The RJafroc Froc Book

Dev P. Chakraborty, PhD

2023-12-26

Contents

Preamble

Chapter 1

Please ignore preface

TBA

1.1 Please ignore issues

1.2 Please ignore following sections

- They are intended for my convenience and will be deleted in final version

1.3 TBA Rationale and Organization

- 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, 2023).
- Since its publication in 2017 `RJafroc`, on which the `R` code examples in the print book depend, has evolved considerably causing many of the examples to "break" if one uses the most current version of `RJafroc`. The code will still run if one uses `RJafroc` 0.0.1 but this is inconvenient and misses out on many of the software improvements made since the print book appeared.
- 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.

1.4 TBA Acknowledgements

Dr. Xuetong Zhai

Dr. Peter Phillips

Online Latex Editor at

Dataset contributors: Nico especially 10

1.5 TBA Nearly finished chapters

- Chapter 1 The FROC paradigm and search
- Chapter 2 Empirical plots from FROC data
- Chapter 3 Visual Search
- Chapter 4 The radiological search model (RSM)
- Chapter 5 ROC curve implications of the RSM
- Chapter 6 Search and classification performances
- Chapter 7 RSM fitting
- Chapter 8 Three proper ROC fits
- Chapter 9 Standalone CAD vs. Radiologists
- Chapter 10 Optimal operating point
- Chapter 11 Optimal operating point appendices

1.6 The pdf file of the book

Go here and then click on `Download` to get the `RJafrocFrocBook.pdf` file. The pdf version may not be as aesthetically pleasing as the HTML version, in particular the layout of figures and tables is sometimes disjointed from the citing text.

1.7 Please ignore: TBA How much finished HMF

- HMF approximately 30%
- This book is currently (as of August 2022) in preparation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.
- Un-comment links like `\@ref(froc-paradigm-solar-analogy)` etc. Search for `\@ref`

1.8 Please ignore: A note on the online distribution mechanism of the book

- In the hard-copy version of my book (Chakraborty, 2017) the online distribution mechanism was `BitBucket`.
- `BitBucket` allows code sharing within a *closed* group of a few users (e.g., myself and a grad student).
- Since the purpose of open-source code is to encourage collaborations, this was, in hindsight, an unfortunate choice. Moreover, as my experience with R-packages grew, it became apparent that the vast majority of R-packages are shared on `GitHub`, not `BitBucket`.
- For these reasons I have switched to `GitHub`. All previous instructions pertaining to `BitBucket` are obsolete.
- In order to access `GitHub` material one needs to create a (free) `GitHub` account.
- Go to this link and click on `Sign Up`.

1.9 Please ignore: Structure of the book

1.10 Please ignore Contributing to this book

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

1.11 Please ignore: Is this book relevant to you and what are the alternatives?

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

1.12 Please ignore: Chapters needing heavy edits

1.13 Please ignore: Shelved vs. removed vs. parked folders needing heavy edits

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

1.14 Please ignore: Coding aids (for me)

- weird error with knitr not responding to changes in Rmd file: traced to upper case lower case confusion: 13A-empirical1.Rmd which should be 13a-empirical1.Rmd

1.14.1 formatting

- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:

1.14.2 tables

- https://github.com/haozhu233/kableExtra/issues/624
- kbl(dfA, caption = "....", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

1.14.3 tinytex problems

- dont update in response to messages; breaks everything
- DONT DO THIS: When tinytex::install_tinytex() hangs up try
- DONT DO THIS: tinytex::install_tinytex(repository = "http://mirrors.tuna.tsinghua.edu.cn/CTAN/", version = "latest")
- Getting very long builds: looping certain commands
- First uninstall tinytex then reinstall:

```
#uninstall_tinytex(force = FALSE, dir = tinytex_root())
#tinytex::install_tinytex()
```

- get very long build first time with looping certain commands
- fixed on subsequent pdf builds

FROC paradigm

Chapter 2

The FROC paradigm and visual search

2.1 How much finished 100%

2.2 Introduction

For diagnostic tasks such as detecting diffuse interstitial lung disease[1] *where disease location is either irrelevant or implicit*, the receiver operating characteristic (ROC) paradigm is appropriate because 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 to also offer guidance to subsequent care-givers (e.g., the surgeon responsible for resecting a malignant lesion) by identifying other lesion characteristics, e.g., location, type, size and extent.

For localized disease the ROC paradigm limits the collected information to a single rating that categorizes the probability that there is disease *somewhere* in the patient's imaged anatomy. "Somewhere" begs the question: if the radiologist believes the disease is "somewhere", why not have them point to it? In fact they do "point to it" by recording the location(s) of suspicious regions in their clinical report, but the ROC paradigm cannot use the location information.

From the data analyst's point of view the most troubling issue with ROC analysis when applied to a localization task is that neglect of location information leads to loss of statistical power. That this is a problem can be appreciated from the following simple example comparing expert and non-expert radiologists.

> Recall that an ROC paradigm true positive event occurs anytime a diseased patient is diagnosed as diseased: lesion location, if provided, is not considered. Therefore two types of true positive events are possible on diseased cases: those with correct localizations, expected to be associated with expert radiologists, and those with incorrect localizations, expected to be associated with non-experts. The indistinguishability between the two types of true positive events leads to reduced ability to detect a difference between experts and non-experts. The resulting loss of statistical power is highly undesirable since counteracting it would lead to inflated sample size requirements (numbers of readers anc cases) for a contemplated ROC study. Obtaining participating radiologists and finding truth-proven cases are both expensive in radiological observer performance studies [2].

[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.*

[2] Numerical examples of the loss of statistical power of ROC analysis as compared to a method that credits correct localizations are here.

2.3 Location specific paradigms

The term "location-specific" is used for any observer performance paradigm that accounts for lesion location. These paradigms are sometimes incorrectly referred to as lesion-specific (or lesion-level) paradigms. All observer performance methods involve detecting the presence of true lesions and ROC methodology is, in this sense, also lesion-specific. On the other hand *location* is a characteristic of true and perceived lesions, and methods that account for location are more accurately termed *location-specific* than lesion-specific.

There are three location-specific paradigms that take into account, to varying degrees, information regarding the locations of perceived lesions:

- 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).

Together with the ROC paradigm they constitute four currently-used observer performance paradigms.

> In this book *lesion* always refers to a true or real lesion. The term *suspicious region* or *perceived lesion* is used 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. 2.1 show a schematic mammogram interpreted according to these paradigms. The panels are as follows:

- upper left – ROC,
- upper right – FROC,
- lower left – LROC,
- lower right – ROI.

With reference to the top-left panel, the arrows point to two lesions and the three light-shaded crosses indicate suspicious regions [3]. A marked suspicious region is indicated by a dark-shaded cross. Evidently the radiologist perceived one of the lesions (the light-shaded cross near the left most arrow), missed the other lesion and mistook two normal structures for lesions, the two light-shaded crosses that are far from any of the lesions. In this example there are three suspicious regions, one of which is close to a real lesion, and one missed lesion.

- In the ROC paradigm, Fig. 2.1 (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 and if the left-most light-shaded cross is a highly suspicious region then the ROC rating for the image might be 5 (highest confidence for presence of disease somewhere in the image). There are no dark-shaded crosses on this panel as no marking occurs in the ROC paradigm.
- In the free-response (FROC) paradigm, Fig. 2.1 (top-right panel), the two dark-shaded crosses indicate suspicious regions that were *marked*, and the adjacent numbers are the corresponding ratings. *Unlike the ROC paradigm where the rating applies to the whole image, in this example each rating applies to a specific suspicious region.* 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 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.
- In the LROC paradigm, Fig. 2.1 (bottom-left panel), the radiologist rates the confidence that there is at least one lesion somewhere in the image (just as in the ROC paradigm) and then marks the *most suspicious* region. In this example the rating is LROC-5, the five rating is the same as in the ROC paradigm panel, and

[3] These were obtained using an eye-tracking apparatus

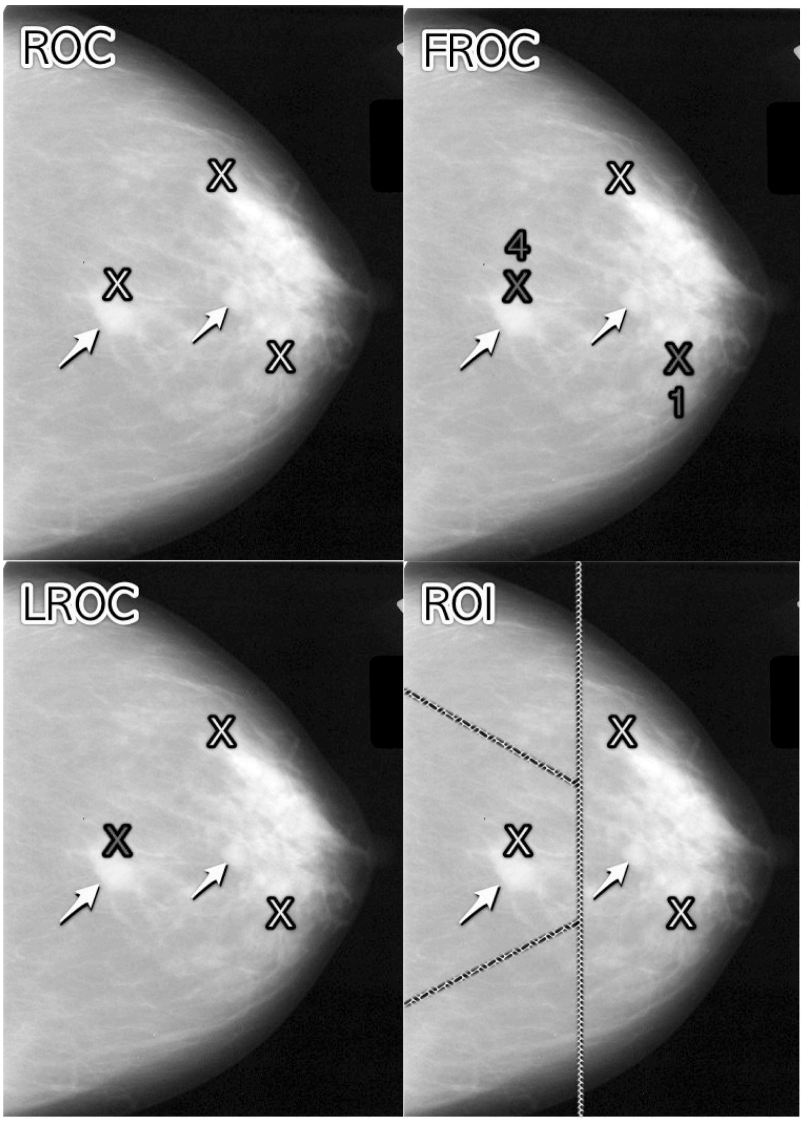


Figure 2.1: Schematic showing the four observer performance paradigms. Arrows point to two lesions and light-shaded crosses indicate suspicious regions. Marked suspicious regions are indicated by dark-shaded crosses.

the mark is the same mark rated FROC-4 in the FROC paradigm panel. Since it is close to a true lesion in LROC terminology this mark 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 per image is allowed, 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, Fig. 2.1 (bottom-right panel) 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 the ROI at ~9 o'clock might be rated ROI-5 as it contains the most suspicious light-shaded cross, 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[4].

Why does the integer FROC rating scale extend from 1 to 4 while the remaining paradigm scales range from 1 to 5? The **absence* of any marked region in FROC conveys information that the case had no reportable suspicious region. In the other paradigms this would need to be indicated using the 1-rating.

2.4 Visual search

Any search task has two components: finding things while not finding irrelevant things, a subtle but important point, and acting on each finding. Two 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 without finding too many extraneous objects. Acting on the finding could be driving to work or drinking milk from the carton. There is expertise associated with any search task. Husbands are notoriously bad at finding the milk carton in the refrigerator (analogy due to Dr. Krupinski at an SPIE course taught jointly with the author).

Likewise, a medical imaging search task has two components: finding lesions and acting on each finding. "Finding" is the actual term used by radiologists in their clinical reports. Acting on the finding involves determining if it is sufficiently suspicious for cancer to warrant reporting and further patient follow-up. Such a region is marked and rated for confidence that it is a malignant lesion.

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[5]. 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 (typically integers 1 through 5) 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 1-5 BIRADS rating) or BIRADS-1 for normal or no abnormality detected (the patient comes back about a year later for the next screening examination).

The FROC paradigm in medical imaging is a visual search task.

[4]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.

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

2.4.1 Proximity criterion and scoring the data

In the first quasi-clinical application of the FROC paradigm (Chakraborty 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[6], was given only if a mark was sufficiently close (as per the adopted 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:

- 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

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. Clinical considerations should be used to determine if a mark actually localized the lesion with sufficient accuracy. 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 in a study involving both volumetric (CT) and 2D chest images.[7]

2.4.2 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. [8] 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. [9]

2.4.3 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 noise. The tone(s) could occur at any instant within an active listening interval,

[6]The 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 paradigm term that does not involve search and localization with one that does.

[7]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.

[8]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.

[9]The highest rating gives full and deserved credit for the correct 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.

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 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 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 analogous to medical imaging search tasks.

2.5 The 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 coordinates are denoted $(\mathrm{NLF}_{\max}, \mathrm{LLF}_{\max})$.

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

If *integer ratings* are used 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. [10].

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 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.

Empirical plot:

> A plot is termed *empirical* if is based on the observed operating points: one simply connects adjacent operating points (including the origin) with straight lines.

Chapter 3 describes the empirical FROC and other empirical operating characteristics in more detail.

[10] 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 FROC data collection interface should present an explicit "next-image" option and reserve the "1" rating to mean the lowest reportable confidence level.

2.5.1 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 one could generate 20 plots. In this example I have selected treatment 1 and reader 1.

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

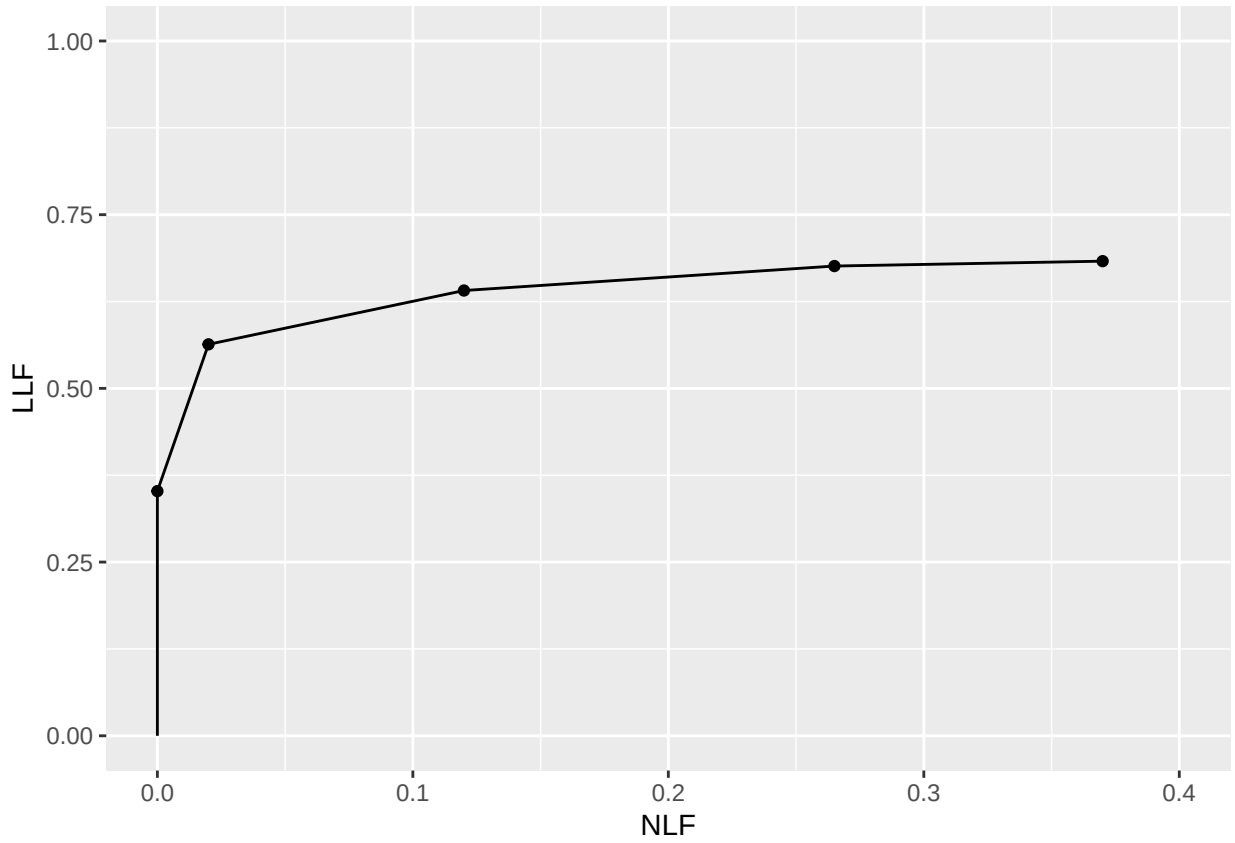


Figure 2.2: Empirical FROC plot for 'dataset04', treatment 1 and reader 1.

The study in question was a 5 rating FROC study. The lowest non-trivial point (i.e., not counting the origin which is common to all FROC plots) corresponds to 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, namely the operating point cannot move downward or to the left as one cumulates lower confidence level marks.

The above plot is termed an *empirical plot* as it consists of the empirical (observed) operating points connected by straight line segments. A model based plot would be termed a *predicted plot*.

2.6 The Astronomical 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" as being the sun.

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 achieve $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$. LLF and NLF subscripted "max" because by randomly choosing to *not mark* the position of the sun, even though it is visible, the observer can "walk down" the y-axis of the FROC plot, eventually reaching $LLF = 0$ and $NLF = 0$, demonstrating 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 will still correctly locate 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 peripheral vision, so that even if the observer did not start out looking in the correct direction, peripheral vision will guide the observer's gaze to the correct location.

The implication of this is that a fundamentally different mechanism is involved from that considered in conventional (i.e., ROC) observer performance methodology, namely *search*.

> Search describes the process of *finding* lesions while *not finding* non-lesions; search performance is the ability to find lesions while not finding non-lesions.

Think of the eye as two cameras: a low-resolution camera (peripheral vision) with a wide field-of-view plus a high-resolution camera (foveal vision) with a narrow field-of-view. If one were limited to viewing with the high-resolution camera one would spend so much time steering the high-resolution narrow field-of-view camera from spot-to-spot that one would have a hard time finding the desired stellar object. Having a single high-resolution narrow field of view vision would also have negative evolutionary consequences as one would spend so much time scanning and processing the surroundings 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 surroundings 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 appropriate location. 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. 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 $\text{LLF}_{\text{max}} = 0$ and $\text{NLF}_{\text{max}} = 0$, i.e., the operating point is "stuck" at the origin. On the other hand if 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 signal-to-noise-ratio (SNR) object and therefore atypical. Consider finding stars or planets. In clear skies 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 TBA the observer will not fill up the sky with false sightings of a desired star. Why?

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 or incorrect localizations.

2.7 Implications for models of visual search

[This section will make more sense after reading Chapter 6 on the Radiological Search Model (RSM).]

The Astronomical Analogy elucidates some crucial features of visual search. In the medical imaging context visual search is defined as finding lesion(s) given that their locations are unknown while minimizing finding non-lesions. As shown in the previous section, if lesion contrast is high then the observer's visual system will guide the eye to the correct location(s). The result is that incorrect localizations and missed lesions will rarely occur. In terms of the RSM, as lesion contrast, i.e., μ, increases the number of non-lesion localizations, i.e., λ, decreases, and the fraction of detected lesions, i.e., ν, approaches unity. This tandem behavior will be accounted for in the formulation of the RSM, in particular the distinction made in Chapter 6 between physical and intrinsic RSM λ and ν parameters, see in particular Sections 6.5 and 6.6.

2.8 Discussion

The FROC paradigm has been confused by loose terminology and misconceptions about visual search, the FROC paradigm and the FROC curve. Some examples follow

- Loose terminology:
  - 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.
  - 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.
- Misconceptions:
  - A fundamental misunderstanding of search performance is embodied in the statement *"CAD is perfect at search because it looks at everything"*.
  - Showing FROC curves as reaching the unit ordinate – which is the exception rather than the rule.
  - The belief that FROC curves extend to very large (potentially infinite) values along the abscissa and all the observer has to do to access this region is to lower the 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.

Chapter 3

Empirical plots from FROC data

3.1 How much finished 100%

3.2 Introduction

FROC data consists of mark-rating pairs. An important 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 and the wAFROC1. Formulae for coordinates of each plot are given in terms of the underlying mark-rating 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.

3.3 FROC data and notation

3.3.1 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].

3.3.2 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.

3.3.3 z-samples vs. ratings

z-samples are conceptual quantities that can range from $-\infty$ to $+\infty$. Ratings are observed values typically collected as integers but any ordered set of values will do where larger values correspond to greater suspicion for disease. The conversion from z-samples to ratings is accomplished by adopting a binning rule.

3.3.4 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 $\zeta_r \le z < \zeta_{r+1}$ 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:

$$\left.\begin{aligned} \text{if}\,(\zeta_r \le z < \zeta_{r+1}) \Rightarrow \text{rating} = r+1 \\ r = 0, 1, ..., R_{ROC} - 1 \\ \zeta_0 = -\infty \\ \zeta_{R_{ROC}} = \infty \end{aligned}\right\} \tag{3.1}$$

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 $\zeta_r \le z < \zeta_{r+1}$. 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

[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 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 3.1: 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	$r_{k_t t l_s s}$	rating for case $k_t t$ and LL/NL mark $l_s s$
11	R_{FROC}	Number of FROC bins
12	ζ_1	Lowest non-dummy reporting threshold
13	ζ_r	$r = 2, 3, ...,$ non-dummy reporting thresholds
14	$\zeta_0, \zeta_{R_{FROC}+1}$	Dummy thresholds, negative and positive infinity
15	$W_{k_2 l_2}$	Weight of lesion $l_2 2$ in case $k_2 2$, explained later
16	L_{max}	Maximum number of lesions per case in dataset
17	L_T	Total number of lesions in dataset

bins is one less than the number of ROC bins. For example, 4 non-dummy cutoffs $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ can correspond to a 5-rating ROC study or to a 4-rating FROC study. To summarize:

$$\left.\begin{aligned} \text{if}\,(\zeta_r \le z < \zeta_{r+1}) \Rightarrow \text{rating} = r \\ r = 1, 2, ..., R_{FROC} \\ \zeta_0 = -\infty \\ \zeta_{R_{FROC}+1} = \infty \end{aligned}\right\} \tag{3.2}$$

3.3.5 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 3.1 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).

3.3.6 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: The rating of a mark is $r_{k_2 2 l_2 2}$. Unmarked NLs are unobservable events. Unmarked lesions are assigned negative infinity ratings.
- Row 11: R_{FROC} is the number of bins in the FROC study.
- Rows 12, 13 and 14: 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 15: $W_{k_2 l_2}$ 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 16: L_{max} is the maximum number of lesions per case in the dataset.
- Row 17: L_T is the total number of lesions in the dataset.

3.3.7 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 6, 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.

3.4 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 3.1 and assuming binned data[4] and $n(x)$ denotes the number of events x:

$$\mathrm{NLF}_r = \frac{n\,(\text{NLs rated} \geq \zeta_r)}{K_1 + K_2} \tag{3.3}$$

and

$$\mathrm{LLF}_r = \frac{n\,(\text{LLs rated} \geq \zeta_r)}{L_T} \tag{3.4}$$

The allowed values of r are:

$$r = 1, 2, ..., R_{FROC} \tag{3.5}$$

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 (3.3) and (3.4) are equivalent to:

$$\mathrm{NLF}_r = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right) \tag{3.6}$$

and

$$\mathrm{LLF}_r = \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{3.7}$$

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

$$\left.\begin{array}{rcl} \mathbb{I}\,(\text{True}) & = & 1 \\ \mathbb{I}\,(\text{False}) & = & 0 \end{array}\right\} \tag{3.8}$$

In Eqn. (3.6) $\mathbb{I}\left(N_{k_t t} > 0\right)$ 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

[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.

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. (3.7) 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. (3.6) and Eqn. (3.7) $r = R_{FROC} + 1$ yields the trivial operating point (0,0).

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

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

$$\text{NLF}_1 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \tag{3.9}$$

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:

$$\left.\begin{aligned} \text{LLF}_1 =& \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_1\right) \\ \leq& 1 \end{aligned}\right\} \tag{3.10}$$

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.

3.4.2 Futility of extrapolation outside the observed end-point

To understand this consider the expression for NLF_0, i.e., using Eqn. (3.6) with $r = 0$:

$$\text{NLF}_0 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq -\infty\right) \tag{3.11}$$

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

$$\left.\begin{aligned} \text{1st term} =& \left(\frac{1}{K_1 + K_2}\right) \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \\ =& \text{NLF}_1 \\ \text{2nd term} =& \left(\frac{1}{K_1 + K_2}\right) \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} > 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(-\infty \leq z_{k_t t l_1 1} < \zeta_1\right) \\ =& \frac{\text{unknown number}}{K_1 + K_2} \end{aligned}\right\} \tag{3.12}$$

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:

$$\left.\begin{aligned} \mathrm{LLF}_0 =& \frac{\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq -\infty\right)}{L_T} \\ =& 1 \end{aligned}\right\} \tag{3.13}$$

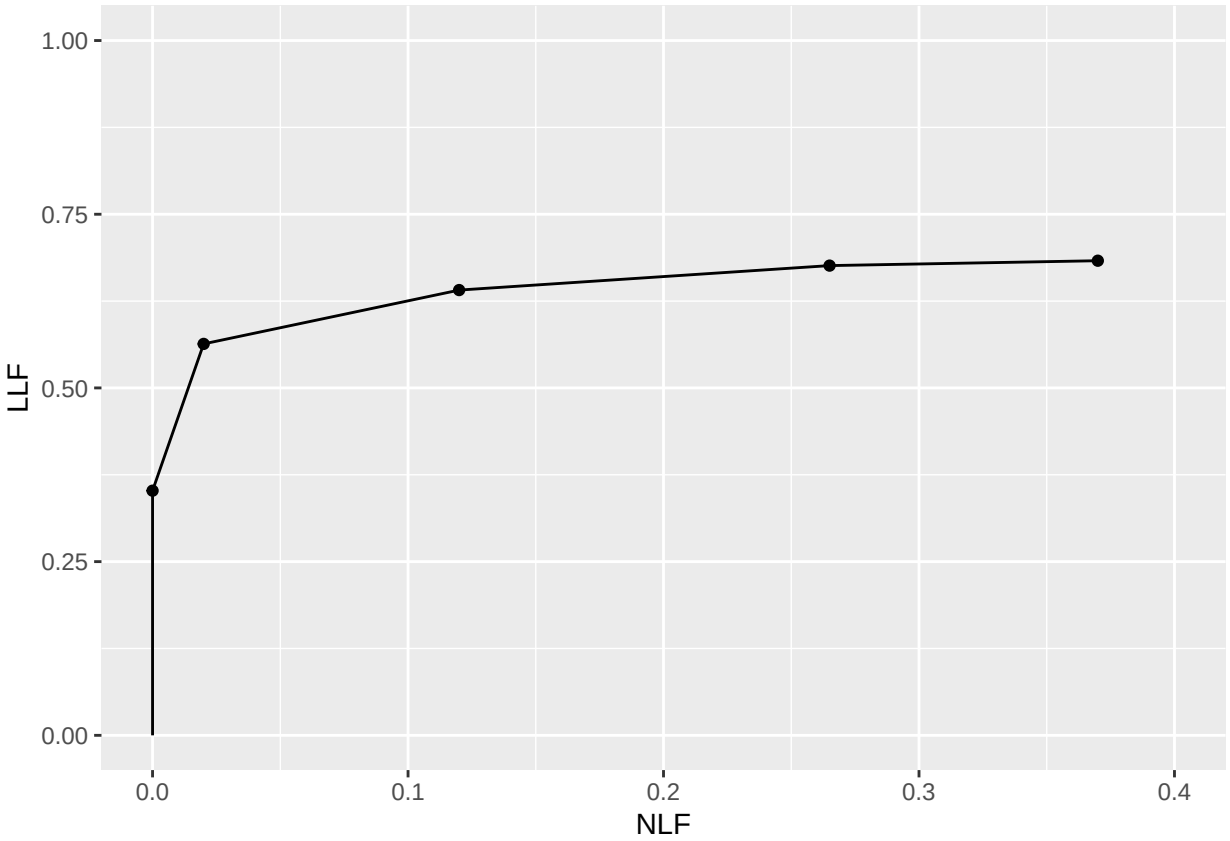
Unlike unmarked latent NLs, *unmarked lesions can safely be assigned the $-\infty$ rating, because an unmarked lesion is an observable event.* The right hand side of Eqn. (3.13) evaluates to unity. However, since the corresponding abscissa NLF_0 is undefined, one cannot plot this point. It follows that 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.

3.4.3 Illustration with a dataset

The following plot uses `dataset04` (Zanca et al., 2009) 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. (3.45). 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.

3.5 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.

3.5.1 The inferred-ROC z-sample

The highest ROC z-sample of a case, denoted $h_{k_t t}$, is the z-sample of the highest rated latent mark on the case or $-\infty$ if the case has no latent marks. For non-diseased cases $t = 1$ the maximum is over all latent NLs on the case. For diseased cases $t = 2$ the maximum is over all latent NLs *and* latent LLs on the case.

[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.

When there is little possibility for confusion, the prefix "inferred" is suppressed. ROC z-samples on non-diseased cases are referred to as FP z-samples and those on diseased cases as TP z-samples.

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 $h_{k_1 1} \geq \zeta$ divided by total number of non-diseased cases.
- $TPF(\zeta)$ = cumulated inferred TP counts with $h_{k_2 2} \geq \zeta$ 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).*

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

$$\left.\begin{array}{lcl} FP_{k_1 1} = & \max_{l_1}\left(z_{k_1 1 l_1 1}\right) & \text{if} \quad l_1 \neq \emptyset \\ FP_{k_1 1} = & -\infty & \text{if} \quad l_1 = \emptyset \end{array}\right\} \tag{3.14}$$

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:

$$\text{FPF}_r \equiv \text{FPF}\left(\zeta_r\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_r\right) \tag{3.15}$$

3.5.2 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:

$$TP_{k_2 2} = \max_{l_1 l_2}\left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}\right) \quad \text{if} \quad l_1 \neq \emptyset \tag{3.16}$$

or

$$TP_{k_2 2} = \max_{l_2}\left(z_{k_2 2 l_2 2}\right) \quad \text{if} \quad \left(l_1 = \emptyset\right) \wedge \left(\max_{l_2}\left(z_{k_2 2 l_2 2}\right) > -\infty\right) \tag{3.17}$$

or

$$TP_{k_2 2} = -\infty \quad \text{if} \quad \left(l_1 = \emptyset \wedge \left(\max_{l_2}\left(z_{k_2 2 l_2 2}\right) = -\infty\right)\right) \tag{3.18}$$

Here $\wedge$ is the logical AND operator. An explanation is in order. Consider Eqn. (3.16). 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. (3.16) 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. (3.17) applies, i.e., one takes the maximum z-sample over all marked LLs.
- If $l_1 = \emptyset$ and no lesions are marked, then Eqn. (3.18) applies; this represents an unmarked diseased case; the $-\infty$ z-sample assignment is justified because an unmarked diseased case is an observable event.

The inferred true positive fraction TPF_r is defined by:

$$\text{TPF}_r \equiv \text{TPF}(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I}\left(TP_{k_2 2} \geq \zeta_r\right) \tag{3.19}$$

3.5.3 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}.

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

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

$$\text{FPF}_1 \equiv \text{FPF}\left(\zeta_1\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_1\right) \tag{3.20}$$

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

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

$$\text{TPF}_1 \equiv \text{TPF}(\zeta_1) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I}\left(TP_{k_2 2} \geq \zeta_1\right) \tag{3.21}$$

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

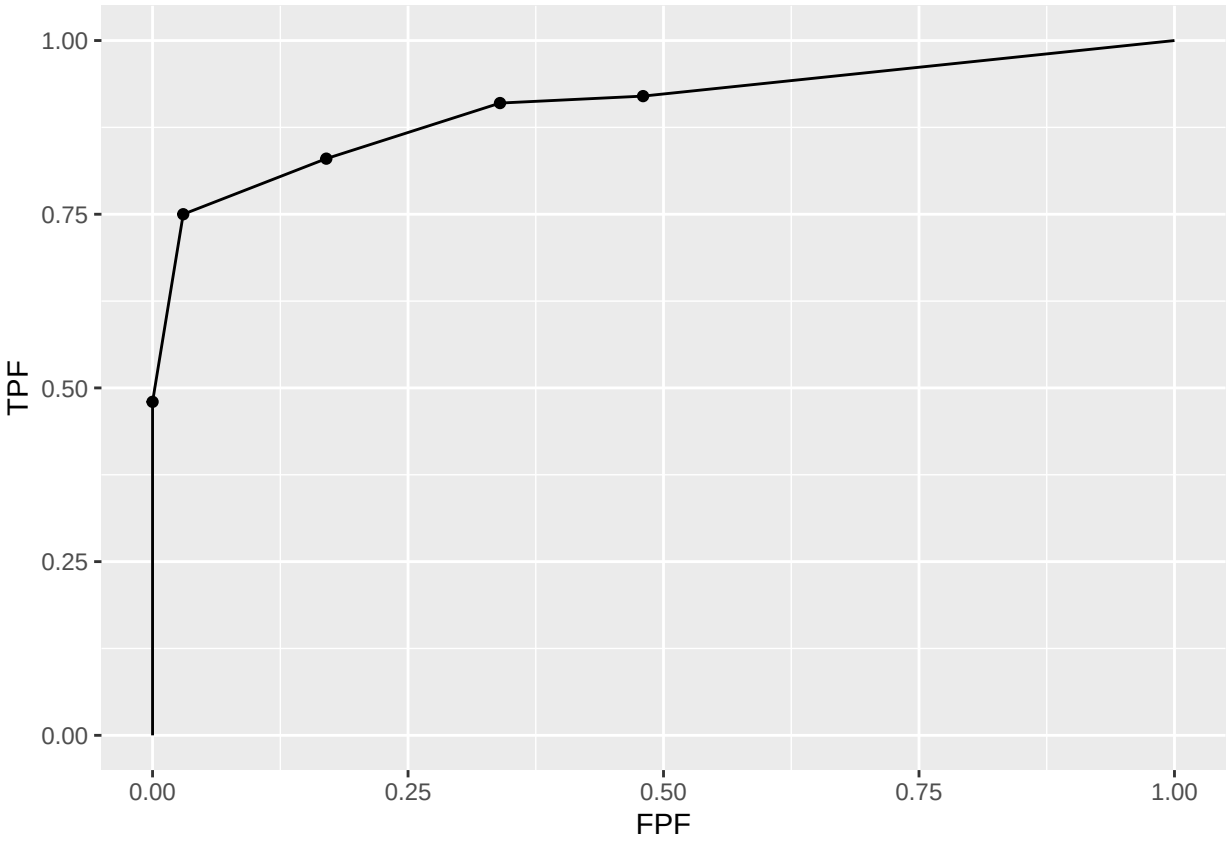
It follows that the observed end-point of the ROC (as is well known) satisfies the constrained end-point property: it lies below-left the (1,1) corner of the plot.

