The RJafroc Froc Book

Dev P. Chakraborty, PhD

2022-09-16

Contents

Preface

- Intended as an online update to my print book (Chakraborty, 2017).
- All references in this book to `RJafroc` refer to the R package with that name (case sensitive) (Chakraborty and Zhai, 2022).
- Since its publication in 2017 `RJafroc`, on which the `R` code examples in the book depend, has evolved considerably, causing many of the examples to "break".
- This gives me the opportunity to update the print book.
- The online book has been divided into 3 books.
  - The RJafrocQuickStartBook book.
  - The RJafrocRocBook book.
  - This book RJafrocFrocBook.

FROC paradigm

Chapter 1

The FROC paradigm and search

1.1 TBA How much finished

85%

1.2 Introduction

For diagnostic tasks such as detecting diffuse interstitial lung disease[1], or diseases similar to it, *where disease location is either irrelevant or implicit*, this is an appropriate paradigm in the sense that essential information is not being lost by limiting the radiologist's response to a single rating per case.

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. *Diffuse interstitial lung disease is spread through and confined to the lung.*

For focal disease the ROC paradigm constrains the collected information to a single rating 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 information (Black and Dwyer, 1990; Black, 2000).

From our point of view the most important problem is that neglect of location information leads to loss of statistical power. The reason for this is the additional noise introduced into the AUC measurement due to crediting a radiologist for correctly detecting the diseased condition (a ROC paradigm True Positive or TP event) when the radiologist may have pointed to the wrong location, in which case *the radiologist is credited for a TP when in fact two mistakes were made: the true localized disease was missed and a false location was incorrectly identified as diseased*. Not discriminating between 2 types of TP events, one with correct localizations (associated with an expert radiologist) and the other with incorrect localizations (a radiologist with lesser expertise) leads to reduced ability to distinguish between them (the ROC performance difference will be smaller) as compared to the FROC paradigm described in this book (the FROC performance difference will be larger). This and other problems with using the ROC paradigm in localization tasks are described in (Bunch et al., 1977), a publication that spurred my initial interest in this field (Chakraborty et al., 1986).

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. Key differences between FROC ratings and ROC ratings are noted. The FROC plot is introduced. A "solar" analogy is introduced which yields a good intuitive feel for the FROC paradigm.

1.3 Location specific paradigms

Location-specific paradigms[2] take into account, to varying degrees, information regarding the locations of perceived lesions.

[2]Location-specific paradigms 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.

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).

> In this book *lesion* always refers to a true or real lesion. The term *suspicious region* or *perceived lesion* is reserved for any region that, as far as the observer is concerned, has "lesion-like" characteristics. *A lesion is a real entity while a suspicious region is a perceived entity.*

The 4 panels in Fig. 1.1 show a schematic mammogram interpreted according to the 4 current observer performance paradigms. The arrows point to two lesions and the three light crosses indicate suspicious regions. A marked suspicious region is indicated by a dark cross. Evidently the radiologist found one of the lesions (the light-shaded cross near the left most arrow in the top-left panel), missed the other lesion and mistook two normal structures for lesions (the two light-shaded crosses that are relatively far from any of the lesions). In this example there are three suspicious regions one of which is close to a real lesion.

- In the ROC paradigm, Fig. 1.1 (top-left panel), 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 light-shaded cross is a highly suspicious region then the ROC rating might be 5 (highest confidence for presence of disease). There are no dark-shaded crosses on this panel as no marking occurs in the ROC paradigm.
- In the free-response (FROC) paradigm, Fig. 1.1 (top-right panel), the dark-shaded crosses indicate suspicious regions that were *marked*, and the adjacent numbers are the corresponding ratings. In this example the two ratings shown apply to specific suspicious regions, unlike the ROC paradigm where the single rating applies to the whole image. Assuming the allowed FROC ratings are integers 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 light-shaded cross, was not marked, implying its confidence level did not exceed the lowest reporting threshold for a FROC-1 rating. The marked region rated FROC-4 (the highest FROC confidence level) is likely what caused the radiologist to assign the ROC-5 rating to this image in the ROC paradigm panel.
- In the LROC paradigm, Fig. 1.1 (bottom-left panel), the radiologist rates the 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

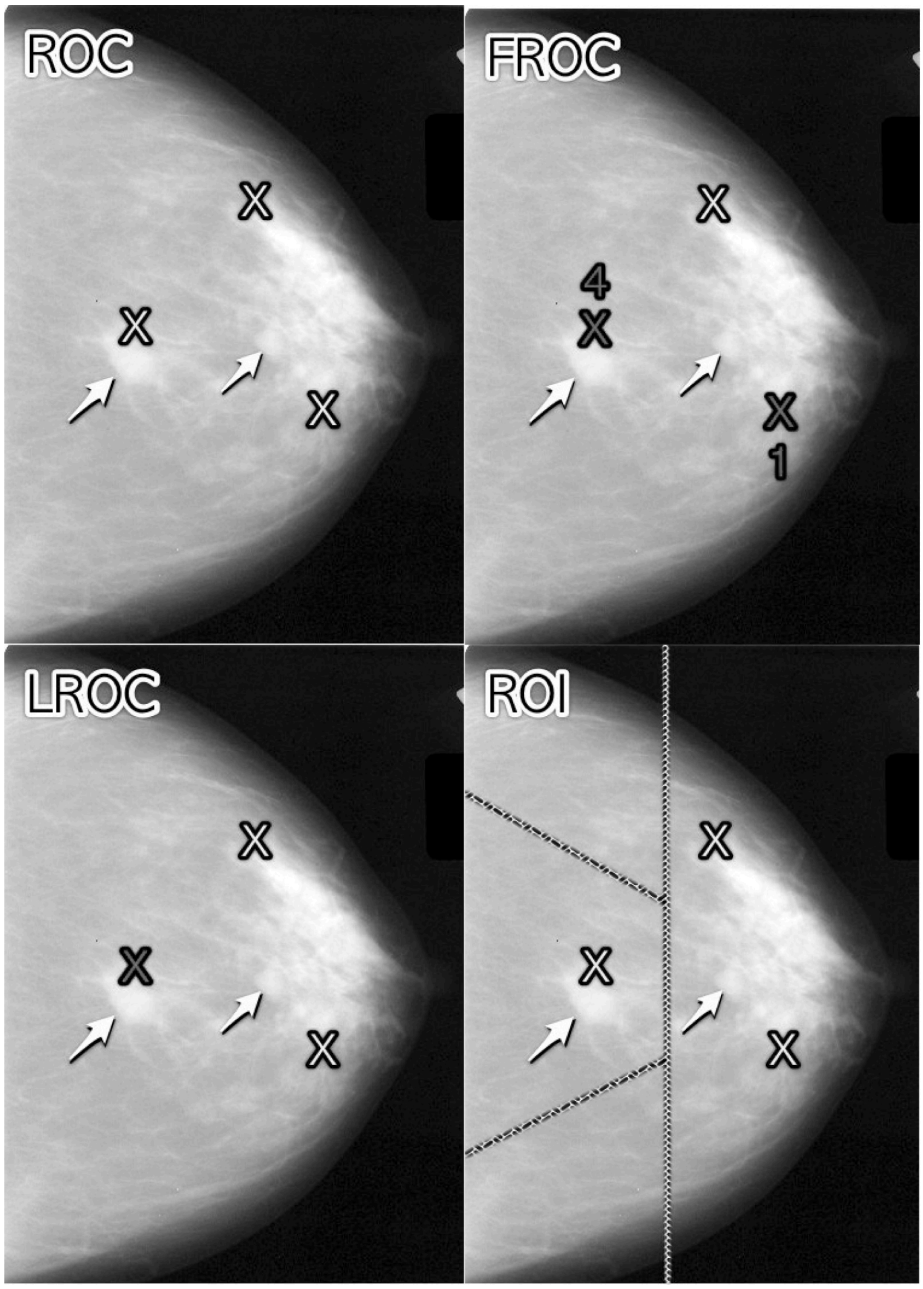


Figure 1.1: Upper Left: ROC, Upper Right: FROC, Lower Left: LROC, Lower Right: ROI

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 panel, 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.

- 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 within the ROI. The rating is similar to the ROC rating, except it applies to the ROI, not the whole image. Assuming a $1 - 5$ positive directed integer rating scale in Fig. 1.1 (bottom-right panel) there are four ROIs. The ROI at ~9 o'clock might be rated ROI-5 as it contains the most suspicious light-shaded cross (the region that was rated FROC-4), the one at ~11 o'clock might be rated ROI-1 as it does not contain any light-shaded crosses, the one at ~3 o'clock might be rated LROC-2 or LROC-3 (the unmarked light-shaded cross would tend to increase the confidence level) and the one at ~7 o'clock might be rated ROI-1[3].

1.4 Visual search

The FROC paradigm in medical imaging is equivalent to a visual search task. Any search task has two components: finding something[4] and acting on 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[5]. Acting on it could be driving to work or drinking milk from the carton[6].

Likewise, a medical imaging search task has two components: finding lesions, without finding too many extraneous suspicious regions, and acting on each finding[7]. Acting on a finding involves determining if it is sufficiently suspicious for cancer to warrant reporting and further patient follow-up. If a suspicious region is found and provided it is sufficiently suspicious for malignancy the region is marked and rated for confidence that it is a malignant lesion.

[3]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 the same number of ROIs.

[4]while not finding irrelevant stuff, a subtle but important point

[5]without finding too many extraneous objects

[6]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).

[7]"Finding" is the actual term used by radiologists in their clinical reports

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 that about 5 out of 1000 patients have cancers, so 99.5% of the time odds are that the patient has no malignant lesions[8]. Considerably search expertise is needed for the radiologist to mark malignant lesions with high probability while not generating too many false marks.

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 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 (Dr. 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 screening BIRADS-0 rating. The screening rating itself is binary: BIRADS-0 for recall (the patient is recalled for a diagnostic workup to determine the final BIRADS rating) or BIRADS-1 for normal or no abnormality detected (the patient comes back about a year later for the next screening appointment).

1.4.1 Proximity criterion and scoring the data

In the first two quasi-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 of an anthropomorphic chest phantom with superposed simulated lesions. Credit for a correct detection and localization, termed a lesion-localization or LL-event[9], 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 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:

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

[9] 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.

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

What is meant by sufficiently close? One adopts an acceptance radius (for spherical lesions) or *proximity criterion* (the more general case). What constitutes "close enough" is a clinical decision the answer to which depends on the application. 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 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. [10]

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. [11] 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. [12]

[10] 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.

[11] 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.

[12] 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.

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 (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.

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.

The rating can be any real number, as long as higher values are associated with higher confidence levels.

If *integer ratings* are used for each recorded mark then in a four-rating FROC study at most 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 [13].

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.

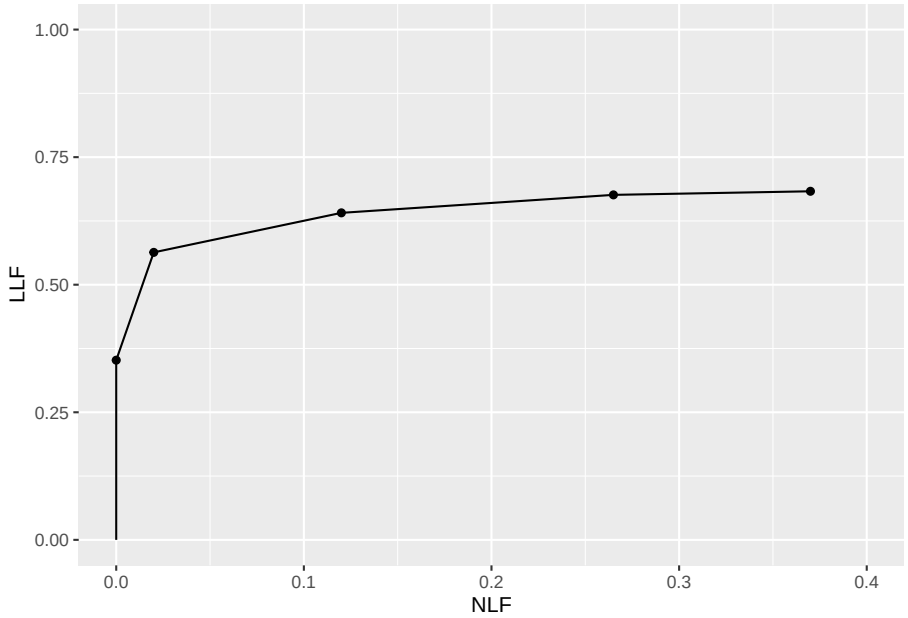
Chapter 2 describes the FROC, and other possible operating characteristics, in more detail.

1.5.1 Illustration with a dataset

The following code uses `dataset04` (Zanca et al., 2009) in `RJafroc` to illustrate an empirical FROC plot. The dataset has 5-treatments and 4 readers, so in principle one can generate 20 plots. In this example I have selected treatment 1 and reader 1 to produce the plot. The reader should experiment by running, for example `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 2, rdrs = 1, opChType = "FROC")$Plot`, i.e., with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "FROC")
print(ret$Plot)
```

[13] I have seen publications that describe a data collection process where the "1" rating is used to mean, in effect, that the observer sees nothing to report in the image, i.e., to mean "let's move on to the next image". This amounts to wasting a confidence level. The user interface should present an explicit "next-image" option and reserve the "1" rating to mean the lowest reportable confidence level.



The study in question was a 5 rating FROC study. The lowest non-trivial point corresponds to the marks rated 5, the next higher one corresponds to marks rated 4 or 5, etc. FROC plots may vary widely in shape but they share the common characteristic that the operating point cannot move downward or to the left as one cumulates lower confidence level marks (i.e., it can only move upward and to the right; the plot can flatten out or it can end at a finite value of NLF).

The plot shown above is termed an *empirical plot* as it consists of the empirical (observed) operating points connected by straight line segments.

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 in the sky. Furthermore assume the observer knows his GPS coordinates and has 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 rational 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, eventually 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 rational 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* lesions while *not finding* non-lesions and search performance is the ability to find more lesions while minimizing 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 false 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

The FROC paradigm is often misunderstood. Some of this has to do with loose terminology and some to misconceptions regarding the meaning of search, the paradigm and the FROC curve. These are summarized below:

- Loose terminology:

  - Using the term "lesion-specific" to describe location-specific paradigms.
  - Using the term "lesion" when one means a "suspicious region" that may or may not be a true lesion.
  - Using ROC paradigm 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 localizations, that apply to regions of the image.
  - 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.

- Misconception
  - A fundamental misunderstanding of search performance embodied in the commonly heard statement "CAD is perfect at search because it looks at everything".
  - Showing FROC curves as reaching the unit ordinate as this is the exception rather than the rule.
  - Believing that FROC curves extend to very large values (potentially infinite) along the abscissa and all the observer has to do to access this region is to lower their reporting threshold.

The FROC plot is historically 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 that have evolved over the years.

1.8 References

Chapter 2

Empirical plots from FROC data

2.1 How much finished

100%

2.2 Introduction

FROC data consists of mark-rating pairs. A distinction is made between *latent* marks (suspicious regions perceived by the visual system but not necessarily marked) and *actual* marks. A key table (used in later chapters) summarizing FROC notation is introduced which allows unambiguous description of the data.

Empirical plots refer to those generated directly from the data. Empirical operating characteristics (empirical plots) introduced in this chapter are the FROC, the (inferred) ROC, the alternative FROC (AFROC), the weighted AFROC (wAFROC), the AFROC1, wAFROC1. Formulae for x and y coordinates of each plot are given in terms of the underlying mark-rating FROC data.

Plots are *visual* depictions of performance. Scalar measures derived from plots can serve as *quantitative* measures of performance. Empirical area under curve (AUC) measures associated with all plots are illustrated with a small FROC dataset. Except for the FROC plot all of the other plots include a straight line extension from the uppermost observed operating point to (1,1).

If one ignores localization information and simply considers the highest rating on each case as representing its ROC rating, one can define the empirical ROC plot and associated area measure ROC-AUC from FROC data. Since ROC-AUC

is a fundamental measure of classification accuracy between non-diseased and diseased cases any other proposed area measure that does not ignore location information should, if it is to be useful, correlate with ROC-AUC. These correlations are explored using the small dataset and it is shown that FROC-AUC is a poor measure of performance. While ways of circumventing FROC-AUC have been proposed and have been used by some investigators none are satisfactory and the claim of this book is that **the FROC should never be used to quantify performance**. The basic reason is simple: unlike all of the other plots defined in this chapter the FROC plot is not constrained to lie within the unit square and the area under a straight line extension to (1,1) is meaningless.

Some of the other empirical plots and AUCs are less familiar as compared to the well-known ROC plots and ROC-AUC. As an aid to understanding them I have included numerical ("hand") calculations of the empirical plots and AUCs for the small dataset. The calculations also illustrate the advantage of using *weighted* versions implemented in some of the empirical plots (lesion weights are a way of allowing one to model the clinical importance (i.e., morbidity/mortality) associated with different type of lesions present in a clinical dataset; a weighted plot assures that each case gets the same importance in determining AUC regardless of the number of lesions in it).

Computing the AUCs from plots can be tedious at best; computational formulae are needed which would allow any of the AUCs to be calculated directly from the FROC ratings. Appendix 1 proves a formula for the wAFROC-AUC, Appendix 2 provides a physical interpretation of the area under the straight line extension for this plot. Appendix 3 summarizes, without proofs, the computational formulae for AUCs for all plots introduced in this chapter.

2.3 FROC data and notation

2.3.1 LLs vs. NLs

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 a *lesion localization* (LL) if it is sufficiently close to a true lesion and otherwise it is recorded as a *non-lesion localization* (NL).

In an FROC study the number of marks on a case is an a-priori unknown non-negative random integer. It is incorrect and naive to estimate it by dividing the anatomically-relevant 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, radiologists don't assign equal attention units to all areas of the image [1].

[1] Currently the best insight into the numbers and locations of marks per case is obtained from eye-tracking studies (Duchowski and Duchowski, 2017), but the information is incomplete

2.3.2 Latent vs. actual marks

To distinguish between suspicious regions that were considered for marking but not necessarily marked 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*. This is an important point.
- In contrast unmarked lesions are observable events – one knows (trivially) which lesions were not marked.

2.3.3 Binning rule

Recall 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 $r = 0, 1, ..., R_{ROC} - 1$, where R_{ROC} is the number of ROC study bins [2] and a binning 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:

$$\left.\begin{aligned} \text{if}\,(\zeta_r \leq z < \zeta_{r+1}) \Rightarrow \text{rating} = r + 1 \\ r = 0, 1, ..., R_{ROC} - 1 \\ \zeta_0 = -\infty \\ \zeta_{R_{ROC}} = \infty \end{aligned}\right\} \tag{2.1}$$

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 binning 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

as eye-tracking studies can only measure *foveal* gaze and not lesions found by *peripheral* vision. Moreover, such studies are near impossible to conduct in a clinical setting (at least with the eye-tracking apparatus that I am familiar with).

[2] The subscript is used to make explicit the paradigm used as otherwise it leads to confusion.

Table 2.1: FROC notation; all marks refer to latent marks.

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

$\zeta_1, \zeta_2, \zeta_3, \zeta_4$ can correspond to a 5-rating ROC study or to a 4-rating FROC study. To summarize:

$$\left.\begin{aligned} \text{if}\,(\zeta_r \le z < \zeta_{r+1}) \Rightarrow \text{rating} = r \\ r = 1, 2, ..., R_{FROC} \\ \zeta_0 = -\infty \\ \zeta_{R_{FROC}+1} = \infty \end{aligned}\right\} \tag{2.2}$$

2.3.4 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 truths. *The notation also has to account for cases with no marks.*

FROC notation is summarized in Table 2.1 in which *marks refer to latent marks.* 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).

2.3.5 Comments

- 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*.

- Row 4: s location-level truth state: 1 for non-diseased region (NL) and 2 for lesion (LL).

- Row 5: Similar to row 3, two indices $l_s s$ are needed to select latent mark l_s in location-level truth state s. As a useful mnemonic, l is for *location*.

- Row 6: $N_{k_t t}$ is the total number of latent NL marks in case $k_t t$. Latent NL marks are possible on non-diseased and diseased cases (i.e., both values of t are allowed).

- Row 7: $L_{k_2 2}$ is the number of lesions in diseased case $k_2 2$.

- Row 8: The z-sample for case $k_t t$ and NL mark $l_1 1$ is denoted $z_{k_t t l_1 1}$. 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 9: The z-sample of a latent LL is $z_{k_2 2 l_2 2}$. Unmarked lesions are observable events assigned negative infinity ratings (the null-set notation is unnecessary).

- Row 10: R_{FROC} is the number of bins in the FROC study.

- Rows 11, 12 and 13: 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 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 in a case sum to unity: $\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1$.

- 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.

2.3.6 A conceptual and notatonal issue

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 [3]. In fact the first model 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, actual FROC datasets do not agree with this thinking. This is one reason I introduced the radiological search model (RSM) (Chakraborty, 2006b). I will have more to say about this in Chapter 4, but for now I state one 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 distribution. This means that the actual number of latent NL marks per case can be 0, 1, 2, .., whose average (over all cases) is a finite number. It is highly unlikely that any case will have an infinite number of NLs.

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 cases with no latent NL marks. Basically it involves restricting summations over cases to those cases which have at least one latent NL mark, i.e., $N_{k_t t} > 0$, as in the following:

- $l_1 = \{1, 2, ..., N_{k_t t}\}$ indexes latent NL marks, provided the case has at least one latent NL mark; 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").

- $l_2 = \left\{1, 2, ..., L_{k_2 2}\right\}$ indexes latent LL marks. Unmarked LLs are assigned negative infinity ratings as these are observable events. The null set notation is not needed because for every diseased case $L_{k_2 2} > 0$.

[3] 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.

2.4 The FROC plot

Definitions:

- $NLF_r \equiv NLF(\zeta_r)$ = cumulated NL counts with z-sample $\geq$ threshold ζ_r divided by total number of cases.
- $LLF_r \equiv LLF(\zeta_r)$ = cumulated LL counts with z-sample $\geq$ threshold ζ_r divided by total number of lesions.

Definitions:

> The empirical FROC plot connects adjacent operating points $(\mathrm{NLF}_r, \mathrm{LLF}_r)$, 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}. **Warning: this is a particularly dangerous figure of merit, as will shortly become clear.**

Using the notation of Table 2.1 and assuming binned data[4] and $n(x)$ denotes the number of events x:

$$\mathrm{NLF}_r = \frac{n\left(\text{NLs rated} \geq \zeta_r\right)}{K_1 + K_2} \tag{2.3}$$

and

$$\mathrm{LLF}_r = \frac{n\left(\text{LLs rated} \geq \zeta_r\right)}{L_T} \tag{2.4}$$

The allowed values of r are:

$$r = 1, 2, ..., R_{FROC} \tag{2.5}$$

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 (2.3) and (2.4) are equivalent to:

$$\mathrm{NLF}_r = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right) \tag{2.6}$$

[4]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.

and

$$\mathrm{LLF}_r = \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{2.7}$$

The indicator function is defined as unity if the argument is true and zero otherwise:

$$\left.\begin{array}{rcl} \mathbb{I}\,(\text{True}) & = & 1 \\ \mathbb{I}\,(\text{False}) & = & 0 \end{array}\right\} \tag{2.8}$$

In Eqn. (2.6) $\mathbb{I}\left(N_{k_t t} > 0\right)$ ensures that *only cases with at least one latent NL* are included in the summation (recall that $N_{k_t t}$ is the total number of latent NLs in case $k_t t$). 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 right hand side yields 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. (2.7) a summation over t is not needed as only diseased cases contribute to LLF. A term like $\mathbb{I}\left(L_{k_2 2} > 0\right)$ 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.

Since $\zeta_{R_{FROC}+1} = \infty$ according to Eqn. (2.6) and Eqn. (2.7) $r = R_{FROC} + 1$ yields the trivial operating point (0,0).

2.4.1 The observed FROC end-point and its semi-constrained property

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

$$\mathrm{NLF}_1 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \tag{2.9}$$

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

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

$$\left.\begin{aligned}\mathrm{LLF}_1 =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_1\right)\\ \leq& 1\end{aligned}\right\} \tag{2.10}$$

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 *ordinate* is constrained to the range (0,1) the *abscissa* is not.

2.4.2 Futility of extrapolation outside the observed end-point

To understand this consider the expression for NLF_0, i.e., using Eqn. (2.6) with $r = 0$:

$$\mathrm{NLF}_0 = \frac{1}{K_1 + K_2}\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right)\sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq -\infty\right) \tag{2.11}$$

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. (2.9) but the 2nd term cannot be evaluated:

$$\left.\begin{aligned}\text{1st term} =& \left(\frac{1}{K_1 + K_2}\right)\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right)\sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right)\\ =& \mathrm{NLF}_1\\ \text{2nd term} =& \left(\frac{1}{K_1 + K_2}\right)\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right)\sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(-\infty \leq z_{k_t t l_1 1} < \zeta_1\right)\\ =& \frac{\text{unknown number}}{K_1 + K_2}\end{aligned}\right\} \tag{2.12}$$

The 2nd term 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:

$$\left.\begin{aligned}\mathrm{LLF}_0 =& \frac{\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq -\infty\right)}{L_T} \\ =& 1\end{aligned}\right\} \tag{2.13}$$

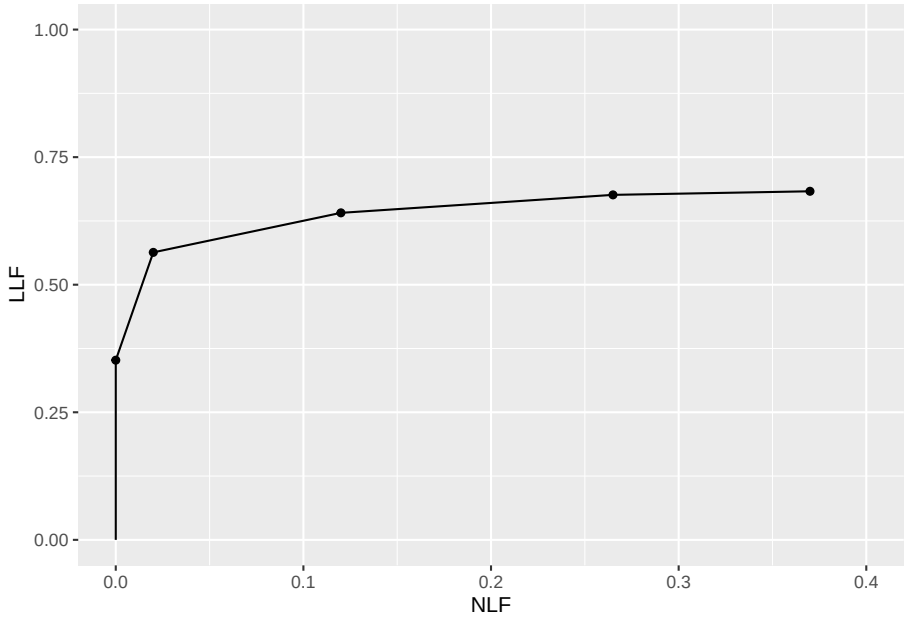
Unlike unmarked latent NLs, *unmarked lesions can safely be assigned the* $-\infty$ *rating, because an unmarked lesion is an observable event.* The right hand side of Eqn. (2.13) evaluates to unity. However, since the corresponding abscissa NLF_0 is undefined, one cannot plot this point. It follows that one cannot extrapolate outside the observed end-point.

The above formalism should not obscure the fact that the futility of extrapolation outside the observed end-point of the FROC is obvious for scientific reasons: extrapolating outside the range of the observed data is generally not a good idea.

2.4.3 Illustration with a dataset

The following plot uses `dataset04` (Zanca et al., 2009) in `RJafroc` to illustrate an empirical FROC plot. This dataset has $L_{max} = 3$, $\max(N_{k_t t}) = 3$ and a 5-point rating scale was employed. The following plot applies to reader 1 in modality (treatment) 1 only. The full dataset has 5 modalities and 4 readers.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "FROC")
print(ret$Plot)
```



Shown next are FROC-AUCs for this dataset calculated using the formula in Eqn. (2.45). All 20 modality-reader combinations are shown.

```
auc_froc <- as.data.frame(UtilFigureOfMerit(dataset04, FOM = "FROC"))
print(auc_froc)
#>           rdr1      rdr3      rdr4       rdr5
#> trt1 0.2361972 0.1085035 0.2268486 0.09922535
#> trt2 0.2192077 0.2231338 0.4793310 0.18450704
#> trt3 0.1947359 0.1063028 0.2543662 0.15137324
#> trt4 0.2198768 0.1307394 0.3293662 0.13882042
#> trt5 0.1800528 0.1097535 0.3015141 0.16563380
```

The value 0.2361972 for `trt1` and `rdr1` is the area under the FROC plot shown above.

2.5 The inferred-ROC plot

By adopting a rule for converting the mark-rating data per case to a single rating per case, and commonly the highest rating rule is used [5], it is possible to infer ROC data from FROC mark-rating data.

[5] 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.

2.5.1 The inferred-ROC rating

The rating of the highest rated mark in 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 end-point to (1,1).*

2.5.2 Inferred FPF

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

$$\left.\begin{array}{rcl} FP_{k_1 1} = & \max_{l_1}\left(z_{k_1 1 l_1 1}\right) & \text{if} \quad l_1 \neq \emptyset \\ FP_{k_1 1} = & -\infty & \text{if} \quad l_1 = \emptyset \end{array}\right\} \tag{2.14}$$

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 (this is allowed because a non-diseased case with no marks is an observable event). The corresponding false positive fraction is defined by:

$$\mathrm{FPF}_r \equiv \mathrm{FPF}\left(\zeta_r\right) = \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_r\right) \tag{2.15}$$

2.5.3 Inferred TPF

The inferred true positive (TP) z-sample for diseased case $k_2 2$ is defined by one of the following three equations, as explained below:

$$TP_{k_2 2} = \max_{l_1 l_2} \left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2} \right) \quad \text{if} \quad l_1 \neq \emptyset \tag{2.16}$$

or

$$TP_{k_2 2} = \max_{l_2} \left(z_{k_2 2 l_2 2} \right) \quad \text{if} \quad (l_1 = \emptyset) \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) > -\infty \right) \tag{2.17}$$

or

$$TP_{k_2 2} = -\infty \quad \text{if} \quad \left(l_1 = \emptyset \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) = -\infty \right) \right) \tag{2.18}$$

Here $\wedge$ is the logical AND operator. An explanation is in order. Consider Eqn. (2.16). There are two z-samples inside the max operator: $z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}$. The first z-sample is from a NL on a diseased case, as per the $l_1 1$ subscripts, while the second is from a LL on the same diseased case, as per the $l_2 2$ subscripts.

- If $l_1 \neq \emptyset$ then Eqn. (2.16) applies, i.e., one takes the maximum over all z-samples, NLs and LLs, whichever is higher, on the diseased case.
- If $l_1 = \emptyset$ and at least one lesion is marked, then Eqn. (2.17) applies, i.e., one takes the maximum z-sample over all marked LLs.
- If $l_1 = \emptyset$ and no lesions are marked, then Eqn. (2.18) 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:

$$\text{TPF}_r \equiv \text{TPF}(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I} \left(TP_{k_2 2} \geq \zeta_r \right) \tag{2.19}$$

2.5.4 The empirical ROC plot and AUC

Definitions:

> The inferred empirical ROC plot connects adjacent points $(\text{FPF}_r, \text{TPF}_r)$, 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}.

2.5.5 The observed end-point of the ROC and its constrained property

The abscissa of the observed end-point FPF_1, is defined by:

$$\text{FPF}_1 \equiv \text{FPF}\left(\zeta_1\right) = \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_1\right) \tag{2.20}$$

Since each case gets a single FP rating, and only unmarked cases get the $-\infty$ rating, $\text{FPF}_1 \leq 1$.

The ordinate of the observed end-point TPF_1, is defined by:

$$\text{TPF}_1 \equiv \text{TPF}(\zeta_1) = \frac{1}{K_2}\sum_{k_2=1}^{K_2} \mathbb{I}\left(TP_{k_2 2} \geq \zeta_1\right) \tag{2.21}$$

Since each case gets a single TP rating, and only unmarked cases get the $-\infty$ rating, $\text{TPF}_1 \leq 1$.

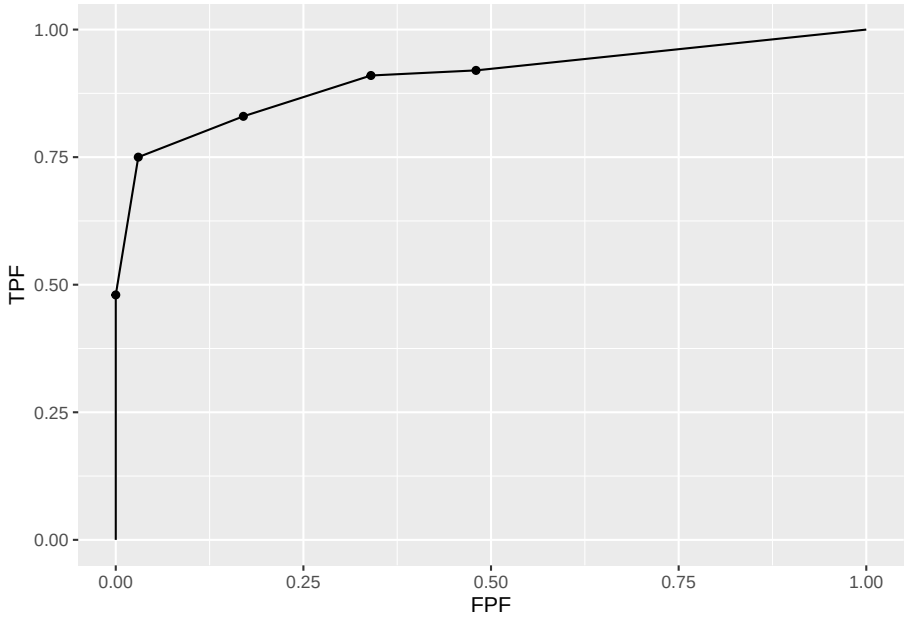
It follows that the observed end-point of the ROC (as is well known) satisfies the constrained end-point property: it lies below-left the (1,1) corner of the plot.

> The upper-right corner (reached by counting all ratings $\geq -\infty$) of the ROC plot is not to be confused by the observed end-point (reached by counting all ratings $\geq \zeta_1$).

