric model one does not need the four subscript notation needed to account for case and location dependence, as necessary in the empirical description in Chapter 5. This allows for a simpler notation, as the reader may have noticed, unencumbered by the 4 subscripts in Table 5.4.

$$\text{pmf}_{Poi}\left(l_1, \lambda'\right) = exp\left(-\lambda'\right)\frac{\left(\lambda'\right)^{l_1}}{\left(l_1'\right)!} \tag{3.2}$$

Assumption 2: The number of latent LLs, l_2, where $0 \leq l_2 \leq L$, is sampled from the binomial distribution Bin with success probability ν' and trial size L:

$$l_2 \sim \text{Bin}\left(L, \nu'\right) \tag{3.3}$$

The pmf of the binomial distribution is defined by:

$$\text{pmf}_{Bin}\left(l_2, L, \nu'\right) = \binom{L}{l_2}\left(\nu'\right)^{l_2}\left(1-\nu'\right)^{L-l_2} \tag{3.4}$$

Assumption 3: Each latent mark is associated with a z-sample. That for a latent NL is denoted z_{l_1} while that for a latent LL is denoted z_{l_2}. Latent NLs can occur on non-diseased and diseased cases while latent LLs can only occur on diseased cases.

Assumption 4: For latent NLs the z-samples are obtained by sampling $N\left(0,1\right)$:

$$z_{l_1} \sim N\left(0,1\right) \tag{3.5}$$

Assumption 5: For latent LLs the z-samples are obtained by sampling $N\left(\mu,1\right)$:

$$z_{l_2} \sim N\left(\mu,1\right) \tag{3.6}$$

The probability density function $\phi\left(z|\mu\right)$ of the normal distribution $N\left(\mu,1\right)$ is defined by:

$$\phi\left(z|\mu\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(z-\mu\right)^2}{2}\right) \tag{3.7}$$

Binning rule: In an FROC study with R ratings, the observer adopts R ordered cutoffs ζ_r, where $(r = 1, 2, ..., R)$. Defining $\zeta_0 = -\infty$ and $\zeta_{R+1} = \infty$, then if $\zeta_r \leq z_{l_s} < \zeta_{r+1}$ the corresponding latent site is marked and rated in bin r, and if $z_{l_s} \leq \zeta_1$ the site is not marked.

Mark location: The location of the mark is assumed to be at the exact center of the latent site that exceeded a cutoff and an infinitely precise proximity criterion is adopted. Consequently, there is no confusing a mark made because of a latent LL z-sample exceeding the cutoff with one made because of a latent NL z-sample exceeding the cutoff. Therefore, any mark made because of a latent NL z-sample that satisfies $\zeta_r \leq z_{l_1} < \zeta_{r+1}$ will be scored as a non-lesion localization (NL) and rated r. Likewise, any mark made because of a latent

LL z-sample that satisfies $\zeta_r \leq z_{l_2} < \zeta_{r+1}$ will be scored as a lesion-localization (LL) and rated r.

Rating assigned to unmarked sites: Unmarked LLs are assigned the zero rating: even lesions that were not flagged by the search stage, and therefore do not qualify as latent LLs, are assigned the zero rating. This is because they represent observable events. In contrast, unmarked latent NLs are unobservable events (unlike lesions, there is no a-priori reader-independent list of non-lesion locations; in fact, what constitutes a NL is reader dependent).

By choosing R large enough, the preceding discrete rating model is applicable to continuous z-samples.

3.5 Physical interpretation of RSM parameters

The parameters μ, λ' and ν' have the following meanings:

3.5.1 The μ parameter

The μ parameter is the lesion contrast-to-noise-ratio, or more accurately, the perceptual signal to noise ratio *pSNR* introduced in (book) Chapter 12, between latent NLs and latent LLs. It is not the pSNR of the latent LL relative to its immediate surround. For structured backgrounds - as opposed to homogeneous backgrounds - pSNR is determined by the competition for the observer's foveal attention from other regions, outside the immediate surround, that could be mistaken for lesions.

The μ parameter is similar to detectability index d', which is the separation parameter of two unit normal distributions required to achieve the observed probability of correct choice (PC) in a two alternative forced choice (2AFC) task between cued (i.e., pointed to by toggle-able arrows) NLs and cued LLs. Individually and for each reader one determines the locations of the latent marks using eye-tracking apparatus and then runs a 2AFC study as follows: pairs of images are shown, each with a cued location, one a latent NL and the other a latent LL, where all locations were recorded in prior eye-tracking sessions for the specific radiologist. The radiologist's task is to pick the image with the latent LL. The probability correct PC in this task is related to the d' parameter by:

$$\mu = \sqrt{2}\Phi^{-1}(\text{PC}) \tag{3.8}$$

The radiologist on whom the eye-tracking measurements are performed and the one who performs the two alternative forced choice tasks must be the same, as two radiologists may not agree on latent NL marks. A complication in conducting such a study is that because of memory effects a lesion can only be shown

once to each reader: clinical images are distinctive - once a radiologist has found a lesion in a clinical image, that event becomes imprinted in long-term memory; one cannot repeatedly compare this lesion to other NLs in the 2AFC task as the radiologist will always pick the remembered lesion.

3.5.2 The λ' parameter

The λ' parameter determines the tendency of the observer to generate latent NLs. The mean number of latent NLs per case is an estimate of λ'. [2]

Consider two observers, one with $\lambda' = 1$ and the other with $\lambda' = 2$. While one cannot predict the exact number of latent NLs on any specific case, one can predict the average number of latent NLs on a given case set.

In the following examples the number of samples has been set to $K_1 = 100$ (the first argument to `rpois()`; the second argument is λ').

3.5.2.1 Example 1

```
seed <- 1;set.seed(seed)
K1 <- 100
samples1 <- rpois(K1,1)
```

```
## mean(samples1) =  1.01
```

```
## samples1[1:10] =  0 1 1 2 0 2 3 1 1 0
```

For this observer, $\lambda' = 1$, the first case generated zero latent NLs, the 2nd and 3rd cases generated one NL each, the third case generated 2 NLs, etc.

3.5.2.2 Example 2

```
seed <- 1;set.seed(seed)
samples2 <- rpois(K1,2)
```

```
## mean(samples2) =  2.02
```

[2] It can be measured via eye-tracking apparatus. This time it is only necessary to cluster the marks and classify each mark as a latent NL or latent LL according to the adopted acceptance radius. An eye-tracking based estimate would be the total number of latent NLs in the dataset divided by the total number of cases.

```
## samples2[1:10] =  1 1 2 4 1 4 4 2 2 0
```

For the second observer $\lambda' = 2$, the first and second case generated one latent NL each, the third generated two, etc. The average number of latent NL marks per case for the 1^{st} observer is 1.01 and that for the 2^{nd} one is 2.02.

3.5.2.3 Confidence intervals

The following code illustrates Poisson sampling and estimation of an exact confidence interval for the mean for 100 samples from two Poisson distributions.

```
K1 <- 100
lambdaP <- c(1,2)
seed <- 1;set.seed(seed);samples1 <- rpois(K1,lambda = lambdaP[1])
seed <- 1;set.seed(seed);samples2 <- rpois(K1,lambda = lambdaP[2])

ret11 <- poisson.exact(sum(samples1),K1)
ret21 <- poisson.exact(sum(samples2),K1)
```

```
## K1 =  100 , lambdaP 1st reader =  1 , lambdaP 2nd reader =  2
```

```
## obs. mean, reader 1 =  1.01
```

```
## obs. mean, reader 2 =  2.02
```

```
## Rdr. 1: 95% CI =  0.8226616 1.227242
```

```
## Rdr. 2: 95% CI =  1.751026 2.318599
```

For reader 1 the estimate of the Poisson parameter (the mean parameter of the Poisson distribution is frequently referred to as the Poisson parameter) is 1.01 with 95% confidence interval (0.823, 1.227); for reader 2 the corresponding estimates are 2.02 and (1.751, 2.319). As the number of cases increases, the confidence interval shrinks. For example, with 10000 cases, i.e., 100 times the value in the previous example:

```
## K1 =  10000 , lambdaP 1st reader =  1 , lambdaP 2nd reader =  2
```

```
## obs. mean, reader 1 =  1.0055
```

```
## obs. mean, reader 2 =  2.006
```

```
## Rdr. 1: 95% CI =  0.9859414 1.025349

## Rdr. 2: 95% CI =  1.978335 2.033955
```

This time for reader 1, the estimate of the Poisson parameter is 1.01 with 95% confidence interval (0.986, 1.025); for reader 2 the corresponding estimate is 2.01 with 95% confidence interval (1.978, 2.034). The width of the confidence interval is inversely proportional to the square root of the number of cases (the example below is for reader 1):

```
ret11$conf.int[2] - ret11$conf.int[1]
```

```
## [1] 0.40458
```

```
ret12$conf.int[2] - ret12$conf.int[1]
```

```
## [1] 0.03940756
```

Since the number of cases was increased by a factor of 100, the width decreased by a factor of 10, the square-root of the ratio of the numbers of cases.

3.5.3 The ν' parameter

The ν' parameter determines the ability of the observer to find lesions. Assuming the same number of lesions per diseased case, the mean fraction of latent LLs per diseased case is an estimate of ν'. [3] Consider two observers, one with $\nu' = 0.5$ and the other with $\nu' = 0.9$. Again, while one cannot predict the precise number of latent LLs on any specific diseased case, or which specific lesions will be correctly localized, one can predict the average number of latent LLs per diseased case.

The following code also uses $K_2 = 100$ samples, the number of diseased cases, each with one lesion.

```
K2 <- 100
L <- 1
nuP1 <- 0.5;nuP2 <- 0.9;
seed <- 1;set.seed(seed);samples1 <- rbinom(K2,L,nuP1)
seed <- 1;set.seed(seed);samples2 <- rbinom(K2,L,nuP2)

ret1 <- binom.exact(sum(samples1),K2*L)
ret2 <- binom.exact(sum(samples2),K2*L)
```

[3] It too can be measured via eye-tracking apparatus performed on a radiologist. An eye-tracking based estimate would be the total number of latent LLs in the dataset divided by the total number of lesions.

```
## K2 =  100 , nuP 1st reader =  0.5 , nuP 2nd reader =  0.9

## mean, reader 1 =  0.48

## mean, reader 2 =  0.94

## Rdr. 1: 95% CI =  0.3790055 0.5822102

## Rdr. 2: 95% CI =  0.8739701 0.9776651
```

The result shows that for reader 1 the estimate of the binomial success rate parameter is 0.48 with 95% confidence interval (0.38, 0.58). For reader 2 the corresponding estimates are 0.94 and (0.87, 0.98). As the number of diseased cases increases, the confidence interval shrinks in inverse proportion to the square root of cases.

As a more complicated but clinically realistic example, consider a dataset with 100 cases in all where 97 have one lesion per case, two have two lesions per case and one has three lesions per case (these are typical lesion distributions observed in screening mammography). The code follows:

```
K2 <- c(97,2,1);Lk <- c(1,2,3);nuP1 <- 0.5;nuP2 <- 0.9;
samples1 <- array(dim = c(sum(K2),length(K2)))
seed <- 1;set.seed(seed)
for (l in 1:length(K2)) {
  samples1[1:K2[l],l] <- rbinom(K2[l],Lk[l],nuP1)
}

samples2 <- array(dim = c(sum(K2),length(K2)))
seed <- 1;set.seed(seed)
for (l in 1:length(K2)) {
  samples2[1:K2[l],l] <- rbinom(K2[l],Lk[l],nuP2)
}

ret1 <- binom.exact(sum(samples1[!is.na(samples1)]),sum(K2*Lk))
ret2 <- binom.exact(sum(samples2[!is.na(samples2)]),sum(K2*Lk))
```

```
## K2[1] = 97 , K2[2] = 2 , K2[3] = 1 , nuP1 = 0.5 , nuP2 = 0.9

## obsvd. mean, reader 1 =  0.4903846

## obsvd. mean, reader 2 =  0.9326923

## Rdr. 1: 95% CI =  0.3910217 0.5903092
```

```
## Rdr. 2: 95% CI =  0.8662286 0.9725125
```

For reader 1, the estimate of the binomial success probability is 0.490 with 95% confidence interval (0.391, 0.590); for reader 2 the corresponding estimates are 0.933 and (0.866, 0.973).

3.6 Model re-parameterization

While the parameters μ, λ' and ν' are physically meaningful a little thought reveals that they cannot be varied independently of each other. Rather, μ is the *intrinsic* parameter whose value, together with two other intrinsic parameters λ and ν, determine λ' and ν', respectively. The following is a convenient re-parameterization [4]:

$$\left.\begin{aligned} \nu' &= 1 - exp\left(-\mu\nu\right) \\ \lambda' &= \frac{\lambda}{\mu} \end{aligned}\right\} \tag{3.9}$$

The inverse transformations are:

$$\left.\begin{aligned} \nu &= -\frac{\ln\left(1-\nu'\right)}{\mu} \\ \lambda &= \mu\lambda' \end{aligned}\right\} \tag{3.10}$$

The parameter limits are as follows: $0 \leq \nu' \leq 1$, $\lambda' \geq 0$, $\mu \geq 0$, $\lambda \geq 0$ and $\nu \geq 0$.

Since it determines ν', the ν parameter can be considered as the intrinsic (i.e., μ-independent) ability to find lesions; specifically, it is the rate of increase of ν' with μ at small μ:

$$\nu = \left(\frac{\partial\nu'}{\partial\mu}\right)_{\mu=0} \tag{3.11}$$

The colloquial term *find* is used as shorthand for *flagged for further inspection by the holistic 1st stage of the search mechanism, thus qualifying as a latent site.* In other words, *finding* a lesion means the lesion was perceived as a suspicious region, which makes it a latent site, independent of whether or not the region was actually marked. Finding refers to the search stage. Marking refers to the

[4] The need for the first re-parameterization, involving ν', was foreseen in the original search model papers (Chakraborty, 2006b,a) but the need for the second re-parameterization (involving λ') was discovered more recently.

decision stage, where the region's z-sample is determined and compared to a marking threshold.

According to Eqn. (3.9), as $\mu \to \infty$, $\nu' \to 1$ and conversely, as $\mu \to 0$, $\nu' \to 0$. The dependence of ν' on μ is consistent with the fact that higher contrast lesions are easier to find. An observer without special expertise may find a high contrast lesion. Conversely, lower contrast lesions will be more difficult to find even by expert observers.

The analogy to finding the sun 1.6 is instructive: objects with very high perceptual SNR are certain to be found.

According to Eqn. (3.9) the value of μ also determines λ': as $\mu \to \infty$, $\lambda' \to 0$, and conversely, as $\mu \to 0$, $\lambda' \to \infty$. Here too the sun analogy is instructive. Since the sun has very high contrast, there is no reason for the observer to searcch for other suspicious regions which have no possibility of resembling the sun. On the other hand, attempting to locate a faint star, possibly hidden by clouds, can generate latent NLs, because the expected small SNR from the faint real star could be comparable to that from a number of regions in the near background.

The re-parameterization used here is not unique, but is simple and has the right limiting behaviors.

3.7 Discussion / Summary

This chapter has described a statistical parameterization of the Nodine-Kundel model. The 3-parameter model of search in the context in the medical imaging accommodates key aspects of the process: search, the ability to find lesions while minimizing finding non-lesions, is described by two parameters, specifically, λ' and ν' . The ability to correctly mark a found lesion (while not marking found non-lesions) is characterized by the third parameter of the model, μ. While the primed parameters have relatively simple physical meaning, they depend on μ. Consequently, it is necessary to define them in terms of intrinsic parameters.

The next chapter explores the predictions of the radiological search model.

3.8 References

1. 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.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. Medical Physics. 1995;22(2):133-143.

3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic Radiology. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys Med Biol. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol. 2006;51:3449–3462.

Chapter 4

ROC curve implications of the RSM

4.1 TBA How much finished

90%

4.2 TBA Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM. The starting point is a general characteristic of all RSM prediced operating characteristics, namely they have the constrained end-point property. Derived next is the predicted *inferred ROC* curve followed by the predicted FROC and AFROC curves.

Shown next is how *search performance* and *lesion-classification* performance can be measured from the inferred ROC curve. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it. Lesion-classification is different from (case) classification performance, i.e., distinguishing between diseased and non-diseased cases, which is measured by the area AUC under the ROC curve.

TBA Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, leading to an important and perhaps surprising conclusion, *the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC curves are preferred*. Most applications of FROC methods, particularly in CAD, have relied on the FROC curve to measure performance.

In this chapter formulae for RSM quantities are given in terms of the physical search parameters λ' and ν'. The formulae can be transformed to intrinsic RSM parameters λ and ν using Eqn. (3.10).

4.3 Inferred ROC ratings

Consider a $R_{\text{FROC}} \geq 1$ rating FROC study with allowed ratings $r = 1, 2, ..., R_{\text{FROC}}$. To be clearer one precedes the rating with the applicable paradigm: e.g., the ratings of marks range from FROC:1 to FROC:R_{FROC}. **The inferred-ROC z-sample (continuous rating) of a case is defined as the z-sample of the highest rated mark on the case or $-\infty$ if the case has no marks.** The inferred-ROC rating ROC:1 is reserved for cases with no marks. Since the ratings are ordered labels no ordering information is lost provided every other ROC rating is also "bumped up" by unity. The ROC ratings scale therefore extends from 1 to $R_{\text{FROC}} + 1$. Thus, a R_{FROC} rating FROC study corresponds to a $R_{\text{FROC}} + 1$ rating ROC study. The symbol $h_{k_t t}$ is used to denote the rating of the highest rated z-sample on case $k_t t$ with truth state t. Thus $h_{k_1 1}$ refers to the highest rating on a non-diseased case $k_1 1$ and $h_{k_2 2}$ refers to the highest rating on diseased case $k_2 2$. For non-diseased cases, the maximum is over all latent NLs on the case. For diseased cases, the maximum is over all latent NLs *and* latent LLs on the case. Define the set of ordered thresholds $\zeta_r < \zeta_{r+1}$ and dummy thresholds $\zeta_0 = -\infty, \zeta_{R_{\text{FROC}}+1} = \infty$. Then, if $\zeta_r \leq h_{k_t t} < \zeta_{r+1}$, where $r = 1, 2, ..., R_{\text{FROC}}$, the case is rated ROC:$(r + 1)$ and if $h_{k_t t} < \zeta_1$ the case is rated ROC:1. The lowest possible ROC rating on a case with at least one mark is ROC:2. A case with no latent sites *or* if the highest rated latent site satisfies $h_{k_t t} < \zeta_1$ is rated ROC:1. Note that one cannot distinguish between whether the ROC:1 rating was the result of the case not having any latent sites or the case had at least one latent site, but none of the z-samples exceeded ζ_1.

4.4 End-point of the ROC

A consequence of the possibility that some cases have no marks is that the ROC curve has the *constrained end-point property*, namely the full range of ROC space, i.e., $0 \leq \text{FPF} \leq 1$ and $0 \leq \text{TPF} \leq 1$, is not continuously accessible to the observer. In fact, $0 \leq \text{FPF} \leq \text{FPF}_{max}$ and $0 \leq \text{TPF} \leq \text{TPF}_{max}$ where FPF_{max} and TPF_{max} are generally less than unity.

Starting from a finite value as ζ_1 is lowered to $-\infty$ some of the previously ROC:1 rated cases that had at least one latent site but whose z-sample did not exceed ζ_1 will now generate marks and therefore the case will be "bumped-up" to the ROC:2 bin, until eventually *only cases with no latent sites* remain in the ROC:1 bin - these cases will never be rated ROC:2. An observer who finds no suspicious

regions, literally nothing to report, will assign the lowest available bin to the case, which happens to be ROC:1. The finite number of cases in the ROC:1 bin at infinitely low threshold has the consequence that the uppermost non-trivial continuously accessible operating point – that obtained by cumulating ratings ROC:2 and above - is below-left of (1,1). The (1,1) point is "trivially" reached when one cumulates the counts in all bins, i.e., ROC:1 and above. This behavior is distinct from traditional ROC models where the entire curve, extending from (0, 0) to (1, 1), is continuously accessible to the observer. This is because in conventional ROC models every case yields a finite decision variable, no matter how small. Lowering the lowest threshold to $-\infty$ eventually moves all cases in the previously ROC:1 bin to the ROC:2 bin, and one is eventually left with zero counts in the ROC:1 bin and the operating point, obtained by cumulating bins ROC:2 and above, is (1,1).

As another way of describing this unusual behavior, as the observer is encouraged to be more "aggressive in reporting lesions", the ROC point moves continuously upwards and to the right from (0, 0) to the end-point, $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$, and no further. The ROC curve cannot just "hang there" since cumulating all cases always yields the (1,1) operating point. **Therefore, the complete ROC curve is obtained by extending the end-point using a dashed line that connects it to (1,1).** The observer cannot operate along the dashed line. In the language of "moving up the ROC curve" there is a discontinuous jump from the end-point to (1,1). At the end-point the shape of the ROC curve changes from concave-down to a dashed straight line. It can be shown TBA that the limiting slope of the continuous ROC at the end-point is equal to the slope of the dashed straight line connecting the end-point to (1,1).

How closely the operating point approaches the limiting point $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$ is unrelated to the number of bins; rather, it depends on ζ_1. As the latter is lowered the observed end-point approaches $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$ from below-left. As will be shown below, how closely $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$ approaches (1,1) depends on the λ' and ν' RSM parameters: namely, as λ' and ν' increase, $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$ approaches (1,1) from below-left.

4.4.1 The abscissa of the ROC end-point

Consider the probability that a non-diseased case has at least one latent NL. Such a case will generate a finite value of $h_{k_1 1}$ and with an appropriately low ζ_1 it will be rated ROC:2 or higher. The probability of *zero* latent NLs, see Eqn. (3.2), is:

$$\mathrm{pmf}_{Poi}(0, \lambda') = \exp(-\lambda')$$

.

Therefore the probability that the case has *at least one* latent NL is the complement of the above probability. At sufficiently low ζ_1 each of these cases yields a

FP. Therefore, the maximum continuously accessible abscissa of the ROC, i.e., FPF_{max}, is:

$$\mathrm{FPF}_{max} = 1 - \exp\left(-\lambda'\right) \tag{4.1}$$

4.4.2 The ordinate of the ROC end-point

A diseased case has no marks, even for very low ζ_1, if it has zero latent NLs, the probability of which is $\exp(-\lambda')$, and it has zero latent LLs, the probability of which is, see Eqn. (3.4), $\mathrm{pmf}_{Bin}\left(0, L, \nu'\right) = (1-\nu')^L$.

Here L is the number of lesions in each diseased case, assumed constant.

- Assumption 1: occurrences of latent LLs are independent of each other, i.e., the probability that a lesion is found is independent of whether other lesions are found on the same case.
- Assumption 2: occurrences of latent NLs are independent of each other; i.e., the probability of a NL is independent of whether other NLs are found on the same case.
- Assumption 3: occurrence of a latent NL is independent of the occurrence of a latent LL on the same case.

By these assumptions the probability of zero latent NLs *and* zero latent LLs on a diseased case is the product of the two probabilities, namely

$$\exp(-\lambda')(1-\nu')^L$$

.

Therefore, the probability that there exists *at least one* latent site is the complement of the above expression, which equals TPF_{max}, i.e.,

$$\mathrm{TPF}_{max}\left(\mu, \lambda', \nu', L\right) = 1 - \exp\left(-\lambda'\right)\left(1-\nu'\right)^L \tag{4.2}$$

4.4.3 Variable number of lesions per case

Defining f_L the fraction of diseased cases with L lesions and L_{max} the maximum number of lesions per diseased case in the dataset, then:

$$\sum_{L=1}^{L_{max}} f_L = 1 \tag{4.3}$$

By restricting attention to the set of diseased cases with L lesions each, Eqn. (4.2) for TPF_{max} applies. Since TPF is a probability and probabilities of independent processes add it follows that:

$$\mathrm{TPF}_{max}\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right) = \sum_{L=1}^{Lmax} f_L \mathrm{TPF}_{max}\left(\mu, \lambda', \nu', L\right) \tag{4.4}$$

In other words the ordinate of the end-point is a weighted summation of $\mathrm{TPF}_{max}\left(\mu, \lambda', \nu', L\right)$ over the lesion distribution vector $\overrightarrow{f_L}$.

The expression for FPF_{max} is unaffected.

4.5 ROC curve

On the continuous ROC curve each case has at least one mark and therefore the inferred ROC decision variable is the rating of the highest rated mark $h_{k_t t}$ on the case. Therefore, FPF is the probability that $h_{k_t t}$ on a non-diseased case exceeds ζ and TPF is the probability that $h_{k_t t}$ on a diseased case exceeds ζ:

$$\left.\begin{aligned}\mathrm{FPF}\left(\zeta\right) &= P\left(h_{k_1 1} \geq \zeta\right)\\ \mathrm{TPF}\left(\zeta\right) &= P\left(h_{k_2 2} \geq \zeta\right)\end{aligned}\right\} \tag{4.5}$$

Varying the threshold parameter ζ from ∞ to $-\infty$ sweeps out the continuous section of the predicted ROC curve from (0,0) to $(\mathrm{FPF}_{max}, \mathrm{TPF}_{max})$.

4.5.1 Derivation of FPF

- Assumption 4: the z-samples of NLs on the same case are independent of each other.

Consider the set of non-diseased cases with n latent NLs each, where $n > 0$. According to 3.4 each latent NL yields a z sample from $N(0,1)$. The probability that a z-sample from a latent NL is smaller than ζ is $\Phi(\zeta)$. By the independence assumption the probability that all n samples are smaller than ζ is $(\Phi(\zeta))^n$. If all z-samples are smaller than ζ, then the highest z-sample $h_{k_1 1}$ must be smaller than ζ. Therefore, the probability that $h_{k_1 1}$ exceeds ζ is:

$$\left.\begin{aligned}\mathrm{FPF}\left(\zeta \mid n\right) &= P\left(h_{k_1 1} \geq \zeta \mid n\right)\\ &= 1 - \left[\Phi\left(\zeta\right)\right]^n\end{aligned}\right\} \tag{4.6}$$

The conditioning notation in Eqn. (4.6) reflects the fact that this expression applies specifically to non-diseased cases with n latent NLs each. To obtain

FPF_{max} one performs a Poisson-weighted summation of $\text{FPF}(\zeta \mid n)$ over n from 0 to ∞ (the inclusion of the $n = 0$ term is explained below):

$$\text{FPF}(\zeta, \lambda') = \sum_{n=0}^{\infty} \text{pmf}_{Poi}(n, \lambda') \, \text{FPF}(\zeta \mid n) \tag{4.7}$$

The infinite summations, see below, are easier performed using symbolic algebra software such as Maple^{TM}. Inclusion in the summation of $n = 0$, which term evaluates to zero, is done to make it easier for Maple to evaluate the summation in closed form. Otherwise one may need to simplify the Maple-generated result. The `Maple` code is shown below (Maple 17, Waterloo Maple Inc.), where `lambda` and `nu` refer to the primed quantities.

```
# Maple Code
restart;
phi := proc (t, mu) exp(-(1/2)*(t-mu)^2)/sqrt(2*Pi) end:
PHI := proc (c, mu) local t; int(phi(t, mu), t = -infinity .. c) end:
Poisson := proc (n, lambda) lambda^n*exp(-lambda)/factorial(n) end:
Bin := proc (l, L, nu) binomial(L, l)*nu^l*(1-nu)^(L-l) end:
FPF := proc(zeta,lambda) sum(Poisson(n,lambda)*(1 - PHI(zeta,0)^n), n=0..infinity);end
FPF(zeta, lambda);
```

The above code yields:

$$\text{FPF}(\zeta, \lambda') = 1 - \exp\left(-\frac{\lambda'}{2}\left[1 - \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \tag{4.8}$$

The error function in Eqn. (4.8) is related to the unit normal CDF function $\Phi(x)$ by:

$$\text{erf}(x) = 2\Phi\left(\sqrt{2}x\right) - 1 \tag{4.9}$$

Using this transformation yields the following simpler expression for FPF:

$$\text{FPF}(\zeta, \lambda') = 1 - \exp\left(-\lambda'\Phi(-\zeta)\right) \tag{4.10}$$

The software implementation follows:

```
# lambdaP is the physical lambda' parameter
FPF <- function (zeta, lambdaP) {
  x = 1 - exp(-lambdaP * pnorm(-zeta))
  return(x)
}
```

Because Φ ranges from 0 to 1, $\text{FPF}(\zeta, \lambda')$ ranges from 0 to $\exp(-\lambda')$.

4.5.2 Derivation of TPF

The derivation of the true positive fraction TPF(ζ) follows a similar line of reasoning except this time one needs to consider the highest of the latent NLs and latent LL z-samples. Consider a diseased case with L lesions, n latent NLs and l latent LLs. Each latent NL yields a decision variable sample from $N(0,1)$ and each latent LL yields a sample from $N(\mu,1)$. The probability that all n latent NLs have z-samples less than ζ is $[\Phi(\zeta)]^n$. The probability that all l latent LLs have z-samples less than ζ is $[\Phi(\zeta-\mu)]^l$. Using the independence assumptions, the probability that all latent marks have z-samples less than ζ is the product of these two probabilities. The probability that $h_{k_2 2}$ (the highest z-sample on diseased case $k_2 2$) is larger than ζ is the complement of the product probabilities, i.e.,

$$\mathrm{TPF}_{n,l}\left(\zeta,\mu,n,l,L\right)=P\left(h_{k_2 2}\geq\zeta\mid\mu,n,l,L\right)=1-\left[\Phi\left(\zeta\right)\right]^{n}\left[\Phi\left(\zeta-\mu\right)\right]^{l}$$

One averages over the distributions of n and l to obtain the desired ROC-ordinate:

$$\left.\begin{array}{l}\mathrm{TPF}\left(\zeta,\mu,\lambda',\nu'\right)\times\\=\sum_{n=0}^{\infty}\mathrm{pmf}_{Poi}(n,\lambda')\times\\\sum_{l=0}^{L}\mathrm{pmf}_{Bin}(l,\nu',L)\mathrm{TPF}_{n,l}\left(\zeta,\mu,n,l\right)\end{array}\right\}\tag{4.11}$$

This can be evaluated using Maple yielding:

$$\left.\begin{array}{l}\mathrm{TPF}\left(\zeta,\mu,\lambda',\nu',L\right)\\=1-\exp\left(-\lambda'\Phi\left(-\zeta\right)\right)\left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^{L}\end{array}\right\}\tag{4.12}$$

4.5.3 Variable number of lesions per case

To extend the results to varying numbers of lesions per diseased case, one averages the right hand side of (4.12) over the fraction of diseased cases with L lesions:

$$\left.\begin{array}{l}\mathrm{TPF}\left(\zeta,\mu,\lambda',\nu',\overrightarrow{f_L}\right)=\\1-\exp\left(-\lambda'\Phi\left(-\zeta\right)\right)\sum_{L=1}^{L_{max}}f_L\left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^{L}\end{array}\right\}\tag{4.13}$$

The expression for FPF, Eqn. (4.10), is unaffected.

The software implementation follows:

```
# lambdaP is the physical lambda' parameter
# nuP is the physical nu' parameter
# lesDistr is the lesion distributin vector f_L
TPF <- function (zeta, mu, lambdaP, nuP, lesDistr){
  Lmax <- length(lesDistr)
  x <- 1
  for (L in 1:Lmax ) {
    x <- x - exp(-lambdaP * pnorm(-zeta)) * lesDistr[L] * (1 - nuP * pnorm(mu - zeta))
  }
  return(x)
}
```

4.6 Proper ROC curve

A proper ROC curve has the property that it never crosses the chance diagonal and its slope never increases as the operating point moves up the ROC curve (Metz and Pan, 1999; Macmillan and Creelman, 2004). *It is shown next that the RSM predicted ROC curve, including the dashed straight line extension, is proper* [1]. We considered first the continuous section which is below-left of the end-point. In 4.13 a proof is presented that the slope is continuous at the end-point transition from a continuous curve to the dashed straight line. In 4.14 the slope near the end-point is examined numerically to resolve an apparent paradox, namely the ROC plot can appear discontinuous at the end-point when in fact no discontinuity exists.

For convenience one abbreviates FPF and TPF to x and y, respectively, and suppresses the dependence on model parameters. From Eqn. (4.10) and Eqn. (4.13) one can express the ROC coordinates as:

$$\left.\begin{aligned} x(\zeta) &= 1 - G(\zeta) \\ y(\zeta) &= 1 - F(\zeta)\, G(\zeta) \end{aligned}\right\} \tag{4.14}$$

where:

$$\left.\begin{aligned} G(\zeta) &= \exp\left(-\lambda' \Phi(-\zeta)\right) \\ F(\zeta) &= \sum_{L=1}^{L_{max}} f_L \left(1 - \nu' \Phi(\mu - \zeta)\right)^L \end{aligned}\right\} \tag{4.15}$$

[1] The statement in the physical book that the proper property only applies to the continuous section is incorrect.

These equations have exactly the same structure as (Swensson, 1996) Eqns. 1 and 2 and the logic used there to demonstrate that ROC curves predicted by Swensson's LROC model was proper also applies to the present situation. Specifically, since the Φ function ranges between 0 and 1 and $0 \leq \nu' \leq 1$, it follows that $F(\zeta) \leq 1$. Therefore $y(\zeta) \geq x(\zeta)$ and the ROC curve is constrained to the upper half of the ROC space, namely the portion above the chance diagonal. Additionally, the more general constraint shown by Swensson applies, namely the slope of the ROC curve at any operating point (x, y) cannot be less than the slope of the dashed straight line connecting (x, y) and $(\text{FPF}_{max}, \text{TPF}_{max})$, the coordinates of the RSM end-point. This implies that the slope decreases monotonically and also rules out curves with "hooks".

4.7 ROC decision variable pdfs

In TBA (binormal-model-pdf-curves-appendix-1) the pdf functions were derived for non-diseased and diseased cases for the unequal variance binormal ROC model. The procedure was to take the derivative of the appropriate cumulative distribution function (CDF) with respect to ζ. An identical procedure is used for the RSM.

The CDF for non-diseased cases is the complement of FPF. The pdf for non-diseased cases is given by:

$$\text{pdf}_N(\zeta) = \frac{\partial}{\partial \zeta}\left(1 - \text{FPF}(\zeta, \lambda')\right) \tag{4.16}$$

Similarly, for diseased cases,

$$\text{pdf}_D(\zeta) = \frac{\partial}{\partial \zeta}\left(1 - \text{TPF}\left(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}\right)\right) \tag{4.17}$$

Both expressions can be evaluated using Maple. The pdfs are implemented in the `RJafroc` function `PlotRsmOperatingCharacteristics()`.

The integrals of the pdfs (non-diseased followed by diseased) over the entire allowed range are given by (note the vertical bar notation, meaning the difference of two limiting values):

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_N(\zeta)\, d\zeta &= \left(1 - \text{FPF}(\zeta, \lambda')\right)\Big|_{-\infty}^{\infty} \\ &= \text{FPF}_{max}\end{aligned}\right\} \tag{4.18}$$

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_D(\zeta)\, d\zeta &= \left(1 - \text{TPF}\left(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}\right)\right)\Big|_{-\infty}^{\infty} \\ &= \text{TPF}_{max}\end{aligned}\right\} \tag{4.19}$$

In other words, they evaluate to the coordinates of the predicted end-point, *each of which is less than unity*. The reason is that the integration is along the *continuous* section of the ROC curve and does not include the contribution along the dashed straight line extension from $(\text{FPF}_{max}, \text{TPF}_{max})$ to (1,1). The latter contributions correspond to cases with no marks, i.e., $1-\text{FPF}_{max}$ for non-diseased cases and $1-\text{TPF}_{max}$ for diseased cases. Adding these contributions to the integrals along the continuous section yields unity for both types of cases. [2]

4.8 ROC AUC

It is possible to numerically perform the integration under the RSM-ROC curve to get AUC:

$$AUC_{RSM}^{ROC}\left(\mu,\lambda,\nu,\zeta_1,\overrightarrow{f_L}\right) = \sum_{L=0}^{L_{max}} f_L \int_0^1 \text{TPF}\left(\zeta,\mu,\lambda,\nu,L\right) d\left(\text{FPF}\left(\zeta,\lambda\right)\right) \tag{4.20}$$

The superscript *ROC* is needed to keep track of the operating characteristic that is being predicted (for RSM other possibilities are AFROC, wAFROC, FROC) and the subscript *RSM* keeps track of the predictive model that is being used (for ROC models - binormal, CBM or PROPROC - the superscript is always ROC).

The right hand side of Eqn. (4.20) can be evaluated using a numerical integration function implemented in `R`, which is used in the `RJafroc` function `UtilAnalyticalAucsRSM()` whose help page follows:

The arguments to `UtilAnalyticalAucsRSM()` are the intrinsic RSM parameters μ, λ, ν and ζ_1. The default value of ζ_1 is $\zeta_1 = -\infty$. The remaining arguments `lesDistr` and `relWeights` are not RSM parameters per se, rather they specify the lesion-richness of the diseased cases and the relative lesion weights (not needed for computing ROC AUC). The dimensions of `lesDistr` and `relWeights` are each equal to the maximum number of lesions per case L_{max}. In the following code $L_{max} = 3$ and `lesDistr <- c(0.5, 0.3, 0.2)`, meaning 50 percent of diseased cases have one lesion per case, 30 percent have two lesions and 20 percent have three lesions.

The function returns a list containing the AUCs under the ROC and other operating characteristics.

[2] The original RSM publications (Chakraborty, 2006b,a) unnecessarily introduced Dirac delta functions to force the integrals to be unity. The explanation given here should clarify the issue.

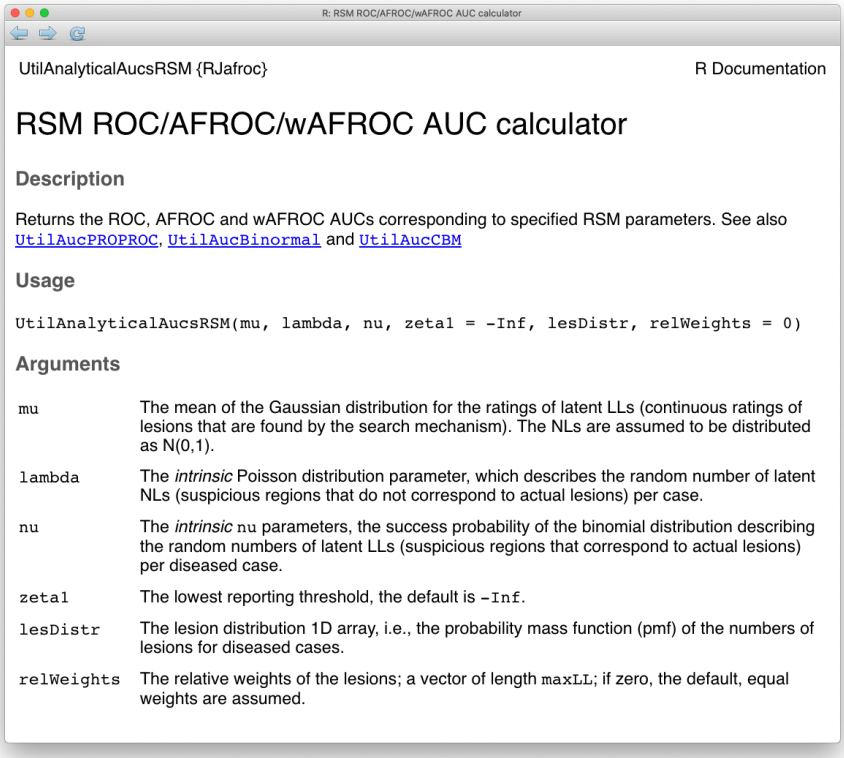


Figure 4.1: Help page for ‘RJafroc‘ function ‘UtilAnalyticalAucsRSM‘.

```
mu <- 1; lambda <- 1; nu <- 1
lesDistr <- c(0.5, 0.3, 0.2) # implies L_max = 3
aucs <- UtilAnalyticalAucsRSM(mu = mu,
                              lambda = lambda,
                              nu = nu,
                              lesDistr = lesDistr)
cat("mu = ", mu,
", lambda = ", lambda,
", nu = ", nu,
", AUC ROC = ", aucs$aucROC, "\n")
```

```
## mu =  1 , lambda =  1 , nu =  1 , AUC ROC =  0.7802109
```

Experimenting with different parameter combinations reveals the following behavior for ROC AUC.

- AUC is an increasing functions of μ. Increasing perceptual signal-to-noise-ratio leads to improved performance: for background on this important dependence see 1.6. Increasing μ increases the separation between the two pdfs defining the ROC curve, which increases AUC. Furthermore, the number of NLs decreases because $\lambda' = \lambda/\mu$ decreases, which increases performance. Finally, ν' increases approaching unity, which leads to more LLs and increased performance. *Because all three effects reinforce each other, a change in μ results in a large effect on performance.*

- AUC increases as λ decreases. Decreasing λ results in fewer NLs which results in increased performance. This is a relatively weak effect.

- AUC increases as ν increases. Increasing ν results in more LLs being marked, which increases performance. This is a relatively strong effect.

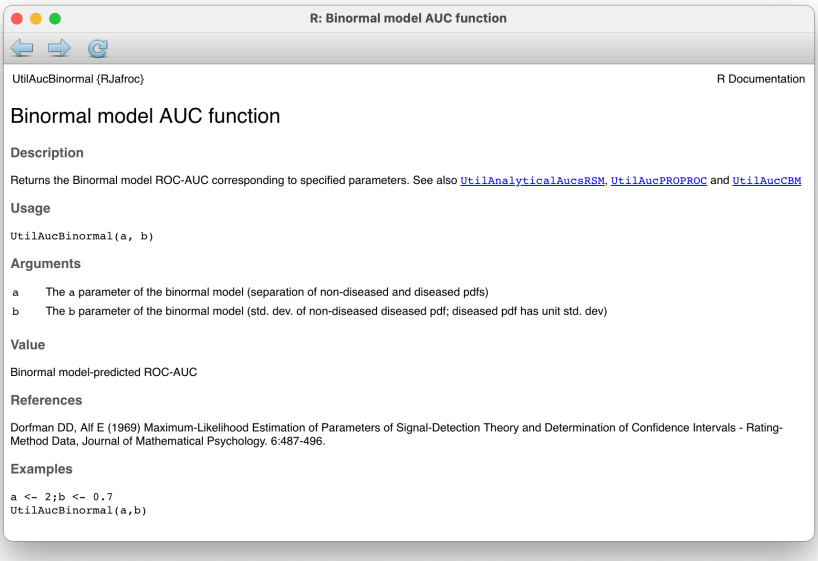
- AUC decreases as ζ_1 increases. This important effect is discussed in the next section.

- ROC AUC increases with L_{max}. With more lesions per case, there is increased probability that that at least one of them will result in a LL, and the diseased case pdf moves to the right, both of which result in increased performance.

- ROC AUC increases as `lesDistr` is weighted towards more lesions per case. For example, `lesDistr <- c(0, 0, 1)` (all cases have 3 lesions per case) will yield higher performance than `lesDistr <- c(1, 0, 0)` (all cases have one lesion per case).

4.9 ζ_1 dependence of ROC AUC

When it comes to predicted ROC AUC there is an important difference between conventional ROC models and the RSM. The former has no dependence on ζ_1. This is because in the ROC model every case yields a rating, no matter how low the z-sample, implying that effectively $\zeta_1 = -\infty$. The lack of ζ_1 dependence is demonstrated by the help page for function `UtilAucBinormal`, shown below, which depends on only two parameters, a and b (the two-parameter dependence is also true for other ROC models implemented in `RJafroc`, e.g., `UtilAucCBM` and `UtilAucPROPROC`).


R: Binormal model AUC function

UtilAucBinormal {RJafroc} R Documentation

Binormal model AUC function

Description

Returns the Binormal model ROC-AUC corresponding to specified parameters. See also UtilAnalyticalAucsRSM, UtilAucPROPROC and UtilAucCBM

Usage

```
UtilAucBinormal(a, b)
```

Arguments

a	The a parameter of the binormal model (separation of non-diseased and diseased pdfs)
b	The b parameter of the binormal model (std. dev. of non-diseased diseased pdf; diseased pdf has unit std. dev)

Value

Binormal model-predicted ROC-AUC

References

Dorfman DD, Alf E (1969) Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals - Rating-Method Data, Journal of Mathematical Psychology. 6:487-496.

Examples

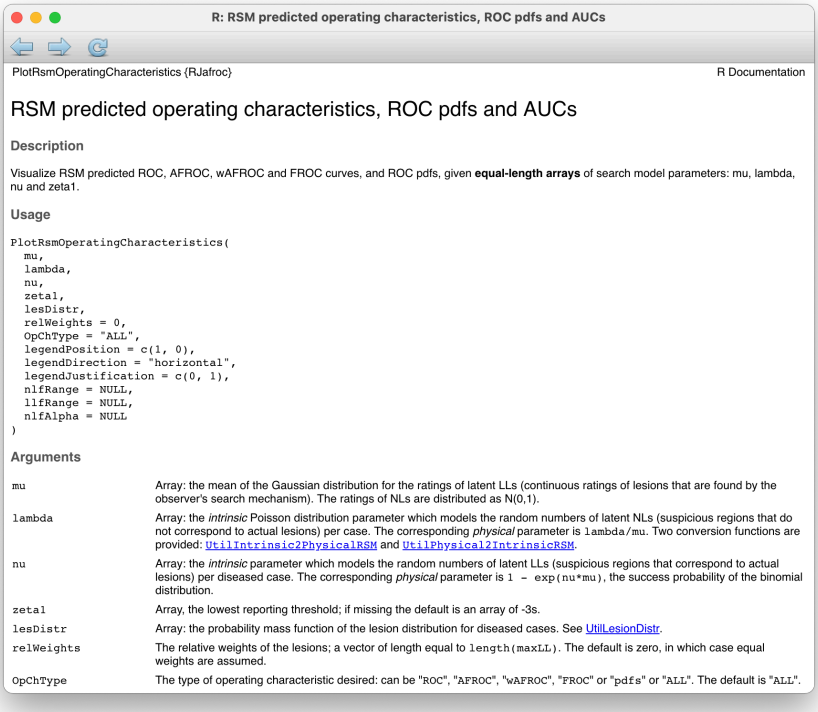
```
a <- 2;b <- 0.7
UtilAucBinormal(a,b)
```

Figure 4.2: Help page for 'RJafroc' function 'UtilAucBinormal'.

In contrast, in addition to the basic RSM parameters, i.e., μ, λ and ν, the rsm-pred have an additional dependence on ζ_1. This is because the value of ζ_1 determines the location of the end-point. The ζ_1 dependence is demonstrated next for the ROC plots, but it is true for all RSM predictions.

The dependence is demonstrated next for two values: $zeta_1 = -10$ and $zeta_1 = 1$. The common parameter values are $\mu = 2$, $\lambda = 1$, $\nu = 1$, as shown in the following code-chunk.

```
roc <- PlotRsmOperatingCharacteristics(
     mu = c(2,2),
     lambda = c(1,1),
```



R: RSM predicted operating characteristics, ROC pdfs and AUCs

PlotRsmOperatingCharacteristics {RJafroc} R Documentation

RSM predicted operating characteristics, ROC pdfs and AUCs

Description

Visualize RSM predicted ROC, AFROC, wAFROC and FROC curves, and ROC pdfs, given **equal-length arrays** of search model parameters: mu, lambda, nu and zeta1.

Usage

```
PlotRsmOperatingCharacteristics(
  mu,
  lambda,
  nu,
  zeta1,
  lesDistr,
  relWeights = 0,
  OpChType = "ALL",
  legendPosition = c(1, 0),
  legendDirection = "horizontal",
  legendJustification = c(0, 1),
  nlfRange = NULL,
  llfRange = NULL,
  nlfAlpha = NULL
)
```

Arguments

`mu`	Array: the mean of the Gaussian distribution for the ratings of latent LLs (continuous ratings of lesions that are found by the observer's search mechanism). The ratings of NLs are distributed as N(0,1).
`lambda`	Array: the *intrinsic* Poisson distribution parameter which models the random numbers of latent NLs (suspicious regions that do not correspond to actual lesions) per case. The corresponding *physical* parameter is `lambda/mu`. Two conversion functions are provided: `UtilIntrinsic2PhysicalRSM` and `UtilPhysical2IntrinsicRSM`.
`nu`	Array: the *intrinsic* parameter which models the random numbers of latent LLs (suspicious regions that correspond to actual lesions) per diseased case. The corresponding *physical* parameter is `1 - exp(nu*mu)`, the success probability of the binomial distribution.
`zeta1`	Array, the lowest reporting threshold; if missing the default is an array of -3s.
`lesDistr`	Array: the probability mass function of the lesion distribution for diseased cases. See UtilLesionDistr.
`relWeights`	The relative weights of the lesions; a vector of length equal to `length(maxLL)`. The default is zero, in which case equal weights are assumed.
`OpChType`	The type of operating characteristic desired: can be "`ROC`", "`AFROC`", "`wAFROC`", "`FROC`" or "`pdfs`" or "`ALL`". The default is "`ALL`".

Figure 4.3: Help page for ‘RJafroc‘ function ‘PlotRsmOperatingCharacteristics‘.

```
      nu = c(1,1),
      zeta1 = c(-10, 1),
      lesDistr = c(0.5, 0.5),
      relWeights = c(0.5, 0.5),
      OpChType = "ROC",
      legendPosition = "null"
)
```

Clearly the red curve has higher AUC. The specific values are 0.9386603 for the red curve and 0.9031788 for the green curve.

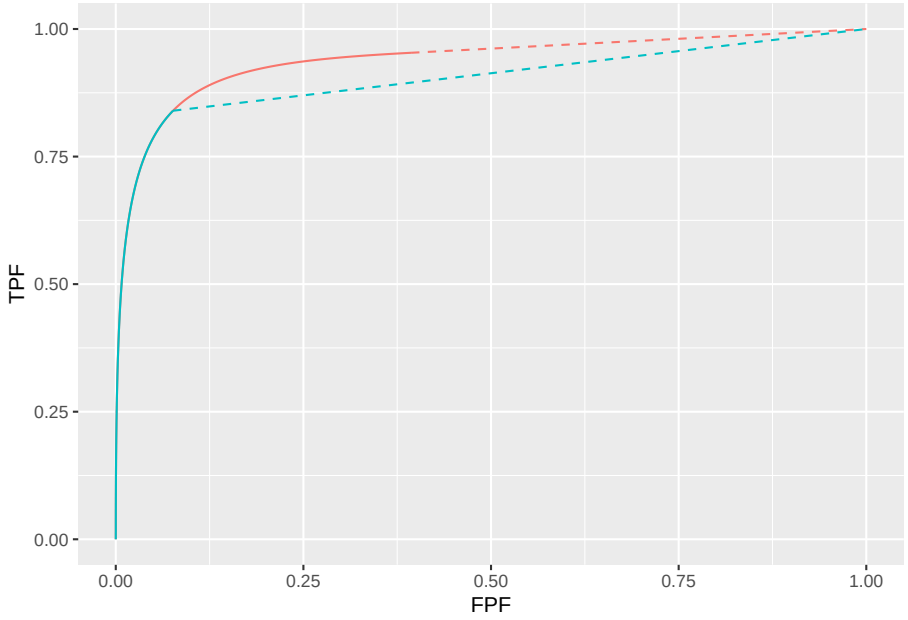


Figure 4.4: ROC curves for two values of ζ_1: both curves correspond to $\mu = 2$, $\nu = 1$ and $\lambda = 1$. The red curve corresponds to $\zeta_1 = -10$ and the blue curve to $\zeta_1 = 1$.

A consequence of the ζ_1 dependence is that if one uses ROC AUC as the measure of performance, the optimal threshold is $\zeta_1 = -\infty$. In particular, a CAD algorithm that generates FROC data should show all generated marks to the radiologist. This is not the strategy adopted by any CAD designer that I am aware of. I will address this issue, i.e., what is the optimal value of the reporting threshold, in Chapter 14.