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

3.5.5 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
```

3.6 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. (3.7).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$, the abscissa of the ROC plot, was defined in Eqn. (3.15).

3.6.1 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).

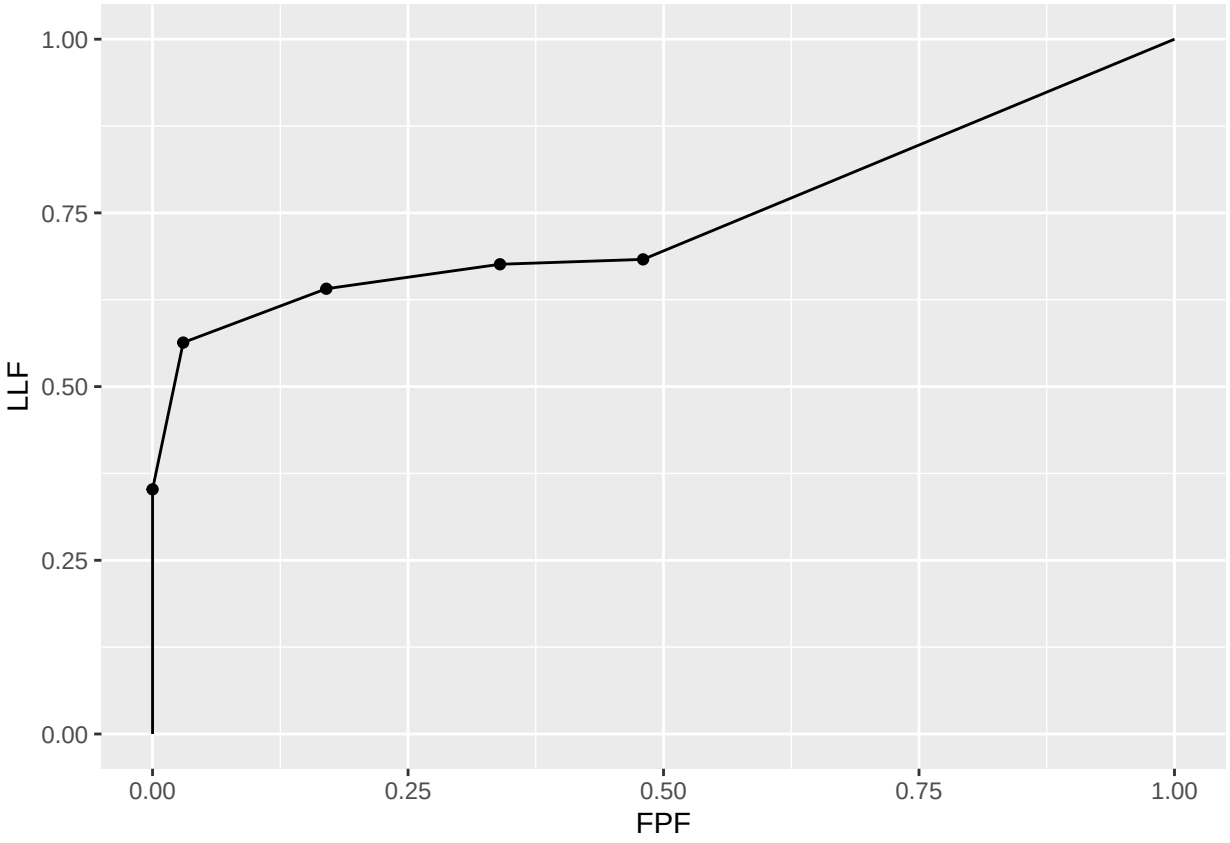
3.6.2 The observed end-point of the AFROC and its constrained property

According to Eqn. (3.15) the abscissa of the observed end-point $FPF_1 \leq 1$ and according to Eqn. (3.10) 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.

3.6.3 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
```

3.7 The weighted-AFROC plot (wAFROC) plot

The AFROC ordinate defined in Eqn. (3.7) 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]

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:

$$\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1 \tag{3.22}$$

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

$$\text{wLLF}_r \equiv \text{wLLF}\left(\zeta_r\right) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{3.23}$$

3.7.1 The empirical wAFROC plot and AUC

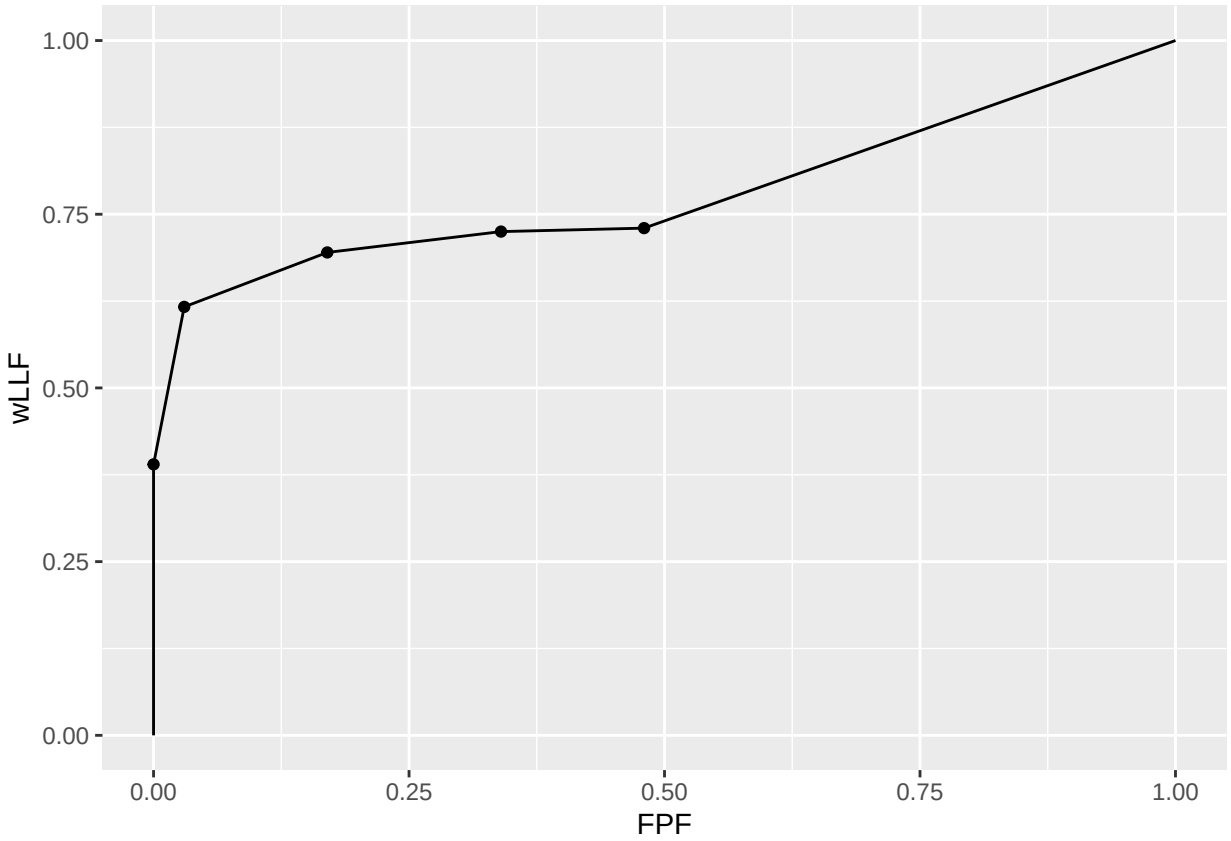
The empirical wAFROC plot connects adjacent operating points $(\text{FPF}_r, \text{wLLF}_r)$, including the origin (0,0), with straight lines plus a straight-line segment connecting the observed end-point to (1,1). The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC}.

3.7.2 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)
```

[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 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
```

3.8 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 z-samples (`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 z-samples 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.

3.8.1 NL and LL z-samples

Shown next is the `NL` z-samples 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 z-samples:
#>       [,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` z-samples 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 z-samples:
#>      [,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
```

3.8.2 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.

3.8.3 FPF

Shown next is the FP z-samples 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 z-samples for the corresponding non-diseased case. As an example, for case #3 the maximum of the two NL values is 0.7.

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

Here are the sorted FP z-samples.

```
#> [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.

3.8.4 LLF

Here are the sorted LL z-samples.

```
#>  [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.

3.8.5 wLLF

The sorted LL z-samples 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. 3.1 are the empirical AFROC and wAFROC plots.

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

[7] The jumps need not be contiguous: they will be contiguous only if the three lesion z-samples 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.

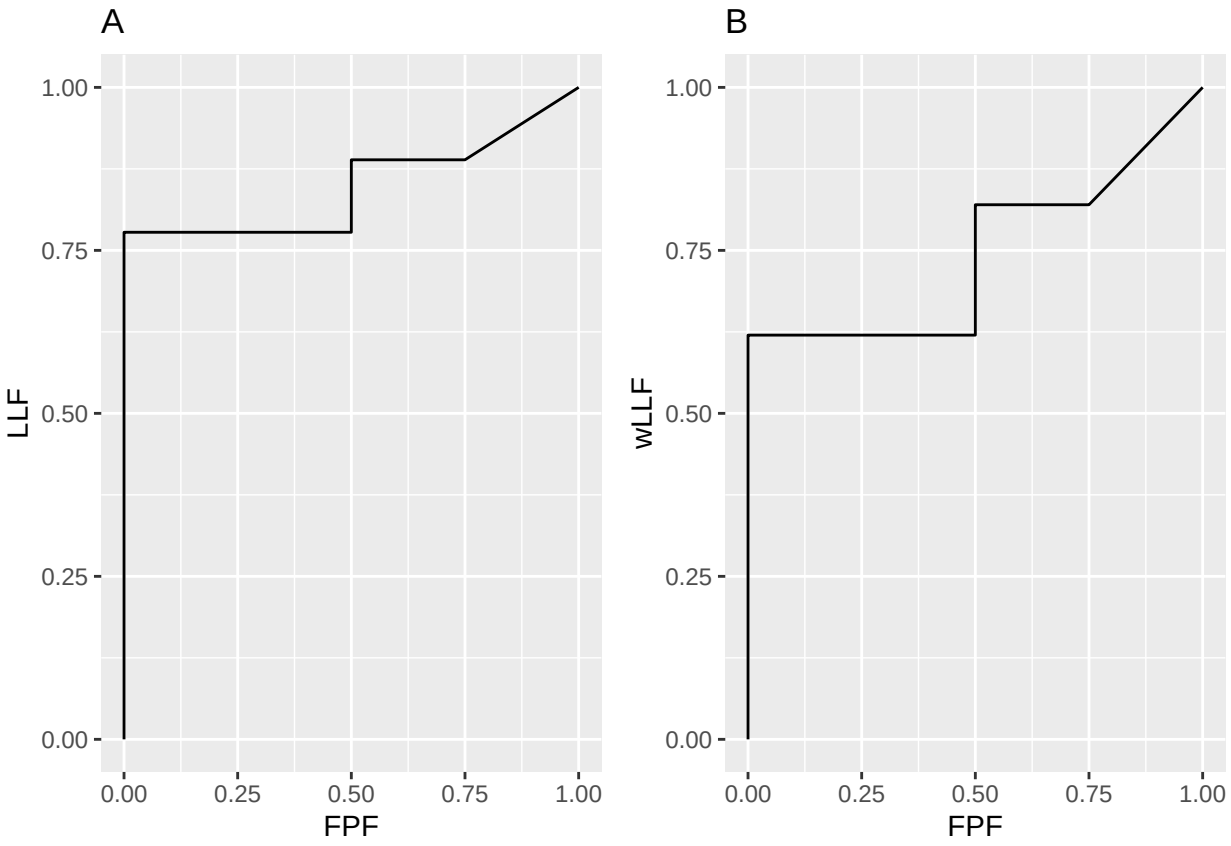


Figure 3.1: Left: AFROC plot; Right: corresponding wAFROC plot.

```
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 +
  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` z-sample exceeding a virtual threshold. Each rightward jump is associated with a `FP` z-sample 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.

3.9 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.

3.10 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.

3.10.1 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 $A_{\mathrm{FROC}} = 0$.

3.10.2 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 $A_{\mathrm{ROC}} = 0.5$.

3.10.3 The AFROC

3.10.3.1 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 $A_{\mathrm{AFROC}} = 0$.

3.10.3.2 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.

3.10.4 The wAFROC

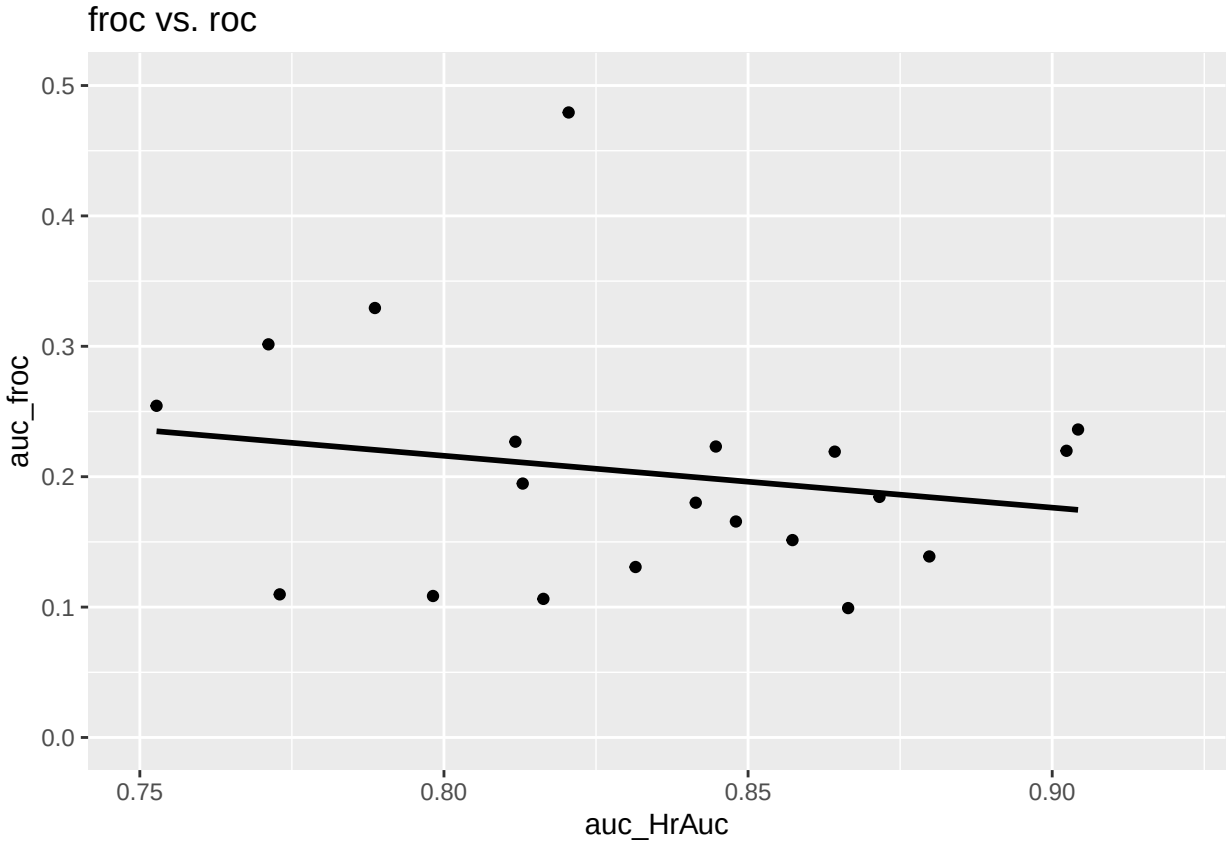
Similar comments apply to the wAFROC as already described above for AFROC.

3.11 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.

3.11.1 Plot of FROC AUC vs. ROC AUC

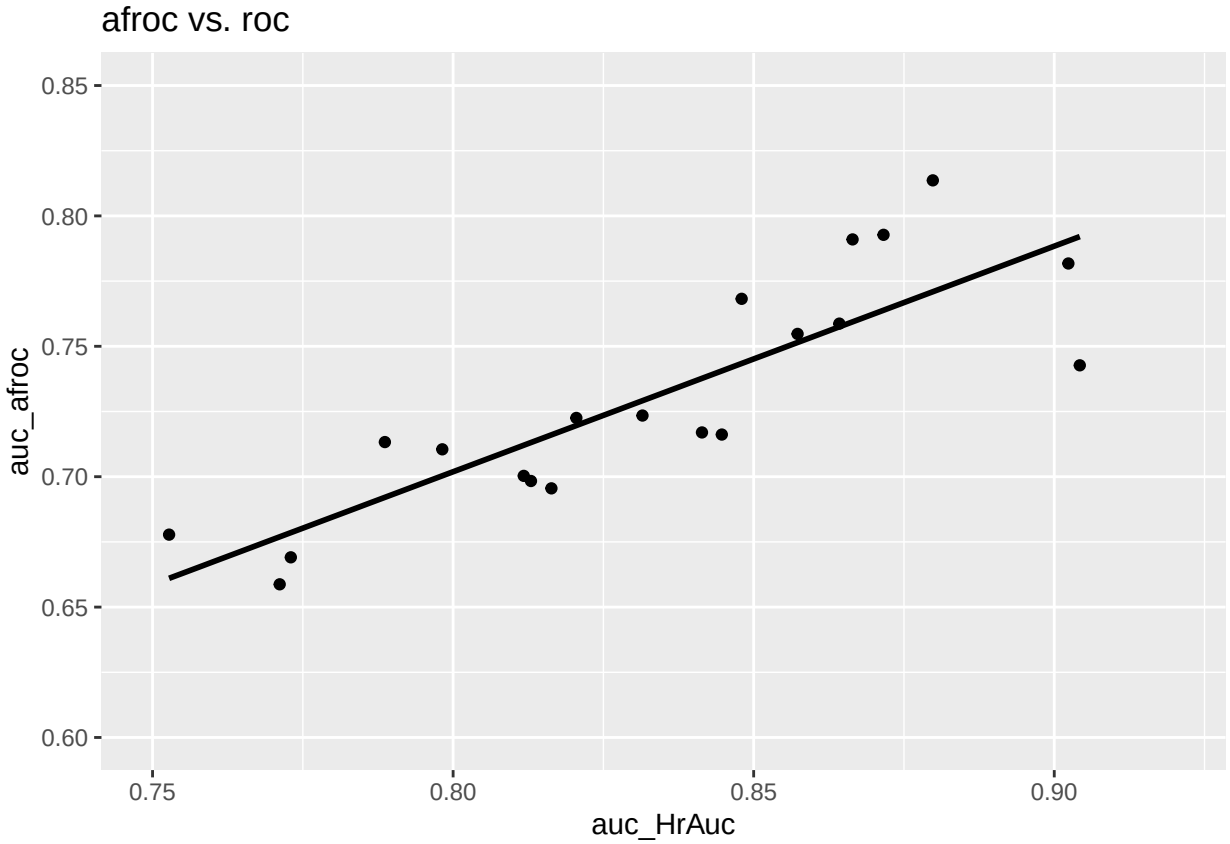
The following is the plot of A_{FROC} vs. A_{ROC}. There are 20 points on the plot corresponding to 5 treatments and 4 readers. The straight line is a least squares fit. Note the poor correlation and negative slope between A_{FROC} and A_{ROC}, $R^2 = 0.0347791$, slope = -0.3978636.



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.

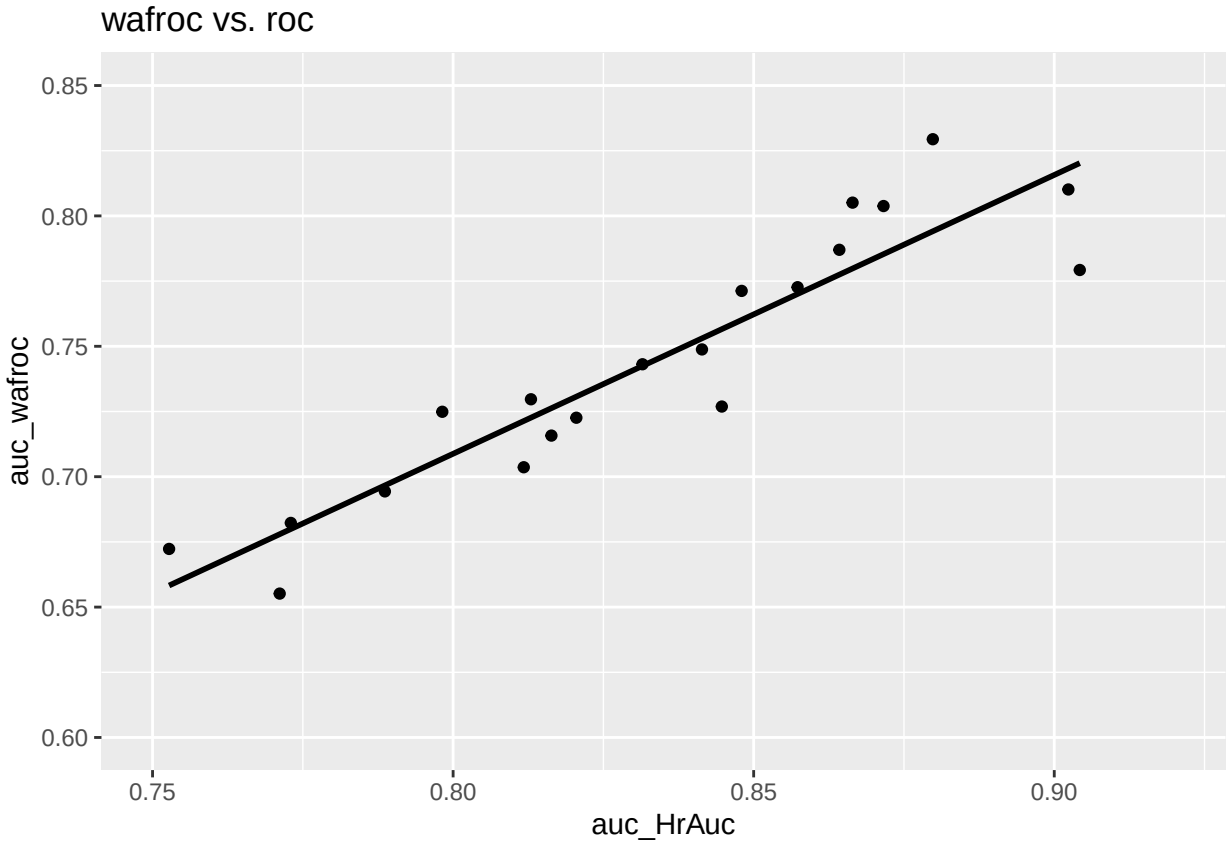
3.11.2 Plot of AFROC AUC vs. ROC AUC

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.



3.11.3 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.



3.12 The AFROC1 plot

Historically the AFROC originally used a different definition of FPF resulting in what is now termed the AFROC1 plot.

Since NLs can occur on diseased cases it is possible to define an inferred "FP" z-sample on a *diseased case* as the maximum of all NL z-samples on the case, or $-\infty$ if the case has no NLs. The quotes emphasize that this is non-standard usage of ROC terminology since 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 $k_t t$, where $t = 1, 2$, is:

$$\left.\begin{array}{rcl} FP^1_{k_1 t} = & \max_{l_1}\left(z_{k_t t l_1 1}\right) & \text{if} \quad l_1 \neq \emptyset \\ FP^1_{k_t t} = & -\infty & \text{if} \quad l_1 = \emptyset \end{array}\right\} \tag{3.24}$$

The "1" superscript below is necessary to distinguish the above definition from that in Eqn. (3.14).

$FP^1_{k_t t}$ 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:

$$\left.\begin{array}{l} FPF^1_r \equiv FPF^1\left(\zeta_r\right) \\ \qquad = \dfrac{1}{K_1 + K_2} \displaystyle\sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(FP^1_{k_t t} \geq \zeta_r\right) \end{array}\right\} \tag{3.25}$$

Note the differences between Eqn. (3.15) and Eqn. (3.25). The latter counts "FPs" on non-diseased *and* diseased cases while Eqn. (3.15) counts FPs on *only* non-diseased cases. The denominators in the two equations are different and, unlike the first equation, the second equation is valid even when $K_1 = 0$. This definition, resulting in the AUCs

described next, is useful in applications where all (or almost all) cases are diseased (i.e., all cases have "targets"). Most machine language applications may fall into this category: for example, a face-recognition algorithm may be used to search for target faces (e.g., known criminals) to be localized in crowd images; there mau be no (or very few) crowd images without any target faces. For these applications the following two empirical characteristics (AFROC1 and wAFROC1) are relevant.

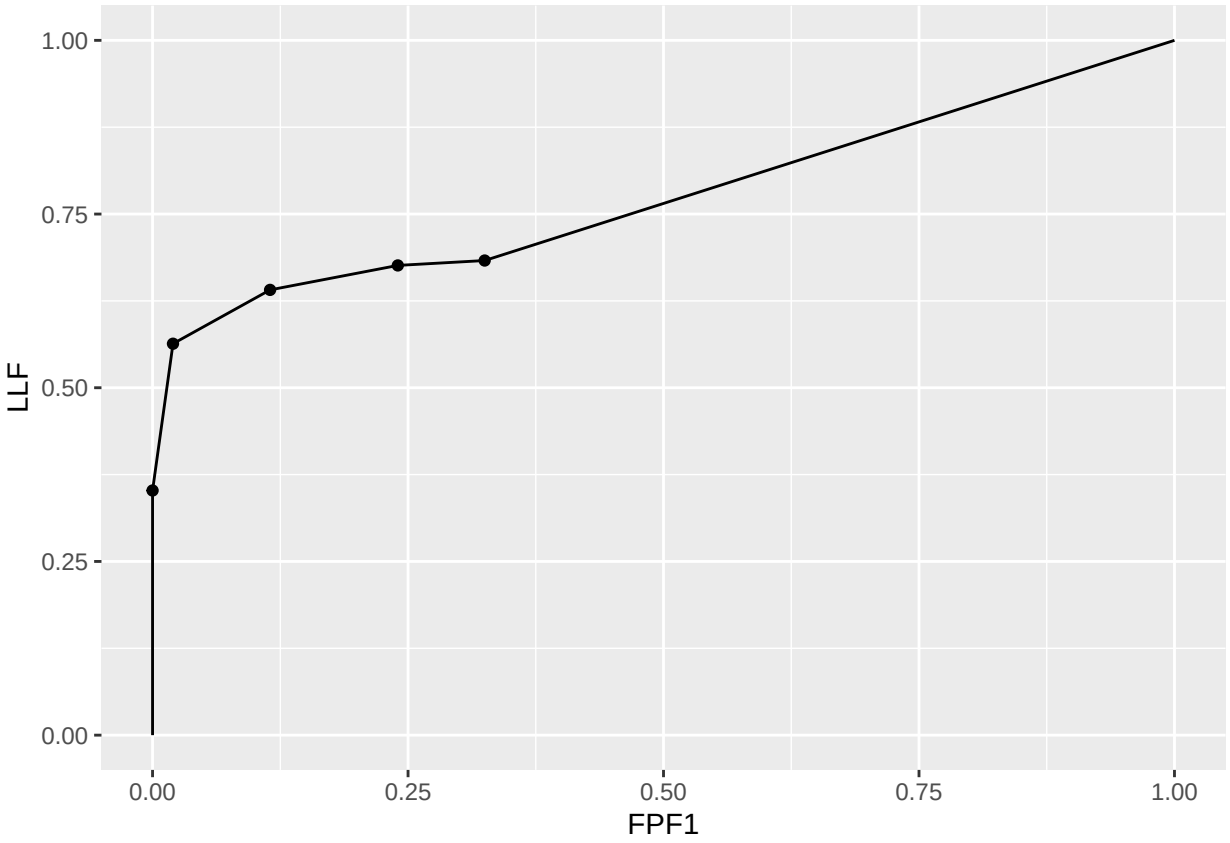
3.12.1 Empirical AFROC1 plot and AUC

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

3.12.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical AFROC1 plot for treatment 1 and reader 1. Note that the only difference from an empirical AFROC plot is in the abscissa.

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



Shown next is calculation of AFROC1-AUC 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
```

3.13 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. (3.23). The abscissa is FPF1, defined in Eqn. (3.25).

3.13.1 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} or wAFROC1-AUC.

The wAFROC1-AUC may be preferable as it gives equal importance to each case (or crowd image) regardless of the number of targets contained in it.

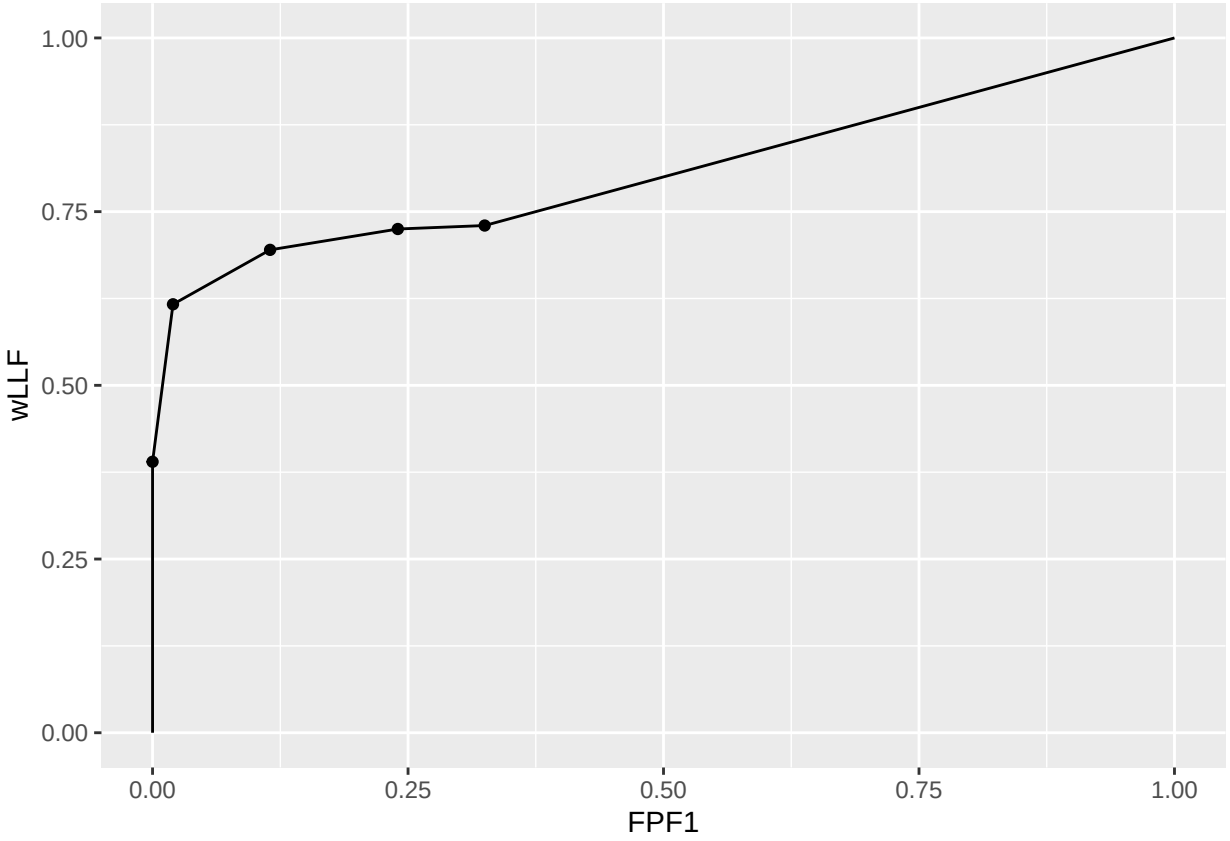
3.13.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical wAFROC1 plot for treatment 1 and reader 1. Note that the only difference from an empirical wAFROC plot is in the abscissa.

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

Table 3.2: Summary of plots from FROC data. OC = Operating Characteristic. All empirical plots except FROC include a straight line extension from the uppermost observed point to (1,1). Each figure of merit is defined by appending "-AUC" to the name of the corresponding OC

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 $K_1 \ll K_2$



Shown next is calculation of wAFROC1-AUC 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
```

3.14 Summary

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

3.15 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. 3.2.

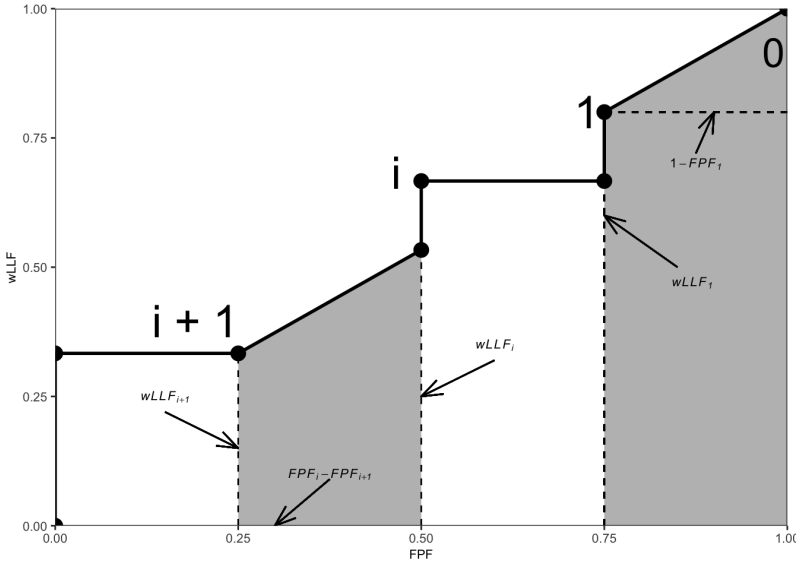


Figure 3.2: An example wAFROC plot; from left to right, the two shaded areas correspond to A_i and A_0, respectively, defined below.

The operating point labeled i has coordinates $(\mathrm{FPF}_i, \mathrm{wLLF}_i)$ given by Eqn. (3.15) and Eqn. (3.23).

The area A_i of the leftmost shaded trapezoid in Fig. 3.2 is:

$$A_i = \frac{\left(\mathrm{FPF}_i - \mathrm{FPF}_{i+1}\right)\left(\mathrm{wLLF}_i + \mathrm{wLLF}_{i+1}\right)}{2} \tag{3.26}$$

The weighted lesion localization fraction wLLF_r corresponding to threshold ζ_r is defined by Eqn. (3.23). It follows that:

$$\left.\begin{aligned} A_i = & \frac{\left(\mathrm{FPF}_i - \mathrm{FPF}_{i+1}\right)}{2} \times \\ & \frac{1}{K_2}\left[\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_i\right)\right. \\ & \left. + \sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right)\right] \end{aligned}\right\} \tag{3.27}$$

Using the probabilistic relation:

$$\mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_i\right) = \mathbb{I}\left(\zeta_i \leq z_{k_2 2 l_2 2} < \zeta_{i+1}\right) + \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_{i+1}\right) \tag{3.28}$$

we can expand the first term inside the square bracket:

$$\left.\begin{aligned} A_i =& \frac{(\mathrm{FPF}_i - \mathrm{FPF}_{i+1})}{2K_2} \times \\ & \left[\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(\zeta_i \le z_{k_2 2 l_2 2} < \zeta_{i+1}\right) \right. \\ & + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \ge \zeta_{i+1}\right) \\ & \left. + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \ge \zeta_{i+1}\right) \right] \end{aligned}\right\} \tag{3.29}$$

The last two terms are equal, therefore:

$$\left.\begin{aligned} A_i =& \frac{(\mathrm{FPF}_i - \mathrm{FPF}_{i+1})}{K_2} \times \\ & \left[\frac{1}{2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(\zeta_i \le z_{k_2 2 l_2 2} < \zeta_{i+1}\right) \right. \\ & \left. + \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 2 l_2 2} \ge \zeta_{i+1}\right) \right] \end{aligned}\right\} \tag{3.30}$$

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 $\zeta_i \le z_{k_t t l_s s} < \zeta_{i+1}$, 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. (3.15) it follows that:

$$\left.\begin{aligned} \mathrm{FPF}_i - \mathrm{FPF}_{i+1} =& \frac{1}{K_1} \left[\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1} \left(z_{k_1 1 l_1 1}\right) \ge \zeta_i \right) - \sum_{k_1=1}^{K_1} \mathbb{I}\left(z_{k_1 1 l_1 1} \ge \zeta_{i+1}\right) \right] \\ =& \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(\zeta_i \le \max_{l_1} \left(z_{k_1 1 l_1 1}\right) < \zeta_{i+1} \right) \end{aligned}\right\} \tag{3.31}$$

Because of the binning rule, $\mathbb{I}\left(\zeta_i \le \max_{l_1}\left(z_{k_1 1 l_1 1}\right) < \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = i\right)$, $\mathbb{I}\left(\zeta_i \le z_{k_2 2 l_2 2} < \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(r_{k_2 2 l_2 2} = i\right)$ and $\mathbb{I}\left(z_{k_2 2 l_2 2} \ge \zeta_{i+1}\right)$ can be replaced by $\mathbb{I}\left(r_{k_2 2 l_2 2} > i\right)$. Then Eqn. (3.27) can be re-written as:

$$\left.\begin{aligned} \mathrm{A}_i =& \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} \\ & \left[\frac{1}{2} W_{k_2 l_2} \mathbb{I}\left(\max_{l_1} \left(r_{k_1 1 l_1 1}\right) = i \right) \mathbb{I}\left(r_{k_2 2 l_2 2} = i\right) \right. \\ & \left. + \mathbb{I}\left(\max_{l_1} \left(r_{k_1 1 l_1 1}\right) = i \right) \mathbb{I}\left(r_{k_2 2 l_2 2} > i\right) \right] \end{aligned}\right\} \tag{3.32}$$

Eqn. (3.32) 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:

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} W_{k_2 l_2} (A + B) \tag{3.33}$$

where A and B are defined by:

$$\left.\begin{aligned} A =& \mathbb{I}\left(r_{k_2 2 l_2 2} = \max_{l_1}\left(r_{k_1 1 l_1 1}\right)\right) \\ B =& \mathbb{I}\left(r_{k_2 2 l_2 2} > \max_{l_1}\left(r_{k_1 1 l_1 1}\right)\right) \end{aligned}\right\} \tag{3.34}$$

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

$$\left.\begin{aligned} \psi(x, y) &= 1 \quad x < y \\ \psi(x, y) &= 0.5 \quad x = y \\ \psi(x, y) &= 0 \quad x > y \end{aligned}\right\} \tag{3.35}$$

It follows that:

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} W_{k_2 l_2} \psi\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right), r_{k_2 2 l_2 2}\right) \tag{3.36}$$

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

$$\mathrm{A}_{ROC} = \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \psi\left(r_{k_1 1}, r_{k_2 2}\right) \tag{3.37}$$

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.

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

We prove that the contribution of the $i = 0$ term in Eqn. (3.32) is identical to the area under the extension of the wAFROC from the uppermost empirical operating point to (1,1).

According to Eqn. (3.32),

$$\left.\begin{aligned} \mathrm{A}_0 =& \frac{1}{K_1 K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{l_{k_2}} \sum_{k_1=1}^{K_1} \\ & \left[\frac{1}{2} W_{k_2 l_2} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \mathbb{I}\left(r_{k_2 2 l_2 2} = 0\right)\right. \\ & \left. + \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right) = 0\right) \mathbb{I}\left(r_{k_2 2 l_2 2} > 0\right)\right] \end{aligned}\right\} \tag{3.38}$$

Rearranging the summations:

$$\left.\begin{aligned} \mathrm{A}_0 =& \frac{1}{2}\frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right)=0\right)\frac{1}{K_2}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{l_{k_2}} W_{k_2 l_2}\mathbb{I}\left(r_{k_2 2 l_2 2}=0\right) \\ &+\frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right)=0\right)\frac{1}{K_2}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{l_{k_2}} W_{k_2 l_2}\mathbb{I}\left(r_{k_2 2 l_2 2}>0\right) \end{aligned}\right\} \tag{3.39}$$

Consider the term:

$$\frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right)=0\right) \tag{3.40}$$

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:

$$\frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(\max_{l_1}\left(r_{k_1 1 l_1 1}\right)=0\right) = 1-\mathrm{FPF}_1 \tag{3.41}$$

Similarly,

$$\frac{1}{K_2}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{l_{k_2}} W_{k_2 l_2}\mathbb{I}\left(r_{k_2 2 l_2 2}=0\right) = 1-\mathrm{wLLF}_1 \tag{3.42}$$

Using these expressions, Eqn. (3.39) reduces to:

$$\mathrm{A}_0 = \frac{\left(1-\mathrm{FPF}_1\right)\left(1+\mathrm{wLLF}_1\right)}{2} \tag{3.43}$$

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 $(1-\mathrm{wLLF}_1)$:

$$\left.\begin{aligned} \text{Area st. line ext.} &= \left(1-\mathrm{FPF}_1\right)\mathrm{wLLF}_1 + \frac{\left(1-\mathrm{FPF}_1\right)\left(1-\mathrm{wLLF}_1\right)}{2} \\ &= \left(1-\mathrm{FPF}_1\right)\left(\mathrm{wLLF}_1 + \frac{\left(1-\mathrm{wLLF}_1\right)}{2}\right) \\ &= \frac{\left(1-\mathrm{FPF}_1\right)\left(1+\mathrm{wLLF}_1\right)}{2} \end{aligned}\right\} \tag{3.44}$$

which equals the right hand side of Eqn. (3.43).

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. (3.43), 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 $\text{FPF}_1 = 0$ and $\text{wLLF}_1 = 1$ and then, according to Eqn. (3.43), the area under the straight line extension is $A_0 = 1$.

3.17 Appendix 3: Summary of computational formulae

3.17.1 FROC

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

$$\text{A}_{FROC} = \frac{1}{(K_1 + K_2)\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} (A + B) \tag{3.45}$$

where A and B are defined by:

$$\left.\begin{aligned} A &= \sum_{k_1=1}^{K_1} \sum_{l_1=1}^{N_{k_1 1}} \mathbb{I}\left(z_{k_1 1 l_1 1} \neq -\infty\right) \psi\left(z_{k_1 1 l_1 1}, z_{k_2 2 l_2 2}\right) \\ B &= \sum_{k_2'=1}^{K_2} \sum_{l_1=1}^{N_{k_2' 2}} \mathbb{I}\left(z_{z_{k_2' 2 l_1 1}} \neq -\infty\right) \psi\left(z_{k_2' 2 l_1 1}, z_{k_2 2 l_2 2}\right) \end{aligned}\right\} \tag{3.46}$$

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. (3.45) 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. (3.45) implemented in function `UtilFigureOfMerit`.

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

3.17.2 ROC

The ROC-AUC formula is much simpler.

$$\text{A}_{ROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \psi\left(\max_{l_1}\left(z_{k_1 1 l_1 1}\right), \max_{l_1 l_2}\left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}\right)\right) \tag{3.47}$$

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. (3.37).