2.5.6 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot for treatment 1 and reader 1. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = ROC")$Plot` with different treatments `trts` and readers `rdrs` specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "ROC")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset. Note that in function `UtilFigureOfMerit()` the `FOM` argument has to be set to `HrAuc`, for highest rating AUC.].

```
UtilFigureOfMerit(dataset04, FOM = "HrAuc")
#>         rdr1    rdr3    rdr4    rdr5
#> trt1 0.90425 0.79820 0.81175 0.86645
#> trt2 0.86425 0.84470 0.82050 0.87160
#> trt3 0.81295 0.81635 0.75275 0.85730
#> trt4 0.90235 0.83150 0.78865 0.87980
#> trt5 0.84140 0.77300 0.77115 0.84800
```

2.6 The alternative FROC (AFROC) plot

- Fig. 4 in (Bunch et al., 1977) anticipated another way of visualizing FROC data. I subsequently termed 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.
- $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$, the ordinate of the FROC plot, was defined in Eqn. (2.7).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$, the abscissa of the ROC plot, was defined in Eqn. (2.15).

2.6.1 Definition: empirical AFROC plot and AUC

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 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 a hybrid plot incorporating aspects of both ROC and FROC plots.
- The AFROC is constrained to within the unit square.

> 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 plot would later turn out to be, otherwise a more meaningful name might have been proposed. To anticipate the central message of this book, the AUC based on this plot (and weighted versions of it introduced below), are superior to the FROC-AUC and the ROC-AUC in terms of statistical power and reliability (the FROC-AUC is especially unreliable).

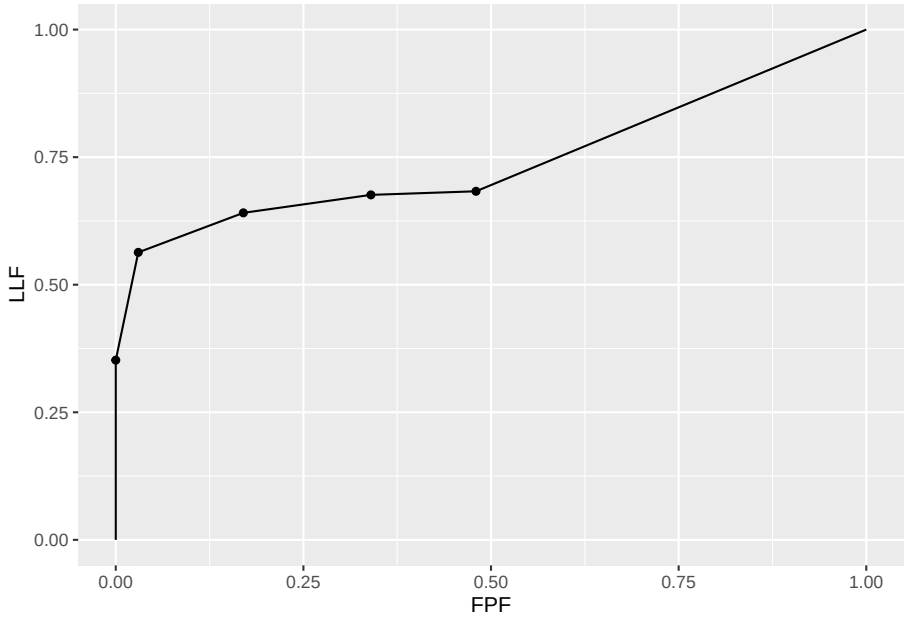
2.6.2 The observed end-point of the AFROC and its constrained property

According to Eqn. (2.15) the abscissa of the observed end-point $FPF_1 \leq 1$ and according to Eqn. (2.10) the ordinate of the observed end-point $\text{LLF}_1 \leq 1$. It follows that the observed end-point of the AFROC satisfies the constrained end-point property, i.e., it lies below-left the (1,1) corner of the plot.

2.6.3 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical AFROC plot for treatment 1 and reader 1.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "AFROC")
print(ret$Plot)
```



Shown next are the figures of merit for this dataset for all treatment reader combinations.

```
UtilFigureOfMerit(dataset04, FOM = "AFROC")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.7427113 0.7104930 0.7003169 0.7909859
#> trt2 0.7586972 0.7161620 0.7225352 0.7927465
#> trt3 0.6983451 0.6955282 0.6777817 0.7547535
#> trt4 0.7817606 0.7234507 0.7132746 0.8136268
#> trt5 0.7169718 0.6690845 0.6587324 0.7682042
```

2.7 The weighted-AFROC plot (wAFROC) plot

The AFROC ordinate defined in Eqn. (2.7) gives equal importance to every lesion in a case. A case with more lesions will have more influence on the AFROC (see next section 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 an equally valid sample from the population. In particular, one does not want the analysis to be skewed towards cases with greater numbers of lesions. [6]

[6]Historical note: I became aware of how serious this issue could be when a researcher contacted me about using FROC methodology for nuclear medicine bone scan images, where

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., short for 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., 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:

$$\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1 \tag{2.22}$$

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

$$\text{wLLF}_r \equiv \text{wLLF}\left(\zeta_r\right) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{2.23}$$

2.7.1 The empirical wAFROC plot and AUC

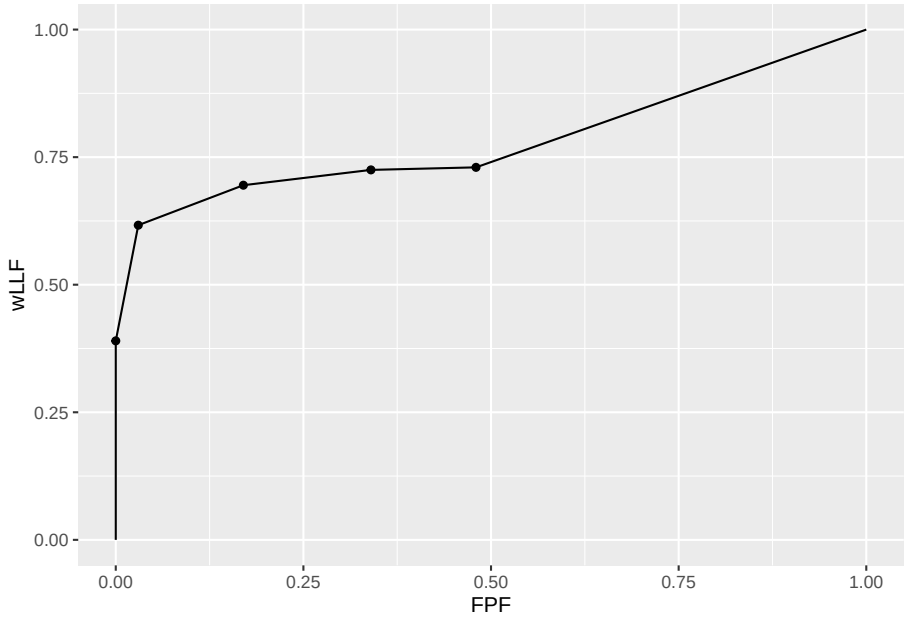
The empirical wAFROC plot connects adjacent operating points $(\text{FPF}_r, \text{wLLF}_r)$, including the origin (0,0), with straight lines plus a straight-line segment connecting the observed end-point to (1,1). The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC}.

2.7.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot for treatment 1 and reader 1.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04, trts = 1, rdrs = 1, opChType = "wAFROC")
print(ret$Plot)
```

the number of lesions on diseased cases can vary from a few to a hundred!



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "wAFROC")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.7792667 0.7248917 0.7036250 0.8050917
#> trt2 0.7870000 0.7269000 0.7226167 0.8037833
#> trt3 0.7296917 0.7157583 0.6723083 0.7726583
#> trt4 0.8101333 0.7431167 0.6943583 0.8294083
#> trt5 0.7488000 0.6822750 0.6551750 0.7712500
```

2.8 AFROC vs. wAFROC

The fact that the wAFROC gives equal importance to each diseased case while the AFROC gives more importance to diseased cases with more lesions can be illustrated with a fictitious small dataset consisting of $K_1 = 4$ non-diseased and $K_2 = 5$ diseased cases. The maximum number of NLs per case is two and the maximum number of lesions per case is three. The first two diseased cases have one lesion each, the third and fourth have two lesions each and the fifth has 3 lesions. Here is how we code the NL and LL ratings (`t()` is the `R` transpose operator). The negative infinities represent unmarked locations. For example, the first non-diseased case has no NL marks, the second has one mark rated 0.5, etc., and the first diseased case has one NL mark rated 1.5, etc. The first lesion in the LL array was rated 0.9. the second was rated -0.2, ..., and the 3 lesions in the fifth diseased case were rated 1, 2.5 and 1, respectively.

```
NL <- t(array(c(-Inf, -Inf,
                 0.5, -Inf,
                 0.7, 0.6,
                -0.3, -Inf,
                 1.5, -Inf,
                -Inf, -Inf,
                -Inf, -Inf,
                -Inf, -Inf,
              -  Inf, -Inf), dim = c(2,9)))
LL <- t(array(c(0.9, -Inf, -Inf,
               -0.2, -Inf,-Inf,
                1.6, -Inf, -Inf,
                  3,    2, -Inf,
                  1,    2.5,  1), dim = c(3,5)))
```

The ratings are converted to a dataset `frocData` as shown next:

```
frocData <- Df2RJafrocDataset(NL, LL, perCase = c(1,1,2,2,3))
```

In the above code `perCase = c(1,1,2,2,3)` specifies the number of lesions per case: 1 in the first diseased case, 1 in the second, 2 in the third, ..., and 3 in the fifth. The function `Df2RJafrocDataset()` generates the dataset object.

The lesion weights are specified in the following lines.

```
frocData$lesions$weights[3,] <- c(0.1, 0.9, -Inf)
frocData$lesions$weights[4,] <- c(0.9, 0.1, -Inf)
frocData$lesions$weights[5,] <- c(0.3, 0.4, 0.3)
```

The first and second diseased cases, which have only one lesion each, are assigned unit weights by default. The first lesion in the third diseased case has weight 0.1 and the second has weight 0.9 – notice that the weights sum to unity. The fourth diseased cases has the lesion weights reversed, 0.9 and 0.1. The three lesions in the fifth diseased case are assigned weights 0.3. 0.4 and 0.3.

2.8.1 NL and LL ratings

Shown next is the `NL` ratings array; it has 9 rows, corresponding to the total number of cases (the first four correspond to non-diseased cases and the rest to diseased cases) and 2 columns, corresponding to the maximum number of NLs per case.

```
#> NL ratings:
```

```
#>       [,1] [,2]
#>  [1,] -Inf -Inf
#>  [2,]  0.5 -Inf
#>  [3,]  0.7  0.6
#>  [4,] -0.3 -Inf
#>  [5,]  1.5 -Inf
#>  [6,] -Inf -Inf
#>  [7,] -Inf -Inf
#>  [8,] -Inf -Inf
#>  [9,] -Inf -Inf
```

Shown next is the `LL` ratings array; it has 5 rows, corresponding to the total number of diseased cases, and 3 columns, corresponding to the maximum number of LLs per case:

```
#> LL ratings:
#>      [,1] [,2] [,3]
#> [1,]  0.9 -Inf -Inf
#> [2,] -0.2 -Inf -Inf
#> [3,]  1.6 -Inf -Inf
#> [4,]  3.0  2.0 -Inf
#> [5,]  1.0  2.5    1
```

2.8.2 Lesion weights

Show next is the lesion weights array:

```
#> lesion weights:
#>      [,1] [,2] [,3]
#> [1,]  1.0 -Inf -Inf
#> [2,]  1.0 -Inf -Inf
#> [3,]  0.1  0.9 -Inf
#> [4,]  0.9  0.1 -Inf
#> [5,]  0.3  0.4  0.3
```

The negative infinities represent missing values.

2.8.3 FPF

Shown next is the `FP` ratings array. Since FPs are only possible on non-diseased cases, this is a length 4 row-vector. Each value is the maximum of the two `NL` ratings for the corresponding non-diseased case. As an example, for case #3 the maximum of the two `NL` values is 0.7.

```
#> FP ratings:
#> [1] -Inf  0.5  0.7 -0.3
```

Here are the sorted FP ratings.

```
#> [1] -Inf -0.3  0.5  0.7
```

The sorting makes it easy to construct the FPF values, shown next.

```
#> FPF values:
#>  0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.250 0.500 0.500 0.750 1.000
```

The first non-zero FPF value is 0.25 = 1/4, which occurs when a conceptual sliding threshold is lowered past the highest FP value, namely 0.7. (The 0.25 comes from 1 FP case divided by 4 non-diseased cases.) The next FPF value is 0.5 = 2/4, which occurs when the sliding threshold is lowered past the next-highest FP value, namely 0.5. The next FPF value is 0.75 and the last FPF value is unity.

2.8.4 LLF

Here are the sorted LL ratings.

```
#>  [1] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -0.2  0.9  1.0  1.0  1.6  2.0  2.5  3.0
```

The LLF values are shown next.

```
#> LLF values:
#>  0.000 0.111 0.222 0.333 0.444 0.667 0.778 0.778 0.778 0.889 0.889 1.000
```

The first non-zero LLF value is 0.111, which occurs when the sliding threshold is lowered past the highest LL value, namely 3. The 0.111 comes from 1 LL divided by 9, the total number of lesions. The next LLF value is 0.222, which occurs when the sliding threshold is lowered past the next-highest LL value, namely 2.5 (2/9 = 0.222). The next LLF value is 0.333, which occurs when the sliding threshold is lowered past 2 (3/9 = 0.333), and so on.

2.8.5 wLLF

The sorted LL ratings array and the weights are used to construct the wLLF values shown next.

```
#> wLLF values:
#>  0.000 0.180 0.260 0.280 0.300 0.420 0.620 0.620 0.620 0.820 0.820 1.000
```

The first non-zero `wLLF` value is 0.18, which occurs when the sliding threshold is lowered past the highest `LL` value, namely 3. Since this comes from lesion #1 on diseased case #4, whose weight is 0.9, the corresponding incremental vertical jump is $1/5 * 0.9 = 0.18$, which is also the net `wLLF` value corresponding to the most suspicious lesion crossing the cutoff. Notice that we are dividing by 5, the total number of diseased cases, not 9 as in the `LLF` example.

The next `wLLF` value is 0.26, which occurs when the sliding threshold is lowered past the next-highest `LL` value, namely 2.5, which comes from the 2nd lesion on the fifth diseased case with weight 0.4. The incremental jump in `wLLF` is $1/5 * 0.4 = 0.08$. The net `wLLF` value corresponding to the two most suspicious lesions crossing the cutoff is $1/5 * 0.9 + 1/5 * 0.4 = 0.26$.

The next `wLLF` value is 0.280, which occurs when the sliding threshold is lowered past 1.6, which comes from lesion #1 on diseased case #3, with weight 0.1, and the net `wLLF` value corresponding to the three most suspicious lesions crossing the cutoff is $1/5 * 0.9 + 1/5 * 0.4 + 1/5 * 0.1 = 0.280$, and so on.

The reader should complete these hand-calculations to reproduce all of the `wLLF` values shown above. The values (FPF, LLF and wLLF) defining the AFROC and wAFROC are summarized here:

```
#>      FPF       LLF wLLF
#> 1  0.00 0.0000000 0.00
#> 2  0.00 0.1111111 0.18
#> 3  0.00 0.2222222 0.26
#> 4  0.00 0.3333333 0.28
#> 5  0.00 0.4444444 0.30
#> 6  0.00 0.6666667 0.42
#> 7  0.00 0.7777778 0.62
#> 8  0.25 0.7777778 0.62
#> 9  0.50 0.7777778 0.62
#> 10 0.50 0.8888889 0.82
#> 11 0.75 0.8888889 0.82
#> 12 1.00 1.0000000 1.00
```

This shows that the empirical AFROC is defined by the following 6 operating points: (0,0), (0,0.7777778), (0.5,0.7777778), (0.5,0.8888889), (0.75, 0.8888889) and (1,1). Likewise, the empirical wAFROC is defined by the following 6 operating points: (0,0), (0,0.62), (0.5,62), (0.5,0.82), (0.75, 0.82) and (1,1). In each case one simply connects neighboring points with straight lines.

The hand-calculations also show why the AFROC gives more importance to diseased cases with more lesions while the wAFROC does not.

- Considering the AFROC, diseased case #5 with three lesions which contributes three vertical jumps to LLF totaling $3/9 = 0.333333$ [7]. This is larger than the contribution to LLF of diseased case #1 with one lesion $1/9 = 0.11111$.
- Considering the wAFROC, the three lesions on diseased case #5 contribute $1/5 * 0.3 + 1/5 * 0.4 + 1/5 * 0.3 = 0.2$ to wLLF, the same as diseased case #1, $1/5 * 1 = 0.2$.

Shown in Fig. 2.1 are the empirical AFROC and wAFROC plots.

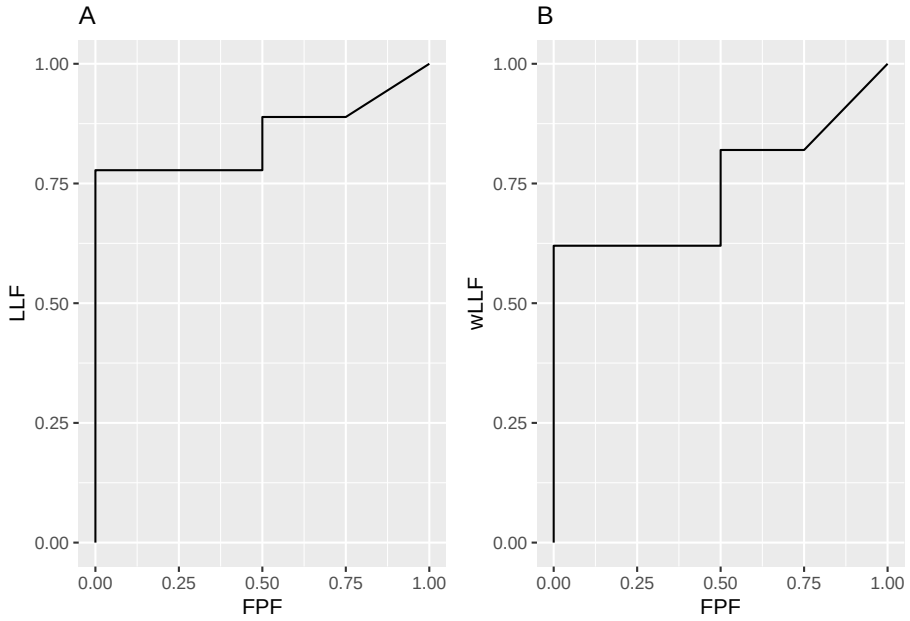


Figure 2.1: Left: AFROC plot; Right: corresponding wAFROC plot.

The operating points can be used to numerically calculate the AUCs under the empirical AFROC and wAFROC plots, as done in the following code:

```
afroc_auc <- 0.5 * 0.7777778 +
  0.25 * 0.8888889 +
  0.25 * 0.8888889 + (1 - 0.8888889) * 0.25 /2

wafroc_auc <- 0.5 * 0.62 +
```

[7] The jumps need not be contiguous: they will be contiguous only if the three lesion ratings are closely spaced such that they are crossed in succession, in any order, by the sliding virtual threshold; otherwise the jumps will be interspersed by jumps from lesions in other cases.

```
  0.25 * 0.82 +
  0.25 * 0.82 +
  (1 - 0.82) * 0.25 /2

cat("afroc_auc =", afroc_auc,"\n")
#> afroc_auc = 0.8472222
cat("wafroc_auc =", wafroc_auc,"\n")
#> wafroc_auc = 0.7425
```

The same AUC results are obtained using the function `UtilFigureOfMerit`:

```
cat("AFROC AUC = ",
    as.numeric(UtilFigureOfMerit(frocData, FOM = "AFROC")),"\n")
#> AFROC AUC =  0.8472222
cat("wAFROC AUC = ",
    as.numeric(UtilFigureOfMerit(frocData, FOM = "wAFROC")),"\n")
#> wAFROC AUC =  0.7425
```

It is seen that the empirical plots consist of upward and rightward jumps starting from the origin (0,0) and ending at (1,1). 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, i.e., correct decisions are rewarded and incorrect ones are penalized. If there are only upward jumps then the empirical 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 of the plot to (1,1) ensures that the net area is unity. If there are only horizontal jumps the operating point moves from the origin to (1,0), where none of the lesions are localized and every non-diseased case has at least one NL mark and despite the straight line extension to (1,1), the net area is zero. This represents worst possible performance.

2.9 Interpretation of AUCs

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

2.10 Instructive examples

I am including a few extreme cases that I have found to be instructive. These include chance level performance and observers who do not generate any marks.

2.10.1 The FROC

The chance level FROC is a "flat-liner" hugging the x-axis except for a possible upturn at large NLF. For an observer who does not generate any marks the FROC plot contains but one point, the origin, and $A_{\mathrm{FROC}} = 0$.

2.10.2 The ROC

The chance level ROC is the positive diagonal connecting (0,0) to (1,1). There could be several operating points on this diagonal (apart from sampling effects) but $A_{\mathrm{ROC}} = 0.5$.

An observer who does not generate any marks the ROC plot consists of two points, the origin and (1,1) and $A_{\mathrm{ROC}} = 0.5$.

2.10.3 The AFROC

2.10.3.1 Chance level performance

The chance level AFROC is not the line connecting (0,0) to (1,1). This is a serious misconception that I have encountered. A chance level observer will generate a "flat-liner" but this time the plot ends at (1,0) and the straight line extension will be a vertical line connecting (1,0) to (1,1) and $A_{\mathrm{AFROC}} = 0$.

2.10.3.2 Case of no marks

This is a highly interesting and instructive example. The AFROC plot is a straight line connecting (0,0) and (1,1) which could be mistakenly termed as representing chance level performance. This is far from the truth.

> An expert radiologist successfully screens out non-diseased cases and sees nothing suspicious in any of them – not mistaking variants of normal anatomy for false lesions on non-diseased cases is a sign of expertise. Suppose the lesions on diseased cases are very difficult to see, even for the expert, so the radiologist does not mark any of them in addition to not marking any NLs on diseased cases. **The expert radiologist therefore does not report anything, i.e.,**

generates no marks, and the operating point is "stuck" at the origin (0,0). 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 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.

2.10.4 The wAFROC

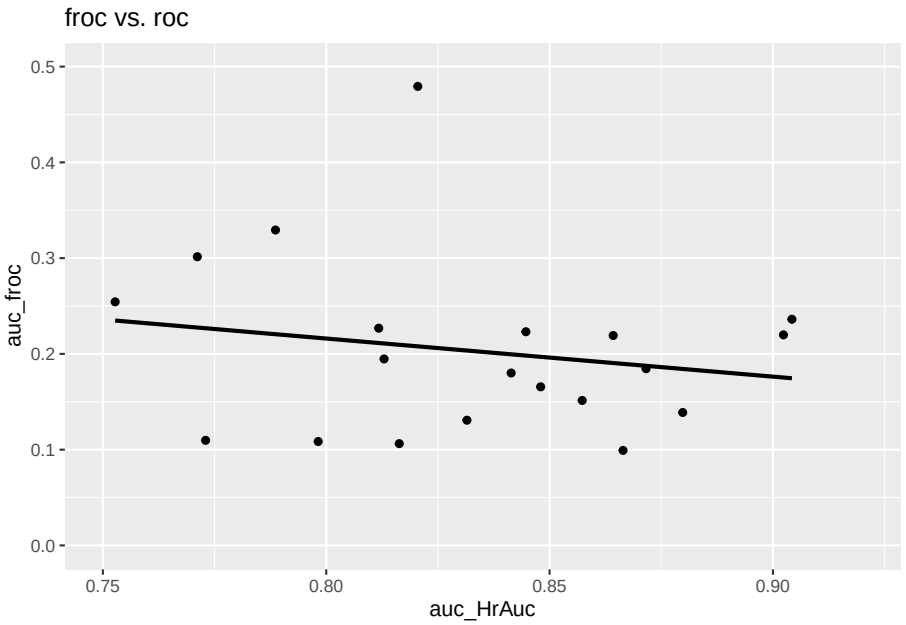
Similar comments apply to the wAFROC as already described above for AFROC.

2.11 FROC-AUC is a poor measure

Regarding the ROC-AUC, i.e., A_{ROC}, as the gold standard against which all other figures of merit should be compared for consistency in orderings, shown next are plots of A_{FROC}, A_{AFROC} and A_{wAFROC} vs. A_{ROC} for the dataset used in the previous illustrations.

2.11.1 Plot of FROC AUC vs. ROC AUC

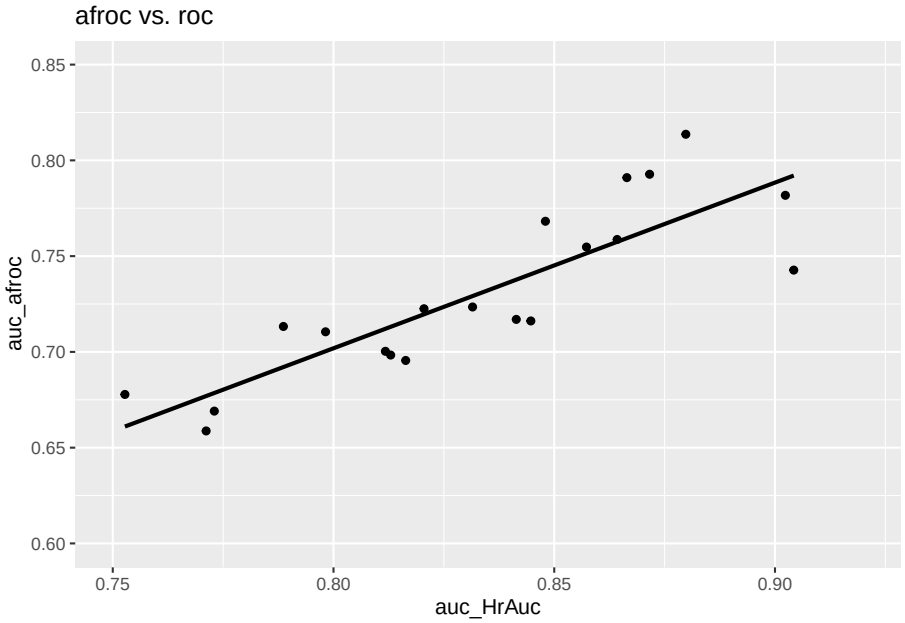
The following is the plot of A_{FROC} vs. A_{ROC}. There are 20 points on the plot corresponding to 5 treatments and 4 readers. The straight line is a least squares fit. Note the poor correlation and negative slope between A_{FROC} and A_{ROC}, $R^2 = 0.0347791$, slope = -0.3978636.



The reason should be fairly obvious. The FROC is unconstrained in the NLF direction and the area under the plot *rewards* an observer who generates more NLs, i.e., as the operating point moves further to the right. (The perfect observer whose FROC plot is the vertical line connecting (0,0) and (0,1) is heavily penalized since $A_{\mathrm{FROC}} = 0$ for this observer.) One can try to try to avoid this problem by limiting the area under the FROC to that between NLF = 0 and NLF = x where x is an arbitrarily chosen fixed value – indeed the partial area procedure has been used by CAD algorithm designers. Since the choice of x is arbitrary the procedure is subjective. The method would fail for any observer with $\mathrm{NLF}_{max} < x$ as then the partial area is undefined. This forces the algorithm designer to chose x as the minimum of all NLF_{max} values over all observers and treatments, which would exclude a lot of data and lead to a statistical power penalty.

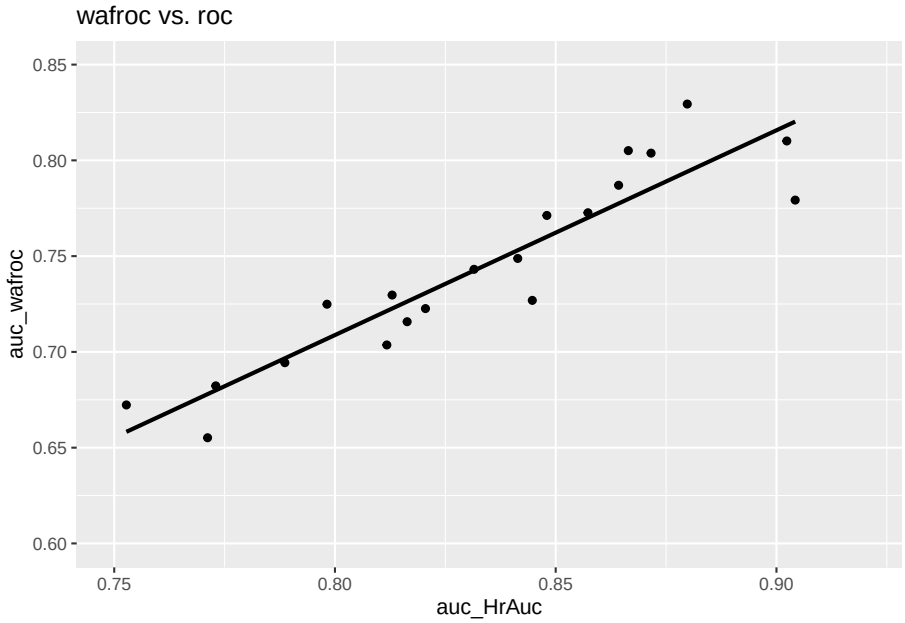
2.11.2 Plot of AFROC AUC vs. ROC AUC

The following is the plot of A_{AFROC} vs. A_{ROC}. This time there is a strong positive correlation between the two, $R^2 = 0.7258723$, slope = 0.8649687. The reason is that the AFROC is fully contained in the unit square. An observer who generates more NL marks will yield smaller A_{AFROC} – as the abscissa of the AFROC approaches unity the restriction to the unit square ensures that AUC will decrease.



2.11.3 Plot of wAFROC AUC vs. ROC AUC

The following is the plot of A_{wAFROC} vs. A_{ROC}. Again, there is a strong positive correlation between the two, $R^2 = 0.8569511$, slope $= 1.0691159$. The reason is that the wAFROC is also fully contained in the unit square.



2.12 The AFROC1 plot

Historically the AFROC originally used a different definition of FPF, resulting in what 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. (2.14)]:

$$\left.\begin{array}{rcl} FP^1_{k_1 t} = & \max_{l_1}\left(z_{k_t t l_1 1}\right) & \text{if} \quad l_1 \neq \emptyset \\ FP^1_{k_t t} = & -\infty & \text{if} \quad l_1 = \emptyset \end{array}\right\} \tag{2.24}$$

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

$$\left.\begin{array}{l} FPF^1_r \equiv FPF^1_r\left(\zeta_r\right) \\ \qquad = \dfrac{1}{K_1 + K_2} \displaystyle\sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(FP^1_{k_t t} \geq \zeta_r\right) \end{array}\right\} \tag{2.25}$$

Note the subtle differences between Eqn. (2.15) and Eqn. (2.25). The latter counts "FPs" on non-diseased and diseased cases while Eqn. (2.15) 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 generate both a TP and a FP may be questionable, however, this plot is useful in applications where all or almost all cases are diseased.

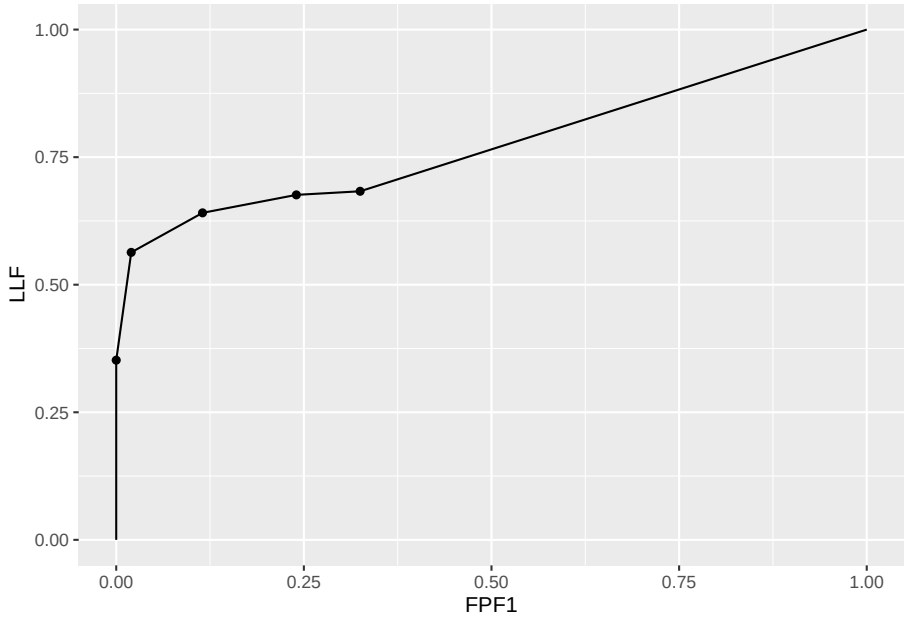
2.12.1 Empirical AFROC1 plot and AUC

The empirical AFROC1 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 AFROC1 plot and the AFROC plot is the x-axis. The area under this plot is the empirical AFROC1 AUC, denoted A_{AFROC1}.

2.12.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot for treatment 1 and reader 1.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "AFROC1")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "AFROC1")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.7744718 0.7157218 0.7229225 0.7913908
#> trt2 0.7826585 0.7278169 0.7364437 0.7897887
#> trt3 0.7412852 0.6868310 0.6946303 0.7573415
#> trt4 0.8087852 0.7346831 0.7343486 0.8155634
#> trt5 0.7580810 0.6825704 0.6643662 0.7742782
```

2.13 The weighted-AFROC1 (wAFROC1) plot

Similar to the logic for introducing the wAFROC plot as a way of giving equal importance to all diseased cases and allowing the clinical importance of lesions to be modeled by appropriate weights, we introduce a weighted version of the AFROC1, termed the wAFROC1. The ordinate of this plot is the weighted lesion localization fraction wLLF_r defined in Eqn. (2.23). The abscissa is FPF1, defined in Eqn. (2.25).

2.13.1 Empirical wAFROC1 plot and AUC

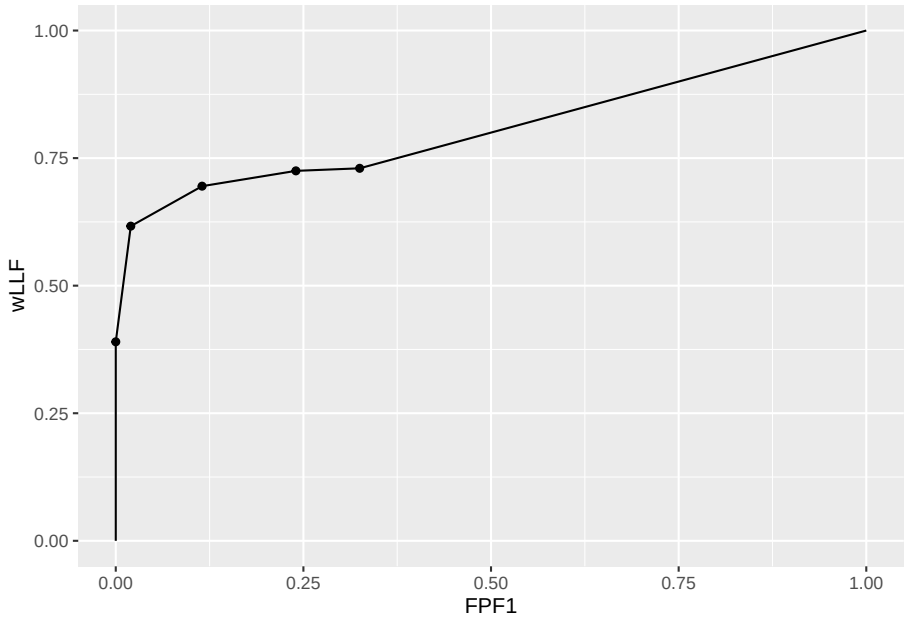
The empirical weighted-AFROC1 (wAFROC1) plot connects adjacent operating points (FPF_r^1, wLLF_r), including the origin (0,0) and

(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}.