4.10 Example ROC curves

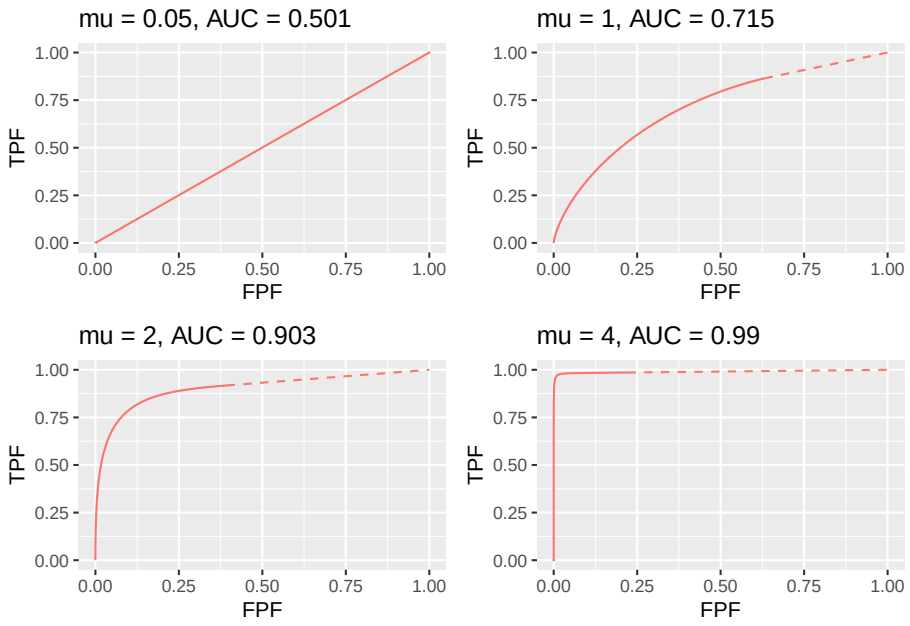


Figure 4.5: ROC curves for indicated values of the μ parameter. Notice the transition, as μ increases, from near chance level performance to almost perfect performancea as the end-point moves from near (1,1) to near (0,1).

Fig. 4.5 displays ROC curves for indicated values of μ. The remaining RSM model parameters are $\lambda = 1$, $\nu = 1$ and $\zeta_1 = -\infty$ and there is one lesion per diseased case.

The following are evident from these figures:

1. As μ increases the ROC curve more closely approaches the upper-left corner of the ROC plot. This signifies increasing performance and the area under the ROC and AFROC curves approach unity. The end-point abscissa decreases, meaning increasing numbers of unmarked non-diseased cases, i.e., more perfect decisions on non-diseased cases. The end-point ordinate increases, meaning decreasing numbers of unmarked lesions, i.e., more good decisions on diseased cases.
2. For μ close to zero the operating characteristic approaches the chance diagonal and the area under the ROC curve approaches 0.5.
3. The area under the ROC increases monotonically from 0.5 to 1 as μ increases from zero to infinity.

4. For large μ the accessible portion of the operating characteristic approaches the vertical line connecting (0,0) to (0,1), the area under which is zero. The complete ROC curve is obtained by connecting this point to (1,1) by the dashed line and in this limit the area under the complete ROC curve approaches unity. Omitting the area under the dashed portion of the curve will result in a severe underestimate of true performance.
5. As L_{max} increases (allowed values are 1, 2, 3, etc.) the area under the ROC curve increases, approaching unity and TPF_{max} approaches unity. With more lesions per diseased case, the chances are higher that at least one of them will be found and marked. However, FPF_{max} remains constant as determined by the constant value of $\lambda' = \frac{\lambda}{\mu}$, Eqn. (4.1)
6. As λ decreases FPF_{max} decreases to zero and TPF_{max} decreases. The decrease in TPF_{max} is consistent with the fact that, with fewer NLs, there is less chance of a NL being rated higher than a LL, and one is completely dependent on at least one lesion being found.
7. As ν increases FPF_{max} stays constant at the value determined by λ and μ, while TPF_{max} approaches unity. The corresponding physical parameter ν' increases approaching unity, guaranteeing every lesion will be found.

4.11 Example RSM pdf curves

Fig. 4.6 shows pdf plots for the same values of parameters as in Fig. 4.5.

Consider the plot of the pdfs for $\mu = 1$. Since the integral of a pdf function over an interval amounts to counting the fraction of events occurring in the interval, it should be evident that the area under the non-diseased pdf equals FPF_{max} and that under the diseased pdf equals TPF_{max}. For the chosen value $\lambda = 1$ one has $\text{FPF}_{max} = 1 - e^{-\lambda} = 0.632$. The area under the non-diseased pdf is less than unity because it is missing the contribution of non-diseased cases with no marks, the probability of which is $e^{-\lambda} = e^{-1} = 0.368$. Equivalently, it is missing the area under the dashed straight line segment of the ROC curve. Likewise, the area under the diseased pdf equals TPF_{max}, Eqn. (4.2), which is also less than unity. For the chosen values of $\mu = \lambda = \nu = L = 1$ it equals $\text{TPF}_{max} = 1 - e^{-\lambda}e^{-\nu} = 0.865$. This area is somewhat larger than that under the non-diseased pdf, as is evident from visual examination of the plot. A greater fraction of diseased cases generate marks than do non-diseased cases, consistent with the presence of lesions in diseased cases. The complement of 0.865 is due to diseased cases with no marks, which account for a fraction 0.135 of diseased cases. To summarize, the pdf's do not integrate to unity for the reason that the integrals account only for the continuous section of the ROC curve and do not include cases with zero latent marks that do not generate z-samples. The effect becomes more exaggerated for higher values of μ as this causes FPF_{max} to further decrease.

The plot in Fig. 4.6 labeled $\mu = 0.05$ may be surprising. Since it corresponds

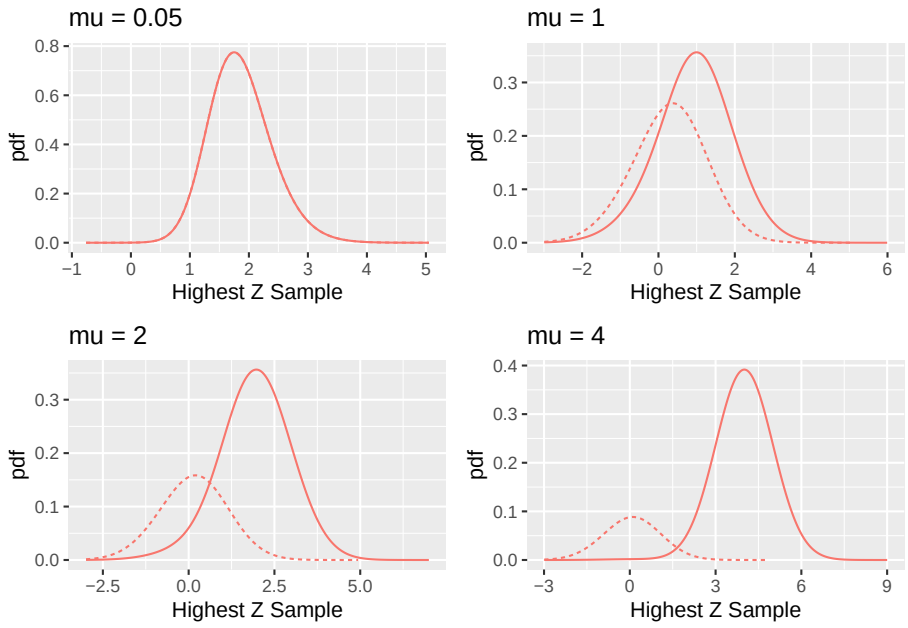


Figure 4.6: RSM pdf curves for indicated values of the μ parameter. The solid curve corresponds to diseased cases and the dotted curve corresponds to non-diseased cases.

to a small value of μ, one may expect both pdfs to overlap and be centered at zero. Instead, while they do overlap, the shape is distinctly non-Gaussian and centered at approximately 1.8. This is because the small value of μ results in a large value of the λ' parameter, since $\lambda' = \lambda/\mu = 20$. The highest of a large number of samples from the unit normal distribution is not normal and is peaked at a value above zero (Fisher and Tippett, 1928).

4.12 TBA Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to –infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are "lost in the mail"; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

ROC, FROC and AFROC curves were derived (wAFROC is implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me TBA 47,48, has used FROC curves to measure performance, this is going to be difficulty for

some to accept. The examples in Fig. 17.6 (A- F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which in my opinion, is slightly misleading as at screening lesions are found not detected ("detection" is "discover or identify the presence or existence of something", correct localization is not necessarily implied; the more precise term is "localize"). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as "diagnostic aids" to optimize performance of a reader. For example, is search performance is low, then training using mainly non-diseased cases is called for, so the resident learns the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large . The RSM explains most empirical results regarding binormal model fits: the common finding that $b < 1$; that b

decreases with increasing lesion pSNR (large and / or); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter < 1 and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects27-29. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

4.12.1 The Wagner review

The two RSM papers12,13 were honored by being included in a list of 25 papers the "Highlights of 2006" in Physics in Medicine and Biology. As stated by the publisher: "I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.

One of the reviewers was the late Dr. Robert ("Bob") Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduces one of his comments with minor edits, as it pertains to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I'm thinking here about the straight-line piece of the ROC curve from the max to (1, 1). 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point–because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a dashed straight line, then the reader would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out–but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space, but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will not knowingly guess. If the radiologist sees nothing to report, nothing will be reported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

4.13 Appendix 1: Proof of continuity of slope at the end-point

The following proof is adapted from a document supplied by Dr. Xuetong Zhai, then (ca. 2017) a graduate student working under the supervision of the author.

The end point coordinates of the continuous part of ROC curve was derived above, Eqn. (4.1) for FPF_{max} and Eqn. (4.2) for TPF_{max}. Therefore, the slope m_{st} of the dashed dashed straight line is:

$$\left.\begin{aligned} m_{st} =& \frac{1-\mathrm{TPF}_{max}}{1-\mathrm{FPF}_{max}} \\ =& \frac{\sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^L \exp\left(-\lambda'\right)}{\exp\left(-\lambda'\right)} \\ =& \sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^L \end{aligned}\right\} \tag{4.21}$$

On the continuous section, $g \equiv \mathrm{FPF}$ and $h \equiv \mathrm{TPF}$ are defined by (4.10) and (4.13), respectively. Therefore,

$$\left.\begin{aligned} g =& 1-\exp\left(-\lambda'\Phi\left(-\zeta\right)\right) \\ h =& 1-\sum_{L=1}^{L_{max}} f_L \exp\left(-\lambda'\Phi\left(-\zeta\right)\right)\left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^L \end{aligned}\right\} \tag{4.22}$$

Taking the differentials of these functions with respect to ζ it follows that the slope of the ROC is given by:

$$\left.\begin{aligned} \frac{dh}{dg} = & \sum_{L=1}^{L_{max}} f_L \left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^{L-1} \times \\ & \left[\frac{L\nu'\phi\left(\mu-\zeta\right)}{\lambda'\phi\left(-\zeta\right)} + \left(1-\nu'\Phi\left(\mu-\zeta\right)\right)\right] \end{aligned}\right\} \tag{4.23}$$

Using the following result:

$$\left.\begin{aligned} & \lim_{\zeta\to-\infty} \frac{\phi\left(\mu-\zeta\right)}{\phi\left(-\zeta\right)} \\ =& \lim_{\zeta\to-\infty} \frac{\frac{1}{\sqrt{2\pi}}\exp\left(\frac{\left(\mu-\zeta\right)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\zeta^2}{2}\right)} \\ =& \lim_{\zeta\to-\infty} \exp\left(\frac{\mu\zeta-\mu^2}{2}\right) \\ =& 0 \end{aligned}\right\} \tag{4.24}$$

it follows that:

$$\left.\begin{aligned} &\lim_{\zeta \to -\infty} \frac{dh}{dg} \\ =&\sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^{L-1} \left(1-\nu'\right) \\ =&\sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^{L} \\ =&m_{st} \end{aligned}\right\} \tag{4.25}$$

This proves that the limiting slope of the continuous section of the ROC curve equals that of the dashed straight line connecting the end-point to (1,1).

4.14 Appendix 2: Numerical illustration of continuity

The code in this section examines the slope of the ROC curve as one approaches the end-point $\zeta_1 = -\infty$. The RSM parameter values are $\mu = 0.5$, $\lambda = 0.1$ and $\nu = 0.8$, and twenty percent of the diseased cases have one lesion and 80 percent have 2 lesions, i.e. `lesDistr` -> c(0.2, 0.8).

```
mu <- 0.5
lambda <- 0.1
nu <- 0.8
lambdaP <- lambda / mu
nuP <- 1 - exp(-mu * nu)
lesDistr <- c(0.2, 0.8)
```

One calculates the coordinates of the end-point and the slope of the line connecting it to (1,1).

```
maxFPF <- FPF (-Inf, lambdaP)
maxTPF <- TPF (-Inf, mu, lambdaP, nuP, lesDistr)
mStLine <- (1 - maxTPF) / (1 - maxFPF)
```

The end-point coordinates are (0.1812692, 0.5959341) and the slope is 0.4935272. Next one calculates and displays the ROC curve.

```
ret <- PlotRsmOperatingCharacteristics(
  mu,
  lambda,
  nu,
```

```
  zeta1 = -Inf, # fixed: this function used to break for -Inf
  OpChType = "ROC",
  lesDistr = lesDistr
)
```

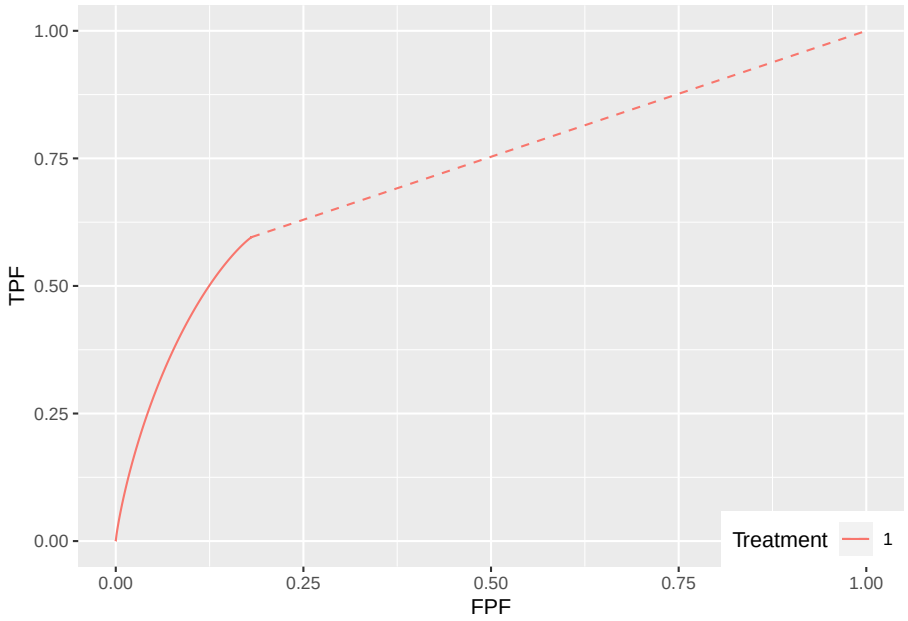


Figure 4.7: ROC curve for selected RSM parameters. The slope of the dashed line is 0.4935272.

At first sight the slope appeared to me to be discontinuous at the end-point [3] but this is not true. In fact the slope decreases as one approaches the end-point, and in the limit equals that of the dashed line. This is demonstrated by the next code section which creates a finely-spaced ζ array ranging from -3 to -20. These are the points at which the slope is numerically calculated. Two types of calculations were performed - one using standard `R` double precision arithmetic and one using multiple precision arithmetic. The R-package `Rmpfr` was used for the latter. For example, the line `zeta_mpr -> mpfr(zeta, 2000)` generates a 2000-bit representation of ζ. All subsequent computations using `zeta_mpr` uses multiple precision arithmetic. The computed slopes are saved in two arrays, `y1`, the standard precision arithmetic slope and `y2`, the multiple precision arithmetic slope.

[3] Others have stated a different visual impression.

```
zeta_arr <- c(seq(-3, -5, -0.2), seq(-5, -20, -0.5))
y1 <- array(0, length(zeta_arr))
y2 <- array(0, length(zeta_arr))
i <- 0
for (zeta in zeta_arr) {
  i <- i + 1
  # normal precision arithmetic
  zeta2 <- zeta + 1e-6
  delta_FPF <- FPF (zeta, lambdaP) - FPF (zeta2, lambdaP)
  delta_TPF <- TPF (zeta, mu, lambdaP, nuP, lesDistr) -
    TPF (zeta2, mu, lambdaP, nuP, lesDistr)
  mAnal <- delta_TPF / delta_FPF
  y1[i] <- mAnal
  # end normal precision arithmetic

  # multiple precision arithmetic
  zeta_mpr <- mpfr(zeta, 2000) # 2000 digit precision
  zeta2_mpr <- zeta_mpr + 1e-12 # small increment
  delta_FPF <- FPF (zeta_mpr, lambdaP) - FPF (zeta2_mpr, lambdaP)
  delta_TPF <- TPF (zeta_mpr, mu, lambdaP, nuP, lesDistr) -
    TPF (zeta2_mpr, mu, lambdaP, nuP, lesDistr)
  mAnalRmpfr <- delta_TPF / delta_FPF
  temp <- as.numeric(mAnalRmpfr)
  if (is.nan(temp)){
    y2[i] <- NA
  } else y2[i] <- temp
  # end multiple precision arithmetic
}
```

The next code section displays 3 plots.

```
m1 <- data.frame(z = zeta_arr, m = y1)
m2 <- data.frame(z = zeta_arr, m = y2)
plots <- ggplot(
  mapping = aes(x = z, y = m)) +
  geom_line(data = m1, linetype = "dashed", color = "blue") +
  geom_line(data = m2) +
  ylim(0, 1) + xlim(-15, -3) +
  geom_hline(yintercept = mStLine, color = "red",linetype = "dashed") +
  xlab(label = "zeta") + ylab(label = "slopes")
suppressWarnings(print(plots))
```

The solid black line is the plot, using multiple precision arithmetic, of slope of the ROC curve vs. ζ. The dashed blue line is the slope using standard precision arithmetic. The horizontal dashed red line is the slope of the straight line

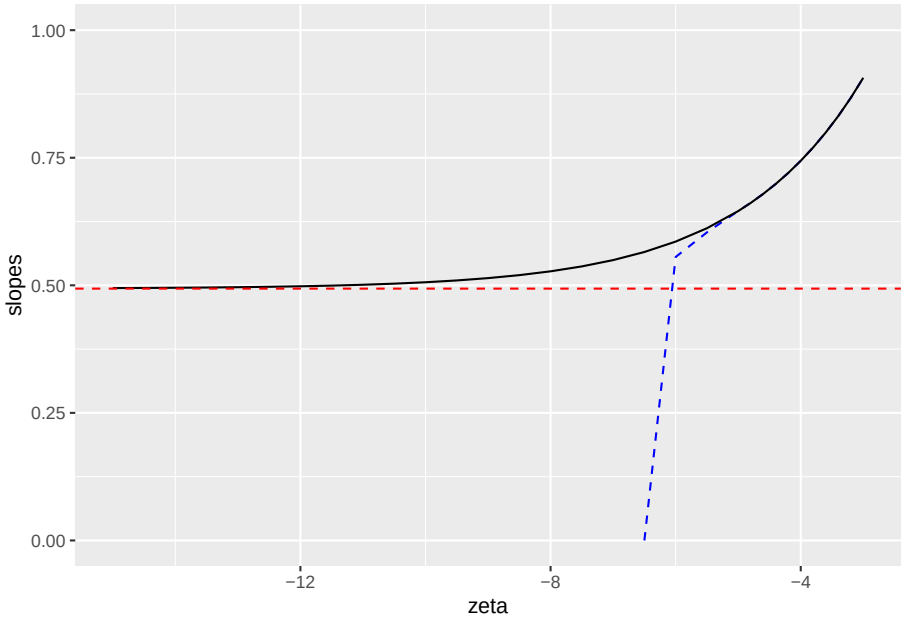


Figure 4.8: Horizontal dashed red line: the value of `mStLine`, the slope of the straight line connecting the ROC end-point to (1,1). Dashed blue line: slope using double precision arithmetic. Solid black line: slope using multiple precision arithmetic - this curve approaches the limiting value `mStLine`.

connecting the end-point to (1,1), i.e., 0.4935272. Standard precision arithmetic breaks down below $\zeta \approx -6$ rapidly falling to illegal values `Nan` (above $\zeta \approx -5$ there is little difference between standard and multiple precision). The multiple precision curve approaches the slope of the straight line as ζ approaches -20. This confirms numerically the continuity of the slope of the ROC at the end-point.

4.15 References

1. 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.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. Medical Physics. 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic Radiology. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys Med Biol. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol. 2006;51:3449–3462.

Chapter 5

Empirical plots

5.1 TBA How much finished

70%

5.2 Introduction

Operating characteristics are visual depicters of performance. If properly defined, scalar quantities derived from operating characteristics can serve as quantitative measures of performance, termed figures of merit (FOMs). The previous chapter defined the FROC curve and suggested the area under this curve as a possible FOM. This chapter introduces mathematical expressions for empirical operating characteristics (FROC and others) possible with FROC data and associated FOMs.

A distinction between latent and actual marks is made followed by a summary of FROC notation applicable to a single modality single reader dataset. This is a key table, which will be used in later chapters. Following this, different empirical operating characteristics proposed for FROC data are described. Formulae are given for calculating each empirical operating characteristic.

The observed end-point of an operating characteristic is defined as that operating point achieved by cumulating all the ratings. For the FROC plot it is demonstrated that the observed FROC curve is not contained in the unit square, unlike the other operating characteristics, which are contained in the unit square.

5.3 Mark rating pairs

FROC data consists of mark-rating pairs. Each mark indicates the location of a region suspicious enough to warrant reporting and the rating is the associated confidence level. A mark is recorded as *lesion localization* (LL) if it is sufficiently close to a true lesion, according to the adopted proximity criterion, and otherwise it is recorded as *non-lesion localization* (NL).

In an FROC study the number of marks on an image is an a-priori unknown modality-reader-case dependent non-negative random integer. It is incorrect to estimate it by dividing the image area by the lesion area because not all regions of the image are equally likely to have lesions, lesions do not have the same size, and perhaps most important, clinicians don't assign equal attention units to all areas of the image. The best insight into the number of marks per case is obtained from eye-tracking studies (Duchowski, 2002), but even here the information is incomplete, as eye-tracking studies can only measure foveal gaze and not lesions found by peripheral vision, not to mention that such studies are very difficult to conduct in a clinical setting.

Experts tend to have smaller numbers of NL marks per case than non-experts while maintaining equal or more LL marks per case. As an example, in screening mammography, the number of marks per case (a case is defined as 4-views, two of each breast) that an expert will consider for marking to typically less than three. About 80% on non-diseased cases have no marks. The reason is that because of the low disease prevalence marking too many cases would result in unacceptably high recall rates.

5.3.1 Latent vs. actual marks

To distinguish between suspicious regions that were considered for marking and regions that were actually marked, it is necessary to introduce the distinction between *latent* marks and *actual* marks.

- A *latent* mark is defined as a suspicious region, regardless of whether or not it was marked. A latent mark becomes an *actual* mark if it is marked.
- A latent mark is a latent LL if it is close to a true lesion and otherwise it is a latent NL.
- A non-diseased case can only have latent NLs. A diseased case can have latent NLs and latent LLs.
- If marked, a latent NL is recorded as an actual NL.
- If not marked, a latent NL is an *unobservable event.*
- In contrast, unmarked lesions are observable events – one knows (trivially) which lesions were not marked.

5.3.2 Binning rule

Recall from Section TBA (binary-task-model-z-sample-model) that ROC data modeling requires the existence of a *case-dependent* decision variable, or z-sample z, and case-independent decision thresholds ζ_r, where $r = 0, 1, ..., R_{ROC} - 1$ and R_{ROC} is the number of ROC study bins, and the rule that if $\zeta_r \leq z < \zeta_{r+1}$ the case is rated $r + 1$. Dummy cutoffs are defined as $\zeta_0 = -\infty$ and $\zeta_{R_{ROC}} = \infty$. The z-sample applies to the whole case. To summarize:

$$\left.\begin{aligned} if\,(\zeta_r \leq z < \zeta_{r+1}) \Rightarrow \text{rating} = r + 1 \\ r = 0, 2, ..., R_{ROC} - 1 \\ \zeta_0 = -\infty \\ \zeta_{R_{ROC}} = \infty \end{aligned}\right\} \tag{5.1}$$

Analogously, FROC data modeling requires the existence of a *case and location-dependent* z-sample for each latent mark and *case and location-independent* reporting thresholds ζ_r, where $r = 1, ..., R_{FROC}$ and R_{FROC} is the number of FROC study bins, and the rule that a latent mark is marked and rated r if $\zeta_r \leq z < \zeta_{r+1}$. Dummy cutoffs are defined as $\zeta_0 = -\infty$ and $\zeta_{R_{FROC}+1} = \infty$. For the same numbers of non-dummy cutoffs, the number of FROC bins is one less than the number of ROC bins. For example, 4 non-dummy cutoffs $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ can correspond to a 5-rating ROC study or to a 4-rating FROC study. To summarize:

$$\left.\begin{aligned} if\,(\zeta_r \leq z < \zeta_{r+1}) \Rightarrow \text{rating} = r \\ r = 1, 2, ..., R_{FROC} \\ \zeta_0 = -\infty \\ \zeta_{R_{FROC}+1} = \infty \end{aligned}\right\} \tag{5.2}$$

5.4 FROC notation

Clear notation is vital to understanding this paradigm. The notation needs to account for case and location dependencies of ratings and the distinction between case-level and location-level ground truth. For example, a diseased case can have several regions that are non-diseased and a few diseased regions (the lesions). The notation also has to account for cases with no marks.

FROC notation is summarized in Table 5.1, in which **all marks are latent marks**. The table is organized into three columns, the first column is the row number, the second column has the symbol(s), and the third column has the meaning(s) of the symbol(s).

Table 5.1: FROC notation; all marks refer to latent marks; see comments

Row	Symbol	Meaning
1	t	Case-level truth: 1 for non-diseased and 2 for diseased
2	K_t	Number of cases with case-level truth t
3	$k_t t$	Case k_t in case-level truth t
4	s	Mark-level truth: 1 for NL and 2 for LL
5	$l_s s$	Mark l_s in mark-level truth s
6	$z_{k_t t l_1 1}$	z-sample for case $k_t t$ and mark $l_1 1$
7	$z_{k_2 2 l_2 2}$	z-sample for case $k_2 2$ and mark $l_2 2$
8	R_{FROC}	Number of FROC bins
9	ζ_1	Lowest reporting threshold
10	ζ_r	Other non-dummy reporting thresholds
11	$\zeta_0, \zeta_{R_{FROC}+1}$	Dummy thresholds
12	$N_{k_t t}$	Number of NLs on case $k_t t$
13	$L_{k_2 2}$	Number of lesions on case $k_2 2$
14	$W_{k_2 l_2}$	Weight of lesion $l_2 2$ on case $k_2 2$
15	L_{max}	Maximum number of lesions per case in dataset
16	L_T	Total number of lesions in dataset

5.4.1 Comments on Table 5.1

- Row 1: The case-truth index t refers to the case (or patient), with $t = 1$ for non-diseased and $t = 2$ for diseased cases. As a useful mnemonic, t is for *truth*.

- Row 2: K_t is the number of cases with truth state t; specifically, K_1 is the number of non-diseased cases and K_2 the number of diseased cases.

- Row 3: Two indices $k_t t$ are needed to select case k_t in truth state t. As a useful mnemonic, k is for *case*.

- Rows 4 and 5: For a similar reason, two indices $l_s s$ are needed to select latent mark l_s in location level truth state s, where $s = 1$ corresponds to a latent NL and $s = 2$ corresponds to a latent LL. One can think of l_s as indexing the locations of different latent marks with location-level truth state s. As a useful mnemonic, l is for *location*.

  - $l_1 = \{1, 2, ..., N_{k_t t}\}$ indexes latent NL marks, provided the case has at least one NL mark, and otherwise $N_{k_t t} = 0$ and $l_1 = \emptyset$, the null set.
  - The possible values of l_1 are $l_1 = \{\emptyset\} \oplus \left\{1, 2, ...N_{k_t t}\right\}$. The null set applies when the case has no latent NL marks and $\oplus$ is the "exclusive-or" symbol ("exclusive-or" is used in the English sense: "one or the

other, but not neither nor both"). In other words, l_1 can *either* be the null set or take on values $1, 2, ...N_{k_t t}$.

- Likewise, $l_2 = \left\{1, 2, ..., L_{k_2 2}\right\}$ indexes latent LL marks. Unmarked LLs are assigned negative infinity ratings. The null set notation is not needed for latent LLs.

- Row 6: The z-sample for case $k_t t$ and **latent NL mark** $l_1 1$ is denoted $z_{k_t t l_1 1}$. Latent NL marks are possible on non-diseased and diseased cases (both values of t are allowed). The range of a z-sample is $-\infty < z_{k_t t l_1 1} < \infty$, provided $l_1 \neq \emptyset$; otherwise, it is an *unobservable event*.

- Row 7: The z-sample of a **latent LL** is $z_{k_2 2 l_2 2}$. Unmarked lesions are assigned negative infinity ratings and are observable events. The null-set notation is unnecessary for them.

- Row 8: R_{FROC} is the number of bins in the FROC study.

- Rows 9, 10 and 11: The cutoffs in the FROC study. The lowest threshold is ζ_1. The other non-dummy thresholds are ζ_r where $r = 2, 3, ..., R_{FROC}$. The dummy thresholds are $\zeta_0 = -\infty$ and $\zeta_{R_{FROC}+1} = \infty$.

- Row 12: $N_{k_t t}$ is the total number of latent NL marks on case $k_t t$.

- Row 13: $L_{k_2 2}$ is the number of lesions in diseased case $k_2 2$.

- Row 14: $W_{k_2 l_2}$ is the weight (i.e., clinical importance) of lesion $l_2 2$ in diseased case $k_2 2$. The weights of lesions on a case sum to one: $\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1$.

- Row 15: L_{max} is the maximum number of lesions per case in the dataset.

- Row 16: L_T is the total number of lesions in the dataset.

5.4.2 Discussion: cases with zero latent NL marks

An aspect of FROC data, **that there could be cases with no NL marks, no matter how low the reporting threshold**, has created problems both from conceptual and notational viewpoints. Taking the conceptual issue first, my thinking (prior to 2004) was that as the reporting threshold ζ_1 is lowered, the number of NL marks per case increases almost indefinitely. I visualized this process as each case "filling up" with NL marks [1]. In fact the first modeling of FROC data (Chakraborty, 1989) predicts that, as the reporting threshold is lowered to $\zeta_1 = -\infty$, the number of NL marks per case approaches ∞. However, observed FROC curves end with a finite value of NLs per case. This

[1] I expected the number of NL marks per image to be limited only by the ratio of image size to lesion size, i.e., larger values for smaller lesions.

mismatch between observation and theory is one reason I introduced the radiological search model (RSM) (Chakraborty, 2006b). I will have much more to say about this in a subsequent chapter, but for now I state one prediction (actually an assumption) of the RSM: the number of latent NL marks is a Poisson distributed random integer with a finite value for the mean parameter of the Poisson distribution. This means that the actual number of latent NL marks per case can be 0, 1, 2, .., whose average (over cases) is a finite number. With this background, let us return to the conceptual issue: why does the observer not keep "filling-up" the image with NL marks? The answer is that **the observer can only mark regions that have a non-zero chance of being a lesion**. For example, if the actual number of latent NLs on a particular case is 2, then, as the reporting threshold is lowered, the observer will make at most two NL marks. Having exhausted these two regions the observer will not mark any more regions because there are no more regions to be marked - *all other regions in the image have, in the perception of the observer, zero chance of being a lesion.*

The notational issue is how to handle images with no latent NL marks. Basically it involves restricting summations over cases $k_t t$ to those cases which have at least one latent NL mark, i.e., $N_{k_t t} \neq 0$. This is illustrated in the next section.

5.5 The empirical FROC

The FROC was defined, Chapter 1, as the plot of LLF (along the ordinate) vs. NLF (along the abscissa).

Using the notation of Table 5.1 and assuming binned data[2], then, corresponding to the operating point determined by threshold ζ_r, the FROC abscissa is $\text{NLF}_r \equiv \text{NLF}(\zeta_r)$, the total number of NLs rated $\geq$ threshold ζ_r divided by the total number of cases, and the corresponding ordinate is $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$, the total number of LLs rated $\geq$ threshold ζ_r divided by the total number of lesions:

$$\text{NLF}_r = \frac{n(\text{NLs rated} \geq \zeta_r)}{n(\text{cases})} \tag{5.3}$$

and

$$\text{LLF}_r = \frac{n(\text{LLs rated} \geq \zeta_r)}{n(\text{lesions})} \tag{5.4}$$

The observed operating points correspond to the following values of r:

[2]This is not a limiting assumption: if the data is continuous, for finite numbers of cases, no ordering information is lost if the number of ratings is chosen large enough. This is analogous to Bamber's theorem in Chapter 05, where a proof, although given for binned data, is applicable to continuous data.

$$r = 1, 2, ..., R_{FROC} \tag{5.5}$$

Due to the ordering of the thresholds, i.e., $\zeta_1 < \zeta_2 ... < \zeta_{R_{FROC}}$, higher values of r correspond to lower operating points. The uppermost operating point, i.e., that defined by $r = 1$, is referred to the as the *observed end-point*.

Equations (5.3) and (5.4) is are equivalent to:

$$\text{NLF}_r = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right) \tag{5.6}$$

and

$$\text{LLF}_r = \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{5.7}$$

Each indicator function, $\mathbb{I}()$, yields unity if the argument is true and zero otherwise.

In Eqn. (5.6) $\mathbb{I}\left(N_{k_t t} \neq 0\right)$ ensures that **only cases with at least one latent NL** are counted. Recall that $N_{k_t t}$ is the total number of latent NLs in case $k_t t$. Not including this term would cause the summation over l_1 to be undefined for cases with zero latent NLs. The term $\mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right)$ counts over all NL marks with ratings $\geq \zeta_r$. The three summations yield the total number of NLs in the dataset with z-samples $\geq \zeta_r$ and dividing by the total number of cases yields NLF_r. This equation also shows explicitly that NLs on both non-diseased ($t = 1$) and diseased ($t = 2$) cases contribute to NLF.

In Eqn. (5.7) a summation over t is not needed as only diseased cases contribute to LLF. Analogous to the first indicator function term in Eqn. (5.6), a term like $\mathbb{I}\left(L_{k_2 2} \neq 0\right)$ would be superfluous since $L_{k_2 2} > 0$, as each diseased case must have at least one lesion. The term $\mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right)$ counts over all LL marks with ratings $\geq \zeta_r$. Dividing by L_T, the total number of lesions in the dataset, yields LLF_r.

5.5.1 Definition

The empirical FROC plot connects adjacent operating points $(\text{NLF}_r, \text{LLF}_r)$, including the origin (0,0) and the observed end-point, with straight lines. The area under this plot is the empirical FROC AUC, denoted A_{FROC}.

5.5.2 The origin, a trivial point

Since $\zeta_{R_{FROC}+1} = \infty$ according to Eqn. (5.6) and Eqn. (5.7), $r = R_{FROC} + 1$ yields the trivial operating point (0,0).

5.5.3 The observed end-point and its semi-constrained property

The abscissa of the observed end-point NLF_1, is defined by:

$$\mathrm{NLF}_1 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \tag{5.8}$$

Since each case could have an arbitrary number of NLs, NLF_1 need not equal unity, except fortuitously.

The ordinate of the observed end-point LLF_1, is defined by:

$$\left.\begin{aligned} \mathrm{LLF}_1 =& \frac{\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_1\right)}{L_T} \\ \leq& 1 \end{aligned}\right\} \tag{5.9}$$

The numerator is the total number of lesions that were actually marked. The ratio is the fraction of lesions that are marked, which is ≤ 1.

This is the **semi-constrained property of the observed end-point**, namely, while the observed end-point *ordinate* is constrained to the range (0,1) the corresponding *abscissa* is not so constrained.

5.5.4 Futility of extrapolation outside the observed end-point

To understand this consider the expression for NLF_0, i.e., using Eqn. (5.6) with $r = 0$:

$$\mathrm{NLF}_0 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq -\infty\right) \tag{5.10}$$

The right hand side of this equation can be separated into two terms, the contribution of latent NLs with z-samples in the range $z \geq \zeta_1$ and those in the range $-\infty \leq z < \zeta_1$. The first term yields the abscissa of the observed end-point, Eqn. (5.8). The 2nd term is:

$$\left.\begin{aligned}\text{2nd term} =& \left(\frac{1}{K_1+K_2}\right)\sum_{t=1}^{2}\sum_{k_t=1}^{K_t}\mathbb{I}\left(N_{k_t t}\neq 0\right)\sum_{l_1=1}^{N_{k_t t}}\mathbb{I}\left(-\infty\leq z_{k_t t l_1 1}<\zeta_1\right)\\ =& \frac{\text{unknown number}}{K_1+K_2}\end{aligned}\right\} \tag{5.11}$$

It represents the contribution of unmarked NLs, i.e., latent NLs whose z-samples were below ζ_1. It determines how much further to the right the observer's NLF would have moved, relative to NLF_1, if one could get the observer to lower the reporting criterion to $-\infty$. **Since the observer may not oblige, this term cannot, in general, be evaluated.** Therefore NLF_0 cannot be evaluated. The basic problem is that **unmarked latent NLs represent unobservable events**.

Turning our attention to LLF_0:

$$\left.\begin{aligned}\text{LLF}_0 =& \frac{\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}}\mathbb{I}\left(z_{k_2 2 l_2 2}\geq -\infty\right)}{L_T}\\ =& 1\end{aligned}\right\} \tag{5.12}$$

Unlike unmarked latent NLs, **unmarked lesions can safely be assigned the $-\infty$ rating, because an unmarked lesion is an observable event**. The right hand side of Eqn. (5.12) evaluates to unity. However, since the corresponding abscissa NLF_0 is undefined, one cannot plot this point. It follows that one cannot extrapolate outside the observed end-point.

The formalism should not obscure the fact that the futility of extrapolation outside the observed end-point of the FROC is a fairly obvious property: one does not know how far to the right the abscissa of the observed end-point might extend if one could get the observer to report every latent NL, no matter how low its z-sample.

5.6 The inferred ROC plot

By adopting a sensible rule for converting the zero or more mark-rating data per case to a single rating per case, and commonly the highest rating rule is used [3], it is possible to infer ROC data from FROC mark-rating data.

[3]The highest rating method was used in early FROC modeling in (Bunch et al., 1977) and in (Swensson, 1996), the latter in the context of LROC paradigm modeling.

5.6.1 Inferred-ROC rating

The rating of the highest rated mark on a case, or $-\infty$ if the case has no marks, is defined as the inferred-ROC rating for the case. Inferred-ROC ratings on non-diseased cases are referred to as inferred-FP ratings and those on diseased cases as inferred-TP ratings.

When there is little possibility for confusion, the prefix "inferred" is suppressed. Using the by now familiar cumulation procedure, FP counts are cumulated to calculate FPF and likewise, TP counts are cumulated to calculate TPF.

Definitions:

- $FPF(\zeta)$ = cumulated inferred FP counts with z-sample $\geq$ threshold ζ divided by total number of non-diseased cases.
- $TPF(\zeta)$ = cumulated inferred TP counts with z-sample $\geq$ threshold ζ divided by total number of diseased cases

Definition of ROC plot:

- The ROC is the plot of inferred $TPF(\zeta)$ vs. inferred $FPF(\zeta)$.
- The plot includes a **straight line extension from the observed endpoint to (1,1)**.

The mathematical definition of the ROC follows.

5.6.2 Inferred FPF

The highest z-sample ROC false positive (FP) rating for non-diseased case $k_1 1$ is defined by:

$$\left.\begin{aligned} FP_{k_1 1} &= \max_{l1} \left(z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset \right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{5.13}$$

If the case has at least one latent NL mark, then $l_1 \neq \emptyset$, where $\emptyset$ is the null set, and the first definition applies. If the case has no latent NL marks, then $l_1 = \emptyset$, and the second definition applies. $FP_{k_1 1}$ is the maximum z-sample over all latent marks occurring on non-diseased case $k_1 1$, or $-\infty$ if the case has no latent marks. The corresponding false positive fraction is defined by:

$$\mathrm{FPF}_r \equiv \mathrm{FPF}\left(\zeta_r\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_r \right) \tag{5.14}$$

5.6.3 Inferred TPF

The inferred true positive (TP) z-sample for diseased case $k_2 2$ is defined by:

$$TP_{k_2 2} = \max_{l_1 l_2} \left(\left(z_{k_2 2 l_1 2}, z_{k_2 2 l_2 2} \right) \mid l_1 \neq \emptyset \right) \tag{5.15}$$

or

$$TP_{k_2 2} = \max_{l_2} \left(z_{k_2 2 l_2 2} \right) \mid \left(l_1 = \emptyset \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) \neq -\infty \right) \right) \tag{5.16}$$

or

$$TP_{k_2 2} == -\infty \mid \left(l_1 = \emptyset \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) = -\infty \right) \right) \tag{5.17}$$

Here $\wedge$ is the logical AND operator.

- If $l_1 \neq \emptyset$ then Eqn. (5.15) applies, i.e., one takes the maximum over all ratings, NLs and LLs, whichever is higher, occurring on the diseased case.
- If $l_1 = \emptyset$ and at least one lesion is marked, then Eqn. (5.16) applies, i.e., one takes the maximum over all marked LLs.
- If $l_1 = \emptyset$ and no lesions are marked, then Eqn. (5.17) applies; this represents an unmarked diseased case; the $-\infty$ rating assignment is justified because an unmarked diseased case is an observable event.

The inferred true positive fraction TPF_r is defined by:

$$\mathrm{TPF}_r \equiv \mathrm{TPF}(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I} \left(TP_{k_2 2} \geq \zeta_r \right) \tag{5.18}$$

5.6.4 Definition

The inferred empirical ROC plot connects adjacent points $(\mathrm{FPF}_r, \mathrm{TPF}_r)$, including the origin (0,0), with straight lines plus a straight-line segment connecting the observed end-point to (1,1). Like a real ROC, this plot is constrained to lie within the unit square. The area under this plot is the empirical inferred ROC AUC, denoted A_{ROC}.

5.7 The alternative FROC (AFROC) plot

- Fig. 4 in (Bunch et al., 1977) anticipated another way of visualizing FROC data. I subsequently termed[4] this the *alternative FROC (AFROC)* plot (Chakraborty, 1989).
- The empirical AFROC is defined as the plot of $\text{LLF}(\zeta_r)$ along the ordinate vs. $\text{FPF}(\zeta_r)$ along the abscissa.
- $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$ was defined in Eqn. (5.7).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$ was defined in Eqn. (5.14).

5.7.1 Definition

The empirical AFROC plot connects adjacent operating points $(\text{FPF}_r, \text{LLF}_r)$, including the origin (0,0) and (1,1), with straight lines. The area under this plot is the empirical inferred AFROC AUC, denoted A_{AFROC}.

Key points:

- The ordinates LLF of the FROC and AFROC are identical.
- The abscissa FPF of the ROC and AFROC are identical.
- The AFROC is, in this sense, a hybrid plot, incorporating aspects of both ROC and FROC plots.
- Unlike the empirical FROC, whose observed end-point has the semi-constrained property, **the AFROC end-point is constrained to within the unit square**.

5.7.2 The constrained observed end-point of the AFROC

Since $\zeta_{R_{FROC}+1} = \infty$, according to Eqn. (5.7) and Eqn. (5.14), $r = R_{FROC} + 1$ yields the trivial operating point (0,0). Likewise, since $\zeta_0 = -\infty$, $r = 0$ yields the trivial point (1,1):

$$\left.\begin{aligned} \text{FPF}_{R_{FROC}+1} =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \infty\right) \\ =& 0 \\ \text{LLF}_{R_{FROC}+1} =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq \infty\right) \\ =& 0 \end{aligned}\right\} \tag{5.19}$$

[4] The late Prof. Richard Swensson did not like my choice of the word "alternative" in naming this operating characteristic. I had no idea in 1989 how important this operating characteristic would later turn out to be, otherwise a more meaningful name might have been proposed.

and

$$\left.\begin{aligned} \mathrm{FPF}_0 =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq -\infty\right) \\ =& 1 \\ \mathrm{LLF}_0 =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq -\infty\right) \\ =& 1 \end{aligned}\right\} \tag{5.20}$$

Because every non-diseased case is assigned a rating, and is therefore counted, the right hand side of the first equation in (5.20) evaluates to unity. This is obvious for marked cases. Since each unmarked case also gets a rating, albeit a $-\infty$ rating, it is also counted (the argument of the indicator function in Eqn. (5.20) is true even when the inferred FP rating is $-\infty$).

5.8 The weighted-AFROC (wAFROC) plot

The AFROC ordinate defined in Eqn. (5.7) gives equal importance to every lesion on a case. Therefore, a case with more lesions will have more influence on the AFROC (see TBA Chapter 14 for an explicit demonstration of this fact). This is undesirable since each case (i.e., patient) should get equal importance in the analysis. As with ROC analysis, one wishes to draw conclusions about the population of cases and each case is regarded as an equally valid sample from the population. In particular, one does not want the analysis to be skewed towards cases with greater than the average number of lesions. [5]

Another issue is that the AFROC assigns equal clinical importance to each lesion in a case. Lesion weights were introduced (Chakraborty and Berbaum, 2004) to allow for the possibility that the clinical importance of finding a lesion might be lesion-dependent (Chakraborty and Yoon, 2009). For example, it is possible that a diseased cases has lesions of two types with differing clinical importance; the figure-of-merit should give more credit to finding the more clinically important one. Clinical importance could be defined as the mortality associated with the specific lesion type; these can be obtained from epidemiological studies (DeSantis et al., 2011).

Let $W_{k_2 l_2} \geq 0$ denote the **weight** (i.e., clinical importance) of lesion l_2 in diseased case k_2 (since weights are only applicable to diseased cases, one can, without ambiguity, drop the case-level and location-level truth subscripts, i.e.,

[5] Historical note: I became aware of how serious this issue could be when a researcher contacted him about using FROC methodology for nuclear medicine bone scan images, where the number of lesions on diseased cases can vary from a few to a hundred!

the notation $W_{k_2 2 l_2 2}$ would be superfluous). For each diseased case $k_2 2$ the weights are subject to the constraint:

$$\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1 \tag{5.21}$$

The constraint assures that the each diseased case exerts equal importance in determining the weighted-AFROC (wAFROC) operating characteristic, regardless of the number of lesions in it (see TBA Chapter 14 for a demonstration of this fact).

The weighted lesion localization fraction wLLF_r is defined by (Chakraborty and Zhai, 2016):

$$\text{wLLF}_r \equiv \text{wLLF}\left(\zeta_r\right) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 l_2 2} \geq \zeta_r\right) \tag{5.22}$$

5.8.1 Definition

The empirical wAFROC plot connects adjacent operating points $(\text{FPF}_r, \text{wLLF}_r)$, including the origin (0,0), with straight lines plus a straight-line segment connecting the observed end-point to (1,1). The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC}.

5.9 The AFROC1 plot

Historically the AFROC originally used a different definition of FPF, which is retrospectively termed the AFROC1 plot. Since NLs can occur on diseased cases, it is possible to define an inferred "FP" rating on a *diseased case* as the maximum of all NL ratings on the case, or $-\infty$ if the case has no NLs. The quotes emphasize that this is non-standard usage of ROC terminology: in an ROC study, a FP can only occur on a *non-diseased case*. Since both case-level truth states are allowed, the highest false positive (FP) z-sample for case $k_t t$ is [the "1" superscript below is necessary to distinguish it from Eqn. (5.13)]:

$$\left.\begin{aligned} FP^1_{k_t t} &= \max_{l_1}\left(z_{k_t t l_1 1} \mid l_1 \neq \emptyset\right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{5.23}$$

$FP^1_{k_t t}$ is the maximum over all latent NL marks, labeled by the location index l_1, occurring on case $k_t t$, or $-\infty$ if $l_1 = \emptyset$. The corresponding false positive

fraction FPF_r^1 is defined by [the "1" superscript is necessary to distinguish it from Eqn. (5.14)]:

$$FPF_r^1 \equiv FPF_r^1(\zeta_r) = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(FP_{k_t t}^1 \geq \zeta_r\right) \tag{5.24}$$

Note the subtle differences between Eqn. (5.14) and Eqn. (5.24). The latter counts "FPs" on non-diseased and diseased cases while Eqn. (5.14) counts FPs on non-diseased cases only, and for that reason the denominators in the two equations are different. The advisability of allowing a diseased case to be both a TP and a FP is questionable from both clinical and statistical considerations. However, this operating characteristic can be useful in applications where all cases contain lesions, for example lesion localization plus classification tasks (See Chapter TBA).

5.9.1 Definition

The empirical AFROC1 plot connects adjacent operating points (FPF_r^1, LLF_r), including the origin (0,0) and (1,1), with straight lines. The only difference between AFROC1 and the AFROC plot is in the x-axis. The area under this plot is the empirical AFROC1 AUC, denoted A_{AFROC1}.

5.10 The weighted-AFROC1 (wAFROC1) plot

5.10.1 Definition

The empirical weighted-AFROC1 (wAFROC1) plot connects adjacent operating points (FPF_r^1, wLLF_r), including the origin (0,0) and (1,1), with straight lines. The only difference between it and the wAFROC plot is in the x-axis. The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC1}.

5.11 The EFROC plot

An *exponentially transformed FROC* (EFROC) plot has been proposed (Popescu, 2011) that, like the AFROC, is contained within the unit square. The EFROC inferred FPF is defined by (this represents another way of inferring ROC data, albeit only FPF, from FROC data):

$$FPF_r = 1 - \exp(NLF(\zeta_r)) \tag{5.25}$$

In other words, one computes NLF_r using NLs rated $\geq \zeta_r$ on all cases and then transforms it to FPF_r using the exponential transformation shown. Note that FPF_r so defined is in the range (0,1).

5.11.1 Definition

The empirical EFROC plot connects adjacent operating points $(FPF_r^1, \mathrm{LLF}_r)$, including the origin (0,0) and (1,1), with straight lines. The only difference between it and the AFROC plot is in the x-axis. The area under this plot is the empirical EFROC AUC, denoted A_{EFROC}.

A_{EFROC} has the advantage, compared to A_{FROC}, of being defined by points contained within the unit square. It has the advantage over the AFROC of using all NL ratings, not just the highest rated ones. In my opinion this is a mixed blessing. The effect on statistical power compared to A_{AFROC} has not been studied, but I expect the advantage to be minimal (because the highest rated NL contains more information than a randomly selected NL mark). A disadvantage is that cases with more LLs get more importance in the analysis; this can be corrected by replacing LLF with wLLF, essentially yielding a weighted version of the EFROC AUC. Another disadvantage is that inclusion of NLs on diseased cases causes the EFROC AUC to depend on diseased prevalence. *The EFROC represents the first recognition by someone other than me, of significant limitations of the FROC curve, and that an operating characteristic for FROC data that is completely contained within the unit square is highly desirable.*

5.12 Discussion

TBA This chapter started with the difference between latent and actual marks and the notation to describe FROC data. The notation is used in deriving formulae for FROC, inferred ROC, AFROC, wAFROC, AFROC1, wAFROC1 and EFROC operating characteristics. In each case an area measure was defined. With the exception of the FROC plot, all operating characteristics defined in this chapter are contained in the unit square. Discussion of the preferred operating characteristic is deferred to a subsequent chapter TBA.