3.17.3 AFROC

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

$$\mathrm{A}_{AFROC} = \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \psi \left(\max_{l_1} \left(z_{k_1 1 l_1 1} \right), z_{k_2 2 l_2 2} \right) \tag{3.48}$$

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.

3.17.4 wAFROC

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

$$\mathrm{A}_{wAFROC} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \psi \left(\max_{l_1} \left(z_{k_1 1 l_1 1} \right), z_{k_2 2 l_2 2} \right) \tag{3.49}$$

This is similar to Eqn. (3.48) 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.

3.17.5 AFROC1

$$\mathrm{A}_{AFROC1} = \frac{1}{(K_1 + K_2) \sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} (A + B) \tag{3.50}$$

where A and B are defined by:

$$\left.\begin{aligned} A &= \sum_{k_1=1}^{K_1} \psi\left(\max_{l_1}\left(z_{k_1 1 l_1 1}\right), z_{k_2 2 l_2 2}\right) \\ B &= \sum_{k_2'=1}^{K_2} \psi\left(\max_{l_1}\left(z'_{k_2 2 l_1 1}\right), z_{k_2 2 l_2 2}\right) \end{aligned}\right\} \tag{3.51}$$

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 $\mathrm{A}_{AFROC1} = 1$:

$$\left.\begin{aligned} \mathrm{A}_{AFROC1} =& \frac{1}{\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} 1 \\ =& \frac{1}{\sum_{k_2=1}^{K_2} L_{k_2 2}} \sum_{k_2=1}^{K_2} L_{k_2 2} \\ =& 1 \end{aligned}\right\} \tag{3.52}$$

3.17.6 wAFROC1

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

$$\mathrm{A}_{wAFROC1} = \frac{1}{(K_1 + K_2) K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} (A + B) \tag{3.53}$$

A and B are as defined in Eqn. (3.51).

Chapter 4

Validity of the highest rating assumption

4.1 How much finished 0%

4.2 Introduction

4.3 The FROC and real ROC datasets

4.4 Code implementation

4.5 Load the three datasets

```
1  # start with original Federica FROC dataset
2  ds <- dataset04
3  # convert it to ROC and extract modalities 4 and 5
4  # infd_ds means the highest rating inferred ROC dataset,
5  # implemented in DfFroc2Roc
6  infd_ds <- DfExtractDataset(DfFroc2Roc(ds), trts = c(4,5))
7
8  # Federica real ROC dataset; this used modalities 4 and 5,
9  # same readers and same cases as the previous FROC study
10 real_ds <- dataset14
11
12 # load a cross modality dataset
13 # This will serve as a template whose list elements will be modified to create
14 # the desired cross modality dataset
15 xds <- datasetX
```

Line 2, `ds <- dataset04`, is the Federica Zanca 5 modality, 4 reader, 200 case FROC dataset.

Line 6 converts this to an **inferred** ROC dataset `infd_ds` containing treatments 4 and 5 only.

Line 10, `real_ds <- dataset14` is the Federica Zanca 2 modality, 4 reader, 200 case **real** ROC dataset. The two modalities correspond to treatments 4 and 5 in the FROC dataset `dataset04`.

Line 15 assigns a pre-loaded **crossed** modality dataset `xds <- datasetX` which serves as a template to be modified to meet our needs.

The original dimensions of `xds$ratings$NL` is `dim(xds$ratings$NL) = 2, 4, 11, 68, 5`. This because it represents a crossed modality dataset with two modality-1 factors (adaptive iterative dose reduction and filtered back

projection) crossed with 4 modality-2 factors (x-ray tube charge = 20 mAs, 40 mAs, 60 mAs and 80 mAs), 11 readers, and 68 cases with a maximum of 5 marks per case.

4.6 Modify the template

Any dataset is a multilevel list containing three list members at level 1, as shown below:

```
str(xds, max.level = 1)
#> List of 3
#>  $ ratings     :List of 3
#>  $ lesions     :List of 3
#>  $ descriptions:List of 8
```

Each of these lists needs to be modified as shown next for the `ratings` list member.

4.6.1 Modify the `ratings` list member

```
1 # modify the ratings list
2 xds$ratings$NL <- array(dim = c(2,2,4,200,1))
3 xds$ratings$NL[1,,,,] <- infd_ds$ratings$NL
4 xds$ratings$NL[2,,,,] <- real_ds$ratings$NL
5
6 xds$ratings$LL <- array(dim = c(2,2,4,100,1))
7 xds$ratings$LL[1,,,,] <- infd_ds$ratings$LL
8 xds$ratings$LL[2,,,,] <- real_ds$ratings$LL
```

Since the desired crossed modality dataset has two modality-1 factors (inferred ROC and real ROC), two modality-2 treatments (the investigated image processing algorithms), 4 readers, 200 cases and a maximum of 1 mark per case (because it is ROC data) we initialize the array with `NA`s, see line 2, `xds$ratings$NL <- array(dim = c(2,2,4,200,1))`.

Line 3, `xds$ratings$NL[1,,,,] <- infd_ds$ratings$NL`, copies the NL ratings from the inferred dataset `infd_ds` to `xds`. The index 1 refers to the modality-1 factor (inferred-ROC).

Line 4, `xds$ratings$NL[2,,,,] <- real_ds$ratings$NL`, copies the NL ratings from the real dataset `real_dstoxds`‘. The index 2 refers to the modality-2 factor (real-ROC).

Lines 6-8 repeats the above steps for the LL events. In the initialization at line 6, `xds$ratings$LL <- array(dim = c(2,2,4,100,1))`, the 100 follows from the fact that the maximum number of diseased cases is 100 each with 1 true lesion per case.

4.6.2 Modify the `lesions` list member

```
1 # modify the lesions list
2 xds$lesions$perCase <- array(1,dim = c(100))
3 xds$lesions$IDs <- array(1,dim = c(100,1))
4 xds$lesions$weights <- array(1,dim = c(100,1))
```

The next three lines, 2-4, modify the lesions list. `xds$lesions$perCase` is set to an array of one-hundred ones, as each diseased case has one lesion. Likewise for the `xds$lesions$IDs` and `xds$lesions$weights` (the redundant dimension is necessary for compatibility with other code in `RJafroc`).

4.6.3 Modify the descriptions list member

```
# modify the descriptions list
xds$descriptions$fileName <- "combined dataset04 & dataset14"
xds$descriptions$type <- "ROC"
xds$descriptions$name <- "FEDERICA-INFERRED-PLUS-REAL"
xds$descriptions$design <- "FCTRL-X-MOD"
xds$descriptions$modalityID1 <- c("infd", "real")
xds$descriptions$modalityID2 <- c("trt4", "trt5")
xds$descriptions$readerID <- c("rdr1","rdr2","rdr3","rdr4")
```

Lines 2-8 update the `descriptions` list. As examples, `xds$descriptions$type <- "ROC"` sets the `type` member to "ROC", `xds$descriptions$design <- "FCTRL-X-MOD"` sets the `design` member to "FCTRL-X-MOD", for factorial crossed modality, `xds$descriptions$modalityID1 <- c("infd", "real")` sets the two levels of the `modalityID1` member to `c("infd", "real")`, corresponding to inferred and real, respectively, `xds$descriptions$modalityID2 <- c("trt4", "trt5")` sets the two levels of the `modalityID2` member to `c("trt4", "trt5")`, corresponding to the two image processing algorithms and `xds$descriptions$readerID <- c("rdr1","rdr2","rdr3","rdr4")` sets the `readerID` member to the indicated labels.

This completes the merging of the two datasets, inferred ROC and real ROC, into a crossed modality dataset.

4.7 Analysis of the crossed modality dataset

This is done as shown next.

```
st <- St(
  xds,
  FOM <- "Wilcoxon",
  analysisOption = "RRRC")
```

```
st
#> $FOMs
#> $FOMs$foms
#> $FOMs$foms$AvgMod1
#>          rdrrdr1  rdrrdr2  rdrrdr3  rdrrdr4
#> trttrt4 0.903125 0.848875 0.825100 0.879300
#> trttrt5 0.857775 0.818575 0.799875 0.848125
#>
#> $FOMs$foms$AvgMod2
#>          rdrrdr1 rdrrdr2  rdrrdr3  rdrrdr4
#> trtinfd 0.871875 0.80225 0.779900 0.863900
#> trtreal 0.889025 0.86520 0.845075 0.863525
#>
#>
#> $FOMs$trtMeans
#> $FOMs$trtMeans$AvgMod1
#>          Estimate
#> trttrt4 0.8641000
#> trttrt5 0.8310875
#>
#> $FOMs$trtMeans$AvgMod2
#>          Estimate
#> trtinfd 0.8294812
#> trtreal 0.8657063
```

```
#>
#>
#> $FOMs$trtMeanDiffs
#> $FOMs$trtMeanDiffs$AvgMod1
#>                   Estimate
#> trttrt4-trttrt5 0.0330125
#>
#> $FOMs$trtMeanDiffs$AvgMod2
#>                   Estimate
#> trtinfd-trtreal -0.036225
#>
#>
#>
#> $ANOVA
#> $ANOVA$TRanova
#> $ANOVA$TRanova$AvgMod1
#>             SS DF           MS
#> T  0.00217965  1 2.179650e-03
#> R  0.00217965  3 1.842759e-03
#> TR 0.00217965  3 3.726552e-05
#>
#> $ANOVA$TRanova$AvgMod2
#>              SS DF          MS
#> T  0.0055282777  1 0.002624501
#> R  0.0055282777  3 0.001842759
#> TR 0.0055282777  3 0.000542624
#>
#>
#> $ANOVA$VarCom
#> $ANOVA$VarCom$AvgMod1
#>            Estimates      Rhos
#> VarR    0.0008321326        NA
#> VarTR -0.0001789411        NA
#> Cov1    0.0003146504 0.5827166
#> Cov2    0.0002531509 0.4688227
#> Cov3    0.0002440363 0.4519429
#> Var     0.0005399715        NA
#>
#> $ANOVA$VarCom$AvgMod2
#>           Estimates      Rhos
#> VarR   0.0005905767        NA
#> VarTR  0.0003041707        NA
#> Cov1   0.0003005249 0.5423688
#> Cov2   0.0002561530 0.4622891
#> Cov3   0.0002410342 0.4350036
#> Var    0.0005540970        NA
#>
#>
#> $ANOVA$IndividualTrt
#> $ANOVA$IndividualTrt$AvgMod1
#>          DF  msREachTrt   varEachTrt  cov2EachTrt
#> trttrt4   3 0.001168930 0.0005165784 0.0002390223
#> trttrt5   3 0.000711094 0.0005633647 0.0002672795
#>
#> $ANOVA$IndividualTrt$AvgMod2
#>          DF   msREachTrt   varEachTrt  cov2EachTrt
```

```
#> trtinfd  3 0.0020605739 0.0005708085 0.0002192931
#> trtreal  3 0.0003248089 0.0005373855 0.0002930128
#>
#>
#> $ANOVA$IndividualRdr
#> $ANOVA$IndividualRdr$AvgMod1
#>         DF   msTEachRdr   varEachRdr  cov1EachRdr
#> rdrrdr1  1 0.0010283113 0.0004861758 0.0003203932
#> rdrrdr2  1 0.0004590450 0.0005679882 0.0003195145
#> rdrrdr3  1 0.0003181503 0.0006937098 0.0003582329
#> rdrrdr4  1 0.0004859403 0.0004120123 0.0002604610
#>
#> $ANOVA$IndividualRdr$AvgMod2
#>         DF   msTEachRdr   varEachRdr  cov1EachRdr
#> rdrrdr1  1 1.470612e-04 0.0004777946 0.0003287743
#> rdrrdr2  1 1.981351e-03 0.0005660321 0.0003214706
#> rdrrdr3  1 2.123890e-03 0.0006664924 0.0003854503
#> rdrrdr4  1 7.031250e-08 0.0005060688 0.0001664044
#>
#>
#>
#> $RRRC
#> $RRRC$FTests
#> $RRRC$FTests$AvgMod1
#>                 DF           MS    FStat            p
#> Treatment  1.00000 2.179650e-03 29.56509 0.0001627221
#> Error     11.74146 7.372378e-05       NA           NA
#>
#> $RRRC$FTests$AvgMod2
#>                DF           MS    FStat         p
#> Treatment 1.00000 0.0026245013 4.351692 0.1108024
#> Error     3.70596 0.0006030991       NA        NA
#>
#>
#> $RRRC$ciDiffTrt
#> $RRRC$ciDiffTrt$AvgMod1
#>                  Estimate      StdErr       DF        t        PrGTt    CILower
#> trttrt4-trttrt5 0.0330125 0.006071399 11.74146 5.437379 0.0001627221 0.01975168
#>                    CIUpper
#> trttrt4-trttrt5 0.04627332
#>
#> $RRRC$ciDiffTrt$AvgMod2
#>                  Estimate     StdErr      DF         t     PrGTt     CILower
#> trtinfd-trtreal -0.036225 0.01736518 3.70596 -2.086071 0.1108024 -0.08598526
#>                    CIUpper
#> trtinfd-trtreal 0.01353526
#>
#>
#> $RRRC$ciAvgRdrEachTrt
#> $RRRC$ciAvgRdrEachTrt$AvgMod1
#>          Estimate     StdErr        DF   CILower   CIUpper         Cov2
#> trttrt4 0.8641000 0.02304897  9.914476 0.8126836 0.9155164 0.0002390223
#> trttrt5 0.8310875 0.02109628 18.802288 0.7869010 0.8752740 0.0002672795
#>
#> $RRRC$ciAvgRdrEachTrt$AvgMod2
#>          Estimate     StdErr        DF   CILower   CIUpper         Cov2
```

```
#> trtinfd 0.8294812 0.02710049  6.097804 0.7634257 0.8955368 0.0002192931
#> trtreal 0.8657063 0.01934464 63.712979 0.8270575 0.9043550 0.0002930128
```

The radiological search model (RSM)

Chapter 5

Visual Search

5.1 How much finished 100%

5.2 Introduction

This chapter draws 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 gratefully acknowledges critical insights gained through conversations with Dr. Claudia Mello-Thoms ca. 2003.

To understand free-response data, specifically how radiologists interpret images, one must understand visual search. Casual usage of everyday terms like "search", "recognition" and "detection" can lead to confusion.

> Visual search is broadly defined as grouping and labeling parts of an image. In the medical imaging context visual search involves finding lesions and correctly classifying them (as benign or malignant).

A schema of how radiologists find perform the search task, termed the Kundel-Nodine search model, is described. This model is the basis of the radiological search model (RSM) described in Chapter 6.

5.3 Grouping and labeling ROIs

Looking at and understanding an image involves grouping and assigning labels to different regions in the image, where the labels correspond to entities that exist in the real world. As an example, if one looks at Fig. 5.1, one would group the image into 8 rectangular regions arranged in two rows and 4 columns and 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 expertise of the observer: if one were ignorant about American history one would be unable to correctly label them.

Image interpretation in radiology is not fundamentally different. It involves grouping and recognizing areas of the image that have correspondences to the radiologist's knowledge of 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. 5.2. This is because they know the underlying anatomy, Fig. 5.3 and have a basic understanding of x-ray image formation physics that relates the anatomy to the image.

5.4 Lesion-localization vs. detection

The process of grouping and labeling parts of an image is termed *recognition*. Recognition is distinct from detection, which is deciding about the presence of something that is unexpected or the absence of something that is expected,

Figure 5.1: Grouping and labeling regions of an image.

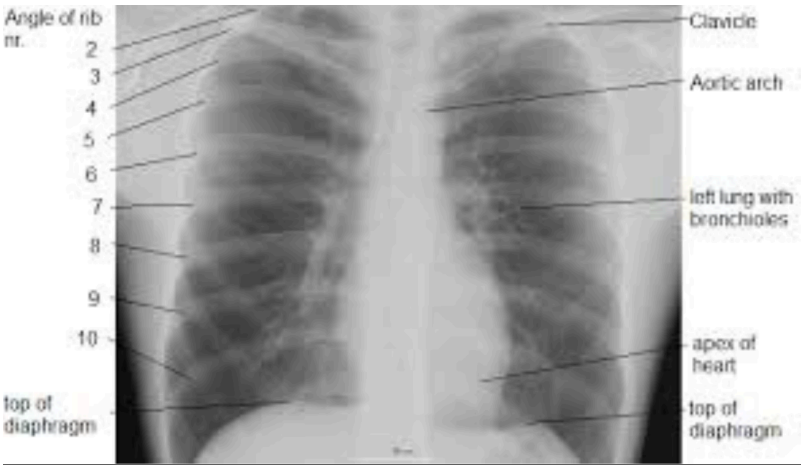


Figure 5.2: Grouping and labeling in radiology.

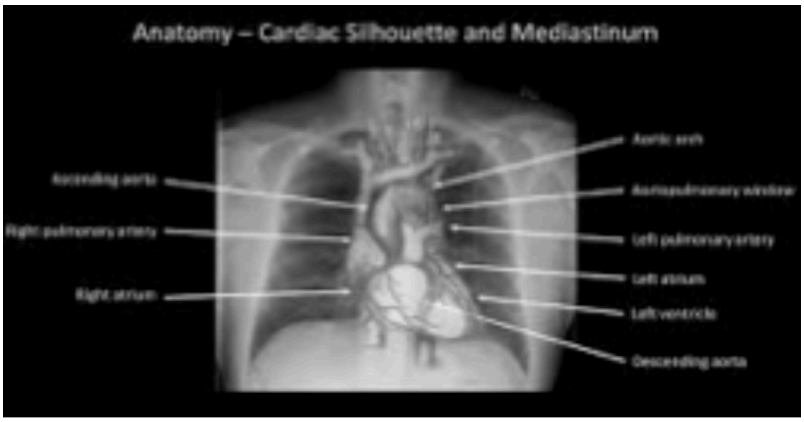


Figure 5.3: Correct grouping and labeling requires knowledge of the underlying anatomy.

in other words, a deviation 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" imply expertise dependent expectations regarding the structure of a generic non-diseased image, which I term a *non-diseased template*, and therefore the ability to recognize clinically relevant perturbations from this template. 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. Detection is the presence or absence of something, i.e., the perturbation, which could be anywhere. For example, in Fig. 5.1, recognizing a face is equivalent to assigning a row and column index in the image. Specifically, recognizing George H.W. Bush implies pointing to row = 2 and column = 3. Detecting George H.W. Bush implies stating that George H.W. Bush is somewhere in the image. Recognition is an FROC paradigm task while detection is an ROC task.

Instead of recognition (as used by Kundel and Nodine) I prefer the term "search", as in "searching for and finding" a lesion.

5.5 Lesion-localization vs. lesion-classification

Since template perturbations can occur at different locations in the images, the ability to selectively recognize them is related to search expertise. [1] Lesion-localization expertise is the selective ability to locate clinically relevant perturbations that are actually present while minimizing false localizations.

Two important terms are introduced using FROC terminology:

> Lesion-localization (or finding) expertise is the ability to find latent LLs while minimizing finding latent NLs.
>
> Lesion-classification (or recognition) expertise is the ability to mark LLs while not marking NLs.

The skills required to find and recognize a nodule in a chest x-ray are different from those required to find and recognize a low-contrast circular or Gaussian shaped artificial nodule against a background of random noise (or even an anthropomorphic phantom). In the former instance the skills of the radiologist are relevant while in the latter they are not. This is the reason why having radiologists interpret random noise images and claiming that this somehow makes it "clinically relevant" is incorrect. One might as well use anyone with good eyesight, motivation and training. This paragraph also argues against 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 skills.

5.6 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 (and based on eye-tracking recordings obtained on radiologists while they interpreted clinical images) Kundel and Nodine proposed the following schema for the diagnostic interpretation process. It consists of two stages:

- Lesion-localization* or finding the locations of suspicious regions.
- Lesion-classification* or determining the classification (malignant or benign) of each found suspicious region.

[1] A non-expert can trivially recognize any and all perturbations that may be present by claiming all regions in the image are perturbed.

5.6.1 Lesion-localization

The search stage is brief, typically lasting about 100 - 300 ms, which is too short for detailed foveal examination. Instead peripheral vision is responsible for identification of perturbations. The result is a global impression or gestalt, that identifies perturbations from the generic non-diseased template. 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). Perturbations are flagged for subsequent feature analysis, described below, in other words *search tells the visual system where to look more closely.* In the computer aided detection (CAD) context this stage is termed *initial detection* (Edwards et al., 2002).

5.6.2 Lesion-classification

Having found a set of suspicious regions the observer analyzes each region for evidence of disease: in principle he calculates the probability of malignancy for each region. In the CAD context this is termed *candidate analysis*, aka the feature analysis stage, where each region found by the initial detection stage is analyzed to calculate a probability of malignancy (and marked if the probability exceeds some algorithm-designer selected value).

An essential point that emerges is that decisions (to mark or not mark) 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.

5.6.3 Example

Eye-tracker recordings for a two-view digital mammogram for two observers are shown in Fig. 5.4, for an inexperienced observer (upper two panels) and an expert mammographer (lower two panels). The small circles indicate individual fixations (dwell time ~ 100 ms). The larger bright (high-contrast) circles are clustered fixations (cumulative dwell time ~ 1 s). These correspond to the latent marks defined in the previous chapter.

The large low-contrast circle is a mass (and so labeled) visible in both views.

The inexperienced observer finds more suspicious regions than does the expert mammographer but misses the lesion in the MLO view. In other words, the inexperienced observer generated more latent NLs but only one latent LL. The mammographer finds the lesion in the MLO (mediolateral oblique) view, which qualifies as a latent LL, without finding suspicious regions in other areas, 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 CC (craniocaudal) view - the decision to recall the patient had already been made.

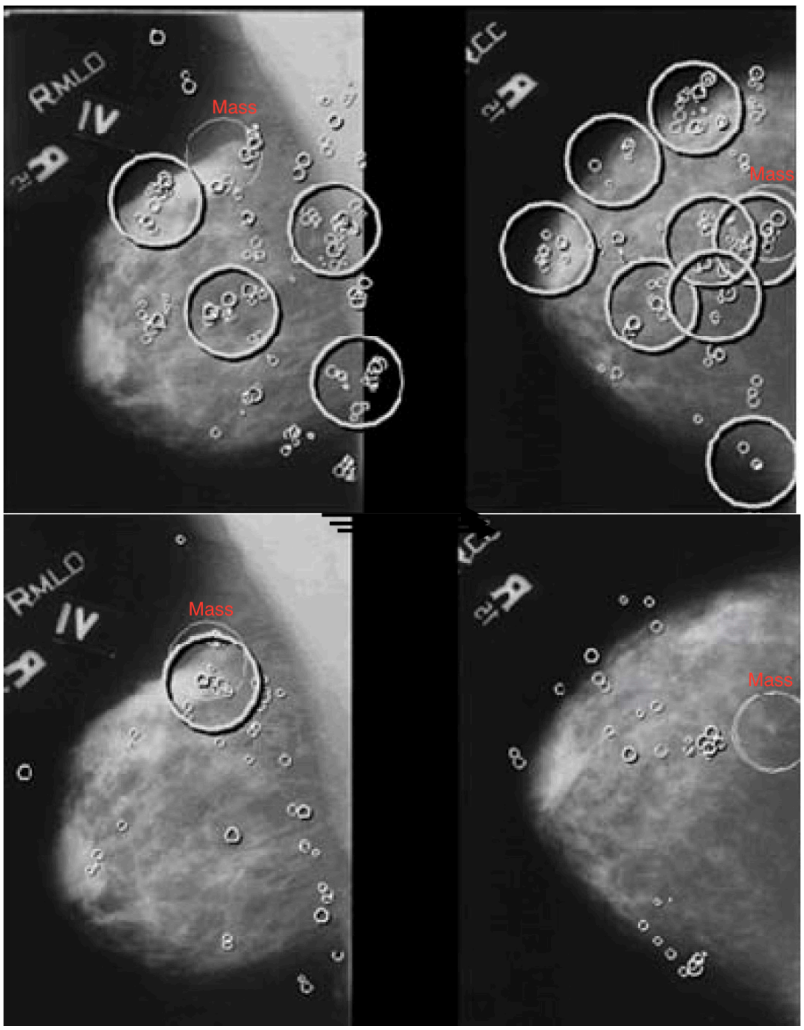


Figure 5.4: Eye-tracking recordings for a two-view digital mammogram. The top row is an inexperienced observer while the bottom row is an expert radiologist. The left column shows MLO views while the right column shows CC views.

Chapter 6

The radiological search model (RSM)

6.1 How much finished 99%

6.2 Introduction

All models of ROC data *that do not incorporate search* involve two fundamental parameters (i.e., not including binning-related threshold parameters). For example, the unequal variance binormal model requires the a, b parameters. Alternative ROC models (e.g., CBM and PROPROC) also require two fundamental parameters.

Two fundamental parameters of ROC models are needed (1) to accommodate the average visibility of lesions in the dataset (e.g., the a or separation parameter) and (ii) the fact that the observed diseased case distribution is usually wider than that of the non-diseased cases (e.g., the $b < 1$ parameter). If one assumes same widths for both distributions, so in effect $b = 1$ is no longer a free parameter, and one allows a varying number of latent marks on all cases, then it becomes possible that the distribution of the highest rating on diseased cases will have greater width than that on non-diseased cases simply due to the fact that latent NLs on diseased cases will have lower z-samples than latent LLs on diseased cases (i.e., a mix of NL and LLs) while on non-diseased cases there will be only NL z-samples. So the basic idea is to have a visibility parameter, a parameter describing the distribution of the number of latent NLs per case and a parameter describing the distribution of the number of latent LLs per case, i.e., a three-parameter model should suffice. And in fact the RSM contains three fundamental parameters: μ, λ and ν. In addition the lowest threshold ζ_1 needs to be included as a parameter as it determines the extent and shape of the RSM predicted operating characteristics. This will become clearer in the next chapter but for now can be illustrated by considering the extreme case $\zeta_1 = \infty$ when the predicted FROC is the single point (0,0).

6.3 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* in which suspicious regions, i.e., the latent marks, are identified via peripheral vision. The total number of latent marks on a case is random non-negative integer and in fact 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.
- If the z-sample exceeds a pre-selected minimum reporting threshold, denoted ζ_1 the location is marked, i.e., the latent mark becomes an actual mark.

- Latent marks can be either latent NLs (corresponding to non-diseased regions) or latent LLs (corresponding to lesions). The number of latent NLs or LLs on a case are denoted l_1, l_2 respectively. Latent NLs can occur on non-diseased or diseased cases but latent LLs can only occur on diseased cases. Assume that every diseased case has L actual lesions (this will later be extended to arbitrary number of lesions per diseased case). [1]

6.4 RSM assumptions

Assumption 1: The number of latent NLs, $l_1 \geq 0$, is sampled from the Poissson distribution Pois() with mean λ:

$$l_1 \sim \text{Pois}(\lambda) \tag{6.1}$$

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

$$\text{pmf}_P(l_1, \lambda) = exp(-\lambda) \frac{(\lambda)^{l_1}}{l_1!} \tag{6.2}$$

Assumption 2: The number of latent LLs, l_2, where $0 \leq l_2 \leq L$ (since the number of latent LLs cannot exceed the number of lesions) is sampled from the binomial distribution B with success probability ν and trial size L:

$$l_2 \sim \text{B}(L, \nu) \tag{6.3}$$

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

$$\text{pmf}_B(l_2, L, \nu) = \binom{L}{l_2} (\nu)^{l_2} (1-\nu)^{L-l_2} \tag{6.4}$$

Collectively λ and ν are termed the *search* parameters.

Assumption 3: Each latent mark is associated with a z-sample. That for a latent NL is denoted $z_{l_1 1}$ while that for a latent LL is denoted $z_{l_2 2}$. 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 $N(0,1)$:

$$z_{l_1 1} \sim N(0,1) \tag{6.5}$$

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

$$z_{l_2 2} \sim N(\mu,1) \tag{6.6}$$

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

$$\phi(z|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right) \tag{6.7}$$

The parameter μ is termed the *classification* parameter.

[1] Since the RSM is a parametric model one does not need the four subscript notation needed to account for case and location dependence necessary to describe observed data, as in Chapter 3. This allows for simpler notation, as the reader may have noticed, unencumbered by 4 subscripts as in $z_{k_t t l_s s}$ in Table 3.3.5.

Bning 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 $\zeta_r \leq z_{l_s s} < \zeta_{r+1}$ the corresponding latent site is marked and rated in bin r, and if $z_{l_s s} \leq \zeta_1$ the site is not marked. (R is the number of FROC bins.)

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 $\zeta_r \leq z_{l_1 1} < \zeta_{r+1}$ 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 $\zeta_r \leq z_{l_2 2} < \zeta_{r+1}$ will be scored as a lesion-localization (LL) and rated r.

Rating assigned to unmarked sites: Unmarked LLs are assigned the zero rating (or any rating lower than the lowest allowed FROC-1 rating). Note that even lesions that were not found by the search stage, and therefore do not qualify as latent LLs, are assigned the zero rating. This is because they represent observable events (and less suspicious than the lowest allowed FROC-1 rating). In contrast, unmarked latent NLs are unobservable events. Unlike lesions there is no a-priori reader-independent list of non-lesion locations; what constitutes a NL is reader dependent, see Fig. 5.4.

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

6.5 Physical meanings of the RSM parameters

The parameters have the following physical meanings:

6.5.1 The μ parameter

The μ parameter is the lesion *perceptual signal to noise ratio pSNR*, as described in (print book) Chapter 12.5.2, between latent NLs and latent LLs. For white noise background this is similar to the physical SNR (Chakraborty, 1997) after correction for the non-linear response of the visual system to visual stimuli (Siddiqui et al., 2005). For clinical backgrounds pSNR is determined by the competition for the observer's foveal attention from other regions 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 task between cued 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 μ parameter by:

$$\mu = \sqrt{2}\Phi^{-1}(\text{PC}) \tag{6.8}$$

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 may become 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. This is a difficult study to conduct as I found out.

6.5.2 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]

[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.

I have found it best to illustrate sampling to non-statistics majors with numerical examples. 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, the value of λ determines the average number of latent NLs.

The following code illustrates Poissson sampling, estimation of the mean and confidence interval for 100 samples from two Poissson distributions. The number of samples has been set to $K_1 = 100$ (the first argument to `rpois()` is the number of non-diseased cases; the second argument is the value of λ).

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

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

```
## K1 =  100 , lambda 1st reader =  1 , lambda 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 Poissson parameter (the mean parameter of the Poissson distribution is frequently referred to as the Poissson 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 , lambda 1st reader =  1 , lambda 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 Poissson 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.

6.5.3 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 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 number of latent LLs on any specific diseased case, or which lesions will be correctly localized, one can predict the average number of latent LLs per diseased case.

The following code uses $K_2 = 100$ samples, the number of diseased cases, each with one lesion. The arguments to `rbinom()` - for random binomial samples - are the number of diseased cases, the number of lesions per case and the value of ν.

```
K2 <- 100
nu <- c(0.5, 0.9)
seed <- 1;set.seed(seed);samples1 <- rbinom(K2,1,nu[1])
seed <- 1;set.seed(seed);samples2 <- rbinom(K2,1,nu[2])

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

```
## K2 =  100 , nu 1st reader =  0.5 , nu 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 (0.379, 0.582). For reader 2 the corresponding estimates are 0.94 and (0.874, 0.978).

As a more complicated but clinically realistic example, consider a dataset with 100 cases where 97 cases 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);nu <- c(0.5, 0.9)
samples1 <- array(dim = c(sum(K2),length(K2)))
seed <- 1;set.seed(seed)
# I am using el instead of l as the latter looks like 1
for (el in 1:length(K2)) {
  samples1[1:K2[el],el] <- rbinom(K2[el],Lk[el],nu[1])
}

samples2 <- array(dim = c(sum(K2),length(K2)))
seed <- 1;set.seed(seed)
for (el in 1:length(K2)) {
  samples2[1:K2[el],el] <- rbinom(K2[el],Lk[el],nu[2])
}

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

[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[1] = 97 , K2[2] = 2 , K2[3] = 1 , nu1 = 0.5 , nu2 = 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).

6.6 Intrinsic RSM parameters

While the parameters λ and ν are physically meaningful a little thought reveals that they must depend on μ. From the solar-analogy described in Section 2.6 we know that if $\mu = 0$ the lesions have zero contrast and therefore cannot be found by the search mechanism implying $\nu = 0$. Moreover attempting to find these zero contrast lesions must generate a large number of non-lesion localizations implying $\lambda = \infty$.

The following is a simple model of the μ dependence of λ and ν. The model re-parameterizes the *physical* parameters λ and ν in terms of *intrinsic* parameters λ_i and ν_i that are μ independent[4]:

$$\left.\begin{aligned} \nu &= 1 - \exp\left(-\mu\nu_i\right) \\ \lambda &= \frac{\lambda_i}{\mu} \end{aligned}\right\} \tag{6.9}$$

The inverse transformations are:

$$\left.\begin{aligned} \nu_i &= -\frac{\ln\left(1-\nu\right)}{\mu} \\ \lambda_i &= \mu\lambda \end{aligned}\right\} \tag{6.10}$$

The intrinsic parameters obey $\lambda_i \geq 0$ and $\nu_i \geq 0$.

Since it determines ν, the ν_i 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*:

$$\nu_i = \left(\frac{\partial\nu}{\partial\mu}\right)_{\mu=0} \tag{6.11}$$

According to Eqn. (6.9), as $\mu \to \infty$, $\nu \to 1$ and conversely, as $\mu \to 0$, $\nu \to 0$. The solar analogy in Section 2.6 is instructive. The dependence of ν on μ is consistent with the fact that higher contrast lesions are easier to find. A non-expert is expected to find a high contrast lesion whereas a low contrast lesion will be more difficult to find even by an expert observer.

According to Eqn. (6.9) 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 solar analogy in Section 2.6 is instructive. Since the sun has very high contrast, there is no reason for the observer to search for other suspicious regions which have no possibility of resembling it. On the other hand, attempting to locate a faint star can generate many false sightings because the expected small contrast from the faint real star could be comparable to that from a number of regions in the nearby background.

[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) became evident more recently.

6.7 Summary

This chapter has described a statistical parameterization of the Nodine-Kundel model of visual search. The model accounts for key aspects of the process:

- Search: finding lesions and finding non-lesions. These are characterized by the two search parameters λ and ν.
- Classification: The ability to correctly rate a lesion higher than a NL is characterized by the third (classification) parameter of the model μ.

While the 2 search parameters have relatively simple physical meanings they depend on μ. Consequently, it is necessary to introduce intrinsic parameters λ_i and ν_i which are independent of μ.

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

Chapter 7

ROC predictions of the RSM

7.1 TBA How much finished 90%

7.2 TBA Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes ROC-related predictions of the RSM. The next chapter will describe the FROC, AFROC and wAFROC curve predictions.

The inferred-ROC z-sample and the end-point of the ROC are defined and expressions in terms of RSM parameters are derived. Derived next is the predicted *inferred ROC* curve and the probability density functions of the inferred-ROC z-samples for non-diseased and diseased cases. Integrating the total area under the predicted ROC yields ROC-AUC.

Since the ROC is a basic measure of performance, numerical examples are given showing the behavior of the operating point as parameters of the RSM are varied.

In this chapter formulae for RSM quantities are given in terms of the RSM search parameters λ and ν.

7.3 Inferred ROC z-sample

The inferred ROC z-sample of a case, denoted h_t, where $t = 1$ for non-diseased cases and $t = 2$ for diseased cases, is the z-sample of the highest rated latent mark on the case or $-\infty$ if the case has no latent marks. The difference from the previous chapter is that in this chapter we are concerned with statistical/probabilistic modeling of the continuous z-samples instead of describing observed ratings for a finite dataset.

Definitions:

- FPF(ζ) = probability that $h_1 \geq \zeta$.
- TPF(ζ) = probability that $h_2 \geq \zeta$.

Accordingly, FPF and TPF are defined by:

$$\mathrm{FPF}(\zeta) = \mathrm{P}(h_1 \geq \zeta) \tag{7.1}$$

$$\mathrm{TPF}(\zeta) = \mathrm{P}(h_2 \geq \zeta) \tag{7.2}$$

Definition of ROC plot:

- The ROC is the plot of TPF(ζ) vs. FPF(ζ).
- *The plot includes a straight line extension from the theoretical end-point to (1,1).*
- The theoretical end-point corresponds to $\zeta = -\infty$.

7.4 End-point of the ROC

A consequence of the possibility that some cases have no marks is that the ROC curve has the *end-point-discontinuity 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 $\text{FPF}_{\max}$ and $\text{TPF}_{\max}$ are generally less than unity.

Starting from ∞ as ζ is lowered to $-\infty$ some of the cases that had at least one latent site but whose z-sample did not exceed ζ will now generate marks and contribute to FPF and TPF resulting in upward and rightward movement of the theoretical operating point until eventually *only cases with no latent sites* remain. These cases cannot generate marks. The finite number of cases with no marks has the consequence that the uppermost continuously accessible operating point is below-left of (1,1). The (1,1) point is "trivially" reached when one cumulates cases with no marks, i.e., those rated $-\infty$.

This behavior is distinct from conventional ROC models where the entire curve, extending from (0, 0) to (1, 1), is continuously accessible. This is because every case yields a finite decision variable, no matter how small. The number of cases with $-\infty$ rating is zero. When $\zeta = -\infty$ the operating point reaches (1,1).

7.4.1 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_1 and with an appropriately low ζ it will be marked. The probability of *zero* latent NLs, see Eqn. (6.2), is:

$$\text{pmf}_P(0, \lambda) = \exp(-\lambda)$$

.

The probability that the case has *at least one* latent NL is the complement of the above probability. At sufficiently low ζ each of these cases yields a marked non-disease case. Therefore, the maximum continuously accessible abscissa of the ROC, i.e., $\text{FPF}_{\max}$, is:

$$\text{FPF}_{\max} = 1 - \exp(-\lambda) \tag{7.3}$$

7.4.2 The ordinate of the ROC end-point

A diseased case has no marks, even for very low ζ, 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. (6.4), $\text{pmf}_B(0, L, \nu) = (1-\nu)^L$.

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

- 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

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

.

The probability that there exists *at least one* latent site is the complement of the above expression, which equals $\mathrm{TPF}_{\max}$, i.e.,

$$\mathrm{TPF}_{\max} = 1 - \exp(-\lambda)(1-\nu)^L \tag{7.4}$$

7.4.3 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:

$$\sum_{L=1}^{L_{max}} f_L = 1 \tag{7.5}$$

By restricting attention to the set of diseased cases with L lesions each, Eqn. (7.4) for $\mathrm{TPF}_{\max}$ applies. Since TPF is a probability and probabilities of independent processes add it follows that:

$$\mathrm{TPF}_{\max} = 1 - \exp(-\lambda) \sum_{L=1}^{L_{max}} f_L (1-\nu)^L \tag{7.6}$$

The ordinate of the end-point is a f_L weighted summation of $\mathrm{TPF}_{\max}$. The expression for $\mathrm{FPF}_{\max}$ is unaffected.

7.5 ROC curve

On the continuous ROC curve each case has at least one mark and the ROC decision variable is the rating of the highest rated mark h_t on the case. Therefore Eqn. (7.1) and Eqn. (7.2) apply. Varying the threshold parameter ζ from ∞ to $-\infty$ sweeps out the continuous section of the predicted ROC curve from (0,0) to $(\mathrm{FPF}_{\max}, \mathrm{TPF}_{\max})$.

7.5.1 Derivation of FPF

- Assumption 4: the z-samples of latent 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 6.4 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 $\Phi(\zeta)$. The probability that all n z-samples are smaller than ζ is $(\Phi(\zeta))^n$. If all z-samples are smaller than ζ, then the highest z-sample h_t must be smaller than ζ. Therefore, the probability that h_t exceeds ζ is:

$$\left.\begin{aligned} \mathrm{FPF}(\zeta \mid n) &= P(h_1 \geq \zeta \mid n) \\ &= 1 - [\Phi(\zeta)]^n \end{aligned}\right\} \tag{7.7}$$

The conditioning notation in Eqn. (7.7) reflects the fact that this expression applies specifically to non-diseased cases each with n latent NLs. To obtain $\mathrm{FPF}_{\max}$ one performs a Poissson pmf-weighted summation of $\mathrm{FPF}(\zeta \mid n)$ over n from 0 to ∞ (the inclusion of the $n = 0$ term is explained below):

$$\mathrm{FPF}(\zeta, \lambda) = \sum_{n=0}^{\infty} \mathrm{pmf}_{\mathrm{Pois}}(n, \lambda)\, \mathrm{FPF}(\zeta \mid n) \tag{7.8}$$

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