2.13.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical wAFROC1 plot for treatment 1 and reader 1.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "wAFROC1")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "wAFROC1")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.8068333 0.7298917 0.7262042 0.8058542
#> trt2 0.8084625 0.7379917 0.7363083 0.8010167
#> trt3 0.7680875 0.7075583 0.6890208 0.7743875
#> trt4 0.8348750 0.7533917 0.7160250 0.8308333
#> trt5 0.7857708 0.6953292 0.6605167 0.7774000
```

Table 2.2: Summary of plots from FROC data. All empirical plots except FROC include a straight line extension from the uppermost observed point to (1,1).

OC	Abscissa	Ordinate	Comments
FROC	NLF	LLF	Not recommended
ROC	FPF	TPF	
AFROC	FPF	LLF	
wAFROC	FPF	wLLF	Recommended when $K_1 \approx K_2$
AFROC1	FPF1	LLF	
wAFROC1	FPF1	wLLF	Recommended when $K_1 \ll K_2$

2.14 Summary

Here is a summary of the plots defined from FROC data along with my recommendations:

2.15 Appendix 1: Proof of formula for wAFROC-AUC

The area A_{wAFROC} under the empirical wAFROC plot is obtained by summing the areas of individual trapezoids defined by dropping vertical lines from each pair of adjacent operating points to the x-axis. A sample plot is shown Fig. 2.2.

The operating point labeled i has coordinates $(\mathrm{FPF}_i, \mathrm{wLLF}_i)$ given by Eqn. (2.15) and Eqn. (2.23).

The area A_i of the leftmost shaded trapezoid in Fig. 2.2 is:

$$A_i = \frac{\left(\mathrm{FPF}_i - \mathrm{FPF}_{i+1}\right)\left(\mathrm{wLLF}_i + \mathrm{wLLF}_{i+1}\right)}{2} \tag{2.26}$$

The weighted lesion localization fraction wLLF_r corresponding to threshold ζ_r is defined by Eqn. (2.23). It follows that:

$$\left.\begin{aligned} A_i = & \frac{\left(\mathrm{FPF}_i - \mathrm{FPF}_{i+1}\right)}{2} \times \\ & \frac{1}{K_2}\left[\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_i\right)\right. \\ & \left. + \sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right)\right] \end{aligned}\right\} \tag{2.27}$$

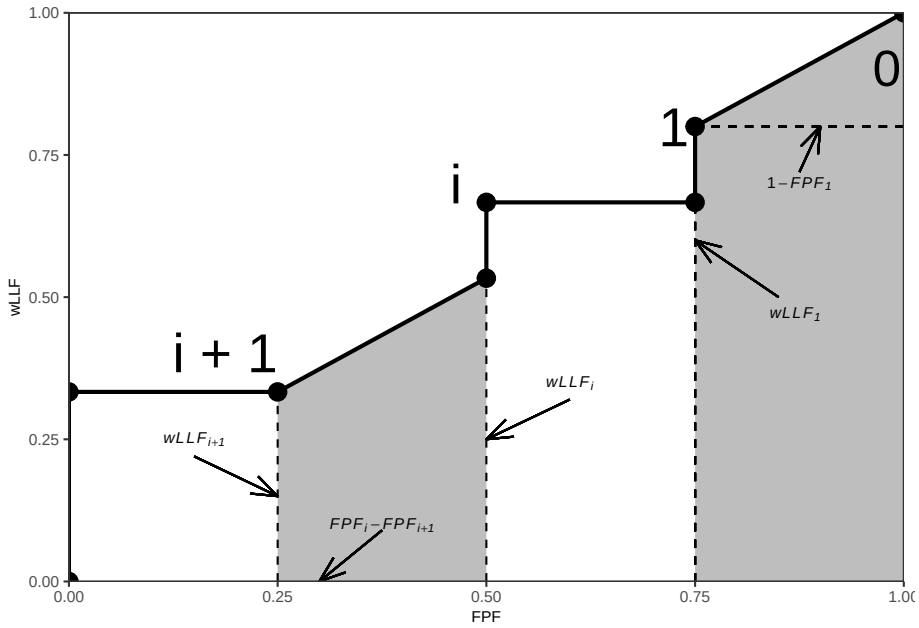


Figure 2.2: An example wAFROC plot; from left to right, the two shaded areas correspond to A_i and A_0, respectively, defined below.

Using the probabilistic relation:

$$\mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_i\right) = \mathbb{I}\left(\zeta_i \leq z_{k_2 2 l_2 2} < \zeta_{i+1}\right) + \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right) \tag{2.28}$$

we can expand the first term inside the square bracket:

$$\left.\begin{aligned} A_i = & \frac{(\mathrm{FPF}_i - \mathrm{FPF}_{i+1})}{2K_2} \times \\ & \left[\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(\zeta_i \leq z_{k_2 2 l_2 2} < \zeta_{i+1}\right)\right. \\ & + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right) \\ & \left. + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right)\right] \end{aligned}\right\} \tag{2.29}$$

The last two terms are equal, therefore:

$$\left.\begin{aligned} A_i = & \frac{(\mathrm{FPF}_i - \mathrm{FPF}_{i+1})}{K_2} \times \\ & \left[\frac{1}{2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(\zeta_i \leq z_{k_2 2 l_2 2} < \zeta_{i+1}\right)\right. \\ & \left. + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right)\right] \end{aligned}\right\} \tag{2.30}$$

The final steps of the proof require that the z-samples be converted to integer ratings, which can be done without loss of ordering information if the number of bins is sufficiently large. Let $r_{k_t t l_s s}$ denote the integer rating of mark $k_t t l_s s$, which implies that marks with z-samples satisfying $\zeta_i \leq z_{k_t t l_s s} < \zeta_{i+1}$, where $i = 0, 1, ...R$, are rated i (dummy thresholds ζ_0 and ζ_{R+1} are defined as $-\infty$ and $+\infty$, respectively).

From Eqn. (2.15) it follows that:

$$\left.\begin{aligned} \mathrm{FPF}_i - \mathrm{FPF}_{i+1} = & \frac{1}{K_1}\left[\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(z_{k_1 1 l_1 1}\right) \geq \zeta_i\right) - \sum_{k_1=1}^{K_1} \mathbb{I}\left(z_{k_1 1 l_1 1} \geq \zeta_{i+1}\right)\right] \\ = & \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(\zeta_i \leq \max_{l_1}\left(z_{k_1 1 l_1 1}\right) < \zeta_{i+1}\right) \end{aligned}\right\} \tag{2.31}$$

Because of the binning rule, $\mathbb{I}\left(\zeta_i \leq \max_{l_1}\left(z_{k_1 1 l_1 1}\right) < \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = i\right)$, $\mathbb{I}\left(\zeta_i \leq z_{k_2 2 l_2 2} < \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(r_{k_2 2 l_2 2} = i\right)$ and $\mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(r_{k_2 2 l_2 2} > i\right)$. Then Eqn. (2.27) can be re-written as:

$$\left.\begin{aligned} \mathrm{A}_i =& \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} \\ & \left[\frac{1}{2} W_{k_2 l_2} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = i\right) \mathbb{I}\left(r_{k_2 2 l_2 2} = i\right)\right. \\ & \left. + \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = i\right) \mathbb{I}\left(r_{k_2 2 l_2 2} > i\right)\right] \end{aligned}\right\} \tag{2.32}$$

Eqn. (2.32) follows from the property of the indicator function, which constrains i in the indicator functions inside the square bracket in Eqn. (17) to $\max_{l_1}\left(r_{k_1 1 l_1 1}\right)$, where the functions are unity and otherwise they are zero.

Summing over all values of i, one gets for the total area under the empirical wAFROC plot:

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} W_{k_2 l_2} (A + B) \tag{2.33}$$

where A and B are defined by:

$$\left.\begin{aligned} A =& \mathbb{I}\left(r_{k_2 2 l_2 2} = \max_{l_1}\left(r_{k_1 1 l_1 1}\right)\right) \\ B =& \mathbb{I}\left(r_{k_2 2 l_2 2} > \max_{l_1}\left(r_{k_1 1 l_1 1}\right)\right) \end{aligned}\right\} \tag{2.34}$$

Defining the Wilcoxon kernel function $\psi(x, y)$ by:

$$\left.\begin{aligned} \psi(x, y) &= 1 \quad x < y \\ \psi(x, y) &= 0.5 \quad x = y \\ \psi(x, y) &= 0 \quad x > y \end{aligned}\right\} \tag{2.35}$$

It follows that:

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} W_{k_2 l_2} \psi\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right), r_{k_2 2 l_2 2}\right) \tag{2.36}$$

This formula is the wAFROC analog of the familiar Bamber theorem (Bamber, 1975) relating the empirical AUC under the ROC to the ratings:

$$\mathrm{A}_{ROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \psi\left(r_{k_1 1}, r_{k_2 2}\right) \tag{2.37}$$

where $r_{k_1 1}$ and $r_{k_2 2}$ are the ROC ratings of non-diseased case $k_1 1$ and diseased case $k_2 2$ respectively.

2.16 Appendix 2: Interpretation of area under straight line extension of wAFROC

We prove that the contribution of the $i = 0$ term in Eqn. (2.32) is identical to the area under the extension of the wAFROC from the uppermost empirical operating point to (1,1).

According to Eqn. (2.32),

$$\left.\begin{aligned} \mathrm{A}_0 = & \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} \\ & \left[\frac{1}{2} W_{k_2 l_2} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \mathbb{I}\left(r_{k_2 2 l_2 2} = 0\right)\right. \\ & \left. + \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \mathbb{I}\left(r_{k_2 2 l_2 2} > 0\right)\right] \end{aligned}\right\} \tag{2.38}$$

Rearranging the summations:

$$\left.\begin{aligned} \mathrm{A}_0 = & \frac{1}{2} \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} W_{k_2 l_2} \mathbb{I}\left(r_{k_2 2 l_2 2} = 0\right) \\ & + \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} W_{k_2 l_2} \mathbb{I}\left(r_{k_2 2 l_2 2} > 0\right) \end{aligned}\right\} \tag{2.39}$$

Consider the term:

$$\frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \tag{2.40}$$

Because the indicator function and the summation over k_1 counts the numbers of unmarked non-diseased cases and the division by K_1 yields the corresponding contribution to FPF, the above term equals the complement of the largest observed FPF value, FPF_1, obtained by cumulating all non-zero ratings, i.e, 1 and above. It follows that:

$$\frac{1}{K_1}\sum_{k_1=1}^{K_1}\mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right)=0\right)=1-\mathrm{FPF}_1 \tag{2.41}$$

Similarly,

$$\frac{1}{K_2}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{l_{k_2}}W_{k_2 l_2}\mathbb{I}\left(r_{k_2 2 l_2 2}=0\right)=1-\mathrm{wLLF}_1 \tag{2.42}$$

Using these expressions, Eqn. (2.39) reduces to:

$$\mathrm{A}_0=\frac{\left(1-\mathrm{FPF}_1\right)\left(1+\mathrm{wLLF}_1\right)}{2} \tag{2.43}$$

The area under the straight line extension of the wAFROC from the observed end-point $(\mathrm{FPF}_1, \mathrm{wLLF}_1)$ to (1,1) equals the area of a rectangle with base $(1-\mathrm{FPF}_1)$ and height wLLF_1 plus the area of a triangle with base $(1-\mathrm{FPF}_1)$ and height $(1-\mathrm{wLLF}_1)$:

$$\left.\begin{aligned}\text{Area st. line ext.}&=\left(1-\mathrm{FPF}_1\right)\mathrm{wLLF}_1+\frac{\left(1-\mathrm{FPF}_1\right)\left(1-\mathrm{wLLF}_1\right)}{2}\\&=\left(1-\mathrm{FPF}_1\right)\left(\mathrm{wLLF}_1+\frac{\left(1-\mathrm{wLLF}_1\right)}{2}\right)\\&=\frac{\left(1-\mathrm{FPF}_1\right)\left(1+\mathrm{wLLF}_1\right)}{2}\end{aligned}\right\} \tag{2.44}$$

which equals the right hand side of Eqn. (2.43).

> In other words A_0 is the area under the extension of the wAFROC from observed end-point $(\mathrm{FPF}_1, \mathrm{wLLF}_1)$ to (1,1).

According to Eqn. (2.43), 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 performance measure.

- Failure to include the area under the straight-line extension results in not counting the full contribution to the FOM of unmarked non-diseased cases and unmarked lesions. This is best seen by considering the case of a perfect observer.
- For a perfect observer whose plot 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 $\mathrm{FPF}_1 = 0$ and $\mathrm{wLLF}_1 = 1$ and then, according to Eqn. (2.43), the area under the straight line extension is $A_0 = 1$.

2.17 Appendix 3: Summary of computational formulae

2.17.1 FROC

The formula for the area under the empirical FROC plot follows:

$$\mathrm{A}_{FROC} = \frac{1}{(K_1 + K_2)\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} (A + B) \tag{2.45}$$

where A and B are defined by:

$$\left.\begin{aligned} A &= \sum_{k_1=1}^{K_1} \sum_{l_1=1}^{N_{k_1 1}} \mathbb{I}\left(z_{k_1 1 l_1 1} \neq -\infty\right) \psi\left(z_{k_1 1 l_1 1}, z_{k_2 2 l_2 2}\right) \\ B &= \sum_{k_2'=1}^{K_2} \sum_{l_1=1}^{N_{k_2' 2}} \mathbb{I}\left(z_{z_{k_2' 2 l_1 1}} \neq -\infty\right) \psi\left(z_{k_2' 2 l_1 1}, z_{k_2 2 l_2 2}\right) \end{aligned}\right\} \tag{2.46}$$

For term A, $\mathbb{I}\left(z_{k_1 1 l_1 1} \neq -\infty\right)$ ensures that only *finite* NL z-samples on non-diseased cases enter the computation (recall that unmarked NLs are unobservable events). Likewise, for term B, $\mathbb{I}\left(z_{k_2' 2 l_1 1} \neq -\infty\right)$ ensures that only *finite* NL z-samples on diseased cases enter the computation. This is not needed for LLs since unmarked LLs are observable events. In term A the double summation compares using the ψ function all finite NL ratings on *non-diseased* cases $k_1 1$ with all lesion ratings on diseased case $k_2 2$. In term B the double summation compares all finite NL ratings on *diseased cases* $k_2' 2$ with all lesion ratings on diseased case $k_2 2$. The double summation in Eqn. (2.45) sums over all diseased

cases $k_2 2$ and all lesions in each diseased case. The final value is divided by the total number of cases and the total number of lesions.

In term B notice the need to distinguish between two indices for diseased cases $z_{k_2' 2 l_1 1}$ and $z_{k_2 2 l_2 2}$.

The above formula is equivalent to creating two arrays the first containing all finite NL ratings and the second containing all lesion ratings (including unmarked lesions). One cumulates theψ function values, using the ratings in the two arrays, and divides by the total number of cases and by the total number of lesions.

The following example uses the same 9-case FROC dataset used earlier. The AUC is calculated two ways: using geometry and using Eqn. (2.45) implemented in function `UtilFigureOfMerit`.

```
#> numerical integration yields:  0.4074074
#> RJafroc yields:  0.4074074
```

2.17.2 ROC

The ROC-AUC formula is much simpler.

$$\mathrm{A}_{ROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \psi \left(\max_{l_1} \left(z_{k_1 1 l_1 1} \right), \max_{l_1 l_2} \left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2} \right) \right) \tag{2.47}$$

The first argument of the ψ function is the maximum NL rating on a non-diseased case or $-\infty$ if the case has no NL marks. The second argument is the maximum of all marks, NL or LL, on a diseased case, or $-\infty$ if the case has no marks. The value of the ψ function is summed over all non-diseased and diseased cases and divided by K_1 and K_2, analogous to the Bamber theorem Eqn. (2.37).

2.17.3 AFROC

The formula for the area under the empirical AFROC plot follows:

$$\mathrm{A}_{AFROC} = \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \psi \left(\max_{l_1} \left(z_{k_1 1 l_1 1} \right), z_{k_2 2 l_2 2} \right) \tag{2.48}$$

The first argument of the ψ function is the maximum NL rating on a non-diseased case or $-\infty$ if the case has no NL marks. The second argument is the

LL rating on a diseased case. The value of the ψ function is summed over all non-diseased cases and all lesions and divided by K_1 and the total number of lesions.

2.17.4 wAFROC

The formula for the area under the empirical wAFROC plot follows:

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \psi \left(\max_{l_1} \left(z_{k_1 1 l_1 1} \right), z_{k_2 2 l_2 2} \right) \tag{2.49}$$

This is similar to Eqn. (2.48) except for the inclusion of the lesion weight term $W_{k_2 l_2}$ inside the summations.

The FOM-statistic A_{wAFROC} 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 unity. 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, $0 \leq \mathrm{A}_{wAFROC} \leq 1$. This shows that A_{wAFROC} behaves like a probability and its range is *twice* that of A_{ROC}; recall that $0.5 \leq \mathrm{A}_{ROC} \leq 1$ (assuming the observer has equal or better than random performance and the observer does not have the direction of the rating scale reversed). This has the consequence that treatment related differences between A_{wAFROC} (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, see sample size chapter in the `RJafrocQuickStart` book.

The range $0 \leq \mathrm{A}_{wAFROC} \leq 1$ is one reason why the "chance diagonal" of the AFROC, corresponding to $\mathrm{A}_{wAFROC} = 0.5$, does *not* reflect chance-level performance. $\mathrm{A}_{AFROC} = 0.5$ is actually reasonable performance, being exactly in the middle of the allowed range. An example of this was given above for the case of an expert radiologist who does not mark any cases.

Similar comments apply to the AFROC_AUC, i.e. $0 \leq \mathrm{A}_{AFROC} \leq 1$, etc.

2.17.5 AFROC1

$$\mathrm{A}_{AFROC1} = \frac{1}{(K_1 + K_2) \sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} (A + B) \tag{2.50}$$

where A and B are defined by:

$$\left.\begin{aligned} A &= \sum_{k_1=1}^{K_1} \psi\left(\max_{l_1}\left(z_{k_1 1 l_1 1}\right), z_{k_2 2 l_2 2}\right) \\ B &= \sum_{k_2'=1}^{K_2} \psi\left(\max_{l_1}\left(z'_{k_2 2 l_1 1}\right), z_{k_2 2 l_2 2}\right) \end{aligned}\right\} \tag{2.51}$$

The normalization can checked by assuming all NL ratings are less than any LL rating, in which case terms A and B reduce to $K_1 + K_2$ and $\mathrm{A}_{AFROC1} = 1$:

$$\left.\begin{aligned} \mathrm{A}_{AFROC1} =& \frac{1}{\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} 1 \\ =& \frac{1}{\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} L_{k_2 2} \\ =& 1 \end{aligned}\right\} \tag{2.52}$$

2.17.6 wAFROC1

This is similar to the above expression for AFROC1 except for the presence of the weight term $W_{k_2 l_2}$:

$$\mathrm{A}_{wAFROC1} = \frac{1}{(K_1 + K_2) K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} (A + B) \tag{2.53}$$

A and B are as defined in Eqn. (2.51).

2.18 References

Chapter 3

Visual Search

3.1 TBA How much finished

10%

3.2 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.

3.3 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. 3.1, 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.

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.

Figure 3.1: Grouping and labeling ROIs

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., Fig. 3.2. This is because they know the underlying anatomy, Fig. 3.3 and have a basic understanding of the x-ray image formation physics that relates the anatomy to the image.

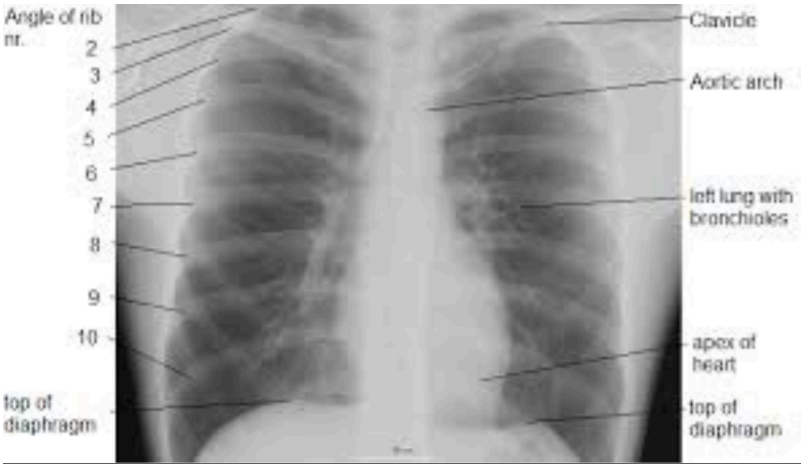


Figure 3.2: Grouping and labeling in radiology.

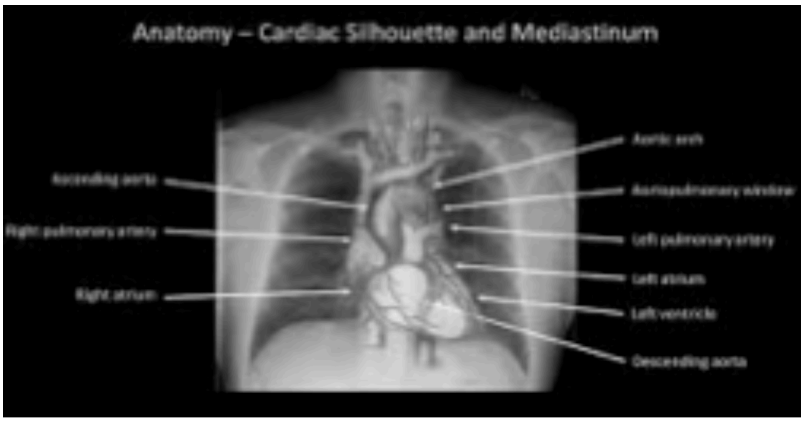


Figure 3.3: Correct grouping and labeling requires knowledge of the underlying anatomy.

3.4 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. 3.1. 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. 3.1, 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.

3.5 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.

3.6 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:

3.6.1 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).

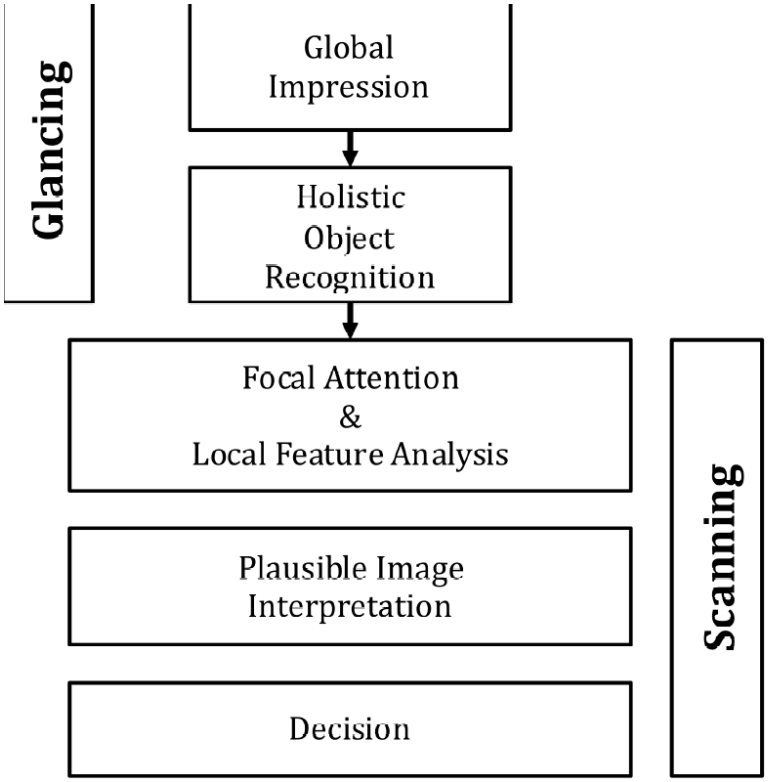


Figure 3.4: The Kundel-Nodine model of radiological search.

Perturbations are flagged for subsequent detailed viewing, in other words *the initial glance tells the visual system where to look more closely.*

3.6.2 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. 3.5, 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

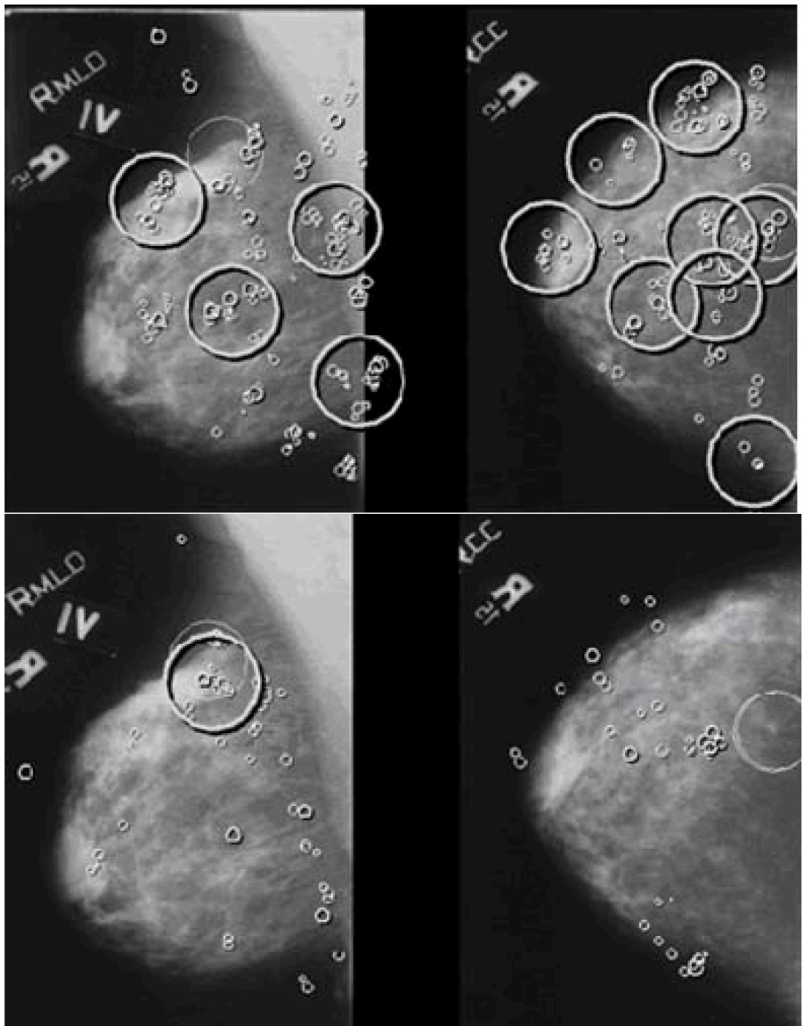


Figure 3.5: Eye-tracking recordings for a two-view digital mammogram.

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.

3.7 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.

3.8 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).

3.9 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.

The radiological search model (RSM)

Chapter 4

The radiological search model (RSM)

4.1 TBA How much finished

70%

4.2 Introduction

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.

4.3 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]

4.4 RSM assumptions

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

$$l_1 \sim \text{Poi}\,(\lambda') \tag{4.1}$$

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 2. This allows for a simpler notation, as the reader may have noticed, unencumbered by the 4 subscripts in Table 2.3.4.

$$\text{pmf}_{Poi}\left(l_1, \lambda'\right) = exp\left(-\lambda'\right)\frac{\left(\lambda'\right)^{l_1}}{\left(l_1'\right)!} \tag{4.2}$$

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:

$$l_2 \sim \text{Bin}\left(L, \nu'\right) \tag{4.3}$$

The pmf of the binomial distribution is defined by:

$$\text{pmf}_{Bin}\left(l_2, L, \nu'\right) = \binom{L}{l_2}\left(\nu'\right)^{l_2}\left(1-\nu'\right)^{L-l_2} \tag{4.4}$$

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 $N\left(0,1\right)$:

$$z_{l_1} \sim N\left(0,1\right) \tag{4.5}$$

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

$$z_{l_2} \sim N\left(\mu,1\right) \tag{4.6}$$

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

$$\phi\left(z|\mu\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(z-\mu\right)^2}{2}\right) \tag{4.7}$$

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.

4.5 Physical interpretation of RSM parameters

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

4.5.1 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:

$$\mu = \sqrt{2}\Phi^{-1}(\text{PC}) \tag{4.8}$$

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.

4.5.2 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').

4.5.2.1 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.

4.5.2.2 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 1st observer is 1.01 and that for the 2nd one is 2.02.

4.5.2.3 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.

4.5.3 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).

4.6 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]:

$$\left.\begin{aligned}\nu' =& 1 - exp\left(-\mu\nu\right)\\ \lambda' =& \frac{\lambda}{\mu}\end{aligned}\right\} \tag{4.9}$$

The inverse transformations are:

$$\left.\begin{aligned}\nu =& -\frac{\ln\left(1-\nu'\right)}{\mu}\\ \lambda =& \mu\lambda'\end{aligned}\right\} \tag{4.10}$$

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:

$$\nu = \left(\frac{\partial\nu'}{\partial\mu}\right)_{\mu=0} \tag{4.11}$$

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. (4.9), 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. (4.9) 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.

4.7 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.

4.8 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 5

ROC curve implications of the RSM

5.1 TBA How much finished

90%

5.2 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. (4.10).

5.3 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.

5.4 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.

5.4.1 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. (4.2), 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:

$$\mathrm{FPF}_{max} = 1 - \exp\left(-\lambda'\right) \tag{5.1}$$

5.4.2 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. (4.4), $\mathrm{pmf}_{Bin}\left(0, L, \nu'\right) = (1-\nu')^L$.

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.,

$$\mathrm{TPF}_{max}\left(\mu, \lambda', \nu', L\right) = 1 - \exp\left(-\lambda'\right)\left(1-\nu'\right)^L \tag{5.2}$$

5.4.3 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:

$$\sum_{L=1}^{L_{max}} f_L = 1 \tag{5.3}$$

By restricting attention to the set of diseased cases with L lesions each, Eqn. (5.2) for TPF_{max} applies. Since TPF is a probability and probabilities of independent processes add it follows that:

$$\text{TPF}_{max}(\mu, \lambda', \nu', L) = 1 - \sum_{L=1}^{L_{max}} f_L \exp(-\lambda')(1-\nu')^L \tag{5.4}$$

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

The expression for FPF_{max} is unaffected.

5.5 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:

$$\left.\begin{aligned}\text{FPF}(\zeta) &= P\left(h_{k_1 1} \geq \zeta\right) \\ \text{TPF}(\zeta) &= P\left(h_{k_2 2} \geq \zeta\right)\end{aligned}\right\} \tag{5.5}$$

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

5.5.1 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 4.4 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:

$$\left.\begin{aligned}\text{FPF}(\zeta \mid n) &= P\left(h_{k_1 1} \geq \zeta \mid n\right) \\ &= 1 - [\Phi(\zeta)]^n\end{aligned}\right\} \tag{5.6}$$

The conditioning notation in Eqn. (5.6) 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):

$$\text{FPF}\,(\zeta, \lambda') = \sum_{n=0}^{\infty} \text{pmf}_{Poi}\,(n, \lambda')\,\text{FPF}\,(\zeta \mid n) \tag{5.7}$$

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:

$$\text{FPF}\,(\zeta, \lambda') = 1 - \exp\left(-\frac{\lambda'}{2}\left[1 - \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \tag{5.8}$$

The error function in Eqn. (5.8) is related to the unit normal CDF function $\Phi(x)$ by:

$$\text{erf}\,(x) = 2\Phi\left(\sqrt{2}x\right) - 1 \tag{5.9}$$

Using this transformation yields the following simpler expression for FPF:

$$\text{FPF}\,(\zeta, \lambda') = 1 - \exp\left(-\lambda'\Phi\left(-\zeta\right)\right) \tag{5.10}$$

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, $\mathrm{FPF}(\zeta, \lambda')$ ranges from 0 to $\exp(-\lambda')$.

5.5.2 Derivation of TPF

The derivation of the true positive fraction $\mathrm{TPF}(\zeta)$ 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.,

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

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

$$\left.\begin{array}{l} \mathrm{TPF}(\zeta,\mu,\lambda',\nu') \times \\ = \sum_{n=0}^{\infty} \mathrm{pmf}_{Poi}(n,\lambda') \times \\ \sum_{l=0}^{L} \mathrm{pmf}_{Bin}(l,\nu',L) \mathrm{TPF}_{n,l}(\zeta,\mu,n,l) \end{array}\right\} \tag{5.11}$$

This can be evaluated using Maple yielding:

$$\left.\begin{array}{l} \mathrm{TPF}(\zeta,\mu,\lambda',\nu',L) \\ = 1 - \exp(-\lambda'\Phi(-\zeta))\left(1-\nu'\Phi(\mu-\zeta)\right)^L \end{array}\right\} \tag{5.12}$$

5.5.3 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 (5.12) over the fraction of diseased cases with L lesions:

$$\left.\begin{array}{l}\mathrm{TPF}\left(\zeta,\mu,\lambda',\nu',\overrightarrow{f_L}\right)= \\ 1-\exp\left(-\lambda'\Phi\left(-\zeta\right)\right)\sum_{L=1}^{L_{max}} f_L\left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^L\end{array}\right\} \tag{5.13}$$

Since $\Phi\left(-\zeta\right)$ tends to unity as $\zeta_1 \to -\infty$, this expression reduces to Eqn. (5.4) for the ROC end-point.

The expression for FPF, Eqn. (5.10), is unaffected.

The software implementation follows:

```
# lambdaP is the physical lambda' parameter
# nuP is the physical nu' parameter
# lesDistr is the lesion distribution 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)
}
```

Two first principle calculations are shown next and compared to the values yielded by the function `RSM_yROC`:

```
zeta_1 <- 1
mu <- 2
lambdaP <- 1
nuP <- 0.9
lesDistr <- c(0.5,0.5)

1-exp(-lambdaP*pnorm(-zeta_1))*(lesDistr[1]*(1-nuP*pnorm(mu-zeta_1))+lesDistr[2]*(1-nu
```

```
## [1] 0.8712655
```

```
RSM_yROC(zeta_1,mu,lambdaP,nuP, lesDistr = lesDistr)
```

```
## [1] 0.8712655
```

```
cat("\n")
```

```
zeta_1 <- 0
mu <- 1
lambdaP <- 2
nuP <- 0.5
lesDistr <- c(0.1,0.9)
1-exp(-lambdaP*pnorm(-zeta_1))*(lesDistr[1]*(1-nuP*pnorm(mu-zeta_1))+lesDistr[2]*(1-nuP*pnorm(mu-
```

```
## [1] 0.8675666
```

```
RSM_yROC(zeta_1,mu,lambdaP,nuP, lesDistr = lesDistr)
```

```
## [1] 0.8675666
```

5.6 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 5.13 a proof is presented that the slope is continuous at the end-point transition from a continuous curve to the dashed straight line. In 5.14 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. (5.10) and Eqn. (5.13) one can express the ROC coordinates as:

$$\left.\begin{aligned} x(\zeta) &= 1 - G(\zeta) \\ y(\zeta) &= 1 - F(\zeta)\, G(\zeta) \end{aligned}\right\} \tag{5.14}$$

where:

$$\left.\begin{aligned} G(\zeta) &= \exp\left(-\lambda' \Phi(-\zeta)\right) \\ F(\zeta) &= \sum_{L=1}^{L_{max}} f_L \left(1 - \nu' \Phi(\mu - \zeta)\right)^L \end{aligned}\right\} \tag{5.15}$$

[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".

5.7 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:

$$\text{pdf}_N(\zeta) = \frac{\partial}{\partial \zeta}\left(1 - \text{FPF}(\zeta, \lambda')\right) \tag{5.16}$$

Similarly, for diseased cases,

$$\text{pdf}_D(\zeta) = \frac{\partial}{\partial \zeta}\left(1 - \text{TPF}\left(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}\right)\right) \tag{5.17}$$

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):

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_N(\zeta)\, d\zeta &= \left(1 - \text{FPF}(\zeta, \lambda')\right)\Big|_{-\infty}^{\infty} \\ &= \text{FPF}_{max}\end{aligned}\right\} \tag{5.18}$$

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_D(\zeta)\, d\zeta &= \left(1 - \text{TPF}\left(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}\right)\right)\Big|_{-\infty}^{\infty} \\ &= \text{TPF}_{max}\end{aligned}\right\} \tag{5.19}$$

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]

5.8 ROC AUC

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

$$AUC_{RSM}^{ROC}\left(\mu, \lambda, \nu, \zeta_1, \overrightarrow{f_L}\right) = \sum_{L=0}^{L_{max}} f_L \int_0^1 \text{TPF}\left(\zeta, \mu, \lambda, \nu, L\right) d\left(\text{FPF}\left(\zeta, \lambda\right)\right) \tag{5.20}$$

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. (5.20) 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.

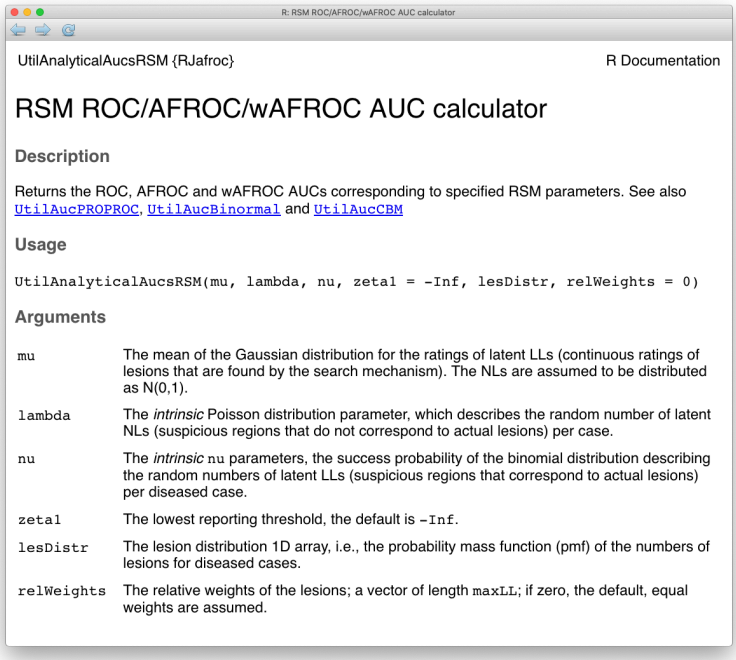


Figure 5.1: Help page for ‘RJafroc‘ function ‘UtilAnalyticalAucsRSM‘.