5.13 References

Chapter 6

Empirical plot examples

6.1 TBA How much finished

50%

6.2 Introduction

The previous chapter introduced definitions and formulae for the various operating characteristics possible with FROC data. This chapter illustrates these definitions with numerical values and plots. The RSM simulator, introduced in TBA Section (froc-paradigm-preview-rsm), is used to generate FROC datasets under controlled conditions. Structure of the FROC dataset. TBA.

The starting point is the FROC plot.

6.3 Raw FROC/AFROC/ROC plots

Raw plots correspond to the actual simulator generated floating-point ratings, prior to any binning operation. If binning is employed the plots are termed *binned plots*. The FROC plots shown below were generated using the data simulator introduced in Chapter 1. The examples are similar to the population FROC curves shown in that chapter but the emphasis here is on understanding the FROC data structure. To this end smaller numbers of cases, not 20,000 as in the previous chapter, are used. Examples are given using smaller datasets. With a very small dataset, the logic of constructing the plot is more transparent but the operating points are more susceptible to sampling variability. The examples illustrate key points distinguishing the free-response paradigm from ROC. TBA

6.3.1 Code for raw plots

```
seed <- 1;set.seed(seed)
mu <- 1
lambda <- 1
nu <- 1
zeta1 <- -1
K1 <- 5
K2 <- 7
Lmax <- 2
Lk2 <- floor(runif(K2, 1, Lmax + 1))

frocDataRaw <- SimulateFrocDataset(
  mu = mu,
  lambda = lambda,
  nu = nu,
  I = 1,
  J = 1,
  K1 = K1,
  K2 = K2,
  perCase = Lk2,
  zeta1 = zeta1,
  seed = seed
)

p1A <- PlotEmpiricalOperatingCharacteristics(
  dataset = frocDataRaw,
  trts= 1, rdrs = 1, opChType = "FROC",
  legend.position = "NULL")$Plot + ggtitle("A")

p1B <- PlotEmpiricalOperatingCharacteristics(
  dataset = frocDataRaw,
  trts= 1, rdrs = 1, opChType = "AFROC",
  legend.position = "NULL")$Plot + ggtitle("B")

p1C <- PlotEmpiricalOperatingCharacteristics(
  dataset = frocDataRaw,
  trts= 1, rdrs = 1, opChType = "ROC",
  legend.position = "NULL")$Plot + ggtitle("C")

frocDataRaw_1_5_7 <- frocDataRaw # seed 1, K1 = 5, K2 = 7
```

6.3.2 Explanation of the code

Line 1 sets the seed of the random number generator. Lines 2-5 set the simulator parameters $\mu = 1, \lambda = 1, \nu = 1, \zeta_1 = -1$. Briefly, μ determines the separation of two unit variance Gaussians, the one centered at zero determines the z-samples of latent NLs, while the one centered at μ determines the z-samples of latent LLs. λ is the mean parameter of a Poisson distribution determining the number (a random non-negative integer) of latent NLs on each case while ν, the success probability of a binomial distribution, determines the number of latent LLs on each diseased case. A latent `NL` or `LL` is marked if its z-sample $\geq \zeta_1$.

Lines 6-7 set the number of non-diseased cases $K_1 = 5$ and the number of diseased cases $K_2 = 7$.

Line 8 sets the maximum number of lesions per diseased case `Lmax` = 2. Line 9 randomly samples a uniform distribution to obtain the actual number of lesions per diseased case `Lk2`. The following code illustrates the process.

6.3.2.1 Number of lesions per diseased case

```
Lk2
#> [1] 1 1 2 2 1 2 2
sum(Lk2)
#> [1] 11
max(floor(runif(1000, 1, Lmax + 1)))
#> [1] 2
```

This shows that the first two diseased cases have one lesion each, the third and fourth have two lesions each, etc. The total number of lesions in the dataset is 11. The last two lines of the code snippet show that, even with a thousand simulations, the number of lesions per diseased case is indeed limited to `Lmax` = 2.

6.3.2.2 The structure of the FROC dataset

Lines 11-21 uses the function `SimulateFrocDataset` to simulate the dataset object `frocDataRaw`. Its structure is examined next:

```
str(frocDataRaw)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1, 1, 1:12, 1:3] -Inf 0.487 0.738 0.576 -Inf ...
#>   ..$ LL   : num [1, 1, 1:7, 1:2] -Inf -Inf -0.238 1.919 -Inf ...
#>   ..$ LL_IL: logi NA
```

```
#>  $ lesions     :List of 3
#>   ..$ perCase: num [1:7] 1 1 2 2 1 2 2
#>   ..$ IDs    : num [1:7, 1:2] 1 1 1 1 1 ...
#>   ..$ weights: num [1:7, 1:2] 1 1 0.5 0.5 1 ...
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "NA"
#>   ..$ type         : chr "FROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: logi NA
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : chr "1"
#>   ..$ readerID     : chr "1"
```

It is seen to consist of three `list` members: `ratings`, `lesions` and `descriptions`.

6.3.2.3 The structure of the `ratings` member

The `ratings` member is itself a `list` of 3, consisting of `NL` the non-lesion localization ratings, `LL` the lesion localization ratings and `LL_IL` the incorrect localization ratings. The last member is needed for LROC datasets and can be ignored for now.

6.3.2.4 The structure of the `NL` member

```
frocDataRaw$ratings$NL[1,1,,]
#>              [,1]       [,2] [,3]
#>  [1,]        -Inf       -Inf -Inf
#>  [2,]  0.48742905       -Inf -Inf
#>  [3,]  0.73832471       -Inf -Inf
#>  [4,]  0.57578135 -0.3053884 -Inf
#>  [5,]        -Inf       -Inf -Inf
#>  [6,]  1.51178117  0.3898432 -Inf
#>  [7,]  1.12493092 -0.6212406 -Inf
#>  [8,] -0.04493361       -Inf -Inf
#>  [9,] -0.01619026       -Inf -Inf
#> [10,]        -Inf       -Inf -Inf
#> [11,]        -Inf       -Inf -Inf
#> [12,]        -Inf       -Inf -Inf
```

- It is seen to be an array with dimensions [1,1,1:12,1:4].

- Note that all listed ratings are greater than $\zeta_1 = -1$. Unmarked locations are assigned the $-\infty$ rating.

- Case 1, the first non-diseased case, has a single `NL` mark rated $-\infty$ and the remaining 3 locations are filled with $-\infty$s.

- Case 6, the first diseased case, has zero `NL` marks and all 4 locations for it are filled with $-\infty$s. [As seen below, this case actually generated a rating in the first location, but it fell below $\zeta_1 = -1$.]

- Case 11, the sixth diseased case, has three `NL` marks rated $-\infty$, $-\infty$, $-\infty$ and the remaining location for it is $-\infty$. As noted below, this case generated a fourth rating that fell below $\zeta_1 = -1$.

- The first dimension corresponds to the number of modalities, one in this example, the second dimension corresponds to the number of readers, also one in this example.

- The third dimension is the total number of cases, $K_1 + K_2 = 12$ in this example, because NLs are possible on *both* non-diseased and diseased cases.

- The fourth dimension is 4, as the simulator generates, over 12 cases, a maximum of 4 latent NLs per case. This can be demonstrated (see below) by running the preceding code where one temporarily sets $\zeta_1 = -\infty$, which results in all latent marks being marked: one sees that case 11, the sixth diseased case, actually generates 4 NLs, but one of them, at position 4, has rating equal to -1.237538, which is less than $\zeta_1 = -1$, and is consequently not marked in the original example, i.e., this location is assigned a rating of $-\infty$.

```
frocDataRaw1$ratings$NL[1,1,,]
#>              [,1]       [,2]    [,3]
#>  [1,]        -Inf       -Inf    -Inf
#>  [2,]  0.48742905       -Inf    -Inf
#>  [3,]  0.73832471       -Inf    -Inf
#>  [4,]  0.57578135 -0.3053884    -Inf
#>  [5,]        -Inf       -Inf    -Inf
#>  [6,]  1.51178117  0.3898432    -Inf
#>  [7,]  1.12493092 -0.6212406 -2.2147
#>  [8,] -0.04493361       -Inf    -Inf
#>  [9,] -0.01619026       -Inf    -Inf
#> [10,]        -Inf       -Inf    -Inf
#> [11,]        -Inf       -Inf    -Inf
#> [12,]        -Inf       -Inf    -Inf
```

6.3.2.5 The structure of the LL member

```
frocDataRaw$ratings$LL[1,1,,]
#>             [,1]      [,2]
#> [1,]       -Inf      -Inf
#> [2,]       -Inf      -Inf
#> [3,] -0.2375384      -Inf
#> [4,]  1.9189774      -Inf
#> [5,]       -Inf      -Inf
#> [6,]  1.0745650      -Inf
#> [7,]  1.5036080 0.9428932
```

- It is seen to be an array with dimensions `[1,1,1:7,1:2]`.
- The first dimension corresponds to the number of modalities, one in this example, the second dimension corresponds to the number of readers, also one in this example.
- The third dimension is the total number of diseased cases, $K_2 = 7$, because LLs are only possible on diseased cases.
- The fourth dimension is 2, as the maximum number of lesions per diseased case is `Lmax` = 2.
- Note that all listed ratings are greater than $\zeta_1 = -1$.
- Case 1, the first diseased case, has zero `LL` marks and both locations are filled with $-\infty$s.
- Case 2, the second diseased case, has one `LL` mark rated $-\infty$ and the remaining location is $-\infty$.
- Case 7, the seventh diseased case, has two `LL` marks rated 1.503608, 0.9428932 and zero locations with $-\infty$.
- The following output shows that setting $\zeta_1 = -\infty$ does not reveal any more latent LLs.

```
frocDataRaw1$ratings$LL[1,1,,]
#>             [,1]      [,2]
#> [1,]       -Inf      -Inf
#> [2,]       -Inf      -Inf
#> [3,] -0.2375384      -Inf
#> [4,]  1.9189774      -Inf
#> [5,]       -Inf      -Inf
#> [6,]  1.0745650      -Inf
#> [7,]  1.5036080 0.9428932
```

- Lines 23 - 25 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the FROC plot `ggplot` object, which is saved to `p1A`. Note the argument `opChType = "FROC"`, for the desired FROC plot.

- Lines 28 - 31 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the AFROC plot object, which is saved to `p1B`. Note the argument `opChType = "AFROC"`.

- Finally, lines 33 - 35 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the ROC plot object, which is saved to `p1C`. Note the argument `opChType = "ROC"`.

In summary, the code generates FROC, AFROC and ROC plots shown in the top row of Fig. 6.1, labeled A, B and C. The discreteness, i.e., the relatively big jumps between data points, is due to the small numbers of cases. Increasing the numbers of cases to $K_1 = 50$ and $K_2 = 70$ yields the lower row of plots in Fig. 6.1, labeled D, E and F. The fact that the upper row left plot does not seem to match the lower row left plot, especially near NLF = 0.25, is due to sampling variability with few cases.

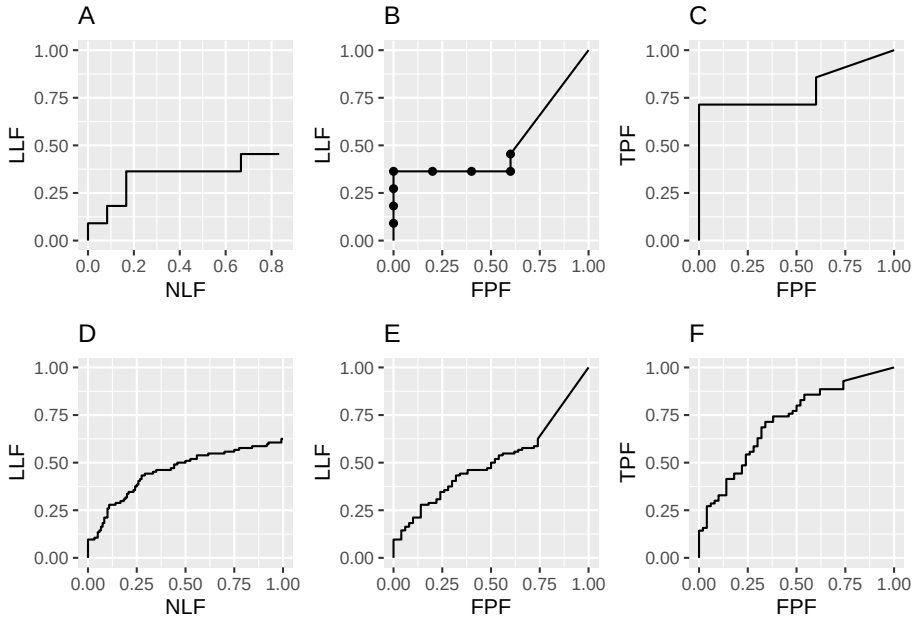


Figure 6.1: Raw FROC, AFROC and ROC plots with seed = 1: Plots A, B, C are for $K_1 = 5$ and $K_2 = 7$ cases while plots D, E, F are for $K_1 = 50$ and $K_2 = 70$ cases.

Fig. 6.1 Raw FROC, AFROC and ROC plots with `seed` = 1: Plots A, B and C are for $K_1 = 5$ and $K_2 = 7$ cases while D, E and F are for $K_1 = 50$ and $K_2 = 70$ cases. Model parameters are $\mu = 1$, $\lambda = 1$, $\nu = 1$ and ζ_1 = -1. The discreteness (jumps) in A, B and C is due to the small number of cases. The decreased discreteness in D, E and F is due to the larger numbers of cases. If the number of cases is increased further, the plots will approach continuous plots, like those shown in Chapter 1. Note that the AFROC (B and E) and ROC plots (C and F), are each contained within unit squares, unlike the semi-constrained FROC plots A and D.

6.3.2.6 Effect of `seed` on raw plots

Shown next are similar plots but this time `seed` = 2.

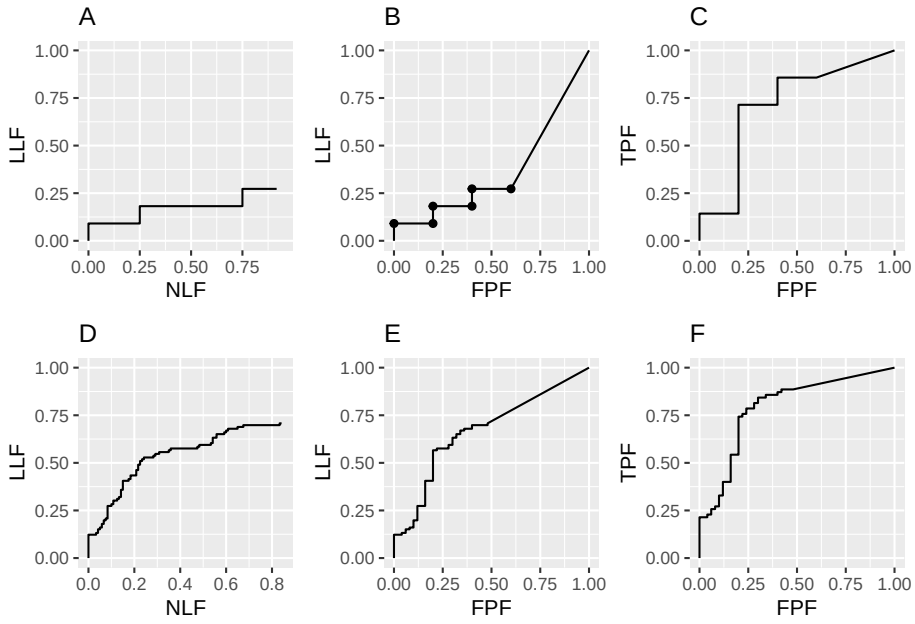


Figure 6.2: Raw FROC, AFROC and ROC plots with seed = 2: Plots A, B, C are for $K_1 = 5$ and $K_2 = 7$ cases while plots D, E, F are for $K_1 = 50$ and $K_2 = 70$ cases.

Fig. 6.2 Raw FROC, AFROC and ROC plots with `seed` = 2: Plots A, B, C are for $K_1 = 5$ and $K_2 = 7$ cases while plots D, E, F are for $K_1 = 50$ and $K_2 = 70$ cases. Model parameters are $\mu = 1$, $\lambda = 1$, $\nu = 1$ and ζ_1 = -1. Note the large variability in the upper row plots as compared to those in Fig. 6.1.

6.3.3 Key differences from the ROC paradigm:

- In a ROC study, each case generates exactly one rating.
- In a FROC study, each case can generate zero or more (0, 1, 2, ...) mark-rating pairs.
- The number of marks per case is a random variable as is the rating of each mark.
- Each mark corresponds to a distinct location on the image and associated with it is a rating, i.e., confidence level in presence of disease at the region indicated by the mark.
- In the ROC paradigm, each non-diseased case generates one FP and each diseased case generates one TP.
- In a FROC study, each non-diseased case can generate zero or more NLs and each diseased case can generate zero or more NLs and zero or more LLs.
- The number of lesions in the case limits the number of LLs.

6.4 The chance level FROC and AFROC

The chance level FROC was addressed in the previous chapter; it is a "flat-liner", hugging the x-axis, except for a slight upturn at large NLF.

Fig. 6.3 shows "near guessing" FROC (plot A) and AFROC (plot B) plots for $\mu = 0.1$. These plots were generated by the code with $\mu = 0.1$, $\lambda = 1$, $\nu = 0.1$, $\zeta_1 =$ -1, $K_1 = 50$, $K_2 = 70$.

The AFROC of a guessing observer is not the line connecting (0,0) to (1,1). A guessing observer will also generate a "flat-liner", but this time the plot ends at FPF = 1, and the straight line extension will be a vertical line connecting this point to (1,1). In the limit $\mu \to 0+$, AFROC-AUC tends to zero.

To summarize, AFROC AUC of a guessing observer is zero. On the other hand, suppose an expert radiologist views screening images and the lesions on diseased cases are very difficult, even for the expert, and the radiologist does not find any of them. Being an expert the radiologist successfully screens out non-diseased cases and sees nothing suspicious in any of them – this is a measure of the expertise of the radiologist, not mistaking variants of normal anatomy for false lesions on non-diseased cases. Accordingly, the expert radiologist does not report anything, and the operating point is "stuck" at the origin. Even in this unusual situation, one would be justified in connecting the origin to (1,1) and claiming area under AFROC is 0.5. The extension gives the radiologist credit for not marking any non-diseased case; of course, the radiologist does not get

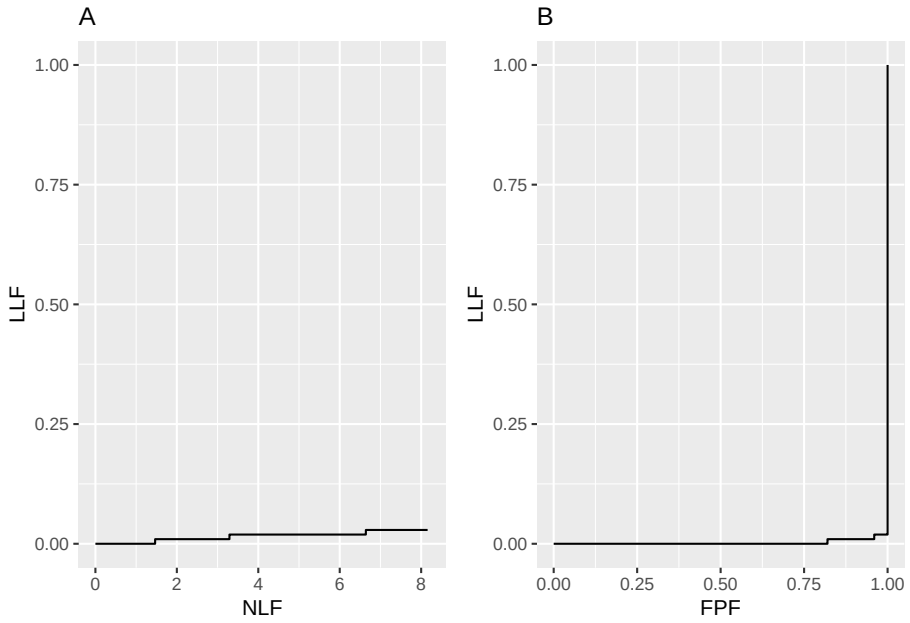


Figure 6.3: Plot A is the near guessing observer's FROC and plot B is the corresponding AFROC for $\mu = 0.01$.

any credit for marking any of the lesions. An even better radiologist, who finds and marks some of the lesions, will score higher, and AFROC-AUC will exceed 0.5. See TBA §17.7.4 for a software demonstration of this unusual situation.

6.5 Location-level "true-negatives"

The quotes are intended to draw attention to confusion that can result when one inappropriately applies ROC terminology to the FROC paradigm. For the 5 / 7 dataset, seed = 1, and reporting threshold set to -1, the first non-diseased case has one `NL` rated $-\infty$. The remaining three entries for this case are filled with $-\infty$.

What really happened is only known if one has access to the internals of the simulator. To the data analyst the following possibilities are indistinguishable:

- Four latent NLs, one of whose ratings exceeded ζ_1, i.e., three location-level "true negatives" occurred on this case.
- Three latent NLs, one of whose ratings exceeded ζ_1, i.e., two location-level "true negatives" occurred on this case.
- Two latent NLs, one of whose ratings exceeded ζ_1, i.e., one location-level "true negative" occurred on this case.
- One latent `NL`, whose rating exceeded ζ_1, i.e., 0 location-level "true negatives" occurred on this case.

The second non-diseased case has one `NL` mark rated 0.4874291 and similar ambiguities occur regarding the number of latent NLs. The third, fourth and fifth non-diseased cases have no marks. All four locations-holders on each of these cases are filled with $-\infty$, which indicates un-assigned values corresponding to either absence of any latent `NL` or presence of one or more latent NLs that did not exceed ζ_1 and therefore did not get marked.

To summarize: absence of an actual `NL` mark, indicated by a $-\infty$ rating, could be due to either (i) non-occurrence of the corresponding latent `NL` or (ii) occurrence of the latent `NL` but its rating did not exceed ζ_1. One cannot distinguish between the two possibilities, as in either scenario, the corresponding rating is assigned the $-\infty$ value and either scenario would explain the absence of a mark.

For those who insist on using ROC terminology to describe FROC data the second possibility would be termed a location level True Negative ("TN"). Their "logic" is as follows: there was the possibility of a `NL` mark, which they term a "FP", but the observer did not make it. Since the complement of a FP event is a TN event, this was a TN event. However, as just shown, one cannot tell if it was a "TN" event or there was no latent event in the first place. Here is the conclusion: there is no place in the FROC lexicon for a location level "TN".

If $\zeta_1 = -\infty$ then all latent marks are actually marked and the ambiguities mentioned above disappear. As noted previously, when this change is made one confirms that there were actually four latent NLs on the sixth diseased case (the eleventh sequential case), but the one rated -1.237538 fell below $\zeta_1 = -1$ and was consequently not marked.

So one might wonder, why not ask the radiologists to report everything they see, no matter now low the confidence level? Unfortunately, that would be contrary to their clinical task, where there is a price to pay for excessive NLs. It would also be contrary to a principle of good experimental design: one should keep interference with actual clinical practice, designed to make the data easier to analyze, to a minimum.

6.6 Binned FROC/AFROC/ROC plots

In the preceding example, continuous ratings data was available and data binning was not employed. Shown next is the code for generating the plots when the data is binned.

6.6.1 Code for binned plots

```
seed <- 1;set.seed(seed)
mu <- 1
zeta1 <- -1
K1 <- 5
K2 <- 7
Lmax <- 2
Lk2 <- floor(runif(K2, 1, Lmax + 1))

frocDataRaw <- SimulateFrocDataset(
  mu = mu,
  lambda = lambda,
  nu = nu,
  I = 1,
  J = 1,
  K1 = K1,
  K2 = K2,
  perCase = Lk2,
  zeta1 = zeta1,
  seed = seed
)

frocDataBinned <- DfBinDataset(
```

```
23   frocDataRaw,
24   desiredNumBins = 5,
25   opChType = "FROC")
26
27 p4A <- PlotEmpiricalOperatingCharacteristics(
28   dataset = frocDataBinned,
29   trts= 1, rdrs = 1, opChType = "FROC",
30   legend.position = "NULL")$Plot + ggtitle("A")
31
32 p4B <- PlotEmpiricalOperatingCharacteristics(
33   dataset = frocDataBinned,
34   trts= 1, rdrs = 1, opChType = "AFROC",
35   legend.position = "NULL")$Plot + ggtitle("B")
36
37 p4C <- PlotEmpiricalOperatingCharacteristics(
38   dataset = frocDataBinned,
39   trts= 1, rdrs = 1, opChType = "ROC",
40   legend.position = "NULL")$Plot + ggtitle("C")
```

This is similar to the code for the raw plots except that at lines 21-24 we have used the function `DfBinDataset` to bin the raw data `frocDataRaw` and the binned data is saved to `frocDataBinned`, which is used in the subsequent plotting routines. Note the arguments `desiredNumBins` and `opChType`. The binning function needs to know the desired number of bins (set to 5 in this example) and the operating characteristic that the binning is aimed at (here set to "FROC").

6.6.2 Effect of `seed` on binned plots

Shown next are corresponding plots with `seed` = 2.

6.7 Structure of the binned data

```
str(frocDataBinnedSeed1$ratings$NL)
#>  num [1, 1, 1:120, 1:4] -Inf 4 2 3 -Inf ...
table(frocDataBinnedSeed1$ratings$NL)
#>
#> -Inf    1    2    3    4
#>  376   35   30   23   16
sum(as.numeric(table(frocDataBinnedSeed1$ratings$NL)))
#> [1] 480
```

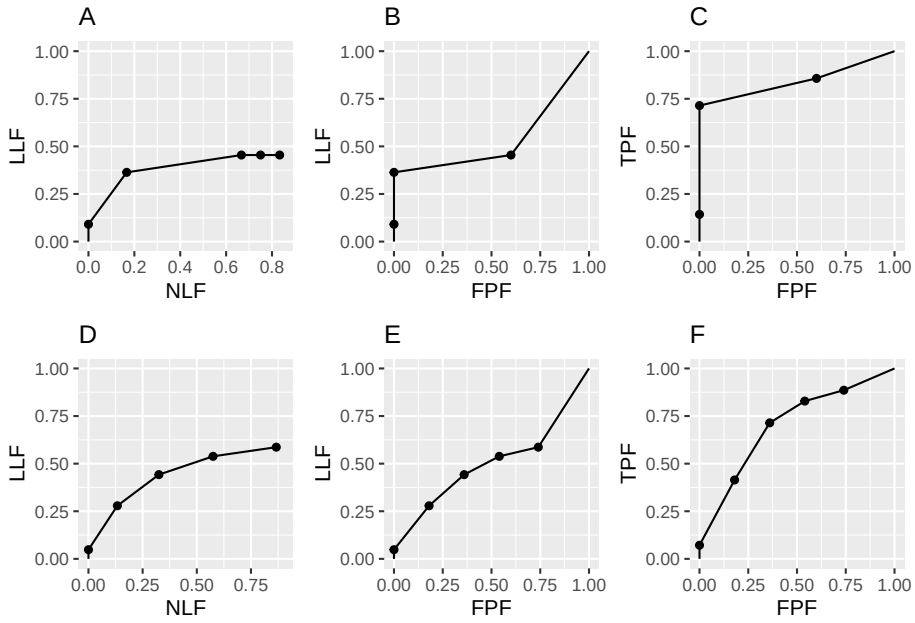


Figure 6.4: Binned FROC, AFROC and ROC plots with seed = 1: Plots A, B, C are for $K_1 = 5$ and $K_2 = 7$ cases while plots D, E, F are for $K_1 = 50$ and $K_2 = 70$ cases

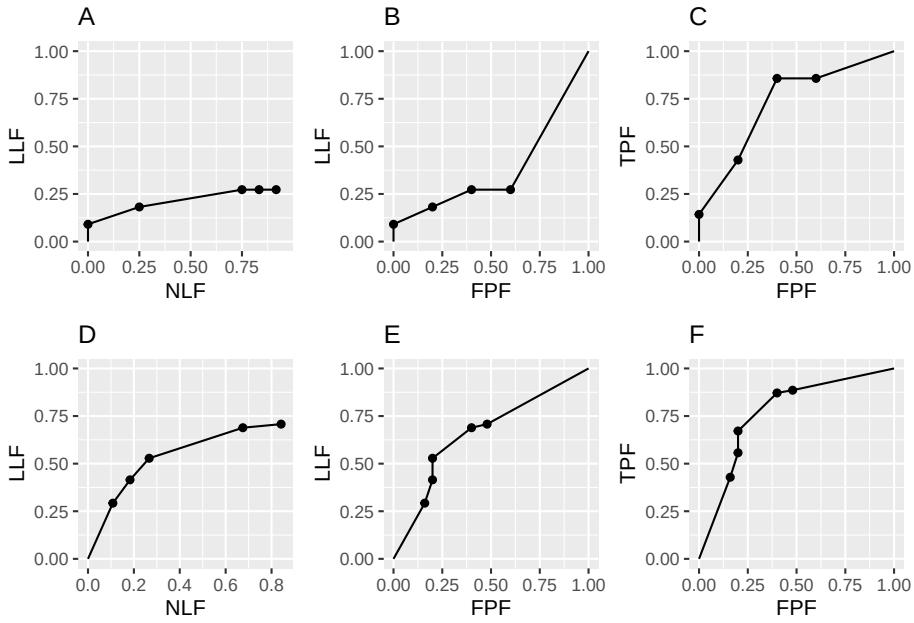


Figure 6.5: Binned FROC, AFROC and ROC plots with seed = 2: Plots A, B, C are for $K_1 = 5$ and $K_2 = 7$ cases while plots D, E, F are for $K_1 = 50$ and $K_2 = 70$ cases

- The `table()` function converts an array into a counts table.
- There are 120 x 4 = 480 elements in the `NL` array to be "tabled".
- From the output one sees that there are 378 entries in the `NL` array that equal $-\infty$, 50 that equal 1, 15 that equal 2, 12 that equal 3, and 25 that equal 4 (none of the NLs were binned into the rating "5" category). These sum to 480 (see code output above).
- Because the fourth dimension of the `NL` array is determined by cases with the *most* NLs, on the *unknown number* (to the data analyst) of cases with *fewer* NLs, this dimension is "padded" with negative-infinities.

- Because of the unknown number of negative-infinity paddings, one does not know how many of the 378 *observed* negative-infinities are *actually* latent NLs. The *actual* number of latent NLs could be considerably smaller - and the number of *marked* NLs even smaller - as this is determined by those latent NLs whose z-samples $\geq \zeta_1$. Notice that in the special case $\zeta_1 = -\infty$ the observer marks all latent NL, in which case the observed count equal the actual count.

```
str(frocDataBinnedSeed1$ratings$LL)
#>  num [1, 1, 1:70, 1:2] 3 4 4 4 3 ...
table(frocDataBinnedSeed1$ratings$LL)
#>
#> -Inf    1    2    3    4    5
#>   79    5   10   17   24    5
sum(as.numeric(table(frocDataBinnedSeed1$ratings$LL)))
#> [1] 140
sum(Lk2Seed1)
#> [1] 104
sum(Lk2Seed1) - sum(as.vector(table(frocDataBinnedSeed1$ratings$LL))[2:6])
#> [1] 43
```

- The `LL` array contains 70 x 2 = 140 values to be "tabled".
- From the output one sees that there are 78 entries in the `LL` array that equal $-\infty$, 10 entries that equal 1, 5 entries that equal 2, 8 entries that equal 3, 35 entries that equal 4, and 4 entries that equal 5. These sum to 140, the product of the lengths of the third and fourth dimensions of the `LL` array.
- The number of negative-infinity counts is 78. This is smaller than 140 because, of the varying numbers of lesions, some of the location-holders are filled with negative infinities.

- The *known* total number of lesions – each of which contributes a latent `LL` – is 104, see 2nd last line of above code output.

- Summing the `LL` counts in bins 1 through 5 (corresponding to table

columns 2-6, since column 1 applies to the negative-infinities) and subtracting from the total number of lesions one gets: 104 – (10+5+8+35+4) = 104 – 62 = 42, see last line of above code output.

- Therefore, the number of unmarked lesions is 42. The listed value (78) is an overestimate because it includes the $-\infty$ counts from the fourth dimension negative-infinity "padding" of the `LL` array.

6.8 Summary

The preceding detailed example illustrates a key point: *The total number of latent `NL`s in the dataset is generally unknown to the data analyst, unlike the total number of latent `LL`s, which is known.* The only exception to this rule is if $\zeta_1 = -\infty$, in which case the observer marks all latent `NL` (and `LL`) sites.

6.9 Discussion

TBA

6.10 References

Chapter 7

FROC vs. wAFROC

7.1 TBA How much finished

50% Need to replace simulation values with analytical values

7.2 Introduction

In the medical imaging context the FROC curve, which was introduced in (Bunch et al., 1977), has been widely used for evaluating performance in the free-response paradigm, particularly in CAD algorithm development. Typically CAD researchers report sensitivity at a stated value of false positives per image, i.e., they report a *pair* of values. (TBA) From basic ROC analysis, see Section TBA (binary-task-model-beam-study), we know that a scalar FOM is preferable to reporting a pair of values. This chapter recommends adoption of the area under the wAFROC as the preferred scalar figure of merit in lieu of sensitivity / false positives per image pairs. operating characteristic in assessing performance in the free-response paradigm, and details simulation-based studies supporting this recommendation.

7.3 FROC vs. wAFROC

Recall, from TBA Section (froc-paradigm-preview-rsm), that the RSM is defined by parameters μ, λ, ν and ζ_1. This section examines RSM-predicted TBA analytical FROC, wAFROC and ROC panels for two observers denoted R1 and R2. The former could be an algorithmic observer while the latter could be a radiologist. For typical threshold ζ_1 parameters, three types of situations are

considered: R2 has moderately better performance than R1, R2 has much better performance than R1 and R2 has slightly better performance than R1. For each type of simulation pairs of FROC, wAFROC and ROC curves are shown, one for each observer. Finally the simulations and panels are repeated for hypothetical R1 and R2 observers who report all suspicious regions, i.e., $\zeta_1 = -\infty$ for each observer. Both R1 and R2 observers share the same λ, ν parameters, and the only difference between them is in the μ and ζ_1 parameters.

7.3.1 Moderate difference in performance

```
1  source(here("R/CH13-CadVsRadPlots/CadVsRadPlots.R"))
2
3  nu <- 1
4  lambda <- 1
5  K1 <- 500
6  K2 <- 700
7  mu1 <- 1.0
8  mu2 <- 1.5
9  zeta1_1 <- -1
10 zeta1_2 <- 1.5
11 Lmax <- 2
12 seed <- 1
13
14 ret <- do_one_figure (
15   seed, Lmax, mu1,
16   mu2, lambda, nu, zeta1_1, zeta1_2, K1, K2)
17
18 froc_plot_1A <- ret$froc_plot_A
19 wafroc_plot_1B <- ret$wafroc_plot_B
20 roc_plot_1C <- ret$roc_plot_C
21 froc_plot_1D <- ret$froc_plot_D
22 wafroc_plot_1E <- ret$wafroc_plot_E
23 roc_plot_1F <- ret$roc_plot_F
24 wafroc_1_1B <- ret$wafroc_1_B
25 wafroc_2_1B <- ret$wafroc_2_B
26 roc_1_1C <- ret$roc_1_C
27 roc_2_1C <- ret$roc_2_C
28 wafroc_1_1E <- ret$wafroc_1_E
29 wafroc_2_1E <- ret$wafroc_2_E
30 roc_1_1F <- ret$roc_1_F
31 roc_2_1F <- ret$roc_2_F
```

The λ and ν parameters are defined at lines 3 and 4 of the preceding code: $\lambda = \nu = 1$. The number of simulated cases is defined, lines 5-6, by $K_1 = 500$

and $K_2 = 700$. The simulated R1 observer μ parameter is defined at line 7 by $\mu_1 = 1$ and that of the simulated R2 observer is defined at line 8 by $\mu_2 = 1.5$. Based on these choices one expect R2 to be moderately better than R1. The corresponding threshold parameters are (lines 9 -10) $\zeta_1 = -1$ for R1 and $\zeta_1 = 1.5$ for R2. The maximum number of lesions per case is defined at line 11 by `Lmax` = 2. The actual number of lesions per case is determined determined by random sampling within the helper function `do_one_figure()` called at lines 14-16. This function returns a large list `ret`, whose contents are as follows:

- `ret$froc_plot_A`: a pair of FROC panels for the thresholds specified above, a red panel labeled "R: 1" corresponding to R1 and a blue panel labeled "R: 2" corresponding to R2. These are shown in panel A.
- `ret$wafroc_plot_B`: a pair of wAFROC panels, similarly labeled. These are shown in panel B.
- `ret$roc_plot_C`: a pair of ROC panels, similarly labeled. These are shown in panel C.
- `ret$froc_plot_D`: a pair of FROC panels for the both thresholds at $-\infty$. These are shown in panel D.
- `ret$froc_plot_E`: a pair of wAFROC panels for the both thresholds at $-\infty$. These are shown in panel E.
- `ret$froc_plot_F`: a pair of ROC panels for the both thresholds at $-\infty$. These are shown in panel F.
- `ret$wafroc_1_B`: the wAFROC AUC for R1, i.e., the area under the curve labeled "R: 1" in panel B.
- `ret$wafroc_2_B`: the wAFROC AUC for R2, i.e., the area under the curve labeled "R: 2" in panel B.
- `ret$roc_1_C`: the ROC AUC for R1, i.e., the area under the curve labeled "R: 1" in panel C.
- `ret$roc_2_C`: the ROC AUC for R2, i.e., the area under the curve labeled "R: 2" in panel C.
- `ret$wafroc_1_E`: the wAFROC AUC for R1, i.e., the area under the curve labeled "R: 1" in panel E.
- `ret$wafroc_2_E`: the wAFROC AUC for R2, i.e., the area under the curve labeled "R: 2" in panel E.
- `ret$roc_1_F`: the ROC AUC for R1, i.e., the area under the curve labeled "R: 1" in panel F.
- `ret$roc_2_F`: the ROC AUC for R2, i.e., the area under the curve labeled "R: 2" in panel F.

The coordinates of the end-point of the R1 FROC in panel A are (0.826, 0.590). Those of the R2 FROC curve in A are (0.049, 0.398). The FROC for the R1 observer extends to much larger NLF values while that for the R2 observer is relatively short and steep. One suspects the R2 observer is performing better than R1: he is better at finding lesions and producing fewer NLs, both of which are desirable characteristics, but he is adopting a too-strict reporting

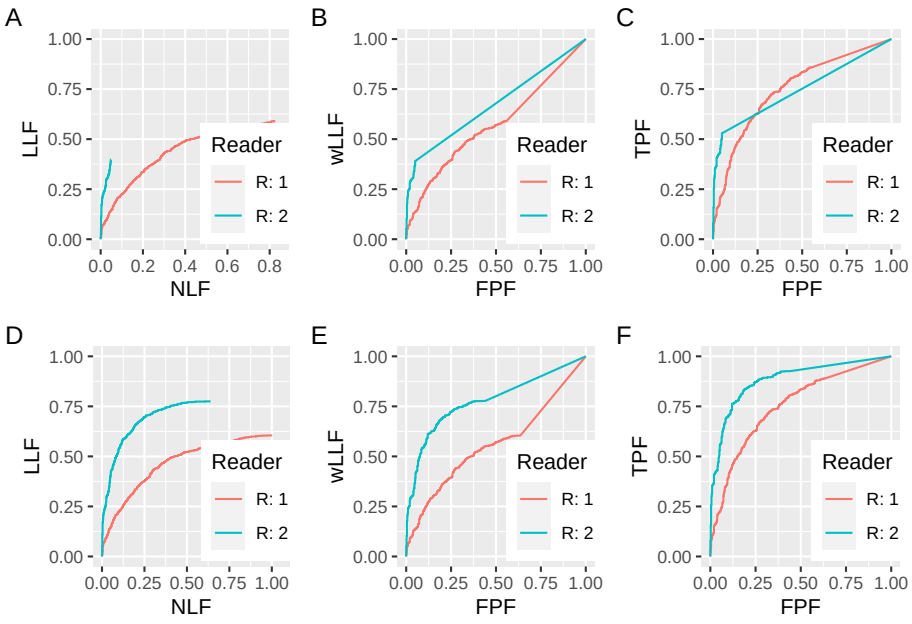


Figure 7.1: Plots A and D: FROC curves for the R1 and R2 observers; B and E are corresponding wAFROC curves and C and F are corresponding ROC curves. All curves in this plot are for $\lambda = \nu = 1$. All RAD_1 curves are for $\mu = 1$ and all RAD_2 curves are for $\mu = 1.5$. For panels A, B and C, $\zeta_1 = -1$ for R1 and $\zeta_1 = 1.5$ for R2. For panels D, E and F, $\zeta_1 = -\infty$ for R1 and R2.

criterion. If he could be induced to relax the threshold and report more NLs, his LLF would exceed that of the R1 observer while still maintaining a lower NLF. However, as this involves a subjective extrapolation, it is not possible to objectively quantify this from the FROC curves. The basic issue is the lack of a common NLF range for the two panels. If a common NLF range is "forced", for example defined as the common NLF range 0 to 0.0492, where both curves contribute, it would ignore most NLs from the R1 observer.

Algorithm developers typically quote LLF at a specified NLF. According to the two panels in A, the R2 observer is better if the NLF value is chosen to less than 0.0492 - this is the maximum NLF value for the R2 curve in A - but there is no basis for comparison for larger values of NLF (because the R2 observer does not provide any data beyond the observed end-point). A similar problem was encountered in ROC analysis when comparing a pair of sensitivity-specificity values, where, given differing choices of thresholds, ambiguous results can be obtained, see Section TBA (binary-task-model-beam-study). Indeed, this was the rationale for using AUC under the ROC curve as an unambiguous measure of performance.

Plot B shows wAFROC curves for the same datasets whose FROC curves are shown in panel A. **The wAFROC is contained within the unit square, a highly desirable characteristic, which solves the lack of a common NLF range problem with the FROC.** The wAFROC AUC under the R2 observer is visibly greater than that for the R1 observer, even though – due to his higher threshold – his AUC estimate is actually biased downward (because the R2 observer is adopting a high threshold, his LLF_{max} is smaller than it would have been with a lower threshold, and consequently the area under the large straight line segment from the uppermost non-trivial operating point to (1,1) is smaller). AUCs under the two wAFROC panels in B are 0.5731 for R1 and 0.6737 for R2.

Plot C shows ROC curves. Since the curves cross, it is not clear which has the larger AUC. AUCs under the two curves in C are 0.7499 for R1 and 0.7453 for R2, which are close, but here is an example where the ordering given by the wAFROC is opposite to that given by the ROC.

Plots D, E and F correspond to A, B and C with this important difference: the two threshold parameters are set to $-\infty$. The coordinates of the end-point of the R1 FROC in panel D are (1.002, 0.605). Those of the R2 FROC in panel D are (0.639, 0.775). The R2 observer has higher LLF at lower NLF, and there can be no doubt that he is better. Panels E and F confirm that R2 is actually the better observer *over the entire FPF range*. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.7780 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.8826 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC. This is because the probability of a lesion localization (case is declared positive *and* a lesion is correctly localized) is smaller than the probability of a true positive (case is declared positive). In

other words, the ROC is everywhere above the wAFROC.

7.3.2 Large difference in performance

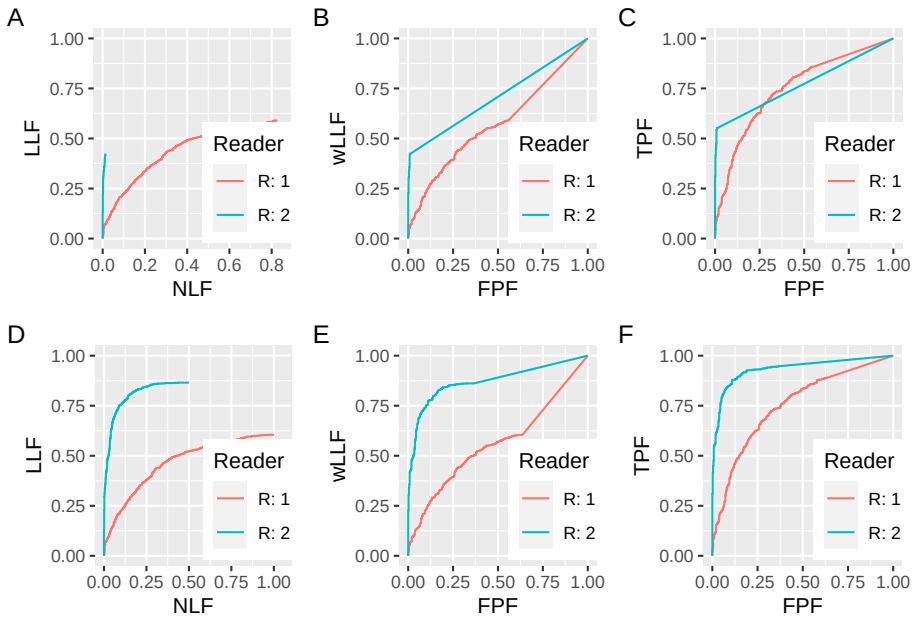


Figure 7.2: Similar to preceding figure but with the following changes. All RAD_2 curves are for $\mu = 2$ and for panels A, B and C $\zeta_1 = 2$ for R2.

In Fig. 7.2 panel A, the R1 parameters are the same as in Fig. 7.1, but the R2 parameters are $\mu_2 = 2$ and $\zeta_1 = +2$. Doubling the separation parameter over that of R1 ($\mu_1 = 1$) has a huge effect on performance. The end-point coordinates of the FROC for R1 are (0.826, 0.590). The end-point coordinates of the FROC for R2 are (0.015, 0.421). The common NLF region defined by NLF = 0 to NLF = 0.0150 *would exclude almost all of the marks made by R1.* The wAFROC panels in panel B show the markedly greater performance of R2 over R1 (the AUCs are 0.5731 for R1 and 0.7075 for R2). The inter-reader difference is larger (compared to Fig. 7.1 panel B), despite the greater downward bias working against the R2 observer. Panel C shows ROC panels for the two observers. Although the curves cross, it is evident that R2 has the greater AUC. The AUCs are 0.7499 for R1 and 0.7722 for R2.

Plots D, E and F correspond to A, B and C with the difference that the two threshold parameters are set to $-\infty$. The coordinates of the end-point of the R1 FROC in panel D are OpPtStr(nlf_1_2D, llf_1_2D). `Those of the R2`

`FROC in panel D are OpPtStr(nlf_2_2D, llf_2_2D)`. The R2 observer has higher LLF at lower NLF, and there can be no doubt that he is better. Panels E and F confirm that R2 is actually the better observer *over the entire FPF range*. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.8720 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.9343 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC.

7.3.3 Small difference in performance and identical thresholds

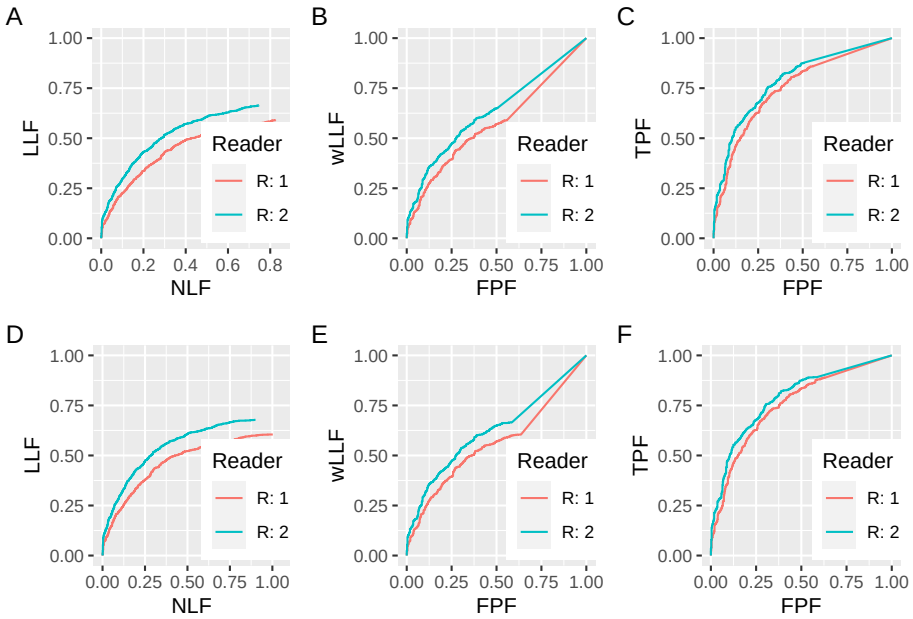


Figure 7.3: Similar to preceding figure but with the following changes. All RAD_2 curves are for $\mu = 1.1$ and for panels A, B and C, $\zeta_1 = -1$ for R2.

The final example, Fig. 7.3 shows that *when there is a small difference in performance*, there is less ambiguity in using the FROC as a basis for measuring performance. The R1 parameters are the same as in Fig. 7.1 but the R2 parameters are $\mu = 1.1$ and $\zeta_1 = -1$. In other words, the μ parameter is 10% larger and the thresholds are identical. This time there is much more common NLF range overlap in panel A and one is counting most of the marks for the R1 reader. The end-point coordinates of the FROC for R1 are (0.826, 0.590). The end-point coordinates of the FROC for R2 are ((0.746, 0.664). The common

Table 7.1: Summary of R1 simulations: A refers to panel A, B refers to panel B, etc.

wAFROC-B	wAFROC-E	ROC-C	ROC-F
0.5731	0.5605	0.7499	0.7513

NLF region defined by NLF = 0 to NLF = 0.7458 includes almost all of the marks made by R1. The wAFROC panels in panel B show the slight greater performance of R2 over R1 (the AUCs are 0.5731 for R1 and 0.6341 for R2). Panel C shows ROC panels for the two observers. Although the curves cross, it is evident that R2 has the greater AUC. The AUCs are 0.7499 for R1 and 0.7722 for R2.

Plots D, E and F correspond to A, B and C with the difference that the two threshold parameters are set to $-\infty$. The coordinates of the end-point of the R1 FROC in panel D are ((1.002, 0.605). Those of the R2 FROC in panel D are ((0.901, 0.678). Panels E and F confirm that R2 is actually the better observer over the entire FPF range. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.6238 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.7857 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC.

7.4 Summary of simulations

The following tables summarize the numerical values from the plots in this chapter. Table 7.1 refers to the R1 observer, and Table 7.2 refers to the R2 observer.

7.4.1 Summary of R1 simulations

- The first column is labeled "wAFROC-B", meaning the R1 wAFROC AUC in panel B, which are identical for the three figures (one may visually confirm that the red curves in panels A, B ad C in the three figures are identical; likewise for the red curves in panels D, E and F).
- The second column is labeled "wAFROC-E", meaning the R1 wAFROC AUC in panel E, which are identical for the three figures.
- The third column is labeled "ROC-C", meaning the R1 ROC AUC in panel C, which are identical for the three figures.
- The fourth column is labeled "ROC-F", meaning the R1 ROC AUC in panel F, which are identical for the three figures.

Table 7.2: Summary of R2 simulations: Fig refers to the figure number in this chapter, A refers to panel A, B refers to panel B, etc.

Fig	wAFROC-B	wAFROC-E	ROC-C	ROC-F
1	0.6737	0.778	0.7453	0.8826
2	0.7075	0.872	0.7722	0.9343
3	0.6341	0.6238	0.7868	0.7857

7.4.2 Summary of R2 simulations

- The first column refers to the figure number, for example, "1" refers to Fig. 7.1, "2" refers to Fig. 7.2, and "3" refers to Fig. 7.3.
- The second column is labeled "wAFROC-B", meaning the R2 wAFROC AUC corresponding to the blue curve in panel B.
- The third column is labeled "wAFROC-E", meaning the R2 wAFROC AUC corresponding to the blue curve in panel E.
- The fourth column is labeled "ROC-C", meaning the R2 ROC AUC corresponding to the blue curve in panel C.
- The fifth column is labeled "ROC-F", meaning the R2 ROC AUC corresponding to the blue curve in panel F.