```
# 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:
Poissson := proc (n, lambda) lambda^n*exp(-lambda)/factorial(n) end:
B := proc (l, L, nu) binomial(L, l)*nu^l*(1-nu)^(L-l) end:
FPF := proc(zeta,lambda) sum(Poissson(n,lambda)*
(1 - PHI(zeta,0)^n), n=0..infinity);end:
FPF(zeta, lambda);
```

The above code yields:

$$\mathrm{FPF}\left(\zeta,\lambda\right)=1-\exp\left(-\frac{\lambda}{2}\left[1-\mathrm{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \tag{7.9}$$

The error function in Eqn. (7.9) is related to the unit normal CDF function $\Phi(x)$ by:

$$\mathrm{erf}\left(x\right)=2\Phi\left(\sqrt{2}x\right)-1 \tag{7.10}$$

Using this transformation yields the following simpler expression for FPF:

$$\mathrm{FPF}\left(\zeta,\lambda\right)=1-\exp\left(-\lambda\Phi\left(-\zeta\right)\right) \tag{7.11}$$

The `R` implementation follows:

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

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

7.5.2 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 $[\Phi(\zeta-\mu)]^l$. The probability that all latent marks have z-samples less than ζ is the product of these two probabilities. The probability that h_2 (the highest z-sample on a diseased case) is larger than ζ is the complement of the product probabilities, i.e.,

$$\left.\begin{aligned}\mathrm{TPF}\left(\zeta,\mu,n,l,L\right)=&P\left(h_2\geq\zeta\mid\mu,n,l,L\right)\\=&1-\left[\Phi\left(\zeta\right)\right]^n\left[\Phi\left(\zeta-\mu\right)\right]^l\end{aligned}\right\} \tag{7.12}$$

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

$$\left.\begin{aligned}\mathrm{TPF}\left(\zeta,\mu,\lambda,\nu\right) = &\sum_{n=0}^{\infty}\mathrm{pmf}_P(n,\lambda)\\ &\times\sum_{l=0}^{L}\mathrm{pmf}_B(l,\nu,L)\mathrm{TPF}_{n,l}\left(\zeta,\mu,n,l\right)\end{aligned}\right\} \tag{7.13}$$

This can be evaluated using Maple^{TM} yielding:

$$\left.\begin{aligned}&\mathrm{TPF}\left(\zeta,\mu,\lambda,\nu,L\right)\\ &=1-\exp\left(-\lambda\Phi\left(-\zeta\right)\right)\left(1-\nu\Phi\left(\mu-\zeta\right)\right)^{L}\end{aligned}\right\} \tag{7.14}$$

7.5.3 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 (7.14) over the fraction of diseased cases with L lesions:

$$\left.\begin{aligned}&\mathrm{TPF}\left(\zeta,\mu,\lambda,\nu,\overrightarrow{f_L}\right) =\\ &1-\exp\left(-\lambda\Phi\left(-\zeta\right)\right)\sum_{L=1}^{L_{max}}f_L\left(1-\nu\Phi\left(\mu-\zeta\right)\right)^{L}\end{aligned}\right\} \tag{7.15}$$

Since $\Phi\left(-\zeta\right)$ tends to unity as $\zeta\rightarrow-\infty$, this expression reduces to Eqn. (7.6) for the ROC end-point. The expression for FPF, Eqn. (7.11), is unaffected.

The `R` implementation follows:

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

7.5.4 ROC decision variable pdfs

In the ROC book, 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:

$$\mathrm{pdf}_N\left(\zeta\right) = \frac{\partial}{\partial\zeta}\left(1-\mathrm{FPF}\left(\zeta,\lambda\right)\right) \tag{7.16}$$

For diseased cases,

$$\mathrm{pdf}_D\left(\zeta\right) = \frac{\partial}{\partial\zeta}\left(1-\mathrm{TPF}\left(\zeta,\mu,\lambda,\nu,\overrightarrow{f_L}\right)\right) \tag{7.17}$$

Both expressions can be evaluated using Maple^{TM}. 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 of ζ):

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_N(\zeta)\, d\zeta = (1-\text{FPF}(\zeta,\lambda))\Big|_{-\infty}^{\infty}\\ =\text{FPF}_{\max}\end{aligned}\right\} \tag{7.18}$$

$$\left.\begin{aligned}\int_{-\infty}^{\infty} \text{pdf}_D(\zeta)\, d\zeta = \left(1-\text{TPF}\left(\zeta,\mu,\lambda,\nu,\overrightarrow{f_L}\right)\right)\Big|_{-\infty}^{\infty}\\ =\text{TPF}_{\max}\end{aligned}\right\} \tag{7.19}$$

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. [1]

7.5.5 ROC AUC

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

$$AUC_{RSM}^{ROC}\left(\mu,\lambda,\nu,\zeta,\overrightarrow{f_L}\right) = \sum_{L=0}^{L_{max}} f_L \int_0^1 \text{TPF}(\zeta,\mu,\lambda,\nu,L)\, d\left(\text{FPF}(\zeta,\lambda)\right) \tag{7.20}$$

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).

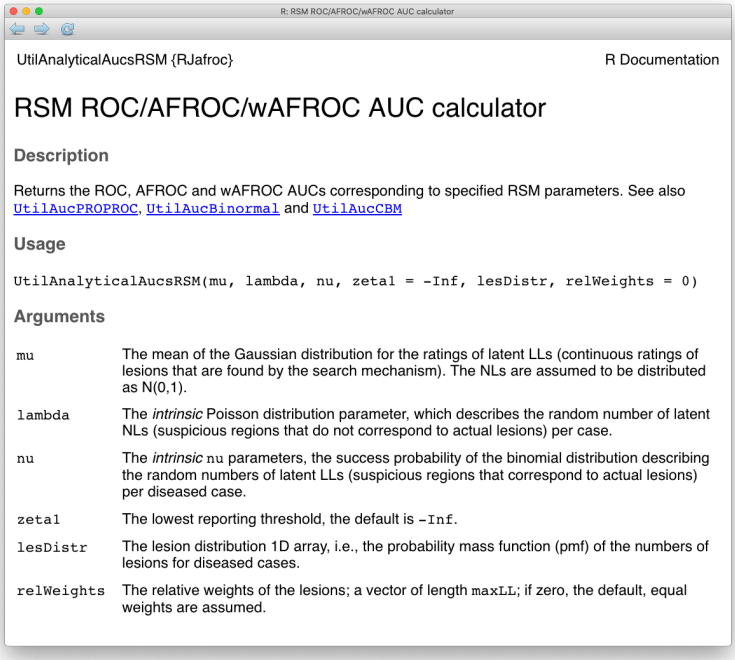
The right hand side of Eqn. (7.20) can be evaluated using a numerical integration function implemented in `R`, which is used in the `RJafroc` function `UtilAnalyticalAucsRSM()` whose help page follows:

The arguments to `UtilAnalyticalAucsRSM()` are the intrinsic RSM parameters μ, λ, ν and ζ. The default value of ζ is $\zeta = -\infty$. 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.

```
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")
```

[1] 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.


R: RSM ROC/AFROC/wAFROC AUC calculator

UtilAnalyticalAucsRSM {RJafroc} R Documentation

RSM ROC/AFROC/wAFROC AUC calculator

Description

Returns the ROC, AFROC and wAFROC AUCs corresponding to specified RSM parameters. See also UtilAucPROPROC, UtilAucBinormal and UtilAucCBM

Usage

```
UtilAnalyticalAucsRSM(mu, lambda, nu, zeta1 = -Inf, lesDistr, relWeights = 0)
```

Arguments

mu	The mean of the Gaussian distribution for the ratings of latent LLs (continuous ratings of lesions that are found by the search mechanism). The NLs are assumed to be distributed as N(0,1).
lambda	The *intrinsic* Poisson distribution parameter, which describes the random number of latent NLs (suspicious regions that do not correspond to actual lesions) per case.
nu	The *intrinsic* nu parameters, the success probability of the binomial distribution describing the random numbers of latent LLs (suspicious regions that correspond to actual lesions) per diseased case.
zeta1	The lowest reporting threshold, the default is -Inf.
lesDistr	The lesion distribution 1D array, i.e., the probability mass function (pmf) of the numbers of lesions for diseased cases.
relWeights	The relative weights of the lesions; a vector of length maxLL; if zero, the default, equal weights are assumed.

Figure 7.1: Help page for 'RJafroc' function 'UtilAnalyticalAucsRSM'.

```
## 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 2.6. 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 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).

7.5.6 Comparing TPF formula to `RJafroc` functions

A hand calculation is shown and compared to the value yielded by the function `RSM_TPF`. The RSM parameters and the value of ζ are:

```
zeta <- 1
mu <- 2
lambda <- 1
nu <- 0.9
lesDistr <- c(0.5,0.5)
```

The `lesDistr` vector corresponds to f_L and specifies $L_{max} = 2$ and 50 percent of diseased cases have one lesion per case and the rest have two lesions per case.

Direct implementation of Eqn. (7.15) followed by usage of the function `RSM_TPF` follows:

```
cat(1-
exp(-lambda*pnorm(-zeta))*
(lesDistr[1]*(1-nu*pnorm(mu-zeta))+
lesDistr[2]*(1-nu*pnorm(mu-zeta))^2))
```

```
## 0.8712655
```

```
cat(RSM_TPF(zeta,mu,lambda,nu, lesDistr = lesDistr))
```

```
## 0.8712655
```

The two values are identical.

7.5.7 Effect on operating point of varying RSM parameters

It is instructive to understand the effects of varying the RSM parameters on the operating point on the ROC curve.

7.5.7.1 Vary μ

```
## lesDistr =  0.1 0.9
```

```
## Varying mu only:
## Other parameters are lambda =  2 , nu =  0.5 , zeta =  0
```

```
## mu = 0 , RSM-x = 0.6321 , RSM-y = 0.7862
## mu = 0.5 , RSM-x = 0.6321 , RSM-y = 0.8342
## mu = 1 , RSM-x = 0.6321 , RSM-y = 0.8676
## mu = 1.5 , RSM-x = 0.6321 , RSM-y = 0.8862
## mu = 2 , RSM-x = 0.6321 , RSM-y = 0.8946
## mu = 2.5 , RSM-x = 0.6321 , RSM-y = 0.8977
## mu = 3 , RSM-x = 0.6321 , RSM-y = 0.8986
## mu = 3.5 , RSM-x = 0.6321 , RSM-y = 0.8988
## mu = 4 , RSM-x = 0.6321 , RSM-y = 0.8988
## mu = 4.5 , RSM-x = 0.6321 , RSM-y = 0.8988
## mu = 5 , RSM-x = 0.6321 , RSM-y = 0.8988
```

The abscissa is independent of μ (because this parameter has no effect on non-diseased cases) and the ordinate is an increasing function of μ (as expected for increasing separation of the LL and NL distributions; the LLs on diseased cases are rated higher causing the distribution of h_2 to shift to higher values).

7.5.7.2 Vary λ

```
## lesDistr =  0.1 0.9


## Varying lambda only:
## Other parameters are mu =  1 , nu =  0.5 , zeta =  0


## lambda = 0.5 , RSM-x = 0.2212 , RSM-y = 0.7196
## lambda = 1 , RSM-x = 0.3935 , RSM-y = 0.7817
## lambda = 1.5 , RSM-x = 0.5276 , RSM-y = 0.8300
## lambda = 2 , RSM-x = 0.6321 , RSM-y = 0.8676
## lambda = 2.5 , RSM-x = 0.7135 , RSM-y = 0.8969
## lambda = 3 , RSM-x = 0.7769 , RSM-y = 0.9197
## lambda = 3.5 , RSM-x = 0.8262 , RSM-y = 0.9374
## lambda = 4 , RSM-x = 0.8647 , RSM-y = 0.9513
## lambda = 4.5 , RSM-x = 0.8946 , RSM-y = 0.9621
## lambda = 5 , RSM-x = 0.9179 , RSM-y = 0.9705
```

The abscissa increases with λ (more NLs on non-diseased cases are generated causing the distribution of h_1 to shift to higher values) and the ordinate also increases with λ (more NLs on diseased cases are generated causing the distribution of h_2 to shift to higher values - recall that on diseased cases the highest z-sample is the maximum of NL and LL z-samples, whichever is highest).

7.5.7.3 Vary ν

```
## lesDistr =  0.1 0.9


## Varying nu only:
## Other parameters are mu =  1 , lambda =  2 , zeta =  0


## nu = 0 , RSM-x = 0.6321 , RSM-y = 0.6321
## nu = 0.1 , RSM-x = 0.6321 , RSM-y = 0.6886
## nu = 0.2 , RSM-x = 0.6321 , RSM-y = 0.7404
## nu = 0.3 , RSM-x = 0.6321 , RSM-y = 0.7875
## nu = 0.4 , RSM-x = 0.6321 , RSM-y = 0.8299
## nu = 0.5 , RSM-x = 0.6321 , RSM-y = 0.8676
## nu = 0.6 , RSM-x = 0.6321 , RSM-y = 0.9006
## nu = 0.7 , RSM-x = 0.6321 , RSM-y = 0.9289
## nu = 0.8 , RSM-x = 0.6321 , RSM-y = 0.9526
## nu = 0.9 , RSM-x = 0.6321 , RSM-y = 0.9716
```

No effect on the abscissa as ν increases (this parameter has no effect on non-diseased case sampling) and the ordinate increases with ν (more LLs on diseased cases, as more lesions are localized, causing the distribution of h_2 to shift to higher values).

7.5.7.4 Vary ζ

```
## lesDistr =  0.1 0.9


## Varying zeta only:
## Other parameters are mu =  1 , lambda =  2 , nu =  0.5
```

```
## zeta = -3 , RSM-x = 0.8643 , RSM-y = 0.9627
## zeta = -2.5 , RSM-x = 0.8630 , RSM-y = 0.9623
## zeta = -2 , RSM-x = 0.8584 , RSM-y = 0.9610
## zeta = -1.5 , RSM-x = 0.8453 , RSM-y = 0.9570
## zeta = -1 , RSM-x = 0.8141 , RSM-y = 0.9467
## zeta = -0.5 , RSM-x = 0.7492 , RSM-y = 0.9224
## zeta = 0 , RSM-x = 0.6321 , RSM-y = 0.8676
## zeta = 0.5 , RSM-x = 0.4605 , RSM-y = 0.7568
## zeta = 1 , RSM-x = 0.2719 , RSM-y = 0.5768
## zeta = 1.5 , RSM-x = 0.1251 , RSM-y = 0.3628
## zeta = 2 , RSM-x = 0.0445 , RSM-y = 0.1831
## zeta = 2.5 , RSM-x = 0.0123 , RSM-y = 0.0740
## zeta = 3 , RSM-x = 0.0027 , RSM-y = 0.0241
```

Increasing ζ causes the operating point to move down the ROC.

7.5.7.5 Vary f_L

The `lesDist` vector is defined as $(f, (1-f))$ where f is varied from 1 (only cases with one lesion per case) to 0 (only cases with two lesions per case):

```
## lesDistr =  (f, 1-f)
```

```
## Varying f only:
## Other parameters are mu =  1 , lambda =  2 , nu =  0.5 , zeta =  0
```

```
## f =  1 , RSM-x =  0.6321 , RSM-y =  0.7869
## f =  0.9 , RSM-x =  0.6321 , RSM-y =  0.7958
## f =  0.8 , RSM-x =  0.6321 , RSM-y =  0.8048
## f =  0.7 , RSM-x =  0.6321 , RSM-y =  0.8138
## f =  0.6 , RSM-x =  0.6321 , RSM-y =  0.8227
## f =  0.5 , RSM-x =  0.6321 , RSM-y =  0.8317
## f =  0.4 , RSM-x =  0.6321 , RSM-y =  0.8407
## f =  0.3 , RSM-x =  0.6321 , RSM-y =  0.8496
## f =  0.2 , RSM-x =  0.6321 , RSM-y =  0.8586
## f =  0.1 , RSM-x =  0.6321 , RSM-y =  0.8676
## f =  0 , RSM-x =  0.6321 , RSM-y =  0.8765
```

No effect on FPF but TPF increases as more lesions per case means more LLs per case and the distribution of h_2 moves to higher values.

7.5.8 Sample ROC curves

Fig. 7.2 displays ROC curves for indicated values of μ. The remaining RSM model parameters are $\lambda = 1$, $\nu = 1$ and $\zeta = -\infty$ 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.

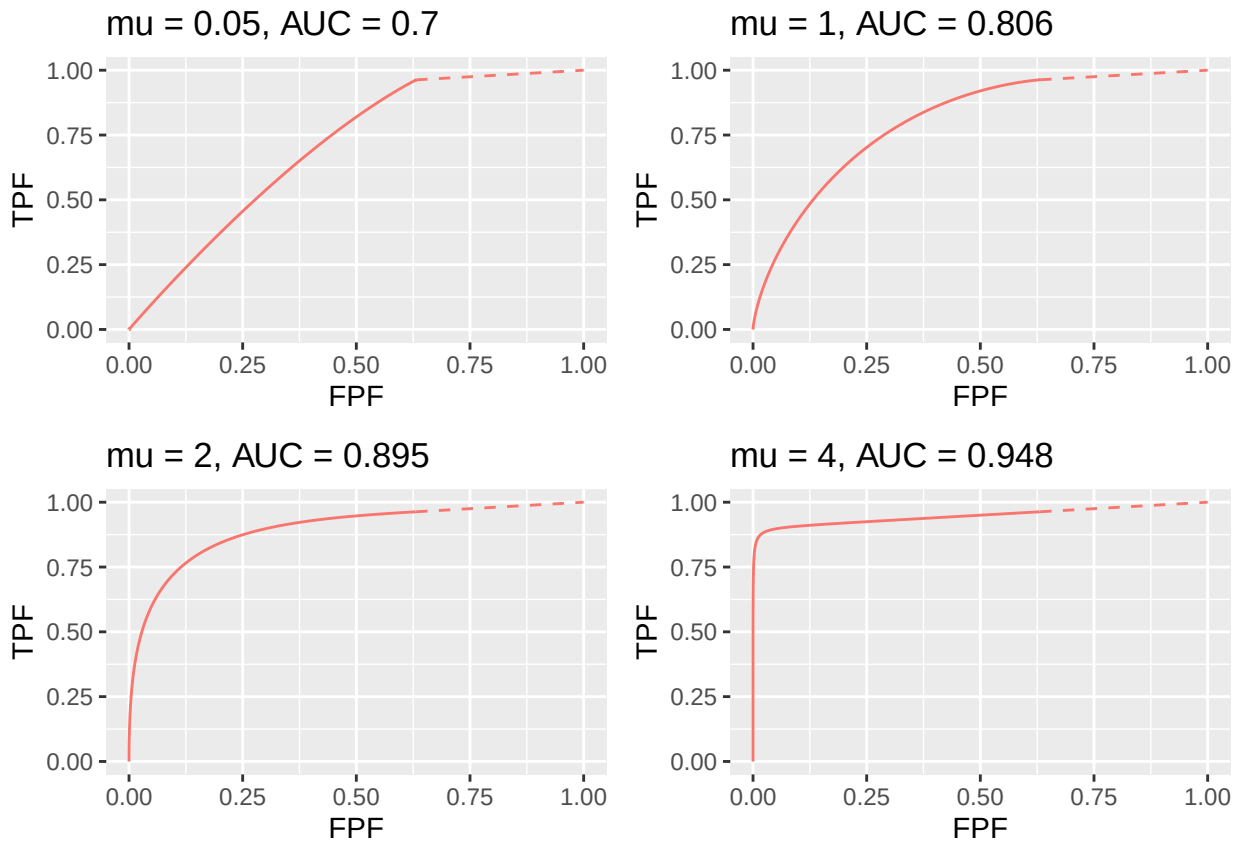


Figure 7.2: ROC curves for indicated values of the μ parameter. Notice the transition, as μ increases, from near chance level performance to almost perfect performancea as the end-point moves from near (1,1) to near (0,1).

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. (7.3)
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.

7.5.9 Sample RSM pdf curves

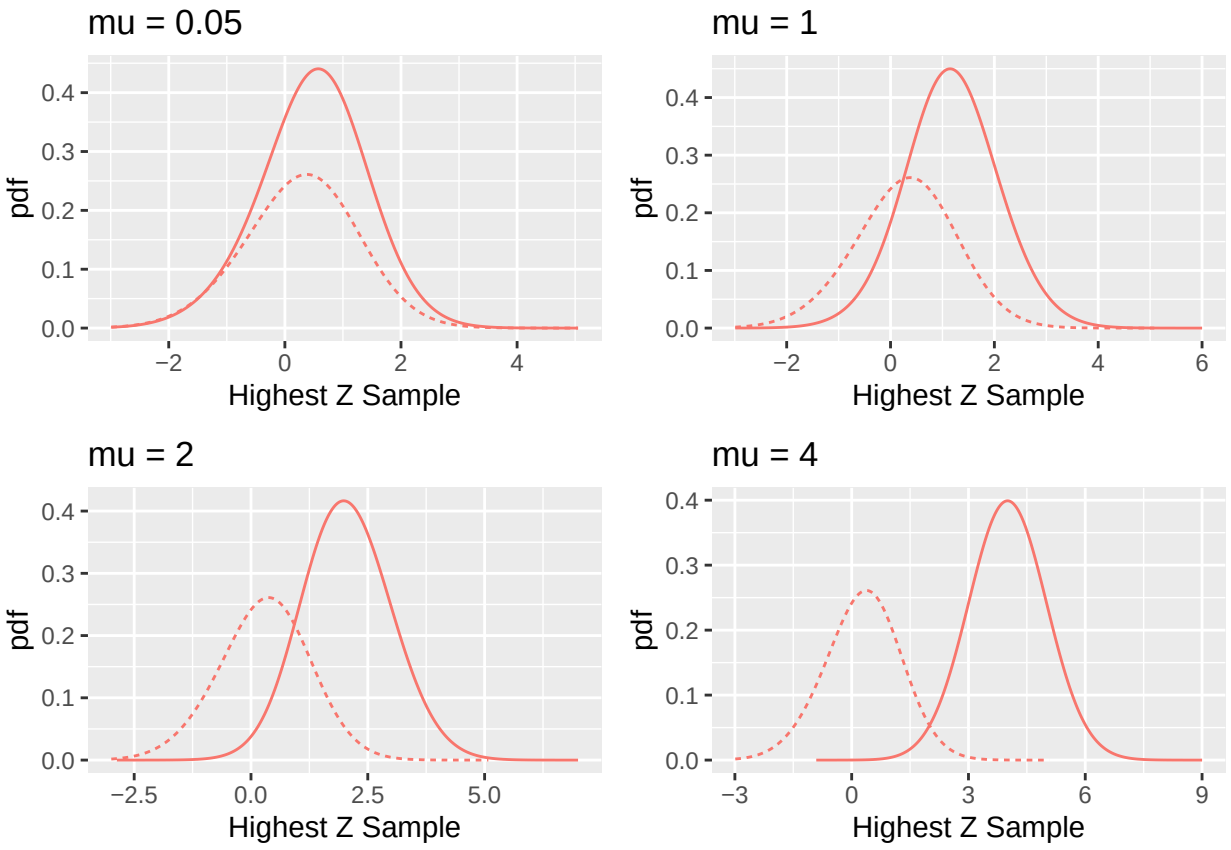


Figure 7.3: RSM pdf curves for indicated values of the μ parameter. The solid curve corresponds to diseased cases and the dotted curve corresponds to non-diseased cases.

Fig. 7.3 shows pdf plots for the same values of parameters as in Fig. 7.2.

Consider the plot of the pdfs for $\mu = 1$. Since the integral of a pdf function over an interval amounts to counting the fraction of events occurring in the interval, it should be evident that the area under the non-diseased pdf equals FPF_{max} and that under the diseased pdf equals TPF_{max}. For the chosen value $\lambda = 1$ one has $\text{FPF}_{\text{max}} = 1 - e^{-\lambda} = 0.632$. 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 $\mathrm{TPF}_{\max}$, Eqn. (7.4), which is also less than unity. For the chosen values of $\mu = \lambda = \nu = L = 1$ it equals $\mathrm{TPF}_{\max} = 1 - e^{-\lambda}e^{-\nu} = 0.865$. 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 $\mathrm{FPF}_{\max}$ to further decrease.

The plot in Fig. 7.3 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).

7.6 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 below that the RSM predicted ROC curve, including the dashed straight line extension, is proper* [2].

Consider first the continuous section which is below-left of the end-point. For convenience one abbreviates FPF and TPF to x and y, respectively, and suppresses the dependence on model parameters. From Eqn. (7.11) and Eqn. (7.15) one can express the ROC coordinates as:

$$\left.\begin{aligned} x(\zeta) &= 1 - G(\zeta) \\ y(\zeta) &= 1 - F(\zeta)\, G(\zeta) \end{aligned}\right\} \tag{7.21}$$

where:

$$\left.\begin{aligned} G(\zeta) &= \exp\left(-\lambda \Phi(-\zeta)\right) \\ F(\zeta) &= \sum_{L=1}^{L_{max}} f_L \left(1 - \nu \Phi(\mu - \zeta)\right)^L \end{aligned}\right\} \tag{7.22}$$

These equations have 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 is proper also applies to the present situation.

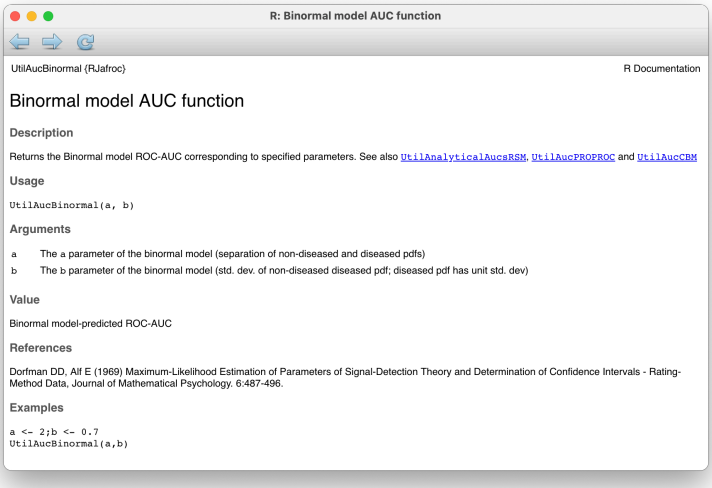
Specifically, since the Φ function ranges between 0 and 1 and $0 \le \nu \le 1$, it follows that $F(\zeta) \le 1$. Therefore $y(\zeta) \ge x(\zeta)$ and the ROC curve is constrained to the upper half of the ROC space, namely the portion above the chance diagonal. Additionally, the more general constraint 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 $(\mathrm{FPF}_{\max}, \mathrm{TPF}_{\max})$, the coordinates of the RSM end-point. This implies that the slope decreases monotonically and also rules out curves with "hooks".

In Appendix 1 7.9 it is shown analytically that the slope is continuous at the end-point transition from the continuous curve to the dashed straight line. In Appendix 2 7.10 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.

[2] The statement in the print book that the "proper" property only applies to the continuous section is incorrect.

7.7 ζ 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. This is because in the ROC model every case yields a rating, no matter how low the z-sample, implying that effectively $\zeta = -\infty$. The lack of ζ dependence is demonstrated by the help page for function `UtilAucBormal`, 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 7.4: Help page for 'RJafroc' function 'UtilAucBormal'.

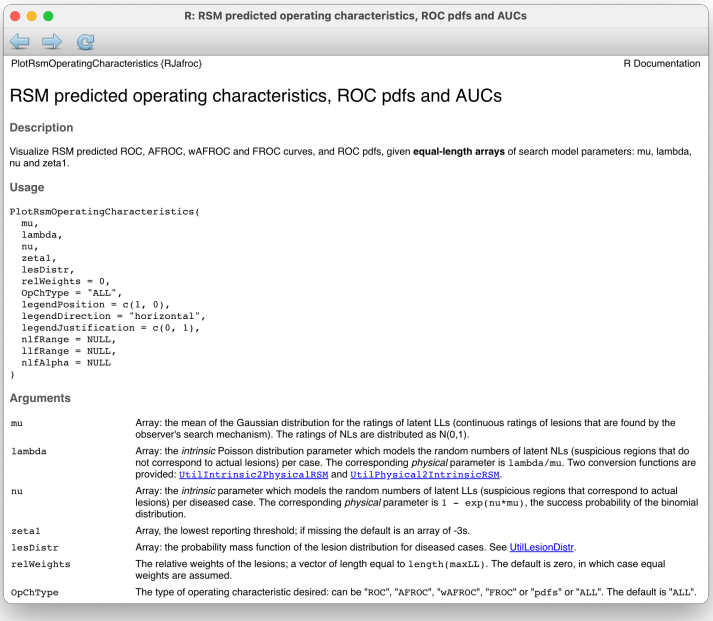
In contrast, in addition to the basic RSM parameters, i.e., μ, λ and ν, the rsm-predictions have an additional dependence on ζ. This is because the value of ζ determines the location of the end-point. The ζ dependence is demonstrated next for the ROC plots, but it is true for all RSM predictions.

The dependence is demonstrated next for two values: $zeta = -10$ and $zeta = 1$. 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),
     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 0.9591597 for the red curve and 0.9337196 for the green curve.

A consequence of the ζ dependence is that if one uses ROC AUC as the measure of performance, the optimal threshold is $\zeta = -\infty$. 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 11.



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 7.5: Help page for 'RJafroc' function 'PlotRsmOperatingCharacteristics'.

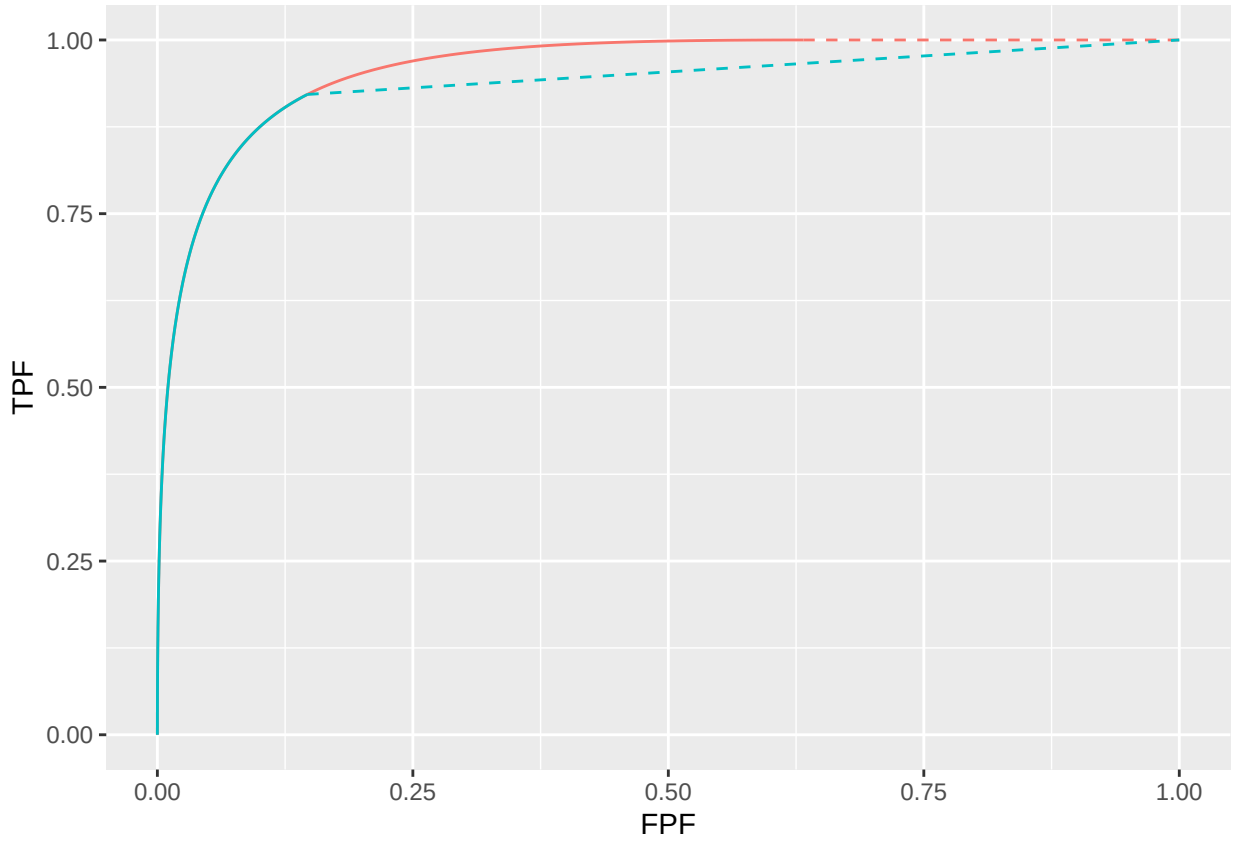


Figure 7.6: ROC curves for two values of ζ: both curves correspond to $\mu = 2$, $\nu = 1$ and $\lambda = 1$. The red curve corresponds to $\zeta = -10$ and the blue curve to $\zeta = 1$.

7.8 TBA Discussion / Summary

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.

7.9 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. (7.3) for $\mathrm{FPF}_{\max}$ and Eqn. (7.6) for $\mathrm{TPF}_{\max}$. Therefore, the slope m_{st} of the dashed straight line is:

$$\left.\begin{aligned} m_{st} =& \frac{1-\mathrm{TPF}_{\max}}{1-\mathrm{FPF}_{\max}} \\ =& \frac{\sum_{L=1}^{L_{max}} f_L (1-\nu)^L \exp(-\lambda)}{\exp(-\lambda)} \\ =& \sum_{L=1}^{L_{max}} f_L (1-\nu)^L \end{aligned}\right\} \tag{7.23}$$

On the continuous section, $g \equiv \mathrm{FPF}$ and $h \equiv \mathrm{TPF}$ are defined by (7.11) and (7.15), respectively. Therefore,

$$\left.\begin{aligned} g =& 1-\exp(-\lambda\Phi(-\zeta)) \\ h =& 1-\exp(-\lambda\Phi(-\zeta)) \sum_{L=1}^{L_{max}} f_L (1-\nu\Phi(\mu-\zeta))^L \end{aligned}\right\} \tag{7.24}$$

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

$$\left.\begin{aligned} \frac{dh}{dg} = & \sum_{L=1}^{L_{max}} f_L (1-\nu\Phi(\mu-\zeta))^{L-1} \times \\ & \left[\frac{L\nu\phi(\mu-\zeta)}{\lambda\phi(-\zeta)} + (1-\nu\Phi(\mu-\zeta))\right] \end{aligned}\right\} \tag{7.25}$$

Using the following result:

$$\left.\begin{aligned} & \lim_{\zeta\to-\infty} \frac{\phi(\mu-\zeta)}{\phi(-\zeta)} \\ =& \lim_{\zeta\to-\infty} \frac{\exp\left(\frac{-(\mu-\zeta)^2}{2}\right)}{\exp\left(\frac{-\zeta^2}{2}\right)} \\ =& \lim_{\zeta\to-\infty} \exp\left(-\frac{1}{2}(\mu^2-2\mu\zeta)\right) \\ =& 0 \end{aligned}\right\} \tag{7.26}$$

it follows that:

$$\left.\begin{aligned} & \lim_{\zeta\to-\infty} \frac{dh}{dg} \\ =& \sum_{L=1}^{L_{max}} f_L (1-\nu)^{L-1} (1-\nu) \\ =& \sum_{L=1}^{L_{max}} f_L (1-\nu)^L \\ =& m_{st} \end{aligned}\right\} \tag{7.27}$$

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).

7.10 Appendix 2: Numerical illustration of continuity

The code in this section examines the slope of the ROC curve as one approaches the end-point $\zeta = -\infty$. 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 <- 0.2
nu <- 0.8
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, lambda)
maxTPF <- TPF (-Inf, mu, lambda, nu, lesDistr)
mStLine <- (1 - maxTPF) / (1 - maxFPF)
```

The end-point coordinates are (0.1812692, 0.9410514) and the slope is 0.072. Next one calculates and displays the ROC curve.

```
ret <- PlotRsmOperatingCharacteristics(
  mu,
  lambda,
  nu,
  zeta1 = -Inf,
  OpChType = "ROC",
  lesDistr = lesDistr,
  legendPosition = "none"
)
```

At first sight the slope appeared to me to be discontinuous at the end-point [3] but this is not true. In fact the slope decreases as one approaches the end-point, and in the limit equals that of the dashed line. This is demonstrated by the next code section which creates a finely-spaced ζ array ranging from -3 to -20. These are the points at which the slope is evaluated numerically. 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, lambda) - FPF (zeta2, lambda)
  delta_TPF <- TPF (zeta, mu, lambda, nu, lesDistr) -
    TPF (zeta2, mu, lambda, nu, lesDistr)
  mAnal <- delta_TPF / delta_FPF
  y1[i] <- mAnal
  # end normal precision arithmetic
```

[3] Others have stated a different visual impression.