```
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 =  0.8817798
```

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.

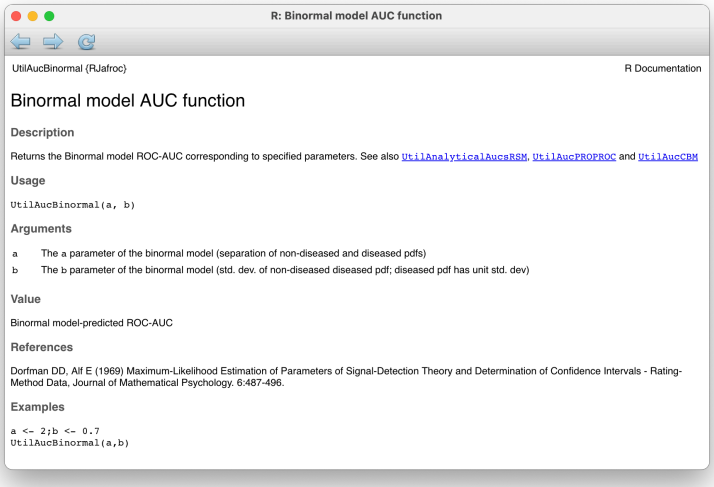
- AUC decreases as ζ_1 increases. This important effect is discussed in the next section.

- 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).

5.9 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

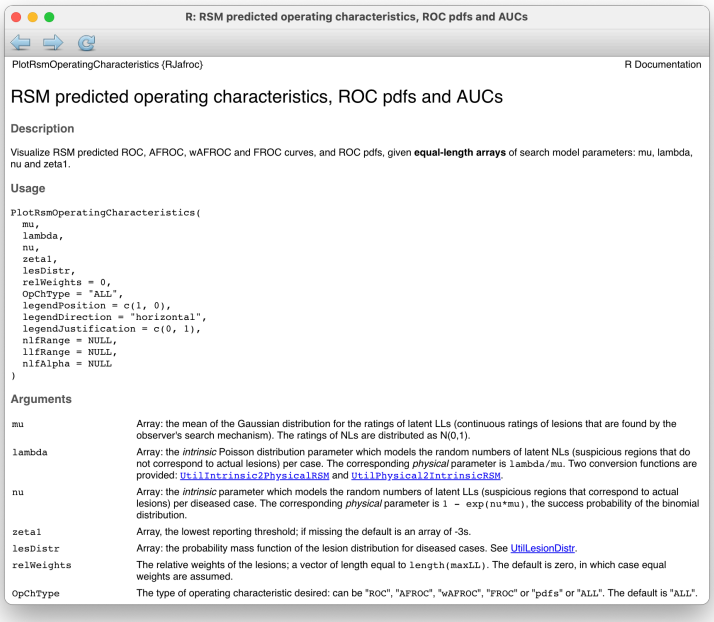
```
a <- 2;b <- 0.7
UtilAucBinormal(a,b)
```

Figure 5.2: 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),
     nu = c(1,1),
     zeta1 = c(-10, 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 5.3: Help page for 'RJafroc' function 'PlotRsmOperatingCharacteristics'.

```
    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.

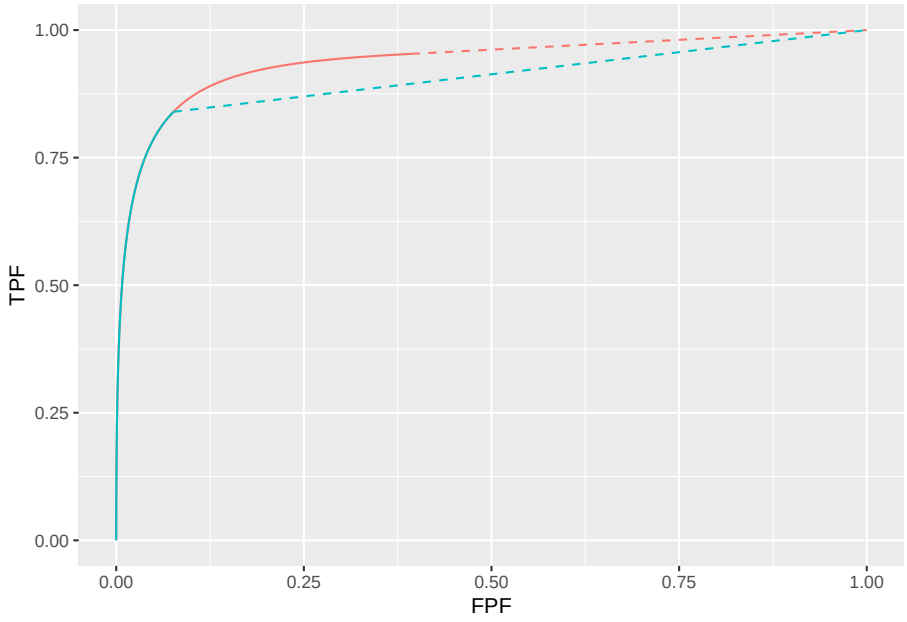


Figure 5.4: ROC curves for two values of ζ_1: both curves correspond to $\mu = 2$, $\nu = 1$ and $\lambda = 1$. The red curve corresponds to $\zeta_1 = -10$ and the blue curve to $\zeta_1 = 1$.

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, which is clearly incorrect and is not adopted by any CAD designer. Selecting the optimal value of the reporting threshold is addressed in Chapter 10.

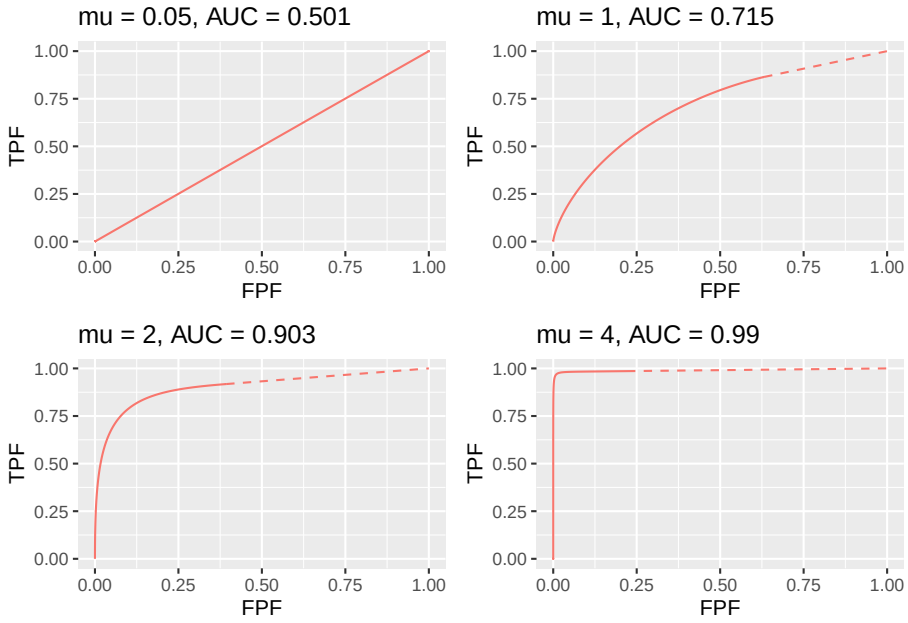


Figure 5.5: ROC curves for indicated values of the μ parameter. Notice the transition, as μ increases, from near chance level performance to almost perfect performancea as the end-point moves from near (1,1) to near (0,1).

5.10 Example ROC curves

Fig. 5.5 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. (5.1)
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.

5.11 Example RSM pdf curves

Fig. 5.6 shows pdf plots for the same values of parameters as in Fig. 5.5.

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

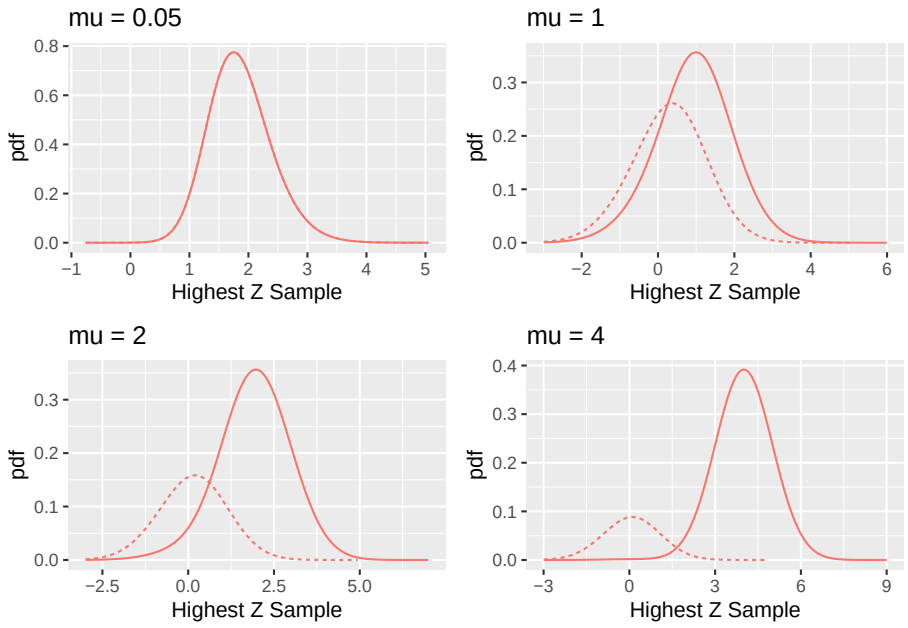


Figure 5.6: RSM pdf curves for indicated values of the μ parameter. The solid curve corresponds to diseased cases and the dotted curve corresponds to non-diseased cases.

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. (5.2), 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. 5.6 labeled $\mu = 0.05$ may be surprising. Since it corresponds 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).

5.12 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.

5.12.1 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.

5.13 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. (5.1) for FPF_{max} and Eqn. (5.2) for TPF_{max}. Therefore, the slope m_{st} of the dashed dashed straight line is:

$$\left.\begin{aligned} m_{st} =& \frac{1-\mathrm{TPF}_{max}}{1-\mathrm{FPF}_{max}} \\ =& \frac{\sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^L \exp\left(-\lambda'\right)}{\exp\left(-\lambda'\right)} \\ =& \sum_{L=1}^{L_{max}} f_L \left(1-\nu'\right)^L \end{aligned}\right\} \tag{5.21}$$

On the continuous section, $g \equiv \mathrm{FPF}$ and $h \equiv \mathrm{TPF}$ are defined by (5.10) and (5.13), respectively. Therefore,

$$\left.\begin{aligned} g =& 1-\exp\left(-\lambda'\Phi\left(-\zeta\right)\right) \\ h =& 1-\sum_{L=1}^{L_{max}} f_L \exp\left(-\lambda'\Phi\left(-\zeta\right)\right)\left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^L \end{aligned}\right\} \tag{5.22}$$

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

$$\left.\begin{aligned} \frac{dh}{dg} = & \sum_{L=1}^{L_{max}} f_L \left(1-\nu'\Phi\left(\mu-\zeta\right)\right)^{L-1} \times \\ & \left[\frac{L\nu'\phi\left(\mu-\zeta\right)}{\lambda'\phi\left(-\zeta\right)} + \left(1-\nu'\Phi\left(\mu-\zeta\right)\right)\right] \end{aligned}\right\} \tag{5.23}$$

Using the following result:

$$\left.\begin{aligned}&\lim_{\zeta\to-\infty}\frac{\phi(\mu-\zeta)}{\phi(-\zeta)}\\ =&\lim_{\zeta\to-\infty}\frac{\frac{1}{\sqrt{2\pi}}\exp\left(\frac{(\mu-\zeta)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\zeta^2}{2}\right)}\\ =&\lim_{\zeta\to-\infty}\exp\left(\frac{\mu\zeta-\mu^2}{2}\right)\\ =&0\end{aligned}\right\} \tag{5.24}$$

it follows that:

$$\left.\begin{aligned}&\lim_{\zeta\to-\infty}\frac{dh}{dg}\\ =&\sum_{L=1}^{L_{max}}f_L\left(1-\nu'\right)^{L-1}\left(1-\nu'\right)\\ =&\sum_{L=1}^{L_{max}}f_L\left(1-\nu'\right)^{L}\\ =&m_{st}\end{aligned}\right\} \tag{5.25}$$

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).

5.14 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
)
```

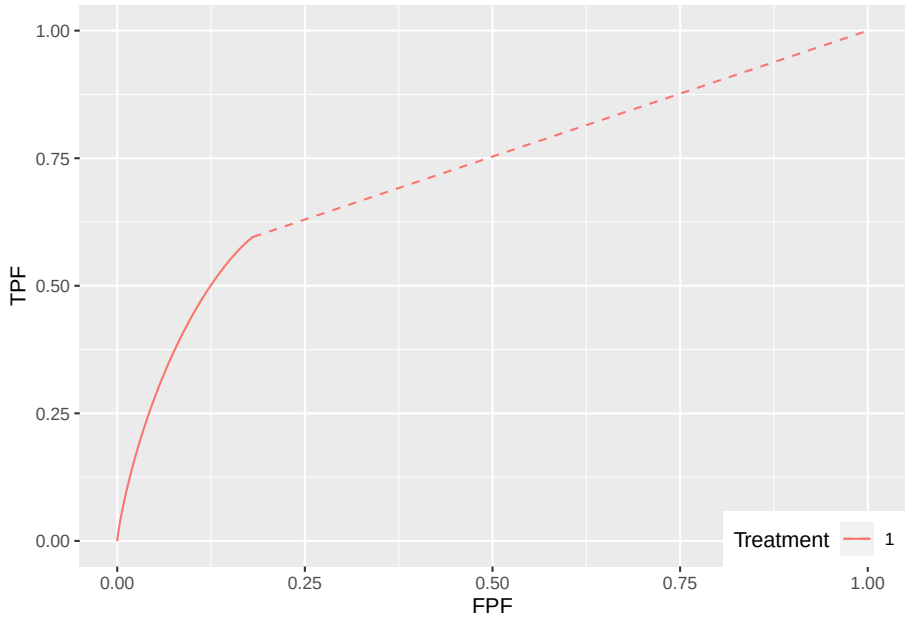


Figure 5.7: ROC curve for selected RSM parameters. The slope of the dashed line is 0.4935272.

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,

[3] Others have stated a different visual impression.

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.

```
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))
```

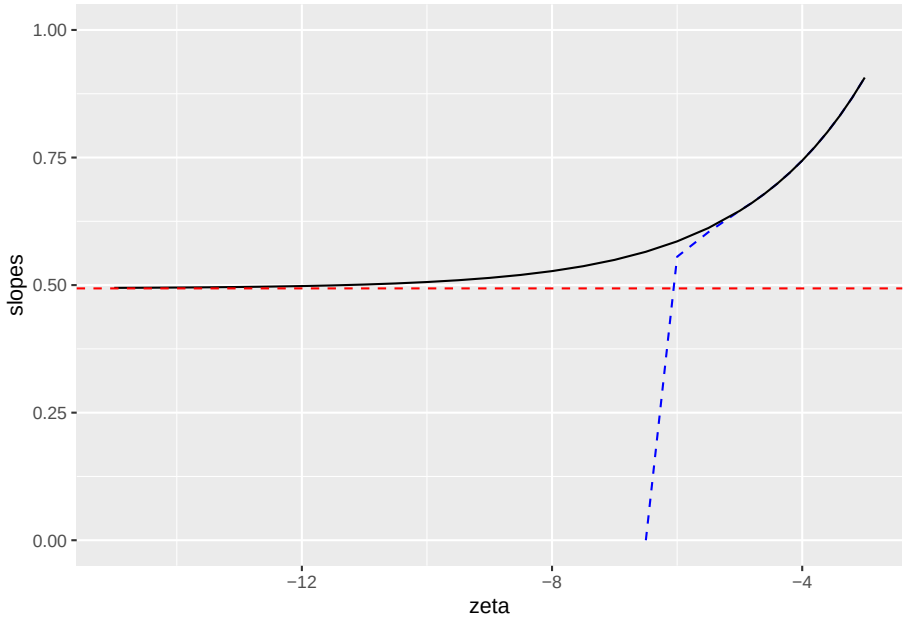


Figure 5.8: Horizontal dashed red line: the value of `mStLine`, the slope of the straight line connecting the ROC end-point to (1,1). Dashed blue line: slope using double precision arithmetic. Solid black line: slope using multiple precision arithmetic - this curve approaches the limiting value `mStLine`.

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 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.

5.15 Appendix 3: wAFROC curve

The wAFROC abscissa is identical to the ROC abscissa, i.e., Eqn. (5.10) applies.

The wAFROC ordinate is calculated using:

$$\text{wLLF}\left(\mu, \lambda, \nu, \overrightarrow{f_L}, \mathbf{W}\right) = \Phi\left(\mu - \zeta_1\right) \sum_{L=1}^{L_{max}} f_L \sum_{l=1}^{L} \mathrm{W}_{Ll}\, l\, \text{pmf}_{Bin}\left(l, L, \nu\right) \quad (5.26)$$

- $\overrightarrow{f_L}$ is the normalized histogram of the lesion distribution for the diseased cases. In the software it is denoted `lesDistr`. For example, the array `lesDistr = c(0.1, 0.4, 0.4, 0.1)` defines a dataset in which 10 percent of the cases contain one lesion, 40 percent contain 2 lesions, 40 percent contain 3 lesions and 10 percent contain 4 lesions.
- L_{max} is the maximum number of lesions per case in the dataset. In the preceding example $L_{max} = 4$.
- $\mathbf{W}$ is the (lower triangular) square matrix with L_{max} rows and columns containing the weights, where each row sums to unity. The relative lesion weights are denoted in the code `relWeights`. For example, `relWeights = c(0.2, 0.3, 0.1, 0.5)` whose meaning is as follows:
  - On cases with one lesion the lesion weight is unity.
  - On cases with two lesions the relative weights are 0.2 and 0.3. Since these do not add up to unity, the actual weights are 0.4 and 0.6.
  - On cases with three lesions the relative weights are 0.2, 0.3 and 0.1. The actual weights are 1/3, 1/2 and 1/6.
  - On cases with four lesions the relative weights are 0.2, 0.3, 0.1 and 0.5. The actual weights are 0.1818182, 0.2727273, 0.09090909, 0.4545455.
- The function `UtilLesionWeightsMatrixLesDistr` calculates the matrix given `lesDistr` and `relWeights`. For example:

```
lesDistr <- c(0.6, 0.2, 0.1, 0.1)
relWeights =  c(0.2, 0.3, 0.1, 0.4)
UtilLesionWeightsMatrixLesDistr(lesDistr, relWeights)[,-1]
```

```
##           [,1] [,2]      [,3] [,4]
## [1,] 1.0000000 -Inf      -Inf -Inf
## [2,] 0.4000000  0.6      -Inf -Inf
## [3,] 0.3333333  0.5 0.1666667 -Inf
## [4,] 0.2000000  0.3 0.1000000  0.4
```

- It is necessary to label the lesions properly so that the correct weights are used. This is done using the `lesionID` field in the Excel input file. For example, `lesionID = 3` for the one with relative weight 0.1. Since **W** is independent of cases, the lesion characteristics (which determine clinical outcome) of `lesionID = 1` on cases with one lesion or on cases with 4 lesions are assumed to be identical. In other words this example assumes that the lesions fall into one of 4 groups, with clinical outcomes as in the weights matrix **W**.

- $\text{pmf}_{Bin}(l, L, \nu)$ is the probability mass function (pmf) of the binomial distribution with success probability ν and trial size L, defined in Eqn. (4.4). W_{Ll} is the weight of lesion l in cases with L lesions; for example $\text{W}_{42} = 0.2727273$.

- The wAFROC-AUC is obtained by numerically integrating the wAFROC curve defined by Eqn. (5.10) and Eqn. (5.26) which is implemented in `UtilAnalyticalAucsRSM`.

- To generate equal weights set `relWeights = 0`, as in following code:

```
lesDistr <- c(0.6, 0.2, 0.1, 0.1)
UtilLesionWeightsMatrixLesDistr(lesDistr, relWeights = 0)[,-1]
```

```
##           [,1]      [,2]      [,3] [,4]
## [1,] 1.0000000      -Inf      -Inf -Inf
## [2,] 0.5000000 0.5000000      -Inf -Inf
## [3,] 0.3333333 0.3333333 0.3333333 -Inf
## [4,] 0.2500000 0.2500000 0.2500000 0.25
```

5.16 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 6

Search and classification performances

6.1 TBA How much finished

10%

6.2 TBA 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.

6.3 Location of ROC end-point

From the previous chapter the coordinates of the end-point are given by:

$$\left.\begin{aligned} &\mathrm{FPF}_{max} = 1 - exp\left(\lambda'\right) \\ &\mathrm{TPF}_{max}\left(\mu, \lambda', \nu', L\right) = 1 - \sum_{L=1}^{L_{max}} f_L \exp\left(-\lambda'\right)\left(1-\nu'\right)^L \end{aligned}\right\} \tag{6.1}$$

6.4 Quantifying search performance

Qualitatively, search performance is the ability to find lesions while not finding non-lesions. To arrive at a quantitative definition of search performance consider the location of the ROC end-point.

In Fig. 6.1, plot (a) is a typical ROC curve predicted by models that do not account for search. 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), the 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), measures 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.

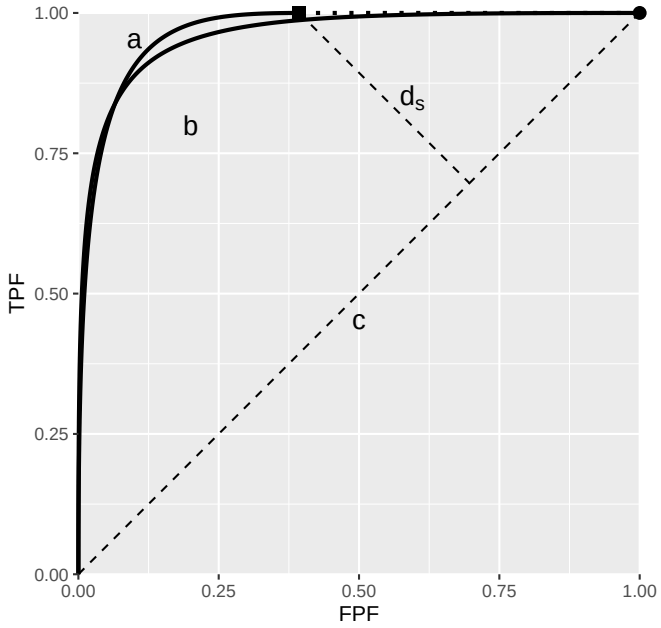


Figure 6.1: Relation of search performance to the end-point of the ROC curve. Plot (a) is for conventional ROC models while plot (b) is for the RSM.

Definition: The perpendicular distance, d_S, from the end-point to the chance diagonal, plot (c), multiplied by $\sqrt{2}$, is a quantitative measure of search performance S.

Using geometry and Eqn. (6.1), it follows that:

$$S = \sqrt{2} d_S = \mathrm{TPF}_{max} - \mathrm{FPF}_{max} \tag{6.2}$$

Therefore, search performance S is given by:

$$S = \exp(-\lambda') \left(1 - \sum_{L=1}^{L_{max}} f_L \left(1 - \nu'\right)^L \right) \tag{6.3}$$

Eqn. (6.3) shows search performance is the product of two terms: the probability $\left(1 - \sum_{L=1}^{L_{max}} f_L \left(1 - \nu'\right)^L\right)$ of finding at least one lesion times the probability $\exp(-\lambda')$ of not finding non-lesions. This puts into mathematical form the qualitative definition of search performance as the ability to find lesions while avoiding finding non-lesions.

Example: consider $\lambda' = 0$ 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 $S = 1$. The same value is obtained using Eqn. (6.3). Since no NLs are found and all lesions are found, the observer never makes a mistake. One cannot improve over perfect performance: the observer simply marks all suspiciuos regions found by search regardless of their z-samples.

6.5 Quantifying lesion-classification performance

Lesion-classification performance C measures the ability, having found a suspicious region, to correctly classify it as a lesion, i.e., mark the location of the lesion resulting in a LL event. It is distinct from *case-classification* performance, ROC AUC, which measures 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. Lesion-classification performance C is determined by the μ parameter of the RSM and is defined by the implied ROC-area of two unit variance normal distributions separated by μ.

$$C = \Phi\left(\frac{\mu}{\sqrt{2}}\right) \tag{6.4}$$

C ranges from 0.5 to 1.

6.5.1 TBA Lesion-classification performance and the 2AFC LKE task

It should be obvious that lesion-classification performance is similar to what is measured using the location-known-exactly (LKE) paradigm. In this paradigm, one uses 2AFC methods as in TBA 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 TBA 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.

6.5.2 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 the observer's performance.

In contrast, the search and lesion-classification measures described above can

be used as an optimization-aid in determining what is limiting performance. If search performance S is poor it indicates that the observer needs to be trained on more *non-diseased* cases to learn the variants of non-diseased anatomy so as 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. In breast CAD since the designer level ROC curve goes almost all the way to (1,1) implying poor search performance. Therefore more research is needed on improving CAD's search performance. In contrast lCAD's esion-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.

To realize these benefits one needs a way of estimating the ROC end-point shown. TBA Chapter 19 describes RSM based curve fitting which determines all parameters of the RSM, thereby determining the location of the end-point TBA.

6.6 Discussion / Summary

TBA 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.

TBA 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.

6.6.1 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.

6.7 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 7

RSM fitting

7.1 TBA How much finished

10%

7.2 TBA 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.

7.3 ROC Likelihood function

In Chapter 5 expressions were derived for the coordinates (x,y) of the ROC curve predicted by the RSM, see Eqn. (5.10) and Eqn. (5.13).

$$x \equiv \text{FPF}\left(\zeta, \lambda'\right) = 1 - \exp\left(-\lambda' \Phi\left(-\zeta\right)\right) \tag{7.1}$$

$$\left.\begin{aligned} y \equiv & \text{TPF}\left(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}\right) = \\ & 1 - \exp\left(-\lambda' \Phi\left(-\zeta\right)\right) \sum_{L=1}^{L_{max}} f_L \left(1 - \nu' \Phi\left(\mu - \zeta\right)\right)^L \end{aligned}\right\} \tag{7.2}$$

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 [2]. 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 [3] :

$$P_{1r} = x\left(\zeta_r\right) - x\left(\zeta_{r+1}\right) \tag{7.3}$$

Likewise, the probability P_{2r} of a count in diseased ROC bin r is:

$$P_{2r} = y\left(\zeta_r\right) - y\left(\zeta_{r+1}\right) \tag{7.4}$$

Ignoring combinatorial factors that do not depend on parameters the likelihood function is:

[2] 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}.

[3] 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.

$$(P_{1r})^{F_r} (P_{2r})^{T_r}$$

The log-likelihood function is:

$$LL_{ROC}\left(\mu, \lambda', \nu', \overrightarrow{f_L}\right) = \sum_{r=0}^{R_{FROC}} \left[F_r log\left(P_{1r}\right) + T_r log\left(P_{2r}\right)\right] \tag{7.5}$$

The total number of parameters to be estimated, including the R_{FROC} thresholds, is $3 + R_{FROC}$. Maximizing the likelihood function yields estimates of the RSM parameters.

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 R-package `bbmle` (Bolker and R Development Core Team, 2022) 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 software also calculates the goodness of fit statistic using the method described in `RJafrocRocBook`. Because of the additional RSM parameter (as compared to conventional ROC models) the degrees-of-freedom (df) of the chisquare goodness of fit statistic is R_{FROC}–3. Calculating goodness of fit for the RSM can fail in situations where the corresponding statistic can be calculated for the binormal model, e.g., three (non – trivial) ROC operating points, corresponding to df = 1. With RSM fitting 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. [4]

7.4 `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

[4] 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 using an ROC model. However, because of the extra RSM parameter, the corresponding RSM df = 0, and the chi-square statistic cannot be calculated.

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

7.5 FitRsmROC usage example

- The following example uses the *first* treatment in `dataset04`; this is a 5 treatment 4 radiologist FROC dataset (Zanca et al., 2009) consisting of 200 cases acquired on a 5-point integer scale, i.e., it is already binned. If not one needs to bin the dataset using `DfBinDataset()`. The number of parameters to be estimated increases with the number of bins: for each additional bin one needs to estimate an additional cutoff parameter.