7.4.3 Comments

- For the same figure label the R1 panels are identical in the three figures. This is the reason why Table 7.1 has only one row. A *fixed* R1 dataset is being compared to *varying* R2 datasets.
- The first R2 dataset, Fig. 7.1 A, B or C, might be considered representative of an average radiologist, the second one, Fig. 7.2 A, B or C, is a super-expert and the third one, Fig. 7.3 A, B or C, is only nominally better than R1.
- Plots D, E and F are for hypothetical R1 and R2 observers that report *all* suspicious regions. The differences between A and D are minimal for the R1 observer, but marked for the R2 observer. Likewise for the differences between B and E.

7.5 Effect size comparison

- The effect size is defined as the AUC – calculated using either wAFROC or ROC – difference between RDR-2 and RDR-1 for the same figure. For example, for Fig. 7.2 and the wAFROC AUC effect size, one takes the difference between the AUCs under the R2 (blue) minus R1 (red) curves in panel B.

Table 7.3: Effect size comparions for R1 simulations: Fig refers to the figure number in this chapter.

Fig	ES-wAFROC	ES-ROC
1	0.1006	-0.004654
2	0.1344	0.02222
3	0.061	0.03685

- In all three figures the wAFROC effect size (ES) is larger than the corresponding ROC effect size.
- For Fig. 7.1 panels B and C:
  - The wAFROC effect size is 0.1006,
  - The ROC effect size is -0.0047.
- For Fig. 7.2 panels B and C:
  - The wAFROC effect size is 0.1344,
  - The ROC effect size is 0.0222.
- For Fig. 7.3 panels B and C:
  - The wAFROC effect size is 0.0610,
  - The ROC effect size is 0.0369.

These results are summarized in Table 7.3.

Since effect size enters as the *square* in sample size formulas, wAFROC yields greater statistical power than ROC. The "small difference" example, corresponding to row number 2, is more typical of modality comparison studies where the modalities being compared are only slightly different. In this case the wAFROC effect size is about twice the corresponding ROC value - see chapter on FROC sample size TBA.

7.6 Performance depends on ζ_1

Consider the wAFROC AUCs for the R2 curves in Fig. 7.2 panels B and E. The wAFROC AUC for R2 in panel B is 0.7075 while that for R2 in panel E is 0.8720. The only difference between the simulation parameters for the two curves are $\zeta_1 = 2$ for panel B and $\zeta_1 = -\infty$ for panel E. Clearly wAFROC AUC depends on the value of ζ_1.

A similar result applies when considering the ROC curves in Fig. 7.2 panels C and F. The ROC AUC for R2 in panel C is 0.7722 while that for R2 in panel F is 0.9343. Clearly ROC AUC also depends on the value of ζ_1.

The reason is that in panels B and C the respective AUCs are depressed due to high value of threshold parameter. The (very good) radiologist is seriously under-reporting and choosing to operate near the origin of a steep wAFROC/ROC curve. It as as if in an ROC study the reader is giving too much importance to specificity and therefore not achieving higher sensitivity.

Since performance depends on threshold, this opens up the possibility of optimizing performance by finding the threshold that maximizes AUC. This is the subject of the next chapter.

7.7 Discussion

7.8 References

Chapter 8

Meanings of FROC figures of merit

8.1 TBA How much finished

50%

8.2 Introduction

Chapter 5 focused on empirical plots possible with FROC data, for example, the FROC, AFROC, wAFROC and inferred ROC plots. Expressions were given for computing *operating points* for each plot from z-samples. Because of the ambiguity in ordering the two values associated with each operating points (e.g., sensitivity-specificity pairs in ROC plots), operating points should not be used as figures of merit. Rather one should use *area measures* derived from operating characteristics. This chapter is devoted to a number of such measures for FROC data.

A generic empirical area under a plot is denoted A_{oc}, where the "oc" subscript denotes the applicable operating characteristic. For example, the area under the empirical wAFROC is denoted A_{wAFROC}. Calculating areas from operating points using planimetry or geometry is tedious at best. *Needed are formulas for calculating them directly from ratings.* In this sense this chapter is analogous to Chapter TBA (empirical-auc) where it was shown that the area under the empirical ROC plot A_{ROC} equaled the Wilcoxon statistic calculated directly from the ratings, i.e., the Bamber theorem (Bamber, 1975).

I make a distinction between *empirical AUC under a plot*, i.e., an area measure, and a *FOM-statistic*, generically denoted θ, that can be computed directly from

the ratings. While any function of the ratings is a possible FOM-statistic, whether it is useful depends upon whether it can be related to the area under an operating characteristic. This chapter derives formulas for FOM-statistics θ_{oc}, which yield the same values as the areas A_{oc} under the corresponding empirical operating characteristics. The meanings of these FOM-statistics are discussed (Chakraborty and Zhai, 2016).

Here is the organization of the chapter.

- Expressions for the empirical AFROC FOM-statistic θ_{AFROC} and the empirical weighted-AFROC FOM-statistic θ_{wAFROC} are presented and their limiting values for chance-level and perfect performances are explored.

- Two important theorems are stated, whose proofs are in [TBA Online Appendix 14.A].

- The first theorem proves the equality between the empirical wAFROC FOM-statistic θ_{wAFROC} and the area A_{wAFROC} under the empirical wAFROC plot. [A similar equality applies to the empirical AFROC FOM-statistic θ_{AFROC} and the area A_{AFROC} under the empirical AFROC plot.]

- The second theorem derives an expression for the area under the straight-line extension of the wAFROC from the observed end-point to (1,1), and explains why it is essential to include this area.

- A small simulated-dataset is used to illustrate how NL and LL ratings and lesion weights determine the wAFROC empirical plot.

- It demonstrates that the wAFROC gives equal importance to all diseased cases, a desirable statistical characteristic.

- Corresponding results, but ignoring the weights, show that the AFROC gives excessive importance to cases with more lesions.

- A physical interpretation of the AUC or FOM-statistics is given. It shows explicitly how the ratings comparisons implied in FOM-statistic properly credit and penalize the observer for correct and incorrect decisions, respectively. The probabilistic meanings of the AFROC and wAFROC AUCs are given.

- Detailed derivations of FOM-statistics, applicable to the areas under the empirical FROC plot, the AFROC1 and wAFROC1 plots are not given. Instead, the results for all plots are summarized in [TBA Online Appendix 14.C], which shows that the definitions "work", i.e., the FOM-statistics yield the correct areas as determined by numerical integration of the relevant curves.

8.3 Empirical AFROC FOM-statistic

A_{AFROC} was defined in 5.8 as the area under the empirical AFROC. The corresponding FOM-statistic θ_{AFROC} is defined as follows: one calculates the rating of the highest rated NL mark $\text{FP}_{k_1 1}$ on each non-diseased case $k_1 1$ (or $-\infty$ if the case has no NL marks) and compares it to each LL rating using the kernel function $\psi(x, y)$ defined in Eqn. TBA (eq:empirical-auc-PsiFunction) [1]. A summation is performed over all cases and all lesions.

The highest rating $\text{FP}_{k_1 1}$ on non-diseased case $k_1 1$ is defined as:

$$\left.\begin{aligned} \text{FP}_{k_1 1} &= \max_{l_1} \left(z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset \right) \\ \text{FP}_{k_1 1} &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{8.1}$$

If the case has at least one latent NL mark, then $l_1 \neq \emptyset$, where $\emptyset$ is the null set, and the first definition applies. If the case has no marks, then $l_1 = \emptyset$, and the second definition applies.

The following equation sums over all cases and lesions:

$$\theta_{\text{AFROC}} = \frac{1}{K_1 L_T} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} \psi \left(\text{FP}_{k_1 1}, z_{k_2 2 l_2 2} \right) \tag{8.2}$$

Since every lesion is assigned a rating, albeit negative infinity for an unmarked lesion, the null set conditioning is not needed.

8.3.1 Upper limit for AFROC FOM-statistic

The FOM-statistic θ_{AFROC} achieves its highest value, unity, if and only if every lesion is rated higher than any mark on non-diseased cases, for then the ψ function always yields unity, and the summations yield :

[1] The kernel function comparison yields 1 if the LL rating is higher, 0.5 if the ratings are identical and zero otherwise.

$$\left.\begin{aligned}\theta_{\text{AFROC}} =& \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} 1 \\ =& \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} L_{k_2} \\ =& \frac{1}{K_1} \sum_{k_1=1}^{K_1} 1 \\ =& 1\end{aligned}\right\} \tag{8.3}$$

8.3.2 Range of AFROC FOM-statistic

If, on the other hand, every lesion is rated lower than every mark on every non-diseased case, the ψ function always yields zero, and the FOM-statistic is zero. Therefore,

$$0 \le \theta_{\text{AFROC}} \le 1 \tag{8.4}$$

Eqn. (8.4) shows that θ_{AFROC} behaves like a probability but its range is *twice* that of θ_{ROC}; recall that $0.5 \le \theta_{\text{ROC}} \le 1$ (assuming the observer has equal or better than random performance and the observer does not have the direction of the rating scale accidentally reversed). This has the consequence that treatment related differences between θ_{AFROC} (i.e., effect sizes) are larger relative to the corresponding ROC effect sizes (just as temperature differences in the Fahrenheit scale are larger than the same differences expressed in the Celsius scale). This has important implications for FROC sample size estimation, Chapter TBA.

Eqn. (8.4) is one reason why the "chance diagonal" of the AFROC, corresponding to AUC = 0.5, does not, in fact, reflect chance-level performance. An area under the AFROC equal to 0.5 is actually reasonable performance, being smack in the middle of the allowed range. An example of this was given in TBA §13.4.2.2 for the case of an expert radiologist who does not mark any cases.

8.4 Empirical weighted-AFROC FOM-statistic

The empirical weighted-AFROC plot and lesion weights were defined in Section 5.8. The empirical weighted-AFROC FOM-statistic (Chakraborty and Berbaum, 2004) is defined by including the lesion weights $W_{k_2 l_2}$ inside the summations (but outside the kernel function):

$$\theta_{\text{wAFROC}} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \psi\left(\text{FP}_{k_1 1}, z_{k_2 2 l_2 2}\right) \tag{8.5}$$

The weights obey the constraint:

$$\sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} = 1 \tag{8.6}$$

This ensures, as will be shown shortly, that each diseased case contributes equally to the FOM, regardless of how many lesions are in it. In the special case of one lesion per diseased case, θ_{AFROC} and θ_{wAFROC} are identical. For equally weighted lesions,

$$W_{k_2 l_2} = \frac{1}{L_{k_2}} \tag{8.7}$$

For example, for equally weighted lesions and a case with three lesions, each weight equals one-third (1/3) [2].

8.5 Two Theorems

The area A_{wAFROC} under the wAFROC plot is obtained by summing the areas of individual trapezoids defined by drawing vertical lines from each pair of adjacent operating points to the x-axis. A sample plot is shown Fig. 8.1.

The operating point labeled i has coordinates $(\text{FPF}_i, \text{wLLF}_i)$ given by Eqn. (5.14) and Eqn. (5.22), respectively, reproduced here for convenience:

$$FPF_i \equiv FPF\left(\zeta_i\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_i\right) \tag{8.8}$$

$$\text{wLLF}_i \equiv \text{wLLF}_{\zeta_i} = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 l_2 2} \geq \zeta_i\right) \tag{8.9}$$

TBA Online Appendix 14.A proves the following theorems:

[2] The `RJafroc` function `DfReadDataFile()` checks that the weights sum to unity to a precision of about 5 decimal places. The easy way to assign equal weights to all lesions on a diseased case is to set the corresponding `lesionWeights` field in the Excel file `Truth` worksheet to zeroes.

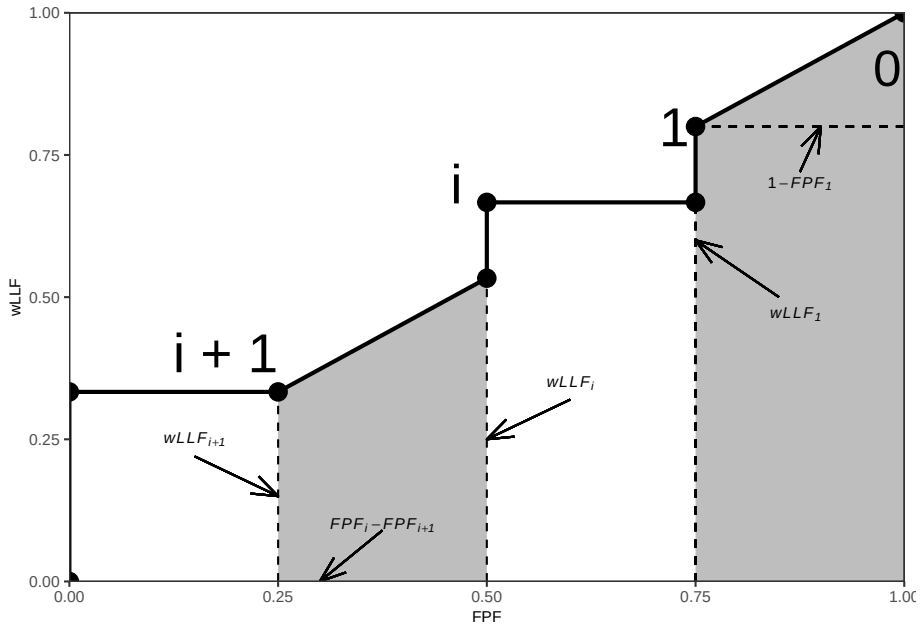


Figure 8.1: An example wAFROC plot; from left to right, the two shaded areas correspond to A_i and A_0, respectively, defined below.

8.5.1 Theorem 1

The area A_{wAFROC} under the empirical wAFROC plot equals the weighted-AFROC FOM-statistic θ_{wAFROC} defined by Eqn. (8.5):

$$\theta_{\text{wAFROC}} = A_{\text{wAFROC}} \tag{8.10}$$

This is the FROC counterpart of Bamber's Wilcoxon vs. empirical ROC area equivalence theorem (Bamber, 1975), derived in Section TBA (empirical-auc-wilcoxon-bamber-theorem).

8.5.2 Theorem 2

The area A_0 under the straight-line extension of the wAFROC from the observed end-point $(\text{FPF}_1, \text{wLLF}_1)$ to (1,1) is given by:

$$A_0 = \frac{(1 - \text{FPF}_1)(1 + \text{wLLF}_1)}{2} \tag{8.11}$$

According to Eqn. (8.11), A_0 increases as FPF_1 decreases, i.e., as more non-diseased cases are *not marked* and as wLLF_1 increases, i.e., as more lesions, especially those with greater weights, *are marked*. Both observations are in keeping with the behavior of a valid FOM.

- Failure to include the area under the straight-line extension results in not counting the full positive contribution to the FOM of unmarked non-diseased cases and marked lesions.
- Each unmarked non-diseased case represents a perfect decision.
- For a perfect observer whose operating characteristic is the vertical line from (0,0) to (0,1) followed by the horizontal line from (0,1) to (1,1), *the area under the straight-line extension comprises the entire AUC*. Excluding it would yield zero AUC for a perfect observer, which is obviously incorrect.
- Stated equivalently, for the perfect observer $\text{FPF}_1 = 0$ and $\text{wLLF}_1 = 1$ and then, according to Eqn. (8.11), the area under the straight line extension is $A_0 = 1$.

8.6 Numerical illustrations

The wAFROC and AFROC concepts are perhaps best illustrated with a numerical simulation-based illustration with very few cases.

Parameters of the simulation are $\mu = 2$, $\lambda = 1$, $\nu = 1$, $\zeta_1 = -1$ and $L_{max} = 2$. One simulates a dataset consisting of $K_1 = 4$ non-diseased cases and $K_2 = 4$ diseased cases. The first two diseased cases have one lesion each, and the remaining two have two lesions each.

```
#> AFROC AUC =  0.7708333
#> wAFROC AUC =  0.7875
```

Shown in Fig. 8.2 are the AFROC and wAFROC plots with operating points.

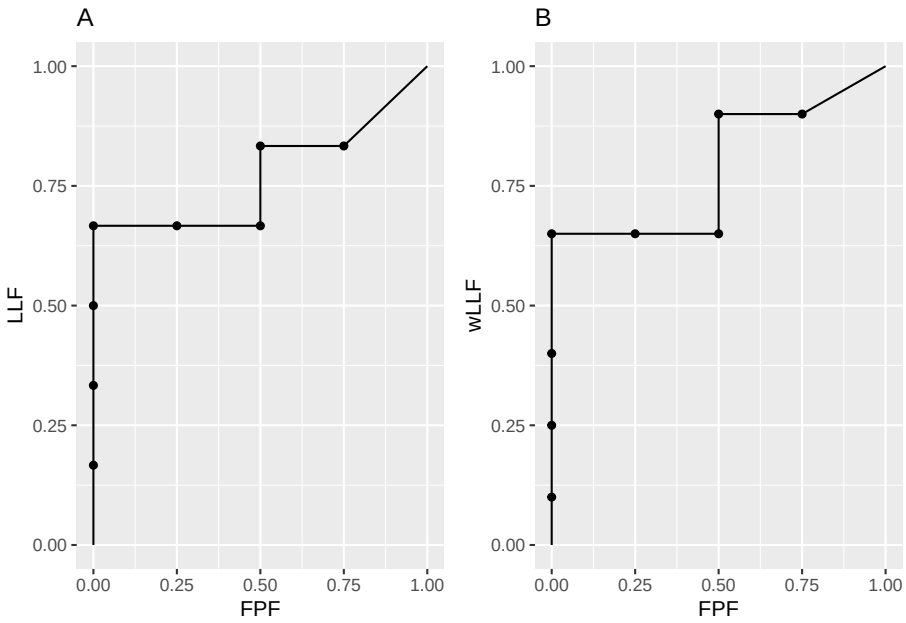


Figure 8.2: Left: AFROC plot; Right: corresponding wAFROC plot.

The number of lesions for diseased cases is shown next. Notice that the first two cases have one lesion each and the next two have two lesions each.

```
Lk2
#> [1] 1 1 2 2
```

The ratings are shown next.

```
x1 <- as.data.frame(frocData$ratings$NL[1,1,,])
colnames(x1) <- c("location1", "location2")
x2 <- as.data.frame(frocData$ratings$LL[1,1,,])
```

```
colnames(x2) <- c("location1", "location2")
x1
#>    location1 location2
#> 1       -Inf      -Inf
#> 2  0.4874291      -Inf
#> 3  0.7383247 0.5757814
#> 4 -0.3053884      -Inf
#> 5  1.5117812      -Inf
#> 6       -Inf      -Inf
#> 7       -Inf      -Inf
#> 8       -Inf      -Inf
x2
#>    location1 location2
#> 1  0.8523430      -Inf
#> 2 -0.2146999      -Inf
#> 3  1.5884892      -Inf
#> 4  2.9438362   1.98381
```

- The length of the third dimension of the NL array is eight (4 non-diseased + 4 diseased cases).
- The fifth sequential case corresponds to NLs on the first diseased case, etc.
- The first non-diseased case has no latent marks.
- The second non-diseased case has one latent mark rated 0.4874291.
- The third non-diseased case has two latent marks rated 0.7383247 and 0.5757814.
- The fourth non-diseased case has one latent mark rated -0.3053884.
- The first diseased case has one latent NL mark rated 1.5117812.
- The remaining diseased case have no latent NL marks.

8.7 Summary tables of ratings

Table 8.1 shows the layout of mark-rating pairs on *non-diseased* cases, illustrating FP ratings, corresponding to the green circles in Fig. 8.3. [UM denotes an unmarked location and blank cells denote unrealized z-samples.]

- Because non-diseased cases have no lesions, all z-samples listed in this table are for NLs.
- The first column lists the case numbers.
- The column labeled $k_t tl_s s$ lists the case-location indexing subscripts.
- The column labeled $z_{k_t tl_s s}$ lists the corresponding z-samples, when realized, and otherwise the cells are blank.

Table 8.1: Layout of mark-rating pairs on non-diseased cases; UM denotes an unmarked non-diseased case.

	$k_t tl_s s$	$z_{k_t tl_s s}$	$FP_{k_t t}$	Label
	1111			
1	1121		$-\infty$	UM
	2111	0.487		
2	2121		0.487	F
	3111	0.738		
3	3121	0.576	0.738	E
	4111	-0.305		
4	4121		-0.305	H

- The column labeled $FP_{k_t t}$ lists the FP rating for each non-diseased case, which is the highest of all realized z-samples on the case or $-\infty$ if none are realized.
- Column 5: the labels **A** - **H** correspond to the operating points shown in Fig. 8.3 and Fig. 8.4.

Table 8.2 shows the layout of mark-rating pairs on *diseased* cases, illustrating LL ratings, corresponding to the red circles in Fig. 8.3. [UM denotes an unmarked location and blank cells denote unrealized z-samples.]

- Because diseased cases can have NLs and LLs, both are shown in this table.
- The first column lists the case numbers.
- The second column lists the number of lesions present.
- Columns 3 and 4 illustrate NL indexing and z-samples.
- Columns 5 and 6 illustrate LL indexing and z-samples.
- Column 7 lists the lesion weights.
- Column 8: the labels **A** - **H** correspond to the operating points shown in Fig. 8.3 and Fig. 8.4.

8.8 AFROC plot from first principles

The following example is based on the same data involving 8 cases that were used to generate Table 8.1 and Table 8.2. It involves use of a linear or "one-dimensional" depiction of the ratings described next.

In Fig. 8.3, plot A, FPs and LLs, represented by green and red circles, respectively, are shown ordered, from left to right, with higher z-samples to the right, henceforth referred to as a *linear plot*. Each circle is labeled using the

Table 8.2: Layout of mark-rating pairs on diseased cases; UM denotes an unmarked lesion.

	L_{k_2}	$k_t tl_s s$	$z_{k_t tl_s s}$	$k_t tl_s s$	$z_{k_t tl_s s}$	weights	Label
		1211	1.512	1212	0.852	1	D
1	1	1221		1222			
		2211		2212	-0.215	1	G
2	1	2221		2222			
		3211		3212	1.588	0.6	C
3	2	3221		3222		0.4	UM
		4211		4212	2.944	0.4	A
4	2	4221		4222	1.984	0.6	B

$k_t tl_s s$ notation. For example, the right-most red circle corresponds to the LL z-sample originating from the first lesion in the fourth diseased case, i.e., z_{4212}. Consistent with the three unique values in the fourth column of Table 8.1, there are three green circles (FPs) [3]. Likewise, consistent with the five unique values in the sixth column of Table 8.2, there are five red circles (LLs) [4].

Starting from ∞, moving a virtual threshold continuously to the left generates the AFROC plot, see plot A in Fig. 8.3. As each FP is crossed, the operating point moves to the right by:

$$\frac{1}{K_1} = 0.025$$

As each LL is crossed, the operating point moves up by:

$$\sum_{k_2=1}^{K_2} L_{k_2} = \frac{1}{6}$$

Since it has one lesion, crossing the z-sample for the first case would result in an upward movement of 1/6, and likewise for the second case. Since the third case contains two lesions, crossing the corresponding z-samples would result in a net upward movement of the operating point by 1/3. *This behavior shows explicitly that the non-weighted method gives greater importance to diseased cases with more lesions, i.e., such case make a greater contribution to AUC.* The jumps from lesions in the same case need not be contiguous – they could be distributed, with intervening jumps from lesions on other cases, but eventually the jumps will occur and contribute to the net upward movement. As an example, the

[3] Not counting FP_{11}, which occurs at $z = -\infty$, representing the first non-diseased case with no marks.

[4] Not counting $z_{3222} = -\infty$ representing the unmarked second lesion on the third diseased case.

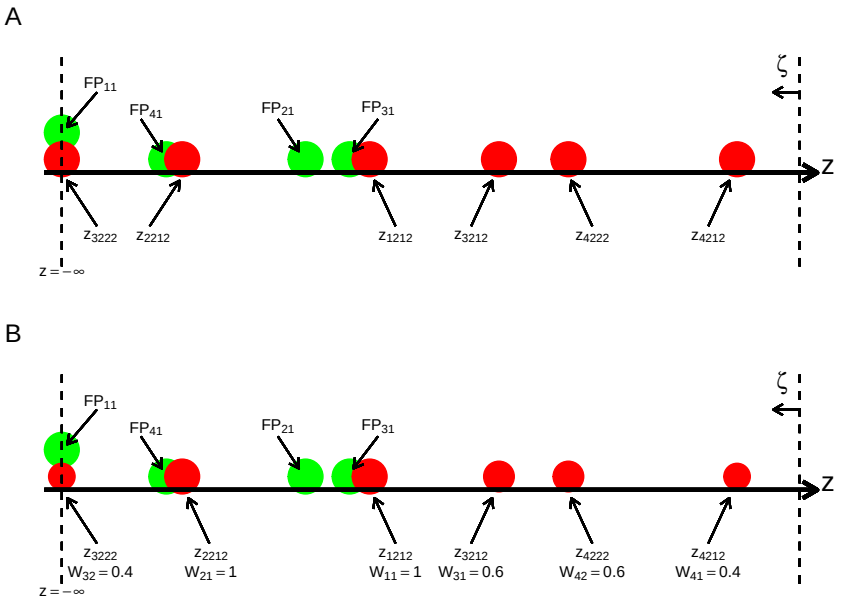


Figure 8.3: Plot A (illustrating generation of the AFROC): a one-dimensional depiction of the data in Table 8.1 and Table 8.2, showing z-samples used for plotting the AFROC; the red circles correspond to lesion localizations (LLs) and the green to false positives (FPs). Plot B (illustrating generation of the wAFROC): Data in same tables but this time including the weights, for plotting the weighted-AFROC plot; the sizes of the red circles code the lesions weights; the weights are shown below each z-sample.

jumps due to the two lesions on the fourth diseased case are contiguous: see points A and B, in Fig. 8.3. However, the jumps due to the two lesions on the third diseased case are not contiguous: the first lesion gives the point C, but the unmarked lesion on this case, indicated by "UM" in Table 8.2, eventually contributes when the operating point moves diagonally from point H to (1,1).

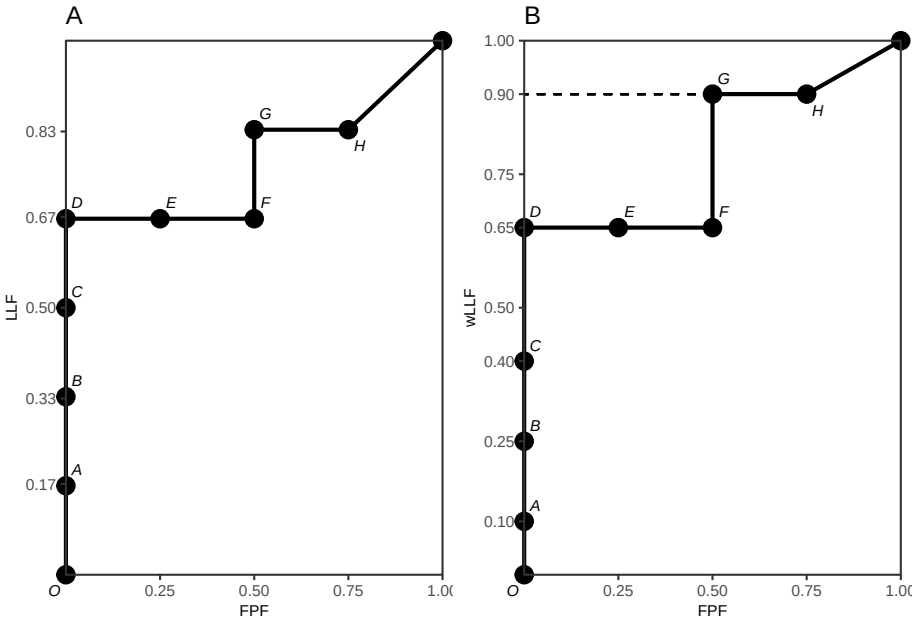


Figure 8.4: Plot A: The empirical AFROC plot for the data shown in Table 8.1 and Table 8.2. The labels correspond to the last columns of the tables. The corresponding one-dimensional depiction is plot A in Fig. 8.3. The area under the empirical plot is 0.7708. Plot B: The empirical weighted-AFROC (wAFROC) plot for the data shown in Table 8.1 and Table 8.2. The corresponding one-dimensional plot is plot B in Fig. 8.3. The area under the wAFROC is 0.7875.

8.9 wAFROC plot from first principles

Plot B in Fig. 8.3, which is the wAFROC analog of plot A in the same figure, is a one-dimensional depiction of the data in Table 8.1 and Table 8.2, but this time the lesion weights, shown in Table 8.2, are incorporated, as indicated by varying the *size* of each red circle (in Fig. 8.3 plot A, all red circles were of the same size). In addition, each lesion is labeled with its rating and weight.

Moving a virtual threshold continuously to the left generates the wAFROC plot, Fig. 8.4 plot B. The movement of the operating point in response to crossing

FPs is the same as before. However, as each LL is crossed the operating point moves up by an amount that depends on the lesion weight:

$$\frac{W_{k_2 l_2}}{K_2} = \frac{W_{k_2 l_2}}{4}$$

Since the first two diseased cases have one lesion each (i.e., unit weights), crossing the corresponding z-samples results in upward jumps of 0.25, Fig. 8.4 plot B – compare the jumps C to D and from F to G. According to the weights in Fig. 14.4, crossing the z-sample of the first lesion in the third diseased case, results in an upward jump of 0.6/4. That from the second lesion in the same case results in an upward jump of 0.4/4, for a net upward jump of the third case of 0.25, the same as for each of the first two diseased cases. Likewise crossing the z-samples of the two lesions in the 4th disease case results in upward jump of 0.4/4 = 0.1 (compare the jump from O to A), for the 1st lesion and 0.6/4 = 0.15 (compare the jump from B to C), for the 2nd lesion, for a net upward jump of 1/4, which is the same as for each of the first three diseased cases. *This shows explicitly that the weighting method gives each diseased case the same importance, regardless of the number of lesions in it, a property not shared by the area under the AFROC.*

8.10 Physical interpretations

From the preceding sections, it is seen that the AFROC-based trapezoidal plots consist of upward and rightward jumps, starting from the origin (0,0) and ending at (1,1). This is true regardless of whether the z-samples are binned or not: i.e., at the "microscopic" level the jumps always exist. Each upward jump is associated with a LL rating exceeding a virtual threshold. Each rightward jump is associated with a FP rating exceeding the threshold. Upward jumps tend to increase the area under the AFROC-based plots and rightward jumps tend to decrease it. This makes physical sense in terms of correct decisions being rewarded and incorrect ones being penalized, and can be seen from two extreme-case examples. If there are only upward jumps, then the trapezoidal plot rises from the origin to (0,1), where all lesions are correctly localized without any generating FPs and performance is perfect – the straight-line extension to (1,1) ensures that the net area is unity. If there are only horizontal jumps, that takes the operating point from the origin to (1,0), where none of the lesions are localized and every non-diseased image has at least one NL mark, representing worst possible performance. Here, despite the straight line extension to (1,1), the net area is zero.

8.10.1 Physical interpretation of area under AFROC

The area under the AFROC has the following physical interpretation: it is the fraction of LL vs. FP z-sample comparisons where the LL sample is equal (counting as half a comparison) or greater (counting as a full comparison) than the FP z-sample. From Tables 1 and 2, there are four FPs and six LLs for 24 possible comparisons. Inspection of the tables reveals that there are 4 x 4 = 16 comparisons contributing ones, two comparisons (from the 2nd diseased case) contributing ones, and one comparison (from the 2nd lesion on the 3rd diseased case) contributing 0.5, which sum to 18.5. Dividing by 24 yields 18.5/24 = 0.7708, the empirical TBA AFROC-AUC, §14.5.1. In probabilistic terms:

The area under the AFROC is the probability that a lesion is rated higher than any mark on a non-diseased case.

8.10.2 Physical interpretation of area under wAFROC

The area under the wAFROC has the following physical interpretation: it is the lesion-weight adjusted fraction of diseased cases vs. non-diseased case comparisons where LL z-samples are equal (counting as half a comparison times the weight of the lesion in question) or greater (counting as a full comparison times the weight of the lesion) than FP z-samples. Note that there are still 24 LL vs. FP comparisons but the counting proceeds differently. The fourth diseased case contributes 0.4 x 4 + 0.6 x 4, i.e., 4 (compared to 8 in the preceding example). The third diseased case contributes 0.6 x 4 + 0.4 x 0.5, i.e., 2.6 (compared to 4.5 in the preceding example). The second diseased case contributes 1 x 2 = 2 (compared to 2 in the preceding example), and the first diseased case contributes 1 x 4 = 4 (compared to 4 in the preceding example). Summing these values and dividing by 16 (the total number of diseased cases vs. non-diseased cases comparisons) one gets 12.6/16 = 0.7875, which is the area under the wAFROC, §14.5.1. In probabilistic terms:

The area under the weighted-AFROC is the lesion-weight adjusted probability that a lesion is rated higher than any mark on a non-diseased case.

8.11 Discussion

TBA TODOLAST

The primary aim of this chapter was to develop expressions for FOMs (i.e., functions of ratings) and show their equivalences to the empirical AUCs under corresponding operating characteristics. Unlike the ROC, the AFROC and wAFROC figures of merit are represented by quasi-Wilcoxon like constructs, not the well-known Wilcoxon statistic5.

I am aware from users of my software that their manuscript submissions have sometimes been held up with the critique that the meaning of the AFROC FOM-statistic is "not intuitively clear"6 TBA. Any critique based on intuitive clarity or lack thereof suffers from a fundamental flaw: it is un-falsifiable. What is "intuitively not clear" to one could be "intuitively very clear" to another, and there is no way of testing either viewpoint. Un-falsifiable claims have no place in science.

An example was given in a previous chapter. This is one reason I have tried to make the meaning clear, perhaps at the risk of making it painfully clear. Clinical interpretations do not always fit into convenient easy to analyze paradigms. Not understanding something is not a reason for preferring a simpler method. Use of the simpler ROC paradigm to analyze location specific tasks results in loss of statistical power and sacrifices better understanding of what is limiting performance. It is unethical to analyze a study with a method with lower statistical power when one with greater power is available7-9. The title of the paper by Halpern et al is "The continuing unethical conduct of under-powered clinical trials". The AFROC FOM-statistic was proposed in 1989 and it has been used, at the time of writing, in over 107 publications .

The subject material of this chapter is not that difficult. However, it does require the researcher to be receptive an unbiased. Dirac addressed an analogous then-existing concern about quantum mechanics, namely it did not provide a satisfying "picture" of what is going on, as did classical mechanics . To paraphrase him, the purpose of science (quantum physics in his case) is not to provide satisfying pictures but to explain data. FROC data is inherently more complex than the ROC paradigm and one should not expect a simple FOM-statistic. The detailed explanations given in this chapter should allow one to understand the wAFROC and AFROC FOMs.

A misconception regarding the wAFROC FOM-statistic is that the weighting may sacrifice statistical power and render the method equivalent to ROC analysis in terms of statistical power. Analysis of clinical datasets and simulation studies suggests that this is not the case; loss of power is minimal. As noted earlier, the highest rating carries more information than a randomly selected rating.

Bamber's equivalence theorem led to much progress in non-parametric analysis of ROC data. The proofs of the equivalences between the areas under the AFROC and wAFROC and the corresponding quasi-Wilcoxon statistics provide a starting point. To realize the full potential of these proofs, similar work like that conducted by DeLong et al10 is needed for the FROC paradigm. This work is not going to be easy; one reason being the relative dearth of researchers working in this area, but it is possible. Indeed work has been published by Popescu11 on non-parametric analysis of the exponentially transformed FROC (EFROC) plot which, like the AFROC and wAFROC, is completely contained within the unit square. This work should be extended to the wAFROC. For reasons stated in Chapter 13, non-parametric analysis of FROC curves12-14 is

not expected to be fruitful.

Current terminology prefixes each of the AFROC-based FOMs with the letter "J" for Jackknife. The author recommends dropping this prefix, which has to do with significance testing procedure rather than the actual definition of the FOM-statistic. For example, the correct way is to refer to the AFROC figure of merit, not the JAFROC figure of merit. For continuity, the software packages implementing the methods are still referred to as JAFROC (Windows) or RJAfroc (cross-platform, open-source).

To gain deeper insight into the FROC paradigm, it is necessary to look at methods used to measure visual search, the subject of the next chapter.

8.12 References

Chapter 9

Search and classification performances

9.1 TBA How much finished

10%

9.2 Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM and how they compare with evidence. The starting point is the inferred ROC curve. While mathematically rather complicated, the results are important because they are needed to derive the ROC-likelihood function, which is used to estimate RSM parameters from ROC data in TBA Chapter 19. The preceding sentence should lead the inquisitive reader to the question: *since the ROC paradigm ignores search, how is it possible to derive parameters of a model of search from the ROC curve?* The answer is that the *shape* of the ROC curve contains information about the RSM parameters. It is fundamentally different from predictions of all conventional ROC models: binormal (Dorfman and Alf, 1969), contaminated binormal model (Dorfman and Berbaum, 2000), bigamma (Dorfman et al., 1997) and proper ROC (Metz and Pan, 1999), namely it has a *constrained end-point property*, while all other models predict that the *end-point*, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The nature of search is such that the limiting end-point is constrained to be below and to the left of (1,1). This key difference, allows one to estimate search parameters from ROC data.

Next, the RSM is used to predict FROC and AFROC curves. Two following sections show how search performance and lesion-classification performance can be quantified from the location of the ROC end-point. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it; if classified as a NL it would not be marked (in the mind of the observer every mark is a potential LL, albeit at different confidence levels). Note that lesion-classification is different from classification between diseased and non-diseased cases, which is measured by the ROC-AUC. Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, and this leads to an important conclusion, namely the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC are preferred. This will come as a surprise (shock?) to most researchers somewhat familiar with this field, since the overwhelming majority of users of FROC methods, particularly in CAD, have relied on the FROC curve. Finally, evidence for the validity of the RSM is presented.

9.3 Location of ROC end-point

From the previous chapter, and restricting to one lesion per diseased case, the coordinates of the end-point are given by:

$$\left.\begin{aligned} \mathrm{FPF}_{max} &= 1 - exp\left(\lambda'\right) \\ \mathrm{TPF}_{max} &= 1 - exp\left(-\lambda'\right)\left(1 - \nu'\right) \end{aligned}\right\} \tag{9.1}$$

9.4 Quantifying search performance

Search performance is qualitatively equivalent to the ability to find lesions while avoiding finding non-lesions.

Fig. 9.1: Plot (a) is a typical ROC curve predicted by models that do not account for search performance. The end-point is at (1,1), the filled circle, i.e., by adopting a sufficiently low reporting threshold the observer can continuously move the operating point to (1,1). The curve labeled (b) is a typical RSM-predicted ROC curve. The end-point is down-left shifted relative to (1,1), see filled square. The observer cannot move the operating point continuously to (1,1). *The location of the end-point, in particular how far it is from (1,1), is a qualitative measure of search performance.* Higher search performance is characterized by the end-point moving upwards and to the left, in the limit to (0,1), corresponding to perfect search performance. The perpendicular distance, d_S, from the end-point to the chance diagonal (c), multiplied by $\sqrt{2}$, is a quantitative measure of search performance S.

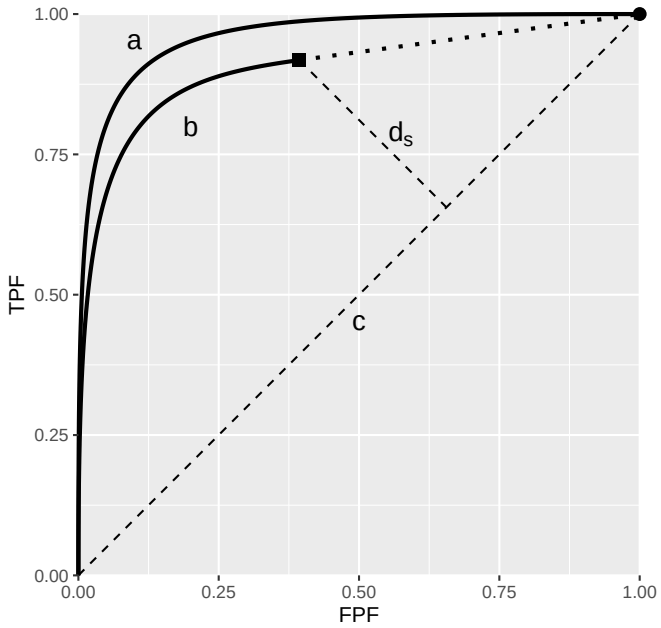


Figure 9.1: Relation of search performance to the end-point of the ROC curve.

Consider the case $\lambda' = 0$ and $\nu' = 1$. The end-point is (0,1). The perpendicular distance from (0,1) to the chance diagonal is $\frac{1}{\sqrt{2}}$, which multiplied by $\sqrt{2}$ yields unity. This observer has perfect search performance, i.e., $S = 1$. Since no NLs are found and all lesions are found, the observer never makes a mistake. One cannot improve over perfection: in this (admittedly extreme) case the observer does not need to make use of the decision variable z-sample information.

Using geometry and Eqn. (9.1), it follows that:

$$d_S = \frac{\mathrm{TPF}_{max} - \mathrm{FPF}_{max}}{\sqrt{2}} = \frac{1}{\sqrt{2}} \left(1 - \exp\left(-\mu\nu\right)\right) \exp\left(-\frac{\lambda}{\mu}\right) \tag{9.2}$$

Therefore, search performance S is given by:

$$\left.\begin{aligned} S &= \left(1 - \exp\left(-\mu\nu\right)\right) \exp\left(-\frac{\lambda}{\mu}\right) \\ S &= \nu' \exp\left(-\lambda'\right) \end{aligned}\right\} \tag{9.3}$$

The second Eqn. (9.3) shows search performance is the product of two terms: the probability ν' of finding lesions times the probability $\exp\left(-\lambda'\right)$ of not finding non-lesions. This puts into a mathematical form the qualitative definition of

search performance as the ability to find lesions while avoiding finding non-lesions.

9.5 Quantifying lesion-classification performance

Lesion-classification performance C is defined as the implied AUC of two unit variance normal distributions separated by the μ parameter of the search model. It measures the ability, having found a suspicious region (or alternatively, if the location of the possible lesion is known in advance), to correctly classify it as a lesion.

Lesion-classification performance C is distinct from case-classification performance AUC. The latter is commonly used in ROC terminology as a measure of the ability to distinguish between diseased and non-diseased cases. In contrast lesion-classification performance is a measure of the ability to distinguish between diseased and non-diseased regions, i.e., between latent NLs and latent LLs. C is determined by the μ parameter, and is defined by the implied ROC-area of two unit variance normal distributions separated by μ.

$$C = \Phi\left(\frac{\mu}{\sqrt{2}}\right) \tag{9.4}$$

C ranges from 0.5 to 1.

9.5.1 Lesion-classification performance and the 2AFC LKE task

It should be obvious that lesion-classification performance is similar to what is commonly measured in model-observer research using the location-known-exactly (LKE) paradigm. In this paradigm, one uses 2AFC methods as in Fig. 4.3, but one could use the ratings method as long as the lesion is cued (i.e., pointed to). On diseased cases, the lesion is cued, but to control for false positives, one must also cue a similar region on non-diseased cases, as in Fig. 4.3. In that figure, the lesion, present in one of the two images, is always in the center of one of the two fields. Sometimes cross hairs are used to indicate where the observer should be looking. The probability of a correct choice in the 2AFC task is , i.e., AUC conditioned on the (possible) position of the lesion being cued. Since the lesion is cued, search performance of the observer is irrelevant, and one expects . The reason for the inequality is that on a non-diseased case, the location being cued, in all likelihood, does not correspond to a latent NL found by the observer's search mechanism. Latent NLs are more suspicious for disease than other locations in the case. measures the separation parameter

between latent NLs and LLs. The separation parameter between latent LLs and a researcher chosen location is likely to be larger. This is because latent NLs are more suspicious for disease than a researcher chosen location. It is known that performance under this condition exceeds that in a free-search 2AFC or ROC study, denoted AUC, where the lesion is not cued and it could be anywhere. This should be obvious – pointing to the possible location of the lesion takes out the need for searching the rest of the image, which introduces the possibility of not finding the lesion and / or finding non-lesions. One expects the following ordering: . is expected to be the least, as there is uncertainty about possible lesion location. is expected to be next in order, as now uncertainty has been reduced, and the observer's task is to pick between two cued locations, one a latent NL and the other a latent LL. is expected to be highest, as now the observer's task is to pick between two cued locations, one a latent LL and the other a researcher chosen location, most likely not a latent NL. Data supporting the expected inequality is presented in §19.5.4.6.

9.5.2 Significance of measuring search and lesion-classification performance

The ability to quantify search and lesion-classification performance from a single paradigm (ROC) study is highly significant, going well-beyond modeling the ROC curve. ROC-AUC measures how well an observer is able to separate two groups of patients, a group of diseased patients from a group of non-diseased patients. While important, it does not inform us about how the observer goes about doing this and what is limiting performance (an exception the CBM model which yields information about how good the observer is at finding lesions but does not account for the ability of the observer to avoid NLs on non-diseased cases). In contrast, the search and lesion-classification measures described above can be used as a "diagnostic aid" in determining what is limiting performance. If search performance is poor, it indicates that the observer needs to be trained on many non-diseased cases, and learn the variants of non-diseased anatomy and learn not to confuse them for lesions. On the other hand, if lesion-classification performance is poor, then one needs to train the observer using images where the location of a possible lesion is cued, and the observer's task is to determine if the cued location is a real lesion. The classic example here is breast CAD, where the designer level ROC curve goes almost all the way to (1,1) implying poor search performance, while lesion-classification performance could actually be quite good, because CAD has access to the pixel values and the ability to apply complex algorithms to properly classify lesions as benign or malignant.

Of course, before one can realize these benefits, one needs a way of estimating the end-point shown in Fig. 17.6 plot (b). The observer will generally not oblige by reporting every suspicious region. RSM based curve fitting is needed to estimate the end-point's location, Chapter 19.

9.6 Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM-predicted curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to –infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are "lost in the mail"; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

RSM-predicted ROC, FROC and AFROC curves were derived (wAFROC is implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me 47,48, has used FROC curves to measure performance, this is going to be difficulty for some to accept. The examples in Fig. 17.6 (A- F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the

ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which in my opinion, is slightly misleading as at screening lesions are found not detected ("detection" is "discover or identify the presence or existence of something", correct localization is not necessarily implied; the more precise term is "localize"). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as "diagnostic aids" to optimize performance of a reader. For example, is search performance is low, then training using mainly non-diseased cases is called for, so the resident learns the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large . The RSM explains most empirical results regarding binormal model fits: the common finding that b < 1; that b decreases with increasing lesion pSNR (large and / or); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter < 1

and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects27-29. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

9.6.1 The Wagner review

The two RSM papers12,13 were honored by being included in a list of 25 papers the "Highlights of 2006" in Physics in Medicine and Biology. As stated by the publisher: "I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.

One of the reviewers was the late Dr. Robert ("Bob") Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduce one of his comments with minor edits, as it pertains to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I'm thinking here about the straight-line piece of the ROC curve from the max to (1, 1). 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point–because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a straight line, then the reader would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out–but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space,

but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will never guess. If the radiologist sees nothing to report, nothing will be reported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

9.7 References

1. 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.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. Medical Physics. 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic Radiology. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys Med Biol. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol. 2006;51:3449–3462.

Chapter 10

The FROC should not be used to measure performance

10.1 TBA How much finished

10%

10.2 Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM and how they compare with evidence. The starting point is the inferred ROC curve. While mathematically rather complicated, the results are important because they are needed to derive the ROC-likelihood function, which is used to estimate RSM parameters from ROC data in TBA Chapter 19. The preceding sentence should lead the inquisitive reader to the question: *since the ROC paradigm ignores search, how is it possible to derive parameters of a model of search from the ROC curve?* The answer is that the *shape* of the ROC curve contains information about the RSM parameters. It is fundamentally different from predictions of all conventional ROC models: binormal (Dorfman and Alf, 1969), contaminated binormal model (Dorfman and Berbaum, 2000), bigamma (Dorfman et al., 1997) and proper ROC (Metz and Pan, 1999), namely it has a *constrained end-point property*, while all other models predict that the *end-point*, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The nature of search is such

that the limiting end-point is constrained to be below and to the left of (1,1). This key difference, allows one to estimate search parameters from ROC data. Next, the RSM is used to predict FROC and AFROC curves. Two following sections show how search performance and lesion-classification performance can be quantified from the location of the ROC end-point. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it; if classified as a NL it would not be marked (in the mind of the observer every mark is a potential LL, albeit at different confidence levels). Note that lesion-classification is different from classification between diseased and non-diseased cases, which is measured by the ROC-AUC. Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, and this leads to an important conclusion, namely the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC are preferred. This will come as a surprise (shock?) to most researchers somewhat familiar with this field, since the overwhelming majority of users of FROC methods, particularly in CAD, have relied on the FROC curve. Finally, evidence for the validity of the RSM is presented.

10.3 The FROC curve is a poor descriptor of search performance

Why is the FROC curve is a bad descriptor of performance? The basic reason is that it is unconstrained in the x-direction26. Experts do not "move" as much along the positive x-direction as non-experts and partial area measures lose their meaning. Another reason is that it depends on the marks; unmarked non-diseased cases – representing perfect decisions - are not taken into account. The only meaningful comparison between two FROC curves occurs when they have a common NLF range, but this is rarely the case. As predicted by the RSM, a common range of NLF occurs when the two curves differ only in the parameter: if and are the same, then Eqn. (17.30) predicts the two curves will have identical . As shown below with numerical integration, this is the only situation where the area under the FROC tracks the area under the ROC, where the latter is regarded as the gold standard.

The code in file mainIsFrocGood.R, Online Appendix 17.E, calculates, by numerical integration, the areas under the full FROC, ROC and AFROC curves. Each full curve consists of the continuously accessible part plus any straight-line extension to (1,1), if applicable.

Fig. 7 here

The code is divided into 3 parts: * Part I, lines 15 – 64, calculates , and for varying , with ; * Part II, lines 66 – 115, calculates the same AUCs for varying

, with ; and * Part III, lines 117 – 159, calculates the same AUCs for varying , with .

This code takes a few minutes to complete running. The plots generated by this code are shown in Fig. 17.7 (A - F). The first column indicates which RSM parameter is being varied, ROC-AUC = is plotted along the x-axis, while is plotted along the y-axis in the left plot and is plotted along the y-axis in the right plot. The idea is that is the gold standard as it measures basic classification ability between diseased and non-diseased cases. So, for a valid figure of merit, the quantity plotted along the y-axis should monotonically increase with the gold standard, i.e., the slope should be positive. This is always true for but is not always true for ; it is only true when is varied, which, as was noted above, is the only situation when the range of integration along the NLF axis is constant.

Fig. 17.7 (A- F): Plots of plots along the x-axis, while is plotted along the y-axis in the left plot and is plotted along the y-axis in the right plot. Plots (A) and (B) correspond to varying , with and ; approximate slope AFROC vs. ROC = 2.00; plots (C) and (D) correspond to varying , with and ; approximate slope AFROC vs. ROC = 1.84; and plots (E) and (F) correspond to varying , with and ; approximate slope AFROC vs. ROC = 1.42. Regarding as the gold standard, the quantity plotted along the y-axis should be monotonic with the gold standard. This is always true for but is not always true for : it is only true for the varying . FROC-AUC is not constrained to unity; see plot (C); the AFROC-AUC is always in the range 0 to 1; see plots (B), (D) and (F). These plots were generated by mainIsFrocGood.R.

The plots of the FROC-AUC in (A) and (C) are non-linear and have negative slope. In contrast, the AFROC-AUCs have a quasi linear dependence on ROC-AUC. [The empirically determined slopes are printed by the code. For plot B the slope is 2.00, for plot D the slope is 1.84 and for plot F the slope is 1.42. These slopes indicate how much an ROC-AUC effect-size is amplified in the AFROC FOM. If only is different between two modalities, the amplification is almost exactly a factor of two. In the worst-case scenario, if only is different, the amplification is a factor of 1.42. In general, all three quantities could be different; one expects an intermediate amplification of the effect-size, in the range 1.4 to 2.]