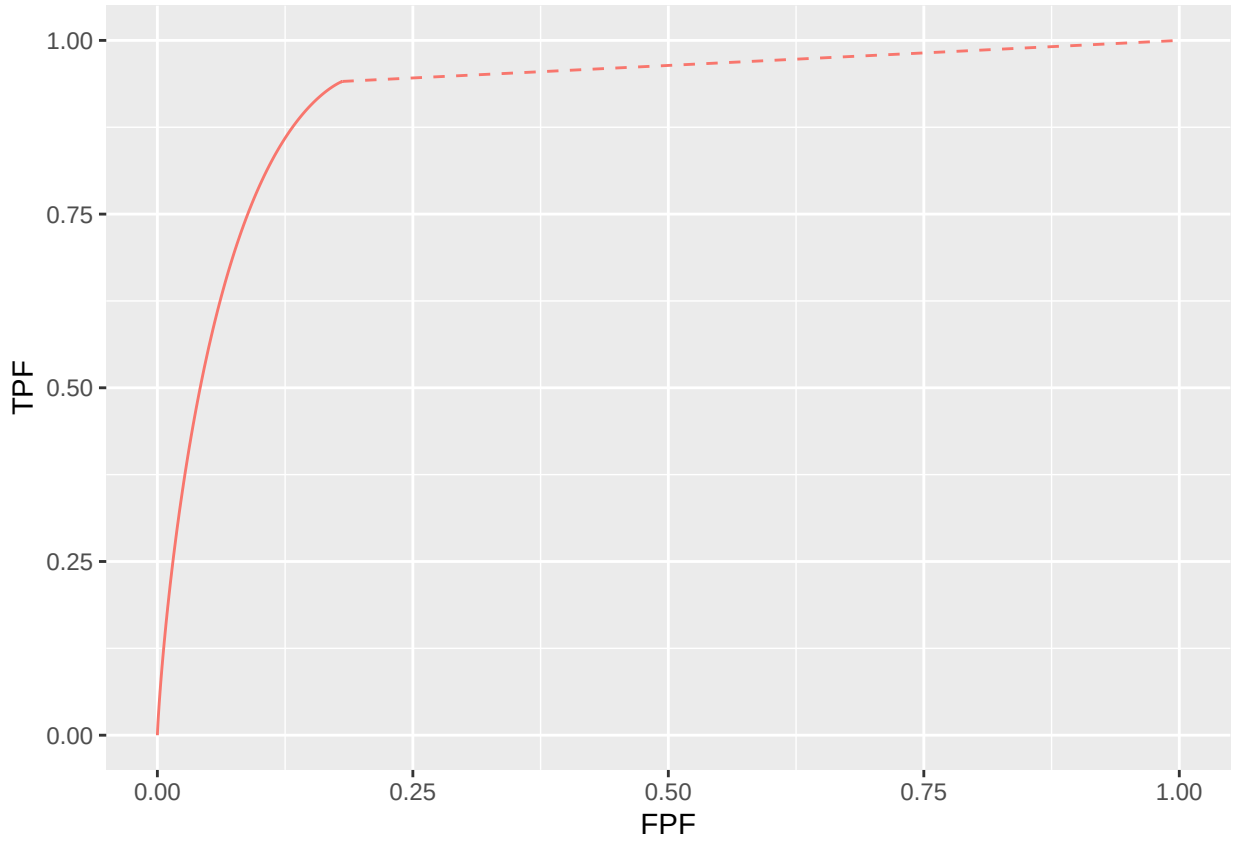


Figure 7.7: ROC curve for selected RSM parameters. The slope of the dashed line is 0.4935272.

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

The next code section displays 3 plots.

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

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

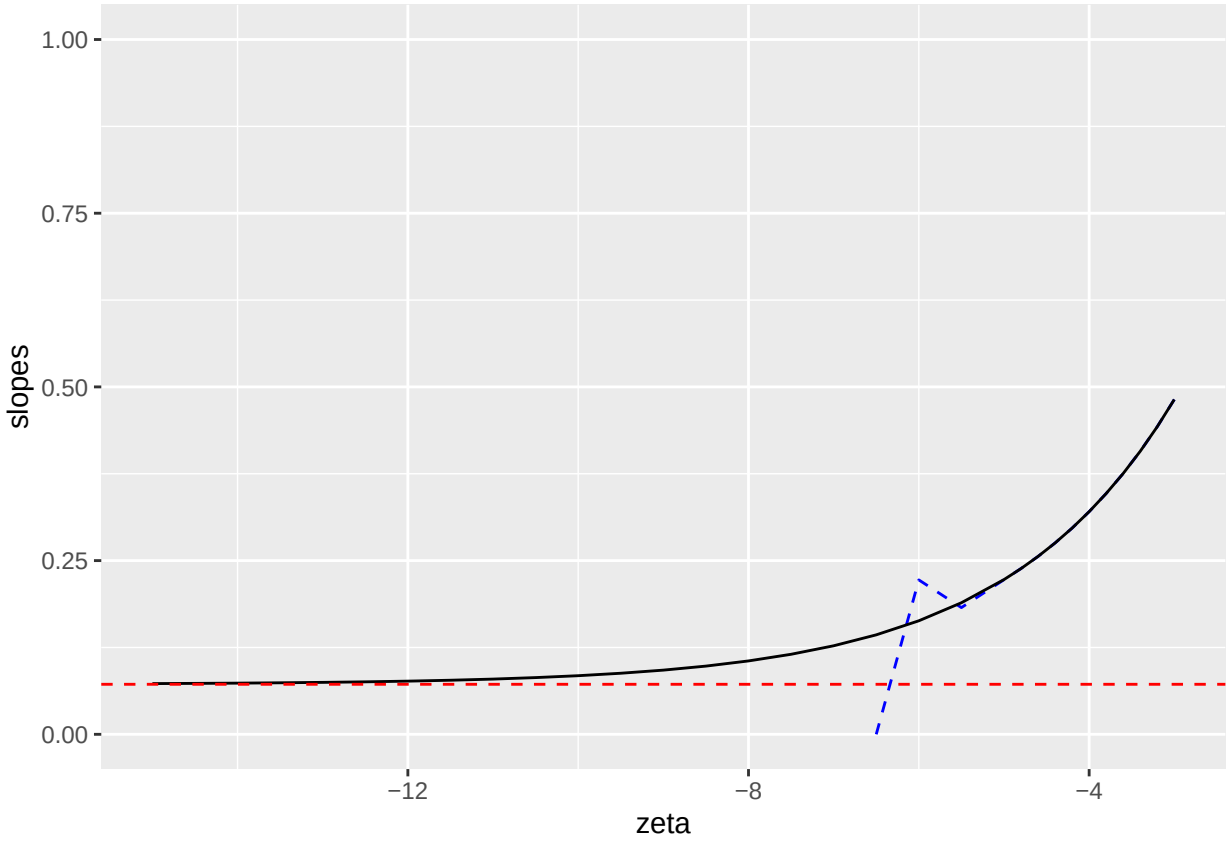


Figure 7.8: Horizontal dashed red line: the value of `mStLine`, the slope of the straight line connecting the ROC end-point to (1,1). Dashed blue line: slope using double precision arithmetic. Solid black line: slope using multiple precision arithmetic - this curve approaches the limiting value `mStLine`.

Chapter 8

Other RSM predictions

8.1 TBA How much finished 95%

Work on intro Need better word than constrained finite? discontinuous? end-point-discontinuity?

8.2 TBA Introduction

Chapter 3 described ROC, FROC, AFROC and wAFROC empirical plots and illustrated them using an actual FROC dataset. Chapter 7 described the ROC curve and related quantities predicted by the radiological search model (RSM). This chapter describes the FROC, AFROC and wAFROC curve predictions of the RSM.

Use of a parametric model allows greater insight into the RSM predictions, for example the limiting slopes of the plots at the origin and the end-point, than is possible with empirical plots. Systematic variation of the parameters quantifies some of the expectations based on the solar analogy in Section 2.6. This chapter also illustrates the need for using reasonable values of the parameters - this is achieved using the intrinsic λ_i, ν_i parameters, described in Section 6.6. While the physical parameters λ, ν are easier to understand as relate immediately to the FROC plot end-point, assigning arbitrary values to them can lead to unrealistic predictions.

8.3 RSM-predicted FROC curve

From a property of the Poisson distribution, namely its mean equals the λ parameter of the distribution, it follows that the expected number of *latent* NLs per case is λ. Recalling that NL z-samples are distributed as N(0,1), one multiplies λ by the probability that a z-sample from $N(0,1)$ exceeds ζ, i.e. by $\Phi(-\zeta)$, to obtain the expected number of *marked* NLs per case, i.e., NLF:

$$\mathrm{NLF}(\zeta, \lambda) = \lambda \Phi(-\zeta) \tag{8.1}$$

We calculate LLF as follows:

$$\left.\begin{aligned} \mathrm{LLF}\left(\mu, \lambda, \nu, \overrightarrow{f_L}\right) =& \Phi(\mu - \zeta) \sum_{L=1}^{L_{max}} f_L \frac{1}{L} \sum_{l_2=0}^{L} l_2 \ \mathrm{pmf}_B(l_2, L, \nu) \\ =& \Phi(\mu - \zeta) \sum_{L=1}^{L_{max}} f_L \frac{1}{L} L \nu \\ =& \nu \Phi(\mu - \zeta) \end{aligned}\right\} \tag{8.2}$$

The inner summation is over all cases with L lesions. One calculates the expected value of l_2 (the number of lesions that are latent LLs) using the binomial distribution of l_2; one divides by L to get the average fraction of LLs on cases with L lesions; one performs an average using the distribution f_L of cases with L lesions; since LL z-samples are distributed as $N(\mu,1)$, one multiplies by the probability that a z-sample from $N(\mu,1)$ exceeds ζ, i.e. by $\Phi(\mu-\zeta)$, to obtain the expected number of *marked* LLs per case, i.e., LLF. Note that the final expression for LLF is independent of the number of lesions in the dataset or their distribution.

The coordinates of the RSM-predicted operating point on the FROC curve for threshold ζ are given by Eqn. (8.1) and Eqn. (8.2). The FROC curve starts at (0,0) and ends at (λ,ν) – the end-point. The end-point is not constrained to lie within the unit-square, rather it is *semi-constrained*: while the maximum ordinate, i.e., ν, cannot exceed unity the maximum abscissa, i.e., λ, can.

The clear connection between λ and ν and the FROC end-point is the reason they are called the *physical* RSM parameters. For reasons explained in Section 2.6 the physical parameters are not the best way of characterizing predicted RSM curves: they hide an inherent μ dependence ignoring which can lead to unreasonable choices of RSM parameters (see Appendix 8.7.1). Intrinsic parameters λ_i,ν_i were introduced in Section 6.6 which are independent of μ. For convenience the transformations between physical and intrinsic parameters are reproduced here:

$$\left.\begin{aligned}\nu &= 1-\exp\left(-\mu\nu_i\right)\\ \lambda &= \frac{\lambda_i}{\mu}\end{aligned}\right\} \tag{8.3}$$

The predicted FROC, AFROC and wAFROC curves that follow use the intrinsic λ_i,ν_i parameters.

8.3.1 FROC plots λ_i,ν_i parameterization

The following code generates FROC plots using the intrinsic $\lambda_i = 2$ and $\nu_i = 0.5$ parameters for 4 values of μ contained in the array `muArr <- c(0.1,1,2,4)` (to avoid a divide by zero error the value $\mu = 0$ is not allowed). A `list` array `p_FROC_lambdai_nui` is created to hold the four plots[1]. The intrinsic λ_i,ν_i parameters are converted to λ,ν using the function `Util2Physical()`which implements Eqn. (8.3)). The parameters are displayed using the `cat()` function. The plots are generated using `PlotRsmOperatingCharacteristics()`. Online help on this function is available. The code-line `p_FROC_lambdai_nui[[i]] <- ret1$FROCPlot` saves the plot to the previously created `list` array.

```
muArr <- c(0.1,1,2,4)
lambda_i <- 2
nu_i <- 0.5
p_FROC_lambdai_nui <- array(list(), dim = c(length(muArr)))
for (i in 1:length(muArr)) {
  mu <- muArr[i]
  ret <- Util2Physical(mu, lambda_i = lambda_i, nu_i = nu_i)
  lambda <- ret$lambda
  nu <- ret$nu
  cat(sprintf("lambda = %6.3f, nu = %4.3f", lambda, nu), "\n")
  ret1 <- PlotRsmOperatingCharacteristics(
    mu, lambda, nu,
    OpChType = "FROC",
    legendPosition  = "none",
    nlfRange = c(0,4),
    llfRange = c(0,1)
  )
  p_FROC_lambdai_nui[[i]] <- ret1$FROCPlot
  #+ ggtitle(paste0("mu = ", as.character(muArr[i])))
}
```

[1]Notation: `p_` stands for a plot array, `FROC` stands for type of plot (also allowed are `AFROC` and `wAFROC`), `lambdai` stands for λ_i and `nui` stands for ν_i (also allowed are `lambda` for λ and `nu` for ν).

```
## lambda = 20.000, nu = 0.049
## lambda =  2.000, nu = 0.393
## lambda =  1.000, nu = 0.632
## lambda =  0.500, nu = 0.865
```

The following code displays the 4 plots.

```
suppressWarnings(grid.arrange(
p_FROC_lambdai_nui[[1]],
p_FROC_lambdai_nui[[2]],
p_FROC_lambdai_nui[[3]],
p_FROC_lambdai_nui[[4]], ncol = 2))
```

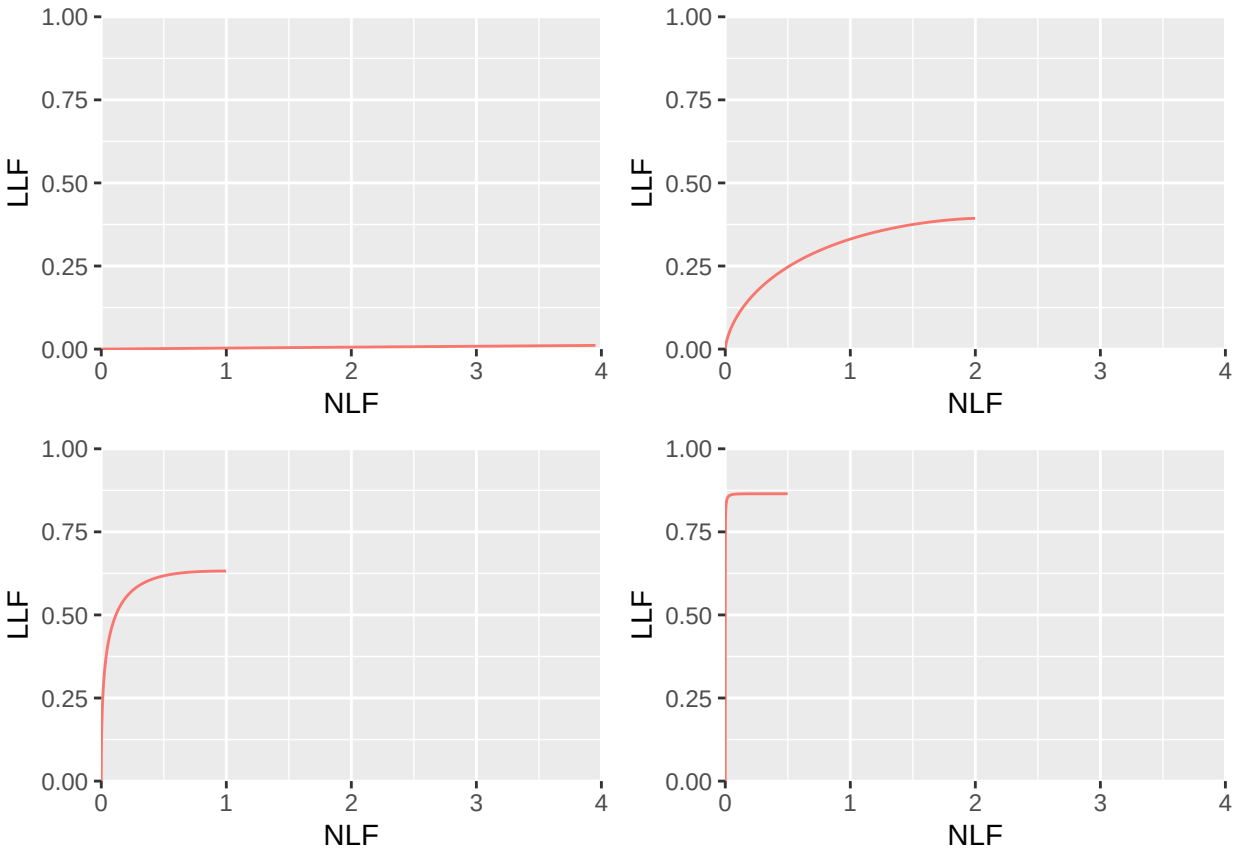


Figure 8.1: RSM-predicted FROC curves using intrinsic parameters $\lambda_i = 2$ and $\nu_i = 0.5$. Top left: $\mu = 0.1$; Top right: $\mu = 1$; Bottom left: $\mu = 2$; Bottom right: $\mu = 4$. Some plots are **not** contained within the unit square which makes it impossible to use the FROC-AUC as a figure of merit.

- In the top-left plot (corresponding to $\mu = 0.1$) because $\lambda = 20$ defines the end-point abscissa and $\nu = 0.049$ defines the end-point ordinate, the FROC curve is close to the x-axis ending at $(20, 0.049)$. For small μ this is close to the chance line FROC. Recall the solar analogy in Section 2.6. When lesion contrast is low the search mechanism has little chance of finding lesions (leading to small LLF) and generates many NLs in attempting to do so (leading to large NLF).
- Increasing μ to 1 decreases λ to 2 and simultaneously increases ν to 0.393. The new end-point $(2, 0.393)$ is evident in the upper-right plot.
- Further increase in μ decreases the abscissa of the end-point and increases the ordinate and the end-point approaches the top-left corner of the FROC plot.

- The perfect performance FROC curve is the vertical line connecting the origin to (0,1). It occurs when $\mu = \infty$.
- Since the FROC end-point is not constrained to lie within the unit square it is not possible, using the FROC-AUC, to defined a valid figure of merit.
- Other characteristics of FROC curves (e.g., slopes at the origin and the end-point) and differences between intrinsic and physical parameterizations of this curve, are explored in Appendix 8.7.1.

8.4 RSM-predicted AFROC curve

The AFROC x-coordinate is the same as the ROC x-coordinate and Eqn. (7.11) applies. The AFROC y-coordinate is identical to the FROC y-coordinate and Eqn. (8.2) applies. Therefore:

$$\left.\begin{aligned} \mathrm{FPF}(\zeta,\lambda) &= 1-\exp(-\lambda\Phi(-\zeta)) \\ \mathrm{LLF}(\zeta,\mu,\nu) &= \nu\Phi(\mu-\zeta) \end{aligned}\right\} \tag{8.4}$$

The end-point of the AFROC is defined by:

$$\left.\begin{aligned} \mathrm{FPF}_{\max} &= 1-\exp(-\lambda) \\ \mathrm{LLF}_{\max} &= \nu \end{aligned}\right\} \tag{8.5}$$

Like the ROC the AFROC has the constrained end-point property (i.e., the end-point lies within the unit square) which makes its AUC a valid performance metric.

8.4.1 AFROC plots λ_i, ν_i parameterization

Shown below are AFROC curves for the same parameter values used to demonstrate the preceding FROC curves.

- As discussed in the previous chapter for the ROC, each AFROC curve needs to be completed by a (dashed) straight line extending from the end-point to (1,1). A dashed line is used to distinguish it from the continous curve that is accessible to the observer by appropriate choice of ζ. The inaccessible dashed line is necessary to fully account for all decisions.
- Since each plot is contained within the unit square its *net* (i.e., continuous line plus dashed line) AUC is a valid performance metric.
- The AFROC plot is independent of the number of lesions per case. This is not true for the wAFROC, as will shortly become clear, or the ROC (since the ROC ordinate increases with increasing numbers of lesions per case).
- As μ increases the AFROC curve more closely approaches the upper-left corner of the plot and the area under the AFROC curve, including that under the straight line extension, approaches 1, which is the best possible performance.
- For $\mu \to 0$ and fixed λ_i and ν_i the operating characteristic approaches the horizontal line extending from the origin to (1,0), which is the continuous section of the curve, followed by the vertical dashed line connecting (1,0) to (1,1) and AFROC-AUC approaches zero. In this limit, no lesion is localized and every case has at least one NL mark, representing worst possible performance.
- For $\mu \to \infty$ 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 AFROC 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 1. As with the ROC, omitting the area under the dashed portion of the curve will result in a severe underestimate of true performance.
- Other characteristics of AFROC curves (e.g., slope and differences between intrinsic and physical parameterizations), are explored in Appendix 8.7.3.

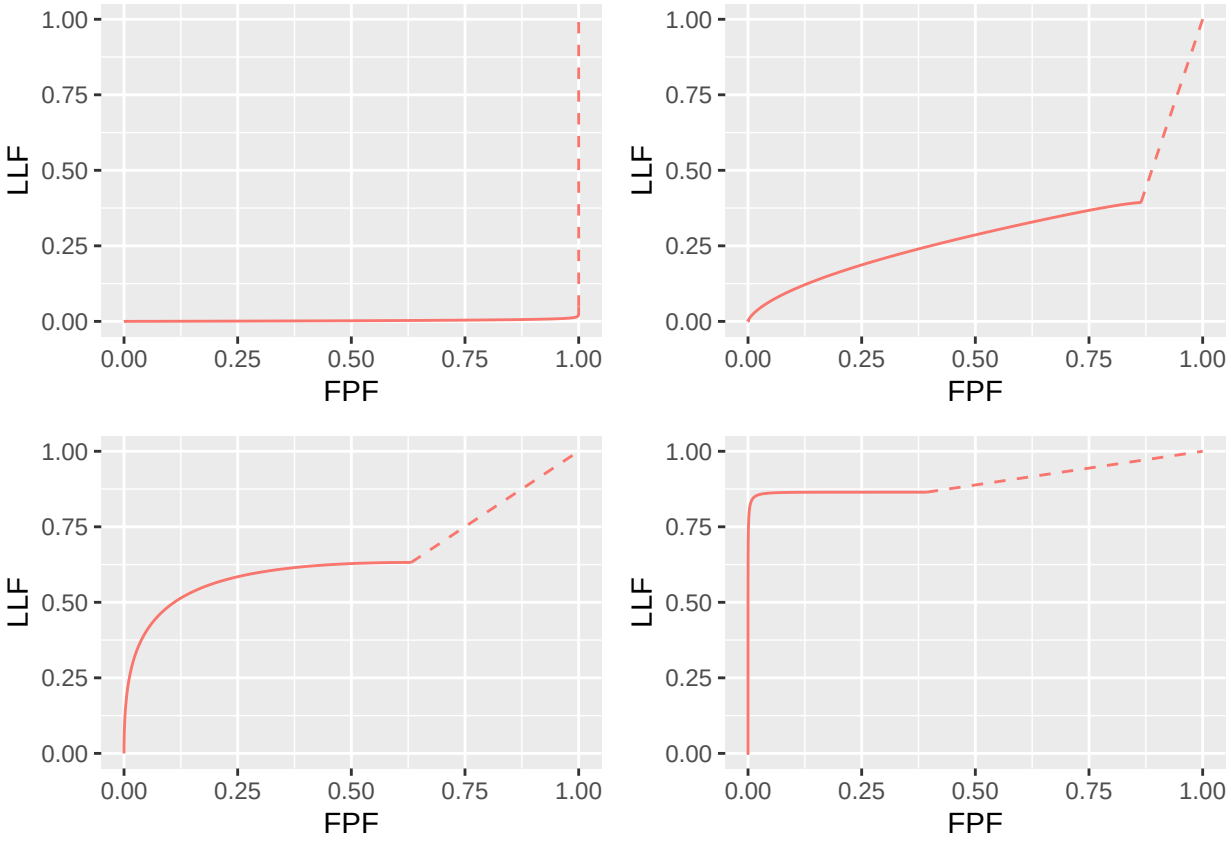


Figure 8.2: RSM-predicted AFROC curves using intrinsic parameters $\lambda_i = 2$ and $\nu_i = 0.5$. Top left: $\mu = 0.1$; Top right: $\mu = 1$; Bottom left: $\mu = 2$; Bottom right: $\mu = 4$. Each curve includes an inaccessible dashed linear extension from the end-point to (1,1). Each plot is contained within the unit square and its AUC is a valid figure of merit.

8.4.2 Case of the reader who does not make any marks

Suppose the radiologist does not mark any case. One possibility is that the radiologist did not interpret the cases and simply "whizzed" through them. Even though the radiologist is not performing the diagnostic task and the AFROC operating point is stuck at the origin one would still be justified in making the straight-line extension to (1,1) which yields AFROC-AUC = 0.5, which implies finite performance (any value greater than zero for AFROC-AUC implies some degree of expertise). This is because the observer is correct in not marking any non-diseased case (any mark on such a case would be incorrect) and deserves credit for correct decisions on non-diseased cases. The situation is somewhat similar to an ROC study in which all cases are diagnosed as non-diseased - the observer is correct on all non-diseased cases and is rewarded with 100 percent specificity. However, the ROC-AUC for this observer is 0.5 (as the operating point is the origin and one needs to connect it via a dashed straight line to the upper right corner of the ROC plot) and the observer is operating at chance level performance, getting no credit for not marking non-diseased cases.

8.5 RSM-predicted wAFROC curve

The wAFROC abscissa is identical to the ROC abscissa, i.e., Eqn. (7.11) applies. The wAFROC ordinate is calculated using:

$$\text{wLLF}\left(\mu,\lambda,\nu,\overrightarrow{f_L},\mathbf{W}\right)=\Phi\left(\mu-\zeta\right)\sum_{L=1}^{L_{max}} f_L \sum_{l_2=1}^{L} \text{W}_{Ll_2}\, l_2\, \text{pmf}_B\left(l_2,L,\nu\right) \tag{8.6}$$

Note that one does not divide by L outside the inner summation as, for each value of L, the weights are already normalized to sum to unit: $\sum_{l_2=1}^{L} \text{W}_{Ll_2}=1$.

Eqn. (8.6) is implemented in `UtilAnalyticalAucsRSM`. Help is available here. A skeleton code is shown below:

```
W <-UtilLesWghtsLD(lesDistr, relWeights)
wLLF <- 0
for (L in 1:L_max){
  wLLF_L <- 0
  for (l_2 in 1:L){
    # W has an extra column that must be skipped, hence W[L, l_2+1]
    wLLF_L <- wLLF_L + W[L, l_2+1] * l_2 * dbinom(l_2, L, nu)
  }
  wLLF <- wLLF +  f_L[L] * wLLF_L
}
wLLF <- wLLF * pnorm(mu - zeta)
```

- $\overrightarrow{f_L}$ 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 weights matrix with L_{max} rows and columns, where each row (excluding cells set to $-\infty$) sums to unity, see example below (the unused matrix elements are set to $-\infty$). In Eqn. (8.6) the index l_2 in W_{Ll_2} ranges from 1 to L.
- 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 0.2, 0.3, 0.1 and 0.4.

- The function `UtilLesWghtsLD` 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)
UtilLesWghtsLD(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 outcome/importance) of the lesion with `lesionID = 1` in cases with one lesion or in cases with 4 lesions are identical. In other words this example assumes that the lesions fall into four classes, with clinical outcomes as specified in `relWeights`.
- $\text{pmf}_B(l_2, L, \nu)$ is the probability mass function (pmf) of the binomial distribution with success probability ν and trial size L. W_{Ll_2} is the weight of lesion l_2 in cases with L lesions; for example $\text{W}_{42} = 0.3$.
- To generate equal weights set `relWeights = 0` as in following code:

```
lesDistr <- c(0.6, 0.2, 0.1, 0.1)
relWeights0 <- 0
UtilLesWghtsLD(lesDistr, relWeights = relWeights0)[,-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
```

8.5.1 wAFROC plots λ_i, ν_i parameterization

- Shown below are wAFROC curves for the same parameter values used to display the AFROC curves shown in Fig. 8.5.
- Note that it is necessary to specify `lesDistr` when requesting a wAFROC plot. A dataset with a maximum of 4 lesions per diseased case is assumed, with `lesDistr <- c(0.6, 0.2, 0.1, 0.1)`.

```
p_wAFROC_lambdai_nui <- array(list(), dim = c(length(muArr)))
for (i in 1:length(muArr)) {
  mu <- muArr[i]
  ret <- Util2Physical(mu, lambda_i = lambda_i, nu_i = nu_i)
  lambda <- ret$lambda
  nu <- ret$nu
  ret1 <- PlotRsmOperatingCharacteristics(
    mu, lambda, nu,
    lesDistr = lesDistr,
    relWeights = relWeights,
    OpChType = "wAFROC",
    legendPosition  = "none"
  )
```

```
  p_wAFROC_lambdai_nui[[i]] <- ret1$wAFROCPlot
  #+ ggtitle(paste0("mu = ", as.character(muArr[i]),
  #+ ", AUC = ", format(ret1$aucAFROC, digits = 3)))
}
```

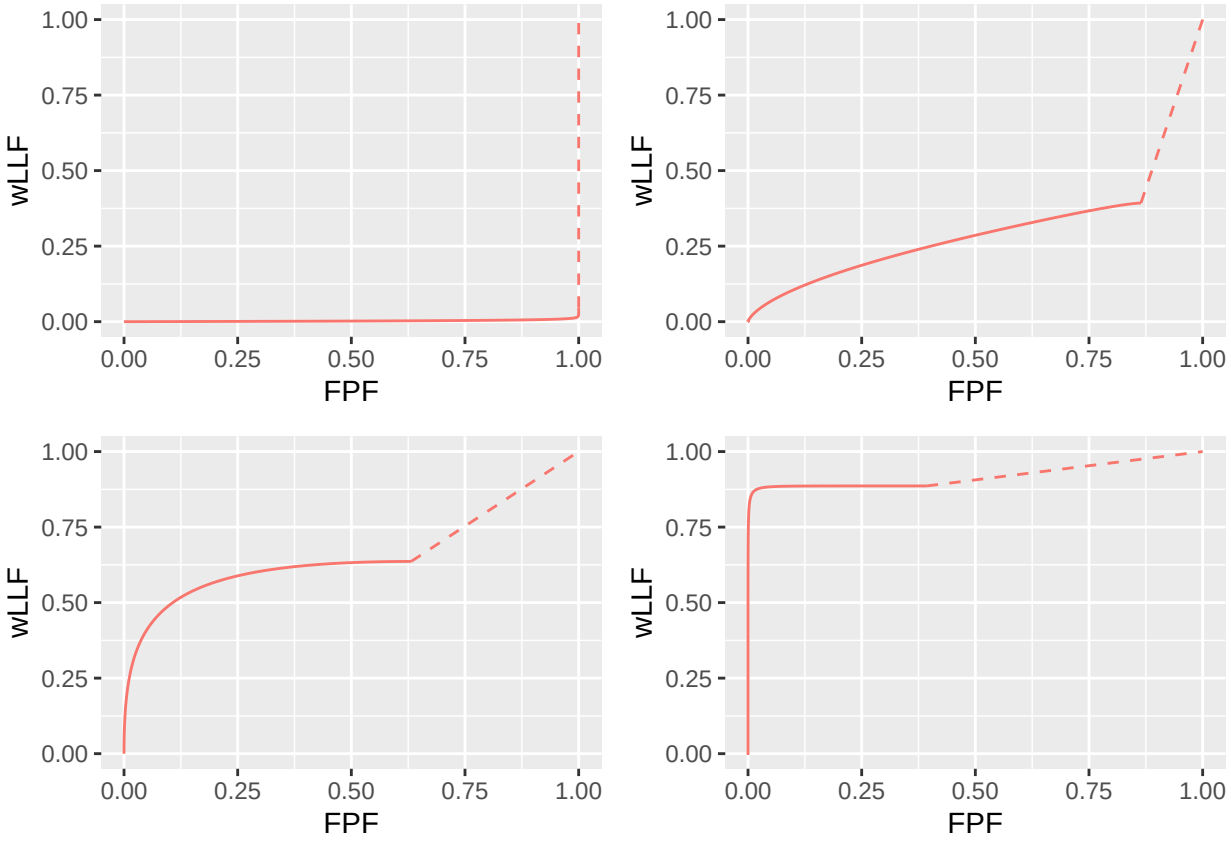


Figure 8.3: RSM-predicted wAFROC curves using intrinsic parameters $\lambda_i = 2$ and $\nu_i = 0.5$. Top left: $\mu = 0.1$. Top right: $\mu = 1$. Bottom left: $\mu = 2$. Bottom right: $\mu = 4$. As μ increases the curve approaches the top-left corner. Each curve includes an inaccessible dashed linear extension to (1,1). Since the plot is contained within the unit square its AUC is a valid figure of merit.

8.6 Comments on end-point-discontinuity property

RSM predicted ROC, AFROC and wAFROC curves share the end-point-discontinuity property (not extending continuously to (1,1)) that makes them qualitatively different from predictions of all other observer performance models that I am aware of. In my experience this is a property that most researchers in this field have difficulty with. There is simply too much history going back to the early 1940s of the ROC curve extending continuously from (0,0) to (1,1).

I am not aware of any empirical evidence that observers can actually operate *continuously* in the range (0,0) to (1,1) *in search tasks*, so the existence of such an ROC is an assumption. In contrast the ROC of an (algorithmic) observer in a *non-search task* can extend continuously to (1,1). Consider 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 $A1C \geq 0.065$ the patient is diagnosed as diabetic. By moving the threshold from $+\infty$ to $-\infty$, 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, will have the end-point-discontinuity property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is more realistic. *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 end-point-discontinuity 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 all cases with no marks. For non-diseased cases these represent correct decisions and for diseased cases they represent incorrect decisions and ignoring all cases with no marks should raise concern regarding validity of the analysis.

8.7 Appendix

Unlike the previous plots which used the *intrinsic* parameters λ_i, ν_i, the plots shown here are for arbitrary choices of RSM *physical* parameters λ, ν. This can lead to peculiar predictions arising from physically unreasonable parameter values.

8.7.1 Slope of the FROC curve

Expressions for LLF and NLF were given above. Taking the derivatives of these functions with respect to ζ the slope of the FROC is given by:

$$\left.\frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(NLF)} = \frac{\nu}{\lambda}\frac{\phi(\mu-\zeta)}{\phi(-\zeta)}\right\} \tag{8.7}$$

With some simplification this yields:

$$\left.\begin{aligned}\frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(NLF)} &= \frac{\nu}{\lambda}\frac{\exp\left(\frac{-(\mu-\zeta)^2}{2}\right)}{\exp\left(\frac{-\zeta^2}{2}\right)} \\ &= \frac{\nu}{\lambda}\exp\left(-\frac{1}{2}\left(\mu^2-2\mu\zeta\right)\right)\end{aligned}\right\} \tag{8.8}$$

Converting to intrinsic parameters leads to the following expression for the slope:

$$\left.\frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(NLF)} = \mu\left(\frac{1-\exp(-\mu\nu_i)}{\lambda_i}\right)\exp\left(-\frac{1}{2}\left(\mu^2-2\mu\zeta\right)\right)\right\} \tag{8.9}$$

Eqn. (8.9) leads to the following conclusions (recall $\mu \geq 0$):

- The slope of the FROC near the end-point, corresponding to $\zeta = -\infty$, is zero.
- The slope near the origin, corresponding to $\zeta = +\infty$, is ∞ provided $\mu \neq 0$.
- For $\mu = 0$ the slope of the FROC is zero regardless of the value of ζ, see top-left panel in Fig. 8.1.

If instead we had used Eqn. (8.8) the last conclusion would change to:

- For $\mu = 0$ the FROC is predicted to be a straight line extending from the origin to (λ, ν), as in the top-left plot in Fig. 8.4 corresponding to $\mu = 0$, $\lambda = 1$ and $\nu = 0.2$. This is unreasonable since for zero contrast lesions the observer should not be able to localize any lesions at finite NLF. The unreasonable prediction is occurring because one is using unrealistic values for the RSM parameters. For zero μ one expects $\lambda = \infty$ and $\nu = 0$, not $\lambda = 1$ and $\nu = 0.2$.

8.7.2 FROC plots λ, ν parameterization

FROC plots are shown below illustrating the statements just made.

```
## mu =  0.100, lambda =  1.000, nu = 0.200
## mu =  1.000, lambda =  2.000, nu = 0.500
## mu =  2.000, lambda =  3.000, nu = 0.700
## mu =  4.000, lambda =  4.000, nu = 0.900
```

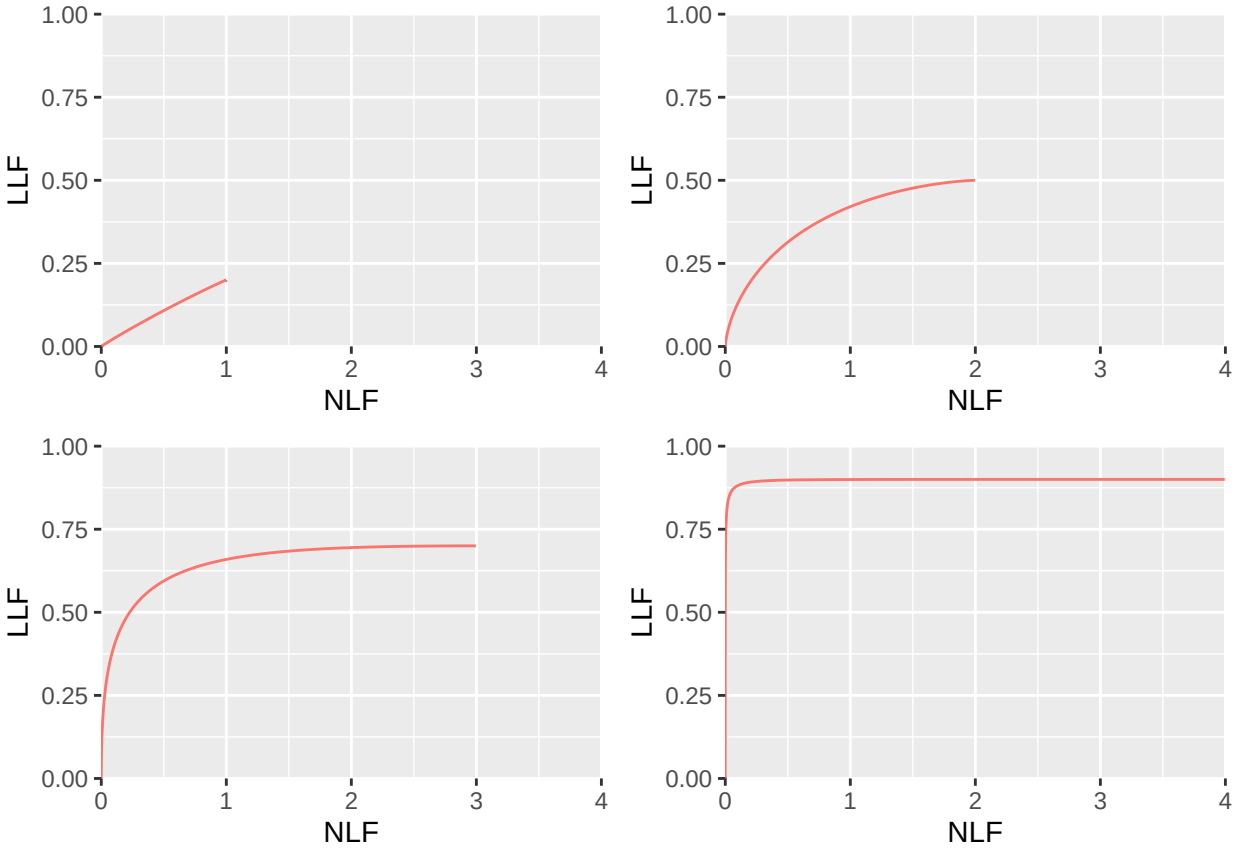


Figure 8.4: RSM-predicted FROC curves using λ, ν paramterization. Top left: $\mu = 0.1$, $\lambda = 1$ and $\nu = 0.2$. Top right: $\mu = 1$, $\lambda = 2$ and $\nu = 0.5$. Bottom left: $\mu = 2$, $\lambda = 3$ and $\nu = 0.7$. Bottom right: $\mu = 4$, $\lambda = 4$ and $\nu = 0.9$. The top-left panel is an unrealistic prediction because of unrealistic parameters $\lambda = 1, \nu = 0.2$ for small μ.