```
rocData <- DfFroc2Roc(dataset04)
lesDistr <- UtilLesionDistrVector(dataset04)
ret <- FitRsmRoc(rocData, lesDistr = lesDistr)
```

The lesion distribution vector is 0.69, 0.2, 0.11. This means that fraction 0.69 of each diseased case contains one lesion, fraction 0.2 contains two lesions and

fraction 0.11 contains three lesions. Since the fitted curve depends on the lesion distribution the fitting function needs to know this distribution. [5]

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):

- $\mu = 3.658$
- $\lambda' = 9.935$
- $\nu' = 0.796$
- $\zeta_1 = 1.504$
- AUC = 0.9064
- σ(AUC) = 0.023
- NLLIni = 281.4
- NLLFin = 267.27

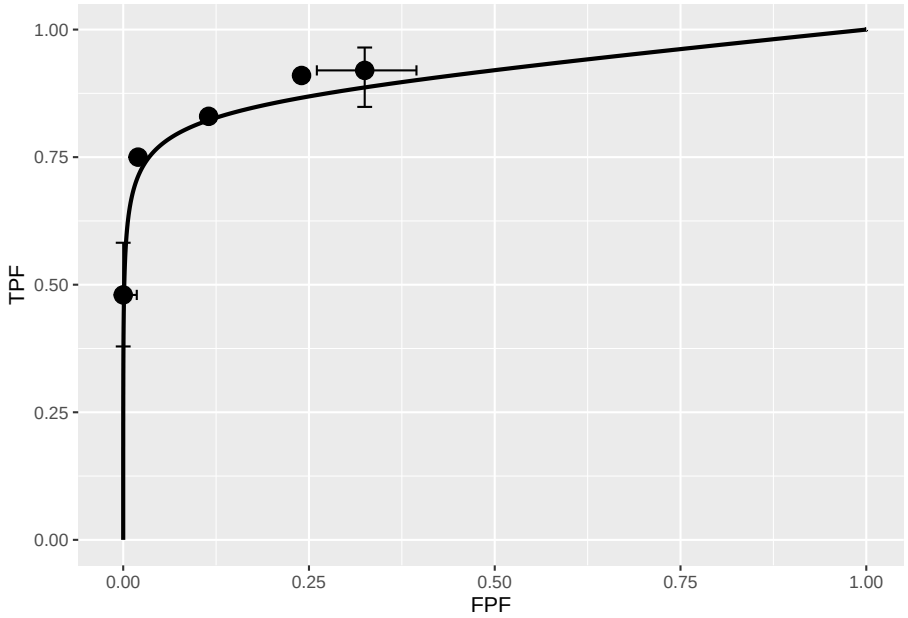
The relatively large separation parameter μ implies good lesion-classification performance. The large λ' parameter implies poor search performance. On the average the observer generates 9.94 latent NL marks per image. However, because of the relatively large value of ζ_1, i.e., 1.5, only fraction 0.066 of these are actually marked, resulting in 0.66 actual marks per image. Search performance depends on the numbers of latent marks, i.e., λ' and ν', not the actual numbers of marks.

The fitting program decreased the negative of the log-likelihood function from 281.4 to 267.27. 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 conventionql 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 shown for the lowest and highest operating points.

[5] For example, all else being equal, if each diseased case contains one lesion the ROC curve will be lower than if each diseased case contains three lesions.



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. TBA The area under the proper ROC is 0.906 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.

7.6 TBA 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.

7.7 Appendix 1: 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:

$$\left.\begin{aligned} p_r =& \Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right) \\ q_r =& \Phi\left(\zeta_{r+1} - \mu\right) - \Phi\left(\zeta_r - \mu\right) \end{aligned}\right\} \tag{7.6}$$

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.

7.7.1 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):

$$P\left(\vec{n} \mid n, \vec{\zeta}\right) = n! \prod_{r=0}^{R_{FROC}} \frac{p_r^{n_r}}{n_r!} \tag{7.7}$$

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:

$$P\left(\vec{n} \mid \lambda', \vec{\zeta}\right) = \text{pmf}_{\text{Poi}}\left(n, K\lambda'\right) P\left(\vec{n} \mid n, \vec{\zeta}\right) \tag{7.8}$$

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. (7.8) 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.

7.7.2 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:

$$P\left(\vec{l} \mid l, \mu, \vec{\zeta}\right) = l! \prod_{r=0}^{R_{FROC}} \frac{q_r^{l_r}}{l_r!} \tag{7.9}$$

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

$$P\left(\vec{l} \mid l, \mu, \nu', \vec{\zeta}\right) = \sum_{l=L}^{L_{tot}} \text{pmf}_{\text{Bin}}\left(l, L_T, \nu'\right) P\left(\vec{l} \mid l, \mu, \vec{\zeta}\right) \tag{7.10}$$

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):

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

(7.11)

7.7.3 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.

7.8 Appendix 2: 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:

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

(7.12)

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 `\@ref(optim-op-point)` for my solution to this problem, the location of each candidate lesion satisfying the criterion is shown to the radiologist, Fig. 7.1.

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:

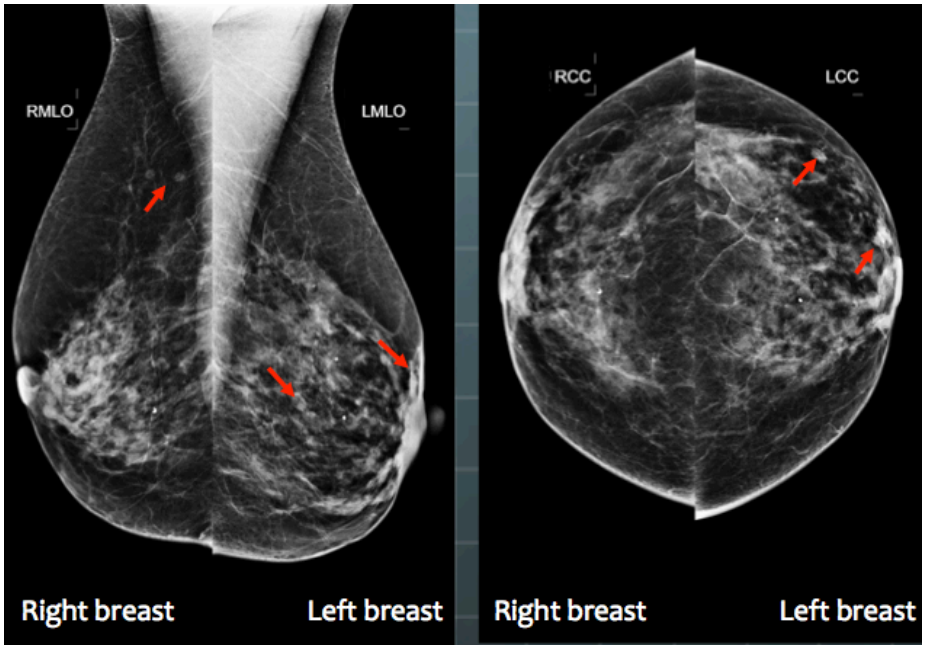


Figure 7.1: A typical 4-view display of a patient mammogram with the CAD cues (the red arrows) turned on.

$$\left.\begin{array}{l}\lambda' = NLF_{max} \\ \nu' = LLF_{max}\end{array}\right\} \tag{7.13}$$

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. (7.12), reduces to:

$$LL_{FROC}^{IDCA} = \sum_{r=1}^{R_{FROC}} \{n_r log(p_r) + l_r log(q_r)\} + ... \tag{7.14}$$

Since the ignored terms in Eqn. (7.14) 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. (7.6).

Eqn. (7.14) 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 [6]:

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} log((\Phi(\zeta_{r+1}) - \Phi(\zeta_r))) + K_{2r} log((\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)))\} \tag{7.15}$$

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. (7.6) as follows:

$$\left.\begin{array}{l}p'_r = \Phi(\zeta_{r+1}) - \Phi(\zeta_r) \\ q'_r = \Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)\end{array}\right\} \tag{7.16}$$

Here (a, b) are the parameters the unequal variance binormal model. Then Eqn. (7.15) becomes,

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} log(p'_r) + K_{2r} log(q'_r)\} \tag{7.17}$$

- With the identifications $K_{1r} \rightarrow n_r$ and $K_{2r} \rightarrow l_r$, Eqn. (7.15) looks exactly like Eqn. (7.14). This implies that binormal ROC fitting method can be used to determine a and b. Notice that instead of fitting an equal

[6] 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. 7.2.

Fig. 7.2: 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.

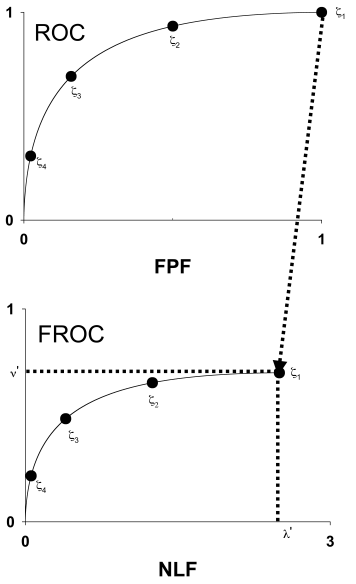


Figure 7.2: The IDCA method of fitting designer-level CAD FROC data.

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. 7.2 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):

$$\left.\begin{aligned} NLF(\zeta) =& \lambda' \Phi(-\zeta) \\ LLF(\zeta) =& \nu' \Phi(a - b\zeta) \end{aligned}\right\} \tag{7.18}$$

The RSM predicted FROC curve is repeated below for convenience,

$$\left.\begin{aligned} NLF(\zeta) =& \lambda' \Phi(-\zeta) \\ LLF(\zeta) =& \nu' \Phi(\mu - \zeta) \end{aligned}\right\} \tag{7.19}$$

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 $\zeta_1 \rightarrow -\infty$ 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.

7.9 References

Chapter 8

Three proper ROC fits

8.1 TBA How much finished

85%

8.2 TBA 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 7,
- The PROPROC (proper ROC) model described in TBA Chapter 20.
- The CBM (contaminated binormal model) described in TBA Chapter 20.

This chapter compares these methods by fitting the to 14 multiple-treatment multiple-reader datasets described in Chapter 12. [1]

Both RSM and CBM are implemented in `RJafroc`. `PROPROC` is implemented in Windows software [2] available here, last accessed 1/4/21.

[1] Comparing the RSM to the binormal model would be inappropriate, as the latter does not predict proper ROCs.

[2] OR DBM-MRMC 2.5, Sept. 04, 2014; the version used in this chapter is no longer distributed but is available from me upon request.

8.3 Application to two datasets

The RSM, PROPROC and CBM algorithms were applied to the 14 datasets described in Chapter 12.

```
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.

```
results <- array(list(), dim = 2)

results[[1]] <- Compare3ProperRocFits(datasetNames, 2) # VD dataset
results[[2]] <- Compare3ProperRocFits(datasetNames, 3) # FR dataset

resultsArr <- plotArr <- array(list(), dim = 2)

for (i in 1:2) {
  plotArr[[i]] <- results[[i]]$allPlots
  resultsArr[[i]] <- results[[i]]$allResults
}
```

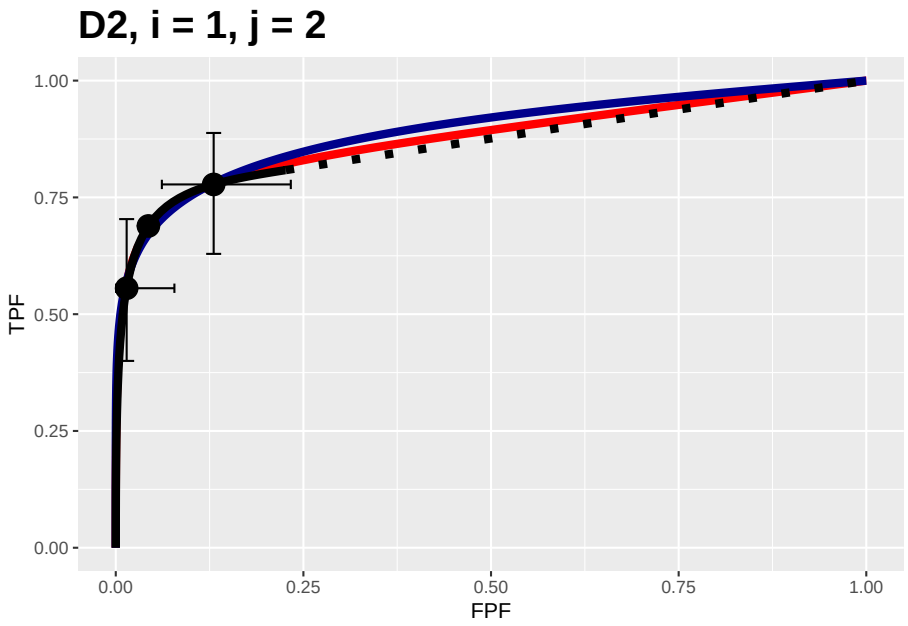
- The supporting code is in the function `Compare3ProperRocFits()` located at `R/compare-3-fits/Compare3ProperRocFits.R`.
- The analyzed results file locations are shown in Section 8.10.2.
- The fitted parameter results are contained in `resultsArr` and the composite plots (i.e., 3 combined plots corresponding to the three proper ROC fitting algorithms for each treatment and reader) are contained in `plotArr`.

8.4 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 double bracket is R-usage to index `lists`.
- The Van Dyke dataset contains $I \times J = 2 \times 5 = 10$ composite plots.
- The Franken dataset contains $I \times J = 2 \times 4 = 8$ composite plots.

- 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 in the `datasetNames` array, i.e., `datasetNames[2]`, the second treatment and the second reader. It contains 3 plots:

- The RSM fitted curve is in black.
- 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 95% confidence intervals for the lowest and uppermost operating points.

All 10 composite plots for the Van Dyke dataset are shown in Appendix 8.10.3.

8.5 RSM parameters

The parameters corresponding to the RSM plots are accessed as shown next.

- `resultsArr[[1]][[1,2]]$retRsm$mu` is the RSM μ parameter for the Van Dyke dataset for treatment 1 and reader 2,

- `resultsArr[[1]][[1,2]]$retRsm$lambdaP` is the RSM λ' parameter;

- `resultsArr[[1]][[1,2]]$retRsm$nuP` is the RSM ν' parameter;
- `resultsArr[[1]][[1,2]]$retRsm$zeta1` is the RSM ζ_1 parameter;
- In general the values are accessed as `[[f]][[i,j]]`, where `f` is the dataset index, `i` is the treatment index and `j` is the reader index;
- For the Van Dyke dataset `f = 1` and for the Franken dataset `f = 2`.

The following displays RSM parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
## 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 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].

Displayed next are RSM parameters for the Franken dataset, treatment 2 and reader 3:

```
## RSM parameters, Franken dataset, treatment 2, reader 3:
## mu =  2.641412
## lambdaP =  2.137379
## nuP =  0.784759
## zeta_1 =  -1.858565
## AUC =  0.8552573
## sigma_AUC =  0.03809136
## NLLini =  132.6265
## NLLfin =  127.9418
```

8.6 CBM parameters

The parameters of the CBM plots are accessed as shown next.

[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.

- `resultsArr[[f]][[i,j]]$retCbm$mu` is the CBM μ parameter;
- `resultsArr[[f]][[i,j]]$retCbm$alpha` is the CBM α parameter;

- `as.numeric(resultsArr[[f]][[i,j]]$retCbm$zetas[1])` is the CBM ζ_1 parameter, the threshold corresponding to the highest non-trivial operating point;
- `resultsArr[[f]][[i,j]]$retCbm$AUC` is the CBM AUC;
- `as.numeric(resultsArr[[f]][[i,j]]$retCbm$StdAUC)` is the standard deviation of the CBM AUC;
- `resultsArr[[f]][[i,j]]$retCbm$NLLIni` is the initial negative log-likelihood value;
- `rresultsArr[[f]][[i,j]]$retCbm$NLLFin)` is the final negative log-likelihood value.

The next example displays CBM parameters and AUC etc. for the Van Dyke dataset, treatment 1 and reader 2:

```
## 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:

```
## CBM parameters, Franken dataset, treatment 2, reader 3:
## mu =  2.340719
## alpha =  0.7860465
## zeta_1 =  -1.144089
## AUC =  0.8545476
## sigma_AUC =  0.03825439
## NLLini =  131.8453
## NLLfin =  128.0437
```

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.

8.7 PROPROC parameters

`PROPROC` displayed parameters are accessed as follows:

- `resultsArr[[f]][[i,j]]$c1` is the PROPROC c parameter;
- `resultsArr[[f]][[i,j]]$da` is the PROPROC d_a parameter;

- `resultsArr[[f]][[i,j]]$aucProp` is the PROPROC AUC;

Other statistics, such as standard error of AUC, are not provided by PROPROC software.

The next example displays PROPROC parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
## 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. 8.6.

The next example displays PROPROC parameters for the Franken dataset, treatment 2 and reader 3:

```
## PROPROC parameters, Franken dataset, treatment 2, reader 3:
## c =  -0.3551936
## d_a =  1.401807
## AUC =  0.8541372
```

The next section provides an overview of the most salient findings from analyzing the datasets.

8.8 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 fitting algorithms were applied.

It is easy to be overwhelmed by the numbers and this section summarizes an important conclusion:

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 each slope m_{PR} and m_{CR} equal to unity:

$$\left.\begin{aligned} \mathrm{AUC}_{PRO} &= \mathrm{m}_{PR}\mathrm{AUC}_{PRO} \\ \mathrm{AUC}_{CBM} &= \mathrm{m}_{CR}\mathrm{AUC}_{PRO} \\ \mathrm{m}_{PR} &= 1 \\ \mathrm{m}_{CR} &= 1 \end{aligned}\right\} \tag{8.1}$$

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}, and likewise for the plots of CBM AUC vs. RSM AUC. The reason for the *zero intercept* is that if the AUCs are identical one cannot have an offset (i.e., intercept) term.

8.8.1 Slopes

- Denote PROPROC AUC for dataset f, treatment i and reader j by AUC_{fij}^{PRO}. Likewise, the corresponding RSM and CBM values are denoted by AUC_{fij}^{RSM} and AUC_{fij}^{CBM}, respectively.
- For a given dataset the slope of the PROPROC values vs. the RSM values is denoted $\mathrm{m}_{PR,f}$.
- The (grand) average over all datasets is denoted $m_{\bullet}^{PR}$. Likewise, the (grand) average of the CBM AUC vs. the RSM slopes is denoted $m_{\bullet}^{CR}$.

An analysis was conducted to determine the average slopes and bootstrap confidence intervals.

The code for calculating the average slopes is in `R/compare-3-fits/slopesConvVsRsm.R` and that for the bootstrap confidence intervals is in `R/compare-3-fits/slopesAucsConvVsRsmCI.R`.

```
slopes <- slopesConvVsRsm(datasetNames)
slopeCI <- slopesAucsConvVsRsmCI(datasetNames)
```

The call to function `slopesConvVsRsm()` returns `slopes`, which contains, for each of 14 datasets, four `lists`: two plots and two slopes. For example:

- PRO vs. RSM: `slopes$p1[[2]]` is the plot of $\text{AUC}_{2\bullet\bullet}^{PRO}$ vs. $\text{AUC}_{2\bullet\bullet}^{RSM}$ for all treatments and readers in the Van Dyke dataset. The plot for dataset $f, f = 1, 2, ...14$ is accessed as `slopes$p1[[f]]` which yields the plot of $\text{AUC}_{f\bullet\bullet}^{PRO}$ vs. $\text{AUC}_{f\bullet\bullet}^{RSM}$.
- CBM vs. RSM: `slopes$p2[[2]]` is the plot of $\text{AUC}_{2\bullet\bullet}^{CBM}$ vs. $\text{AUC}_{2\bullet\bullet}^{RSM}$ for for all treatments and readers in the Van Dyke dataset. The plot for dataset f is accessed as `slopes$p2[[f]]`.
- PRO vs. RSM: `slopes$m_pro_rsm` has two columns, each of length 14, the slopes $\text{m}_{PR,f}$ for the datasets (indexed by f) and the corresponding R^2 values, where R^2 is the fraction of variance explained by the constrained straight line fit. The first column is `slopes$m_pro_rsm[[1]]` and the second column is `slopes$m_pro_rsm[[2]]`.
- CBM vs. RSM: `slopes$m_cbm_rsm` has two columns, each of length 14, the slopes $\text{m}_{CR,f}$ for the datasets and the corresponding R^2 values. The first column is `slopes$m_cbm_rsm[[1]]` and the second column is `slopes$m_cbm_rsm[[2]]`.

As an example, for the Van Dyke dataset, `slopes$p1[[2]]` which is shown in the left in Fig. 8.1, is the plot of $\text{AUC}_{2\bullet\bullet}^{PRO}$ vs. $\text{AUC}_{2\bullet\bullet}^{RSM}$. Shown in the right is `slopes$p2[[2]]`, the plot of $\text{AUC}_{2\bullet\bullet}^{CBM}$ vs. $\text{AUC}_{2\bullet\bullet}^{RSM}$. Each plot has the constrained linear fit superposed on the $2 \times 5 = 10$ data points; each data point represents a distinct modality-reader combination.

The next plot shows corresponding plots for the Franken dataset in which there are $2 \times 4 = 8$ points in each plot.

8.8.2 Confidence intervals

The call to `slopesAucsConvVsRsmCI` returns `slopeCI`, containing the results of the bootstrap analysis (the bullet symbols $\bullet$ denote grand averages over 14 datasets):

- `slopeCI$cislopeProRsm` 95-percent confidence interval for $m_{PR\bullet}$
- `slopeCI$cislopeCbmRsm` 95-percent confidence interval for $m_{CR\bullet}$
- `slopeCI$histSlopeProRsm` histogram of 200 bootstrap values of $m_{PR\bullet}$
- `slopeCI$histSlopeCbmRsm` histogram of 200 bootstrap values of $m_{CR\bullet}$
- `slopeCI$ciAvgAucRsm` confidence interval from 200 bootstrap values of $\text{AUC}_{\bullet}^{RSM}$
- `slopeCI$ciAvgAucPro` confidence interval for 200 bootstrap values of $\text{AUC}_{\bullet}^{PRO}$
- `slopeCI$ciAvgAucCbm` confidence interval for 200 bootstrap values of $\text{AUC}_{\bullet}^{CBM}$

As examples,

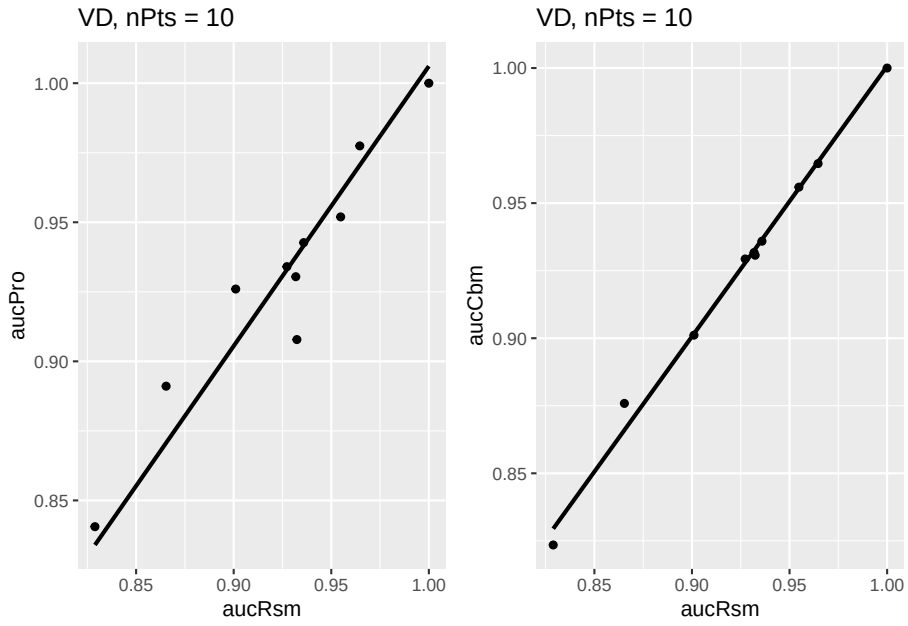


Figure 8.1: Van Dyke dataset: Left plot is PROPROC-AUC vs. RSM-AUC with the superposed constrained linear fit. The number of data points is `nPts` = 10. Right plot is CBM-AUC vs. RSM-AUC.

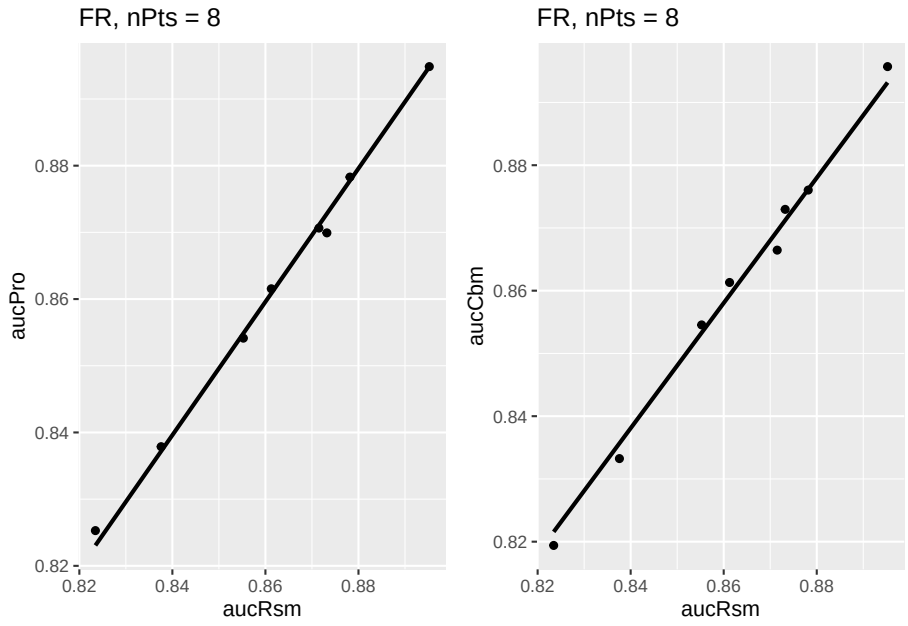


Figure 8.2: Similar to previous plot, for Franken dataset.