One could argue that the above comparison is unfair to FROC as it considers the whole area under the FROC, while most users would use a point measure or a partial area measure, e.g., LLF @ selected NLF. The problem then is that some readers (especially the really good ones) cannot be analyzed as all of their operating points could to the left of the selected NLF value, and one would need to extrapolate outside the range of observed values in order to get the desired LLF @ selected NLF. For other readers, the data lying to the right of the selected NLF value does not contribute to the measure, resulting in loss of measurement accuracy

It is instructive to consider the extreme cases of a perfect observer and the worst

observer to see how the two methods of plotting would deal with defining the average observer. To make the comparison easier, consider that the lesions are small compared to the image area, so that the chance of a random LL is very small.

Fig. 17.8: (A) FROC curves for expert observer: vertical line extending from (0,0) to (0,1) and worst observer: horizontal line over the indicated NLF range. It is not possible to define an average FROC curve, as a common NLF range for the two observers does not exist. (B) Corresponding AFROC curves. AFROC-AUC for a perfect observer is unity (the area includes that under the dashed section extending from (0,1) to (1,1)). The corresponding area for the worst observer is zero, and the average AFROC curve is a straight line parallel to the x-axis at ordinate of 0.5, so the area under the average AFROC-AUC is 0.5 (unlike the ROC-AUC, AFROC-AUC = 0.5 does not denote worst possible performance). This plot was generated by the code in mainBestWorstObserver.R.

The perfect observer (LLF = 1 @ NLF = 0) and the worst observer (LLF = 0 @ NLF < some constrained value) both yield identical areas (zero) under the FROC curves. and it is not possible to define an average FROC curve, Fig. 17.7 (A). Because the two plots do not share a common range of abscissa values one cannot define an average FROC curve. In contrast, the AFROC is contained to the unit square and the area under the AFROC curve, Fig. 17.7 (B), is unity for the perfect observer (the area includes that under the dashed section extending from (0,1) to (1,1)). The corresponding area for the worst observer is zero, and the average AFROC curve is a straight line parallel to the x-axis at ordinate of 0.5, so the area under the average AFROC is 0.5 (as already noted, unlike the ROC area, AFROC area = 0.5 does not denote worst possible performance).

The FROC curve depends only on the marks. A valid FOM should reward correct decisions and penalize incorrect ones on all cases (in my judgment, the use of partial area measures, widespread in the literature, needs to reconsidered). Unmarked non-diseased cases are perfect decisions, but these are not accounted for in the FROC curve (they indirectly affect the curve by the leftward movement of the uppermost point, all the way to NLF = 0 for a perfect observer, but these decisions are not accounted for in FROC curve based partial area measures). The area under the horizontal dashed curve in the AFROC curve shown in Fig. 17.7 (B) is due to unmarked images. See §14.4.2 for further discussion of the meaning of the area under the dashed portion of the AFROC plot.

Finally, FP marks on diseased cases don't have the same negative connotation as those on non-diseased cases, since, following diagnostic workup, it is possible that the cancer will be found on the recalled cases, but, unlike the AFROC, both contribute to the FROC x-axis.

The RSM is a first-order model: a lot of interesting science remains to be uncovered. It does not account for the satisfaction of search (SOS) effects27-29 observed in medical imaging. It is as if the radiologist senses that an image is possibly diseased, without being able to pinpoint the specific reason, and

therefore adopts a more cautious reporting style. They are more reluctant to mark NLs on diseased than on non-diseased cases. This means the probability the a LL rating exceeds the rating of a NL on diseased cases is not equal to the probability that a LL rating exceeds the rating of a NL on non-diseased cases:

. (17.40)

Therefore, inclusion of inter-comparisons between LLs and NLs on diseased cases would make the figure of merit depend on disease prevalence, thereby destroying a desirable property of a valid figure of merit. This is another reason for excluding such comparisons on diseased cases in the AFROC/wAFROC figures of merit.

10.3.1 Clinically relevant portion of an operating characteristic #rsm-goodbye-froc-clinically-relevant}

The reason for the quotes is that in my experience this term is used rather loosely in the literature. There is a serious misconception that the "clinically relevant" part of an operating characteristic is the steep portion emanating from the origin. The purpose of this section is to clarify this notion. One needs to go back the definition of the FROC, particularly the linear plot, Fig. 14.2, showing how the raw plot is generated as a virtual threshold is moved from the far right to the far left. While this plot applies to the AFROC, the essential idea is the same. One orders the LL marks (red dots) from left to right in increasing order according to their z-samples. Likewise, the NL marks (green dots) are also ordered from left to right in increasing order according to their z-samples. As the virtual threshold is moved to the left, starting from , mostly red dots and occasional green dots are crossed; each time a red dot is crossed the operating point moves up by 1/(total number of lesions) and each time a greed dot is crossed the operating point moves to the right by 1/(total number of cases). This causes the operating point to rise, starting from the origin and move upward and to the right. The steep portion of the plot corresponds to crossings by LL and NL marks with high z-samples: it is the contribution of mostly easily visible lesions and the occasional NL. All observers are expected to localize the easy lesions, and there is nothing "clinically significant" about this. This argument applies to all operating characteristics. The clinical significance arises from the application. In a screening application, it is important to maintain high sensitivity at a reasonable specificity. In fact Jiang, Metz and Nishikawa30 had it right when they proposed the area above a preselected high sensitivity threshold divided by . Such a measure would emphasize the upper right corner of the ROC curve, not the steep portion near the origin. In the screening context, most of the z-samples (99.5% to be more precise) are from non-diseased cases, and only 0.5% is from diseased cases. This implies the "clinically relevant" part of the plot is near the upper right corner of the ROC plot. With the FROC a

normalized area above a preselected cannot be defined. On the other hand, the AFROC is amenable to such a partial area measure as is, of course, the ROC.

To do it right, one needs to include the costs and benefits of correct and incorrect decisions on diseased and non-diseased cases, the prevalence of disease and the actual population distribution of the z-samples for non-diseased and diseased cases, and perform a weighted average over the entire ROC or AFROC curve. In the screening context, this would tend to weight the upper end of the curve. This is not an easy problem but it can be solved.

10.4 Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM-predicted curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to –infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are "lost in the mail"; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

RSM-predicted ROC, FROC and AFROC curves were derived (wAFROC is

implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me 47,48, has used FROC curves to measure performance, this is going to be difficulty for some to accept. The examples in Fig. 17.6 (A- F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which, in my opinion, is slightly misleading as at screening lesions are found not detected ("detection" is "discover or identify the presence or existence of something", correct localization is not necessarily implied; the more precise term is "localize"). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as "diagnostic aids" to optimize performance of a reader. For example, is search performance is low, then training using mainly non-diseased cases is called for, so the resident learns the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged

by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large . The RSM explains most empirical results regarding binormal model fits: the common finding that b < 1; that b decreases with increasing lesion pSNR (large and / or); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter < 1 and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects27-29. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

10.4.1 The Wagner review

The two RSM papers12,13 were honored by being included in a list of 25 papers the "Highlights of 2006" in Physics in Medicine and Biology. As stated by the publisher: "I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.

One of the reviewers was the late Dr. Robert ("Bob") Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduce one of his comments with minor edits, as it pertains to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I'm thinking here about the straight-line piece of the ROC curve from the max to (1, 1). 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point–because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a straight line, then the reader

would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out–but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space, but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will not knowingly guess. If the radiologist sees nothing to report, nothing will be reported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

10.5 References

1. 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.

2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. Medical Physics. 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. Academic Radiology. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. Phys Med Biol. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. Phys Med Biol. 2006;51:3449–3462.

Chapter 11

RSM fitting

11.1 TBA How much finished

10%

11.2 Introduction

The radiological search model (RSM) is based on what is known, via eye-tracking measurements, about how radiologists look at medical images (Kundel and Nodine, 2004). The ability of this model to predict search and lesion-classification expertise was described in TBA Chapter 17. If one could estimate search and lesion-classification expertise from clinical datasets then one would know which of them is limiting performance. This would provide insight into the decision making efficiency of observers. For this potential to be realized, one has to be able to reliably estimate parameters of the RSM from data, and this turned out to be a difficult problem.

To put progress in this area in context a brief historical background is needed. I have worked on and off on the FROC estimation problem since 2002, and two persons (Dr. Hong-Jun Yoon and Xuetong Zhai) can attest to the effort. Initial attempts focused on fitting the FROC curve, in the (subsequently shown to be mistaken) belief that this was using *all* the data. In fact unmarked non-diseased cases, which are perfect decisions, are not taken into account in the FROC plot. In addition, there are degeneracy issues, which make parameter estimation difficult except in uninteresting situations. Early work involved maximization of the FROC likelihood function. This method was applied to seven designer-level CAD datasets. With CAD data one has a large number of marks and unmarked cases are relatively rare. However, only the CAD designer knows of their existence since in the clinic only a small fraction of the marks, those whose

z-samples exceed a manufacturer-selected threshold, are actually shown to the radiologist. In other words the full FROC curve, extending to the end-point, is available to the CAD algorithm designer, which makes estimation of the end-point defining parameters λ', ν' trivial. Estimating the remaining parameter of the RSM is then also relatively easy.

It was gradually recognized that the FROC curve based method worked only for designer level CAD data, and not for human observer data. Consequently, subsequent effort focused on ROC curve-based fitting, and this proved successful at fitting radiologist datasets, where detailed definition of the ROC curve is not available. A preliminary account of this work can be found in a conference proceeding (Chakraborty and Svahn, 2011).

The reader should be surprised to read that the research eventually turned to ROC curve based fitting, which implies that one does not even need FROC data to estimate RSM parameters. I have previously stated that the ROC paradigm ignores search, so how can one estimate search-model parameters from ROC data? The reason is that the *shape* of the ROC curve and the *position* of the upper-most observed operating point, depend on the RSM parameters, and this information can be used for a successful fitting method that is not susceptible to degeneracy [1].

The chapter starts with fitting FROC curves. This is partly for historical reasons and to make contact with a method used by CAD designers. Then focus shifts to fitting ROC curves and comparing the RSM-based method to existing methods, namely the proper ROC (PROPROC) (Metz and Pan, 1999; Pan and Metz, 1997) and the contaminated binormal model (CBM) (Dorfman and Berbaum, 2000) methods, both of which are proper ROC fitting models. These are described in more detail in TBA Chapter 20. The comparison is based on a large number of interpretations, namely, 14 datasets comprising 43 modalities, 80 readers and 2012 cases, most of which are from my international collaborations. Besides providing further evidence for the validity of the RSM, the estimates of search and lesion-classification performance derived from the fitted parameters demonstrate that there is information in ROC data that is currently ignored by analyses that do not account for search performance. *Specifically, it shows that search performance is the bottleneck that is currently limiting radiologist performance.*

The ability to fit RSM to clinical datasets is critical to sample size estimation – this was the practical reason why the RSM fitting problem had to be solved. Sample size estimation requires relating the wAFROC-AUC FOM to the corresponding ROC-AUC FOM in order to obtain a physically meaningful effect-size. Lacking a mathematical relationship between them, comparing the effect-sizes in the two units would be like comparing "apples and oranges". A mathematical relation is only possible if one has a parametric model that predicts both ROC

[1] Degenerate datasets are defined as those that do not provide any interior data points, i.e., all operating points lie on the edges of the ROC square, i.e., enclosed by the four lines defined by FPF = 0 or 1 and TPF = 0 or 1.

and wAFROC curves, as does the RSM. Therefore, this chapter concludes with sample size estimation for FROC studies using the wAFROC FOM. However, as long as one can predict the appropriate operating characteristic using RSM parameters, the method can be extended to other paradigms, e.g., the location ROC (LROC) (Chakraborty and Yoon, 2008) paradigm.

11.3 FROC likelihood function

Recall that the likelihood function is the probability of observing the data as a function of the parameter values. FROC notation was summarized in TBA Table 13.1. Thresholds $\vec{\zeta} \equiv \left(\zeta_0, \zeta_1, ..., \zeta_{R_{FROC}+1}, \right)$ were defined, where R_{FROC} is the number of FROC bins, and $\zeta_0 = -\infty$ and $\zeta_{R_{FROC}+1} = \infty$. Since each z-sample is obtained by sampling an appropriately centered unit-variance normal distribution, the probability p_r that a latent NL will be marked and rated in FROC bin r and the probability q_r that a latent LL will be marked and rated in FROC bin r are given by:

$$\left.\begin{aligned} p_r =& \Phi(\zeta_{r+1}) - \Phi(\zeta_r) \\ q_r =& \Phi(\zeta_{r+1} - \mu) - \Phi(\zeta_r - \mu) \end{aligned}\right\} \tag{11.1}$$

Understanding these equations is easy: the CDF function evaluated at a threshold is the probability that a z-sample is less than the threshold. The first equation is the difference between the CDF functions of a unit-normal distribution evaluated at the two thresholds. This is the probability that the NL z-sample falls in bin FROC:r. The second equation gives the probability that the LL z-sample falls in bin FROC:r. The probabilities p_r and q_r individually sum to unity when all bins, including the zero bin, are included.

If NL and LL events are assumed independent, the contributions to the likelihood function can be separated, and one need not enumerate counts at the individual case-level; instead, in the description that follows, one enumerates NL and LL counts in the various bins over the whole dataset.

11.3.1 Contribution of NLs

Define n (a random non-negative integer) as the total number of latent NLs in the dataset. The observed NL counts vector is $\vec{n} \equiv \left(n_0, n_1, ..., n_{R_{FROC}}, \right)$. Here n_r is the total number of NL counts in FROC ratings bin r, $n_0 = n - \sum_{r=1}^{R} n_r = n - N$, is the *unknown number of unmarked latent NLs* and N is the total number of observed NLs in the dataset. The probability $P\left(\vec{n} \mid n, \vec{\zeta}\right)$ of observing the NL counts vector $\vec{n}$ is (the factorials come from the multinomial distribution):

$$P\left(\vec{n} \mid n, \vec{\zeta}\right) = n! \prod_{r=0}^{R_{FROC}} \frac{p_r^{n_r}}{n_r!} \tag{11.2}$$

Since n is a random integer, the probability needs to be averaged over its Poisson distribution, i.e., one is calculating the expected value, yielding:

$$P\left(\vec{n} \mid \lambda', \vec{\zeta}\right) = \text{pmf}_{\text{Poi}}\left(n, K\lambda'\right) P\left(\vec{n} \mid n, \vec{\zeta}\right) \tag{11.3}$$

In this expression $K = K_1 + K_2$ is the total number of cases. $\text{pmf}_{\text{Poi}}\left(n, K\lambda'\right)$ of the Poisson distribution yields the probability of n counts from a Poisson distribution with mean $K\lambda'$. The multiplication by the total number of cases is required because one is counting the total number of latent NLs over the entire dataset. The lower limit on n is needed because n cannot be smaller than N, the total number of observed NL counts. The left hand side of Eqn. (11.3) is the probability of observing the NL counts vector $\vec{n}$ as a function of RSM parameters. Not surprisingly, since NLs are sampled from a zero-mean normal distribution, the μ parameter does not enter the above expression.

11.3.2 Contribution of LLs

Likewise, define l (a non-negative random integer) the total number of latent LLs in the dataset and the LL counts vector is $\vec{l} \equiv \left(l_0, l_1, ..., l_{R_{FROC}}, \right)$. Here l_r is the number of LL counts in FROC ratings bin r, $l_0 = l - \sum_{r=1}^{R_{FROC}} l_r = l - L$ is the *known* number of unmarked latent LLs and L is the total number of observed LLs in the dataset. The probability $P\left(\vec{l} \mid l, \mu, \vec{\zeta}\right)$ of observing the LL counts vector $\vec{l}$ is:

$$P\left(\vec{l} \mid l, \mu, \vec{\zeta}\right) = l! \prod_{r=0}^{R_{FROC}} \frac{q_r^{l_r}}{l_r!} \tag{11.4}$$

The above probability needs to be averaged over the binomial distribution of l:

$$P\left(\vec{l} \mid l, \mu, \nu', \vec{\zeta}\right) = \sum_{l=L}^{L_{tot}} \text{pmf}_{\text{Bin}}\left(l, L_T, \nu'\right) P\left(\vec{l} \mid l, \mu, \vec{\zeta}\right) \tag{11.5}$$

In this expression L_{tot} is the total number of lesions in the dataset and the lower limit on l is needed because it cannot be smaller than L, the total number of observed LLs. Performing the two summations using Maple, multiplying the two probabilities and taking the logarithm yields the final expression for the log-likelihood function (Yoon et al., 2007):

$$LL_{FROC} \equiv LL_{FROC}\left(\vec{n},\vec{l}\mid \mu,\lambda',\nu'\right) = \sum_{r=1}^{R_{FROC}} \{n_r log\left(p_r\right) + l_r log\left(q_r\right)\} + Nlog\left(\lambda'\right) + Llog\left(\nu'\right) - K\lambda'\left(1-p_0\right) + (L_T - 1$$

(11.6)

11.3.3 Degeneracy problems

The product $\lambda'\left(1-p_0\right) = \lambda'\Phi(-\zeta_1)$ reveals degeneracy in the sense that two quantities appear as a product, so that they cannot be individually separated. The effect of increasing λ' can be counteracted by increasing ζ_1; increasing λ' yields more latent NLs but increasing ζ_1 results in fewer of them being marked. The two possibilities cannot be distinguished. A similar degeneracy occurs in the term involving the product $-\nu' + \nu'\ q_0 = -\nu'(1-q_0) = -\nu'\Phi(\mu-\zeta_1)$, where increasing ν' can be counter balanced by decreasing $\mu-\zeta_1$, i.e., by increasing ζ_1. Again, the effect of increasing ν' is to produce more latent LLs, but increasing ζ_1 results in fewer of them being marked.

This is the fundamental problem with fitting RSM FROC curves to radiologist FROC data.

11.4 IDCA Likelihood function

In the limit $\zeta_1 \to -\infty$, $p_0 \to 0$ and $q_0 \to 0$, and TBA Eqn. (18.6) reduces to:

$$LL_{FROC}^{IDCA} = \sum_{r=1}^{R_{FROC}} \{n_r log\left(p_r\right) + l_r log\left(q_r\right)\} + Nlog\left(\lambda'\right) + Llog\left(\nu'\right) - K\lambda' + (L_T - L)\log\left(1-\nu'\right)$$

(11.7)

Notice that in the limit $\zeta_1 \to -\infty$ the degeneracy problems just described go away.

The superscript IDCA comes from *"initial detection and candidate analysis"* (Edwards et al., 2002). All CAD algorithms consist of an *initial detection* stage, which identifies possible *lesion candidates*. In the second stage the algorithm analyzes each candidate lesion, *candidate analysis*, to get a probability of malignancy. If the probability of malignancy exceeds a threshold value selected by the CAD manufacturer, and this is accomplished based on a compromise between sensitivity and specificity, and see Chapter 14 for my solution to this problem, the location of each candidate lesion satisfying the criterion is shown to the radiologist, Fig. 11.1.

According to TBA Eqn. (17.30), in the limit $\zeta_1 \to -\infty$ the end-point coordinates of the FROC curve represent estimates of λ', ν' respectively:

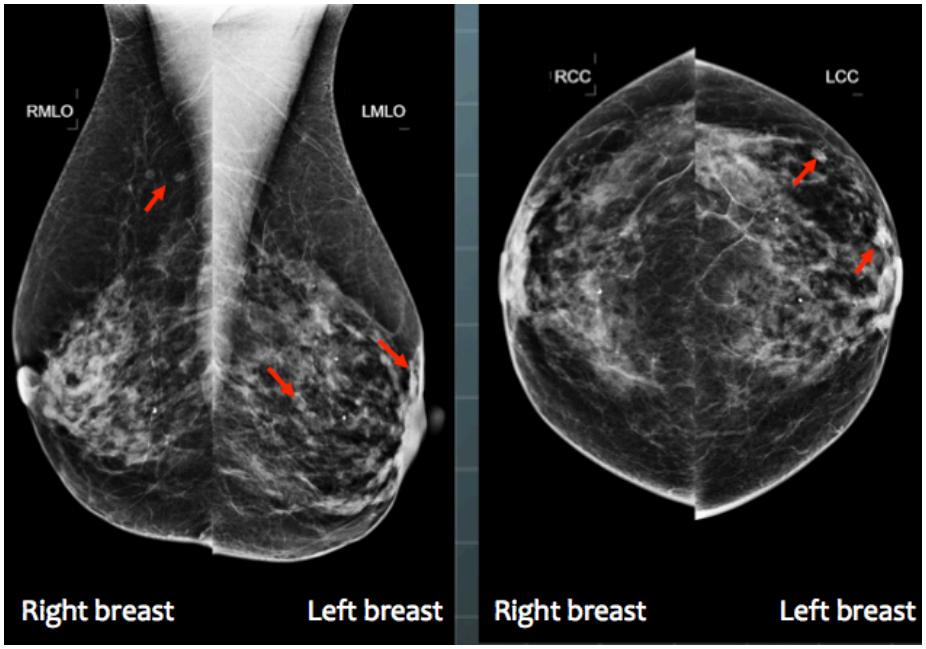


Figure 11.1: A typical 4-view display of a patient mammogram with the CAD cues (the red arrows) turned on.

$$\left.\begin{array}{l}\lambda' = NLF_{max} \\ \nu' = LLF_{max}\end{array}\right\} \tag{11.8}$$

In other words, in this limit two of the three parameters of the RSM are trivially determined from the location of the observed end-point. Suppressing all parameter independent terms, the log-likelihood function, Eqn. (11.7), reduces to:

$$LL_{FROC}^{IDCA} = \sum_{r=1}^{R_{FROC}} \{n_r log(p_r) + l_r log(q_r)\} + ... \tag{11.9}$$

Since the ignored terms in Eqn. (11.9) are independent of model parameters they do not affect the maximization. The equation contains only one parameter, namely μ, which is implicit in the definition of q_r, Eqn. (11.1).

Eqn. (11.9) resembles the log-likelihood function for the binormal model, since, according to TBA Eqn. (6.37), the LL function for the binormal model with R_{FROC} bins, is [2]:

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} log((\Phi(\zeta_{r+1}) - \Phi(\zeta_r))) + K_{2r} log((\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)))\} \tag{11.10}$$

In this equation K_{1r} is the number of counts in bin r of an ROC study consisting of R_{FROC} bins. Define the unequal-variance binormal model versions of Eqn. (11.1) as follows:

$$\left.\begin{array}{l}p'_r = \Phi(\zeta_{r+1}) - \Phi(\zeta_r) \\ q'_r = \Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)\end{array}\right\} \tag{11.11}$$

Here (a, b) are the parameters the unequal variance binormal model. Then Eqn. (11.10) becomes,

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} log(p'_r) + K_{2r} log(q'_r)\} \tag{11.12}$$

- With the identifications $K_{1r} \rightarrow n_r$ and $K_{2r} \rightarrow l_r$, Eqn. (11.10) looks exactly like Eqn. (11.9). This implies that binormal ROC fitting method can be used to determine a and b. Notice that instead of fitting an equal

[2] The number of ROC bins exceeds the number of FROC bins by one.

variance binormal model to determine the remaining single remaining μ parameter of the RSM, one is using an unequal-variance binormal model with two parameters, a and b. It turns out that the extra parameter helps. It gives some flexibility to the fitting curve to match the data.

- This method of fitting FROC data was well known to CAD researchers but was first formalized in (Edwards et al., 2002).

- Regard the NL marks as non-diseased "cases" ($K_{1r} \rightarrow n_r$) and the LL marks as diseased "cases" ($K_{2r} \rightarrow l_r$). Construct a pseudo-ROC counts table, analogous to TBA Table 4.1, where n_r is defined as the pseudo-FP counts in ratings bin r, and likewise, l_r is defined as the pseudo-TP counts in ratings bin r. The pseudo-ROC counts table has the same structure as the ROC counts table and can be fitted by the binormal model (or other alternatives).

- The pseudo-FP and pseudo-TP counts can be used to define pseudo-FPF and pseudo-TPF in the usual manner; the respective denominators are the total number of NL and LL counts, respectively. These probabilities define the pseudo-ROC operating points.

- The prefix "pseudo" is needed because one is regarding localized regions in a case as independent "cases". Since the fitting algorithm assumes each rating is from an independent case, one is violating a basic assumption, but with CAD data it appears one can get away with it, because the method yields good fits, especially with the extra parameter.

- The fitted FROC curve is obtained by scaling (i.e., multiplying) the ROC curve along the y-axis by LLF_{max} and along the x-axis by NLF_{max}. The method is illustrated in Fig. 11.2.

Fig. 11.2: The IDCA method of fitting designer-level CAD FROC data. In the upper half of the figure, the y-axis of the pseudo-ROC is pseudo-TPF and the x-axis is pseudo-FPF. The method is illustrated for a dataset with four FROC bins. Regarding the NLs and LLs as non-diseased and diseased cases, respectively, one constructs a table similar to Table 4.1, but this time with only four ROC bins (i.e., three non-trivial operating points). This defines the four operating points, the filled circles, including the trivial one at the upper right corner, shown in the upper half of the plot. One fits the ratings counts data using, for example, the binormal model, yielding the continuous line (based on experience the unequal variance binormal model is needed; the equal variance model does not fit as well). In practice, the operating points will not fall exactly on the fitted line. Finally, one scales (or "stretches", or multiplies) the y-axis by ν'. Likewise, the x-axis is scaled by λ'. This yields the continuous line shown in the lower half of the figure. Upon adding the FROC operating points one finds that they are magically fitted by the line, which is a scaled replica of the ROC fit in the upper curve.

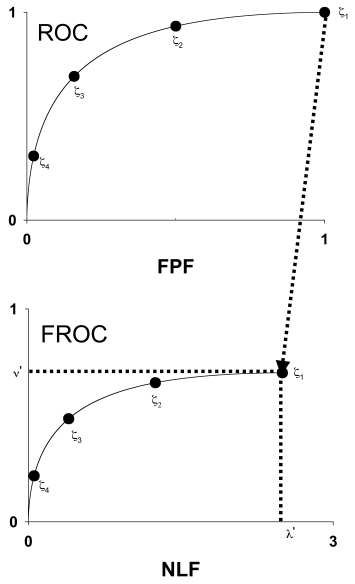


Figure 11.2: The IDCA method of fitting designer-level CAD FROC data.

Reference has already been made to the fact that it is necessary to assume $\zeta_1 = -\infty$ in order to remove the degeneracy problem. This is also evident from the fact that the uppermost point in Fig. 11.2 is at (1,1). A point at the upper-right corner must correspond to $\zeta_1 = -\infty$, another confirmation of this assumption.

Assuming binormal fitting is employed, yielding parameters a and b, the equations defining the IDCA fitted FROC curve are, see TBA Eqn. (6.19) and Eqn. (6.20):

$$\left.\begin{aligned} NLF(\zeta) =& \lambda' \Phi(-\zeta) \\ LLF(\zeta) =& \nu' \Phi(a - b\zeta) \end{aligned}\right\} \tag{11.13}$$

The RSM predicted FROC curve is repeated below for convenience,

$$\left.\begin{aligned} NLF(\zeta) =& \lambda' \Phi(-\zeta) \\ LLF(\zeta) =& \nu' \Phi(\mu - \zeta) \end{aligned}\right\} \tag{11.14}$$

IDCA uses the *unequal variance* binormal model to fit the pseudo-ROC, which of course opens up the possibility of an inappropriate chance-line crossing and a predicted FROC curve that is non-monotonically increasing with NLF (this is always present with IDCA fits, but one would need to examine the curve near

the end-point very closely to see it). In practice the unequal variance model gives visually good fits for CAD datasets.

In fact, IDCA yields excellent fits to some designer-level FROC datasets. However, the issue is not with the quality of the fits, rather the appropriateness of the FROC curve as a measure of performance, especially for human observers. For CAD the method works, so if one wished one could use IDCA to fit designer level CAD FROC data. However, with closely spaced operating points, the empirical FROC would also work and it does not involve any fitting assumptions. The issue is not fitting designer level CAD data but comparing stand-alone performance of designer level CAD to radiologists, and this is not solved by IDCA, which works for designer level CAD, but not for human observers. The latter do not report every suspicious region, no matter how low its confidence level, so the IDCA assumption $\zeta_1 \to -\infty$ is invalid. The problem of analyzing standalone performance of CAD against a group of radiologists interpreting the same cases is addressed in TBA Chapter 22.

11.5 ROC Likelihood function

The second attempt used the ROC likelihood function. In TBA Chapter 17 expressions were derived for the coordinates (x,y) of the ROC curve predicted by the RSM, see Eqn. (17.8) and Eqn. (17.16).

$$\text{FPF}\left(\zeta,\lambda'\right) = 1 - exp\left(-\frac{\lambda'}{2}\left[1 - erf\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \tag{11.15}$$

$$y \equiv y\left(\zeta,\mu,\lambda',\nu',\overrightarrow{f_L}\right) = 1 - \sum_{L=1}^{L_{max}} f_L \times \left[1 - \left(1 - \frac{\nu'}{2} + \frac{\nu'}{2}\text{erf}\left(\frac{\zeta-\mu}{\sqrt{2}}\right)\right)^L exp\left(-\frac{\lambda'}{2}\left[1 - erf\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right)\right] \tag{11.16}$$

Let (F_r, T_r) denote the number of false positives and true positives, respectively, in ROC rating bin r defined by thresholds $[\zeta_r, \zeta_{r+1})$, for $r = 0, 1, ..., R_{FROC}$. The range of r shows explicitly that R_{FROC} FROC ratings correspond to $R_{FROC}+1$ ROC bins [3]. Note that (F_0, T_0) represent the *known* numbers of non-diseased and diseased cases, respectively, with no marks, (F_1, T_1) represent the numbers of non-diseased and diseased cases, respectively, with highest rating equal to one, etc. The probability P_{1r} of a count in non-diseased ROC bin r is [4] :

[3] The rating bookkeeping can be confusing. Basically, $r = 0$ corresponds to unmarked cases, $r = 1$ corresponds to cases where the highest rated FROC mark was rated 1, etc., and $r = R_{FROC}$ corresponds to cases where the highest rated FROC mark was rated R_{FROC}.

[4] One needs to subtract the CDF evaluated at $r+1$ from that at r; the CDF is the complement of x, which results in the reversal. It should also make sense because the higher indexed x is to the right of the lower indexed one. Recall that the operating points are numbered starting from the top-right and working down.

$$P_{1r} = x\left(\zeta_r\right) - x\left(\zeta_{r+1}\right) \tag{11.17}$$

Likewise, the probability P_{2r} of a count in diseased ROC bin r is:

$$P_{2r} = y\left(\zeta_r\right) - y\left(\zeta_{r+1}\right) \tag{11.18}$$

The likelihood function is, ignoring combinatorial factors that do not depend on parameters:

$$\left(P_{1r}\right)^{F_r}\left(P_{2r}\right)^{T_r}$$

The log-likelihood function is:

$$LL_{ROC}\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right) = \sum_{r=0}^{R_{FROC}} \left[F_r log\left(P_{1r}\right) + T_r log\left(P_{2r}\right)\right] \tag{11.19}$$

The area $AUC_{ROC}^{RSM}\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right)$ under the parametric RSM-ROC curve was obtained by numerical integration:

$$AUC_{ROC}^{RSM}\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right) = \int_{x=0}^{1} y\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right) dx \tag{11.20}$$

The total number of parameters to be estimated, including the R_{FROC} thresholds, is $3 + R_{FROC}$. Maximizing the likelihood function yields parameter estimates. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Shanno and Kettler, 1970; Shanno, 1970; Goldfarb, 1970; Fletcher, 1970, 2013; Broyden, 1970) minimization algorithm, as implemented as function `mle2()` in the R-package `bbmle` (Bolker and R Development Core Team, 2020) was used to minimize the negative of the likelihood function. Since the BFGS algorithm varies each parameter in an unrestricted range $(-\infty, \infty)$, which would cause problems (e.g., RSM physical parameters cannot be negative and thresholds need to be properly ordered), appropriate variable transformations (both "forward" and "inverse") were used so that parameters supplied to the log-likelihood function were always in the valid range, irrespective of values chosen by the BFGS algorithm.

The algorithm calculates the goodness of fit statistic using the method described in TBA §6.4.2. Because of the additional parameter, the degrees-of-freedom (df) of the chisquare goodness of fit statistic is R_{FROC}–3. One can appreciate that calculating goodness of fit for the RSM can fail in situations, where the corresponding statistic can be calculated for binormal model, e.g., three (non – trivial) ROC operating points, corresponding to df = 1 . With FROC data one

needs at least four (non – trivial) ROC operating points, each defined by bins with at least five counts in both non-diseased and diseased categories. [5]

11.6 `FitRsmROC` implementation

The `RJafroc` function `FitRsmROC()` fits an RSM-predicted ROC curve to a binned single-modality single-reader ROC dataset. It is called by `ret <- FitRsmRoc(binnedRocData, lesDistr, trt = 1, rdr = 1)`, where `binnedRocData` is a binned ROC dataset, `lesDistr` is the lesion distribution vector (normalized histogram) in the dataset and `trt` and `rdr` are the desired treatment and reader to extract from the dataset, each of which defaults to one.

The return value `ret` is a `list` with the following elements:

- `ret$mu` The mean of the diseased distribution relative to the non-diseased one
- `ret$lambdP` The Poisson parameter describing the distribution of latent NLs per case
- `ret$nuP` The binomial success probability describing the distribution of latent LLs per diseased case
- `ret$zetas` The RSM cutoffs, zetas or thresholds
- `ret$AUC` The RSM fitted ROC-AUC
- `ret$StdAUC` The standard deviation of AUC
- `ret$NLLIni` The initial value of negative LL
- `ret$NLLFin` The final value of negative LL
- `ret$ChisqrFitStats` The chisquare goodness of fit results
- `ret$covMat` The covariance matrix of the parameters
- `ret$fittedPlot` A `ggplot2` object containing the fitted operating characteristic along with the empirical operating points. Use `print` to display the object

[5] With three operating points, each defined by bins with at least five counts in both non-diseased and diseased categories, the number of usable ROC bins is four. Subtracting three one gets $df = 1$, and the statistic can be calculated. However, because of the extra RSM parameter, the corresponding $df = 0$.

11.7 FitRsmROC usage example

- The following example uses the *first* treatment of the "FED" dataset, `dataset04`, which is a 5 treatment 4 radiologist FROC dataset acquired by Dr. Federica Zanca et. al. (Zanca et al., 2009). The dataset has 5 treatments and 4 readers and 200 cases and was acquired on a 5-point integer scale, i.e., it is already binned. If not one needs to bin the dataset using `DfBinDataset()`. I need to emphasize this point: **if the dataset represents continuous ratings, as with a CAD algorithm, one must bin the dataset to (ideally) about 5-6 bins**. The number of parameters that must be estimated increases with the number of bins (because for each additional bin one needs to estimate an additional cutoff parameter).

```
rocData <- DfFroc2Roc(dataset04)
lesDistr <- UtilLesionDistr(dataset04)[,2]
ret <- FitRsmRoc(rocData, lesDistr = lesDistr)
```

The lesion distribution vector is 0.69, 0.2, 0.11. This means that fraction 0.69 of abnormal cases contain one lesion each, fraction 0.2 contain two lesions each and fraction 0.11 contain three lesions each. The fitting algorithm needs to know the distribution of lesions per case, as the fitted curve depends on this distribution. For example, all else being equal, if all abnormal cases contain one lesion, the ROC curve will be lower than if all abnormal cases contain three lesions. With increased number of lesions per case TPF increases, as there is greater chance that at least one the lesions will be marked.

The fitted parameter values are as follows (all cutoffs excepting ζ_1, the chi-square statistic (NA for this dataset) and the covariance matrix are not shown):

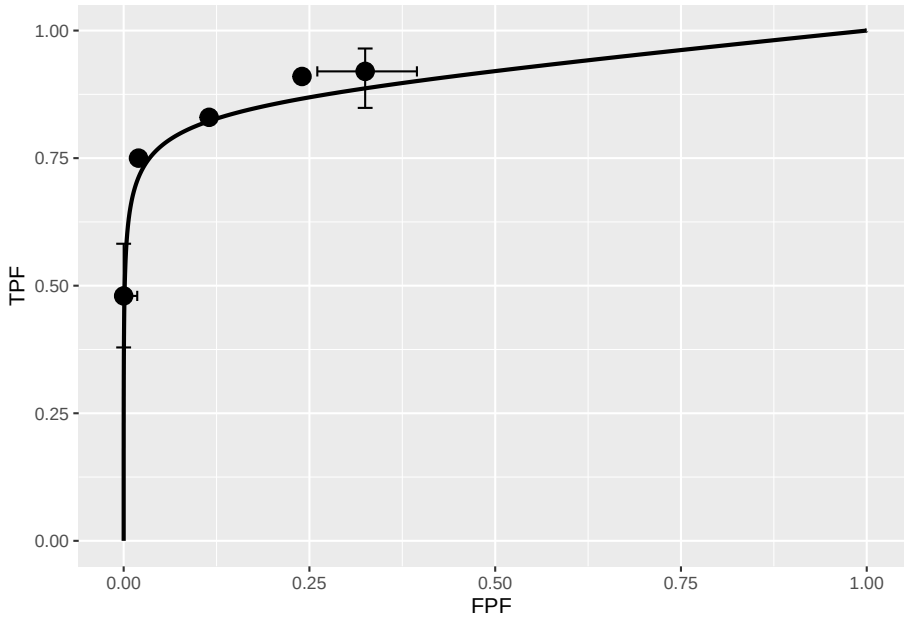
- $\mu = 3.6551363$
- $\lambda' = 9.8734529$
- $\nu' = 0.7963126$
- $\zeta_1 = 1.5006824$
- $AUC = 0.9065157$
- $\sigma(AUC) = 0.0231988$
- $NLLIni = 281.4024966$
- $NLLFin = 267.2673434$

The meaning of the parameters is as follows. The separation parameter μ is 3.66. The relatively large separation will result in good classification performance. The large λ' parameter means that on the average the observer generates 9.87 latent NL marks per image. However, because of the relatively large value of ζ_1, i.e., 1.5, only fraction 0.067 of these are actually marked, resulting in 0.66 actual marks per image. The fitting program decreased the negative of

the log-likelihood function from 281.4024966 to 267.2673434. A decrease in negative log-likelihood is equivalent to an increase in the likelihood, which is as expected, as the function maximizes the log-likelihood. Because the RSM contains 3 parameters, which is one more that other ROC models, the chisquare goodness of fit statistic usually cannot be calculated, except for large datasets - the criterion of 5 counts in each bin for true positives and false positives is usually hard to meet.

Shown next is the fitted plot. Error bars (exact 95% confidence intervals) are only shown for the lowest and highest operating points.

```
print(ret$fittedPlot)
```



The fitted ROC curve is proper: it's slope decreases monotonically as one moves up the curve thereby ruling out hooks such as are predicted by the binormal model. The area under the proper ROC is 0.907 which will be shown in a subsequent chapter to be identical to that yielded by other proper ROC fitting methods and higher than the binormal model fitted value.

11.8 Discussion / Summary

Over the years, there have been several attempts at fitting FROC data. Prior to the RSM-based ROC curve approach described in this chapter, all methods were aimed at fitting FROC curves, in the mistaken belief that this approach

was using all the data. The earliest was my FROCFIT software 36. This was followed by Swensson's approach 37, subsequently shown to be equivalent to my earlier work, as far as predicting the FROC curve was concerned 11. In the meantime, CAD developers, who relied heavily on the FROC curve to evaluate their algorithms, developed an empirical approach that was subsequently put on a formal basis in the IDCA method 12.

This chapter describes an approach to fitting ROC curves, instead of FROC curves, using the RSM. On the face of it, fitting the ROC curve seems to be ignoring much of the data. As an example, the ROC rating on a non-diseased case is the rating of the highest-rated mark on that image, or negative infinity if the case has no marks. If the case has several NL marks, only the highest rated one is used. In fact the highest rated mark contains information about the other marks on the case, namely they were all rated lower. There is a statistical term for this, namely sufficiency 38. As an example, the highest of a number of samples from a uniform distribution is a sufficient statistic, i.e., it contains all the information contained in the observed samples. While not quite the same for normally distributed values, neglect of the NLs rated lower is not as bad as might seem at first.

11.9 References

Chapter 12

Three proper ROC fits

12.1 TBA How much finished

75%

12.2 Introduction

A proper ROC curve is one whose slope decreases monotonically as the operating point moves up the curve, a consequence of which is that a proper ROC does not display an inappropriate chance line crossing followed by a sharp upward turn, i.e., a "hook", usually near the (1,1) upper right corner.

There are three methods for fitting proper curves to ROC datasets:

- The radiological search model (RSM) described in Chapter 11,
- The PROPROC (proper ROC) and CBM (contaminated binormal model) described in TBA Chapter 20.

This chapter compares these methods for a number of datasets. Comparing the RSM to the binormal model would be inappropriate, as the latter does not predict proper ROCs.

- Both RSM and CBM are implemented in `RJafroc`.
- `PROPROC` is implemented in Windows software [1] available here, last accessed 1/4/21.

[1] OR DBM-MRMC 2.5, Sept. 04, 2014; this version, used in this chapter, is no longer distributed but is available from me upon request.

12.3 Applications

The RSM, PROPROC and CBM algorithms were applied to the 14 embedded datasets described in 12.11. The datasets have already been analyzed and the location of pre-analyzed results files are in 12.13.

```
datasetNames <-  c("TONY", "VD", "FR",
            "FED", "JT", "MAG",
            "OPT", "PEN", "NICO",
            "RUS", "DOB1", "DOB2",
            "DOB3", "FZR")
```

In the following we focus on just two ROC datasets, which have been widely used in the literature to illustrate ROC methodological advances, namely the Van Dyke (VD) and the Franken (FR) datasets.

12.3.1 Application to two datasets

- The code uses the function `Compare3ProperRocFits()`.
- The code file is `R/compare-3-fits/Compare3ProperRocFits.R`.
- `startIndx` is the first `index` to analyze and `endIndx` is the last.
- In the current example `startIndx = 2` and `endIndx = 3`; i.e., two datasets are analyzed corresponding to `datasetNames[2]` and `datasetNames[3]`, i.e., the VD and FR datasets. [2]
- `reAnalyze` is set to `FALSE` causing pre-analyzed results (to be found in directory `R/compare-3-fits/RSM6`) to be retrieved. If `reAnalyze` is `TRUE` the analysis is repeated, leading to possibly slightly different results (the maximum likelihood parameter-search algorithm has inherent randomness aimed at avoiding finding false local maxima).
- The fitted parameter results are contained in `ret$allResults` and the *composite plots* (i.e., 3 combined plots corresponding to the three proper ROC fitting algorithms) are contained in `ret$allPlots`.
- These are saved to lists `plotArr` and `resultsArr`.

```
startIndx <-  2
endIndx <- 3
ret <- Compare3ProperRocFits(datasetNames,
                             startIndx = startIndx,
                             endIndx = endIndx,
                             reAnalyze = FALSE)

resultsArr <- plotArr <- array(list(),
```

[2] To analyze all datasets one sets `startIndx <-  1` and `endIndx <-  14`.

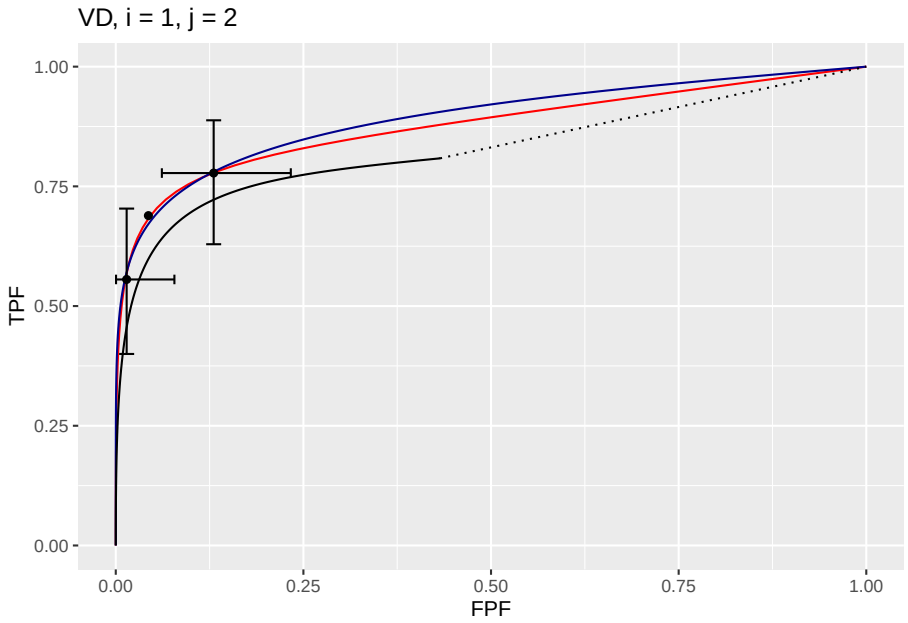
```
                                dim = c(endIndx - startIndx + 1))

for (f in 1:(endIndx-startIndx+1)) {
  plotArr[[f]] <- ret$allPlots[[f]]
  resultsArr[[f]] <- ret$allResults[[f]]
}
```

We show next how to display the composite plots.

12.4 Displaying composite plots

- The `plotArr` list contains plots for the two datasets. The Van Dyke plots are in `plotArr[[1]]` and the Franken in `plotArr[[2]]`.
- The Van Dyke plots contain $I \times J = 2 \times 5 = 10$ composite plots, and similarly for the Franken dataset (both datasets consist of 2 treatments and 5 readers).
- The following shows how to display the composite plot for the Van Dyke dataset for treatment 1 and reader 2.

```
plotArr[[1]][[1,2]]
```

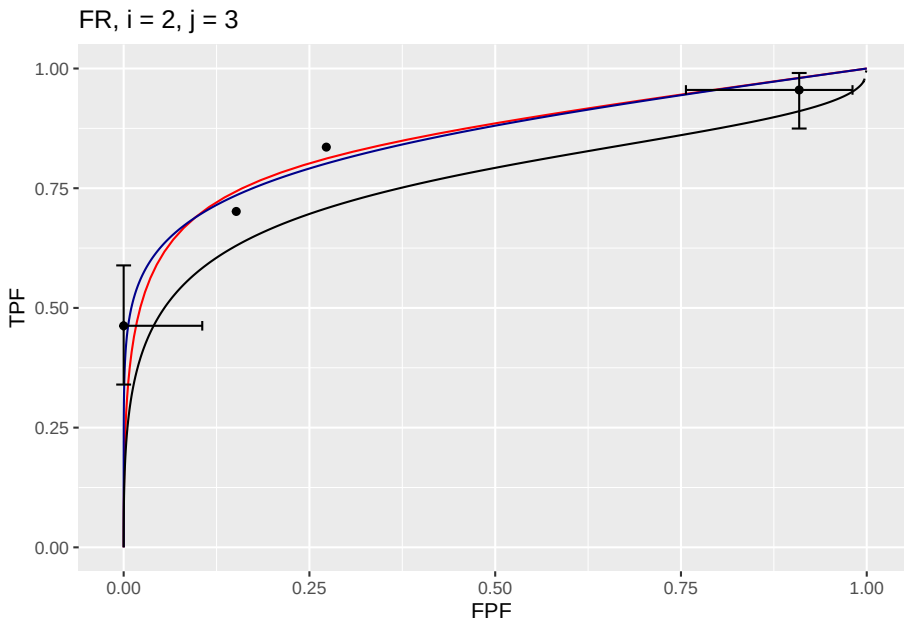


The plot is labeled **D2, i = 1, j = 2**, meaning the second dataset, the first treatment and the second reader. It contains 3 curves:

- The RSM fitted curve is in black. It is the only one with a dotted line connecting the uppermost continuously accessible operating point to (1,1).
- The PROPROC fitted curve is in red.
- The CBM fitted curve is in blue.
- Three operating points from the binned data are shown as well as exact 95% confidence intervals for the lowest and uppermost operating points.

The following example displays the composite plots for the Franken dataset, treatment 2 and reader 3:

```
plotArr[[2]][[2,3]]
```



Shown next is how to display the parameters corresponding to the fitted curves.

12.5 Displaying RSM parameters

The RSM has parameters: μ, λ', ν' and ζ_1. The parameters corresponding to the RSM plots are accessed as shown next.

- `resultsArr[[1]][[2]]$retRsm$mu` is the RSM μ parameter for dataset 1 (i.e., Van Dyke dataset), treatment 1 and reader 2,
- `resultsArr[[1]][[2]]$retRsm$lambdaP` is the corresponding λ' parameter, and

- `resultsArr[[1]][[2]]$retRsm$nuP` is the corresponding ν' parameter.
- `resultsArr[[1]][[2]]$retRsm$zeta1` is the corresponding ζ_1 parameter.
- Treatment 2 and reader 1 values would be accessed as `resultsArr[[1]][[6]]$retRsm$mu`, etc.
- More generally the values are accessed as `[[f]][[(i-1)*J+j]]`, where `f` is the dataset index, `i` is the treatment index, `j` is the reader index and `J` is the total number of readers.
- For the Van Dyke dataset `f = 1` and for the Franken dataset `f = 2`.

The first example displays RSM parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1;i <- 1; j <- 2;J <- 5
cat("RSM parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nmu = ",         resultsArr[[f]][[(i-1)*J+j]]$retRsm$mu,
"\nlambdaP = ",   resultsArr[[f]][[(i-1)*J+j]]$retRsm$lambdaP,
"\nnuP = ",       resultsArr[[f]][[(i-1)*J+j]]$retRsm$nuP,
"\nzeta_1 = ",    as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retRsm$zetas[1]),
"\nAUC = ",       resultsArr[[f]][[(i-1)*J+j]]$retRsm$AUC,
"\nsigma_AUC = ", as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retRsm$StdAUC),
"\nNLLini = ",    resultsArr[[f]][[(i-1)*J+j]]$retRsm$NLLIni,
"\nNLLfin = ",    resultsArr[[f]][[(i-1)*J+j]]$retRsm$NLLFin)
```

```
## RSM parameters, Van Dyke Dataset, treatment 1, reader 2:
## mu =  2.201413
## lambdaP =  0.2569453
## nuP =  0.7524016
## zeta_1 =  -0.1097901
## AUC =  0.8653694
## sigma_AUC =  0.04740562
## NLLini =  96.48516
## NLLfin =  85.86244
```

The next example displays RSM parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2;i <- 2; j <- 3;J <- 5
```

```
## RSM parameters, Franken dataset, treatment 2, reader 3:
## mu =  3.287996
## lambdaP =  9.371198
## nuP =  0.7186006
## zeta_1 =  1.646943
```

```
## AUC =  0.8234519
## sigma_AUC =  0.04054005
## NLLini =  128.91
## NLLfin =  122.4996
```

The first four values are the fitted values for the RSM parameters μ, λ', ν' and ζ_1. The next value is the AUC under the fitted RSM curve followed by its standard error. The last two values are the initial and final values of negative log-likelihood [3].