8.7.3 Slope of the AFROC curve

The AFROC ordinate is LLF and the abscissa is FPF. Expressions for both were given above. Taking the derivatives of these functions with respect to ζ the slope of the continuous section of the AFROC is given by:

$$\left.\frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(FPF)} = \frac{\nu\phi(\mu-\zeta)}{\exp(-\lambda\Phi(-\zeta)(\lambda\phi(-\zeta)))}\right\} \tag{8.10}$$

Using Eqn. (8.7) the slope of the AFROC can be expressed in terms of the slope of the FROC:

$$\left.\begin{aligned}\frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(FPF)} &= \frac{\frac{\partial}{\partial\zeta}(LLF)}{\frac{\partial}{\partial\zeta}(NLF)} \\ &= \frac{\frac{\partial}{\partial\zeta}(LLF)}{1-FPF(\lambda,\zeta)}\end{aligned}\right\} \tag{8.11}$$

The numerator is the slope of the FROC. Since $0 \leq \text{FPF} \leq \text{FPF}_{\max}$ and FPF increases as ζ decreases, the slope of the AFROC equals that of the FROC at the origin and subsequently increases over that of the FROC as the end-point is approached.

This expression leads to the following conclusions, if using intrinsic parameterization:

- The slope of the AFROC near the end-point, corresponding to $\zeta = -\infty$, is zero provided $\mu \neq 0$.
- The slope near the origin, corresponding to $\zeta = +\infty$, is ∞ provided $\mu \neq 0$.
- For $\mu = 0$ the slope of the AFROC is zero regardless of the value of ζ, see top-left panel in Fig. 8.2.

If using physical parameters the last conclusion changes to:

- For $\mu = 0$ the slope of the AFROC curve increases as the end-point is approached, i.e., the FROC curve is concave up, see top-left panel in Fig. 8.5. The unreasonable prediction is due to the unreasonable choice of parameters.

```
## mu =  0.100, lambda =  1.000, nu = 0.200
## mu =  1.000, lambda =  2.000, nu = 0.500
## mu =  2.000, lambda =  3.000, nu = 0.700
## mu =  4.000, lambda =  4.000, nu = 0.900
```

8.7.4 wAFROC plots λ, ν parameterization

```
## mu =  0.100, lambda =  1.000, nu = 0.200
## mu =  1.000, lambda =  2.000, nu = 0.500
## mu =  2.000, lambda =  3.000, nu = 0.700
## mu =  4.000, lambda =  4.000, nu = 0.900
```

8.7.5 ROC curves are above AFROC curves

Since they share a common x-axis one can compare the relative position of ROC and AFROC curves for the same parameter values, i.e., does one lie above or below the other. Using previous equations for the ROC-TPF and the AFROC-LLF, and focusing on cases with L lesions, one has:

$$\left.\begin{aligned}&\text{TPF} - \text{LLF} \\ &= 1 - \exp(-\lambda\Phi(-\zeta))(1-\nu\Phi(\mu-\zeta))^L - \nu\Phi(\mu-\zeta) \\ &= 1 - \nu\Phi(\mu-\zeta) - \exp(-\lambda\Phi(-\zeta))(1-\nu\Phi(\mu-\zeta))^L \\ &= (1-\nu\Phi(\mu-\zeta))\left[1 - \exp(-\lambda\Phi(-\zeta))(1-\nu\Phi(\mu-\zeta))^{L-1}\right] \\ &\geq 0\end{aligned}\right\} \tag{8.12}$$

The final inequality follows from the facts that:

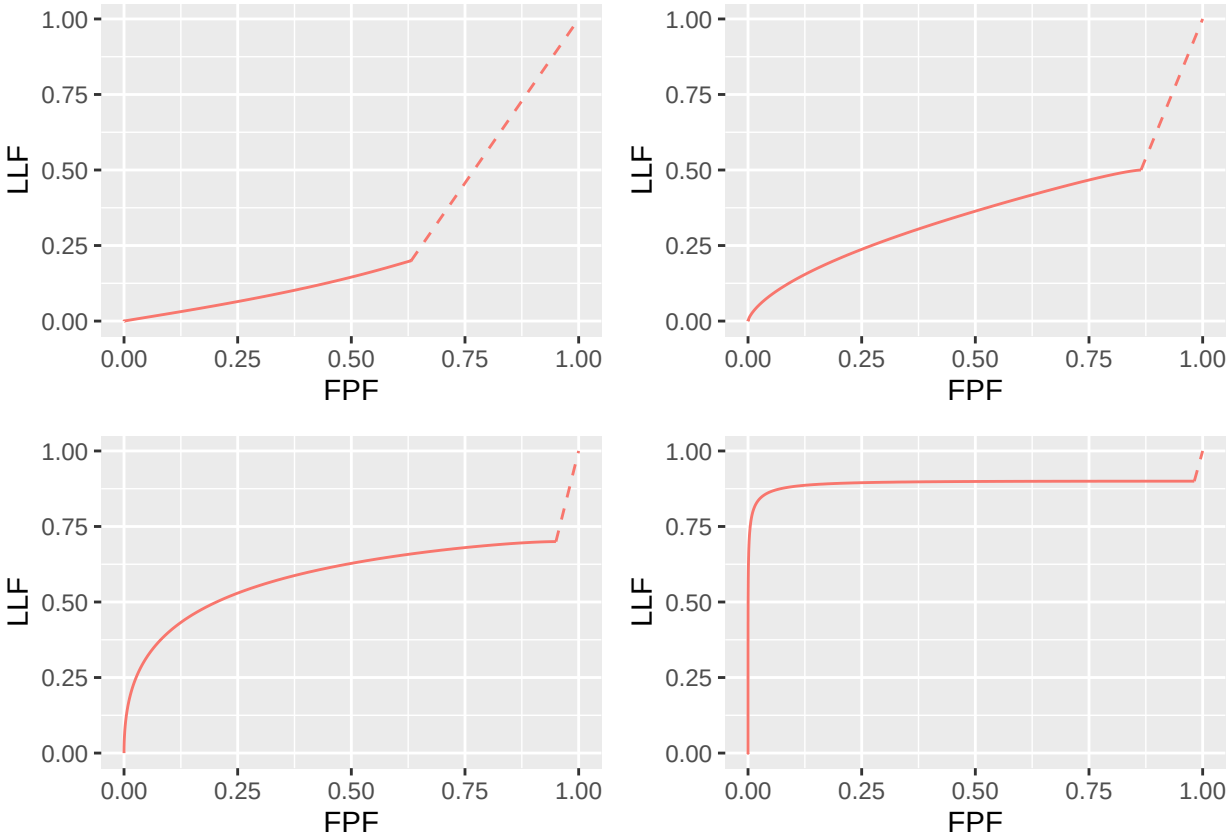


Figure 8.5: RSM-predicted AFROC curves, λ, ν paramterization, using same parameter choices as in preceding plot. Note the unrealistic concave up feature of the top-left plot due to unrealistic choices of parameters.

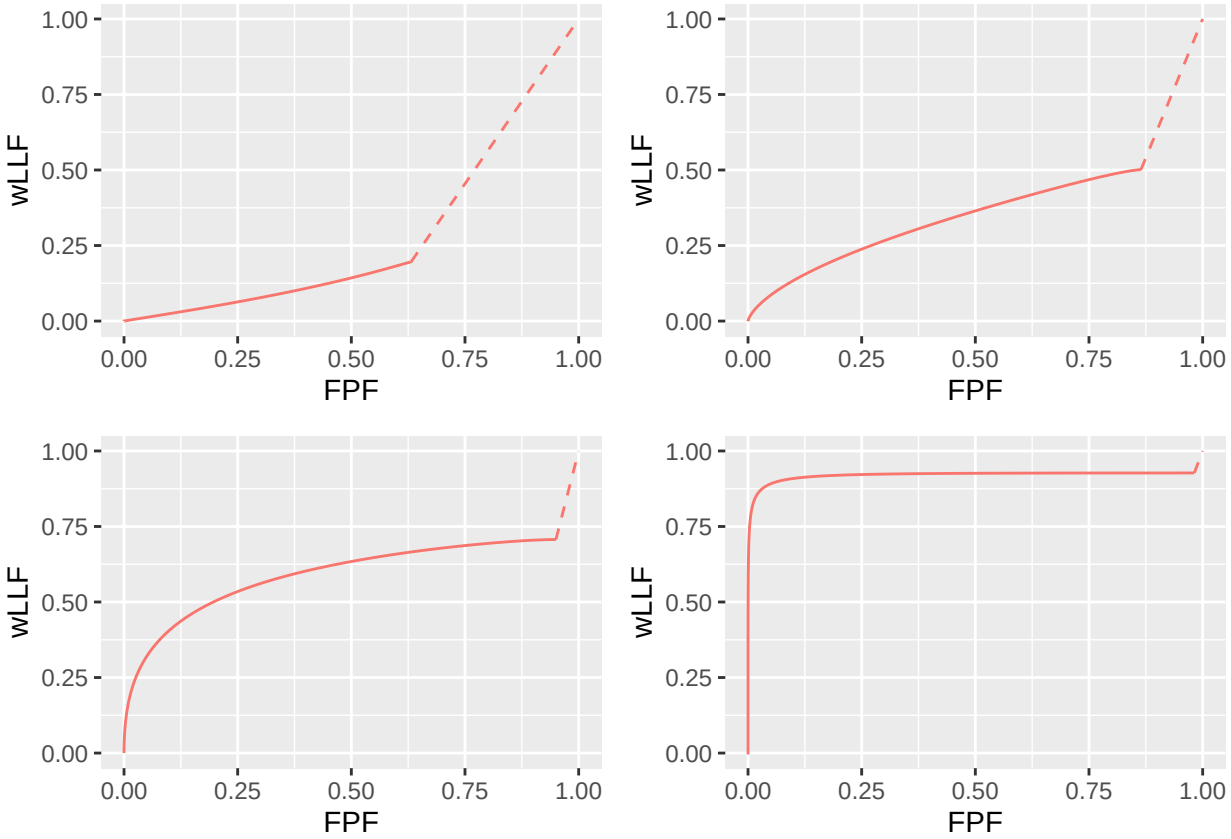


Figure 8.6: RSM-predicted wAFROC curves, λ, ν paramterization, using same parameter choices as in preceding plot. Note the unrealistic concave up feature of the top-left plot due to unrealistic choices of parameters.

- $1 - \nu\Phi(\mu - \zeta)$ is non-negative and less than or equal to one, and so are any integer powers of this quantity.
- $\exp(-\lambda\Phi(-\zeta))$ is non-negative and less than or equal to one.
- The equality is achieved when $\zeta = +\infty$, i.e., at the origin (since the Φ function evaluates to zero).
- Averaging over f_L, the distribution of lesions, does not change the final conclusion.

The basic reason why TPF is greater than LLF is that the ROC gives credit for incorrect localizations on diseased cases while the AFROC does not. This is the well-known "right for wrong reason" argument (Bunch et al., 1977), originally advanced in 1977, against usage of the ROC for localization tasks.

8.7.6 Are wAFROC curves above AFROC curves?

The following expression follows for the difference between wLLF and LLF:

$$\text{wLLF} - \text{LLF} = \Phi(\mu - \zeta) \sum_{L=1}^{L_{max}} f_L \sum_{l_2=1}^{L} \left(\text{W}_{Ll_2} - \frac{1}{L} \right) l_2 \ \text{pmf}_B(l_2, L, \nu) \Bigg\} \tag{8.13}$$

Since for equally weighted lesions each lesion weight is $\frac{1}{L}$, this equation shows immediately that for equally weighted lesions the difference is zero, i.e., *for equally weighted lesions the wAFROC and the AFROC are identical*:

$$|\text{wLLF}|_{\text{equal weights}} - \text{LLF} = 0 \tag{8.14}$$

In the general case the two curves are not identical although, for realistic datasets, the differences tend to be small. For cases with L lesions the probability mass function of the binomial distribution is peaked near $l_2 = L\nu$. If the weights array W_{Ll_2} is likewise peaked near $l_2 = L\nu$ the difference tends to be positive, i.e., the wAFROC is above the AFROC. Otherwise the difference can be negative.

Chapter 9

Lesion localization and classification performances

9.1 How much finished 99%

9.2 Introduction

The preceding two chapters described predictions of the radiological search model (RSM). This chapter describes two performance measures, namely *lesion-localization and lesion-classification performances*, that can be derived from the predicted ROC. These performances were introduced conceptually in Section 5.5 and this chapter deals with relating them to the RSM parameters.

Recall the the search process involves two stages: (1) finding suspicious regions and (2) correctly classifying each suspicious region as either a lesion (in which case the region is marked and rated) or a non-lesion (in which case the region is not marked). The first stage is the lesion-localization task while the second stage is the lesion-classification task. Performance in the lesion-localization task is measured by the ability to mark lesions while minimizing marking non-lesions. Performance in the classification task is the ability, having found a suspicious region, to correctly recognize it as a lesion (to be marked and rated) or a non-lesion (to be ignored). The aim of this chapter is to quantify these two abilities.

9.3 Quantifying lesion-localization performance

From Chapter 7 the coordinates of the RSM-predicted ROC end-point are given by:

$$\left.\begin{aligned} \mathrm{FPF}_{\max} &= 1-\exp(\lambda) \\ \mathrm{TPF}_{\max} &= 1-\exp(-\lambda)\sum_{L=1}^{L_{max}} f_L\,(1-\nu)^L \end{aligned}\right\} \tag{9.1}$$

Qualitatively, lesion-localization performance is the ability to mark lesions while not marking non-lesions. To arrive at a quantitative definition consider the location of the ROC end-point.

In Fig. 9.1 curve (a) is a typical ROC curve predicted by models that do not account for lesion-localization, specifically the binormal model is considered here. The corresponding 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 corresponding end-point, the filled square, is downwards and left shifted relative to (1,1). The chance diagonal is the straight line labeled (c).

The specific parameter values used in the illustration are shown next:

```
a <- 2; b <- 1 # binormal model
mu <- 2; lambda_i <- 2; nu_i <- 1 # rsm
lesDistr <- c(1) # one lesion per dis. case
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

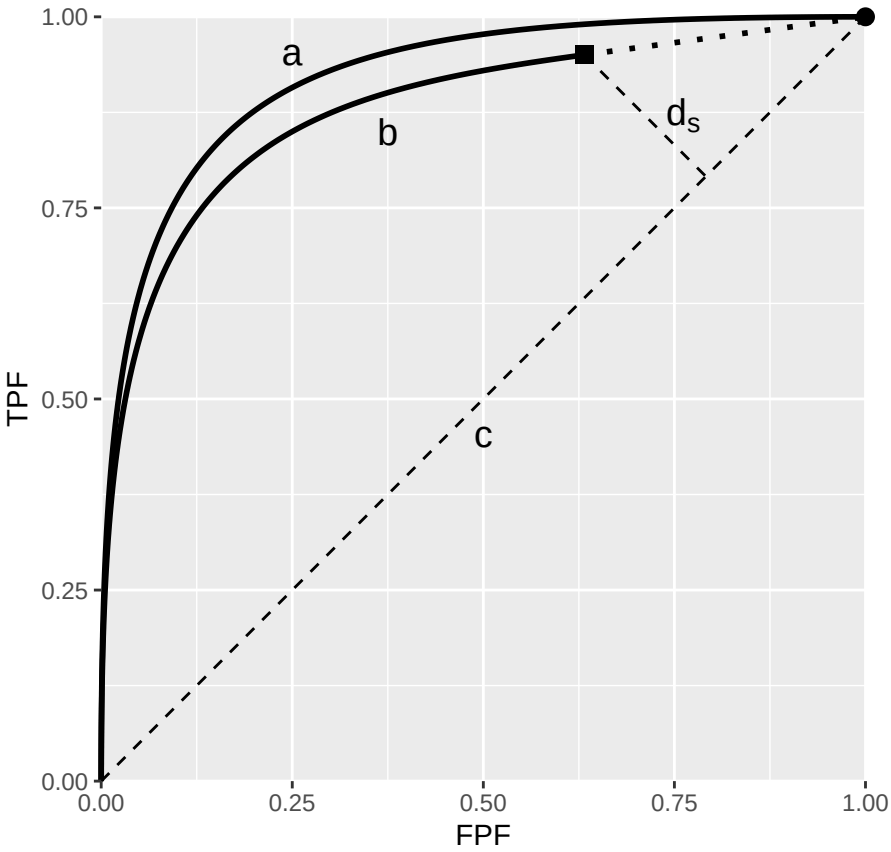


Figure 9.1: Relation of lesion-localization performance to the end-point of the ROC curve. Plot (a) is using the binormal model while plot (b) is using a RSM predicted curve. The chance diagonal is labeled c. The filled square is the end-point of the RSM predicted curve while the filled dot is the end-point of the binormal predicted curve. The distance of the filled square from the chance diagonal, labeled d_S, is a measure of lesion-localization performance.

The location of the end-point, in particular how far it is from (1,1), is a measure of lesion-localization performance. Higher lesion-localization performance is characterized by the end-point moving upwards and to the left, in the limit to (0,1), corresponding to perfect lesion-localization performance. It is more convenient to use a distance measure as defined next:

Definition

> The perpendicular distance, d_S, from the end-point to the chance diagonal, plot (c), multiplied by $\sqrt{2}$, is the quantitative measure of lesion-localization performance denoted by L_L.

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

$$L_L = \sqrt{2} d_S = \mathrm{TPF}_{\max} - \mathrm{FPF}_{\max} \tag{9.2}$$

Therefore, lesion-localization performance L_L is given by:

$$L_L = \exp(-\lambda) \left(1 - \sum_{L=1}^{L_{max}} f_L (1-\nu)^L \right) \tag{9.3}$$

Eqn. (9.3) shows lesion-localization performance is the product of two terms: the probability $\left(1 - \sum_{L=1}^{L_{max}} f_L (1-\nu)^L\right)$ 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 lesion-localization performance as the ability to find lesions while avoiding finding non-lesions.

Example: consider $\lambda = 0$ and $\nu = 1$. (In terms of intrinsic parameters this occurs when $\mu = \infty$.) 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 $L_L = 1$. The same value is obtained using Eqn. (9.3). Since no NLs are found and all lesions are found, the observer never makes a mistake. One cannot improve over perfect performance and the observer does not need to use the z-sample information: he simply marks all suspicious regions found by the search mechanism regardless of their z-samples.

Search performance ranges from zero to one: $0 \le L_L \le 1$. The lower limit is reached if $\lambda = \infty$ or $\nu = 0$. (In terms of intrinsic parameters this occurs when $\mu = 0$.)

9.4 Quantifying lesion-classification performance

Lesion-classification performance L_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. L_C is determined by the μ parameter of the RSM.

Definition

> L_C is defined by the implied ROC-area of two unit variance normal distributions separated by μ (see formula for d' measure in RJafrocRocBook).

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

Since $\mu \ge 0$ it follows that L_C ranges from 0.5 to 1: $0.5 \le L_C \le 1$. The lower limit occurs when $\mu = 0$ and the upper limit occurs when $\mu = \infty$.

9.5 Discussion

We have shown that the RSM parameters determine lesion-localization and lesion-classification performances. In the next chapter it will be shown that these parameters can be estimated from ROC/FROC data. Therefore the results of this chapter should be of interest to researchers in the area of computer aided detection – CAD – algorithm design, because they yield information about which stage – lesion-localization or lesion-classification – is limiting performance. If lesion-localization performance is low then the observer is having difficulty finding lesions while minimizing finding non-lesions [1]. In the CAD context, the *initial detection* stage needs to be further optimized. If lesion-classification performance is low the observer is finding lesions efficiently but is having difficulty correctly classifying the found lesions. In the CAD context, the *candidate analysis* stage needs to be further optimized. Of course, for this to be realized one needs a method for estimating the RSM parameters. This is the subject of the next chapter.

[1] We repeat that a non-expert can trivally "find" all lesions by marking all regions in the image.

CAD applications

Chapter 10

Standalone CAD

10.1 How much finished 99%

10.2 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, 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 FDA-approved 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. Analysis of this type of date is deferred to a following **TBA** chapter. However, a malignancy index (a continuous variable) 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 method for such measurements. This chapter removes that impediment. It describes a method for comparing standalone performance of designer-level CAD to a group of radiologists interpreting the same cases and compares the method to those described in two relevant publications (Hupse et al., 2013; Kooi et al., 2016).

10.3 Overview

This chapter extends the method used in a study of standalone CAD performance (Hupse et al., 2013), termed one-treatment random-reader fixed case or **1T-RRFC** analysis, since CAD is treated as an additional reader within a single treatment and since it only accounts for reader variability but does not account for case-variability.

The extension includes the effect of case-sampling variability and is hence termed one-treatment random-reader random-case or **1T-RRRC** analysis. The method is based on an existing method 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 over CAD as a figure of merit to which the existing work is directly applicable. The 1T-RRRC method is compared to 1T-RRFC.

The 1T-RRRC method is also compared to an unorthodox usage of conventional multiple-treatment multiple-reader method, termed **2T-RRRC** analysis, which involves replicating the CAD ratings as many times as there are radiologists, in effect simulating a second treatment, i.e., CAD is regarded as the second treatment (with identical readers within this treatment) to which existing methods (DBM or OR, as described in RJafrocRocBook) is applied.
‘

10.4 Methods

Summarized are two relevant studies of CAD vs. radiologists in mammography. This is followed by comments on the methods used in the two studies. The second study used multi-treatment multi-reader receiver operating characteristic (ROC) software in an unorthodox way. A statistical model and analysis method is described that avoids the unorthodox usage of ROC software and has fewer model parameters.

10.4.1 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.

10.4.1.1 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 empirical 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. A 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 $PCL_{0.2}$.

10.4.1.2 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 analyzed by a different CAD algorithmic observer from that used to determine the ROIs and by four expert radiologists. In either case the ROC paradigm was used (i.e., a rating was obtained for each ROI) The figure of merit was the empirical area (AUC) under the respective ROC curves (one for each radiologist and one for CAD). The p-value for the difference in AUCs between the average radiologist's AUC and CAD AUC was determined using an unorthodox application of the Dorfman-Berbaum-Metz (Dorfman et al., 1992) multiple-treatment multiple-reader multiple-case (DBM-MRMC) software.

The application was unorthodox in the sense 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 10.1. By this artifice they converted a single-treatment

Table 10.1: 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

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 readers are labeled C1, C2, C3 and C4.

Study – 2 reported a not significant difference between CAD and the radiologists ($p = 0.253$).

10.4.1.3 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 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. 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.

In what follows, the analysis in Study – 1 is referred to as **single-treatment random-reader fixed-case (1T-RRFC)** while that in Study – 2 is referred to as **dual-treatment random-reader random-case (2T-RRRC)**.

10.4.2 The 1T-RRFC analysis model

The sampling model for the FOM is:

$$\left.\begin{array}{r} \theta_j = \mu + R_j \\ (j = 1, 2, ...J) \end{array}\right\} \tag{10.1}$$

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:

$$R_j \sim N\left(0, \sigma_R^2\right) \tag{10.2}$$

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-CAD variability, i.e., CAD applied repeatedly to a fixed dataset does not always produce the same mark-rating data. However, this source of within-CAD variability is much smaller than *inter-reader* variability of radiologists interpreting the same dataset. The *within-reader* variability of radiologists is smaller than *inter-reader* variability and *within-CAD* variability is even smaller. 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 reflecting case sampling variability which needs to be accounted for as done in the following analyses.

10.4.3 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 10.1.

The following approach, termed 2T-RRRC, uses the Obuchowski and Rockette (OR) figure of merit sampling model (Obuchowski and Rockette, 1995). The OR model is:

$$\theta_{ij\{c\}} = \mu + \tau_i + (\tau \mathrm{R})_{ij} + \epsilon_{ij\{c\}} \tag{10.3}$$

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 the figure of merit in treatment i for reader j 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. (10.3) are fixed effects (average performance and treatment effect, respectively). The next two terms are random effect variables that, by assumption, are sampled as follows:

$$\left.\begin{array}{r} R_j \sim N\left(0, \sigma_R^2\right) \\ (\tau R)_{ij} \sim N\left(0, \sigma_{\tau R}^2\right) \end{array}\right\} \tag{10.4}$$

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:

$$\epsilon_{ij\{c\}} \sim N_{I \times J}\left(\vec{0}, \Sigma\right) \tag{10.5}$$

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:

$$\mathrm{Cov}\left(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}\right) = \left\{\begin{array}{l} \mathrm{Var}\ (i = i', j = j') \\ \mathrm{Cov1}\ (i \neq i', j = j') \\ \mathrm{Cov2}\ (i = i', j \neq j') \\ \mathrm{Cov3}\ (i \neq i', j \neq j') \end{array}\right\} \tag{10.6}$$

Software {U of Iowa and `RJafroc`} yields estimates of all terms appearing on the right hand side of Eqn. (10.6). Excluding fixed effects the model represented by Eqn. (10.3) contains six parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \text{Var}, \text{Cov}_1, \text{Cov}_2, \text{Cov}_3 \tag{10.7}$$

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:

$$\text{Var} \geq \text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \tag{10.8}$$

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.

10.4.4 The 1T-RRRC analysis model

The difference from the approach in Study - 2, and the main contribution of this work, is to regard standalone CAD as a different reader, not as a different treatment. This section describes a single treatment method for analyzing readers and CAD, where CAD is regarded as an additional reader and artificially replicated CAD data becomes unnecessary. Accordingly the proposed method is termed **single-treatment random-reader random-case (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:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{10.9}$$

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

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{10.10}$$

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

$$\Sigma_{jj'} = \text{Cov}\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \text{Var} & j = j' \\ \text{Cov}_2 & j \neq j' \end{cases} \tag{10.11}$$

In practice the terms Var and Cov_2 are estimated using the jackknife method.

10.4.4.1 Single treatment analysis for radiologists

Hillis (Hillis et al., 2005; Hillis, 2007) has described how to use the single treatment model (10.9) to compare a groups of radiologists' average performance to a fixed value, in effect the NH : $\mu = \mu_0$, where μ_0 is a pre-specified constant.

One might be tempted to 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.

10.4.4.2 Adaptation of single treatment analysis to accommodate CAD

Instead, the following model is used for the figure of merit of the radiologists **and** CAD (note that $j = 0$ is used to denote the CAD algorithmic reader):

$$\theta_{j\{c\}} = \theta_{0\{c\}} + \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{10.12}$$

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

$$\psi_{j\{c\}} = \theta_{j\{c\}} - \theta_{0\{c\}} \tag{10.13}$$

Then Eqn. (10.12) reduces to:

$$\psi_{j\{c\}} = \Delta\theta + R_j + \epsilon_{j\{c\}} \tag{10.14}$$

Eqn. (10.14) is identical in form to Eqn. (10.9) excepting that the figure of merit on the left hand side of Eqn. (10.14) is a *difference FOM*, that between the radiologist's and CAD, i.e., describing 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. (10.14).

Apart from fixed effects, the model in Eqn. (10.14) contains three parameters:

$$\sigma_R^2, \text{Var}, \text{Cov}_2 \tag{10.15}$$

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 an identical estimate of this parameter using 1T-RRRC analyses.

10.5 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`.

The following code shows usage of the software to generate the results. Note that `RJafroc::datasetCadLroc` is the LROC dataset and `RJafroc::dataset09` is the corresponding ROC dataset.

```
RRFC_1T_PCL_0_05 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRFC")
RRRC_2T_PCL_0_05 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "2T-RRRC")
RRRC_1T_PCL_0_05 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRRC")

RRFC_1T_PCL_0_2 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRFC")
RRRC_2T_PCL_0_2 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "2T-RRRC")
RRRC_1T_PCL_0_2 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRRC")

RRFC_1T_PCL_1 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
```

```
FOM = "PCL", FPFValue = 1, method = "1T-RRFC")
RRRC_2T_PCL_1 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "2T-RRRC")
RRRC_1T_PCL_1 <- RJafroc::StCadVsRad (RJafroc::datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRRC")

RRFC_1T_AUC <- RJafroc::StCadVsRad (RJafroc::dataset09,
FOM = "Wilcoxon", method = "1T-RRFC")
RRRC_2T_AUC <- RJafroc::StCadVsRad (RJafroc::dataset09,
FOM = "Wilcoxon", method = "2T-RRRC")
RRRC_1T_AUC <- RJafroc::StCadVsRad (RJafroc::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 = $\text{PCL}_{0.05}$.
- `RRRC_2T_PCL_0_05` contains the results of 2T-RRRC analysis for figure of merit = $\text{PCL}_{0.05}$.
- `RRRC_1T_PCL_0_05` contains the results of 1T-RRRC analysis for figure of merit = $\text{PCL}_{0.05}$.
- `RRFC_1T_PCL_0_2` contains the results of 1T-RRFC analysis for figure of merit = $\text{PCL}_{0.2}$.
- `RRRC_2T_PCL_0_2` contains the results of 2T-RRRC analysis for figure of merit = $\text{PCL}_{0.2}$.
- `RRRC_1T_PCL_0_2` contains the results of 1T-RRRC analysis for figure of merit = $\text{PCL}_{0.2}$.
- `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.

10.6 Results

The three methods, 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), Table 10.2.

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 non-zero radiologist index j = 1,2,..., 9) and its 95% confidence interval $CI_{\theta_\bullet}$, columns 7 and 8 list the average difference FOM $\psi_\bullet$, i.e., radiologist average 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 is unity).

> **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. Important caveats are noted below.

Other results evident in Table 10.2:

Table 10.2: Significance testing results for an LROC dataset. For each figure of merit (FOM) shown are results of RRRC, 2T-RRRC and 1T-RRRC analyses. Because it is accounting for an additional source of variability, each of the rows labeled RRRC yields a larger p-value and wider confidence interval than the corresponding row labeled 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	0.45	NA	0.493	(0.42,0.57)	0.0433	(-0.032,0.12)	1.8	8	0.22
	2T-RRRC		(0.26,0.64)		(0.38,0.61)		(-0.16,0.24)	0.18	784	0.67
	1T-RRRC		NA		(0.29,0.69)		(-0.16,0.24)	0.18	784	0.67
PCL_0_2	1T-RRFC	0.592	NA	0.71	(0.67,0.75)	0.119	(0.078,0.16)	45	8	0.00015
	2T-RRRC		(0.48,0.71)		(0.63,0.79)		(0.0044,0.23)	4.2	937	0.042
	1T-RRRC		NA		(0.6,0.82)		(0.0044,0.23)	4.2	937	0.042
PCL_1	1T-RRFC	0.675	NA	0.783	(0.74,0.83)	0.108	(0.065,0.15)	33	8	0.00043
	2T-RRRC		(0.57,0.78)		(0.71,0.85)		(0.0045,0.21)	4.2	493	0.041
	1T-RRRC		NA		(0.68,0.89)		(0.0045,0.21)	4.2	493	0.041
Wilcoxon	1T-RRFC	0.817	NA	0.849	(0.83,0.87)	0.0317	(0.009,0.055)	10	8	0.012
	2T-RRRC		(0.75,0.88)		(0.81,0.89)		(-0.031,0.094)	0.99	878	0.32
	1T-RRRC		NA		(0.79,0.91)		(-0.031,0.094)	0.99	878	0.32

1. Where a direct comparison is possible, namely 1T-RRFC analysis using $\text{PCL}_{0.05}$ and $\text{PCL}_{0.2}$ as FOMs, the p-values in Table 10.2 are very close to those reported in Study – 1.
2. All FOMs (i.e., θ_0, $\theta_\bullet$ and $\psi_\bullet$) in Table 10.2 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 the CAD figure of merit is a constant no confidence interval is appropriate for it for either 1T-RRFC or 1T-RRRC analysis and the listed values are NA (not applicable). Since 2T-RRRC analysis assumes CAD is a different treatment the analysis lists a confidence interval that is correctly centered on the CAD value but is otherwise meaningless, i.e., it is an artifact of the unintended usage of the OR analysis method.
4. 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.
5. The LROC FOMs increase as the value of FPF (the subscript) increases, a general feature of any partial curve based figure of merit, as is the observation that the area (AUC) under the ROC is larger than the largest PCL value.
6. Using either RRRC analyses ignoring localization information (i.e., using the AUC FOM) leads to a not-significant difference between CAD and the radiologists ($p = 0.32$) while using localization information via the PCL_1 FOM yields a significant difference ($p = 0.041$), consistent with the expectation that using localization information leads to increased statistical power.
7. Partial curve-based FOMs, such as PCL_{FPF}, lead, depending on the choice of FPF, to different conclusions on whether to reject the NH. Using either RRRC analyses the p-values decrease as FPF increases (e.g., \$ 0.67 $>$ 0.042 $>$ 0.041\$). This trend is not observed for 1T-RRFC analysis which shows a "sweet-spot" effect where the p-value has a minimum for FPF = 0.2

Shown next, Table 10.3, are the model-parameters corresponding to the three analyses.

From Table 10.3 some inconsistencies are evident for 2T-RRRC analysis:

1. For 2T-RRRC analyses the listed values for σ_R^2 are smaller than machine accuracy, therefore one concludes that in fact $\sigma_R^2 = 0$ which is **clearly an incorrect result as the radiologists do not have identical performances**. In contrast, 1T-RRRC analyses yields the expected non-zero values, identical to those obtained by 1T-RRFC analyses (see comment following Eqn. (10.15)).
2. For the 2T_RRRC method the expected ordering of the inequalities, Eqn. (10.8) is not observed: one expects $\text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3$ but instead one observes $\text{Cov}_1 = \text{Cov}_3$ and $\text{Cov}_2 > \text{Cov}_1$.

Table 10.3: Significance testing results for an LROC dataset. For each figure of merit (FOM) shown are results of RRRC, 2T-RRRC and 1T-RRRC analyses. Because it is accounting for an additional source of variability, each of the rows labeled RRRC yields a larger p-value and wider confidence interval than the corresponding row labeled 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	σ_R^2	$\sigma_{\tau R}^2$	Cov1	Cov2	Cov3	Var
PCL_0_05	1T-RRFC	0.0095	NA	NA	NA	NA	NA
	2T-RRRC	-1.1e-19	-0.00571	0.00131	0.00601	0.00131	0.0165
	1T-RRRC	0.0095	NA	NA	0.0094	NA	0.0303
PCL_0_2	1T-RRFC	0.00281	NA	NA	NA	NA	NA
	2T-RRRC	-4.9e-19	0.000265	0.000761	0.00229	0.000761	0.00343
	1T-RRRC	0.00281	NA	NA	0.00307	NA	0.00534
PCL_1	1T-RRFC	0.0032	NA	NA	NA	NA	NA
	2T-RRRC	6e-19	0.001	0.000643	0.00186	0.000643	0.00246
	1T-RRRC	0.0032	NA	NA	0.00244	NA	0.00364
Wilcoxon	1T-RRFC	0.000878	NA	NA	NA	NA	NA
	2T-RRRC	7.9e-19	0.000201	0.000262	0.000724	0.000262	0.000962
	1T-RRRC	0.000878	NA	NA	0.000924	NA	0.0014

The design of a ratings simulator to statistically match a given dataset is addressed in Chapter 23 of my print book (Chakraborty, 2017). Using this simulator, the 1T-RRRC method had the expected null hypothesis behavior (Table 23.5, ibid).

10.7 Discussion

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 subject to case-sampling variability only, and the desire is to compare performance of this special reader to the average of the remaining readers. Usage of the method along with two other methods is illustrated using an LROC dataset.

The proposed method, 1T-RRRC analyses, yields identical "overall" results (specifically the F-statistic, degrees of freedom and p-value) to those yielded by the unorthodox application of commonly available software, termed 2T-RRRC analyses, where the CAD reader is regarded as a second treatment (specifically the CAD ratings are replicated to match the number of radiologists). If interest is in just these values one is justified in using the 2T-RRRC method. However, 2T-RRRC model parameter estimates were unrealistic: for example, it yields zero between-reader variance. 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 $\sigma_R^2 = 0$. Additionally, the covariance estimates are incorrect. Since sample-size estimation requires some of the covariance values the 2T-RRRC method should never be used to perform sample-size estimation for a prospective study.

The 1T-RRRC method described here is applicable to any scalar figure of merit. The paradigm used to collect the observer performance data - ROC, FROC, LROC or ROI - is irrelevant.

Assessing CAD utility 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.

In my opinion standalone performance is the most direct measure of CAD performance. Lack of a clear-cut method for assessing standalone CAD performance may have limited past CAD research. The current work hopefully

removes that impediment. Going forward, assessment of standalone performance of CAD vs. expert radiologists is strongly encouraged.

10.8 Appendix 1

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

10.8.1 Example 1

The first example shows the structure of RRFC_1T_PCL_0_2.

```
x <- RRFC_1T_PCL_0_2
fom_individual_rad <- as.data.frame(t(x$fomRAD))
colnames(fom_individual_rad) <- paste0("rdr", seq(1:9))

stats <- data.frame(fomCAD = x$fomCAD, avgRadFom = x$avgRadFom, avgDiffFom = x$avgDiffFom, varR = x$varR,

ConfidenceIntervals <- data.frame(CIAvgRadFom = x$CIAvgRadFom, CIAvgDiffFom = x$CIAvgDiffFom)
rownames(ConfidenceIntervals) <- c("Lower", "Upper")
```

```
print(fom_individual_rad)
#>        rdr1 rdr2    rdr3  rdr4      rdr5      rdr6   rdr7  rdr8  rdr9
#> 1 0.6945313 0.65 0.80625 0.725 0.6598214 0.7684524 0.7375 0.675 0.675
print(stats)
#>      fomCAD avgRadFom avgDiffFom        varR    Tstat df         pval
#> 1 0.5916667 0.7101728  0.1185061 0.002808612 6.708357  8 0.0001513966
print(ConfidenceIntervals)
#>       CIAvgRadFom CIAvgDiffFom
#> Lower   0.6694362   0.07776953
#> Upper   0.7509094   0.15924271
```

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.