```
##           m-PR      m-CR
## 2.5%  1.005092 0.9919886
## 97.5% 1.012285 0.9966149
```

The CI for $m_{PR\bullet}$ is slightly above unity, while that for $m_{CR\bullet}$ is slightly below. Shown next is the histogram plot for $m_{PR\bullet}$ (left plot) and $m_{CR\bullet}$ (right plot). Quantiles of these histograms were used to compute the confidence intervals cited above.

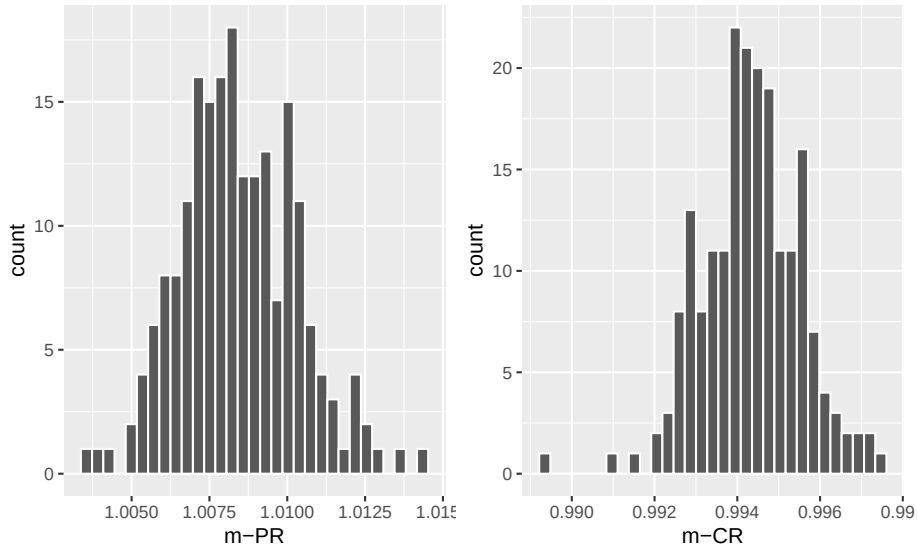


Figure 8.3: Histograms of slope PROPROC AUC vs. RSM AUC (left) and slope CBM AUC vs. RSM AUC (right).

8.8.3 Summary of slopes and confidence intervals

Table 8.1: 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

In Table 8.1 the second column, labeled m_{PR}, shows slopes of straight lines, constrained to go through the origin, to PROPROC AUC vs. RSM AUC values, for each of the 14 datasets, as labeled in the fits column. The third column, labeled R^2_{PR}, lists the square of the correlation coefficient for each fit. The fourth and fifth columns list the corresponding values for the CBM AUC vs. RSM AUC fits. The second last row lists the grand averages (AVG) and the last row lists the 95 percent confidence intervals.

8.9 TBA Discussion / Summary

8.10 Appendices

8.10.1 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` [4] 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.

The contents of `R/compare-3-fits/MRMCRuns/VD/VDproprocnormareapooled.csv` are shown next, see Fig. 8.6. [5] 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).

8.10.2 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.).

8.10.3 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.

The RSM parameter values for the treatment 2 plot are: $\mu = 5.9346513$, $\lambda' = 0.3809031$, $\nu' = 0.9292484$, $\zeta_1 = 0.479145$. The corresponding CBM values are $\mu = 5.9356142$, $\alpha = 0.9292952$, $\zeta_1 = 1.20877$. 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.]

[4] In accordance with R-package policies white-spaces in the original `PROPROC` output file names have been removed.

[5] The `VD.lrc` file in this directory is the Van Dyke data formatted for input to OR DBM-MRMC 2.5.

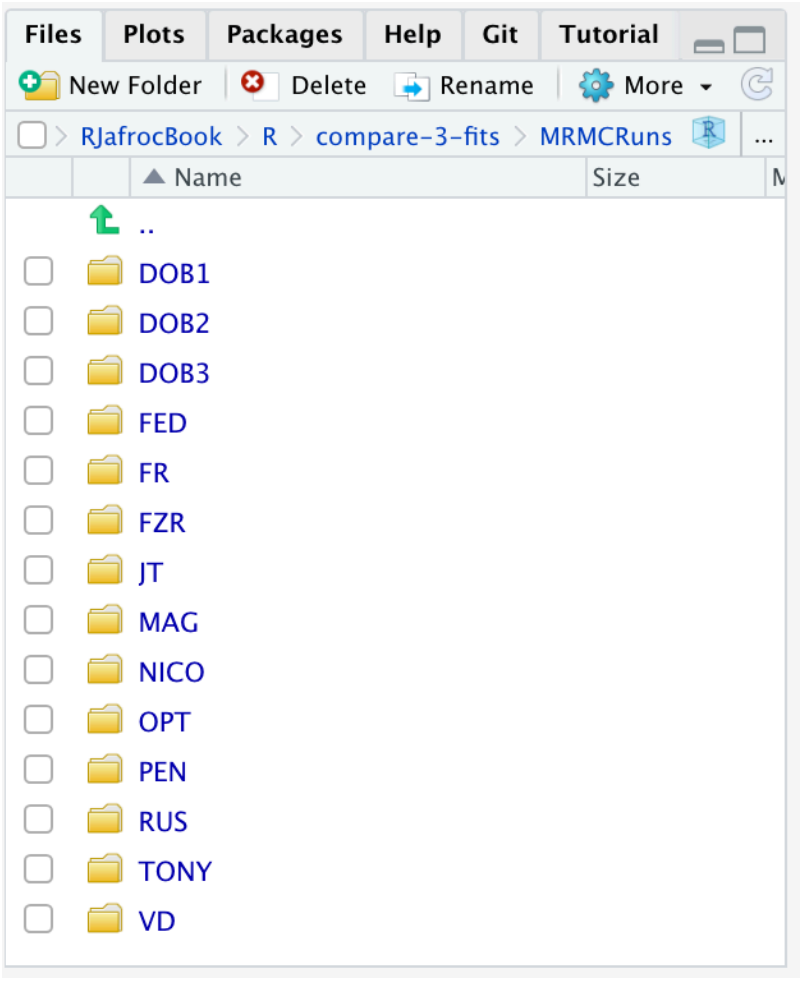


Figure 8.4: Screen shot (1 of 2) of 'R/compare-3-fits/MRMCRuns' showing the folders containing the results of PROPROC analysis on 14 datasets.

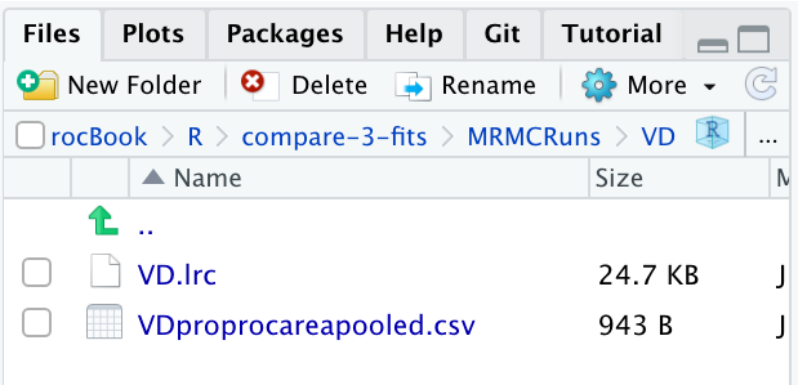


Figure 8.5: Screen shot (2 of 2) of 'R/compare-3-fits/MRMCRuns/VD' showing files containing the results of PROPROC analysis for the Van Dyke dataset.

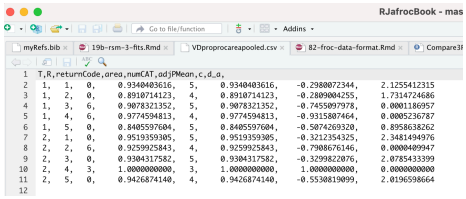


Figure 8.6: PROPROC output for the Van Dyke ROC data set. The first column is the treatment, the second is the reader, the fourth is the AUC and the last two columns are the c and d_a parameters.

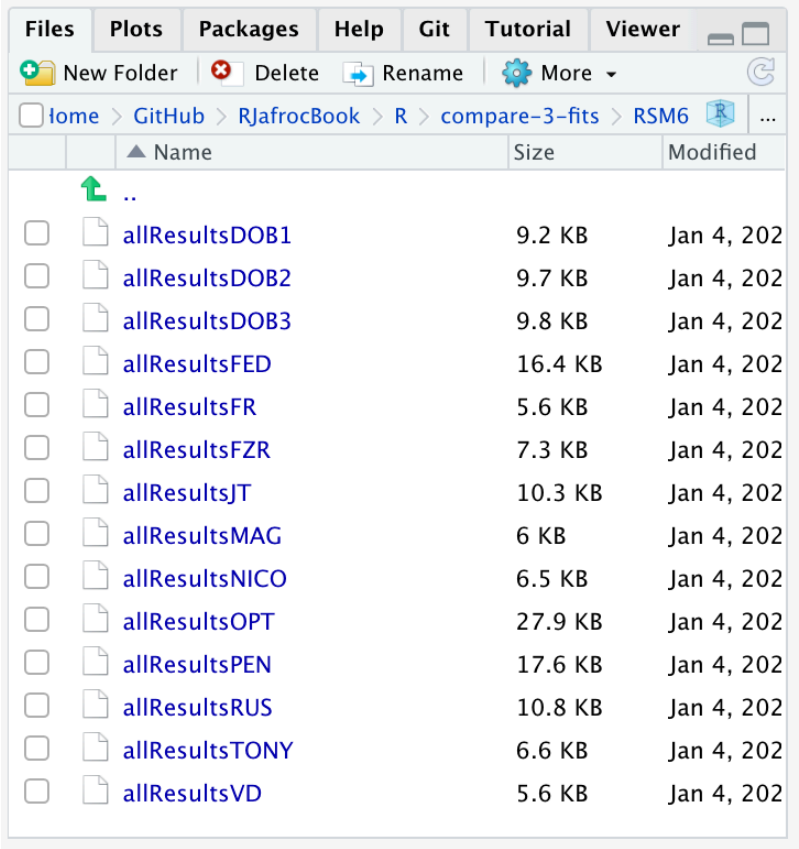


Figure 8.7: Screen shot of 'R/compare-3-fits/RSM6' showing the results files created by 'Compare3ProperRocFits()'.

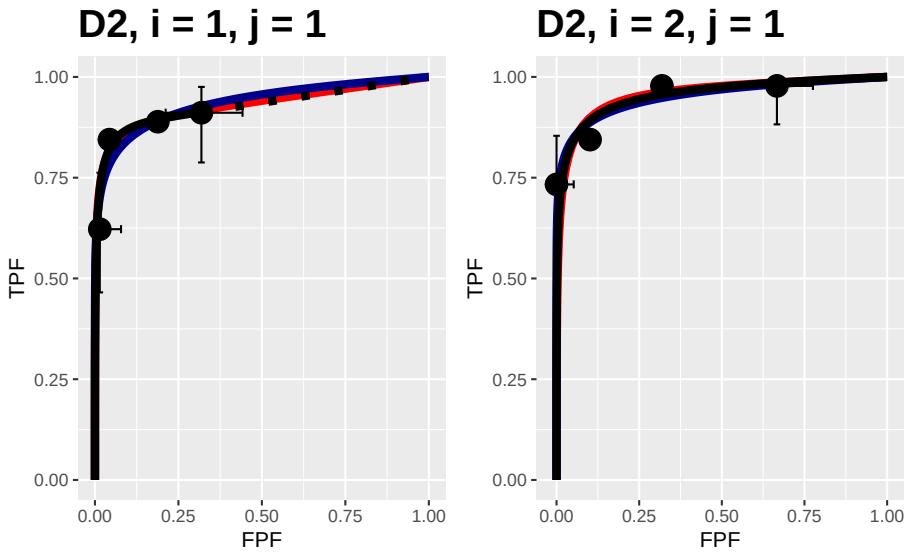


Figure 8.8: Composite plots in both treatments for Van Dyke dataset, reader 1.

The RSM parameters for the treatment 1 plot are: $\mu = 2.2014133$, $\lambda' = 0.2569453$, $\nu' = 0.7524016$, $\zeta_1 = -0.1097901$. The corresponding CBM values are $\mu = 2.7457914$, $\alpha = 0.7931264$, $\zeta_1 = 1.1250285$.

8.11 References

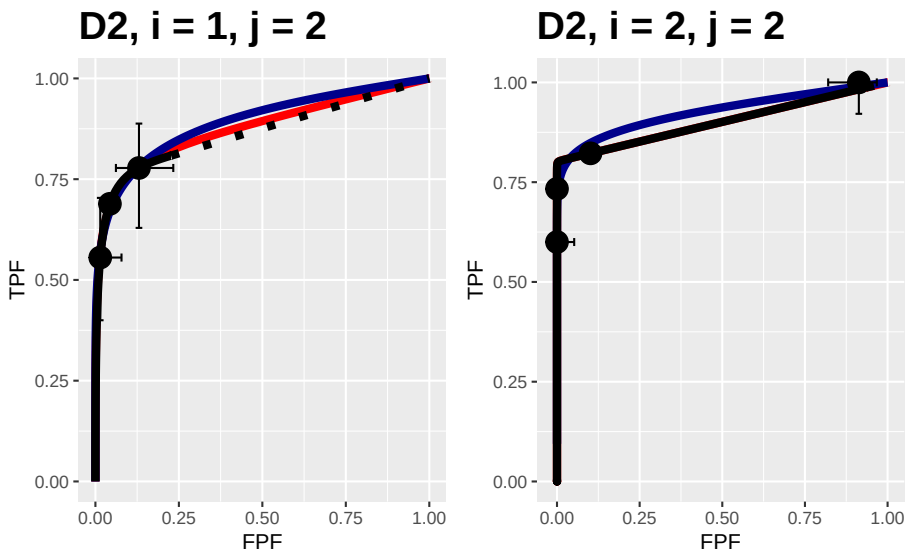


Figure 8.9: Composite plots in both treatments for Van Dyke dataset, reader 2. For treatment 2 the RSM and PROPROC fits are indistinguishable.

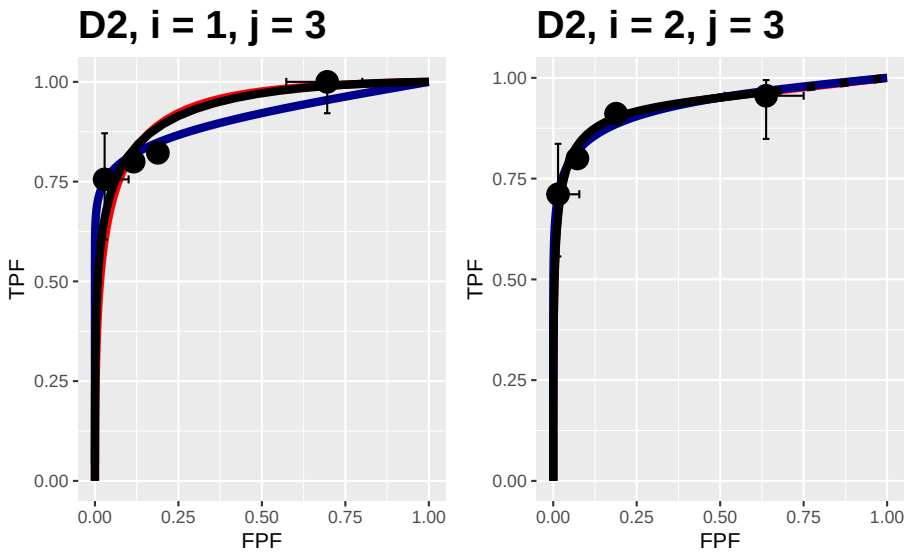


Figure 8.10: Composite plots in both treatments for Van Dyke dataset, reader 3.

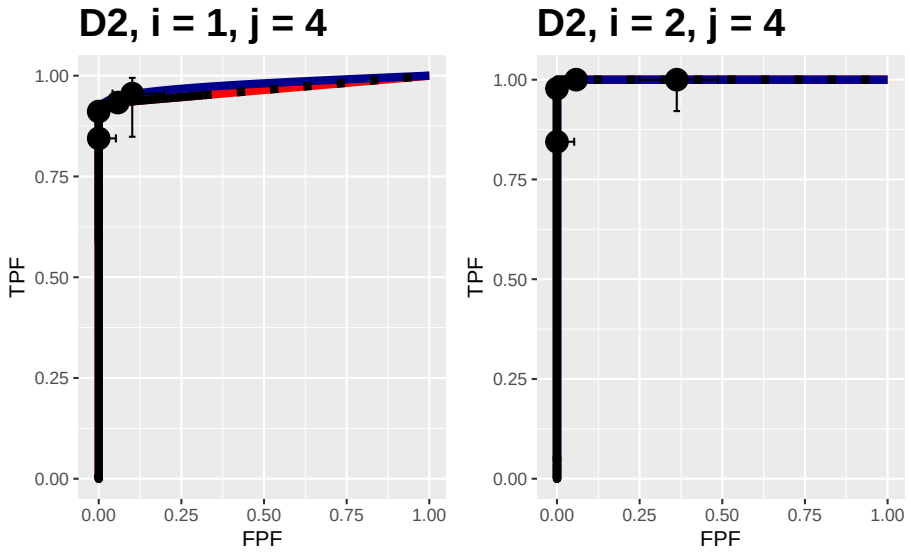


Figure 8.11: Composite plots in both treatments for Van Dyke dataset, reader 4. For treatment 2 the 3 plots are indistinguishable and each one has AUC = 1. The degeneracy is due to all operating points being on the axes of the unit square.

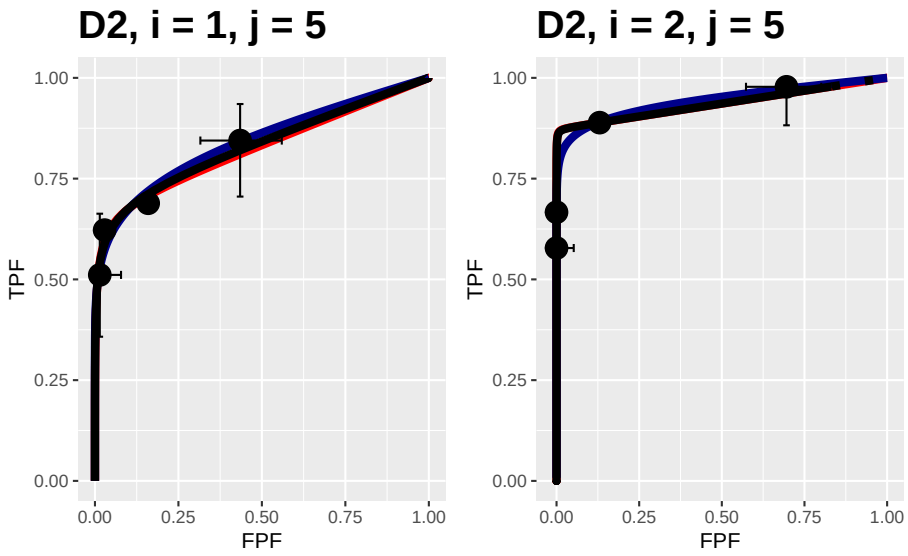


Figure 8.12: Composite plots in both treatments for Van Dyke dataset, reader 5.

CAD applications

Chapter 9

Standalone CAD vs. Radiologists

9.1 TBA How much finished

10%

9.2 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.`

9.3 Keywords

Technology assessment, computer-aided detection (CAD), screening mammography, standalone performance, single-treatment multi-reader ROC analysis.

9.4 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).

9.5 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.

9.5.1 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.

9.5.1.1 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 $PCL_{0.2}$.

9.5.1.2 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 9.1. 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).

9.5.1.3 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 9.1: 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).

9.5.2 The 1T-RRFC analysis model

The sampling model for the FOM is:

$$\left.\begin{array}{r} \theta_j = \mu + R_j \\ (j = 1, 2, ...J) \end{array}\right\} \tag{9.1}$$

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:

$$R_j \sim N\left(0, \sigma_R^2\right) \tag{9.2}$$

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.

9.5.3 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:

$$\theta_{ij\{c\}} = \mu + \tau_i + (\tau \mathrm{R})_{ij} + \epsilon_{ij\{c\}} \tag{9.3}$$

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. (9.3) are fixed effects (average performance and treatment effect, respectively). The next two terms are random effect variables that, by assumption, are sampled as follows:

$$\left.\begin{aligned} R_j &\sim N\left(0, \sigma_R^2\right) \\ (\tau R)_{ij} &\sim N\left(0, \sigma_{\tau R}^2\right) \end{aligned}\right\} \tag{9.4}$$

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:

$$\epsilon_{ij\{c\}} \sim N_{I \times J}\left(\vec{0}, \Sigma\right) \tag{9.5}$$

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:

$$\mathrm{Cov}\left(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}\right) = \left\{\begin{array}{l} \mathrm{Var}\ (i = i', j = j') \\ \mathrm{Cov1}\ (i \neq i', j = j') \\ \mathrm{Cov2}\ (i = i', j \neq j') \\ \mathrm{Cov3}\ (i \neq i', j \neq j') \end{array}\right\} \tag{9.6}$$

Software {U of Iowa and `RJafroc`} yields estimates of all terms appearing on the right hand side of Eqn. (9.6). Excluding fixed effects, the model represented by Eqn. (9.3) contains six parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \text{Var}, \text{Cov}_1, \text{Cov}_2, \text{Cov}_3 \tag{9.7}$$

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:

$$\text{Var} \geq \text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \tag{9.8}$$

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.

9.5.4 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:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{9.9}$$

$\theta_{j\{c\}}$ 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:

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{9.10}$$

The $J \times J$ covariance matrix Σ is defined by two parameters, Var and Cov_2, as follows:

$$\Sigma_{jj'} = \mathrm{Cov}\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \mathrm{Var} & j = j' \\ \mathrm{Cov}_2 & j \neq j' \end{cases} \tag{9.11}$$

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):

$$\Sigma_{jj'}|_{\mathrm{jack}} = \frac{K-1}{K} \sum_{k=1}^{K} \left(\psi_{j(k)} - \psi_{j(\bullet)}\right)\left(\psi_{j'(k)} - \psi_{j'(\bullet)}\right) \tag{9.12}$$

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:

$$\left.\begin{aligned} \mathrm{Var} &= \left\langle \Sigma_{jj'}|_{\mathrm{jack}} \right\rangle_{j=j'} \\ \mathrm{Cov}_2 &= \left\langle \Sigma_{jj'}|_{\mathrm{jack}} \right\rangle_{j\neq j'} \end{aligned}\right\} \tag{9.13}$$

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):

$$\theta_{j\{c\}} = \theta_{0\{c\}} + \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{9.14}$$

$\theta_{0\{c\}}$ 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.,

$$\psi_{j\{c\}} = \theta_{j\{c\}} - \theta_{0\{c\}} \tag{9.15}$$

Then Eqn. (9.14) reduces to:

$$\psi_{j\{c\}} = \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{9.16}$$

Eqn. (9.16) is identical in form to Eqn. (9.9) with the difference that the figure of merit on the left hand side of Eqn. (9.16) is a *difference FOM*, that between the radiologist's and CAD. Eqn. (9.16) 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. (9.16).

Apart from fixed effects, the model in Eqn. (9.16) contains three parameters:

$$\sigma_R^2, \text{Var}, \text{Cov}_2 \tag{9.17}$$

Setting $\text{Var} = 0, \text{Cov}_2 = 0$ 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.]

9.6 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 and Zhai, 2022).

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.

9.7 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 9.2, are the significance testing results corresponding to the three analyses.

Table 9.2: 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
PCL_0_05	2T-RRRC	4.5e-01	(2.58e-01,6.42e-01)	4.93e-01	(3.76e-01,6.11e-01)	4.33e-02	(-1.57e-01,2.44e-01)	1.79e-01	7.84e+02	6.7e-01
PCL_0_05	1T-RRRC	4.5e-01	NA	4.93e-01	(2.93e-01,6.94e-01)	4.33e-02	(-1.57e-01,2.44e-01)	1.79e-01	7.84e+02	6.7e-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
PCL_0_2	2T-RRRC	5.92e-01	(4.78e-01,7.05e-01)	7.1e-01	(6.33e-01,7.87e-01)	1.19e-01	(4.45e-03,2.33e-01)	4.16e+00	9.37e+02	4.2e-02
PCL_0_2	1T-RRRC	5.92e-01	NA	7.1e-01	(5.96e-01,8.24e-01)	1.19e-01	(4.45e-03,2.33e-01)	4.16e+00	9.37e+02	4.2e-02
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
PCL_1	2T-RRRC	6.75e-01	(5.71e-01,7.79e-01)	7.83e-01	(7.12e-01,8.54e-01)	1.08e-01	(4.5e-03,2.12e-01)	4.2e+00	4.93e+02	4.1e-02
PCL_1	1T-RRRC	6.75e-01	NA	7.83e-01	(6.8e-01,8.87e-01)	1.08e-01	(4.5e-03,2.12e-01)	4.2e+00	4.93e+02	4.1e-02
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
Wilcoxon	2T-RRRC	8.17e-01	(7.52e-01,8.82e-01)	8.49e-01	(8.07e-01,8.9e-01)	3.17e-02	(-3.1e-02,9.45e-02)	9.86e-01	8.78e+02	3.2e-01
Wilcoxon	1T-RRRC	8.17e-01	NA	8.49e-01	(7.86e-01,9.11e-01)	3.17e-02	(-3.1e-02,9.45e-02)	9.86e-01	8.78e+02	3.2e-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 9.2 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 9.2.

1. Where a direct comparison is possible, namely 1T-RRFC analysis using and as FOMs, the p-values in Table 9.2 are similar to those reported in Study – 1.
2. All FOMs (i.e., θ_0, $\theta_\bullet$ and $\psi_\bullet$) in Table 9.2 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 9.2.
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. (9.16).

Shown next, Table 9.3, are the model-parameters corresponding to the three analyses.

Table 9.3: 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 9.3.

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 9.2. $\sigma_R^2 = 0$ 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. (9.8), 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).

9.8 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.

9.9 Appendix 1

The structures of theR objects generated by the software are illustrated with three examples.

9.9.1 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.

9.9.2 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. (9.3) – Eqn. (9.6).

- `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.

9.9.3 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] 0.0053445377
#>
#> $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. (9.17).
- `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.

9.10 Appendix 2

TBA

```
source(here("R/standalone-cad/DfReadLrocDataFile.R"))
lrocDataset <- DfReadLrocDataFile()
```

9.11 References

Chapter 10

Optimal operating point

10.1 TBA How much finished

95%

Discussion needs more work

10.2 Introduction

A familiar problem for the computer aided detection or artificial intelligence (CAD/AI) algorithm designer is how to determine the optimal reporting threshold of the algorithm. Assuming that designer level mark-rating FROC data is available for the algorithm, 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 (or the next stage of the AI algorithm – in what follows references to CAD apply equally to AI algorithms).

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). Typically it is maximized at the point that is closest to the (0,1) corner of the ROC.

Lacking a procedure for determining it analytically CAD designers, 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) of the wAFROC curve. The method is compared to the Youden-index based method.

10.3 Methods

Terminology: Non-lesion localizations = NLs, i.e., location level "false positives". Lesion localizations = LLs, i.e., location level "true positives". Latent marks = perceived suspicious regions that are not necessarily marked. There is a distinction, see below, between perceived and actual marks.

Background on the radiological search model (RSM) is provided in Chapter 4. The model predicts ROC, FROC and wAFROC curves and is characterized by the four parameters – $\mu, \lambda, \nu, \zeta_1$ – with the following meanings:

- The μ parameter, $\mu \geq 0$, is the perceptual signal-to-noise-ratio of lesions. Higher values of μ lead to increasing separation of two unit variance normal distributions determining the ratings of perceived NLs and LL. As μ increases performance of the algorithm increases.
- The λ parameter, $\lambda \geq 0$, determines the mean number of latent NLs per case. Higher values lead to more latent NL marks per case and decreased performance.
- The ν parameter, $0 \leq \nu \leq 1$, determines the probability of latent LLs, i.e., the probability that any present lesion will be perceived. Higher values of ν lead to more latent LL marks and increased performance.
- If its rating exceeds ζ_1 the latent mark is actually marked. Higher values of ζ_1 correspond to more stringent reporting criteria and fewer actual marks. As will be shown next performance, as measured by wAFROC-AUC or the Youden-index, peaks at an optimal value of ζ_1. The purpose of this chapter is to investigate this effect, i.e., given the other RSM parameters and the figure of merit to be optimized (i.e., wAFROC-AUC or the Youden-index), to determine the optimal value of 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. The value maximizing wAFROC-AUC is denoted $\zeta_1(1, \mu, \lambda, \nu)$ and that maximizing the Youden-index is denoted $\zeta_1(2, \mu, \lambda, \nu)$.

The wAFROC figure of merit is implemented in the `RJafroc` function `UtilAnalyticalAucsRSM`. It is calculated using Eqn. (5.26).

The Youden-index is defined as sensitivity plus specificity minus 1. Sensitivity is implemented in function `RSM_yROC` and specificity is the complement of `RSM_xROC`.

10.4 Varying λ optimizations

In the following $f = 1$ denotes wAFROC-AUC optimization and $f = 2$ denotes Youden-index optimization.

```
muArr <- c(2)
lambdaPArr <- c(1, 2, 5, 10)
nuPArr <- c(0.9)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

For $\mu = 2$ and $\nu = 0.9$ wAFROC-AUC and Youden-index optimizations were performed for $\lambda = 1, 2, 5, 10$. Half of the diseased cases contained one lesion and the rest contained two lesions. On cases with two lesions the lesions were assigned equal weights (i.e., equal clinical importance).

The following quantities were calculated:

- $\zeta_1(f, \mu, \lambda, \nu)$: the optimal thresholds;
- wAFROC(f, μ, λ, ν): the value of the wAFROC-AUC. For consistency we always report wAFROC-AUC even when the optimized quantity is the Youden-index;
- ROC(f, μ, λ, ν): the AUCs under the ROC curves;
- NLF(f, μ, λ, ν) and LLF(f, μ, λ, ν): the coordinates of the operating point on the FROC curve corresponding to $\zeta_1(f, \mu, \lambda, \nu)$.

10.4.1 Summary table

Table 10.1 summarizes the results. The column labeled FOM shows the quantity being maximized (wAFROC-AUC or the Youden-index), the column labeled λ lists the 4 values of λ, ζ_1 is the optimal value of ζ_1 that maximizes the chosen figure of merit. The column labeled wAFROC is the AUC under the wAFROC curve, the column labeled ROC is the AUC under the ROC curve, and (NLF, LLF) is the operating point on the FROC curve corresponding to

the value of ζ_1 in the third column. All quantities in columns 3 through 6 are functions of f, μ, λ, ν.

Table 10.1: Summary of optimization results for $\mu = 2$, $\nu = 0.9$ and 4 values of λ. FOM = figure of merit. wAFROC = wAFROC-AUC, ROC = ROC-AUC, (NLF,LLF) = operating point on FROC.

FOM	λ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	1	-0.007	0.864	0.929	(0.503, 0.880)
	2	0.474	0.809	0.900	(0.636, 0.843)
	5	1.272	0.715	0.840	(0.509, 0.690)
	10	1.856	0.645	0.774	(0.317, 0.502)
Youden	1	1.095	0.831	0.899	(0.137, 0.735)
	2	1.362	0.781	0.865	(0.173, 0.664)
	5	1.695	0.705	0.811	(0.225, 0.558)
	10	1.934	0.644	0.766	(0.265, 0.474)

Inspection of this table reveals the following effects:

1. For either FOM, as λ increases the optimal threshold $\zeta_1(f, \mu, \lambda, \nu)$ increases and $\text{wAFROC}(f, \mu, \lambda, \nu)$, $\text{ROC}(f, \mu, \lambda, \nu)$ and $\text{LLF}(f, \mu, \lambda, \nu)$ decrease. Equivalently, CAD performance decreases, regardless of how it is measured (i.e., wAFROC-AUC or ROC-AUC).

2. The wAFROC based based optimal thresholds are smaller than the corresponding Youden-index based optimal thresholds, i.e., $\zeta_1(1, \mu, \lambda, \nu) < \zeta_1(2, \mu, \lambda, \nu)$. A small threshold corresponds to a less strict reporting criterion.

3. For fixed μ, λ, ν the operating point on the FROC for $f = 2$ is below that corresponding to $f = 1$:

   - $\text{NLF}(2, \mu, \lambda, \nu) < \text{NLF}(1, \mu, \lambda, \nu)$ and $\text{LLF}(2, \mu, \lambda, \nu) < \text{LLF}(1, \mu, \lambda, \nu)$.
   - The difference decreases with increasing λ.
   - These effects are illustrated in Fig. 10.1.

4. For fixed μ, λ, ν the Youden-index based optimization yields lesser performance than the corresponding wAFROC-AUC based optimization:

- wAFROC $(2, \mu, \lambda, \nu)$ < wAFROC $(1, \mu, \lambda, \nu)$ and ROC $(2, \mu, \lambda, \nu)$ < ROC $(1, \mu, \lambda, \nu)$.
- The difference decreases with increasing λ.
- These effects are illustrated in Fig. 10.2.

10.4.2 FROC

The third effect is illustrated by the FROC plots with superimposed operating points for varying λ shown in Fig. 10.1. The black dots correspond to $f = 1$ and the red dots correspond to $f = 2$. The black dots are consistently above the red dots and the separation of the dots is greatest for $\lambda = 1$.

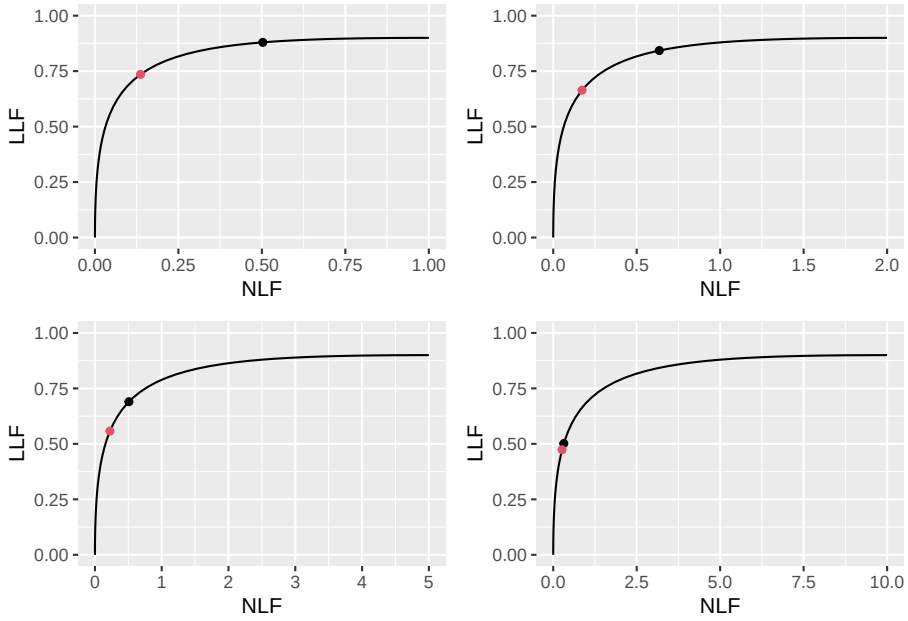


Figure 10.1: FROC plots with superimposed operating points for varying λ. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

10.4.3 wAFROC

The decrease in wAFROC (f, μ, λ, ν) with increasing λ (contained in the first effect) is illustrated in Fig. 10.2 which shows wAFROC plots for the two optimization methods. Each plot consists of a continuous curve followed by a

dashed line. The red curve, which appears as a "green red red-dashed" curve [1] corresponds to wAFROC-AUC optimization $f = 1$ and the green green-dashed curve corresponds to Youden-index optimization $f = 2$.

The transition from continuous to dashed is determined by the value of ζ_1. The transition occurs at a higher value of ζ_1 for the Youden-index optimization. The stricter Youden-index based reporting threshold sacrifices some of the area under the wAFROC. This results in lower performance particularly for the lower values of λ. At the highest value of λ the values of optimal ζ_1 are similar and both methods make similar predictions, as evident in Fig. 10.2.

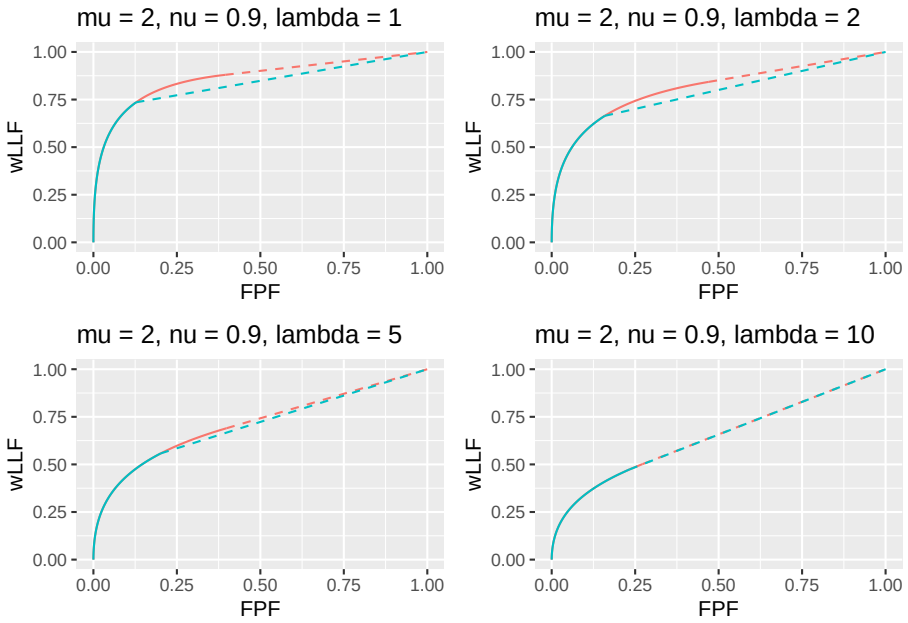


Figure 10.2: wAFROC plots for the two optimization methods: the green red red-dashed curve curve corresponds to wAFROC-AUC optimization and the green green-dashed curve corresponds to Youden-index optimization. The wAFROC optimizations yield greater performance than do Youden-index optimizations and the difference decreases with increasing λ.

10.4.4 ROC

The decrease in ROC (f, μ, λ, ν) with increasing λ (also contained in the first effect) is illustrated in Fig. 10.3 which shows RSM-predicted ROC plots for the

[1] The curve for $f = 1$ is in fact a red curve, complicated by superposition of the green curve over part of its traverse.

two optimization methods. Again, each plot consists of a continuous curve followed by a dashed curve and a similar color-coding convention is used as in Fig. 10.2. The ROC plots show similar dependencies as described for the wAFROC plots: specifically, the stricter Youden-index based reporting threshold sacrifices some of the area under the ROC resulting in lower performance, particularly for the lower values of λ.

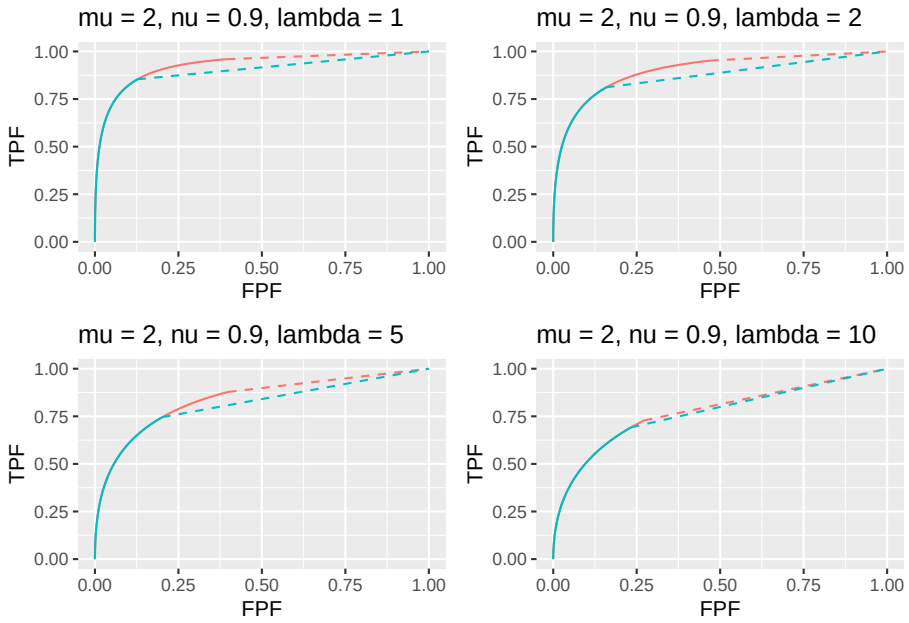


Figure 10.3: ROC plots for the two optimization methods: the green-red-red-dashed curve corresponds to wAFROC-AUC optimization and the green-green curve corresponds to Youden-index optimization. The wAFROC optimizations yield greater performance than do Youden-index optimizations and the difference decreases with increasing λ.

10.4.5 Why not maximize ROC-AUC?

Since the ROC curves show a similar dependence as the wAFROC curves why not maximize ROC-AUC instead of wAFROC-AUC? It can be shown that as long as one restricts to proper ROC models, this will always result in $\zeta_1 = -\infty$.

For a proper ROC curve the slope decreases monotonically as the operating point moves up the curve and at each point the slope is greater than that of the straight curve connecting the point to (1,1). This geometry ensures that AUC under any curve with a finite ζ_1 is smaller than that under the full curve.

Therefore maximum AUC can only be attained by choosing $\zeta_1 = -\infty$. This is illustrated in Fig. 10.4 which shows a binormal ROC curve corresponding to $a = 2$ and $b = 1$, which is a proper ROC curve. The dot is the operating point corresponding to $\zeta_1 = 1.5$. In the region above the dot the continuous curve is above the dotted line, meaning AUC performance of an observer who adopts a finite ζ_1 is less than performance of an observer who rates all cases, i.e., adopts $\zeta_1 = -\infty$.

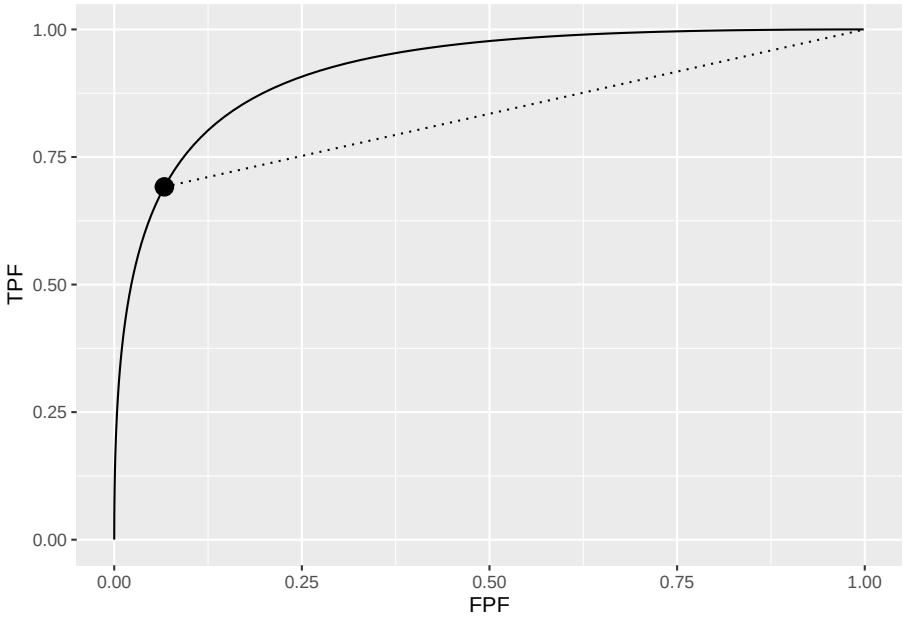


Figure 10.4: In the region above the dot the proper curve is above the dotted line, meaning performance of an observer who adopts a finite ζ_1 is less than performance of an observer who adopts $\zeta_1 = -\infty$.

10.5 Varying ν and μ optimizations

Details of varying ν, including tables and figures, are in Appendix 11.1. The results are similar to those just described for varying λ but, since unlike λ increasing ν results in increasing performance, the directions of the effects are reversed. As ν increases wAFROC-AUC and ROC-AUC performances increase and the reporting threshold ζ_1 decreases. The Youden-index based optimal threshold is almost independent of ν which results in relatively constant NLF while LLF increases with increasing ν. As before wAFROC optimization yields

lower reporting threshold and higher performance than Youden-index optimization.

Details of varying μ are in Appendix 11.2. Increasing μ results in increasing performance and is accompanied by increasing ζ_1: LLF is relatively constant while NLF decreases for both optimization methods. Again wAFROC optimization yields lower reporting threshold and higher performance than Youden-index optimization.

10.6 Very high or very low performance

Limiting situations covering very high and very low performances are described in Appendix 11.3.

For very high performance, defined as ROC-AUC > 0.9, both methods place the optimal operating point on the sharp bend near the upper-left corner of all operating characteristics. The wAFROC based method chooses a lower threshold than the Youden-index method resulting in a higher operating point on the FROC and higher wAFROC-AUC and ROC-AUC performance. The difference between the two methods decreases as ROC-AUC $\rightarrow 1$.