12.6 Displaying CBM parameters

CBM has parameters μ_{CBM}, α and ζ_1. The next example displays CBM parameters and AUC etc. for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1;i <- 1; j <- 2;J <- 5
cat("CBM parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nmu = ",         resultsArr[[f]][[(i-1)*J+j]]$retCbm$mu,
"\nalpha = ",      resultsArr[[f]][[(i-1)*J+j]]$retCbm$alpha,
"\nzeta_1 = ",     as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retCbm$zetas[1]),
"\nAUC = ",        resultsArr[[f]][[(i-1)*J+j]]$retCbm$AUC,
"\nsigma_AUC = ",  as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retCbm$StdAUC),
"\nNLLini = ",     resultsArr[[f]][[(i-1)*J+j]]$retCbm$NLLIni,
"\nNLLfin = ",     resultsArr[[f]][[(i-1)*J+j]]$retCbm$NLLFin)
```

```
## CBM parameters, Van Dyke Dataset, treatment 1, reader 2:
## mu =  2.745791
## alpha =  0.7931264
## zeta_1 =  1.125028
## AUC =  0.8758668
## sigma_AUC =  0.03964492
## NLLini =  86.23289
## NLLfin =  85.88459
```

The next example displays CBM parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2;i <- 2; j <- 3;J <- 5
```

[3] The initial value is calculated using initial estimates of parameters and the final value is that resulting from the log-likelihood maximization procedure. Since negative log-likelihood is being *minimized*, the final value is smaller than the initial value.

```
## CBM parameters, Franken dataset, treatment 2, reader 3:
## mu =  2.533668
## alpha =  0.6892561
## zeta_1 =  0.3097191
## AUC =  0.8194009
## sigma_AUC =  0.03968962
## NLLini =  122.6812
## NLLfin =  122.5604
```

The first three values are the fitted values for the CBM parameters μ, α and ζ_1. The next value is the AUC under the fitted CBM curve followed by its standard error. The last two values are the initial and final values of negative log-likelihood.

12.7 Displaying PROPROC parameters

`PROPROC` displayed parameters are c and d_a. The next example displays PROPROC parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1;i <- 1; j <- 2;J <- 5
cat("PROPROC parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nc = ",     resultsArr[[f]][[(i-1)*J+j]]$c1,
"\nd_a = ",   resultsArr[[f]][[(i-1)*J+j]]$da,
"\nAUC = ",   resultsArr[[f]][[(i-1)*J+j]]$aucProp)
```

```
## PROPROC parameters, Van Dyke Dataset, treatment 1, reader 2:
## c =  -0.2809004
## d_a =  1.731472
## AUC =  0.8910714
```

The values are identical to those listed for treatment 1 and reader 2 in Fig. 12.7. Other statistics, such as standard error of AUC, are not provided by PROPROC software.

The next example displays PROPROC parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2;i <- 2; j <- 3;J <- 5
```

```
## PROPROC parameters, Franken dataset, treatment 2, reader 3:
## c =  -0.4420007
## d_a =  0.9836615
## AUC =  0.8252824
```

All 10 composite plots for the Van Dyke dataset are shown in the Appendix to this chapter, 12.14.

The next section provides an overview of the most salient findings from analyzing the datasets.

12.8 Overview of findings

With 14 datasets the total number of individual modality-reader combinations is 236: in other words, there are 236 datasets to each of which three algorithms were applied. It is easy to be overwhelmed by the numbers and this section summarizes the most important conclusion: *for each dataset, treatment and reader, the three fitting methods are consistent with a single method-independent AUC.*

If the AUCs of the three methods are identical the following relations hold with slopes equal to unity:

$$\left.\begin{aligned} AUC_{PRO} &= m_{PR} AUC_{PRO} \\ AUC_{CBM} &= m_{CR} AUC_{PRO} \\ m_{PR} &= 1 \\ m_{CR} &= 1 \end{aligned}\right\} \tag{12.1}$$

The abbreviations are as follows:

- PRO = PROPROC
- PR = PROPROC vs. RSM
- CR = CBM vs. RSM.

For each dataset the plot of PROPROC AUC vs. RSM AUC should be linear with zero intercept and slope m_{PR}. The reason for the *zero intercept* is that if one of the AUCs indicates zero performance the other AUC must also be zero. Likewise, chance level performance (AUC = 0.5) must be common to all method of estimating AUC. Finally, perfect performance must be common to all methods. All of these conditions require a zero-intercept linear fit.

12.8.1 Slopes

Denote PROPROC AUC for dataset f, treatment i and reader j by θ_{fij}^{PRO}. Likewise, the corresponding RSM and CBM values are denoted by θ_{fij}^{RSM} and θ_{fij}^{CBM}, respectively. For a given dataset the slope of the PROPROC values vs. the RSM values is denoted $m_{PR,f}$. The (grand) average over all datasets

is denoted $m_{\bullet}^{PR}$. Likewise, the average of the CBM AUC values vs. the RSM value is denoted $m_{\bullet}^{CR}$.

An analysis was conducted to determine the average slopes and a bootstrap analysis was conducted to determine the corresponding confidence intervals.

The code for calculating the average slopes is in `R/compare-3-fits/slopesConvVsRsm.R` and that for calculating the bootstrap confidence intervals is in `R/compare-3-fits/slopesAucsConvVsRsmCI.R`.

```
ret <- slopesConvVsRsm(datasetNames)
retCI <- slopesAucsConvVsRsmCI(datasetNames)
```

The call to function `slopesConvVsRsm()` returns `ret`, which contains, for each of 14 datasets, two plots and two slopes. For example:

- `ret$p1[[2]]` is the plot of θ_{2ij}^{PRO} vs. θ_{2ij}^{RSM} for the Van Dyke dataset.
- `ret$p2[[2]]` is the plot of θ_{2ij}^{CBM} vs. θ_{2ij}^{RSM} for the Van Dyke dataset.
- `ret$m_pro_rsm` has two columns, each of length 14, the slopes $m_{PR,f}$ for the datasets (indexed by `f`) and the corresponding R^2 values. The first column is `ret$m_pro_rsm[[1]]` and the second is `ret$m_pro_rsm[[2]]`.
- `ret$m_cbm_rsm` has two columns, each of length 14, the slopes $m_{CR,f}$ for the datasets and the corresponding R^2 values.

Likewise,

- `ret$p1[[3]]` is the plot of θ_{3ij}^{PRO} vs. θ_{3ij}^{RSM} for the Franken dataset.
- `ret$p2[[3]]` is the plot of θ_{3ij}^{CBM} vs. θ_{3ij}^{RSM} for the Franken dataset.

As examples, `ret$p1[[2]]` is the plot of θ_{2ij}^{PRO} vs. θ_{2ij}^{RSM} for the Van Dyke dataset and `ret$p1[[3]]` is the plot of θ_{2ij}^{CBM} vs. θ_{2ij}^{RSM} for the Van Dyke dataset, shown next. Each plot has the constrained linear fit superposed on the data points; each data point represents a distinct modality-reader combination.

The next plot shows corresponding plots for the Franken dataset.

The average slopes and R^2 values (R^2 is the fraction of variance explained by the constrained straight line fit) are listed in Table 12.1.

The slopes and R^2 values for the Van Dyke dataset are shown next:

```
##         m-PR    R2-PR     m-CR     R2-CR
## VD 1.006127 0.999773 1.000699 0.9999832
```

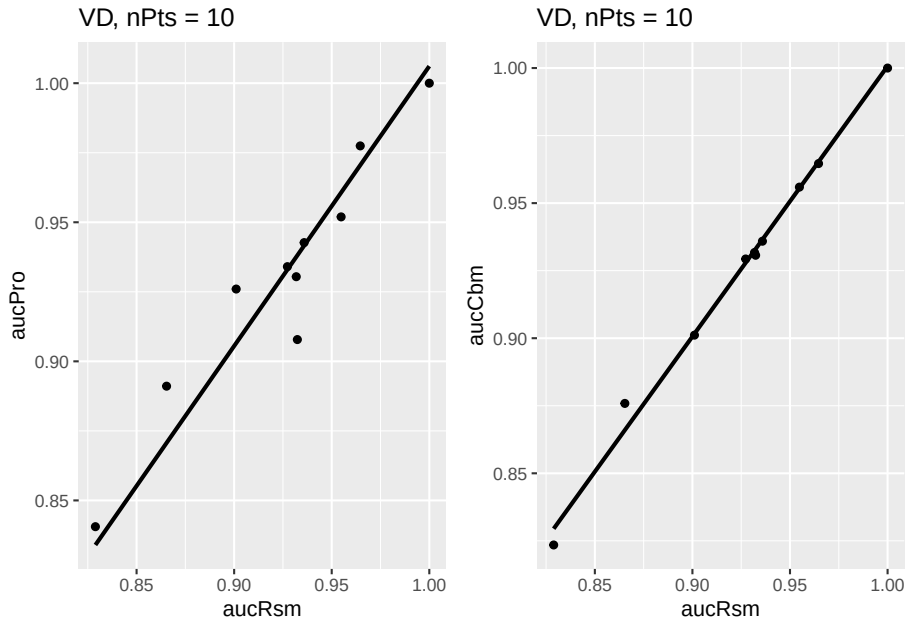


Figure 12.1: Van Dyke dataset: Left plot is PROPROC-AUC vs. RSM-AUC with the superposed constrained linear fit. The number of data points is `nPts` = 10. Right plot is CBM-AUC vs. RSM-AUC.

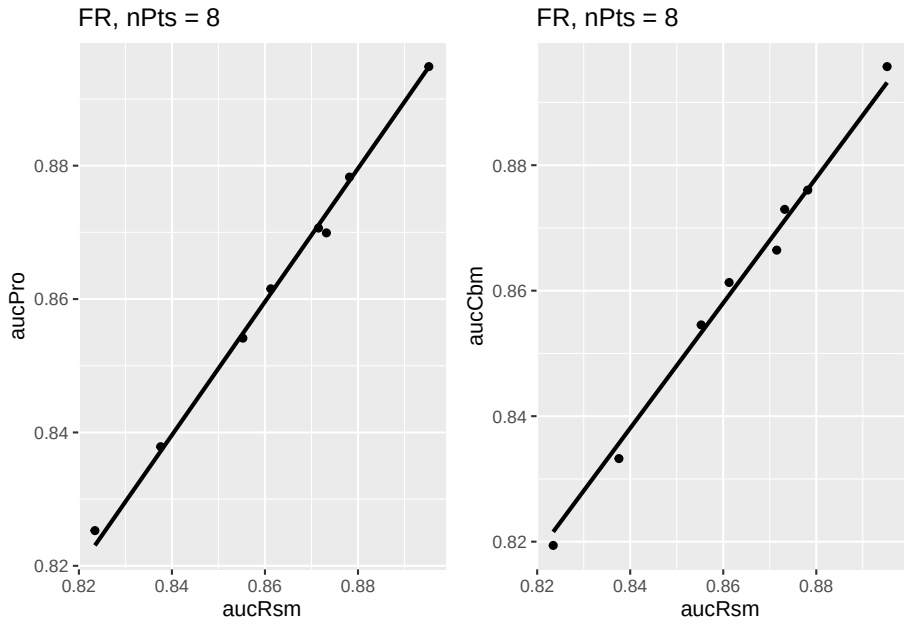


Figure 12.2: Similar to previous plot, for Franken dataset.

12.8.2 Confidence intervals

The call to `slopesAucsConvVsRsmCI` returns `retCI`, containing the results of the bootstrap analysis (note the bullet symbols • denoting averages over 14 datasets):

- `retCI$cislopeProRsm` confidence interval for $m_{\bullet}^{PR}$
- `retCI$cislopeCbmRsm` confidence interval for $m_{\bullet}^{CR}$
- `retCI$histSlopeProRsm` histogram plot for 200 bootstrap values of $m_{\bullet}^{PR}$
- `retCI$histSlopeCbmRsm` histogram plot for 200 bootstrap values of $m_{\bullet}^{CR}$
- `retCI$ciAvgAucRsm` confidence interval for 200 bootstrap values of $\theta_{\bullet}^{RSM}$
- `retCI$ciAvgAucPro` confidence interval for 200 bootstrap values of $\theta_{\bullet}^{PRO}$
- `retCI$ciAvgAucCbm` confidence interval for 200 bootstrap values of $\theta_{\bullet}^{CBM}$

As examples,

```
##            m-PR      m-CR
## 2.5%  1.005092 0.9919886
## 97.5% 1.012285 0.9966149
```

The CI for $m_{\bullet}^{PR}$ is slightly above unity, while that for $m_{\bullet}^{CR}$ is slightly below. Shown next is the histogram plot for $m_{\bullet}^{PR}$ (left plot) and $m_{\bullet}^{CR}$ (right plot). Quantiles of these histograms were used to compute the confidence intervals cited above.

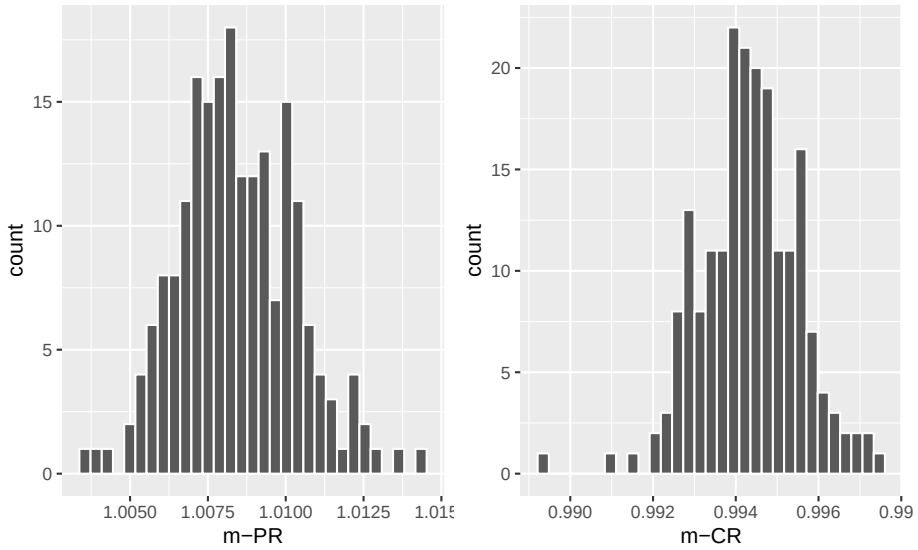


Figure 12.3: Histograms of slope PROPROC AUC vs. RSM AUC (left) and slope CBM AUC vs. RSM AUC (right).

12.8.3 Summary of slopes and confidence intervals

Table 12.1: Summary of slopes and correlations for the two constrained fits: PROPROC AUC vs. RSM AUC and CBM AUC vs. RSM AUC. The average of each slope equals unity to within 0.6 percent.

	m_{PR}	R^2_{PR}	m_{CR}	R^2_{CR}
TONY	1.0002	0.9997	0.9933	0.9997
VD	1.0061	0.9998	1.0007	1
FR	0.9995	1	0.9977	1
FED	1.0146	0.9998	0.9999	0.9999
JT	0.9964	0.9995	0.9972	1
MAG	1.036	0.9983	0.9953	1
OPT	1.0184	0.9997	1.0059	0.9997
PEN	1.0081	0.9996	0.9976	1
NICO	0.9843	0.9998	0.997	1
RUS	0.9989	0.9999	0.9921	0.9999
DOB1	1.0262	0.9963	0.9886	0.9962
DOB2	1.0056	0.9987	0.971	0.9978
DOB3	1.0211	0.998	0.9847	0.9986
FZR	1.0027	0.9999	0.9996	1
AVG	1.0084	0.9992	0.9943	0.9994
CI	(1.005, 1.012)	NA	(0.992, 0.997)	NA

Table 12.1: The first column, labeled m_{PR}, shows results of fitting straight lines, constrained to go through the origin, to fitted PROPROC AUC vs. RSM AUC results, for each of the 14 datasets, as labeled. The second column, labeled R^2_{PR}, lists the square of the correlation coefficient for each fit. The third and fourth columns list the corresponding values for the CBM AUC vs. RSM AUC fits. The second last row lists the averages (AVG) and the last row lists the 95 percent confidence intervals (CI) for the average slopes.

12.9 Discussion / Summary

Over the years, there have been several attempts at fitting FROC data. Prior to the RSM-based ROC curve approach described in this chapter, all methods were aimed at fitting FROC curves, in the mistaken belief that this approach was using all the data. The earliest was my FROCFIT software TBA 36. This was followed by Swensson's approach37, subsequently shown to be equivalent to my earlier work, as far as predicting the FROC curve was concerned TBA 11. In the meantime, CAD developers, who relied heavily on the FROC curve to evaluate their algorithms, developed an empirical approach that was subsequently put on a formal basis in the IDCA method12.

This chapter describes an approach to fitting ROC curves, instead of FROC curves, using the RSM. Fits were described for 14 datasets, comprising 236 distinct treatment-reader combinations. All fits and parameter values are viewable in the online "Supplemental Material" directory corresponding to this chapter. Validity of fit was assessed by the chisquare goodness of fit p-value; unfortunately using adjacent bin combining this could not be calculated in most instances; ongoing research at other ways of validating the fits is underway. PROPROC and CBM were fitted to the same datasets, yielding further validation and insights. One of the insights was the finding that the AUCS were almost identical, with PROPROC yielding the highest value, followed by CBM and closely by the RSM. The PROPROC-AUC / CBM-AUC, vs. RSM-AUC straight-line fits, constrained to go through the origin, had slopes 1.0255 (1.021, 1.030) and 1.0097 (1.006, 1.013), respectively. The R^2 values were generally in excess of 0.999, indicative of excellent fits.

On the face of it, fitting the ROC curve seems to be ignoring much of the data. As an example, the ROC rating on a non-diseased case is the rating of the highest-rated mark on that image, or negative infinity if the case has no marks. If the case has several NL marks, only the highest rated one is used. In fact the highest rated mark contains information about the other marks on the case, namely they were all rated lower. There is a statistical term for this, namely sufficiency38. As an example, the highest of a number of samples from a uniform distribution is a sufficient statistic, i.e., it contains all the information contained in the observed samples. While not quite the same for normally distributed values, neglect of the NLs rated lower is not as bad as might seem at first. A similar argument applies to LLs and NLs on diseased cases. The advantage of fitting to the ROC is that the coupling of NLs and LLs on diseased cases breaks the degeneracy problem described in §18.2.3.

The reader may wonder why I chose not to fit the wAFROC TBA. After all, it is the recommended figure of merit for FROC studies. While the methods described in this chapter are readily adapted to the wAFROC, they are more susceptible to degeneracy issues. The reason is that the y-axis is defined by LL-events, in other words by the parameters, while the x-axis is defined by the highest rated NL on non-diseased cases, in other words by the parameter. The

consequent decoupling of parameters leads to degeneracy of the type described in §18.2.3. This is avoided in ROC fitting because the y-axis is defined by LLs and NLs, in other words all parameters of the RSM are involved. The situation with the wAFROC is not quite as severe as with fitting FROC curves but it does have a problem with degeneracy. There are some ideas on how to improve the fits, perhaps by simultaneously fitting ROC and wAFROC-operating points, which amounts to putting constraints on the parameters, but for now this remains an open research subject. Empirical wAFROC, which is the current method implemented in RJafroc, is expected to have the same issues with variability of thresholds between treatments as the empirical ROC-AUC, as discussed in §5.9. So the fitting problem has to be solved. There is no need to fit the FROC, as it should never be used as a basis of a figure of merit for human observer studies; this is distinct from the severe degeneracy issues encountered with fitting it for human observers.

The application to a large number (236) of real datasets revealed that PROPROC has serious issues. These were apparently not revealed by the millions of simulations used to validate it39. To quote the cited reference, "The new algorithm never failed to converge and produced good fits for all of the several million datasets on which it was tested". This is a good illustration of why simulations studies are not a good alternative to the method described in §18.5.1.3. In my experience this is a common misconception in this field, and is discussed further in the following chapter. Fig. 18.5, panels (J), (K) and (L) show that PROPROC, and to a lesser extent CBM, can, under some circumstances, severely overestimate performance. Recommendations regarding usage of PROPROC and CBM are deferred to Chapter 20.

The current ROC-based effort led to some interesting findings. The near equality of the AUCs predicted by the three proper ROC fitting methods, summarized in Table 18.4, has been noted, which is explained by the fact that proper ROC fitting methods represent different approaches to realizing an ideal observer, and the ideal observer must be unique, §18.6.

This chapter explores what is termed inter-correlations, between RSM and CBM parameters. Since they have similar physical meanings, the RSM and CBM separation parameters were found to be significantly correlated, = 0.86 (0.76, 0.89), as were the RSM and CBM parameters corresponding to the fraction of lesions that was actually visible, = 0.77 (0.68, 0.82). This type of correspondence between two different models can be interpreted as evidence of mutually reinforcing validity of each of the models.

The CBM method comes closest to the RSM in terms of yielding meaningful measures, but the fact that it allows the ROC curve to go continuously to (1,1) implies that it is not completely accounting for search performance, §17.8. There are two components to search performance: finding lesions and avoiding non-lesions. The CBM model accounts for finding lesions, but it does not account for avoiding suspicious regions that are non-diseased, an important characteristic of expert radiologists.

An important finding is the inverse correlation between search performance and lesion-classification performance, which suggest there could be tradeoffs in attempts to optimize them. As a simplistic illustration, a low-resolution gestalt-view of the image1, such as seen by the peripheral viewing mechanism, is expected to make it easier to rapidly spot deviations from the expected normal template described in Chapter 15. However, the observer may not be able to switch effectively between this and the high-resolution viewing mode necessary to correctly classify found suspicious region.

The main scientific conclusion of this chapter is that search-performance is the primary bottleneck in limiting observer performance. It is unfortunate that search is ignored in the ROC paradigm, usage of which is decreasing, albeit at an agonizingly slow rate. Evidence presented in this chapter should convince researchers to reconsider the focus of their investigations, most of which is currently directed at improving classification performance, which has been shown not to be the bottleneck. Another conclusion is that the three method of fitting ROC data yield almost identical AUCs. Relative to the RSM the PROPROC estimates are about 2.6% larger while CBM estimates are about 1% larger. This was a serendipitous finding that makes sense, in retrospect, but to the best of my knowledge is not known in the research community. PROPROC and to a lesser extent CBM are prone to severely overestimating performance in situations where the operating points are limited to a steep ascending section at the low end of false positive fraction scale. This parallels an earlier comment regarding the FROC, namely measurements derived from the steep part of the curve are unreliable, §17.10.1.

12.10 Appendices

12.11 Datasets

The datasets are embedded in ther `RJafroc` package. They can be viewed in the help file of the package, a partial screen-shot of which is shown next [4].

The datasets are identified in the code by dataset`dd` (where `dd` is an integer in the range `01` to `14`) as follows:

- `dataset01` "TONY" FROC dataset (Chakraborty and Svahn, 2011)

```
## List of 3
##  $ NL   : num [1:2, 1:5, 1:185, 1:3] 3 -Inf 3 -Inf 4 ...
```

[4] The raw datasets (Excel files) are in folder `R/compare-3-fits/Datasets` and file `R/compare-3-fits/loadDataFile.R` shows the correspondence between `datasetNames` and a dataset: for example, the Van Dyke dataset corresponds to file `VanDykeData.xlsx` in the `R/compare-3-fits/Datasets` folder.

dataset01	TONY FROC dataset
dataset02	Van Dyke ROC dataset
dataset03	Franken ROC dataset
dataset04	Federica Zanca FROC dataset
dataset05	John Thompson FROC dataset
dataset06	Magnus FROC dataset
dataset07	Lucy Warren FROC dataset
dataset08	Monica Penedo ROC dataset
dataset09	Nico Karssemeijer ROC dataset (CAD vs. radiologists)
dataset10	Marc Ruschin ROC dataset
dataset11	Dobbins 1 FROC dataset
dataset12	Dobbins 2 ROC dataset
dataset13	Dobbins 3 FROC dataset
dataset14	Federica Zanca real (as opposed to inferred) ROC dataset

Figure 12.4: Partial screen shot of 'RJafroc' help file showing the datasets included with the current distribution (v2.0.1).

```
##  $ LL   : num [1:2, 1:5, 1:89, 1:2] 4 4 3 -Inf 3.5 ...
##  $ LL_IL: logi NA
```

- `dataset02` "VAN-DYKE" Van Dyke ROC dataset (Van Dyke et al., 1993)

```
## List of 3
##  $ NL   : num [1:2, 1:5, 1:114, 1] 1 3 2 3 2 2 1 2 3 2 ...
##  $ LL   : num [1:2, 1:5, 1:45, 1] 5 5 5 5 5 5 5 5 5 5 ...
##  $ LL_IL: logi NA
```

- `dataset03` "FRANKEN" Franken ROC dataset (Franken et al., 1992)

```
## List of 3
##  $ NL   : num [1:2, 1:4, 1:100, 1] 3 3 4 3 3 3 4 1 1 3 ...
##  $ LL   : num [1:2, 1:4, 1:67, 1] 5 5 4 4 5 4 4 5 2 2 ...
##  $ LL_IL: logi NA
```

- `dataset04` "FEDERICA" Federica Zanca FROC dataset (Zanca et al., 2009)

```
## List of 3
##  $ NL   : num [1:5, 1:4, 1:200, 1:7] -Inf -Inf 1 -Inf -Inf ...
##  $ LL   : num [1:5, 1:4, 1:100, 1:3] 4 5 4 5 4 3 5 4 4 3 ...
##  $ LL_IL: logi NA
```

- `dataset05` "THOMPSON" John Thompson FROC dataset (Thompson et al., 2014)

```
## List of 3
##  $ NL   : num [1:2, 1:9, 1:92, 1:7] 4 5 -Inf -Inf 8 ...
##  $ LL   : num [1:2, 1:9, 1:47, 1:3] 5 9 -Inf 10 8 ...
##  $ LL_IL: logi NA
```

- `dataset06` "MAGNUS" Magnus Bath FROC dataset (Vikgren et al., 2008)

```
## List of 3
##  $ NL   : num [1:2, 1:4, 1:89, 1:17] 1 -Inf -Inf -Inf 1 ...
##  $ LL   : num [1:2, 1:4, 1:42, 1:15] -Inf -Inf -Inf -Inf -Inf ...
##  $ LL_IL: logi NA
```

- `dataset07` "LUCY-WARREN" Lucy Warren FROC dataset (Warren et al., 2014)

```
## List of 3
##  $ NL   : num [1:5, 1:7, 1:162, 1:4] 1 2 1 2 -Inf ...
##  $ LL   : num [1:5, 1:7, 1:81, 1:3] 2 -Inf 2 -Inf 1 ...
##  $ LL_IL: logi NA
```

- `dataset08` "PENEDO" Monica Penedo FROC dataset (Penedo et al., 2005)

```
## List of 3
##  $ NL   : num [1:5, 1:5, 1:112, 1] 3 2 3 2 3 0 0 4 0 2 ...
##  $ LL   : num [1:5, 1:5, 1:64, 1] 3 2 4 3 3 3 3 4 4 3 ...
##  $ LL_IL: logi NA
```

- `dataset09` "NICO-CAD-ROC" Nico Karssemeijer ROC dataset (Hupse et al., 2013)

```
## List of 3
##  $ NL   : num [1, 1:10, 1:200, 1] 28 0 14 0 16 0 31 0 0 0 ...
##  $ LL   : num [1, 1:10, 1:80, 1] 29 12 13 10 41 67 61 51 67 0 ...
##  $ LL_IL: logi NA
```

- `dataset10` "RUSCHIN" Mark Ruschin ROC dataset (Ruschin et al., 2007)

```
## List of 3
##  $ NL   : num [1:3, 1:8, 1:90, 1] 1 0 0 0 0 0 1 0 0 0 ...
##  $ LL   : num [1:3, 1:8, 1:40, 1] 2 1 1 2 0 0 0 0 0 3 ...
##  $ LL_IL: logi NA
```

- `dataset11` "DOBBINS-1" Dobbins I FROC dataset (Dobbins III et al., 2016)

```
## List of 3
##  $ NL   : num [1:4, 1:5, 1:158, 1:4] -Inf -Inf -Inf -Inf -Inf ...
##  $ LL   : num [1:4, 1:5, 1:115, 1:20] -Inf -Inf -Inf -Inf -Inf ...
##  $ LL_IL: logi NA
```

- `dataset12` "DOBBINS-2" Dobbins II ROC dataset (Dobbins III et al., 2016)

```
## List of 3
##  $ NL   : num [1:4, 1:5, 1:152, 1] -Inf -Inf -Inf -Inf -Inf ...
##  $ LL   : num [1:4, 1:5, 1:88, 1] 3 4 4 -Inf -Inf ...
##  $ LL_IL: logi NA
```

- `dataset13` "DOBBINS-3" Dobbins III FROC dataset (Dobbins III et al., 2016)

```
## List of 3
##  $ NL   : num [1:4, 1:5, 1:158, 1:4] -Inf 3 -Inf 4 5 ...
##  $ LL   : num [1:4, 1:5, 1:106, 1:15] -Inf -Inf -Inf -Inf -Inf ...
##  $ LL_IL: logi NA
```

- `dataset14` "FEDERICA-REAL-ROC" Federica Zanca *real* ROC dataset (Zanca et al., 2012)

```
## List of 3
##  $ NL   : num [1:2, 1:4, 1:200, 1] 2 2 2 2 1 3 2 2 3 1 ...
##  $ LL   : num [1:2, 1:4, 1:100, 1] 6 5 6 4 5 5 5 5 5 4 ...
##  $ LL_IL: logi NA
```

12.12 Location of PROPROC files

For each dataset PROPROC parameters were obtained by running the Windows software with PROPROC selected as the curve-fitting method. The results are saved to files that end with `proprocnormareapooled.csv` [5] contained in "R/compare-3-fits/MRMCRuns/C/", where `C` denotes the name of the dataset (for example, for the Van Dyke dataset, `C` = "VD"). Examples are shown in the next two screen-shots.

[5] In accordance with R-package policies white-spaces in the original `PROPROC` output file names have been removed.

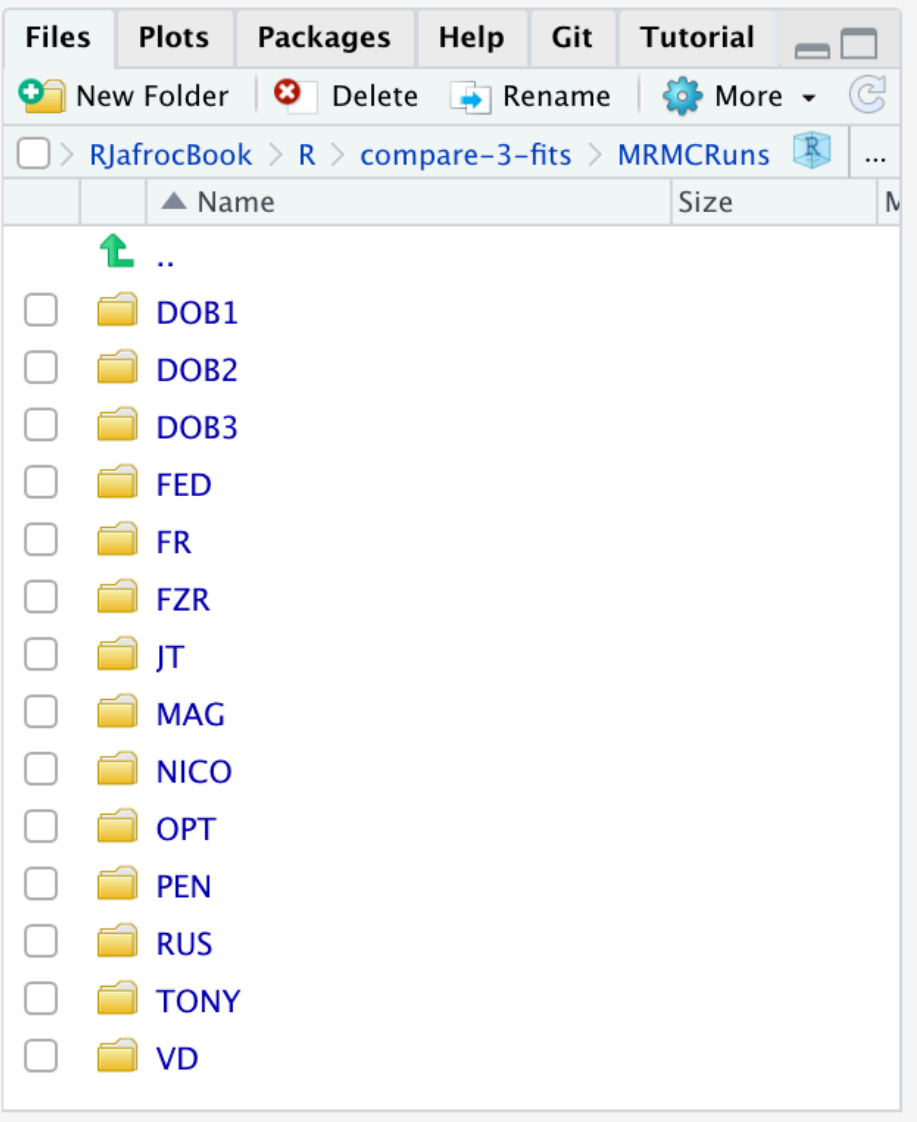


Figure 12.5: Screen shot (1 of 2) of 'R/compare-3-fits/MRMCRuns' showing the folders containing the results of PROPROC analysis on 14 datasets.

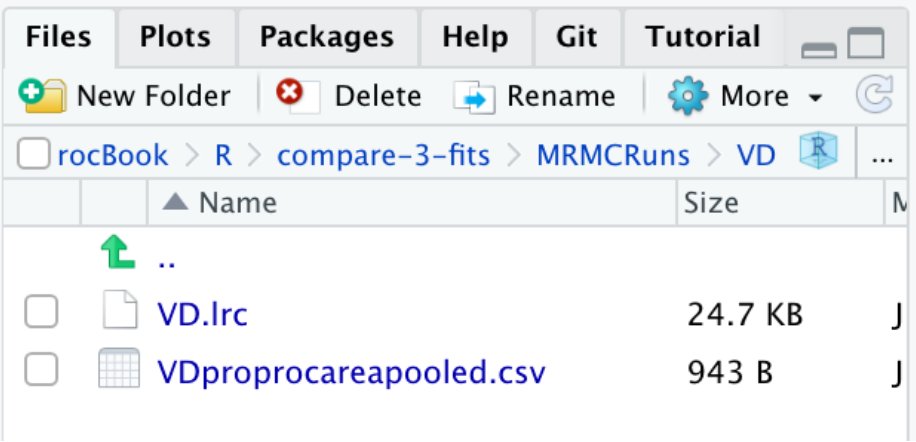
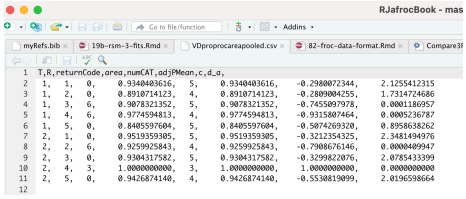


Figure 12.6: Screen shot (2 of 2) of 'R/compare-3-fits/MRMCRuns/VD' showing files containing the results of PROPROC analysis for the Van Dyke dataset.

The contents of `R/compare-3-fits/MRMCRuns/VD/VDproprocnormareapooled.csv` are shown next, see Fig. 12.7. [6] The PROPROC parameters c and d_a are in the last two columns. The column names are `T` = treatment; `R` = reader; `return-code` = undocumented value, `area` = PROPROC AUC; `numCAT` = number of ROC bins; `adjPMean` = undocumented value; `c` = c and `d_a` = d_a, are the PROPROC parameters defined in (Metz and Pan, 1999).



```
T,R,returnCode,area,numCAT,adjPMean,c,d_a,
1,  1,  0,  0.9340403616,  5,  0.9340403616,  -0.2980072344,  2.1255412315
1,  2,  0,  0.8910714123,  4,  0.8910714123,  -0.2809004255,  1.7314724686
1,  3,  6,  0.9078321352,  5,  0.9078321352,  -0.7455097978,  0.0001186957
1,  4,  6,  0.9774594813,  4,  0.9774594813,  -0.9315807464,  0.0005236787
1,  5,  0,  0.8405597604,  5,  0.8405597604,  -0.5074269320,  0.8958638262
2,  1,  0,  0.9519359305,  5,  0.9519359305,  -0.3212354325,  2.3481494976
2,  2,  6,  0.9259925843,  4,  0.9259925843,  -0.7908676146,  0.0000409947
2,  3,  0,  0.9304317582,  5,  0.9304317582,  -0.3299822076,  2.0785433399
2,  4,  3,  1.0000000000,  3,  1.0000000000,   1.0000000000,  0.0000000000
2,  5,  0,  0.9426874140,  4,  0.9426874140,  -0.5530819099,  2.0196598664
```

Figure 12.7: PROPROC output for the Van Dyke ROC data set.

12.13 Location of pre-analyzed results

The following screen shot shows the pre-analyzed files created by the function `Compare3ProperRocFits()` described below. Each file is named `allResultsC`, where `C` is the abbreviated name of the dataset (uppercase C denotes one or more uppercase characters; for example, `C` = `VD` denotes the Van Dyke dataset.).

[6] The `VD.lrc` file in this directory is the Van Dyke data formatted for input to OR DBM-MRMC 2.5.

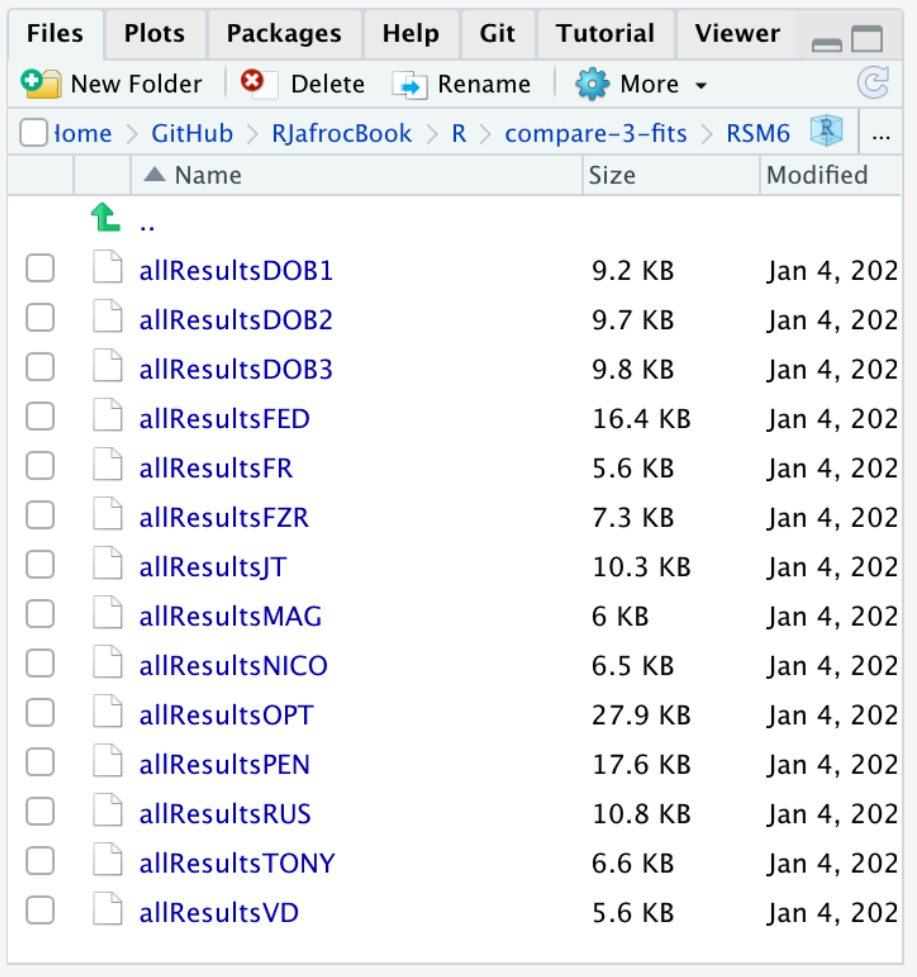


Figure 12.8: Screen shot of 'R/compare-3-fits/RSM6' showing the results files created by 'Compare3ProperRocFits()' .

12.14 Plots for Van Dyke dataset

The following plots are arranged in pairs, with the left plot corresponding to treatment 1 and the right to treatment 2.

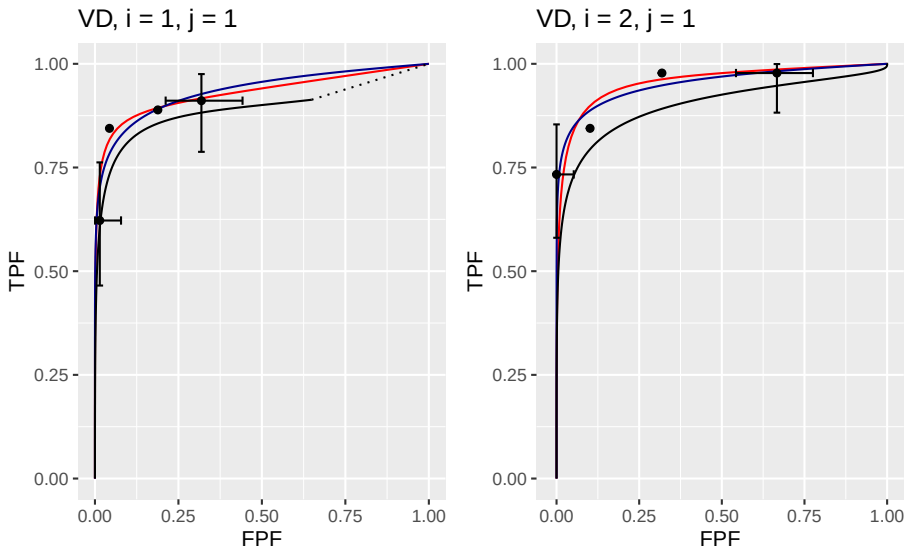


Figure 12.9: Composite plots in both treatments for Van Dyke dataset, reader 1.

The RSM parameter values for the treatment 2 plot are: μ = 5.767237, λ' = 2.7212621, ν' = 0.8021718, ζ_1 = -1.5717303. The corresponding CBM values are μ = 5.4464738, α = 0.8023609, ζ_1 = -1.4253826. The RSM and CBM μ parameters are very close and likewise the RSM ν' and CBM α parameters are very close - this is because they have similar physical meanings, which is investigated later in this chapter TBA. [The CBM does not have a parameter analogous to the RSM λ' parameter.]

The RSM parameters for the treatment 1 plot are: μ = 3.1527627, λ' = 9.9986154, ν' = 0.9899933, ζ_1 = 1.1733988. The corresponding CBM values are μ = 2.1927712, α = 0.98, ζ_1 = -0.5168848.

12.15 References

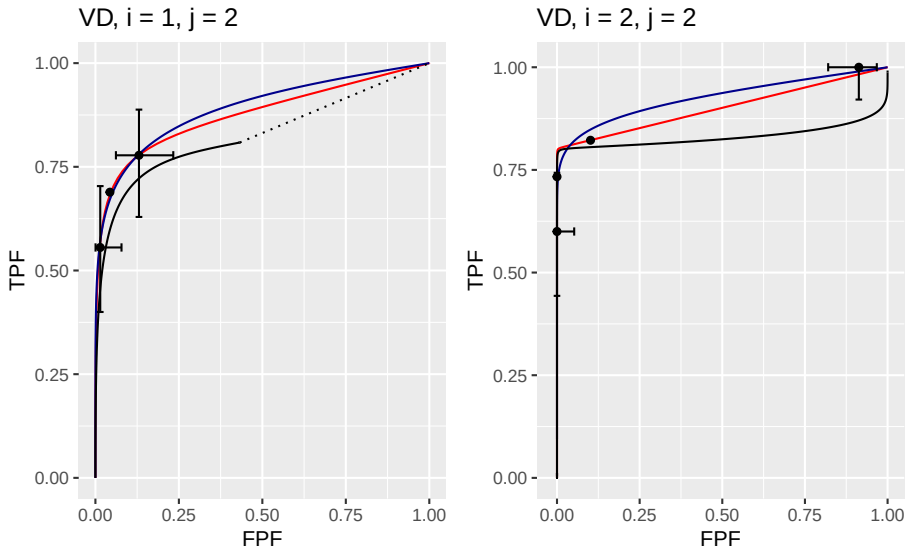


Figure 12.10: Composite plots in both treatments for Van Dyke dataset, reader 2. For treatment 2 the RSM and PROPROC fits are indistinguishable.

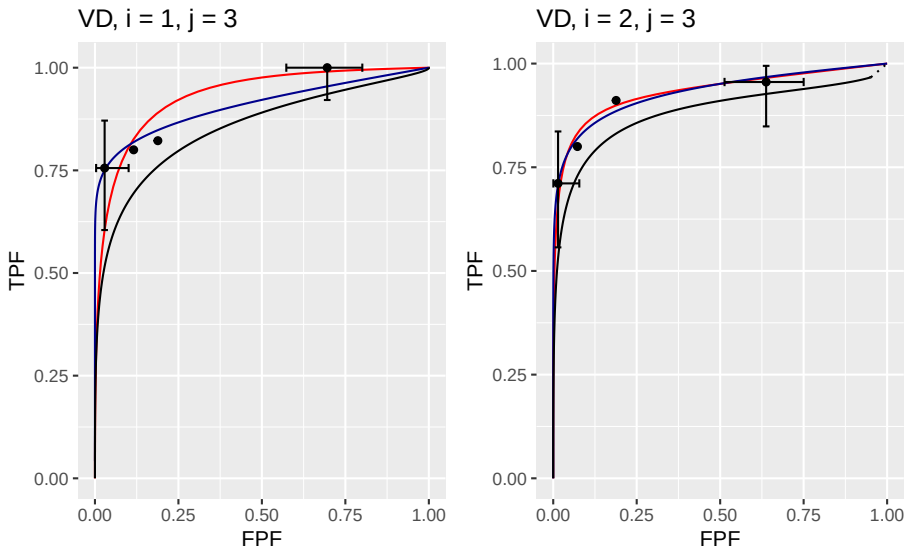


Figure 12.11: Composite plots in both treatments for Van Dyke dataset, reader 3.

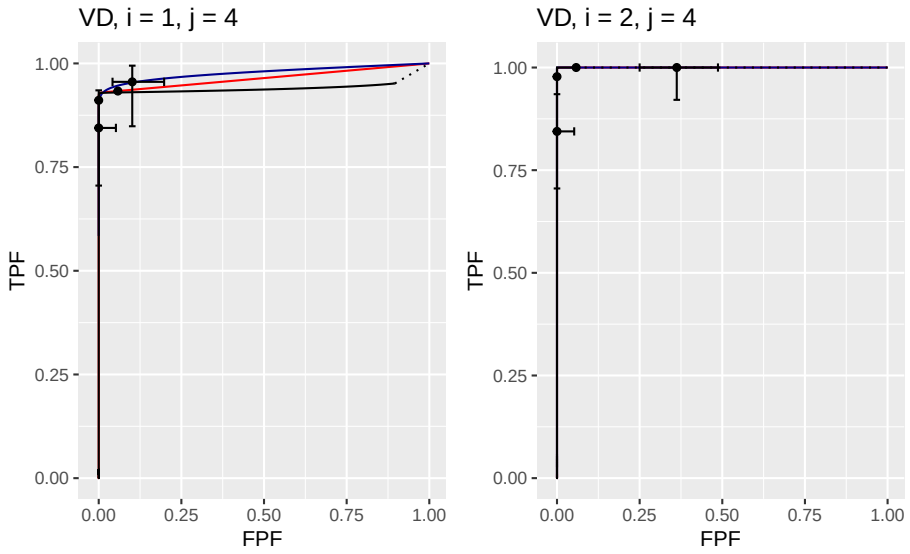


Figure 12.12: Composite plots in both treatments for Van Dyke dataset, reader 4. For treatment 2 the 3 plots are indistinguishable and each one has AUC = 1. The degeneracy is due to all operating points being on the axes of the unit square.

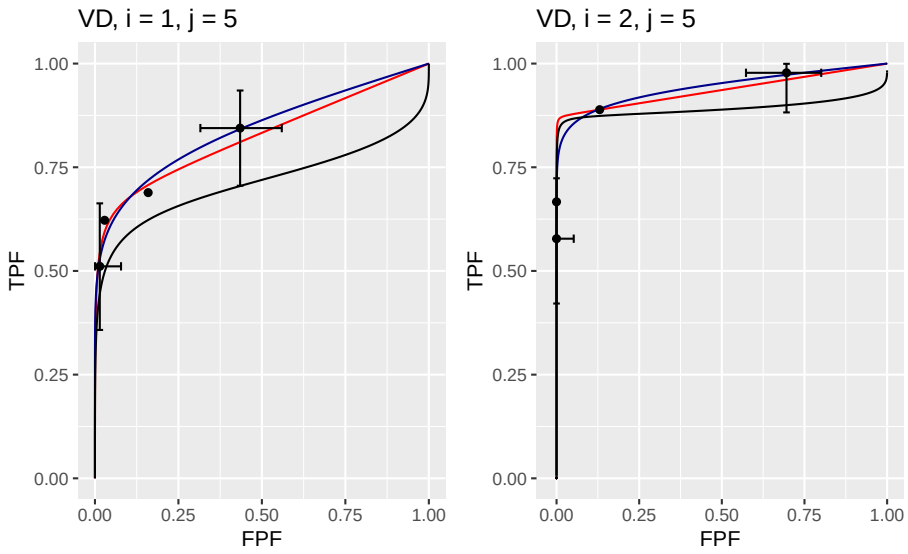


Figure 12.13: Composite plots in both treatments for Van Dyke dataset, reader 5.

CAD

Chapter 13

Standalone CAD vs. Radiologists

13.1 TBA How much finished

10%

13.2 Abstract

Computer aided detection (CAD) research for screening mammography has so far focused on measuring performance of radiologists with and without CAD. Typically a group of radiologists interpret a set of images with and without CAD assist. Standalone performance of CAD algorithms is rarely measured. The stated reason for this is that in the clinic CAD is never used alone, rather it is always used with radiologists. For this reason interest has been focused on the incremental improvement afforded by CAD.

Another reason for the lack of focus on standalone CAD performance is the lack of clear methodology for measuring standalone CAD performance. This chapter extends the methodology used in a recent study of standalone performance. The method is termed random-reader fixed case (1T-RRFC), since it only accounts for reader variability but does not account for case-variability. The extension includes the effect of case-sampling variability. Since in the proposed method CAD is treated as an additional reader within a single treatment, the method is termed one-treatment random-reader random-case (1T-RRRC) analysis. The new method is based on existing methodology allowing comparison of the average performance of readers in a single treatment to a specified value. The key modification is to regard the difference in performance between radiologists and

CAD as a figure of merit, to which the existing work is then directly applicable. The 1T-RRRC method was compared to 1T-RRFC. It was also compared to an unorthodox usage of conventional ROC (receiver operating characteristic) analysis software, termed 2T-RRRC analysis, which involves replicating the CAD ratings as many times as there are radiologists, to in effect simulate a second treatment, i.e., CAD is regarded as the second treatment. The proposed 1T-RRRC analysis has 3 random parameters as compared to 6 parameters in 2T-RRRC and one parameter in 1T-RRFC. As expected, since one is including an additional source of variability, both RRRC analyses (1T and 2T) yielded larger p-values and wider confidence intervals as compared to 1T-RRFC. For the F-statistic, degrees of freedom and p-value, both 1T-RRRC and 2T-RRRC analyses yielded exactly the same results. However, 2T-RRRC model parameter estimates were unrealistic; for example, it yields zero between-reader variance, whereas 1T-RRRC yielded the expected non-zero value. All three methods are implemented in an open-source `R` package `RJafroc.`

13.3 Keywords

Technology assessment, computer-aided detection (CAD), screening mammography, standalone performance, single-treatment multi-reader ROC analysis.

13.4 Introduction

In the US the majority of screening mammograms are analyzed by computer aided detection (CAD) algorithms (Rao et al., 2010). Almost all major imaging device manufacturers provide CAD as part of their imaging workstation display software. In the United States CAD is approved for use as a second reader (Petrick and Pastel, 2018), i.e., the radiologist first interprets the images (typically 4 views, 2 views of each breast) without CAD and then CAD information (i.e., cued suspicious regions, possibly shown with associated probabilities of malignancies) is shown and the radiologist has the opportunity to revise the initial interpretation. In response to the second reader usage, the evolution of CAD algorithms has been guided mainly by comparing observer performance of radiologists with and without CAD.