10.8.2 Example 2

The next example shows the structure of `RRRC_2T_PCL_0_2`.

```
x <- RRRC_2T_PCL_0_2

fom_individual_rad <- as.data.frame(t(x$fomRAD))
colnames(fom_individual_rad) <- paste0("rdr", seq(1:9))

stats1 <- data.frame(fomCAD = x$fomCAD, avgRadFom = x$avgRadFom, avgDiffFom = x$avgDiffFom)

stats2 <- data.frame(varR = x$varR, varTR = x$varTR,
                     cov1 = x$cov1, cov2 = x$cov2 ,
                     cov3 = x$cov3 , Var = x$varError,
                     FStat = x$FStat, df = x$df, pval = x$pval)
```

```
print(fom_individual_rad)
#>        rdr1 rdr2    rdr3  rdr4      rdr5      rdr6   rdr7  rdr8  rdr9
#> 1 0.6945313 0.65 0.80625 0.725 0.6598214 0.7684524 0.7375 0.675 0.675
print(stats1)
#>      fomCAD avgRadFom avgDiffFom
#> 1 0.5916667 0.7101728  0.1185061
print(stats2)
#>           varR        varTR         cov1        cov2         cov3         Var
#> 1 -4.87891e-19 0.0002648898 0.0007613684 0.002294221 0.0007613684 0.003433637
#>     FStat       df       pval
#> 1 4.15768 937.2437 0.04172626
```

In addition to the quantities defined previously, the output contains the covariance matrix for the Obuchowski-Rockette model, summarized in Eqn. (10.3) – Eqn. (10.6).

- `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 0.5916667 0.05802835      Inf 0.4779332 0.7054001 0.003367289
#> trt2 0.7101728 0.03915636 193.1083 0.6329437 0.7874018 0.001221153
```

- `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 0.1185061 0.05811861 937.2437 2.039039 0.04172626 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.

10.8.3 Example 3

The last example shows the structure of `RRRC_1T_PCL_0_2`.

```
RRRC_1T_PCL_0_2
#> $fomCAD
#> [1] 0.5916667
#>
#> $fomRAD
#> [1] 0.6945313 0.6500000 0.8062500 0.7250000 0.6598214 0.7684524 0.7375000
#> [8] 0.6750000 0.6750000
#>
#> $avgRadFom
#> [1] 0.7101728
#>
#> $CIAvgRad
#> [1] 0.5961151 0.8242305
#>
#> $avgDiffFom
#> [1] 0.1185061
#>
#> $CIAvgDiffFom
#> [1] 0.004448434 0.232563801
```

```
#>
#> $varR
#> [1] 0.002808612
#>
#> $varError
#> [1] 0.005344538
#>
#> $cov2
#> [1] 0.003065705
#>
#> $Tstat
#>     rdr2
#> 2.039039
#>
#> $df
#>     rdr2
#> 937.2437
#>
#> $pval
#>       rdr2
#> 0.04172626
```

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. (10.15).
- `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., 0.0417263, is identical to that of `RRRC_2T_PCL_0_2`, i.e., 0.0417263.

10.9 Appendix 2

Two text files `R/standalone-cad/jaf_truth.txt` and `R/standalone-cad/jaf_truth.txt` were provided by Prof. Nico Karssemeijer. These are read into a dataset object by the following code.

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

Chapter 11

CAD optimal operating point

11.1 TBA How much finished 98%

Handling diseased-only datasets Discussion needs more work

11.2 Introduction

A familiar problem for the computer aided detection or artificial intelligence (CAD/AI) algorithm designer is how to set the reporting threshold of the algorithm. Assuming 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 where its slope equals a specific value 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 operating point is unity. For a proper ROC curve this corresponds to the point that maximizes the Youden-index (Youden, 1950). Typically this index is maximized at the point that is closest to the (0,1) corner of the ROC.

Lacking a procedure for determining it analytically currently CAD designers (in consultation with radiologists) set imaging site-specific reporting thresholds. For example, if radiologists at an imaging 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 based on maximizing AUC (area under curve) of the wAFROC curve. For comparison the Youden-index based method was also used.

11.3 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 6. The model predicts ROC, FROC and wAFROC curves and is characterized by the three parameters – μ, λ, ν – 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.

Additionally, there is a threshold parameter ζ_1 with the property that only if the rating of a latent mark exceeds ζ_1 the latent mark is actually marked. Therefore higher values of ζ_1 correspond to more stringent reporting criteria and fewer actual marks. As will be shown next **net performance as measured by wAFROC$_{\mathbf{AUC}}$ 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 3 RSM parameters and the figure of merit to be optimized (i.e., wAFROC$_{\text{AUC}}$ or the Youden-index), to determine the optimal value of ζ_1.

In the following sections the RSM λ parameter is varied (for fixed μ and ν) and the corresponding optimal ζ_1 determined by maximizing either wAFROC$_{\text{AUC}}$ or the Youden-index.

For organizational reasons only the summary results for varying μ or ν are shown in the body of this chapter. Detailed results are in Appendix 11.10 which also has results for limiting cases of high and low ROC performance.

The wAFROC$_{\text{AUC}}$ figure of merit is implemented in the `RJafroc` function `UtilAnalyticalAucsRSM`. The Youden-index is defined as sensitivity plus specificity minus 1. Sensitivity is implemented in function `RSM_TPF` and specificity is the complement of `RSM_FPF`.

11.4 Varying λ optimizations

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

For $\mu = 2$ and $\nu = 0.9$ wAFROC$_{\text{AUC}}$ 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:

- ζ_1: the optimal threshold;
- wAFROC$_{\text{AUC}}$; the wAFROC figure of merit;
- ROC$_{\text{AUC}}$; the ROC figure of merit;
- NLF and LLF: the coordinates of the operating point on the FROC curve corresponding to ζ_1.

Table 11.1: Results for $\mu = 2$, $\nu = 0.9$ and 4 values of λ. FOM = figure of merit used in optimization.

FOM	λ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

11.4.1 Summary table

Table 11.1: The FOM column lists the quantity being maximized, the λ column lists the values of λ, the ζ_1 column lists the optimal values that maximize the chosen figure of merit. The $\text{wAFROC}_{\text{AUC}}$ column lists the AUCs under the wAFROC curves, the ROC_{AUC} column lists the AUCs under the ROC curves, and the (NLF, LLF) column lists the operating point on the FROC curves.

Inspection of this table reveals the following:

1. FROC plots, Fig. 11.1: The $\text{wAFROC}_{\text{AUC}}$ based optimal thresholds are smaller (i.e., corresponding to laxer reporting criteria) than the corresponding Youden-index based optimal thresholds. The Youden-index based operating point (black dot) is left of the $\text{wAFROC}_{\text{AUC}}$ based FROC operating point (red dot). The abscissa difference between the two points decreases with increasing λ.

2. wAFROC, Fig. 11.2, and ROC plots, Fig. 11.3: The Youden-index based optimizations yield lower performance than the corresponding $\text{wAFROC}_{\text{AUC}}$ based optimizations and the difference decreases with increasing λ.

3. For either FOM as λ increases ζ_1 increases (i.e., stricter reporting threshold). **When CAD performance decreases the algorithms adopt stricter reporting criteria.** This should make sense to the CAD algorithm designer: with decreasing performance one has to be more careful about showing CAD generated marks to the radiologist.

11.4.2 FROC

11.4.3 wAFROC

Each wAFROC plot consists of a continuous curve followed by a dashed line. The "red" curve, corresponding to $\text{wAFROC}_{\text{AUC}}$ optimization, appears as a "solid-green solid-red dashed-red" curve (the curve is in fact a true red curve complicated by superposition of the green curve over part of its traverse). The "solid-green dashed-green" curve corresponds to Youden-index optimization. As before the black dot denotes the Youden-index based operating point and the red dot denotes the $\text{wAFROC}_{\text{AUC}}$ based operating point.

The transition from continuous to dashed is determined by the value of ζ_1. It occurs at a higher value of ζ_1 (lower transition point) for the Youden-index optimization. In other words the stricter Youden-index based threshold sacrifices some of the area under the wAFROC resulting 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.

11.4.4 ROC

The decrease in ROC_{AUC} with increasing λ is illustrated in Fig. 11.3 which shows RSM-predicted ROC plots for the two optimization methods for the 4 values of λ. Again, each plot consists of a continuous curve followed by a dashed

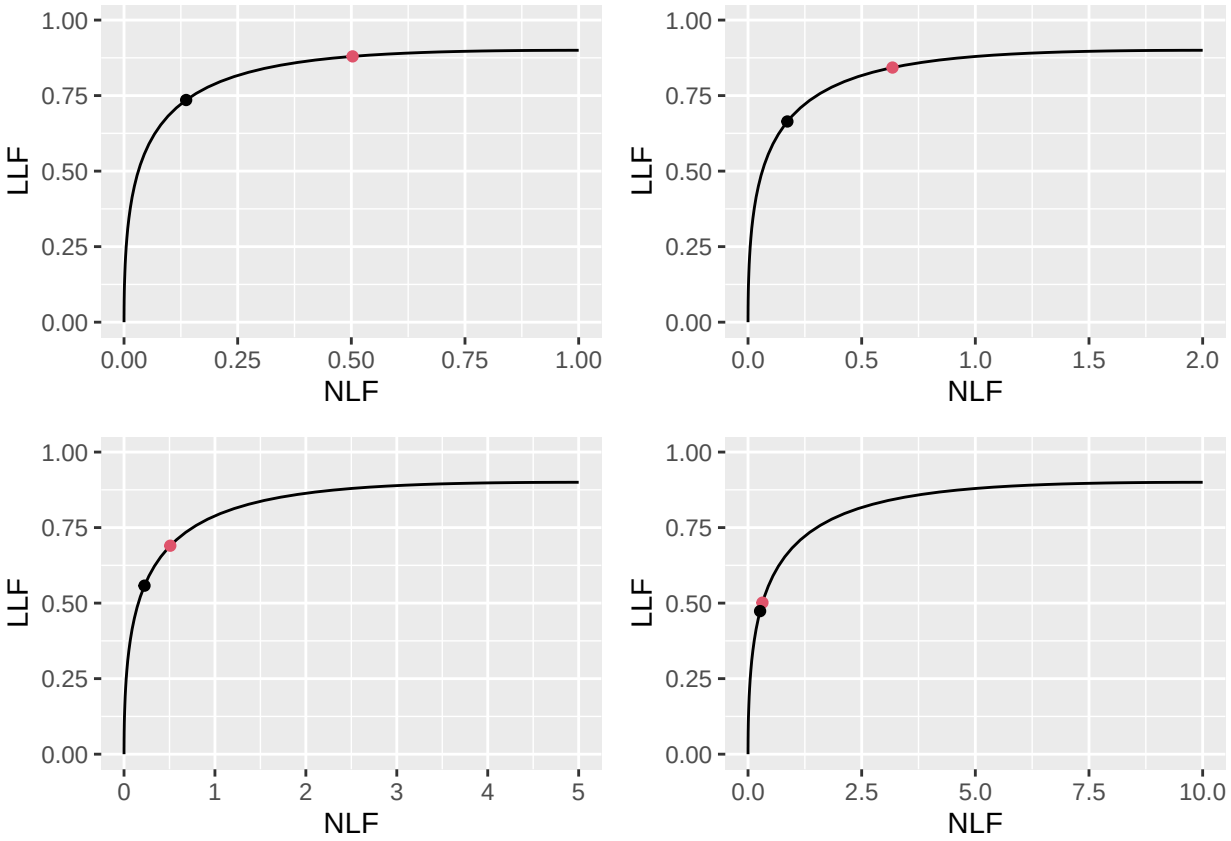


Figure 11.1: FROC plots with superimposed operating points for varying λ. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization.

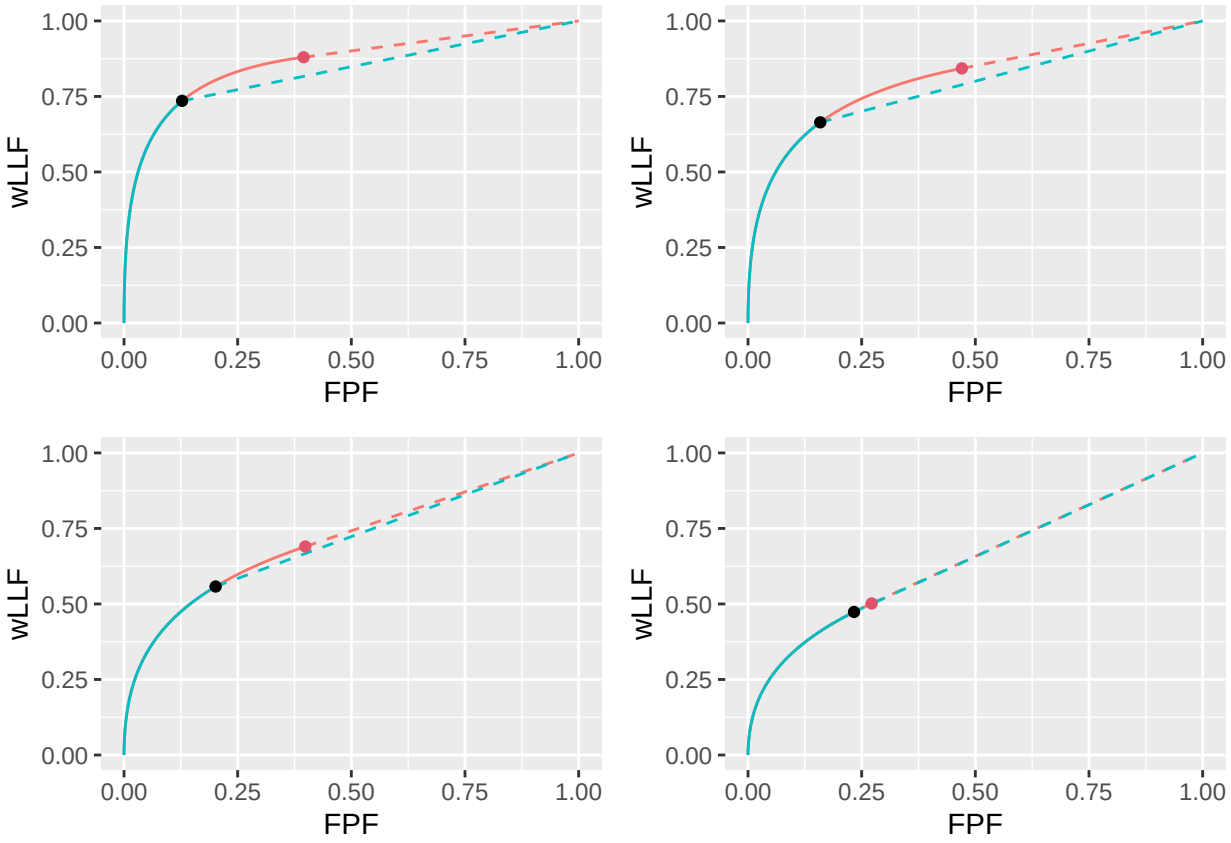


Figure 11.2: wAFROC plots for the two optimization methods: the "solid-green solid-red dashed-red" curve corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the "solid-green dashed-green" curve corresponds to Youden-index optimization. The $\text{wAFROC}_{\text{AUC}}$ optimizations yield greater performance than do Youden-index optimizations and the difference decreases with increasing λ.

curve and a similar color-coding convention is used as in Fig. 11.2. The ROC plots show similar dependencies as the wAFROC plots: the stricter Youden-index based reporting thresholds sacrifice some of the area under the ROC resulting in lower performance, particularly for the lower values of λ.

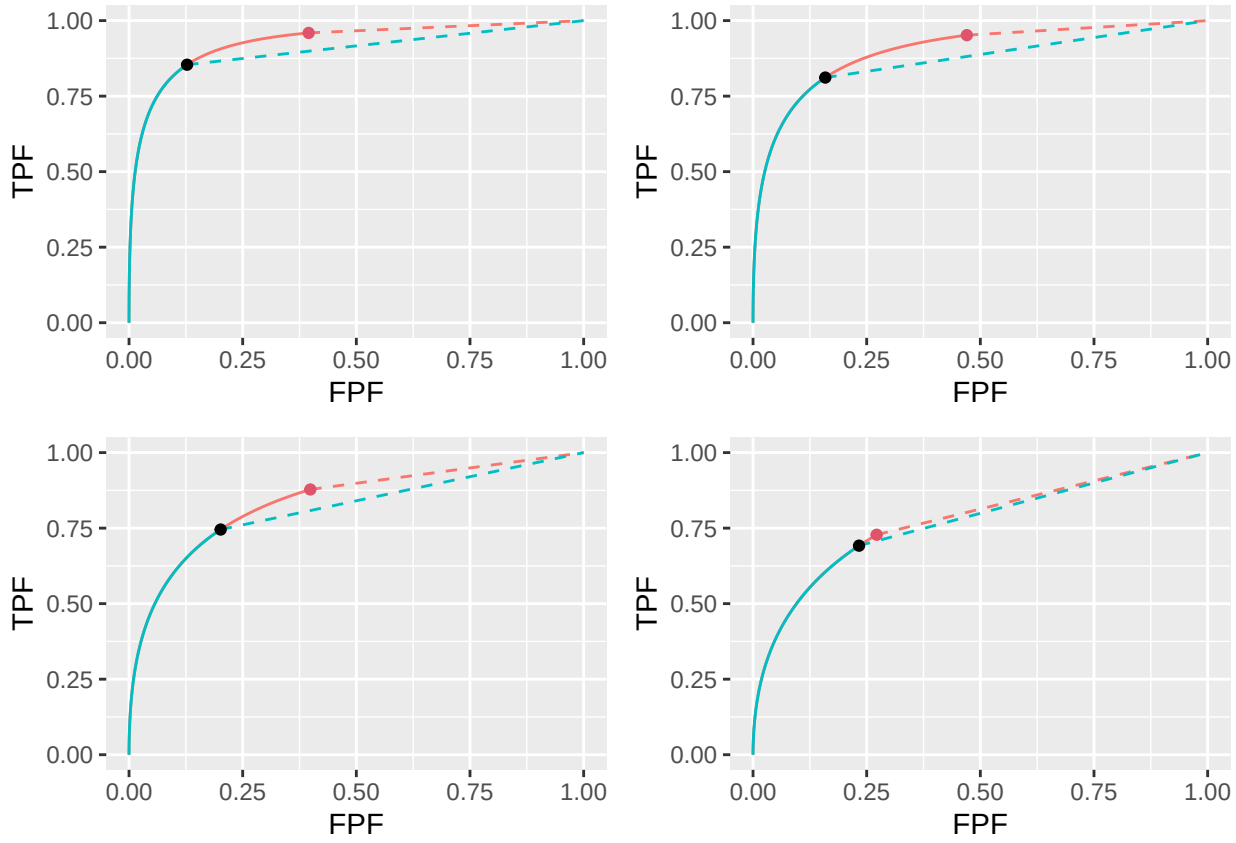


Figure 11.3: ROC plots for the two optimization methods: the "solid-green solid-red dashed-red" curve corresponds to $ ext{wAFROC}_ ext{AUC}$ optimization and the "solid-green dashed-green" curve corresponds to Youden-index optimization. The $\text{wAFROC}_{\text{AUC}}$ optimizations yield greater performance than Youden-index optimizations and the difference decreases with increasing λ.

11.4.5 Why not maximize ROC-AUC?

Since the ROC curves show similarities to the wAFROC curves, why not maximize ROC_{AUC} instead of $\text{wAFROC}_{\text{AUC}}$? It can be shown that as long as one restricts to proper ROC models this always results in $\zeta_1 = -\infty$, i.e., all latent marks are to be shown to the radiologist, an obviously incorrect strategy. This result can be understood from the following geometrical argument.

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 $\zeta_1 = -\infty$, see Fig. 11.4.

11.5 Varying ν and μ optimizations

Details of varying ν (with μ and λ held constant) are in Appendix 11.10.1. The results, summarized in Table 11.3, are similar to those just described for varying λ but, since unlike as was the case with increasing λ, increasing ν results in increasing performance, the *directions of the effects are reversed*. For $\text{wAFROC}_{\text{AUC}}$ optimization the

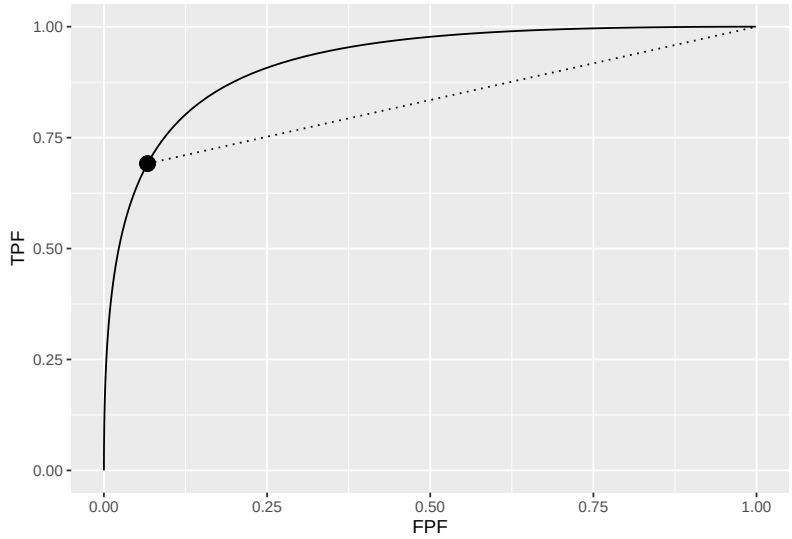


Figure 11.4: In the region above the dot the proper curve is above the dotted line, meaning that performance of an observer who adopts a finite ζ_1 is less than performance of an observer who adopts $\zeta_1 = -\infty$.

optimal reporting threshold ζ_1 decreases with increasing ν. In contrast the Youden-index based optimal threshold is almost independent of ν. For $\text{wAFROC}_{\text{AUC}}$ optimization the FROC operating point moves to higher NLF values while the Youden-index based operating point stays at a near constant NLF value, see Fig. 11.8). As before, $\text{wAFROC}_{\text{AUC}}$ optimizations yield higher performances than Youden-index optimizations (particularly for larger ν): see Fig. 11.9 for the wAFROC and Fig. 11.10 for the ROC. The difference between the two optimization methods *increases* with increasing ν (for comparison the difference between the methods decreases with increasing λ – this is what I meant by "reversed effects").

Details of varying μ (with λ and ν held constant) are in Appendix 11.10.2. The results are summarized in Table 11.4. Increasing μ is accompanied by increasing ζ_1 (i.e., stricter reporting threshold) and increasing $\text{wAFROC}_{\text{AUC}}$ and ROC_{AUC}. Performance measured either way is higher for $\text{wAFROC}_{\text{AUC}}$ optimizations but the difference tends to shrink at the larger values of μ. LLF is relatively constant for $\text{wAFROC}_{\text{AUC}}$ optimizations while it increases slowly with μ for Youden-index optimizations. NLF decreases with increasing μ for both optimization methods, i.e, the FROC operating point shifts leftward, see Fig. 11.11). Again, $\text{wAFROC}_{\text{AUC}}$ optimization yields a lower reporting threshold and higher performance than Youden-index optimization, see Fig. 11.12 for the wAFROC and Fig. 11.13 for the ROC. The difference between the two optimization methods decreases with increasing μ.

11.6 Limiting situations

Limiting situations covering high and low performances are described in 11.10.3.

For high performance, defined as $\text{ROC}_{\text{AUC}} > 0.9$, both methods place the optimal operating point near the inflection point on the upper-left corner of the wAFROC or ROC. The $\text{wAFROC}_{\text{AUC}}$ based method chooses a lower threshold than the Youden-index method resulting in a higher operating point on the FROC and higher $\text{wAFROC}_{\text{AUC}}$ and ROC_{AUC}. The difference between the two methods decreases as $\text{ROC}_{\text{AUC}} \to 1$.

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

11.7 Trends

No matter how the RSM parameters are varied the trend is that $\text{wAFROC}_{\text{AUC}}$ optimizations result in lower optimal thresholds ζ_1 (i.e.,laxer reporting criteria that result in more displayed marks) than Youden-index optimizations. Accordingly the $\text{wAFROC}_{\text{AUC}}$ optimizations yield FROC operating points at higher NLF values (i.e., red dots to the right of the black dots in FROC plots), greater $\text{wAFROC}_{\text{AUC}}$s (red curves above the green curves in wAFROC plots) and greater ROC_{AUC}s (red curves above the green curves in ROC plots). These trends are true no matter how the RSM parameters are varied provided CAD performance is not too low.

If CAD performance is very low there are instructive exceptions where $\text{wAFROC}_{\text{AUC}}$ optimizations yield *greater* ζ_1 (i.e., stricter reporting criteria that result in *fewer* displayed marks) than Youden-index optimizations. This finding is true no matter how the RSM parameters are varied.

Consider for example the low performance varying ν optimizations described in Appendix 11.10.3.6. The FROC plots, Fig. 11.29, corresponding to $\mu = 1$, $\lambda = 10$, $\nu = 0.1, 0.2, 0.3, 0.4$, show that the $\text{wAFROC}_{\text{AUC}}$ optimal operating points are very close to the origin NLF = 0, i.e., very few marks are displayed. In contrast the Youden-index optimal operating points are shifted towards larger NLF values allowing more marks to be displayed. The wAFROC plots, Fig. 11.30, show a large difference in AUCs between the two methods, especially for the smaller values of ν: for example, for $\nu = 0.1$, the $\text{wAFROC}_{\text{AUC}}$ corresponding to $\text{wAFROC}_{\text{AUC}}$ optimization is 0.5000002 while that corresponding to Youden-index optimization is 0.2923394. Clearly the $\text{wAFROC}_{\text{AUC}}$ optimization yields a larger $\text{wAFROC}_{\text{AUC}}$ relative to Youden-index optimization, which it must as $\text{wAFROC}_{\text{AUC}}$ is the quantity being optimized.

While Youden-index optimizations yield smaller $\text{wAFROC}_{\text{AUC}}$ values they do yield larger ROC_{AUC} values as is evident by comparing the ROC plots, Fig. 11.31. For $\nu = 0.1$ the ROC_{AUC} corresponding to $\text{wAFROC}_{\text{AUC}}$ optimization is 0.5000024 while that corresponding to Youden-index optimization is 0.5143474. Clearly $\text{wAFROC}_{\text{AUC}}$ optimization yields a very close to chance-level ROC_{AUC} while Youden-index optimization yields a slightly larger ROC_{AUC}.

Keep in mind that ROC_{AUC} measures classification accuracy performance between non-diseased and diseased cases: it does not care about lesion localization accuracy. In contrast $\text{wAFROC}_{\text{AUC}}$ measures both lesion localization accuracy and lesion classification accuracy. By choosing an optimal operating point close to the origin the low performance CAD does not get credit for missing almost all the lesions on diseased cases but it does get credit for not marking non-diseased cases.

11.8 Applying 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. These steps are illustrated in the following example.

11.8.1 A CAD application

Not having access to any CAD FROC datasets the standalone CAD LROC dataset described in (Hupse et al., 2013) was used to create a simulated FROC (i.e., ROC_{AUC} equivalent) dataset which is 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, are ignored). The function `DfFroc2Roc` converts `dsCad` 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

Table 11.2: 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)

search model (RSM). Object `fit` contains the RSM parameters required to perform the optimizations described in previous sections.

```
ds <- RJafroc::datasetCadSimuFroc
dsCad <- RJafroc::DfExtractDataset(ds, rdrs = 1)
dsCadRoc <- RJafroc::DfFroc2Roc(dsCad)
dsCadRocBinned <- RJafroc::DfBinDataset(dsCadRoc, opChType = "ROC")
lesDistrCad <- c(1) # LROC dataset has one lesion per diseased case
relWeightsCad <- c(1)
fit <- RJafroc::FitRsmRoc(dsCadRocBinned, lesDistrCad)
cat(sprintf("fitted values: mu = %5.3f,", fit$mu),
    sprintf("lambda = %5.3f,", fit$lambda),
    sprintf("nu = %5.3f.", fit$nu))
#> fitted values: mu = 2.756, lambda = 6.778, nu = 0.803.
```

11.8.1.1 Summary table

Table 11.2 summarizes the results. As compared to Youden-index optimization the $\text{wAFROC}_{\text{AUC}}$ based optimization results in a lower reporting threshold ζ_1, larger figures of merit – see Fig. 11.6 for $\text{wAFROC}_{\text{AUC}}$ and Fig. 11.7 for ROC_{AUC} – and a higher operating point on the FROC, see Fig. 11.5. These results match the trends shown in Table 11.1.

11.8.1.2 FROC

Fig. 11.5 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.

11.8.1.3 wAFROC

Fig. 11.6 shows wAFROC curves using the two methods. The red curve is using $\text{wAFROC}_{\text{AUC}}$ optimization and the green curve is using Youden-index optimization. The difference in AUCs is small - following the trend described in Appendix 11.5 for the larger values of λ.

11.8.1.4 ROC

Fig. 11.7 shows ROC curves using the two methods. The red curve is using $\text{wAFROC}_{\text{AUC}}$ optimization and the green curve is using Youden-index optimization.

11.9 TBA Discussion

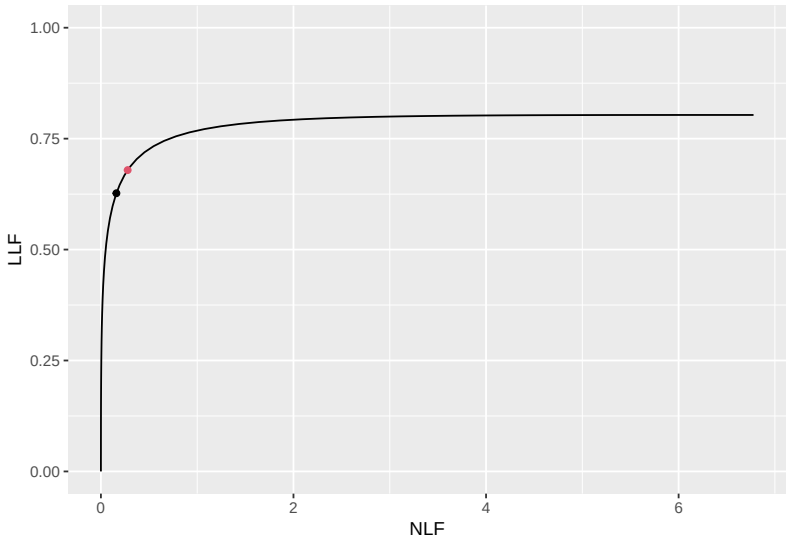


Figure 11.5: FROC plots with superposed optimal operating points. The red dot is using $\text{wAFROC}_{\text{AUC}}$ optimization and black dot is using Youden-index optimization.

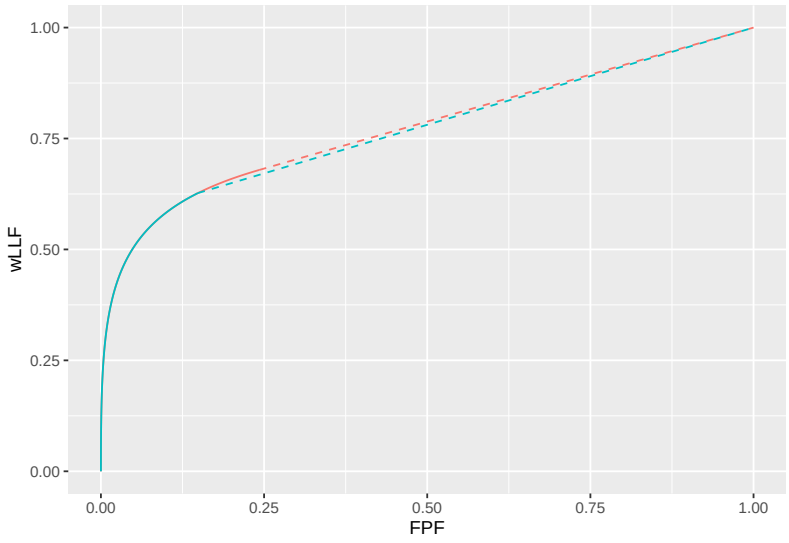


Figure 11.6: The color coding is as in previous figures. The two $\text{wAFROC}_{\text{AUC}}$s are 0.774 (wAFROC optimization) and 0.770 (Youden-index optimization).

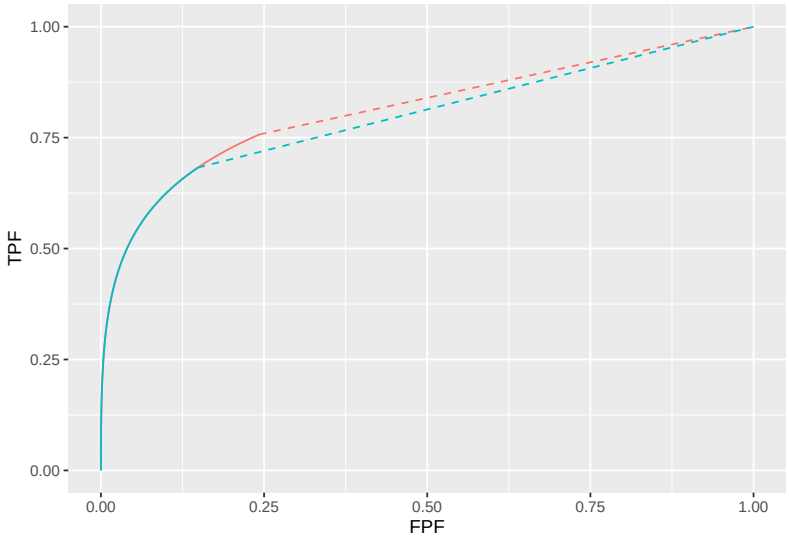


Figure 11.7: The color coding is as in previous figures. The two ROC_{AUC}s are 0.815 (wAFROC optimization) and 0.798 (Youden-index optimization).