For very low performance, defined as $0.5 <$ ROC-AUC < 0.6, the Youden-index method chooses a lower threshold compared to wAFROC optimization, resulting in a higher operating point on the FROC, greater ROC-AUC but sharply lower wAFROC-AUC. The difference between the two methods increases as ROC-AUC $\rightarrow 0.5$. In this limit the wAFROC method severely limits the numbers of marks shown to the radiologist as compared to the Youden-index based method.

10.7 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, as describe here. 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 this chapter to determine 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.

10.8 A CAD 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 reader data, 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. Each diseased case contains one lesion: `lesDistr = c(1)`. `FitRsmRoc` fits the binned ROC dataset to the radiological search model (RSM). Object `fit` contains the RSM 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")
lesDistrCad <- c(1)
relWeightsCad <- c(1)
fit <- FitRsmRoc(dsCadRocBinned, lesDistrCad)
cat("fitted values: \nmu = ", fit$mu,
    "\nlambda = ", fit$lambdaP,
    "\nnu = ", fit$nuP, "\n")
#> fitted values:
#> mu =  2.755784
#> lambda =  6.778332
#> nu =  0.8033886
```

10.8.1 Summary table

Table 10.2 summarizes the results. As compared to Youden-index optimization the wAFROC-AUC based optimization results in a lower reporting threshold ζ_1, larger figures of merit – see Fig. 10.6 for wAFROC-AUC and Fig. 10.7 for ROC-AUC – and a higher operating point on the FROC, see Fig. 10.5. These results match the trends shown in Table 10.1.

Table 10.2: Summary of optimization results for example CAD FROC dataset. Table header row as in the previous table.

FOM	λ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	6.778	1.739	0.774	0.815	(0.278, 0.679)
Youden		1.982	0.770	0.798	(0.161, 0.627)

10.8.2 FROC

Fig. 10.5 shows FROC curves with superimposed optimal operating points. With NLF = 0.278, a four-view mammogram would show about 1.2 false CAD marks per patient and lesion-level sensitivity would be about 68 percent.

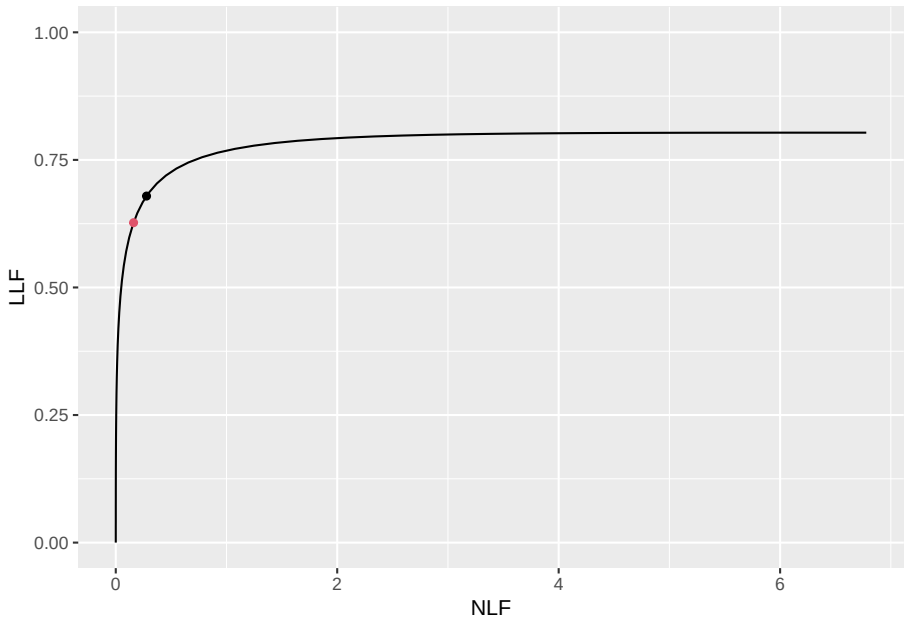


Figure 10.5: FROC plots with superposed optimal operating points. Black dot is using wAFROC optimization and red dot is using Youden-index optimization.

10.8.3 wAFROC

Fig. 10.6 shows wAFROC curves using the two methods. The red curve is using wAFROC-AUC optimization and the green curve is using Youden-index optimization. The difference in AUCs is small - following the trend described in Section 10.5 for the larger values of λ.

10.8.4 ROC

Fig. 10.7 shows ROC curves using the two methods. The red curve is using wAFROC-AUC optimization and the green curve is using Youden-index optimization. The difference in AUCs is larger, but recall that ROC-AUC performance is not being optimized.

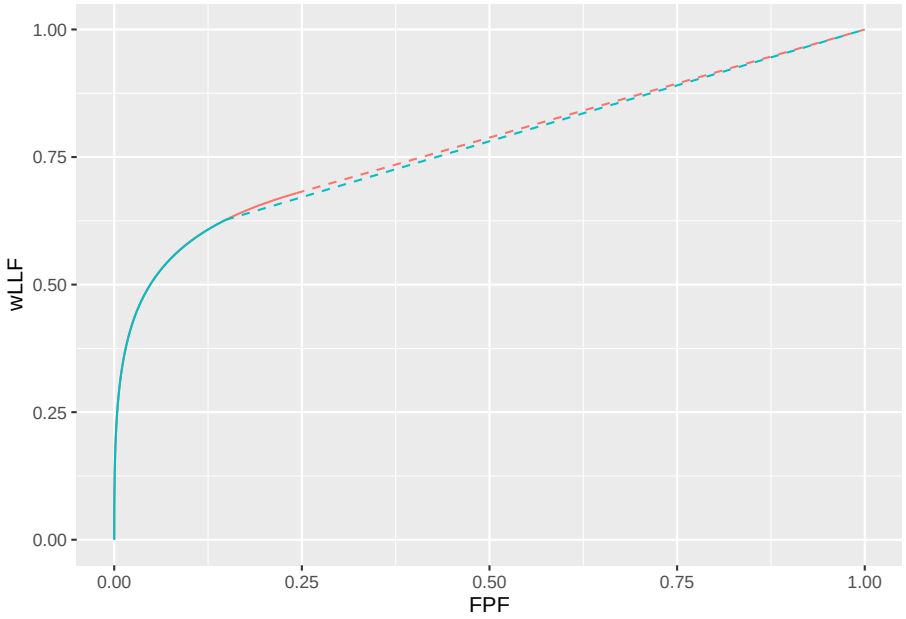


Figure 10.6: The color coding is as in previous figures. The two wAFROC-AUCs are 0.774 (wAFROC optimization) and 0.770 (Youden-index optimization).

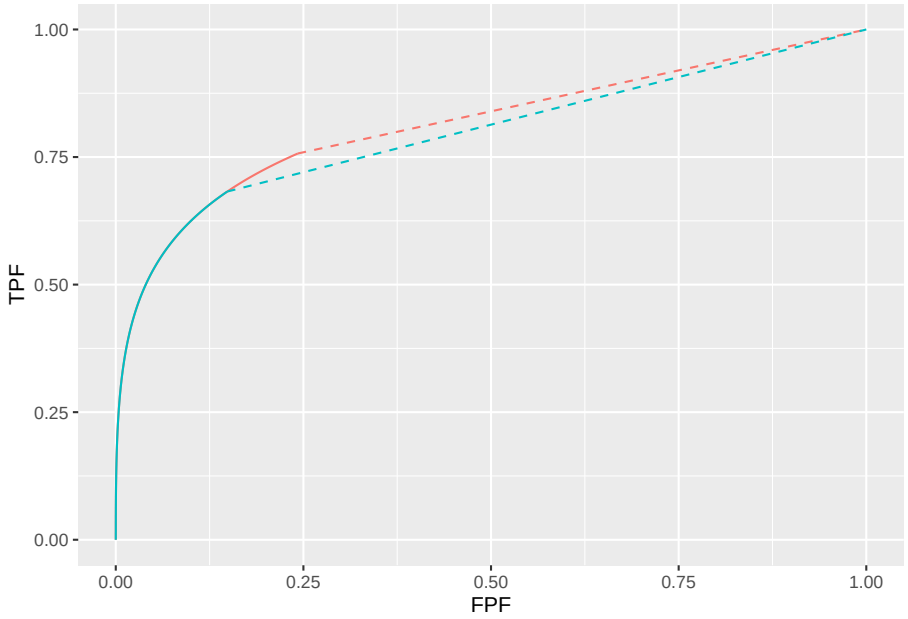


Figure 10.7: The color coding is as in previous figures. The two ROC-AUCs are 0.815 (wAFROC-AUC optimization) and 0.798 (Youden-index optimization).

10.9 TBA Discussion

In Table 10.1 the λ parameter controls the average number of perceived NLs per case. For $\lambda = 1$ there is, on average, one perceived NL for every non-diseased case and the optimal wAFROC-based threshold is TBA $\zeta_{1;1,\mu,\lambda=1,\nu}$ = -0.007. For $\lambda = 10$ there are ten perceived NLs for every non-diseased case and the optimal wAFROC-based threshold is $\zeta_{1;1,\mu,\lambda=10,\nu}$ = -0.007. The increase in ζ_1 should make sense to CAD algorithm designers: with increasing numbers of NLs per case it is necessary to increase the reporting threshold (i.e., adopt a stricter criteria) if only because otherwise the reader would be subjected to 10 times the number of NLs/case for the same number of LLs/case.

The ROC-AUCs are reported as a check of the less familiar wAFROC-AUC figure of merit. The ordering of the two optimization methods is independent of whether it is measured via the wAFROC-AUC or the ROC-AUC: either way the wAFROC-AUC optimizations yield higher AUC values and higher operating points on the FROC than the corresponding Youden-index optimizations.

In this example the difference in wAFROC-AUC, ROC-AUC and the operating points between the two methods decreases as performance *increases*, which is the opposite of that found when λ or ν were varied. With constant λ and ν the *numbers* of NLs and LLs are unchanging; all that happens is the *values* of the z-samples from LLs increase as μ increases, which allows the optimal threshold to increase (this can be understood as a pure "ROC-type" effect: as the normal distributions are more widely separated, the optimal threshold will increase, approaching, in the limit, half the separation, since in that limit TPF = 1 and FPF = 0).

This is due to two reinforcing effects: performance goes down with increasing numbers of NLs per case and performance goes down with increasing optimal reporting threshold (see Section 5.9 for explanation of the ζ_1 dependence of AUC performance). It is difficult to unambiguously infer performance based on the FROC operating points: as λ increases LLF decreases but for $f = 1$ NLF peaks while for $f = 2$ it increases.

The FROC plots also illustrate the decrease in LLF (f, μ, λ, ν) with increasing λ: the black dots move to smaller ordinates, as do the red dots, which would seem to imply decreasing performance. However, the accompanying change in NLF (f, μ, λ, ν) rules out an unambiguous determination of the direction of the change in overall performance based on the FROC.

TBA For very low performance, defined as $0.5 <$ ROC-AUC < 0.6, the Youden-index method chooses a lower threshold compared to wAFROC optimization, resulting in a higher operating point on the FROC, greater ROC-AUC but sharply lower wAFROC-AUC. The difference between the two methods increases as ROC-AUC $\rightarrow 0.5$. In this limit the wAFROC method severely limits the numbers of marks shown to the radiologist as compared to the Youden-index based method.

10.10 References

Chapter 11

Optimal operating point appendices

11.1 Appendix I: Varying ν optimizations

For $\mu = 2$ and $\lambda = 1$ optimizations were performed for $\nu = 0.6, 0.7, 0.8, 0.9$.

```
muArr <- c(2)
lambdaPArr <- c(1)
nuPArr <- c(0.6, 0.7, 0.8, 0.9)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.1.1 Summary table

Table 11.1: Summary of optimization results for $\mu = 2$, $\lambda = 1$ and 4 values of ν.

FOM	ν	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	0.6	0.888	0.701	0.804	(0.187, 0.520)
	0.7	0.674	0.751	0.851	(0.250, 0.635)
	0.8	0.407	0.805	0.893	(0.342, 0.756)
	0.9	-0.007	0.864	0.929	(0.503, 0.880)
Youden	0.6	1.022	0.700	0.797	(0.153, 0.502)
	0.7	1.044	0.745	0.835	(0.148, 0.581)
	0.8	1.069	0.788	0.868	(0.143, 0.659)
	0.9	1.095	0.831	0.899	(0.137, 0.735)

Table 11.1 summarizes the results.

1. For wAFROC-AUC FOM as ν increases the optimal threshold *decreases* and both $\text{wAFROC}(1, \mu, \lambda, \nu)$ and $\text{ROC}(1, \mu, \lambda, \nu)$ *increase*. CAD performance increases, regardless of how it is measured. Performance increases with increasing numbers of LLs per case and this effect is reinforced by performance going up with decreasing optimal reporting threshold. [Since both $\text{LLF}(f, \mu, \lambda, \nu)$ and $\text{NLF}(f, \mu, \lambda, \nu)$ increase with increasing ν, neither FROC-curve based measure has an unambiguous interpretation.

2. The wAFROC based based optimal thresholds are smaller than the corresponding Youden-index based optimal thresholds, i.e., $\zeta_1(1, \mu, \lambda, \nu) < \zeta_1(2, \mu, \lambda, \nu)$. A smaller threshold corresponds to a less strict reporting criterion.

3. For fixed μ, λ, ν the operating point on the FROC for $f = 2$ is below that corresponding to $f = 1$:

   - $\text{NLF}(2, \mu, \lambda, \nu) < \text{NLF}(1, \mu, \lambda, \nu)$ and
   - $\text{LLF}(2, \mu, \lambda, \nu) < \text{LLF}(1, \mu, \lambda, \nu)$.
   - The difference increases with increasing ν.
   - These effects are illustrated in Fig. 11.1.

4. For fixed μ, λ, ν the Youden-index based optimization yields lesser performance than the corresponding wAFROC-AUC based optimization:

- $\text{wAFROC}\,(2, \mu, \lambda, \nu) < \text{wAFROC}\,(1, \mu, \lambda, \nu)$ and
- $\text{ROC}\,(2, \mu, \lambda, \nu) < \text{ROC}\,(1, \mu, \lambda, \nu)$.
- The difference decreases with decreasing ν.
- These effects are illustrated in Fig. 11.2.

11.1.2 FROC

The third effect is illustrated by the FROC plots with superimposed operating points for varying ν shown in Fig. 11.1. The black dots are consistently above the red dots and the separation of the dots is greatest for $\nu = 0.9$ and smallest for $\nu = 0.6$. The difference in optimal thresholds found by the two optimization methods is greatest for poor performance.

The FROC plots also illustrate the decrease in $\text{LLF}\,(f, \mu, \lambda, \nu)$ with increasing ν (the black dots move to larger ordinates, as do the red dots). However, the accompanying change in $\text{NLF}\,(f, \mu, \lambda, \nu)$ rules out an FROC curve based unambiguous determination of the direction of the change in overall performance.

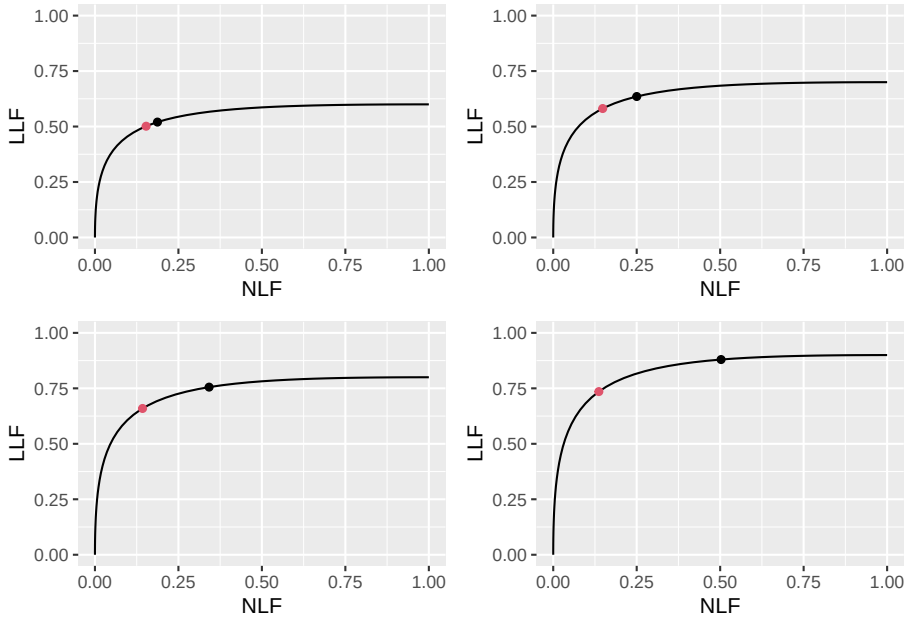


Figure 11.1: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

11.1.3 wAFROC

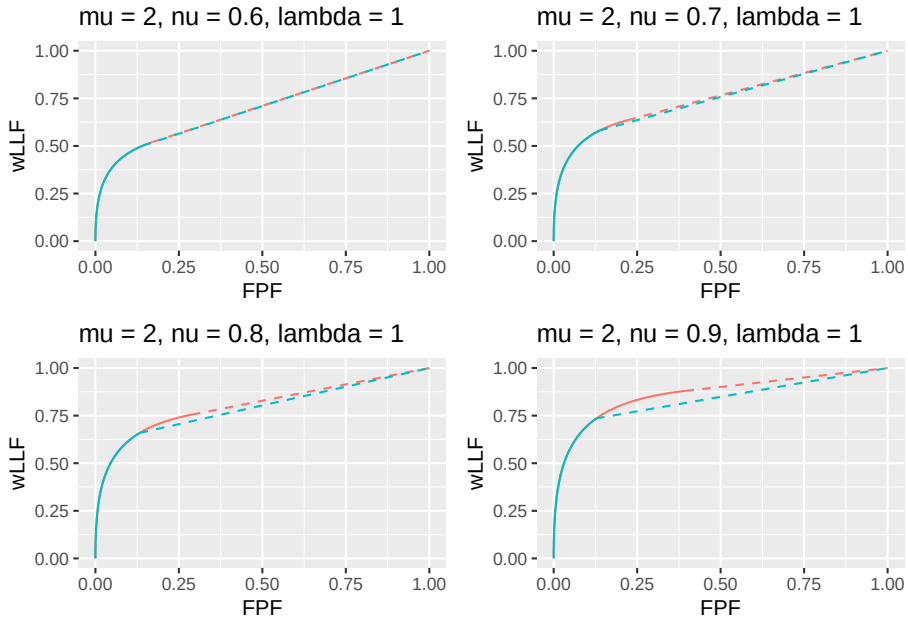


Figure 11.2: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

11.1.4 ROC

11.2 Appendix II: Varying μ optimizations

For $\lambda = 1$ and $\nu = 0.9$ optimizations were performed for $\mu = 1, 2, 3, 4$.

```
muArr <- c(1, 2, 3, 4)
lambdaPArr <- 1
nuPArr <- 0.9
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

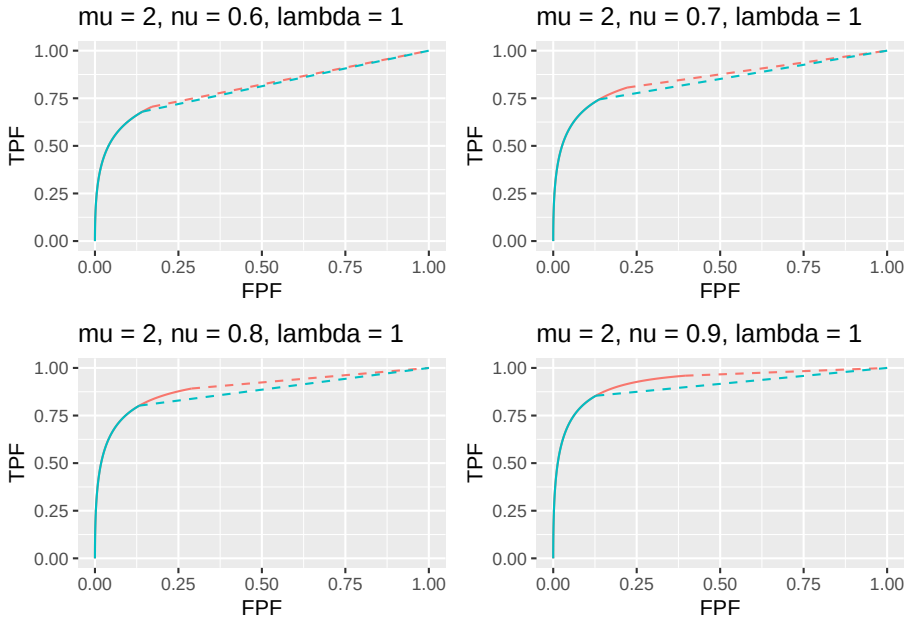


Figure 11.3: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.2.1 Summary table

Table 11.2: Summary of optimization results for 4 values of μ, $\lambda = 1$, $\nu = 0.9$.

FOM	μ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	1	-1.663	0.745	0.850	(0.952, 0.897)
	2	-0.007	0.864	0.929	(0.503, 0.880)
	3	0.808	0.922	0.961	(0.210, 0.887)
	4	1.463	0.942	0.970	(0.072, 0.895)
Youden	1	0.462	0.704	0.815	(0.322, 0.634)
	2	1.095	0.831	0.899	(0.137, 0.735)
	3	1.629	0.903	0.945	(0.052, 0.823)
	4	2.124	0.935	0.964	(0.017, 0.873)

Table 11.2 summarizes the results.

1. For either FOM as μ increases the optimal threshold *increases* and both wAFROC(f, μ, λ, ν) and ROC(f, μ, λ, ν) *increase*. CAD performance increases, regardless of how it is measured. Performance increases with increasing separation of the sampling distributions of NLs and LLs and the negative effect of increasing optimal reporting thresholds is not enough to overcome this. [Since LLF(f, μ, λ, ν) is relatively constant while NLF(f, μ, λ, ν) decreases sharply with increasing μ, this is one example where an FROC-curve based measure does have an unambiguous interpretation, namely performance is higher for the larger values of μ.

2. The wAFROC based based optimal thresholds are smaller than the corresponding Youden-index based optimal thresholds. A smaller threshold corresponds to a less strict reporting criterion and greater wAFROC-AUC and ROC-AUC performance.

3. For fixed μ, λ, ν the operating point on the FROC for $f = 2$ is below that corresponding to $f = 1$. The difference decreases with increasing μ. These effects are illustrated in Fig. 11.4. The black dots are consistently above the red dots and the separation of the dots is greatest for $\mu = 1$ and smallest for $\mu = 4$.

4. For fixed μ, λ, ν the Youden-index based optimization yields lesser performance than the corresponding wAFROC-AUC based optimization. The

difference decreases with increasing μ. These effects are illustrated in Fig. 11.5.

11.2.2 FROC

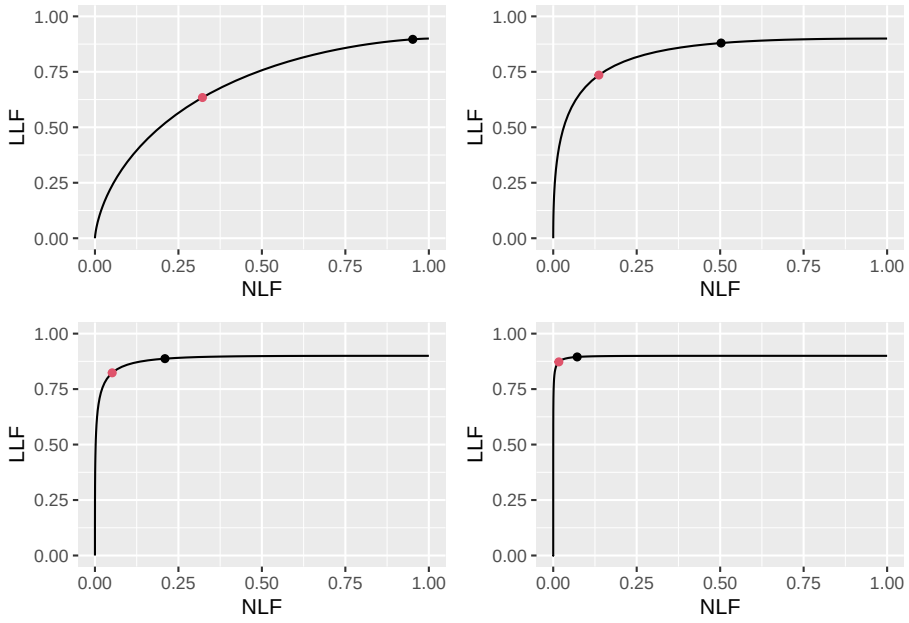


Figure 11.4: FROC plots with superimposed operating points for varying μ. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

11.2.3 wAFROC

TBA The continuous section of each curve ends at the optimal threshold listed in Table 11.2, namely ζ_1 = -1.663 for the green-red-red-dashed curve and ζ_1 = 0.462 for the green curve. The lower performance represented by the green curve, based on Youden-index maximization, is due to the adoption of an overly strict threshold.

11.2.4 ROC

The continuous section of each curve ends at the optimal threshold listed in Table 11.2. The lower performance represented by the green curve, based on

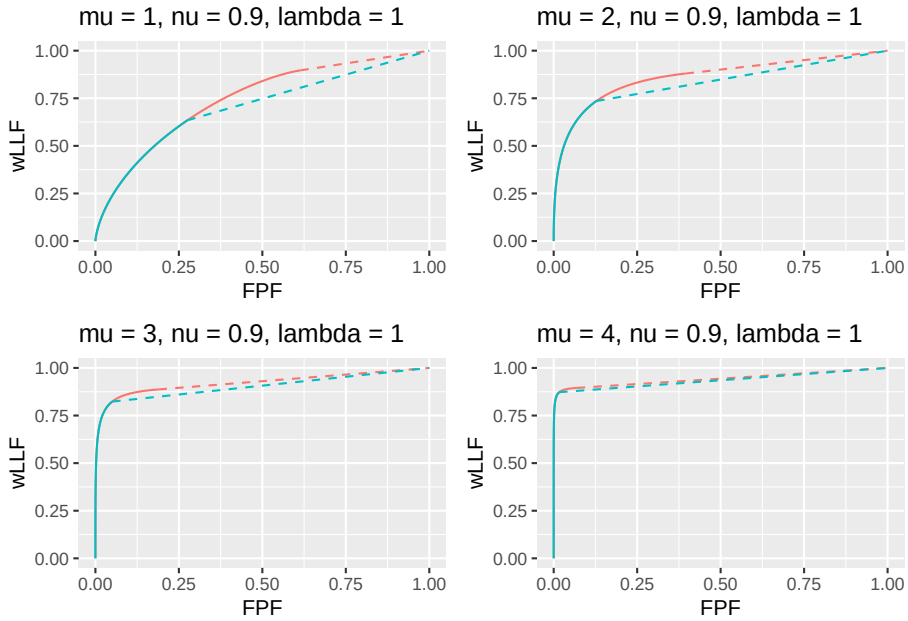


Figure 11.5: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

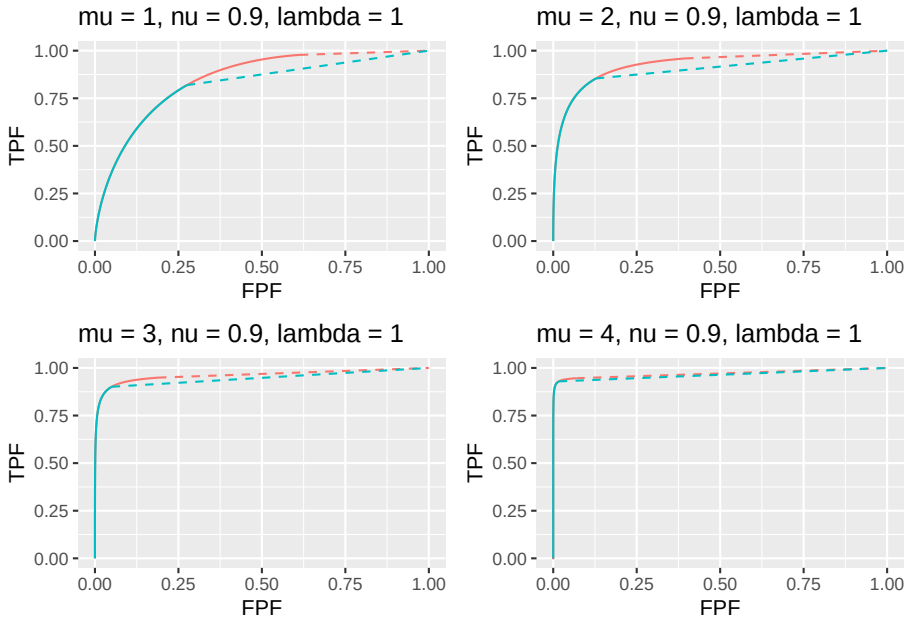


Figure 11.6: ROC plots for the two optimization methods. The color coding is as in previous figures.

Youden-index maximization, is due to the adoption of an overly strict threshold.

11.3 Appendix III: Limiting situations

11.3.1 High performance vary mu

```
muArr <- c(2, 3, 4, 5)
nuPArr <- c(0.9)
lambdaPArr <- c(1)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.1.1 Summary table

Table 11.3: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $nu = 0.9$. Row labeling as in previous tables.

FOM	μ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	2	-0.007	0.864	0.929	(0.503, 0.880)
	3	0.808	0.922	0.961	(0.210, 0.887)
	4	1.463	0.942	0.970	(0.072, 0.895)
	5	2.063	0.948	0.972	(0.020, 0.899)
Youden	2	1.095	0.831	0.899	(0.137, 0.735)
	3	1.629	0.903	0.945	(0.052, 0.823)
	4	2.124	0.935	0.964	(0.017, 0.873)
	5	2.608	0.946	0.970	(0.005, 0.892)