Clinical CAD systems sometimes only report the locations of suspicious regions, i.e., it may not provide ratings. However, a (continuous variable) malignancy index for every CAD-found suspicious region is available to the algorithm designer (Edwards et al., 2002). Standalone performance, i.e., performance of designer-level CAD by itself, regarded as an algorithmic reader, vs. radiologists, is rarely measured. In breast cancer screening I am aware of only one study (Hupse et al., 2013) where standalone performance was measured. [Standalone

performance has been measured in CAD for computed tomography colonography, chest radiography and three dimensional ultrasound (Hein et al., 2010; Summers et al., 2008; Taylor et al., 2006; De Boo et al., 2011; Tan et al., 2012)].

One possible reason for not measuring standalone performance of CAD is the lack of an accepted assessment methodology for such measurements. The purpose of this work is to remove that impediment. It describes a method for comparing standalone performance of designer-level CAD to radiologists interpreting the same cases and compares the method to those described in two recent publications (Hupse et al., 2013; Kooi et al., 2016).

13.5 Methods

Summarized are two recent studies of CAD vs. radiologists in mammography. This is followed by comments on the methodologies used in the two studies. The second study used multi-treatment multi-reader receiver operating characteristic (ROC) software in an unorthodox or unconventional way. A statistical model and analysis method is described that avoids unorthodox, and perhaps unjustified, use of ROC software and has fewer model parameters.

13.5.1 Studies assessing performance of CAD vs. radiologists

The first study (Hupse et al., 2013) measured performance in finding and localizing lesions in mammograms, i.e., visual search was involved, while the second study (Kooi et al., 2016) measured lesion classification performance between non-diseased and diseased regions of interest (ROIs) previously found on mammograms by an independent algorithmic reader, i.e., visual search was not involved.

13.5.1.1 Study - 1

The first study (Hupse et al., 2013) compared standalone performance of a CAD device to that of 9 radiologists interpreting the same cases (120 non-diseased and 80 with a single malignant mass per case). It used the LROC (localization ROC) paradigm (Starr et al., 1975; Metz et al., 1976; Swensson, 1996), in which the observer gives an overall rating for presence of disease (an integer 0 to 100 scale was used) and indicates the location of the most suspicious region. On a non-diseased case the rating is classified as a false positive (FP) but on a diseased case it is classified as a *correct localization* (CL) if the location is sufficiently close to the lesion, and otherwise it is classified as an *incorrect localization*. For a given reporting threshold, the number of correct localizations divided by the number of diseased cases estimates the probability of correct localization (PCL)

at that threshold. On non-diseased cases the number of false positives (FPs) divided by the number of non-diseased cases estimates the probability of a false positive, or false positive fraction (FPF), at that threshold. The plot of PCL (ordinate) vs. FPF defines the LROC curve. Study - 1 used as figures of merit (FOMs) the interpolated PCL at two values of FPF, specifically FPF = 0.05 and FPF = 0.2, denoted $PCL_{0.05}$ and $PCL_{0.2}$, respectively. The t-test between the radiologist PCL_{FPF} values and that of CAD was used to compute the two-sided p-value for rejecting the NH of equal performance. Study - 1 reported p-value = 0.17 for $PCL_{0.05}$ and p-value ≤ 0.001, with CAD being inferior, for $PCL_{0.2}$.

13.5.1.2 Study - 2

The second study (Kooi et al., 2016) used 199 diseased and 199 non-diseased ROIs extracted by an independent CAD algorithm. These were interpreted using the ROC paradigm (i.e., rating only, no localization required) by a different CAD algorithmic observer from that used to determine the ROIs, and by four expert radiologists. The figure of merit was the empirical area (AUC) under the respective ROC curves (one per radiologist and one for CAD). The p-value for the difference in AUCs between the average radiologist and CAD was determined using an unorthodox application of the Dorfman-Berbaum-Metz (Dorfman et al., 1992) multiple-treatment multiple-reader multiple-case (DBM-MRMC) software with recent modifications (Hillis et al., 2008). The unorthodox application was that in the input data file *radiologists and CAD were entered as two treatments*. In conventional (or orthodox) DBM-MRMC each reader provides two ratings per case and the data file would consist of paired ratings of a set of cases interpreted by 4 readers. To accommodate the paired data structure assumed by the software, the authors of Study - 2 *replicated the CAD ratings four times in the input data file*, as explained in the caption to Table 13.1. By this artifice they converted a single-treatment 5-reader (4 radiologists plus CAD) data file to a two-treatment 4-reader data file, in which the four readers in treatment 1 were the radiologists, and the four "readers" in treatment 2 were CAD replicated ratings. Note that for each case the four readers in the second treatment had identical ratings. In Table 1 the replicated CAD observers are labeled C1, C2, C3 and C4.

Study – 2 reported a not significant difference between CAD and the radiologists (p = 0.253).

13.5.1.3 Comments

For the purpose of this work, which focuses on the respective analysis methods, the difference in observer performance paradigms between the two studies, namely a search paradigm in Study - 1 vs. an ROI classification paradigm in Study – 2, is inconsequential. The paired t-test used in Study - 1 treats the case-sample as fixed. In other words, the analysis is not accounting for case-sampling

Table 13.1: The differences between the data structures in conventional DBM-MRMC analysis and the unorthodox application of the software used in Study - 2. There are four radiologists, labeled R1, R2, R3 and R4 interpreting 398 cases labeled 1, 2, ..., 398, in two treatments, labeled 1 and 2. Sample ratings are shown only for the first and last radiologist and the first and last case. In the first four columns, labeled "Standard DBM-MRMC", each radiologist interprets each case twice. In the next four columns, labeled "Unorthodox DBM-MRMC", the radiologists interpret each case once. CAD ratings are replicated four times to effectively create the second "treatment". The quotations emphasize that there is, in fact, only one treatment. The replicated CAD observers are labeled C1, C2, C3 and C4.

Standard DBM-MRMC				Unorthodox DBM-MRMC			
Reader	Treatment	Case	Rating	Reader	Treatment	Case	Rating
R1	1	1	75	R1	1	1	75
...	...	...	...	...	...	...	...
R1	1	398	0	R1	1	398	0
...	...	...	...	...	...	...	...
R4	1	1	50	R4	1	1	50
...	...	...	...	...	...	...	...
R4	1	398	25	R4	1	398	25
R1	2	1	45	C1	2	1	55
...	...	...	...	...	...	...	...
R1	2	398	25	C1	2	398	5
...	...	...	...	...	...	...	...
R4	2	1	95	C4	2	1	55
...	...	...	...	...	...	...	...
R4	2	398	20	C4	2	398	5

variability but it is accounting for reader variability. While not explicitly stated, the reason for the unorthodox analysis in Study – 2 was the desire to include case-sampling variability. [1]

In what follows, the analysis in Study – 1 is referred to as random-reader fixed-case (1T-RRFC) while that in Study – 2 is referred to as dual-treatment random-reader random-case (2T-RRRC).

13.5.2 The 1T-RRFC analysis model

The sampling model for the FOM is:

$$\left.\begin{array}{r} \theta_j = \mu + R_j \\ (j = 1, 2, ...J) \end{array}\right\} \tag{13.1}$$

Here μ is a constant, θ_j is the FOM for reader j, and R_j is the random contribution for reader j distributed as:

$$R_j \sim N\left(0, \sigma_R^2\right) \tag{13.2}$$

Because of the assumed normal distribution of R_j, in order to compare the readers to a fixed value, that of CAD denoted θ_0, one uses the (unpaired) t-test, as done in Study – 1. As evident from the model, no allowance is made for case-sampling variability, which is the reason for calling it the 1T-RRFC method.

Performance of CAD on a fixed dataset does exhibit within-reader variability. The same algorithm applied repeatedly to a fixed dataset does not always produce the same mark-rating data. However, this source of CAD FOM variability is much smaller than inter-reader FOM variability of radiologists interpreting the same dataset. In fact the within-reader variability of radiologists is smaller than their inter-reader variability, and within-reader variability of CAD is even smaller still. For this reason one is justified in regarded θ_0 as a fixed quantity for a given dataset. Varying the dataset will result in different values for θ_0, i.e., its case sampling variability needs to be accounted for, as done in the following analyses.

13.5.3 The 2T-RRRC analysis model

This could be termed the conventional or the orthodox method. There are two treatments and the study design is fully crossed: each reader interprets each case in each treatment, i.e., the data structure is as in the left half of Table 1.

[1] Prof. Karssemeijer (private communication, 10/27/2017) had consulted with a few ROC experts to determine if the procedure used in Study – 2 was valid, and while the experts thought it was probably valid they were not sure.

The following approach, termed 2T-RRRC, uses the Obuchowski and Rockette (OR) figure of merit sampling model (Obuchowski and Rockette, 1995) instead of the pseudovalue-based model used in the original DBM paper (Dorfman et al., 1992). For the empirical FOM, Hillis has shown the two to be equivalent (Hillis et al., 2005).

The OR model is:

$$\theta_{ij\{c\}} = \mu + \tau_i + (\tau \mathrm{R})_{ij} + \epsilon_{ij\{c\}} \tag{13.3}$$

Assuming two treatments, i ($i = 1, 2$) is the treatment index, j ($j = 1, ..., J$) is the reader index, and k ($k = 1, ..., K$) is the case index, and $\theta_{ij\{c\}}$ is a figure of merit for reader j in treatment i and case-sample $\{c\}$. A case-sample is a set or ensemble of cases, diseased and non-diseased, and different integer values of c correspond to different case-samples.

The first two terms on the right hand side of Eqn. (13.3) are fixed effects (average performance and treatment effect, respectively). The next two terms are random effect variables that, by assumption, are sampled as follows:

$$\left.\begin{aligned} R_j &\sim N\left(0, \sigma_R^2\right) \\ (\tau R)_{ij} &\sim N\left(0, \sigma_{\tau R}^2\right) \end{aligned}\right\} \tag{13.4}$$

The terms R_j represents the random treatment-independent contribution of reader j, modeled as a sample from a zero-mean normal distribution with variance σ_R^2, $(\tau R)_{ij}$ represents the random treatment-dependent contribution of reader j in treatment i, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$. The sampling of the last (error) term is described by:

$$\epsilon_{ij\{c\}} \sim N_{I \times J}\left(\vec{0}, \Sigma\right) \tag{13.5}$$

Here $N_{I \times J}$ is the $I \times J$ variate normal distribution and $\vec{0}$, a $I \times J$ length zero-vector, represents the mean of the distribution. The $\{I \times J\} \times \{I \times J\}$ dimensional covariance matrix Σ is defined by 4 parameters, Var, Cov_1, Cov_2, Cov_3, defined as follows:

$$\mathrm{Cov}\left(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}\right) = \left\{\begin{array}{l} \mathrm{Var}\ (i = i', j = j') \\ \mathrm{Cov1}\ (i \neq i', j = j') \\ \mathrm{Cov2}\ (i = i', j \neq j') \\ \mathrm{Cov3}\ (i \neq i', j \neq j') \end{array}\right\} \tag{13.6}$$

Software {U of Iowa and `RJafroc`} yields estimates of all terms appearing on the right hand side of Eqn. (13.6). Excluding fixed effects, the model represented by Eqn. (13.3) contains six parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \text{Var}, \text{Cov}_1, \text{Cov}_2, \text{Cov}_3 \tag{13.7}$$

The meanings the last four terms are described in (Hillis, 2007; Obuchowski and Rockette, 1995; Hillis et al., 2005; Chakraborty, 2017). Briefly, Var is the variance of a reader's FOMs, in a given treatment, over interpretations of different case-samples, averaged over readers and treatments; Cov_1/Var is the correlation of a reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment same-reader pairings; Cov_2/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in the same treatment, averaged over all same-treatment different-reader pairings and finally, Cov_3/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment different-reader pairings. One expects the following inequalities to hold:

$$\text{Var} \geq \text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \tag{13.8}$$

In practice, since one is usually limited to one case-sample, i.e., $c = 1$, resampling techniques (Efron and Tibshirani, 1994) – e.g., the jackknife – are used to estimate these terms.

13.5.4 The 1T-RRRC analysis model

This is the contribution of this work. The key difference from the approach in Study - 2 is to regard standalone CAD as a different reader, not as a different treatment. Therefore, needed is a single treatment method for analyzing readers and CAD, where the latter is regarded as an additional reader. Accordingly the proposed method is termed single-treatment RRRC (1T-RRRC) analysis.

The starting point is the (Obuchowski and Rockette, 1995) model for a single treatment, which for the radiologists (i.e., *excluding* CAD) interpreting in a single-treatment reduces to the following model:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{13.9}$$

$\theta_{j\{c\}}$ is the figure of merit for radiologist j ($j = 1, 2, ..., J$) interpreting case-sample $\{c\}$; R_j is the random effect of radiologist j and $\epsilon_{j\{c\}}$ is the error term. For single-treatment multiple-reader interpretations the error term is distributed as:

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{13.10}$$

The $J \times J$ covariance matrix Σ is defined by two parameters, Var and Cov_2, as follows:

$$\Sigma_{jj'} = \mathrm{Cov}\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \mathrm{Var} & j = j' \\ \mathrm{Cov}_2 & j \neq j' \end{cases} \tag{13.11}$$

The terms Var and Cov_2 are estimated using resampling methods. Using the jackknife, and denoting the FOM with case k removed by $\psi_{j(k)}$ (the index in parenthesis denotes deleted case k, and since one is dealing with a single case-sample, the case-sample index c is now superfluous). The covariance matrix is estimated using (the dot symbol represents an average over the replaced index):

$$\Sigma_{jj'}|_{\mathrm{jack}} = \frac{K-1}{K} \sum_{k=1}^{K} \left(\psi_{j(k)} - \psi_{j(\bullet)}\right)\left(\psi_{j'(k)} - \psi_{j'(\bullet)}\right) \tag{13.12}$$

The final estimates of Var and Cov_2 are averaged (indicated in the following equation by the angular brackets) over all pairings of radiologists satisfying the relevant equalities/inequalities shown just below the closing angular bracket:

$$\left.\begin{aligned} \mathrm{Var} &= \left\langle \Sigma_{jj'}|_{\mathrm{jack}} \right\rangle_{j=j'} \\ \mathrm{Cov}_2 &= \left\langle \Sigma_{jj'}|_{\mathrm{jack}} \right\rangle_{j\neq j'} \end{aligned}\right\} \tag{13.13}$$

Hillis' formulae (Hillis et al., 2005; Hillis, 2007) permit one to test the NH: $\mu = \mu_0$, where μ_0 is a pre-specified constant. One could set μ_0 equal to the performance of CAD, but that would not be accounting for the fact that the performance of CAD is itself a random variable, whose case-sampling variability needs to be accounted for.

Instead, the following model was used for the figure of merit of the radiologists and CAD ($j = 0$ is used to denote the CAD algorithmic reader):

$$\theta_{j\{c\}} = \theta_{0\{c\}} + \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{13.14}$$

$\theta_{0\{c\}}$ is the CAD figure of merit for case-sample $\{c\}$ and $\Delta\theta$ is the average figure of merit increment of the radiologists over CAD. To reduce this model to one to which existing formulae are directly applicable, one subtracts the CAD figure of merit from each radiologist's figure of merit (for the same case-sample), and defines this as the difference figure of merit $\psi_{j\{c\}}$, i.e.,

$$\psi_{j\{c\}} = \theta_{j\{c\}} - \theta_{0\{c\}} \tag{13.15}$$

Then Eqn. (13.14) reduces to:

$$\psi_{j\{c\}} = \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{13.16}$$

Eqn. (13.16) is identical in form to Eqn. (13.9) with the difference that the figure of merit on the left hand side of Eqn. (13.16) is a *difference FOM*, that between the radiologist's and CAD. Eqn. (13.16) describes a model for J radiologists interpreting a common case set, each of whose performances is measured relative to that of CAD. Under the NH the expected difference is zero: NH:$\Delta\theta = 0$. The method (Hillis et al., 2005; Hillis, 2007) for single-treatment multiple-reader analysis is now directly applicable to the model described by Eqn. (13.16).

Apart from fixed effects, the model in Eqn. (13.16) contains three parameters:

$$\sigma_R^2, \mathrm{Var}, \mathrm{Cov}_2 \tag{13.17}$$

Setting $\mathrm{Var} = 0, \mathrm{Cov}_2 = 0$ yields the 1T-RRFC model, which contains only one random parameter, namely σ_R^2. [One expects identical estimates of σ_R^2 using 1T-RRFC, 2T-RRRC or 1T-RRRC analyses.]

13.6 Software implementation

The three analyses, namely random-reader fixed-case (1T-RRFC), dual-treatment random-reader random-case (2T-RRRC) and single-treatment random-reader random-case (1T-RRRC), are implemented in `RJafroc`, an R-package (Chakraborty et al., 2020).

The following code shows usage of the software to generate the results corrsponding to the three analyses. Note that `datasetCadLroc` is the LROC dataset and `dataset09` is the corresponding ROC dataset.

```
RRFC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRFC")
RRRC_2T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "2T-RRRC")
RRRC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRRC")

RRFC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRFC")
RRRC_2T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "2T-RRRC")
RRRC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRRC")

RRFC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
```

```
FOM = "PCL", FPFValue = 1, method = "1T-RRFC")
RRRC_2T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "2T-RRRC")
RRRC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRRC")

RRFC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRFC")
RRRC_2T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "2T-RRRC")
RRRC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRRC")
```

The results are organized as follows:

- `RRFC_1T_PCL_0_05` contains the results of 1T-RRFC analysis for figure of merit = $PCL_{0.05}$.
- `RRRC_2T_PCL_0_05` contains the results of 2T-RRFC analysis for figure of merit = $PCL_{0.05}$.
- `RRRC_1T_PCL_0_05` contains the results of 1T-RRFC analysis for figure of merit = $PCL_{0.05}$.
- `RRFC_1T_PCL_0_2` contains the results of 1T-RRFC analysis for figure of merit = $PCL_{0.2}$.
- `RRRC_2T_PCL_0_2` contains the results of 2T-RRRC analysis for figure of merit = $PCL_{0.2}$.
- `RRRC_1T_PCL_0_2` contains the results of 1T-RRRC analysis for figure of merit = $PCL_{0.2}$.
- `RRFC_1T_AUC` contains the results of 1T-RRFC analysis for the Wilcoxon figure of merit.
- `RRRC_2T_AUC` contains the results of 2T-RRRC analysis for the Wilcoxon figure of merit.
- `RRRC_1T_AUC` contains the results of 1T-RRRC analysis for the Wilcoxon figure of merit.

The structures of these objects are illustrated with examples in the Appendix.

13.7 Results

The three methods, in historical order 1T-RRFC, 2T-RRRC and 1T-RRRC, were applied to an LROC dataset similar to that used in Study – 1 (I thank Prof. Karssemeijer for making this dataset available).

Shown next, Table 13.2, are the significance testing results corresponding to the three analyses.

Table 13.2: Significance testing results of the analyses for an LROC dataset. Three sets of results, namely RRRC, 2T-RRRC and 1T-RRRC, are shown for each figure of merit (FOM). Because it is accounting for an additional source of variability, each of the rows labeled RRRC yields a larger p-value and wider confidence intervals than the corresponding row labeled 1T-RRFC. [θ_0 = FOM CAD; $\theta_\bullet$ = average FOM of radiologists; $\psi_\bullet$ = average FOM of radiologists minus CAD; CI= 95 percent confidence interval of quantity indicated by the subscript, F = F-statistic; ddf = denominator degrees of freedom; p = p-value for rejecting the null hypothesis: $\psi_\bullet = 0$.]

FOM	Analysis	θ_0	CI_{θ_0}	$\theta_\bullet$	$CI_{\theta_\bullet}$	$\psi_\bullet$	$CI_{\psi_\bullet}$	F	ddf	p
PCL_0_05	1T-RRFC	4.5e-01	0	4.93e-01	(4.18e-01,5.68e-01)	4.33e-02	(-3.16e-02,1.18e-01)	1.77e+00	8e+00	2.2e-01
	2T-RRRC		(2.58e-01,6.42e-01)		(3.76e-01,6.11e-01)		(-1.57e-01,2.44e-01)	1.79e-01	7.84e+02	6.7e-01
	1T-RRRC		NA		(2.93e-01,6.94e-01)					
PCL_0_2	1T-RRFC	5.92e-01	0	7.1e-01	(6.69e-01,7.51e-01)	1.19e-01	(7.78e-02,1.59e-01)	4.5e+01	8e+00	1.51e-04
	2T-RRRC		(4.78e-01,7.05e-01)		(6.33e-01,7.87e-01)		(4.45e-03,2.33e-01)	4.16e+00	9.37e+02	4.2e-02
	1T-RRRC		NA		(5.96e-01,8.24e-01)					
PCL_1	1T-RRFC	6.75e-01	0	7.83e-01	(7.4e-01,8.27e-01)	1.08e-01	(6.48e-02,1.52e-01)	3.3e+01	8e+00	4.33e-04
	2T-RRRC		(5.71e-01,7.79e-01)		(7.12e-01,8.54e-01)		(4.5e-03,2.12e-01)	4.2e+00	4.93e+02	4.1e-02
	1T-RRRC		NA		(6.8e-01,8.87e-01)					
Wilcoxon	1T-RRFC	8.17e-01	0	8.49e-01	(8.26e-01,8.71e-01)	3.17e-02	(8.96e-03,5.45e-02)	1.03e+01	8e+00	1.24e-02
	2T-RRRC		(7.52e-01,8.82e-01)		(8.07e-01,8.9e-01)		(-3.1e-02,9.45e-02)	9.86e-01	8.78e+02	3.2e-01
	1T-RRRC		NA		(7.86e-01,9.11e-01)					

Results are shown for the following FOMs: $\text{PCL}_{0.05}$, $\text{PCL}_{0.2}$, PCL_1, and the empirical area (AUC) under the ROC curve estimated by the Wilcoxon statistic. The first two FOMs are identical to those used in Study – 1. Columns 3 and 4 list the CAD FOM θ_0 and its 95% confidence interval CI_{θ_0}, columns 5 and 6 list the average radiologist FOM $\theta_\bullet$ (the dot symbol represents an average over the radiologist index) and its 95% confidence interval $CI_{\theta_\bullet}$, columns 7 and 8 list the average difference FOM $\psi_\bullet$, i.e., radiologist minus CAD, and its 95% confidence interval $CI_{\psi_\bullet}$, and the last three columns list the F-statistic, the denominator degrees of freedom (ddf) and the p-value for rejecting the null hypothesis. The numerator degree of freedom of the F-statistic, not listed, is unity.

In Table 13.2 identical values in adjacent cells in vertical columns have been replaced by the common values. The last three columns show that 2T-RRRC and 1T-RRRC analyses yield *identical F-statistics, ddf and p-values.* So the intuition of the authors of Study – 2, that the unorthodox method of using DBM – MRMC software to account for both reader and case-sampling variability,

turns out to be correct. If interest is solely in these statistics one is justified in using the unorthodox method.

Commented on next are other aspects of the results evident in Table 13.2.

1. Where a direct comparison is possible, namely 1T-RRFC analysis using and as FOMs, the p-values in Table 13.2 are similar to those reported in Study – 1.
2. All FOMs (i.e., θ_0, $\theta_\bullet$ and $\psi_\bullet$) in Table 13.2 are independent of the method of analysis. However, the corresponding confidence intervals (i.e., CI_{θ_0}, $CI_{\theta_\bullet}$ and $CI_{\psi_\bullet}$) depend on the analyses.
3. Since 1T-RRFC analysis ignores case sampling variability, the CAD figure of merit is a constant, with zero-width confidence interval. For compactness the CI is listed as 0, rather than two identical values in parentheses. The confidence interval listed for 2T-RRRC analyses is centered on the corresponding CAD value, as are all confidence intervals in Table 13.2.
4. The LROC FOMs increase as the value of FPF (the subscript) increases. This should be obvious, as PCL increases as FPF increases, a general feature of any partial curve based figure of merit.
5. The area (AUC) under the ROC is larger than the largest PCL value, i.e., $AUC \geq \text{PCL}_1$. This too should be obvious from the general features of the LROC (Swensson, 1996).
6. The p-value for either RRRC analyses (2T or 1T) is larger than the corresponding 1T-RRFC value. Accounting for case-sampling variability increases the p-value, leading to less possibility of finding a significant difference.
7. Partial curve-based FOMs, such as PCL_{FPF}, lead, depending on the choice of FPF, to different conclusions. The p-values generally decrease as FPF increases. Measuring performance on the steep part of the LROC curve (i.e., small FPF) needs to account for greater reader variability and risks lower statistical power.
8. Ignoring localization information (i.e., using the AUC FOM) led to a not-significant difference between CAD and the radiologists ($p = 0.3210$), while the corresponding FOM yielded a significant difference ($p = 0.0409$). Accounting for localization leads to a less "noisy" measurement. This has been demonstrated for the LROC paradigm (Swensson, 1996) and I have demonstrated this for the FROC paradigm (Chakraborty, 2008).
9. For 1T-RRRC analysis, is listed as NA, for not applicable, since is not a model parameter, see Eqn. (13.16).

Shown next, Table 13.3, are the model-parameters corresponding to the three analyses.

Table 13.3: Parameter estimates for the analyses; NA = not applicable.

FOM	Analysis	σ_R^2	$\sigma_{\tau R}^2$	Cov1	Cov2	Cov3	Var
PCL_0_05	1T-RRFC	9.5e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.84e-18	-5.71e-03	1.31e-03	6.01e-03	1.31e-03	1.65e-02
	1T-RRRC	9.5e-03	NA	NA	9.4e-03	NA	3.03e-02
PCL_0_2	1T-RRFC	2.81e-03	NA	NA	NA	NA	NA
	2T-RRRC	-7.59e-19	2.65e-04	7.61e-04	2.29e-03	7.61e-04	3.43e-03
	1T-RRRC	2.81e-03	NA	NA	3.07e-03	NA	5.34e-03
PCL_1	1T-RRFC	3.2e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.63e-18	1e-03	6.43e-04	1.86e-03	6.43e-04	2.46e-03
	1T-RRRC	3.2e-03	NA	NA	2.44e-03	NA	3.64e-03
Wilcoxon	1T-RRFC	8.78e-04	NA	NA	NA	NA	NA
	2T-RRRC	2.98e-19	2.01e-04	2.62e-04	7.24e-04	2.62e-04	9.62e-04
	1T-RRRC	8.78e-04	NA	NA	9.24e-04	NA	1.4e-03

The following characteristics are evident from Table 13.3.

1. For 2T-RRRC analyses $\sigma_R^2 = 0$. Actually, the analysis yielded very small values, of the order of 10^{-18} to 10^{-19}, which, being smaller than double precision accuracy, were replaced by zeroes in Table 13.2. $\sigma_R^2 = 0$ is clearly an incorrect result as the radiologists do not have identical performance. In contrast, 1T-RRRC analyses yielded more realistic values, identical to those obtained by 1T-RRFC analyses, and consistent with expectation – see comment following Eqn. (15).
2. Because 2T analysis found zero reader variability, it follows from the definitions of the covariances (Obuchowski and Rockette, 1995), that $Cov_1 = Cov_3 = 0$, as evident in the table.
3. When they can be compared (i.e., σ_R^2, Cov_2 and Var), all variance and covariance estimates were smaller for the 2T method than for the 1T method.
4. For the 2T method the expected inequalities, Eqn. (13.8), are not obeyed (specifically, $Cov_1 \geq Cov_2 \geq Cov_3$ is not obeyed).

For an analysis method to be considered statistically valid it needs to be tested with simulations to determine if it has the proper null hypothesis behavior. The design of a ratings simulator to statistically match a given dataset is addressed in Chapter 23 of reference (Chakraborty, 2017). Using this simulator, the 1T-RRRC method had the expected null hypothesis behavior (Table 23.5, ibid).

13.8 Discussion

TBA TODOLAST The argument often made for not measuring standalone performance is that since CAD will be used only as a second reader, it is only necessary to measure performance of radiologists without and with CAD. It has been stated (Nishikawa and Pesce, 2011):

> High stand-alone performance is neither a necessary nor a sufficient condition for CAD to be truly useful clinically.

Assessing CAD utility this way, i.e, by measuring performance with and without CAD, may have inadvertently set a low bar for CAD to be considered useful. As examples, CAD is not penalized for missing cancers as long as the radiologist finds them and CAD is not penalized for excessive false positives (FPs) as long as the radiologist ignores them. Moreover, since both such measurements include the variability of radiologists, there is additional noise introduces that presumably makes it harder to determine if the CAD system is optimal.

Described is an extension of the analysis used in Study – 1 that accounts for case sampling variability. It extends (Hillis et al., 2005) single-treatment analysis to a situation where one of the "readers" is a special reader, and the desire is to compare performance of this reader to the average of the remaining readers. The method, along with two other methods, was used to analyze an LROC data set using different figures of merit.

1T-RRRC analyses yielded identical overall results (specifically the F-statistic, degrees of freedom and p-value) to those yielded by the unorthodox application of DBM-MRMC software, termed 2T-RRRC analyses, where the CAD reader is regarded as a second treatment. However, the values of the model parameters of the dual-treatment analysis lacked clear physical meanings. In particular, the result $\sigma_R^2 = 0$ is clearly an artifact. One can only speculate as to what happens when software is used in a manner that it was not designed for: perhaps finding that all readers in the second treatment have identical FOMs led the software to yield $\sigma_R^2 = 0$. The single-treatment model has half as many parameters as the dual-treatment model and the parameters have clear physical meanings and the values are realistic.

The paradigm used to collect the observer performance data - e.g., receiver operating characteristic (ROC) (Metz, 1986), free-response ROC (FROC) (Chakraborty et al., 1986), location ROC (LROC) (Starr et al., 1975) or region of interest (ROI) (Obuchowski et al., 2000) - is irrelevant – all that is needed is a scalar performance measure for the actual paradigm used. In addition to PCL and AUC, RJafroc currently implements the partial area under the LROC, from FPF = 0 to a specified value as well other FROC-paradigm based FOMs.

While there is consensus that CAD works for microcalcifications, for masses its performance is controversial27,28. Two large clinical studies TBA 29,30

(222,135 and 684,956 women, respectively) showed that CAD actually had a detrimental effect on patient outcome. A more recent large clinical study has confirmed the negative view of CAD31 and there has been a call for ending Medicare reimbursement for CAD interpretations32.

In my opinion standalone performance is the most direct measure of CAD performance. Lack of clear-cut methodology to assess standalone CAD performance may have limited past CAD research. The current work hopoefully removes that impediment. Going forward, assessment of standalone performance of CAD vs. expert radiologists is strongly encouraged.

13.9 Appendix

The structures of theR objects generated by the software are illustrated with three examples.

13.9.1 Example 1

The first example shows the structure of 'RRFC_1T_PCL_0_2

```
print(fom_individual_rad)
#>         rdr1 rdr2    rdr3  rdr4       rdr5       rdr6   rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats)
#>       fomCAD  avgRadFom avgDiffFom        varR     Tstat df          pval
#> 1 0.59166667 0.71017278 0.11850612 0.002808612 6.7083568  8 0.00015139664
print(ConfidenceIntervals)
#>       CIAvgRadFom CIAvgDiffFom
#> Lower  0.66943619  0.077769525
#> Upper  0.75090938  0.159242710
```

The results are displayed as three data frames.

The first data frame :

- `fom_individual_rad` shows the figures of merit for the nine radiologists in the study.

The next data frame summarizes the statistics.

- `fomCAD` is the figure of merit for CAD.
- `avgRadFom` is the average figure of merit of the nine radiologists in the study.

- `avgDiffFom` is the average difference figure of merit, RAD - CAD.
- `varR` is the variance of the figures of merit for the nine radiologists in the study.
- `Tstat` is the t-statistic for testing the NH that the average difference FOM `avgDiffFom` is zero, whose square is the F-statistic.
- `df` is the degrees of freedom of the t-statistic.
- `pval` is the p-value for rejecting the NH. In the example shown below the value is highly signficant.

The last data frame summarizes the 95 percent confidence intervals.

- `CIAvgRadFom` is the 95 percent confidence interval, listed as pairs `Lower`, `Upper`, for `avgRadFom`.
- `CIAvgDiffFom` is the 95 percent confidence interval for `avgDiffFom`.
- If the pair `CIAvgDiffFom` excludes zero, the difference is statistically significant.
- In the example the interval excludes zero showing that the FOM difference is significant.

13.9.2 Example 2

The next example shows the structure of `RRRC_2T_PCL_0_2`.

```
print(fom_individual_rad)
#>         rdr1 rdr2    rdr3  rdr4       rdr5       rdr6   rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats1)
#>       fomCAD  avgRadFom avgDiffFom
#> 1 0.59166667 0.71017278 0.11850612
print(stats2)
#>              varR         varTR          cov1         cov2          cov3
#> 1 -7.5894152e-19 0.00026488983 0.00076136841 0.0022942211 0.00076136841
#>            Var     FStat        df        pval
#> 1 0.0034336373 4.1576797 937.24371 0.041726262
```

In addition to the quantities defined previously, the output contains the covariance matrix for the Obuchowski-Rockette model, summarized in Eqn. (13.3) – Eqn. (13.6).

- `varTR` is $\sigma^2_{\tau R}$.
- `cov1` is Cov_1.
- `cov2` is Cov_2.
- `cov3` is Cov_3.

- `Var` is Var.
- `FStat` is the F-statistic for testing the NH.
- `ndf` is the numerator degrees of freedom, equal to unity.
- `df` is denominator degrees of freedom of the F-statistic for testing the NH.
- `Tstat` is the t-statistic for testing the NH that the average difference FOM `avgDiffFom` is zero.
- `pval` is the p-value for rejecting the NH. In the example shown below the value is signficant.

Notice that including the variability of cases results in a higher p-value for 2T-RRRC as compared to 1T-RRFC.

Shown next are the confidence interval statistics `x$ciAvgRdrEachTrt` for the two treatments ("trt1" = CAD, "trt2" = RAD):

```
print(x$ciAvgRdrEachTrt)
#>        Estimate      StdErr        DF    CILower    CIUpper         Cov2
#> trt1 0.59166667 0.058028349       Inf 0.47793319 0.70540014 0.0033672893
#> trt2 0.71017278 0.039156365 193.10832 0.63294372 0.78740185 0.0012211529
```

- `Estimate` contains the difference FOM estimate.
- `StdErr` contains the standard estimate of the difference FOM estimate.
- `DF` contains the degrees of freedom of the t-statistic.
- `t` contains the value of the t-statistic.
- `PrGtt` contains the probability of exceeding the magnitude of the t-statistic.
- `CILower` is the lower confidence interval for the difference FOM.
- `CIUpper` is the upper confidence interval for the difference FOM.

Shown next are the confidence interval statistics `x$ciDiffFom` between the two treatments ("trt1-trt2" = CAD - RAD):

```
print(x$ciDiffFom)
#>             Estimate      StdErr        DF         t       PrGTt     CILower
#> trt2-trt1 0.11850612 0.058118615 937.24371 2.0390389 0.041726262 0.004448434
#>             CIUpper
#> trt2-trt1 0.2325638
```

The difference figure of merit statistics are contained in a dataframe `x$ciDiffFom` with elements:

- `Estimate` contains the difference FOM estimate.
- `StdErr` contains the standard estimate of the difference FOM estimate.
- `DF` contains the degrees of freedom of the t-statistic.

- `t` contains the value of the t-statistic.
- `PrGtt` contains the probability of exceeding the magnitude of the t-statistic.
- `CILower` is the lower confidence interval for the difference FOM.
- `CIUpper` is the upper confidence interval for the difference FOM.

The figures of merit statistic for the two treatments, 1 is CAD and 2 is RAD.

- `trt1`: statistics for CAD.
- `trt2`: statistics for RAD.
- `Cov2`: Cov_2 calculated over individual treatments.

13.9.3 Example 3

The last example shows the structure of `RRRC_1T_PCL_0_2`.

```
RRRC_1T_PCL_0_2
#> $fomCAD
#> [1] 0.59166667
#>
#> $fomRAD
#> [1] 0.69453125 0.65000000 0.80625000 0.72500000 0.65982143 0.76845238 0.73750000
#> [8] 0.67500000 0.67500000
#>
#> $avgRadFom
#> [1] 0.71017278
#>
#> $CIAvgRad
#> [1] 0.59611510 0.82423047
#>
#> $avgDiffFom
#> [1] 0.11850612
#>
#> $CIAvgDiffFom
#> [1] 0.004448434 0.232563801
#>
#> $varR
#> [1] 0.002808612
#>
#> $varError
#> [1] 0.00534453777
#>
#> $cov2
#> [1] 0.0030657054
```

```
#>
#> $Tstat
#>      rdr2
#> 2.0390389
#>
#> $df
#>      rdr2
#> 937.24371
#>
#> $pval
#>        rdr2
#> 0.041726262
```

The differences from `RRFC_1T_PCL_0_2` are listed next:

- `varR` is σ_R^2 of the single treatment model for comparing CAD to RAD, Eqn. (13.17).
- `cov2` is Cov_2 of the single treatment model for comparing CAD to RAD.
- `varError` is Var of the single treatment model for comparing CAD to RAD.

Notice that the `RRRC_1T_PCL_0_2` p value, i.e., 0.04172626, is identical to that of `RRRC_2T_PCL_0_2`, i.e., 0.04172626.

13.10 References

Chapter 14

Optimal operating point on FROC

14.1 TBA How much finished

80%

Discussion and Intro need more work; coding is done

14.2 Introduction

This chapter deals with finding the optimal reporting threshold of an algorithmic observer, such as CAD. We assume that designer level FROC data is available for the algorithm, i.e., the data consists of mark-rating pairs, with continuous-scale ratings, and a decision needs to be made as to the optimal reporting threshold, i.e., the minimum rating of a mark before it is shown to the radiologist. This is a familiar problem faced by a CAD algorithm designer.

The problem has been solved in the context of ROC analysis (Metz, 1978), namely, the optimal operating point on the ROC corresponds to a slope determined by disease prevalence and the cost of decisions in the four basic binary paradigm categories: true and false positives and true and false negatives. In practice the costs are difficult to quantify. However, for equal numbers of diseased and non-diseased cases and equal costs it can be shown that the slope of the ROC curve at the optimal point is unity. For a proper ROC curve this corresponds to the point that maximizes the Youden-index (Youden, 1950), defined as the sum of sensitivity and specificity minus one. Typically it is maximized at the point that is closest to the (0,1) corner of the ROC.

CAD produces FROC data and lacking a procedure for setting it analytically, CAD manufacturers, in consultation with radiologists, set site-specific reporting thresholds. For example, if radiologists at a site are comfortable with more false marks as the price of potentially greater lesion-level sensitivity, the reporting threshold for them is adjusted downward.

This chapter describes an analytic method for finding the optimal reporting threshold. The method is based on maximizing AUC (area under curve) under the wAFROC curve. The method is compared to the Youden-index based method.

14.3 Methods

The ROC, FROC and wAFROC curves are completely defined by the RSM (radiological search model) parameters: λ, ν, μ and ζ_1, which have the following meanings:

- The μ parameter is the perceptual signal to noise ratio of lesions measured under location-known-exactly conditions. Higher values of μ lead to increased overall performance of the algorithm.
- The intrinsic λ parameter determines the number of non-lesion localizations, NLs, per case (location level "false positives"). Lower values lead to fewer NL marks and increased algorithm performance. It is related to the physical λ' parameter by $\lambda' = \lambda/\mu$. The physical parameter λ' equals the mean of the assumed Poisson distribution of NLs per case.
- The intrinsic ν parameter determines the probability of a lesion localizations, LLs, (location level "true positives"). Higher values lead to more LL marks. It is related to the physical ν' parameter by $\nu' = 1 - \exp(-\mu\nu)$. The physical parameter ν' equals the success probability of the assumed binomial distribution of LLs per case.
- The ζ_1 parameter determines if a suspicious region found by the algorithm is actually marked. The higher this value, the fewer the reported marks. The objective is to optimize ζ_1.

In the following sections each of the first three parameters is varied in turn and the corresponding optimal ζ_1 determined by maximizing one of two figures of merit (FOMs), namely, the wAFROC-AUC and the Youden-index.

14.3.1 Functions to be maximized

The functions to be maximized, `wAFROC` and `Youden`, are defined next:

- wAFROC-AUC is computed by `UtilAnalyticalAucsRSM`. Lines 2 - 19 returns `-wAFROC`, the *negative* of wAFROC-AUC. The negative sign is needed because the `optimize()` function, used later, finds the *minimum* of wAFROC-AUC. The first argument is ζ_1, the variable to be varied to find the maximum. The remaining arguments passed to the function, needed to calculate the FOMs, are μ, λ, ν, `lesDistr` and `relWeights`. The last two specify the number of lesions per case and their weights. The following code below uses `lesDistr = c(0.5,0.5)`, i.e., half of the diseased cases contain one lesion and the rest contain two lesions, and `relWeights = c(0.5,0.5)`, which specifies equal weights to all lesions.

- The Youden-index is defined as the sum of sensitivity and specificity minus 1. Sensitivity is computed by `RSM_yROC` and specificity by `(1 - RSM_xROC)`. Lines 22 - 42 returns `-Youden`, the *negative* of the Youden-index.

```
1
2  wAFROC <- function (
3    zeta1,
4    mu,
5    lambda,
6    nu,
7    lesDistr,
8    relWeights) {
9    x <- UtilAnalyticalAucsRSM(
10     mu,
11     lambda,
12     nu, zeta1,
13     lesDistr,
14     relWeights)$aucwAFROC
15   # return negative of aucwAFROC
16   # (as optimize finds minimum of function)
17   return(-x)
18
19 }
20
21
22 Youden <- function (
23   zeta1,
24   mu,
25   lambda,
26   nu,
27   lesDistr,
28   relWeights) {
29   # add sensitivity and specificity
30   # and subtract 1, i.e., Youden's index
```

```
31    x <- RSM_yROC(
32      zeta1,
33      mu,
34      lambda,
35      nu,
36      lesDistr) +
37      (1 - RSM_xROC(zeta1, lambda/mu)) - 1
38    # return negative of Youden-index
39    # (as optimize finds minimum of function)
40    return(-x)
41
42  }
```

14.3.2 Vary lambda

For $\mu = 2$ and $\nu = 1$, wAFROC-AUC and Youden-index based optimizations were performed for $\lambda = 1, 5, 10, 15$. The following quantities were calculated:

- `zetaOptArr`, a [2,4] array, the optimal thresholds ζ_1;
- `fomMaxArr`, a [2,4] array, the maximized values of wAFROC-AUC, using either wAFROC based or Youden-index based optimization; note that in the latter we report wAFROC-AUC even though the optimized quantity is the Youden-index.
- `rocAucArr`, a [2,4] array, the AUCs under the ROC curves corresponding to optimizations based on wAFROC-AUC or Youden-index;
- `nlfOptArr`, a [2,4] array, the abscissa of the optimal reporting point on the FROC curve corresponding to optimizations based on wAFROC-AUC or Youden-index;
- `llfOptArr`, a [2,4] array, the ordinate of the optimal reporting point on the FROC curve corresponding to optimizations based on wAFROC-AUC or Youden-index.

In each of these arrays the first index, `y` in the following code, denotes whether wAFROC-AUC is being maximized (`y` = 1, see lines 14 - 20) - or if Youden-index is being optimized (`y` = 2, see lines 39 - 45). The second index `i` in the following code, corresponds to λ.

```
1  mu <- 2
2  nu <- 1
3  lambdaArr <- c(1,5,10,15)
4  fomMaxArr <- array(dim = c(2,length(lambdaArr)))
5  zetaOptArr <- array(dim = c(2,length(lambdaArr)))
```

```
rocAucArr <- array(dim = c(2,length(lambdaArr)))
nlfOptArr <- array(dim = c(2,length(lambdaArr)))
llfOptArr <- array(dim = c(2,length(lambdaArr)))
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
for (y in 1:2) {
  for (i in 1:length(lambdaArr)) {
    if (y == 1) {
      x <- optimize(wAFROC,
                    interval = c(-5,5),
                    mu,
                    lambdaArr[i],
                    nu,
                    lesDistr,
                    relWeights)
      zetaOptArr[y,i] <- x$minimum
      fomMaxArr[y,i] <- -x$objective # safe to use objective here
      rocAucArr[y,i] <- UtilAnalyticalAucsRSM(
        mu,
        lambdaArr[i],
        nu,
        zeta1 = x$minimum,
        lesDistr,
        relWeights)$aucROC
      nlfOptArr[y,i] <- RSM_xFROC(
        z = x$minimum,
        mu,
        lambda = lambdaArr[i])
      llfOptArr[y,i] <- RSM_yFROC(
        z = x$minimum,
        mu,
        nu)
    } else if (y == 2) {
      x <- optimize(Youden,
                    interval = c(-5,5),
                    mu,
                    lambdaArr[i],
                    nu,
                    lesDistr,
                    relWeights)
      zetaOptArr[y,i] <- x$minimum
      fomMaxArr[y,i] <- UtilAnalyticalAucsRSM(
        mu,
        lambdaArr[i],
        nu,
```

```
        zeta1 = x$minimum,
        lesDistr,
        relWeights)$aucwAFROC
      rocAucArr[y,i] <- UtilAnalyticalAucsRSM(
        mu,
        lambdaArr[i],
        nu,
        zeta1 = x$minimum,
        lesDistr,
        relWeights)$aucROC
      nlfOptArr[y,i] <- RSM_xFROC(
        z = x$minimum,
        mu,
        lambda = lambdaArr[i])
      llfOptArr[y,i] <- RSM_yFROC(
        z = x$minimum, mu, nu)
    } else stop("incorrect y")
  }
}
```

Table 14.1 summarizes the results. The column labeled "FOM" shows the quantity being maximized, "lambda" corresponds to the 4 values of λ, "zeta1" is the optimal value of ζ_1 that maximizes FOM, "wAFROC" is the wAFROC-AUC, "ROC" is the AUC under the ROC curve, i.e., ROC-AUC, and "OptOpPt" is the optimal operating point on the FROC curve.

For the wAFROC-AUC based optimizations (first four rows of table), as λ increases:

- The optimal threshold ζ_1 increases;
- wAFROC-AUC decreases;
- ROC-AUC decreases;
- The optimal operating point moves to lower LLF values, i.e., lower values of location-level "sensitivity".
- The advantage of wAFROC-AUC over Youden-index based optimizations, as measured by the differences between the corresponding wAFROC-AUCs, decreases with increasing λ: `fomMaxArr[1,] - fomMaxArr[2,]` $= 0.024, 0.018, 0.007, 0.001$, where the successive values correspond to $\lambda = 1, 5, 10, 15$.

The λ' Poisson parameter controls the average number of perceived NLs per case. For example, for $\mu = 2$ and $\lambda = 1$, the average number is $\lambda' = \lambda/\mu = 0.5$, i.e., an average of one perceived NL every two non-diseased case. With increasing numbers of NLs per case it is necessary to increase the reporting threshold and LLF consequently decreases. Also, overall CAD performance, regardless of how it is measured (i.e., wAFROC-AUC or ROC-AUC), decreases.

Similar trends are observed for the Youden-index based optimizations (last four rows of table). However, Youden-index based optimizations compared as a group to wAFROC-AUC based optimizations show that Youden yields higher reporting thresholds, lower wAFROC-AUC, lower ROC-AUC and lower LLF values.

Table 14.1: Summary of optimization results for $\mu = 2$, $\nu = 1$ and different values of λ. The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may the Youden-index, as in the last four rows.

FOM	lambda	zeta1	wAFROC	ROC	OptOpPt
wAFROC	1	-0.235	0.880	0.937	(0.296, 0.854)
	5	0.810	0.768	0.875	(0.522, 0.763)
	10	1.373	0.699	0.825	(0.424, 0.635)
	15	1.697	0.660	0.788	(0.336, 0.535)
Youden	1	0.802	0.856	0.915	(0.106, 0.765)
	5	1.438	0.750	0.842	(0.188, 0.616)
	10	1.690	0.693	0.801	(0.227, 0.538)
	15	1.832	0.658	0.776	(0.251, 0.490)

One could display 8 FROC plots, each corresponding to a row of the preceding table, but there is a more efficient method. The FROC curve is defined in terms of the RSM parameters as follows:

$$\left.\begin{aligned} NLF\left(\zeta,\lambda'\right) &= \lambda'\Phi\left(-\zeta\right) \\ LLF\left(\zeta,\mu,\nu',\overrightarrow{f_L}\right) &= \nu'\Phi\left(\mu-\zeta\right) \end{aligned}\right\} \tag{14.1}$$

Here $\overrightarrow{f_L}$ is the lesion-distribution vector, `c(0.5, 0.5)` in the current example.

The *end-point* of the FROC defined by (λ', ν') is not to be confused with the *optimal* value of ζ_1; the former corresponds to $\zeta_1 = -\infty$ while the latter is a finite value of ζ_1 as found by the optimization procedure.

Since the Φ function ranges from one to unity, the *four FROC curves for different values of λ are scaled versions of a single curve whose x-axis ranges from 0 to 1.* The single curve corresponds to $\lambda' = 1$ and the true curves are obtained by scaling this curve along the x-axis by the appropriate λ' factor. With this understanding one can display 4 FROC curves with a single FROC curve where the x-axis is $NLF\left(\zeta,\lambda'=1\right)$. The true FROC curve is defined by:

$$\left.\begin{aligned} NLF\left(\zeta,\lambda'\right) &= \lambda' NLF\left(\zeta,\lambda'=1\right) \\ LLF\left(\zeta,\mu,\nu',\overrightarrow{f_L}\right) &= \nu'\Phi\left(\mu-\zeta\right) \end{aligned}\right\} \tag{14.2}$$

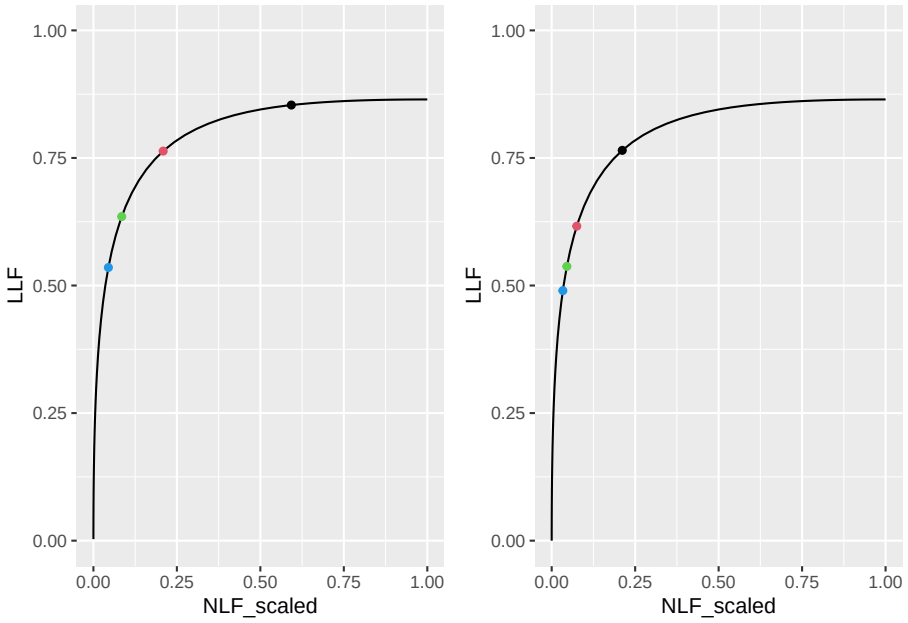


Figure 14.1: Left panel: maximized wAFROC AUC was used to find optimal ζ_1. Right panel: maximized Youden-index was used to find optimal ζ_1. Dot colors: black means $\lambda = 1$, red means $\lambda = 5$, green means $\lambda = 10$ and blue means $\lambda = 15$.

The left panel in 14.1 shows the optimal operating points when wAFROC-AUC is maximized. The 4 operating points are color coded as follows:

- The black dot corresponds to $\lambda = 1$, i.e., $\lambda' = 1/2 = 0.5$. In other words, the true FROC is obtained by *shrinking* the plot shown, including the superposed black dot, along the x-axis by a factor of 2.