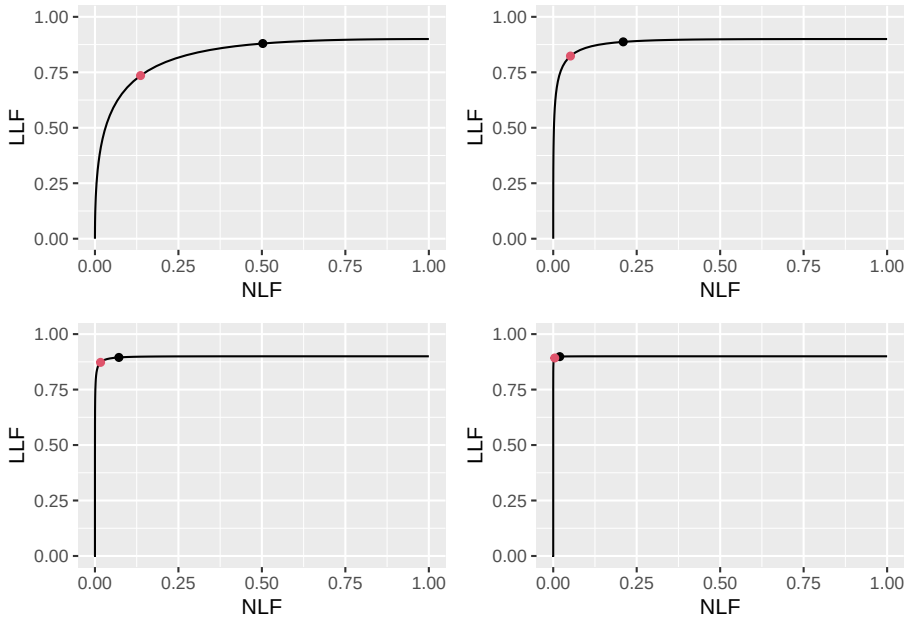


Figure 11.7: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

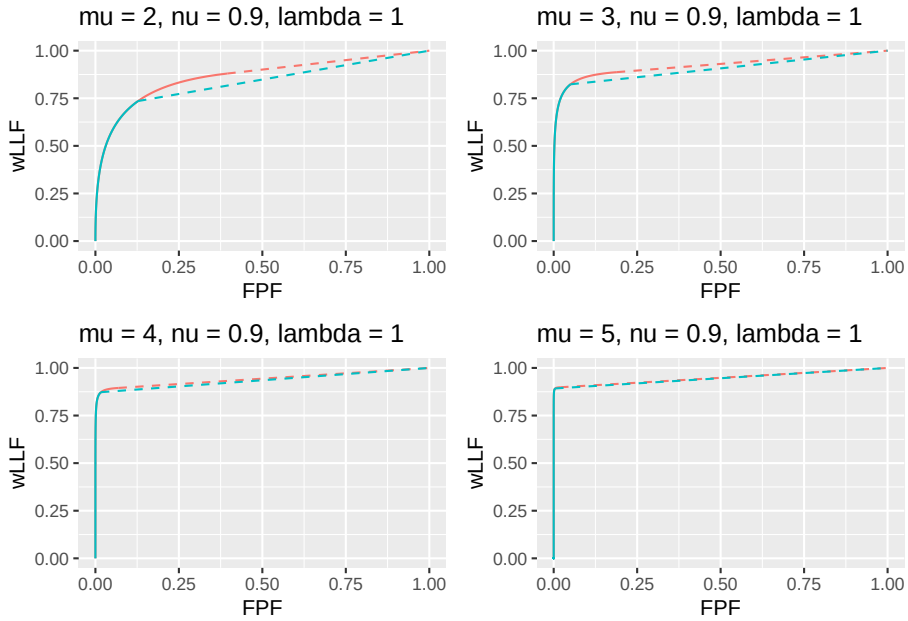


Figure 11.8: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

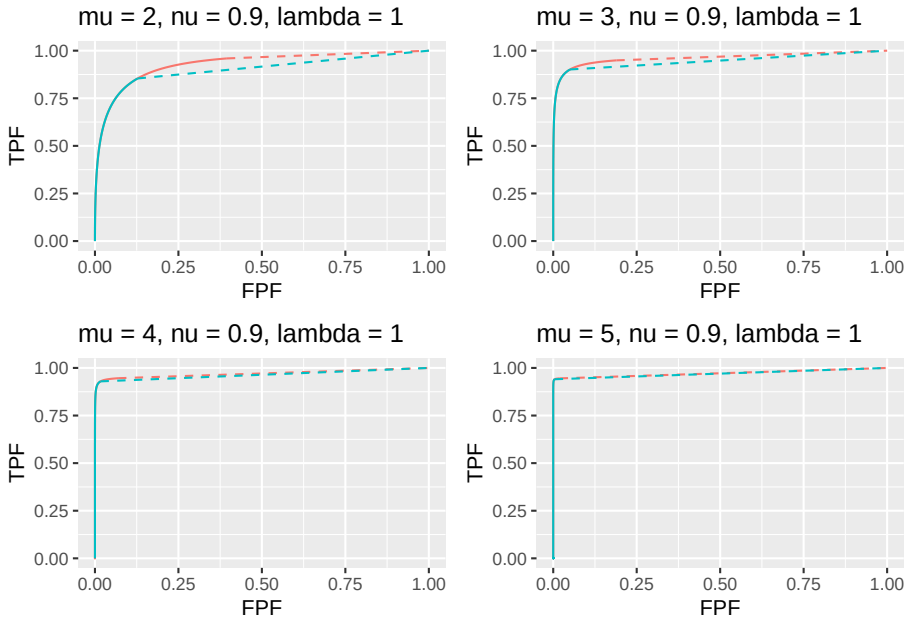


Figure 11.9: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.1.2 FROC

11.3.1.3 wAFROC

11.3.1.4 ROC

11.3.2 Low performance vary mu

```
muArr <- c(1, 2, 3, 4)
nuPArr <- c(0.1)
lambdaPArr <- c(10)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.2.1 Summary table

Table 11.4: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $nu = 0.9$. Row labeling as in previous tables.

FOM	μ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	1	5.000	0.500	0.500	(0.000, 0.000)
	2	3.298	0.502	0.507	(0.005, 0.010)
	3	3.018	0.518	0.536	(0.013, 0.049)
	4	3.130	0.536	0.559	(0.009, 0.081)
Youden	1	1.563	0.292	0.514	(0.590, 0.029)
	2	1.865	0.397	0.535	(0.311, 0.055)
	3	2.198	0.478	0.555	(0.140, 0.079)
	4	2.564	0.523	0.567	(0.052, 0.092)

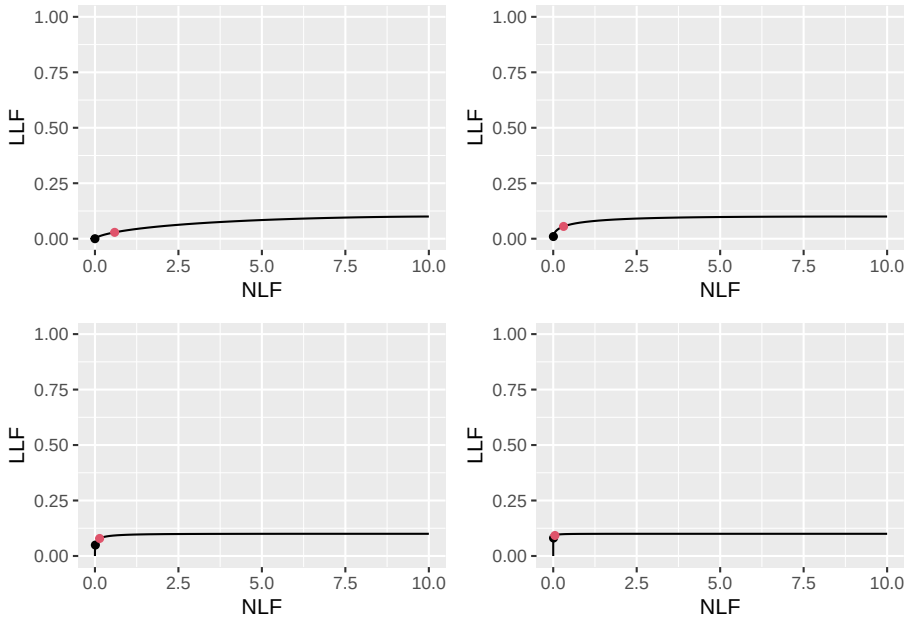


Figure 11.10: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

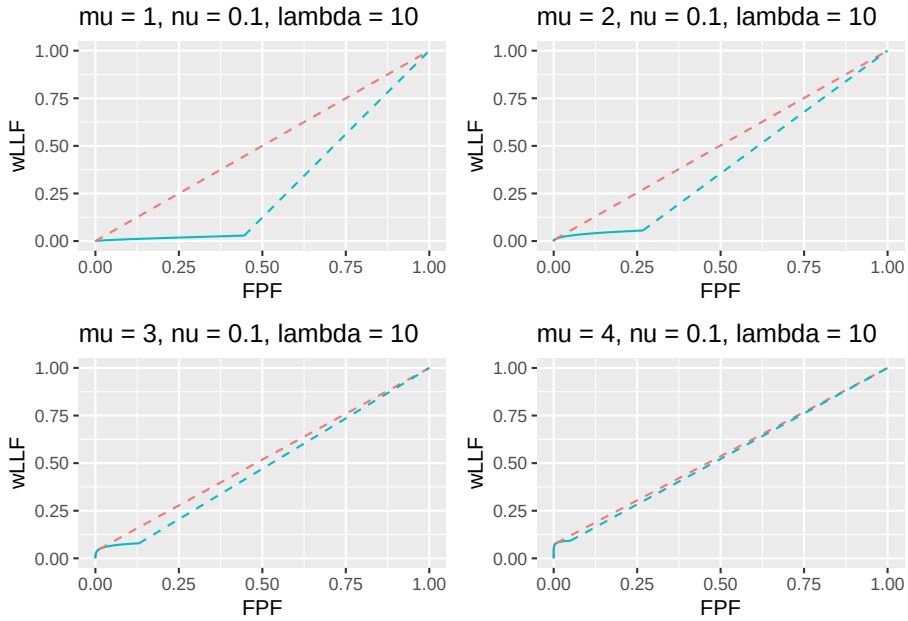


Figure 11.11: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

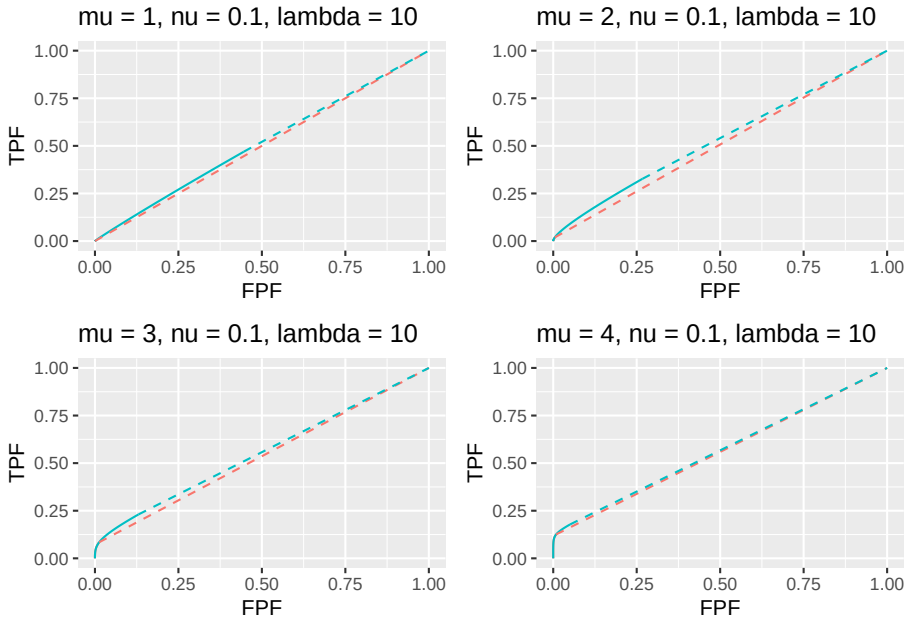


Figure 11.12: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.2.2 FROC

11.3.2.3 wAFROC

11.3.2.4 ROC

11.3.3 High performance vary lambda

```
muArr <- c(4)
nuPArr <- c(0.9)
lambdaPArr <- c(1,2,5,10)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.3.1 Summary table

Table 11.5: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $nu = 0.9$. Row labeling as in previous tables.

FOM	λ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	1	1.463	0.942	0.970	(0.072, 0.895)
	2	1.644	0.938	0.968	(0.100, 0.892)
	5	1.889	0.930	0.965	(0.147, 0.884)
	10	2.082	0.920	0.960	(0.187, 0.875)
Youden	1	2.124	0.935	0.964	(0.017, 0.873)
	2	2.291	0.928	0.960	(0.022, 0.861)
	5	2.508	0.915	0.952	(0.030, 0.839)
	10	2.669	0.903	0.944	(0.038, 0.818)

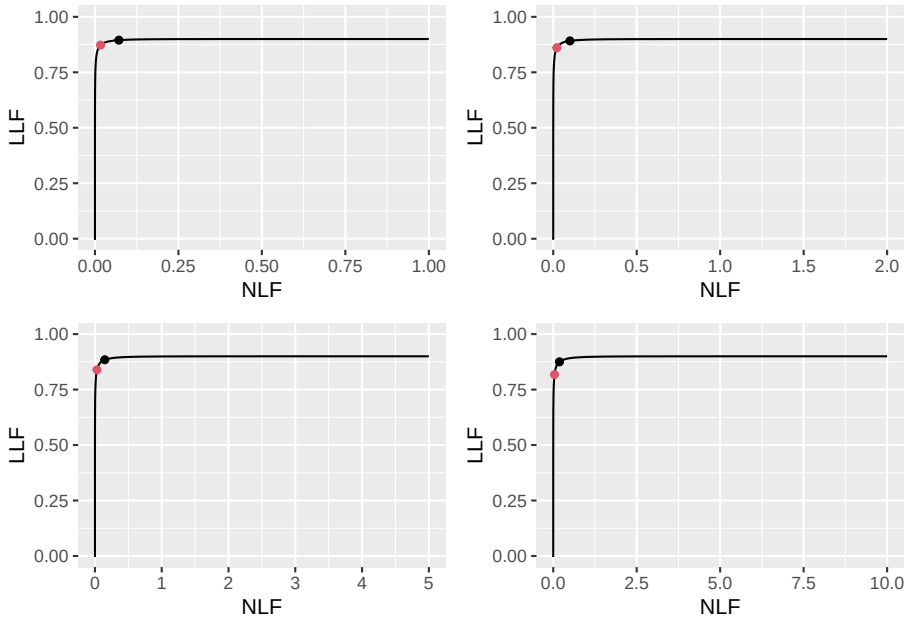


Figure 11.13: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

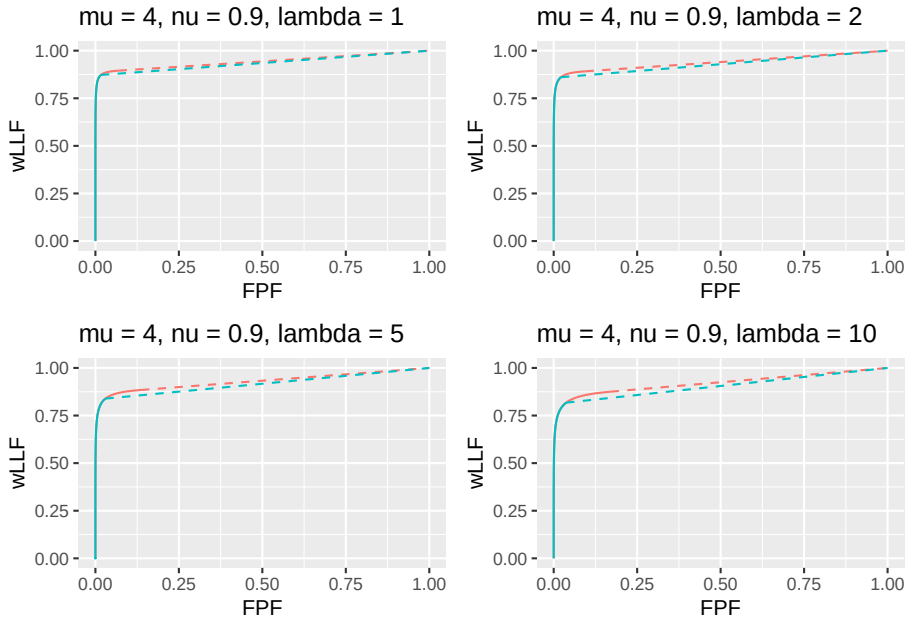


Figure 11.14: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

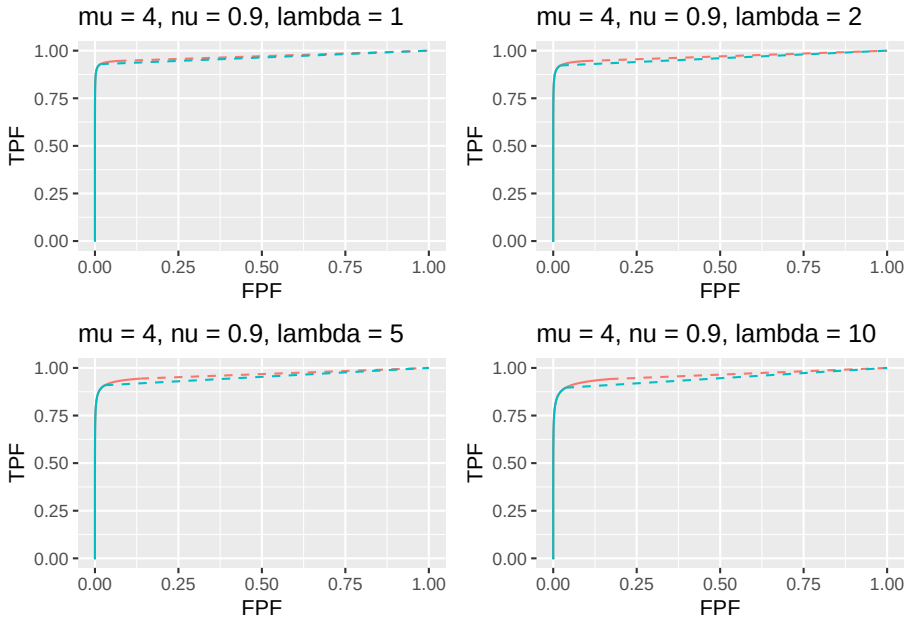


Figure 11.15: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.3.2 FROC

11.3.3.3 wAFROC

11.3.3.4 ROC

11.3.4 Low performance vary lambda

```
muArr <- c(1)
nuPArr <- c(0.2)
lambdaPArr <- c(1, 2, 5, 10)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.4.1 Summary table

Table 11.6: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $nu = 0.9$. Row labeling as in previous tables.

FOM	λ	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	1	2.081	0.505	0.520	(0.019, 0.028)
	2	2.795	0.501	0.505	(0.005, 0.007)
	5	3.718	0.500	0.500	(0.001, 0.001)
	10	4.412	0.500	0.500	(0.000, 0.000)
Youden	1	0.284	0.423	0.587	(0.388, 0.153)
	2	0.734	0.380	0.566	(0.463, 0.121)
	5	1.237	0.335	0.542	(0.540, 0.081)
	10	1.568	0.309	0.528	(0.585, 0.057)

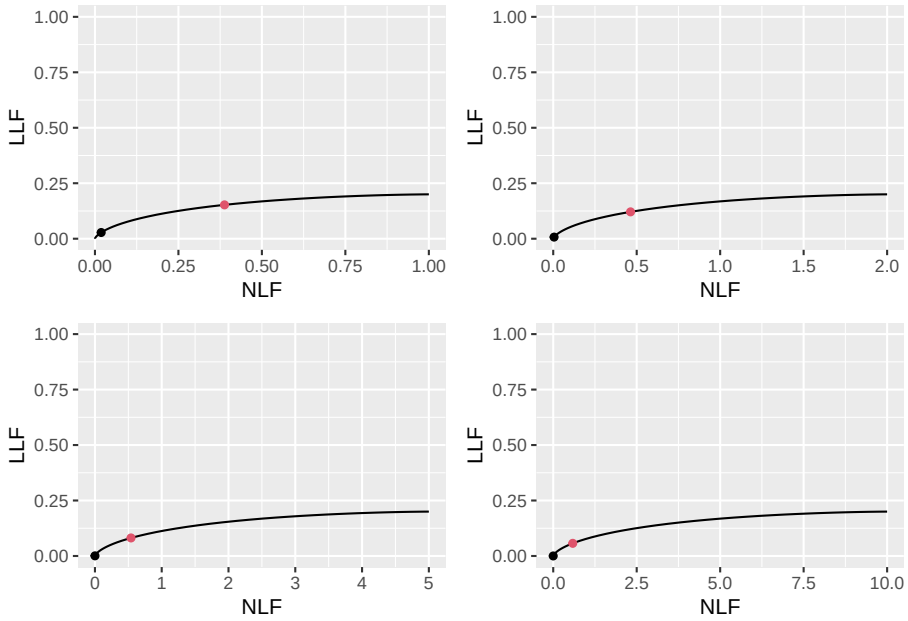


Figure 11.16: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

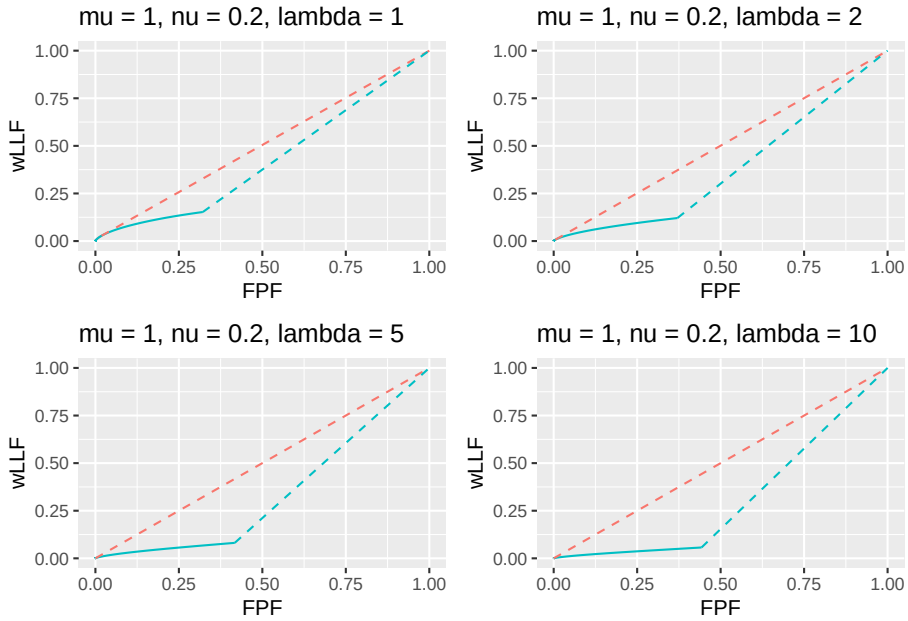


Figure 11.17: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

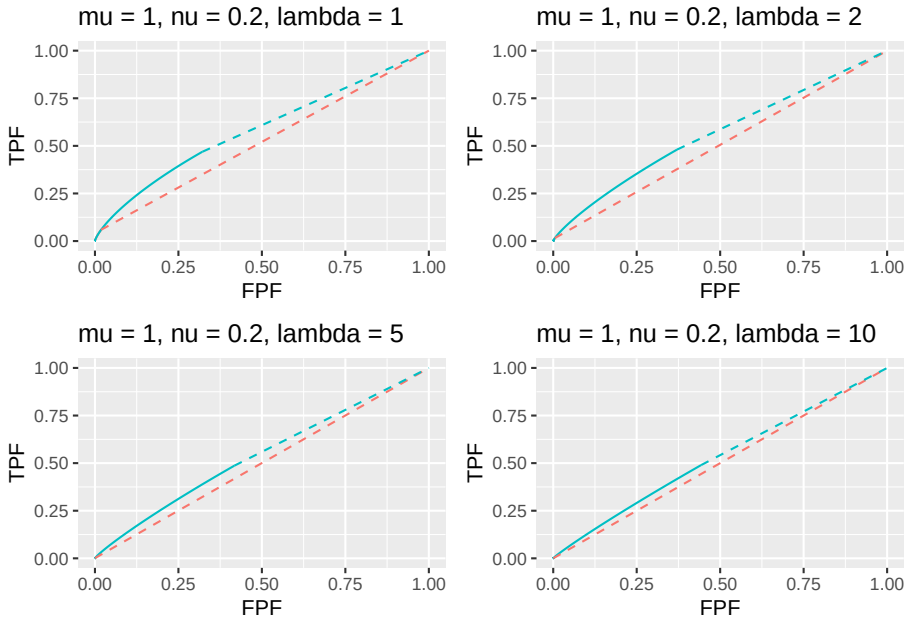


Figure 11.18: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.4.2 FROC

11.3.4.3 wAFROC

11.3.4.4 ROC

11.3.5 High performance vary nu

```
muArr <- c(4)
lambdaPArr <- c(1)
nuPArr <- c(0.6, 0.7, 0.8, 0.9)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.5.1 Summary table

Table 11.7: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $nu = 0.9$. Row labeling as in previous tables.

FOM	ν	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	0.6	1.905	0.788	0.855	(0.028, 0.589)
	0.7	1.796	0.839	0.898	(0.036, 0.690)
	0.8	1.663	0.890	0.936	(0.048, 0.792)
	0.9	1.463	0.942	0.970	(0.072, 0.895)
Youden	0.6	2.063	0.788	0.852	(0.020, 0.584)
	0.7	2.080	0.837	0.894	(0.019, 0.681)
	0.8	2.100	0.886	0.931	(0.018, 0.777)
	0.9	2.124	0.935	0.964	(0.017, 0.873)

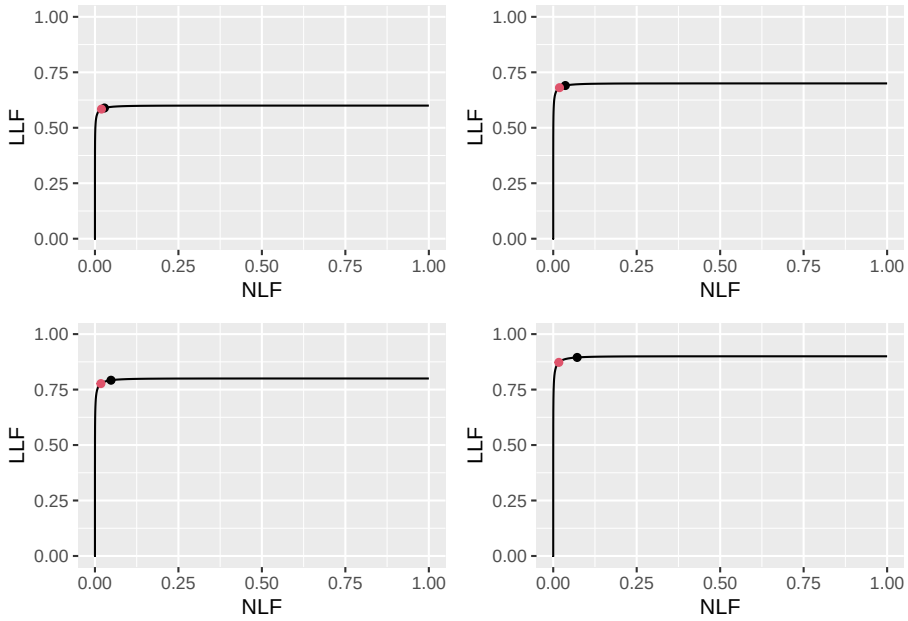


Figure 11.19: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

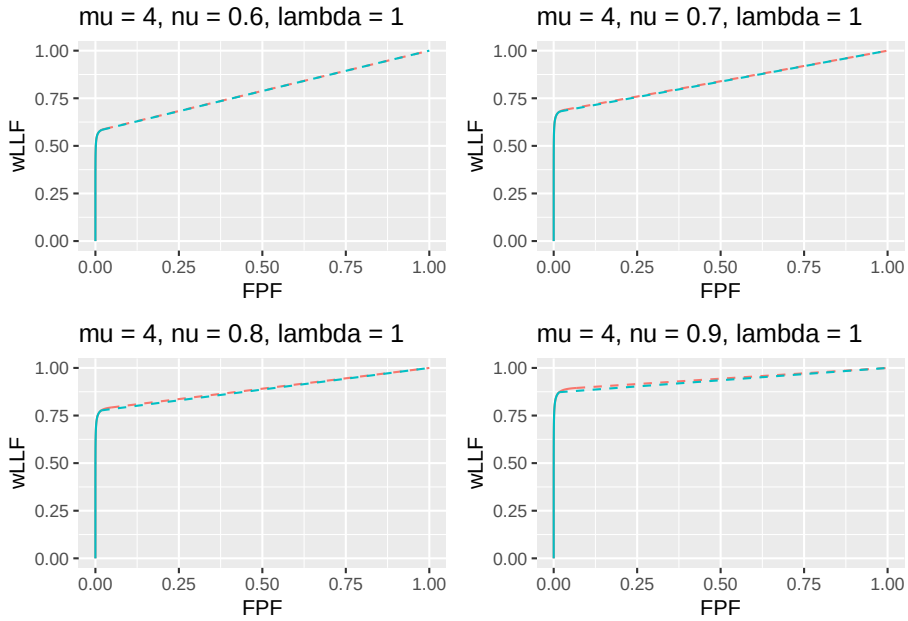


Figure 11.20: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

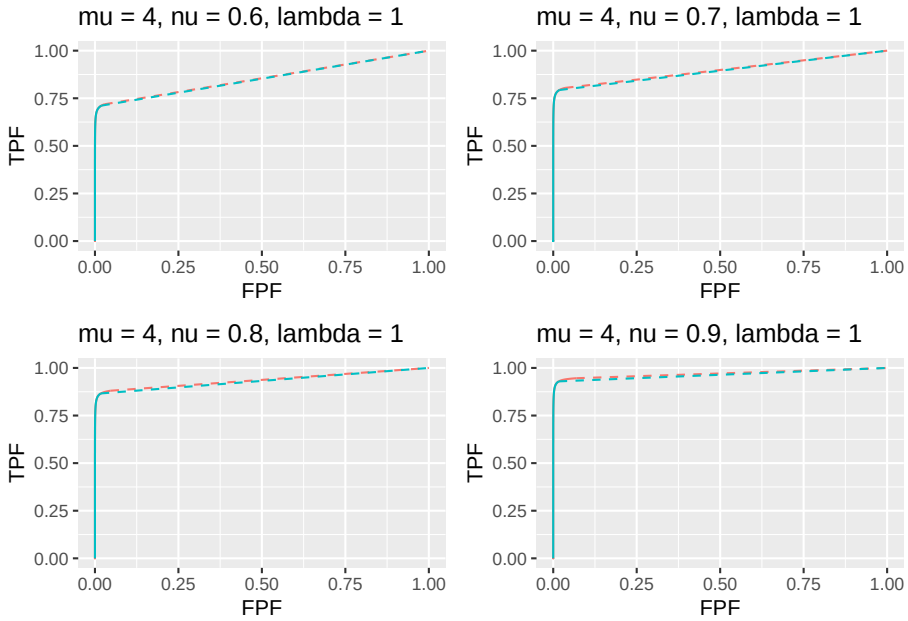


Figure 11.21: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.5.2 FROC

11.3.5.3 wAFROC

11.3.5.4 ROC

11.3.6 Low performance vary nu

```
muArr <- c(1)
lambdaPArr <- c(10)
nuPArr <- c(0.1, 0.2, 0.3, 0.4)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.3.6.1 Summary table

Table 11.8: Summary of optimization results for 4 values of μ, $\lambda = 1$ and $\nu = 0.9$. Row labeling as in previous tables.

FOM	ν	ζ_1	wAFROC	ROC	(NLF, LLF)
wAFROC	0.1	5.000	0.500	0.500	(0.000, 0.000)
	0.2	4.412	0.500	0.500	(0.000, 0.000)
	0.3	4.006	0.500	0.500	(0.000, 0.000)
	0.4	3.718	0.500	0.501	(0.001, 0.001)
Youden	0.1	1.563	0.292	0.514	(0.590, 0.029)
	0.2	1.568	0.309	0.528	(0.585, 0.057)
	0.3	1.572	0.325	0.542	(0.580, 0.085)
	0.4	1.577	0.342	0.556	(0.574, 0.113)

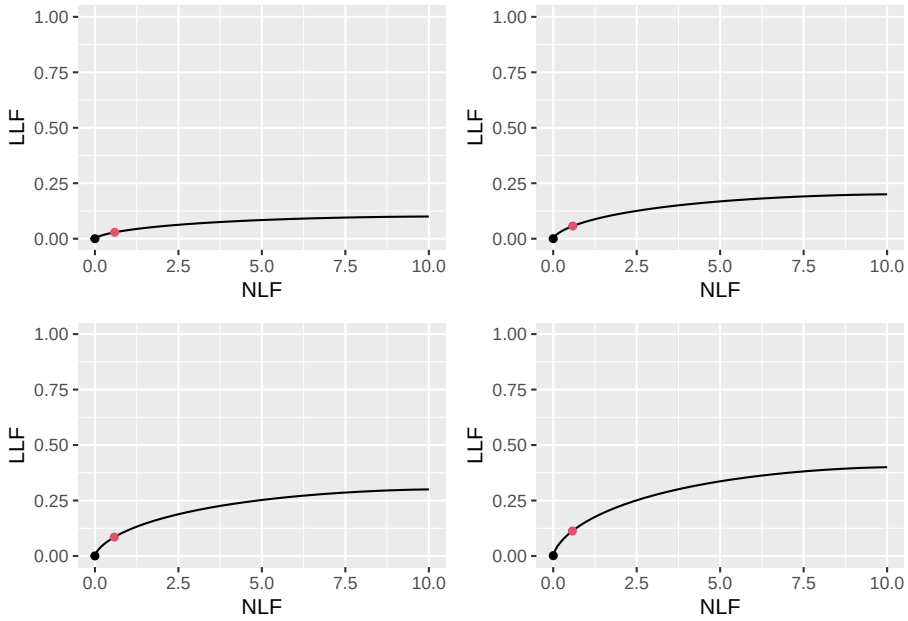


Figure 11.22: FROC plots with superimposed operating points for varying ν. The black dot corresponds to wAFROC AUC optimization and the red dot to Youden-index optimization.

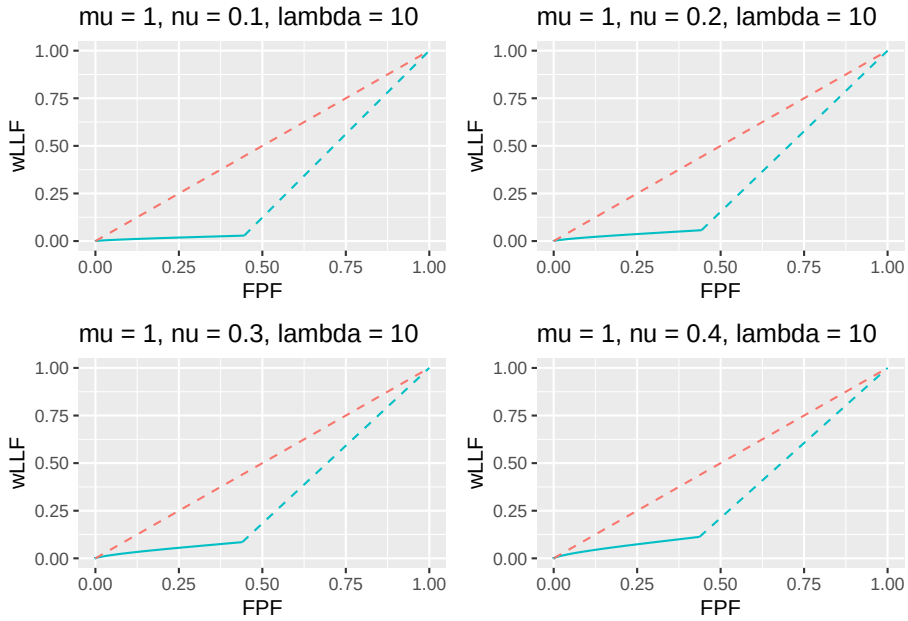


Figure 11.23: wAFROC plots for the two optimization methods. The color coding is as in previous figures.

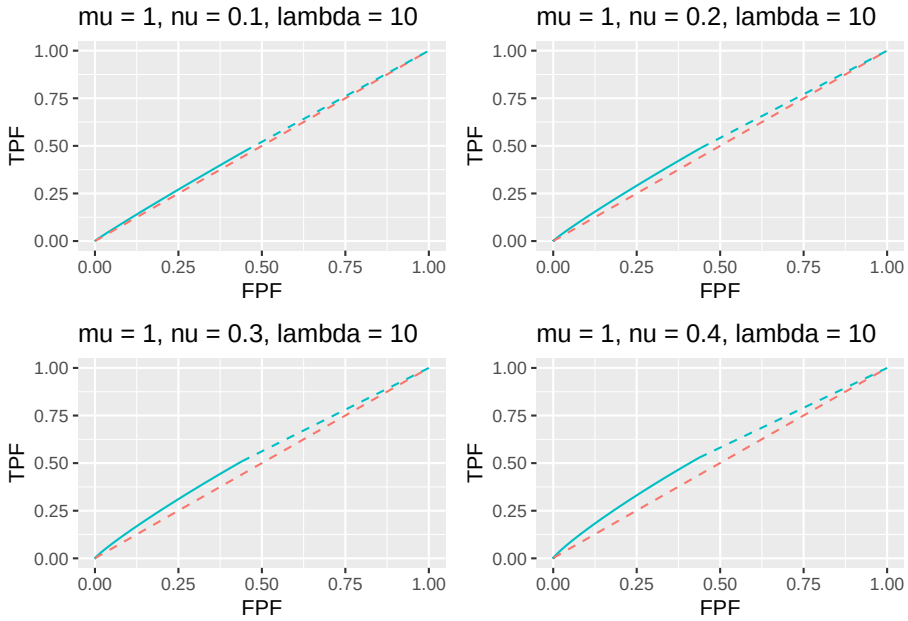


Figure 11.24: ROC plots for the two optimization methods. The color coding is as in previous figures.

11.3.6.2 FROC

11.3.6.3 wAFROC

11.3.6.4 ROC

11.4 References

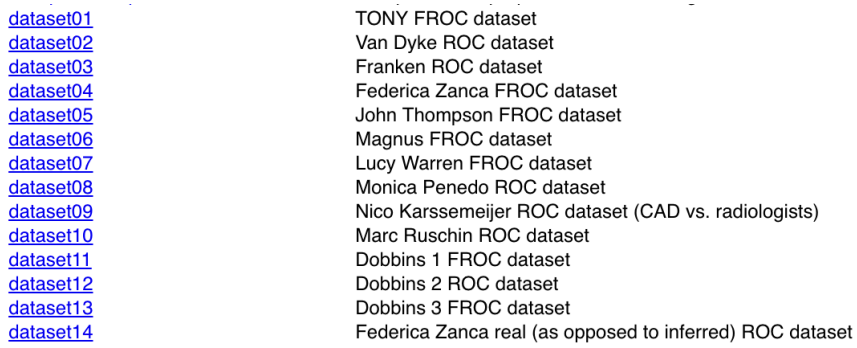
DATASETS

Chapter 12

Datasets

12.1 Datasets

The datasets are embedded in `RJafroc`. They can be viewed in the help file of the package, a partial screen-shot of which is shown next.



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 12.1: Partial screen shot of 'RJafroc' help file showing the datasets included with the current distribution (v2.0.1).

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 ...
##  $ 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
```

12.2 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 (2022). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.25.

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. and Zhai, X. (2022). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.1.1.9000.

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. and Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice*. Springer.

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.

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., 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.

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.