- The red dot corresponds to $\lambda' = 2.5$. In other words, the true FROC is obtained by *magnifying* that shown, including the red dot, along the x-axis by a factor of 2.5.

- The green dot corresponds to $\lambda' = 5$.
- The blue dot corresponds to $\lambda' = 7.5$.

These plots illustrate the previous comments, namely, as λ increases, *the optimal operating point moves down the scaled curve*.

The right panel shows the optimal operating point when the Youden-index is maximized. It shows the same general features as the previous example but the group of four operating points in the right panel are below-left those in the left panel, representing higher values of optimal ζ_1, i.e., a more stringent criteria. As seen in the preceding table the overly strict criteria, using Youden-index based optimization, leads to lower true performance: i.e., lower wAFROC-AUC and lower ROC-AUC, and lower LLF.

The FROC curve does not represent true performance. To visualize true performance one compares wAFROC curves.

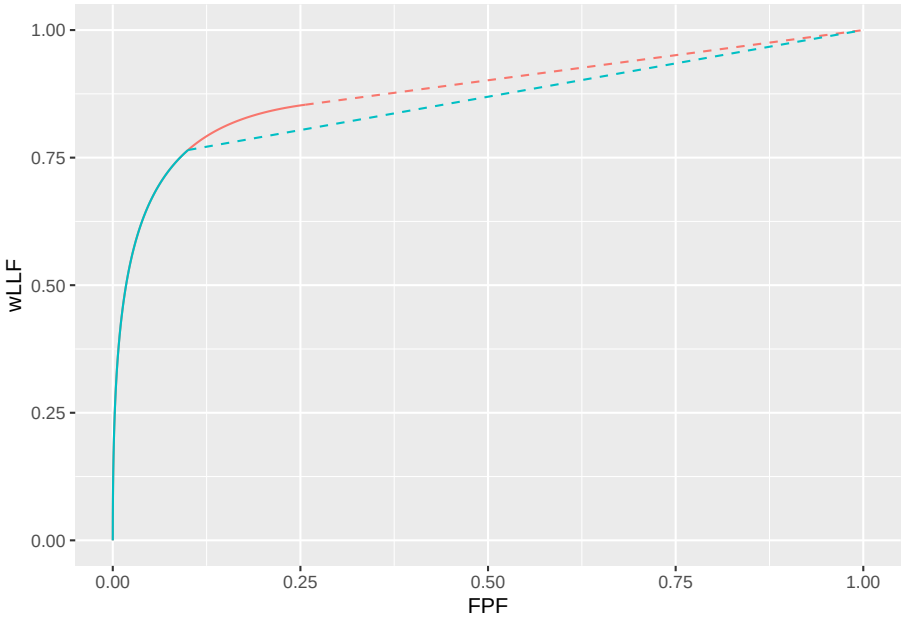


Figure 14.2: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to $\mu = 2$, $\nu = 1$ and $\lambda = 1$. The optimal reporting theshold ζ_1 is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.880 and 0.856, respectively.

Each curve ends at the optimal threshold listed in Table 14.1, namely ζ_1 = -0.235 for the red curve and $\zeta_1 = 0.802$ for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to

the adoption of an overly strict threshold.

14.3.3 Vary nu

For $\mu = 2$ and $\lambda = 5$, wAFROC-AUC and Youden-index based optimizations were performed for $\nu = 0.1, 0.5, 1, 2$. Table 14.2 summarizes the results.

Table 14.2: Summary of optimization results for $\mu = 2$, $\lambda = 5$ and different values of ν. The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may the Youden-index, as in the last four rows.

FOM	nu	zeta1	wAFROC	ROC	OptOpPt
wAFROC	0.1	2.275	0.522	0.551	(0.029, 0.071)
	0.5	1.376	0.660	0.771	(0.211, 0.464)
	1	0.810	0.768	0.875	(0.522, 0.763)
	2	-0.311	0.841	0.915	(1.555, 0.971)
Youden	0.1	1.336	0.473	0.588	(0.227, 0.135)
	0.5	1.398	0.660	0.770	(0.203, 0.459)
	1	1.438	0.750	0.842	(0.188, 0.616)
	2	1.461	0.793	0.874	(0.180, 0.692)

Focusing on the wAFROC-AUC based optimizations (first four rows of table), as ν increases:

- The optimal threshold ζ_1 decreases, resulting in more marks being reported; wAFROC-AUC increases; ROC-AUC increases and the optimal operating point on the FROC moves to higher LLF values, i.e., higher values of lesion-level "sensitivity".

All of these are opposite to the effect of increasing λ. The ν' binomial success probability parameter is the probability of a perceived LL event. For example, for $\mu = 2$ and $\nu = 0.1$, $\nu' = 1 - \exp(-\mu\nu) = 0.1812692$, i.e., an average of 18 percent of lesions present are found by the algorithm at the *initial detection* stage, using terminology in (Edwards et al., 2002).

With one exception similar trends are observed for the Youden-index based optimizations (last four rows of table). As a group Youden-index based optimizations (last four rows of table) compared to wAFROC-AUC based optimizations

show that the former yields higher reporting thresholds, lower wAFROC-AUC, lower ROC-AUC and lower LLF values.

The exception is that as ν increases the optimal threshold increases, but more slowly. The increasing separation of the two underlying probability density functions that generate the ROC causes the optimal threshold to increase (similar to the explanation in Section 14.3.4).

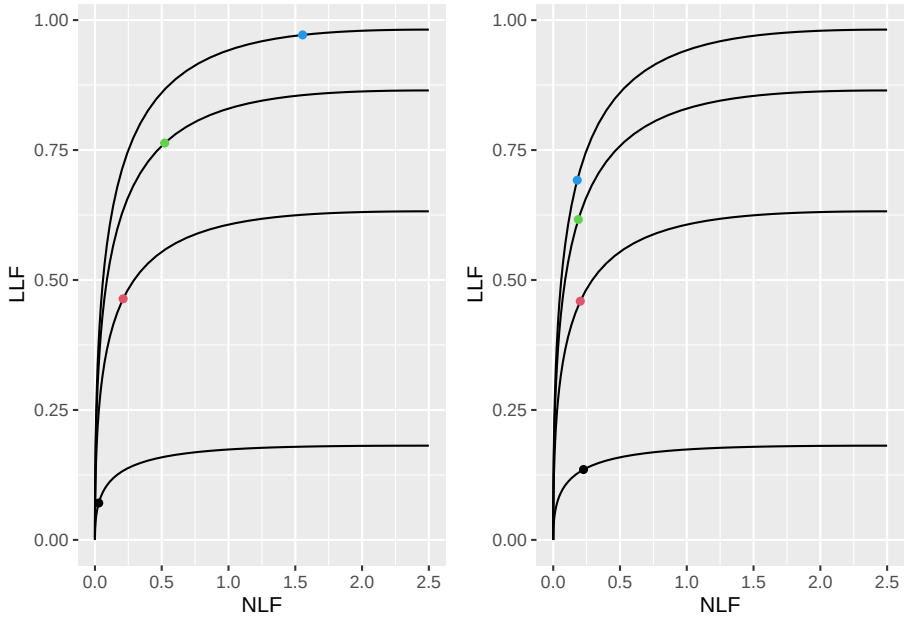


Figure 14.3: Left panel: maximized wAFROC-AUC was used to find optimal ζ_1. Right panel: maximized Youden-index was used to find optimal ζ_1. Dot colors: black means $\nu = 0.1$, red means $\nu = 0.5$, green means $\nu = 1$ and blue means $\nu = 2$.

Fig. 14.3 shows the FROC curves with optimal operating points superimposed. The left panel corresponds to wAFROC-AUC based optimizations while the right panel corresponds to Youden-index based optimizations. These illustrate the previous comments, namely, as ν increases, *the optimal operating point moves up the FROC curve*.

To visualize true performance one compares wAFROC curves.

Each curve ends at the optimal threshold listed in Table 14.2, namely ζ_1 = -0.311 for the red curve and $\zeta_1 = 1.461$ for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to the adoption of an overly strict threshold.

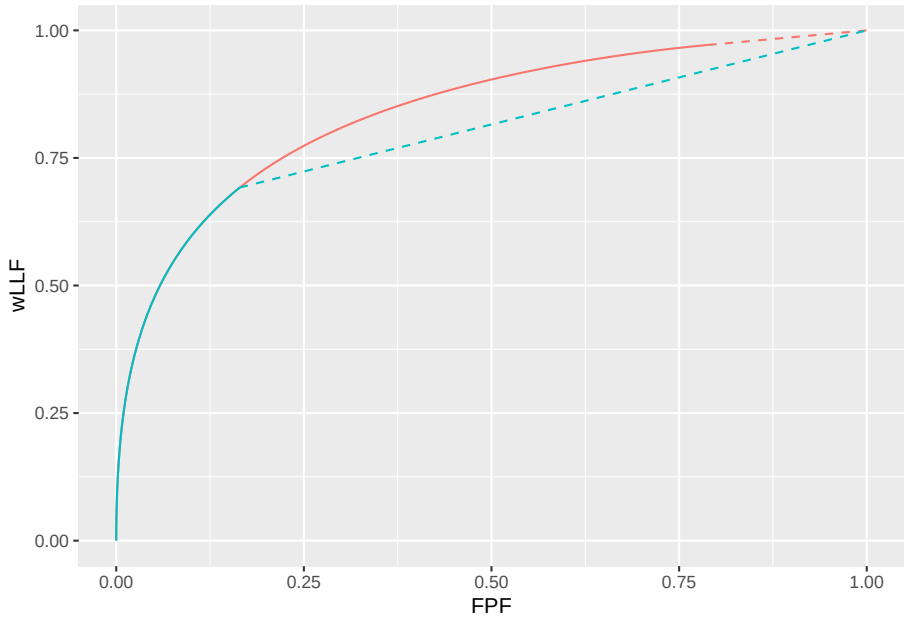


Figure 14.4: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to $\mu = 2$, $\lambda = 5$ and $\nu = 2$. The optimal reporting theshold ζ_1 is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.841 and 0.793, respectively.

14.3.4 Vary mu

For $\nu = 1$ and $\lambda = 1$ wAFROC-AUC and Youden-index based optimizations were performed for 4 values of $\mu = 0.75, 1, 1.25, 1.5$. Table 14.2 summarizes the results.

Table 14.3: Summary of optimization results for $\nu = 1$, $\lambda = 1$ and different values of μ. The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may the Youden-index, as in the last four rows.

FOM	mu	zeta1	wAFROC	ROC	OptOpPt
wAFROC	0.75	1.422	0.518	0.587	(0.103, 0.132)
	1	0.310	0.603	0.745	(0.378, 0.477)
	1.25	-0.132	0.699	0.823	(0.442, 0.654)
	1.5	-0.268	0.777	0.875	(0.404, 0.747)
Youden	0.75	0.367	0.493	0.668	(0.476, 0.343)
	1	0.386	0.603	0.741	(0.350, 0.462)
	1.25	0.461	0.691	0.802	(0.258, 0.560)
	1.5	0.563	0.760	0.850	(0.191, 0.641)

Increasing μ, while holding λ and ν constant, *simultaneously decreases* λ' and increases μ'. As the latter two parameters work in opposite directions (increasing one has a similar effect as decreasing the other) the simultaneous changes result in an amplified effect. The values in the table can be understood from this.

For the wAFROC-AUC based optimizations (first four rows of table), as μ increases the reporting threshold ζ_1 decreases, both wAFROC-AUC and ROC-AUC increase, and the optimal operating point moves to higher LLF values.

For the Youden-index based optimizations (last four rows of table), as μ increases the reporting threshold ζ_1 increases (but the magnitude of the change is smaller than for the first four rows), both wAFROC-AUC and ROC-AUC increase, and the optimal operating point moves to higher LLF values.

The effect of increasing μ can be understood as resulting from the competing effects of *greater search performance*, greater numbers of LLs and fewer NLs, both allowing the threshold to be moved down, and *greater classification performance*, allowing the threshold to be moved up (as the separation of two unit

normal distribution increases, the optimal threshold for discriminating between them increases).

Fig. 14.5 shows FROC curves with superimposed optimal operating points.

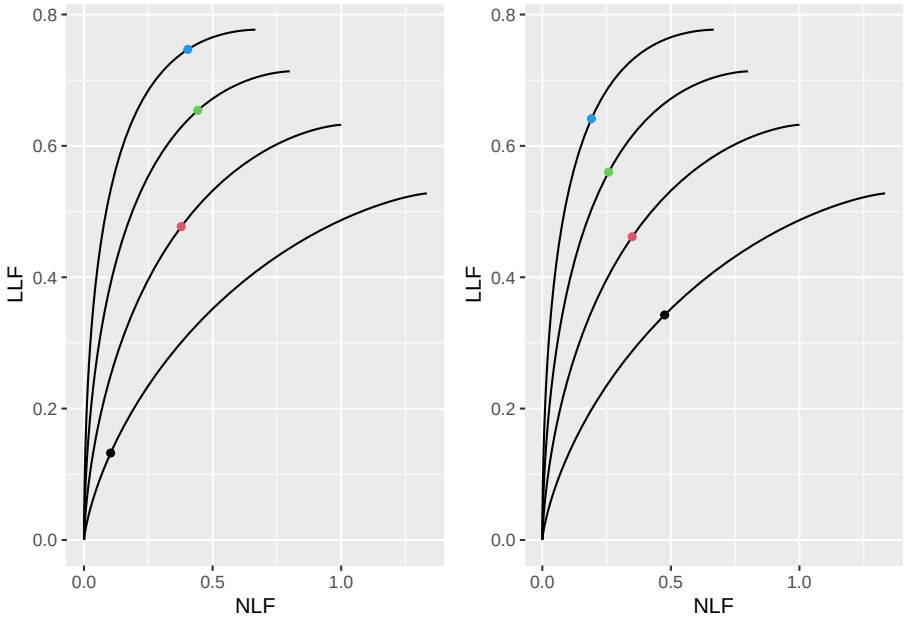


Figure 14.5: Left panel: maximized wAFROC-AUC was used to find optimal ζ_1. Right panel: maximized Youden-index was used to find optimal ζ_1. Dot colors: black means $\mu = 0.75$, red means $\mu = 1$, green means $\lambda = 1.25$ and blue means $\mu = 1.5$.

For each of the four values of μ the left panel in Fig. 14.5 shows the optimal operating point when wAFROC-AUC is maximized. It shows the FROC curves with optimal operating points superimposed. These illustrate the previous comments, namely, as μ increases, *the optimal operating point moves up the FROC curve*.

The right panel in Fig. 14.5 shows the optimal operating point when the Youden-index is maximized.

To visualize true performance one compares wAFROC curves.

Each curve ends at the optimal threshold listed in Table 14.3, namely $\zeta_1 = $ -0.268 for the red curve, and $\zeta_1 = 0.563$ for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to the adoption of an overly strict threshold.

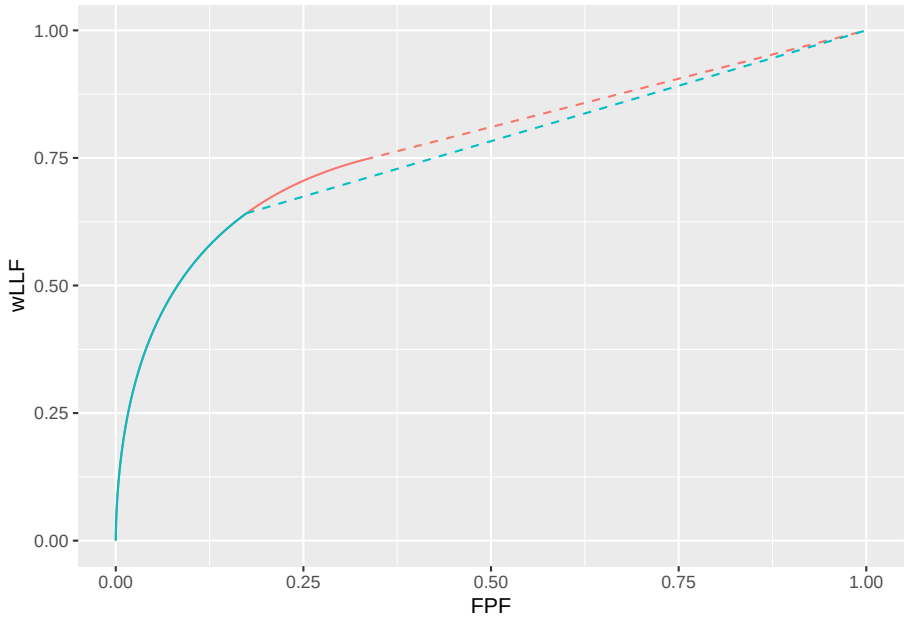


Figure 14.6: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to $\lambda = 1$, $\nu = 1$ and $\mu = 1.5$. The optimal reporting theshold ζ_1 is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.777 and 0.760, respectively.

14.4 Using the method

Assume that one has designed an algorithmic observer that has been optimized with respect to all other parameters except the reporting threshold. At this point the algorithm reports every suspicious region, no matter how low the malignancy index. The mark-rating pairs are entered into a `RJafroc` format Excel input file. The next step is to read the data file – `DfReadDataFile()` – convert it to an ROC dataset – `DfFroc2Roc()` – and then perform a radiological search model (RSM) fit to the dataset using function `FitRsmRoc()`. This yields the necessary λ, μ, ν parameters. These values are used to perform the computations described in the embedded code in this chapter, see for example Section 14.3.2. This determines the optimal reporting threshold. The RSM parameter values and the reporting threshold determine the optimal reporting point on the FROC curve. The designer sets the algorithm to only report marks with confidence levels exceeding this threshold.

14.5 An application

The standalone CAD LROC dataset described in (Hupse et al., 2013) was used to create the quasi-FROC ROC-AUC equivalent dataset embedded in `RJafroc` as object `datasetCadSimuFroc`. In the following code the first reader for this dataset, corresponding to CAD, is extracted using `DfExtractDataset` (the other readers, corresponding to radiologists who interpreted the same cases, are not used here). The function `DfFroc2Roc` converts this to an ROC dataset. The function `DfBinDataset` bins the data to about 7 bins. One lesion per abnormal case is assumed: `lesDistr = c(1)`. `FitRsmRoc` fits the binned ROC dataset to the radiological search model RSM. Object `fit` contains all necessary parameters required to perform the optimizations described in previous sections.

```
ds <- datasetCadSimuFroc
dsCad <- DfExtractDataset(ds, rdrs = 1)
dsCadRoc <- DfFroc2Roc(dsCad)
dsCadRocBinned <- DfBinDataset(dsCadRoc, opChType = "ROC")
lesDistr <- c(1)
fit <- FitRsmRoc(dsCadRocBinned, lesDistr)
```

Table 14.4 summarizes the results.

Table 14.4: Summary of optimization results for example FROC dataset. The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may the Youden-index, as in the last four rows.

FOM	lambda	zeta1	wAFROC	ROC	OptOpPt
wAFROC	18.680	1.739	0.774	0.815	(0.278, 0.679)
Youden		1.982	0.770	0.798	(0.161, 0.627)

The dataset is characterized by a large λ parameter and, consistent with the finding in 14.3.2, the advantage of wAFROC-AUC over Youden-index based optimization, as measured by the difference in corresponding wAFROC-AUCs, is small.

Fig. 14.7 shows FROC curves with superimposed optimal operating points.

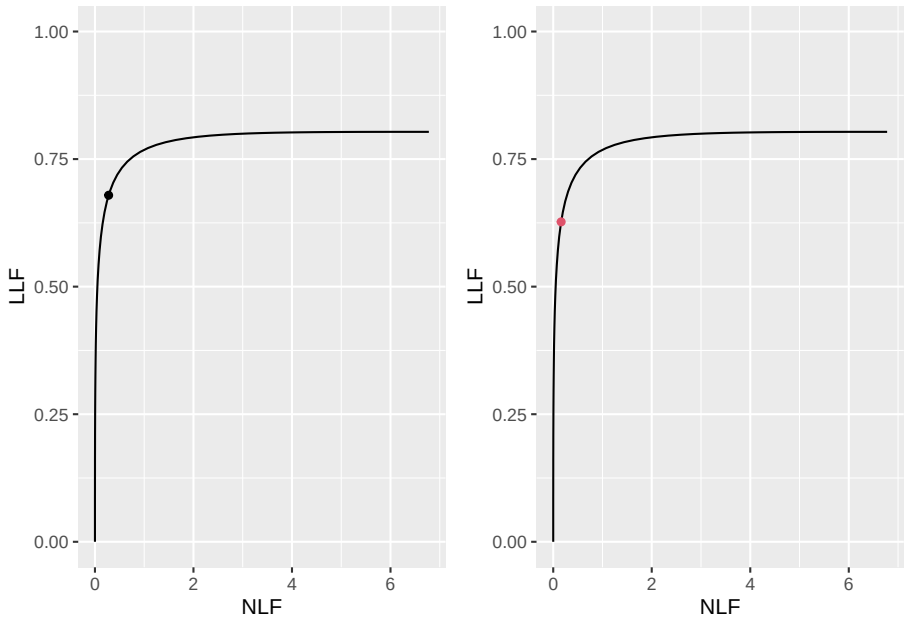


Figure 14.7: Left panel: maximized wAFROC-AUC was used to find optimal ζ_1. Right panel: maximized Youden-index was used.

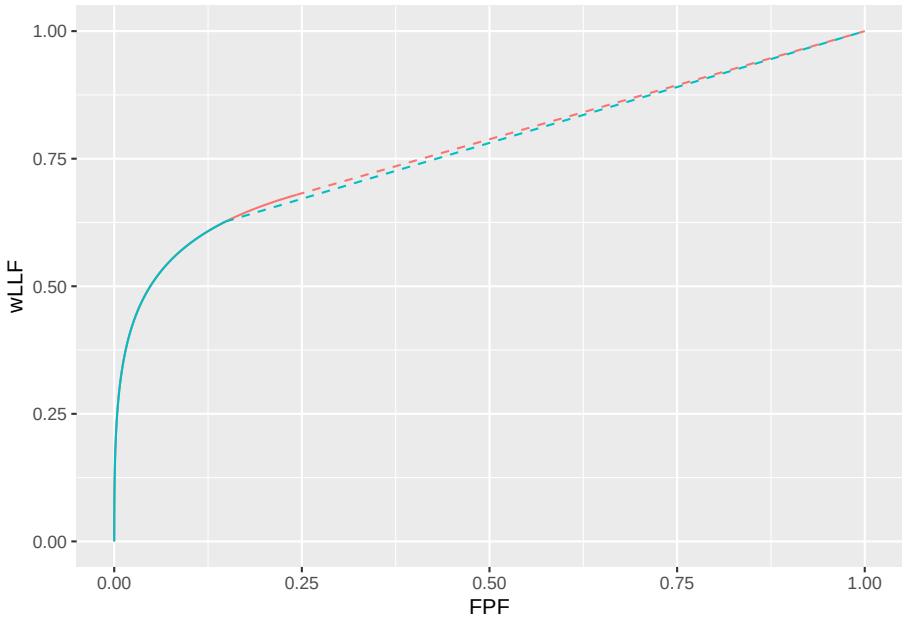


Figure 14.8: Red line and dots: wAFROC-AUC based optimization; blue line and dots: Youden-index based optimization. The two wAFROC-AUCs are 0.774 and 0.770, respectively.

14.6 Discussion

Described is a method for finding the optimal operating point on an FROC curve. The method consists of varying the reporting threshold to maximize the area under the wAFROC. An alternate method, based on maximization of the Youden-index, was also tested. Both methods are illustrated using the radiological search model to parameterize the FROC data. In all cases studied the Youden-index based method selected a stricter reporting threshold than optimal, resulting in lower wAFROC-AUC and ROC-AUC as compared to wAFROC-AUC based optimization. The results are illustrated using FROC curves, which are more familiar to CAD designers.

The method was applied to a quasi-FROC dataset created from an originally LROC dataset. For this dataset the optimized wAFROC-AUC was marginally superior to that using the Youden-index.

With increasing λ every case is guaranteed at least one z-sample, and the model becomes more ROC-like.

14.7 References

Chapter 15

Localization - classification tasks

15.1 TBA How much finished

10%

15.2 Introduction

TBA: This project is a works-in-progress.

15.3 Abbreviations

- Correct-localization correct-classification = **CL-CC**
- Correct-localization incorrect-classification = **CL-IC**
- Incorrect-localization classification not applicable = **IL-NA**

15.4 History and basic idea

This project started with a request to extend localization analysis software `RJafroc` to localization-classification tasks. Since this is new research the required data format is not in the `RJafroc` documentation. Some familiarity with basic localization task analysis is assumed.

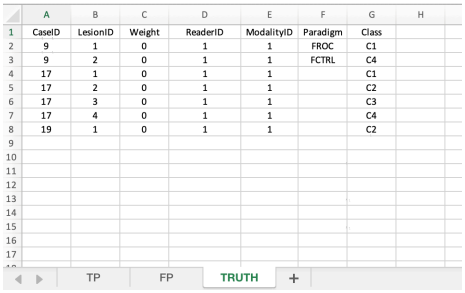
The basic idea is that spatial localization is a special case of localization-with-classification. **CL-CC** marks are put in TP sheet and other are put in FP sheet.

15.5 First example, File1.xlsx

- This example is implemented in file `File1.xlsx`.
- There are four classes of lesions: `C1`, `C2`, `C3`and `C4`.
- The rating scale is 1 - 10 and positive-directed.
- The dataset has 3 cases: 9, 17 and 19.

15.5.1 `Truth` sheet

This has the ground truth of for cases and lesions, and specifies their class types.



	A	B	C	D	E	F	G	H
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm	Class	
2	9	1	0	1	1	FROC	C1	
3	9	2	0	1	1	FCTRL	C4	
4	17	1	0	1	1		C1	
5	17	2	0	1	1		C2	
6	17	3	0	1	1		C3	
7	17	4	0	1	1		C4	
8	19	1	0	1	1		C2	
9								
10								
11								
12								
13								
14								
15								
16								
17								

TP FP TRUTH +

Figure 15.1: Truth worksheet for File1.xlsx

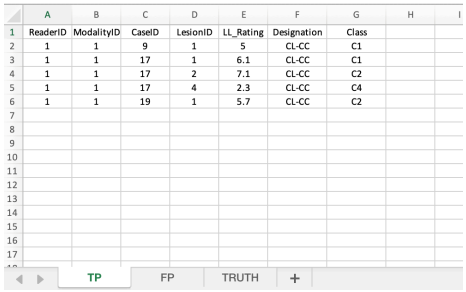
- Case 9 has two lesions, with classes `C1` and `C4`.
- Case 17 has four lesions, with classes `C1`, `C2`, `C3`and `C4`.
- Case 19 has one lesion, with class `C2`.

15.5.2 `TP` sheet

This holds CL-CC marks.

15.5.2.1 Case 9

- Lesion `C1`, `lesionID` = 1, **CL-CC** mark rated 5.



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class		
2	1	1	9	1	5	CL-CC	C1		
3	1	1	17	1	6.1	CL-CC	C1		
4	1	1	17	2	7.1	CL-CC	C2		
5	1	1	17	4	2.3	CL-CC	C4		
6	1	1	19	1	5.7	CL-CC	C2		

Figure 15.2: TP worksheet for File1.xlsx

15.5.2.2 Case 17

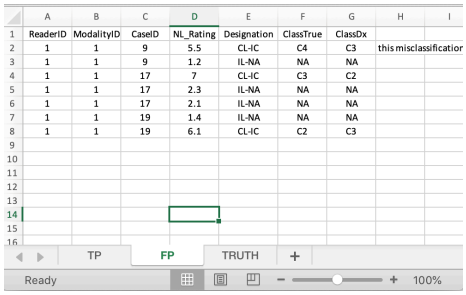
- Lesion `C1`, `lesionID` = 1, **CL-CC** mark rated 6.1.
- Lesion `C2`, `lesionID` = 2, **CL-CC** mark rated 7.1.
- Lesion `C4`, `lesionID` = 4, **CL-CC** mark rated 2.3.

15.5.2.3 Case 19

- Lesion `C2`, `lesionID` = 1, **CL-CC** mark rated 5.7.

15.5.3 `FP` sheet

- This holds **IL-NA** and **CL-IC** marks.
- `ClassTrue` is the true class of the lesion.
- `ClassDx` is the indicated or diagnosed class of the lesion.



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	NL_Rating	Designation	ClassTrue	ClassDx		
2	1	1	9	5.5	CL-IC	C4	C3	this misclassificatior	
3	1	1	9	1.2	IL-NA	NA	NA		
4	1	1	17	7	CL-IC	C3	C2		
5	1	1	17	2.3	IL-NA	NA	NA		
6	1	1	17	2.1	IL-NA	NA	NA		
7	1	1	19	1.4	IL-NA	NA	NA		
8	1	1	19	6.1	CL-IC	C2	C3		

Figure 15.3: FP worksheet for File1.xlsx

15.5.3.1 Case 9

- **CL-IC** mark rated 5.5, `C2` classified as `C3`.

- **IL-NA** mark rated 1.2.

15.5.3.2 Case 17

- **CL-IC** mark rated 7, `C3` classified as `C2`.
- **IL-NA** mark rated 2.3.
- **IL-NA** mark rated 2.1.

15.5.3.3 Case 19

- **IL-NA** mark rated 1.4.
- **CL-IC** mark rated 6.1, `C2` classified as `C3`.

15.5.4 The two ratings arrays

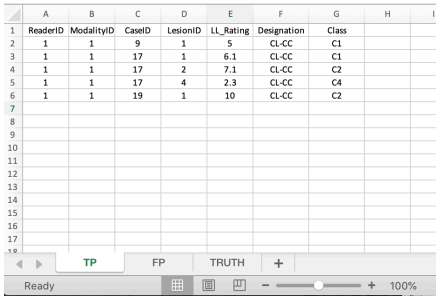
```
fileName <- "R/CH83-ClassificationTask/File1.xlsx"
x <- DfReadDataFile(fileName = fileName)
x$ratings$NL[1,1,,]
#>      [,1] [,2] [,3]
#> [1,]  5.5  1.2 -Inf
#> [2,]  7.0  2.3  2.1
#> [3,]  1.4  6.1 -Inf
x$ratings$LL[1,1,,]
#>      [,1] [,2] [,3] [,4]
#> [1,]  5.0 -Inf -Inf -Inf
#> [2,]  6.1  7.1 -Inf  2.3
#> [3,]  5.7 -Inf -Inf -Inf
```

The FOM is shown next:

```
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>           rdr1
#> trt1 0.2361111
```

15.6 Second example, File2.xlsx

I increased the LL rating for case 19 to 10; this should increase the FOM. This example is implemented in file `File2.xlsx`.



	A	B	C	D	E	F	G
1	ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class
2	1	1	9	1	5	CL-CC	C1
3	1	1	17	1	6.1	CL-CC	C1
4	1	1	17	2	7.1	CL-CC	C2
5	1	1	17	4	2.3	CL-CC	C4
6	1	1	19	1	10	CL-CC	C2

Figure 15.4: TP worksheet for File2.xlsx

```
fileName <- "R/CH83-ClassificationTask/File2.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>           rdr1
#> trt1 0.4583333
```

15.7 Third example, File3.xlsx

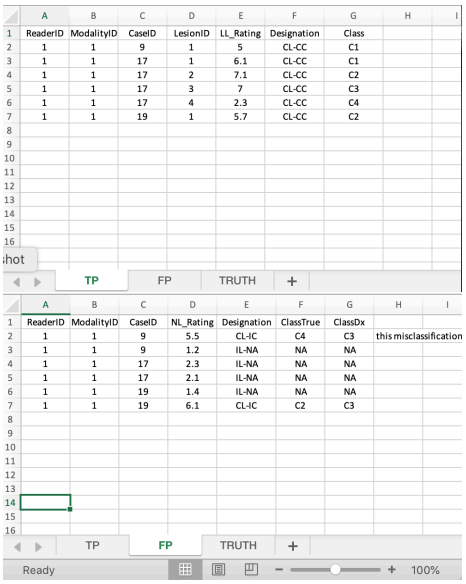
Starting with original file, I transferred a **CL-IC** for case 17 to the TP sheet, where it is a **CL_CC** mark. This should increase the FOM as credit is given for **CL-CC.** This example is implemented in file `File3.xlsx`.

```
fileName <- "R/CH83-ClassificationTask/File3.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>           rdr1
#> trt1 0.5277778
```

15.8 Fourth example, File4.xlsx

So far we have dealt with one modality and one reader.

- Additional algorithmic readers can be added under `readerID`.
- They should not be added as additional treatments (has to do with treatment being regarded as a fixed factor and reader as as random factor in the analysis).
- The starting point is `File3.xlsx`. I duplicated the data from this for two additional readers to create a single-modality three-reader dataset `File4.xlsx`.



ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class
1	1	9	1	5	CL-CC	C1
1	1	17	1	6.1	CL-CC	C1
1	1	17	2	7.1	CL-CC	C2
1	1	17	3	7	CL-CC	C3
1	1	17	4	2.3	CL-CC	C4
1	1	19	1	5.7	CL-CC	C2

TP FP TRUTH

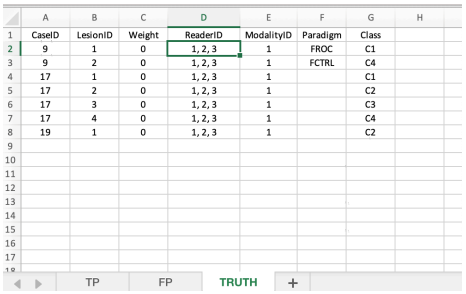
ReaderID	ModalityID	CaseID	NL_Rating	Designation	ClassTrue	ClassDx	
1	1	9	5.5	CL-IC	C4	C3	this misclassification
1	1	9	1.2	IL-NA	NA	NA	
1	1	17	2.3	IL-NA	NA	NA	
1	1	17	2.1	IL-NA	NA	NA	
1	1	19	1.4	IL-NA	NA	NA	
1	1	19	6.1	CL-IC	C2	C3	

TP FP TRUTH

Ready 100%

Figure 15.5: TP and FP worksheets for File3.xlsx

- Shown next are the three worksheets.



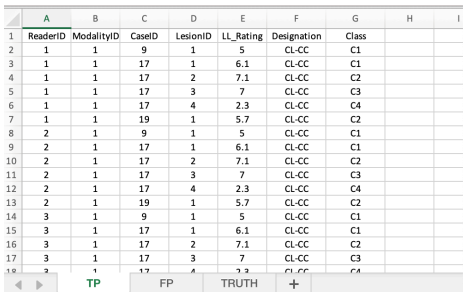
CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm	Class
9	1	0	1, 2, 3	1	FROC	C1
9	2	0	1, 2, 3	1	FCTRL	C4
17	1	0	1, 2, 3	1		C1
17	2	0	1, 2, 3	1		C2
17	3	0	1, 2, 3	1		C3
17	4	0	1, 2, 3	1		C4
19	1	0	1, 2, 3	1		C2

TP FP TRUTH

Figure 15.6: Truth worksheet for File4.xlsx

- Shown next are the three FOMs. Note that they are identical.

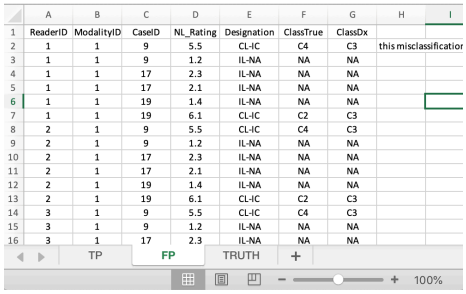
```
fileName <- "R/CH83-ClassificationTask/File4.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>           rdr1      rdr2      rdr3
#> trt1 0.5277778 0.5277778 0.5277778
```



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class		
2	1	1	9	1	5	CL-CC	C1		
3	1	1	17	1	6.1	CL-CC	C1		
4	1	1	17	2	7.1	CL-CC	C2		
5	1	1	17	3	7	CL-CC	C3		
6	1	1	17	4	2.3	CL-CC	C4		
7	1	1	19	1	5.7	CL-CC	C2		
8	2	1	9	1	5	CL-CC	C1		
9	2	1	17	1	6.1	CL-CC	C1		
10	2	1	17	2	7.1	CL-CC	C2		
11	2	1	17	3	7	CL-CC	C3		
12	2	1	17	4	2.3	CL-CC	C4		
13	2	1	19	1	5.7	CL-CC	C2		
14	3	1	9	1	5	CL-CC	C1		
15	3	1	17	1	6.1	CL-CC	C1		
16	3	1	17	2	7.1	CL-CC	C2		
17	3	1	17	3	7	CL-CC	C3		

TP FP TRUTH +

Figure 15.7: TP worksheet for File4.xlsx



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	NL_Rating	Designation	ClassTrue	ClassDx		
2	1	1	9	5.5	CL-IC	C4	C3	this misclassificatior	
3	1	1	9	1.2	IL-NA	NA	NA		
4	1	1	17	2.3	IL-NA	NA	NA		
5	1	1	17	2.1	IL-NA	NA	NA		
6	1	1	19	1.4	IL-NA	NA	NA		
7	1	1	19	6.1	CL-IC	C2	C3		
8	2	1	9	5.5	CL-IC	C4	C3		
9	2	1	9	1.2	IL-NA	NA	NA		
10	2	1	17	2.3	IL-NA	NA	NA		
11	2	1	17	2.1	IL-NA	NA	NA		
12	2	1	19	1.4	IL-NA	NA	NA		
13	2	1	19	6.1	CL-IC	C2	C3		
14	3	1	9	5.5	CL-IC	C4	C3		
15	3	1	9	1.2	IL-NA	NA	NA		
16	3	1	17	2.3	IL-NA	NA	NA		

TP FP TRUTH + 100%

Figure 15.8: FP worksheet for File4.xlsx

15.9 Fifth example, File5.xlsx

- Need to add some randomness to the ratings.
- I randomly added to the ratings from a uniform distribution in the range -0.5 to +0.5.
- This is very crude, as in practice the the number of marks will also vary from reader to reader.
- But file `File5.xlsx` should give one the general idea of how to extend to several algorithmic readers.
- Note that now the FOMs are not identical.

```
fileName <- "R/CH83-ClassificationTask/File5.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>            rdr1      rdr2      rdr3
#> trt1 0.3611111 0.5555556 0.4444444
```

15.10 Precautions

- Unlike regular `RJafroc` analysis, there is no error checking of the classification codes `C1`, etc. For example, if a lesion with class `C1` is recorded in the TP sheet as a **CL-CC** and it is also mistakenly recorded in the FP sheet as a **CL-IC**, the program does not know about the mistake. Multiple FP on the same case are allowed in FROC analysis.
- I suggest that the extra columns in the sample files be recorded for your dataset. This will enable me to subsequently include error-checking code for data entry mistakes.
- For example, the columns `Designation`, `ClassTrue` and `ClassRx` in the `FP` sheet are currently not read by the software.
- To make further progress you need to drastically reduce the file size (once the new method is fully developed you can always add the remaining cases and readers). The current file size makes it impossible to fully develop the system. Most studies in this field are done with 2-3 modalities and about 100-200 cases.

15.11 Discussion

15.12 References

Chapter 16

Split Plot Study Design

16.1 TBA How much finished

10%

16.2 Mean Square R(T)

R(T) is read as "reader nested within treatment" (Hillis, 2014).

$$\text{MS[R(T)]} = \frac{1}{I(J-1)} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\theta_{ij} - \theta_{i\bullet}\right)^2 \tag{16.1}$$

$$\text{MS[R(T)]} = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{J_i - 1} \sum_{j=1}^{J} \left(\theta_{ij} - \theta_{i\bullet}\right)^2 \tag{16.2}$$

16.3 References

Bibliography

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Black, W. C. (2000). Anatomic extent of disease: A critical variable in reports of diagnostic accuracy. *Radiology*, 217(2):319–320.

Black, W. C. and Dwyer, A. J. (1990). Local versus global measures of accuracy: An important distinction for diagnostic imaging. *Med Decis Making*, 10(4):266–273.

Bolker, B. and R Development Core Team (2020). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.23.1.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, volume 127, pages 124–135. International Society for Optics and Photonics.

Chakraborty, D. (2006a). ROC curves predicted by a model of visual search. *Physics in Medicine & Biology*, 51(14):3463.

Chakraborty, D., Breatnach, E., Yester, M., Soto, B., Barnes, G., and Fraser, R. (1986). Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology*, 158:35–39.

Chakraborty, D., Philips, P., and Zhai, X. (2020). *RJafroc: Analyzing Diagnostic Observer Performance Studies*. R package version 1.3.2.9000.

Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response receiver operating characteristic (froc) data. *Medical physics*, 16(4):561–568.

Chakraborty, D. P. (2006b). A search model and figure of merit for observer data acquired according to the free-response paradigm. *Physics in Medicine & Biology*, 51(14):3449.

Chakraborty, D. P. (2008). Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Acad Radiol*, 15(12):1554–1566.

Chakraborty, D. P. (2017). *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*. CRC Press, Boca Raton, FL.

Chakraborty, D. P. and Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis and validation. *Med Phys*, 31(8):2313–2330.

Chakraborty, D. P. and Svahn, T. (2011). Estimating the parameters of a model of visual search from ROC data: an alternate method for fitting proper ROC curves. *Proc. SPIE 7966*, 7966.

Chakraborty, D. P. and Yoon, H. J. (2008). Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Medical Physics*, 35(2):435–445.

Chakraborty, D. P. and Yoon, H. J. (2009). JAFROC analysis revisited: figure-of-merit considerations for human observer studies. *Proc. SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, 7263:72630T.

Chakraborty, D. P. and Zhai, X. (2016). On the meaning of the weighted alternative free-response operating characteristic figure of merit. *Medical physics*, 43(5):2548–2557.

De Boo, D. W., Uffmann, M., Weber, M., Bipat, S., Boorsma, E. F., Scheerder, M. J., Freling, N. J., and Schaefer-Prokop, C. M. (2011). Computer-aided detection of small pulmonary nodules in chest radiographs: an observer study. *Academic radiology*, 18(12):1507–1514.

DeSantis, C., Siegel, R., Bandi, P., and Jemal, A. (2011). Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6):408–418.

Dobbins III, J. T., McAdams, H. P., Sabol, J. M., Chakraborty, D. P., Kazerooni, E. A., Reddy, G. P., Vikgren, J., and Båth, M. (2016). Multi-institutional evaluation of digital tomosynthesis, dual-energy radiography, and conventional chest radiography for the detection and management of pulmonary nodules. *Radiology*, 282(1):236–250.

Dorfman, D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals - rating-method data. *Journal of Mathematical Psychology*, 6:487–496.

Dorfman, D. and Berbaum, K. (2000). A contaminated binormal model for ROC data: Part ii. a formal model. *Acad Radiol.*, 7(6):427–37.

Dorfman, D., Berbaum, K., Metz, C., Lenth, R., Hanley, J., and Abu Dagga, H. (1997). Proper receiving operating characteristic analysis: The bigamma model. *Acad. Radiol.*, 4(2):138–149.

Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Investigative radiology*, 27(9):723–731.

Duchowski, A. T. (2002). *Eye Tracking Methodology: Theory and Practice*. Clemson University, Clemson, SC.

Edwards, D. C., Kupinski, M. A., Metz, C. E., and Nishikawa, R. M. (2002). Maximum likelihood fitting of froc curves under an initial-detection-and-candidate-analysis model. *Medical physics*, 29(12):2861–2870.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Egan, J., Greenburg, G., and Schulman, A. (1961). Operating characteristics, signal detectability and the method of free response. *J Acoust Soc. Am.*, 33:993–1007.

Ernster, V. L. (1981). The epidemiology of benign breast disease. *Epidemiologic reviews*, 3(1):184–202.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical proceedings of the Cambridge philosophical society*, volume 24, pages 180–190. Cambridge University Press.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley and Sons.

Franken, Edmund A., J., Berbaum, K. S., Marley, S. M., Smith, W. L., Sato, Y., Kao, S. C. S., and Milam, S. G. (1992). Evaluation of a digital workstation for interpreting neonatal examinations: A receiver operating characteristic study. *Investigative Radiology*, 27(9):732–737.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.

Halpern, S. D., Karlawish, J. H., and Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3):358–362.

Hein, P. A., Krug, L. D., Romano, V. C., Kandel, S., Hamm, B., and Rogalla, P. (2010). Computer-aided detection in computed tomography colonography with full fecal tagging: comparison of standalone performance of 3 automated polyp detection systems. *Canadian Association of Radiologists Journal*, 61(2):102–108.

Hillis, S. L. (2007). A comparison of denominator degrees of freedom methods for multiple observer (ROC) analysis. *Statistics in medicine*, 26(3):596–619.

Hillis, S. L. (2014). A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. *Statistics in Medicine*, 33(2):330–360.

Hillis, S. L., Berbaum, K. S., and Metz, C. E. (2008). Recent developments in the dorfman-berbaum-metz procedure for multireader roc study analysis. *Academic radiology*, 15(5):647–661.

Hillis, S. L., Obuchowski, N. A., Schartz, K. M., and Berbaum, K. S. (2005). A comparison of the dorfman–berbaum–metz and obuchowski–rockette methods for receiver operating characteristic (ROC) data. *Statistics in medicine*, 24(10):1579–1607.

Hupse, R., Samulski, M., Lobbes, M., Heeten, A., Imhof-Tas, M., Beijerinck, D., Pijnappel, R., Boetes, C., and Karssemeijer, N. (2013). Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *European Radiology*, 23(1):93–100.

Kooi, T., Gubern-Merida, A., Mordang, J.-J., Mann, R., Pijnappel, R., Schuur, K., den Heeten, A., and Karssemeijer, N. (2016). A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In *International Workshop on Breast Imaging*, pages 51–56. Springer.

Kundel, H. and Nodine, C. (1983). A visual concept shapes image perception. *Radiology*, 146(2):363–368.

Kundel, H. L. and Nodine, C. F. (2004). Modeling visual search during mammogram viewing. In *Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, volume 5372, pages 110–115. International Society for Optics and Photonics.

Kundel, H. L., Nodine, C. F., and Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative radiology*, 13(3):175–181.

Kundel, H. L., Nodine, C. F., Conant, E. F., and Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology*, 242(2):396–402.

Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733.

Metz, C. E. and Pan, X. (1999). "proper" binormal roc curves: theory and maximum-likelihood estimation. *Journal of mathematical psychology*, 43(1):1–33.

Metz, C. E., Starr, S. J., and Lusted, L. B. (1976). Observer performance in detecting multiple radiographic signals: prediction and analysis using a generalized roc approach. *Radiology*, 121(2):337–347.

Miller, H. (1969). The FROC curve: a representation of the observer's performance for the method of free response. *The Journal of the Acoustical Society of America*, 46(6(2)):1473–1476.

Niklason, L. T., Hickey, N. M., Chakraborty, D. P., Sabbagh, E. A., Yester, M. V., Fraser, R. G., and Barnes, G. T. (1986). Simulated pulmonary nodules: detection with dual-energy digital versus conventional radiography. *Radiology*, 160:589–593.

Nishikawa, R. M. and Pesce, L. L. (2011). Fundamental limitations in developing computer-aided detection for mammography. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 648:S251–S254.

Nodine, C. F. and Kundel, H. L. (1987). Using eye movements to study visual search and to improve tumor detection. *Radiographics*, 7(6):1241–1250.

Obuchowski, N. A., Lieber, M. L., and Powell, K. A. (2000). Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad. Radiol.*, 7(7):516–525.

Obuchowski, N. A. and Rockette, H. E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Communications in Statistics-simulation and Computation*, 24(2):285–308.

Pan, X. and Metz, C. E. (1997). The "proper" binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Academic radiology*, 4(5):380–389.

Penedo, M., Souto, M., Tahoces, P. G., Carreira, J. M., Villalon, J., Porto, G., Seoane, C., Vidal, J. J., Berbaum, K. S., Chakraborty, D. P., and Fajardo, L. L. (2005). Free-response receiver operating characteristic evaluation of lossy

jpeg2000 and object-based set partitioning in hierarchical trees compression of digitized mammograms. *Radiology*, 237(2):450–457.

Petrick, N. and Pastel, M. (2018). Guidance for industry and fda staff clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data—premarket approval (pma) and premarket notification [510 (k)] submission.

Popescu, L. M. (2011). Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve. *Medical physics*, 38(10):5690–5702.

Rao, V. M., Levin, D. C., Parker, L., Cavanaugh, B., Frangos, A. J., and Sunshine, J. H. (2010). How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802–805.

Ruschin, M., Timberg, P., Bath, M., Hemdal, B., Svahn, T., Saunders, R., Samei, E., Andersson, I., Mattsson, S., Chakraborty, D. P., and Tingberg, A. (2007). Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies. *Medical Physics*, 34:400 – 407.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Shanno, D. F. and Kettler, P. C. (1970). Optimal conditioning of quasi-newton methods. *Mathematics of Computation*, 24(111):657–664.

Starr, S., Metz, C., Lusted, L., Sharp, P., and Herath, K. (1977). Comments on the generalization of receiver operating characteristic analysis to detection and localization tasks. *Physics in Medicine & Biology*, 22(2):376.

Starr, S. J., Metz, C. E., Lusted, L. B., and Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology*, 116(3):533–538.

Summers, R. M., Handwerker, L. R., Pickhardt, P. J., Van Uitert, R. L., Deshpande, K. K., Yeshwant, S., Yao, J., and Franaszek, M. (2008). Performance of a previously validated ct colonography computer-aided detection system in a new patient population. *American Journal of Roentgenology*, 191(1):168–174.

Swensson, R. G. (1996). Unified measurement of observer performance in detecting and localizing target objects on images. *Medical physics*, 23(10):1709–1725.

Tan, T., Platel, B., Huisman, H., Sánchez, C., Mus, R., and Karssemeijer, N. (2012). Computer-aided lesion diagnosis in automated 3-d breast ultrasound using coronal spiculation. *Medical Imaging, IEEE Transactions on*, 31(5):1034–1042.

Taylor, S. A., Halligan, S., Burling, D., Roddie, M. E., Honeyfield, L., McQuillan, J., Amin, H., and Dehmeshki, J. (2006). Computer-assisted reader software versus expert reviewers for polyp detection on ct colonography. *American Journal of Roentgenology*, 186(3):696–702.

Thompson, J. D., Hogg, P., Manning, D. J., Szczepura, K., and Chakraborty, D. P. (2014). A free-response evaluation determining value in the computed tomography attenuation correction image for revealing pulmonary incidental findings: a phantom study. *Academic radiology*, 21(4):538–545.

Van Dyke, C., White, R., Obuchowski, N., Geisinger, M., Lorig, R., and Meziane, M. (1993). Cine MRI in the diagnosis of thoracic aortic dissection. *79th RSNA Meetings*.

Vikgren, J., Zachrisson, S., Svalkvist, A., Johnsson, A. A., Boijsen, M., Flinck, A., Kheddache, S., and Bath, M. (2008). Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: Human observer study of clinical cases. *Radiology*, 249(3):1034–1041.

Warren, L. M., Given-Wilson, R. M., Wallis, M. G., Cooke, J., Halling-Brown, M. D., Mackenzie, A., Chakraborty, D. P., Bosmans, H., Dance, D. R., and Young, K. C. (2014). The effect of image processing on the detection of cancers in digital mammography. *American Journal of Roentgenology*, 203(2):387–393.

Yoon, H. J., Zheng, B., Sahiner, B., and Chakraborty, D. P. (2007). Evaluating computer-aided detection algorithms. *Medical Physics*, 34(6):2024–2038.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Zanca, F., Hillis, S. L., Claus, F., Van Ongeval, C., Celis, V., Provoost, V., Yoon, H.-J., and Bosmans, H. (2012). Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: Results from independently conducted FROC/ROC studies in mammography. *Med Phys*, 39(10):5917–5929.

Zanca, F., Jacobs, J., Van Ongeval, C., Claus, F., Celis, V., Geniets, C., Provost, V., Pauwels, H., Marchal, G., and Bosmans, H. (2009). Evaluation of clinical image processing algorithms used in digital mammography. *Medical Physics*, 36(3):765–775.