```
muArr <- c(2)
lambdaArr <- c(1, 2, 5, 10)
nuArr <- c(0.9)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
source("R/optim-op-point/doOneTable.R", local = knitr::knit_global())
```

In Table 11.1 the λ parameter controls the average number of perceived NLs per case. For $\lambda = 1$ there is, on average, one perceived NL for every case and the optimal $wAFROC_{AUC}$ based threshold is ζ_1 = -0.007. For $\lambda = 10$ there are ten perceived NLs for every case and the optimal $wAFROC_{AUC}$ based threshold is $\zeta_1 = 1.856$. The reason for the increase in ζ_1 should be obvious: with increasing numbers of latent NLs (perceived false marks) per case it is necessary to adopt a stricter criteria because otherwise the reader would be shown 10 times the number of false marks per case.

The ROC_{AUC}s are reported as a check of the less familiar $wAFROC_{AUC}$ figure of merit. With some notable exceptions the behavior 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. The exceptions occur when CAD performance is very low in which situation the .

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 latent 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 "ROC-paradigm" 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 7.7 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 $wAFROC_{AUC}$ optimizations NLF peaks while for Youden-index optimizations it increases.

The FROC plots also illustrate the decrease in LLF 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

Table 11.3: Results for $\mu = 2$, $\lambda = 1$ and varying ν.

FOM	ν	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

rules out an unambiguous determination of the direction of the change in overall performance based on the FROC curve.

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

11.10 Appendices

11.10.1 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)
lambdaArr <- c(1)
nuArr <- c(0.6, 0.7, 0.8, 0.9)
lesDistr <- c(0.5, 0.5)
relWeights <- c(0.5, 0.5)
```

11.10.1.1 Summary table

11.10.1.2 FROC

11.10.1.3 wAFROC

11.10.1.4 ROC

11.10.2 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)
lambdaArr <- 1
nuArr <- 0.9
```

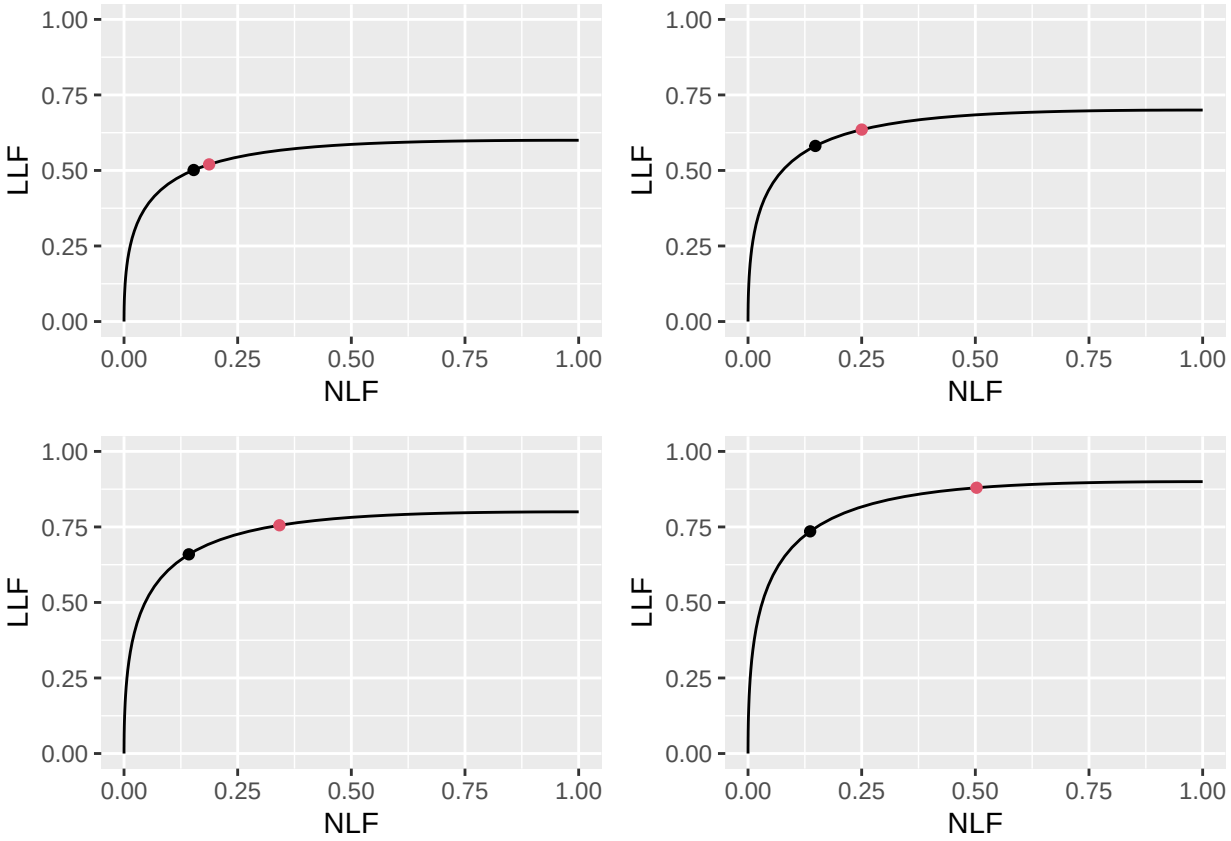


Figure 11.8: Varying ν FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$. Each red dot is above the corresponding black dot and their separation increases as ν increases, i.e., as CAD performance increases.

Table 11.4: Results for $\lambda = 1$, $\nu = 0.9$ and varying μ.

FOM	μ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

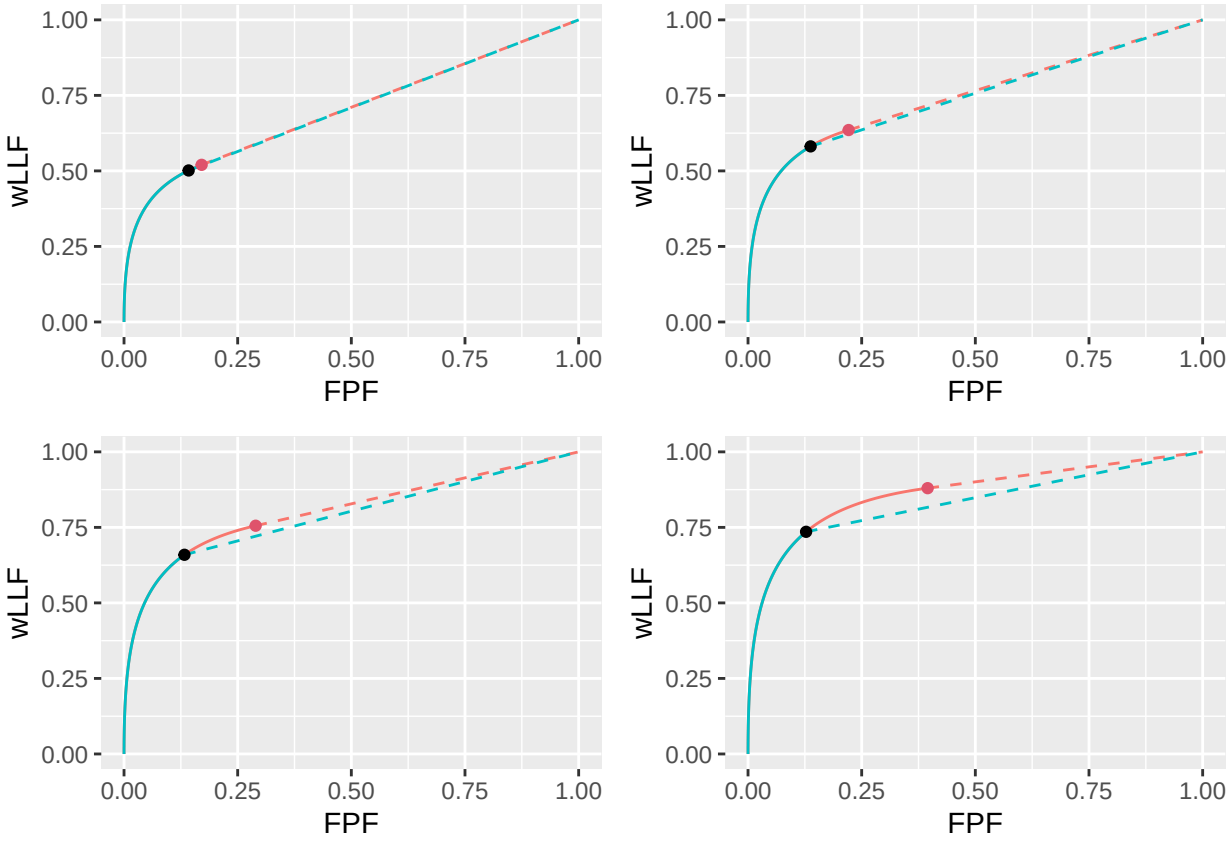


Figure 11.9: Varying ν wAFROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$.

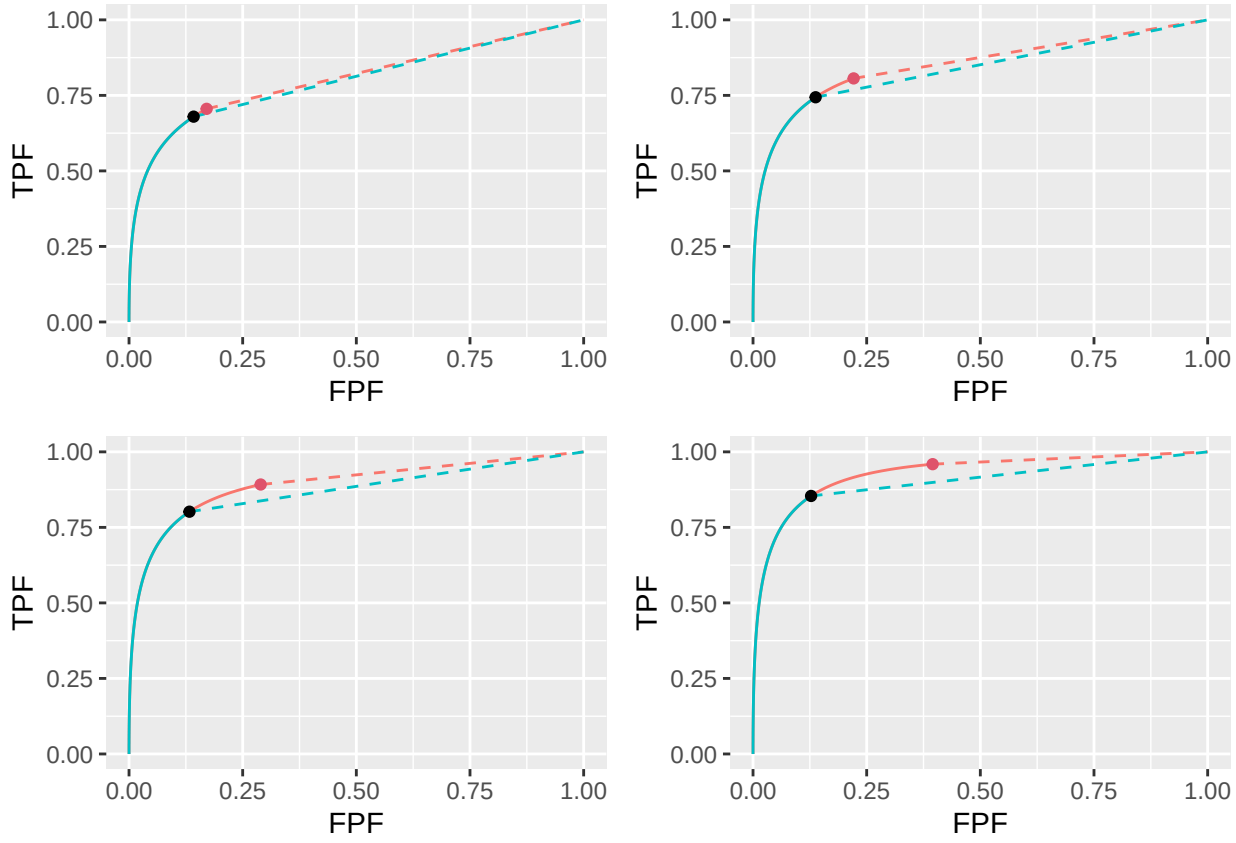


Figure 11.10: Varying ν ROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$.

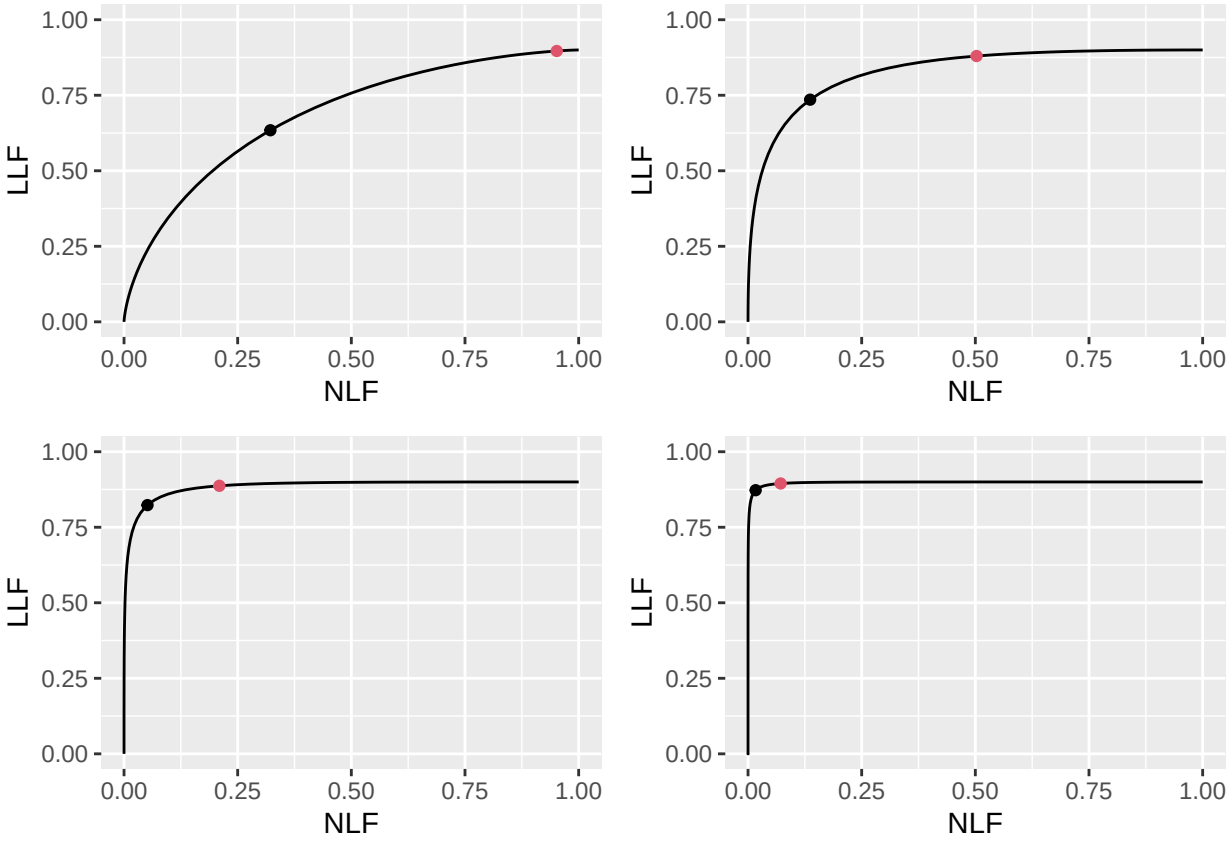


Figure 11.11: Varying μ FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

11.10.2.1 Summary table

11.10.2.2 FROC

11.10.2.3 wAFROC

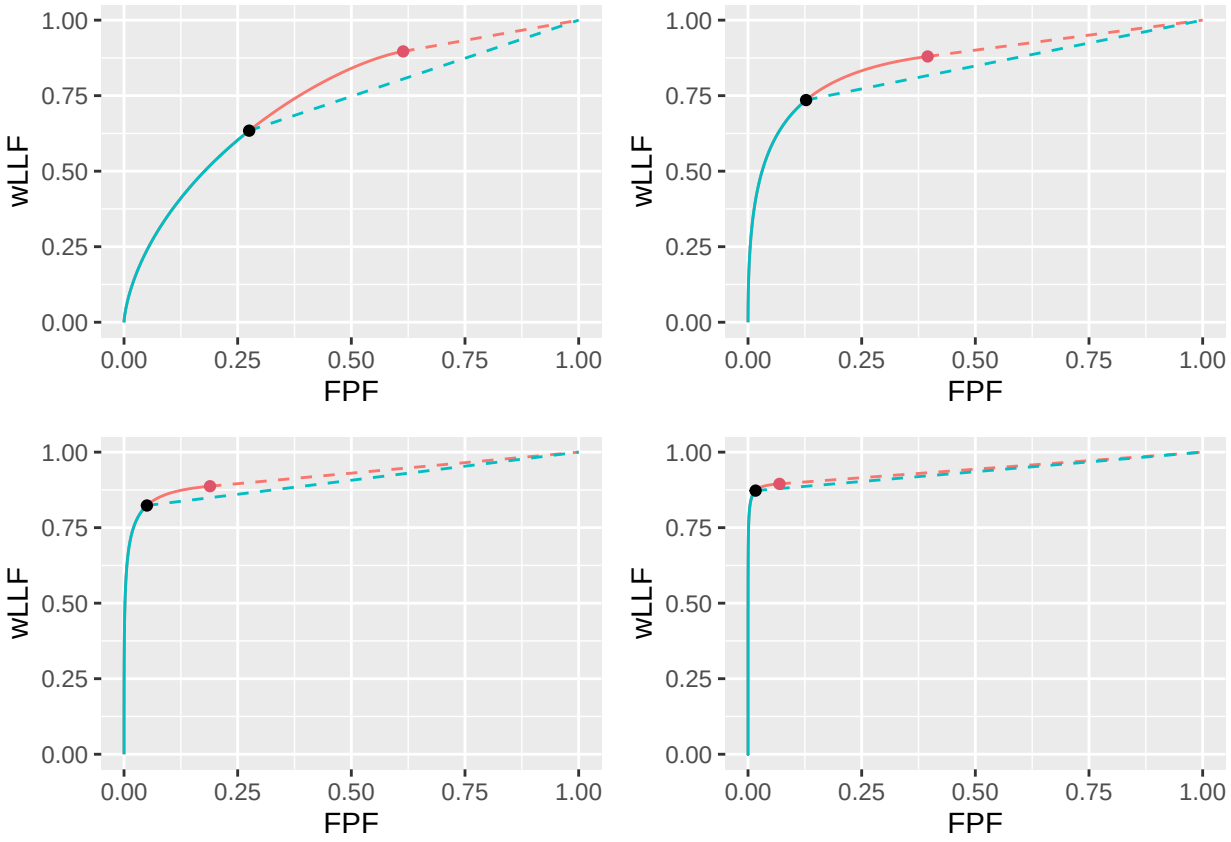


Figure 11.12: Varying μ wAFROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

11.10.2.4 ROC

11.10.3 Limiting cases

11.10.3.1 High performance varying μ

```
muArr <- c(2, 3, 4, 5)
nuArr <- c(0.9)
lambdaArr <- c(1)
```

11.10.3.1.1 Summary table

11.10.3.1.2 FROC

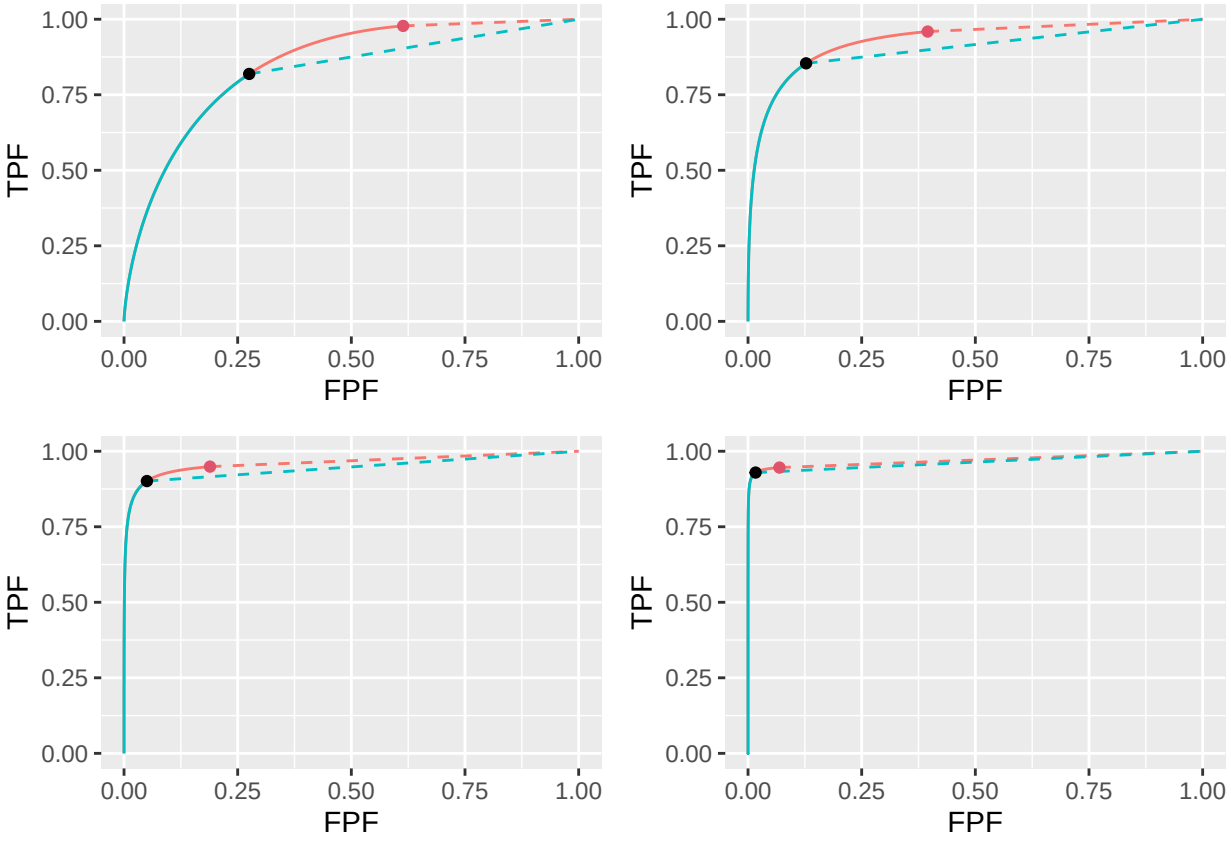


Figure 11.13: Varying μ ROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

Table 11.5: High performance summary of optimization results for $\lambda = 1$ and $\nu = 0.9$ and varying μ.

FOM	μ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

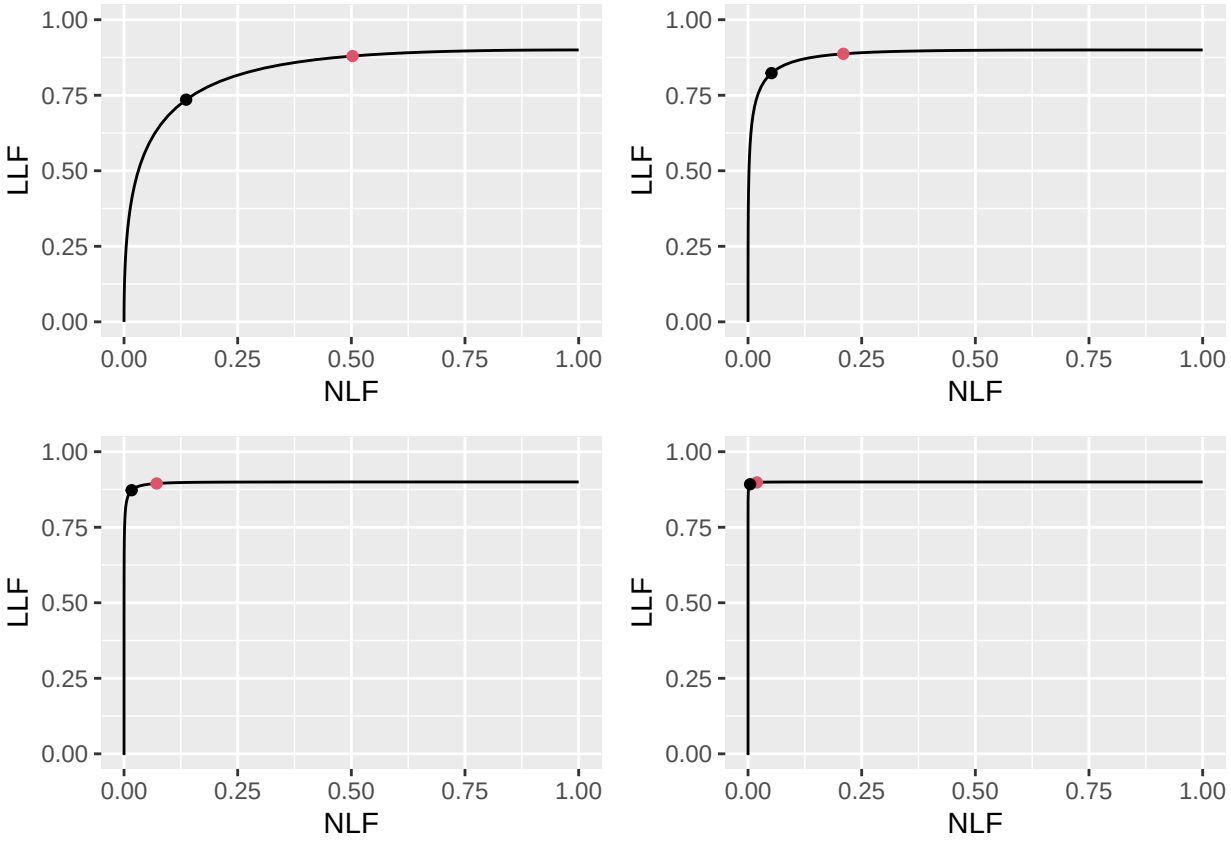


Figure 11.14: High performance varying μ FROC plots with superimposed operating points. The red dot corresponds to wAFROC$_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of μ are: top-left $\mu = 2$, top-right $\mu = 3$, bottom-left $\mu = 4$ and bottom-right $\mu = 5$.

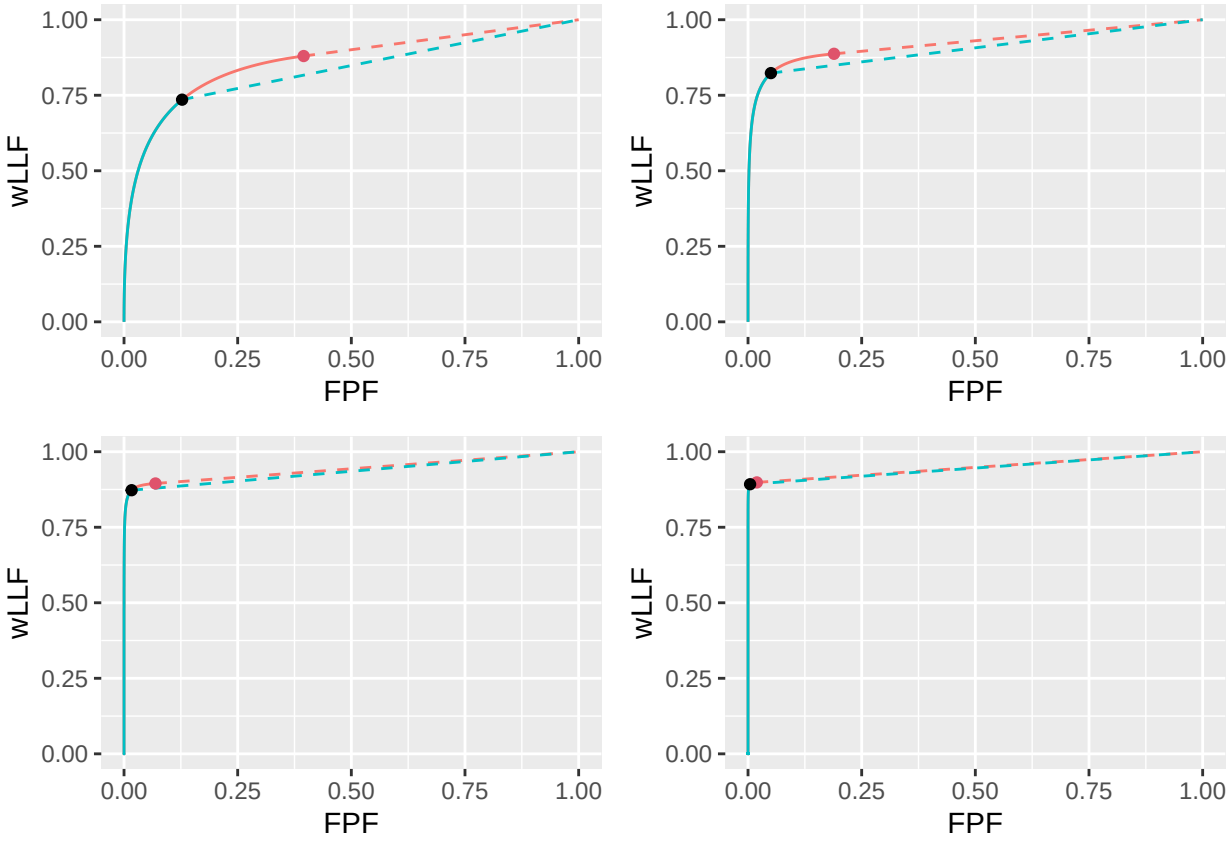


Figure 11.15: High performance varying μ wAFROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 2$, top-right $\mu = 3$, bottom-left $\mu = 4$ and bottom-right $\mu = 5$.

11.10.3.1.3 wAFROC

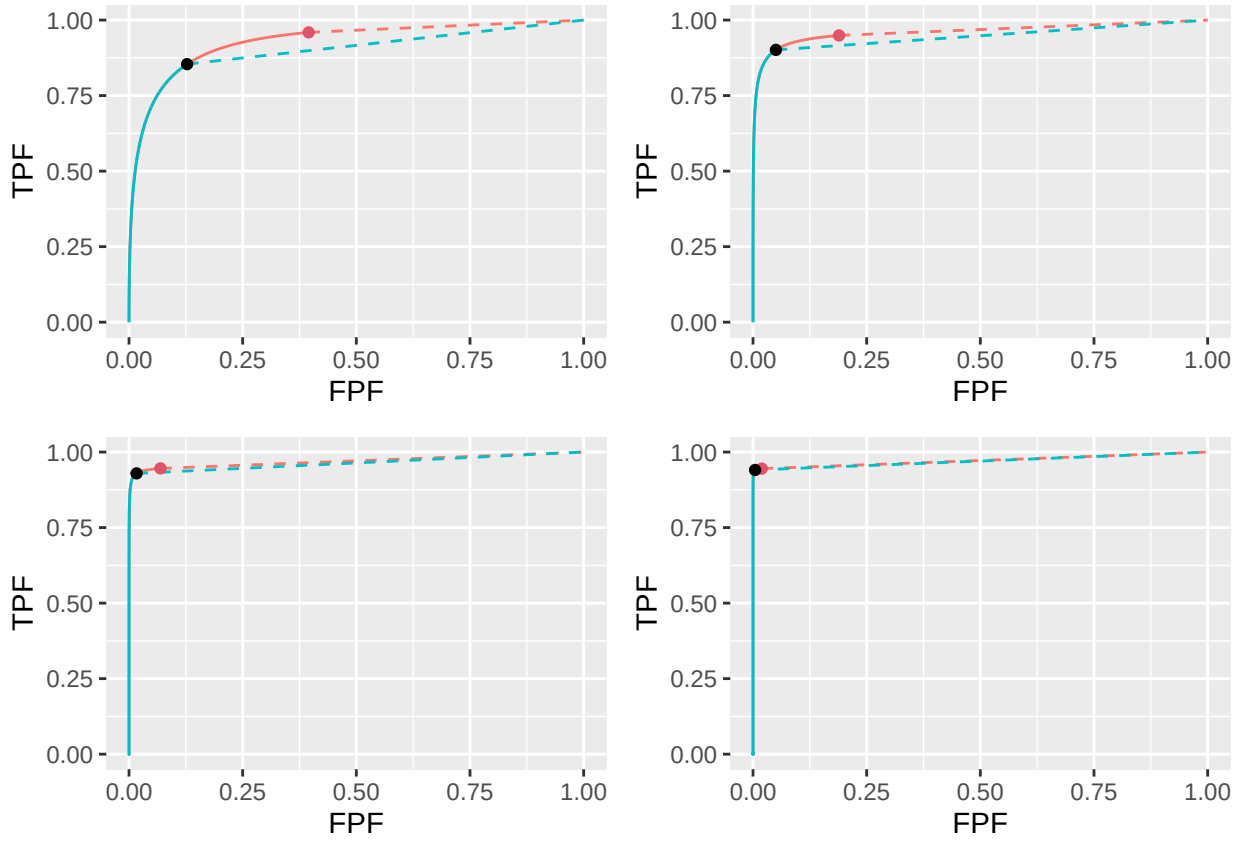


Figure 11.16: High performance varying μ ROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 2$, top-right $\mu = 3$, bottom-left $\mu = 4$ and bottom-right $\mu = 5$.

11.10.3.1.4 ROC

11.10.3.2 Low performance varying μ

```
muArr <- c(1, 2, 3, 4)
nuArr <- c(0.1)
lambdaArr <- c(10)
```

11.10.3.2.1 Summary table

11.10.3.2.2 FROC

11.10.3.2.3 wAFROC

11.10.3.2.4 ROC

Table 11.6: Low performance summary of optimization results for $\lambda = 10$ and $nu = 0.1$ and varying μ. Column labeling as in previous tables.

FOM	μ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

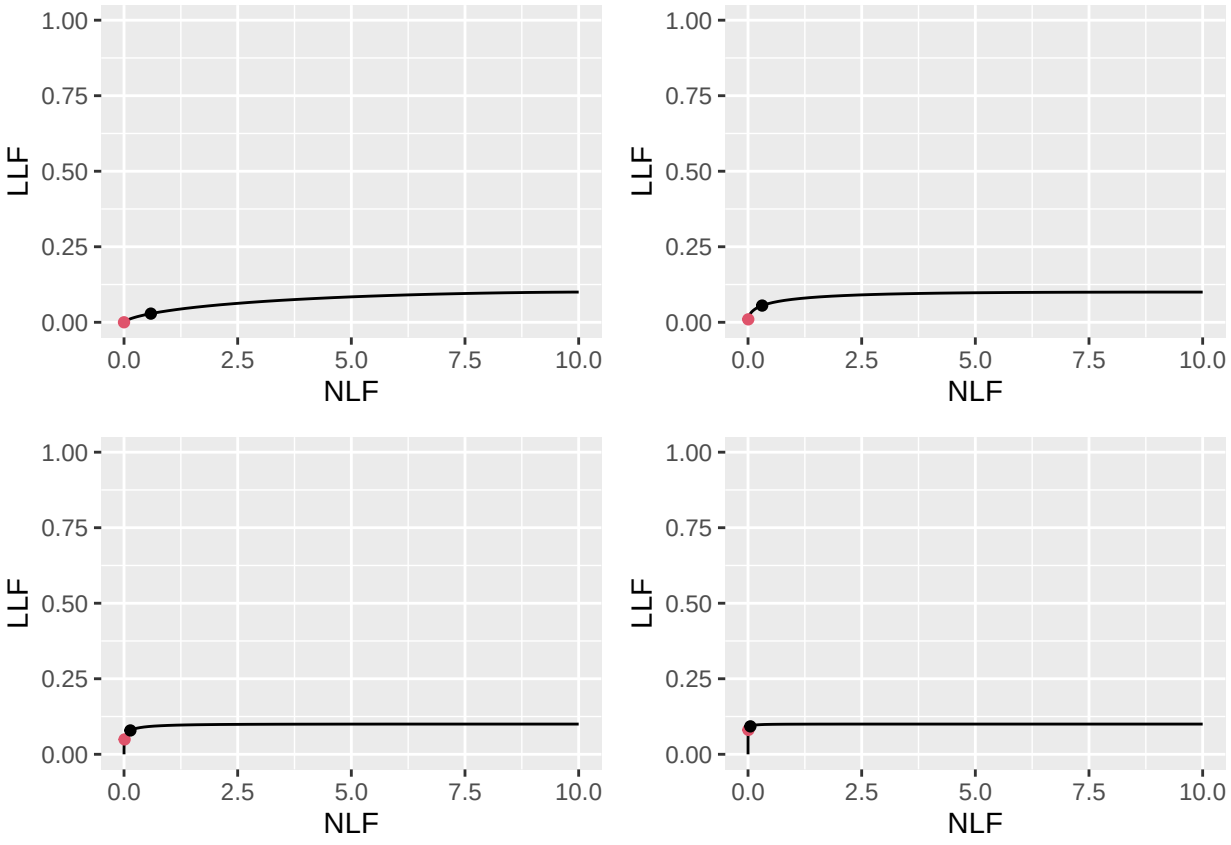


Figure 11.17: Low performance varying μ FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

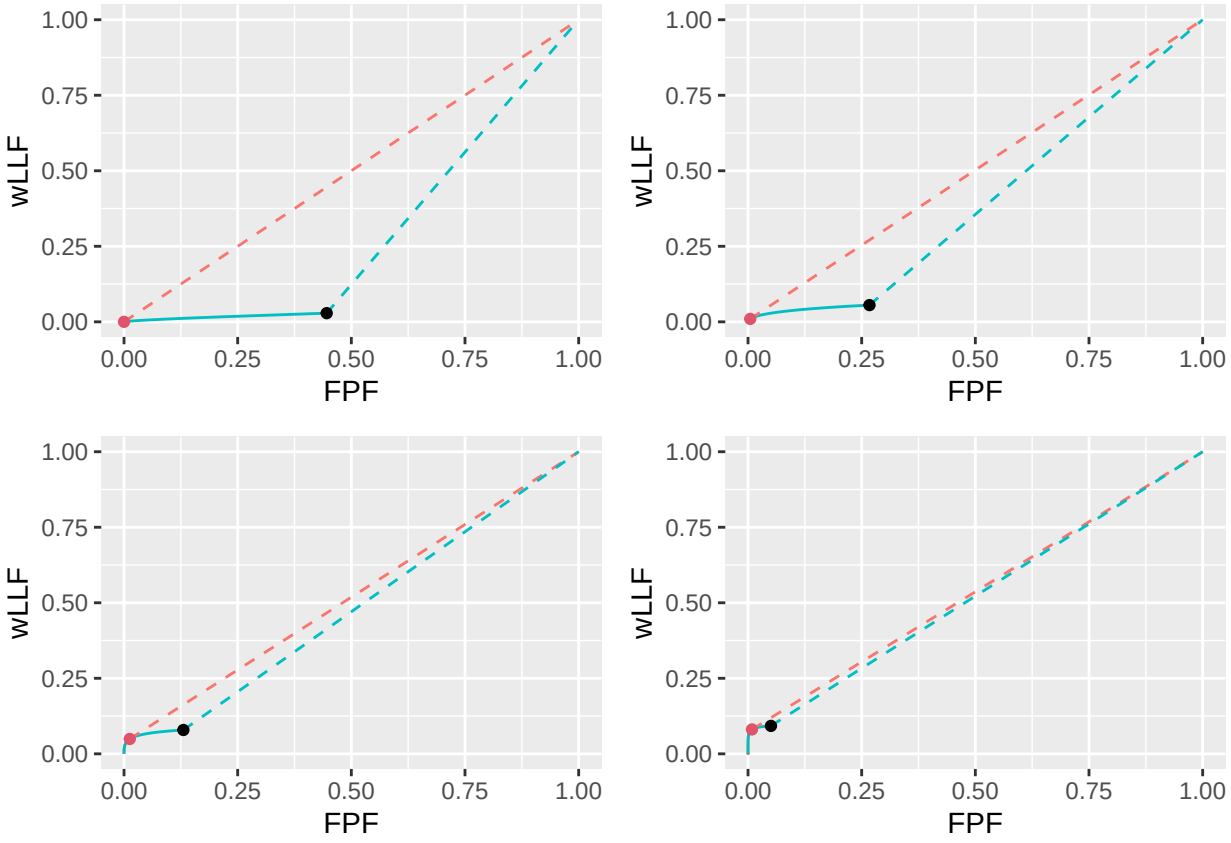


Figure 11.18: Low performance varying μ wAFROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

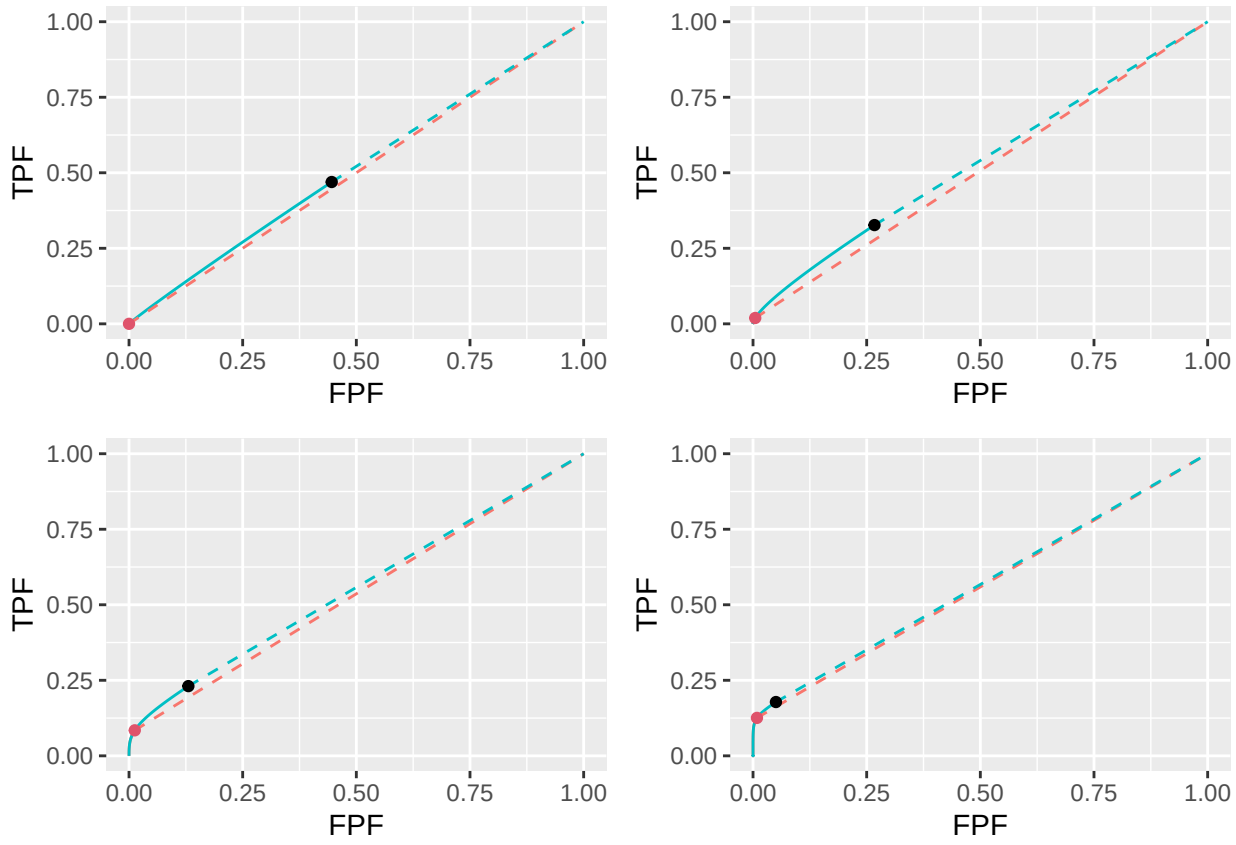


Figure 11.19: Low performance varying μ ROC plots for the two optimization methods with superimposed operating points with superimposed operating points. The color coding is as in previous figures. The values of μ are: top-left $\mu = 1$, top-right $\mu = 2$, bottom-left $\mu = 3$ and bottom-right $\mu = 4$.

Table 11.7: Results for $\mu = 4$, $nu = 0.9$ and varying λ. Column labeling as in previous tables.

FOM	λ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

Table 11.8: Results for $\mu = 1$, $\nu = 0.2$ and varying λ. Column labeling as in previous tables.

FOM	λ	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

11.10.3.3 High performance varying λ

```
muArr <- c(4)
nuArr <- c(0.9)
lambdaArr <- c(1,2,5,10)
```

11.10.3.3.1 Summary table

11.10.3.3.2 FROC

11.10.3.3.3 wAFROC

11.10.3.3.4 ROC

11.10.3.4 Low performance varying λ

```
muArr <- c(1)
nuArr <- c(0.2)
lambdaArr <- c(1, 2, 5, 10)
```

11.10.3.4.1 Summary table

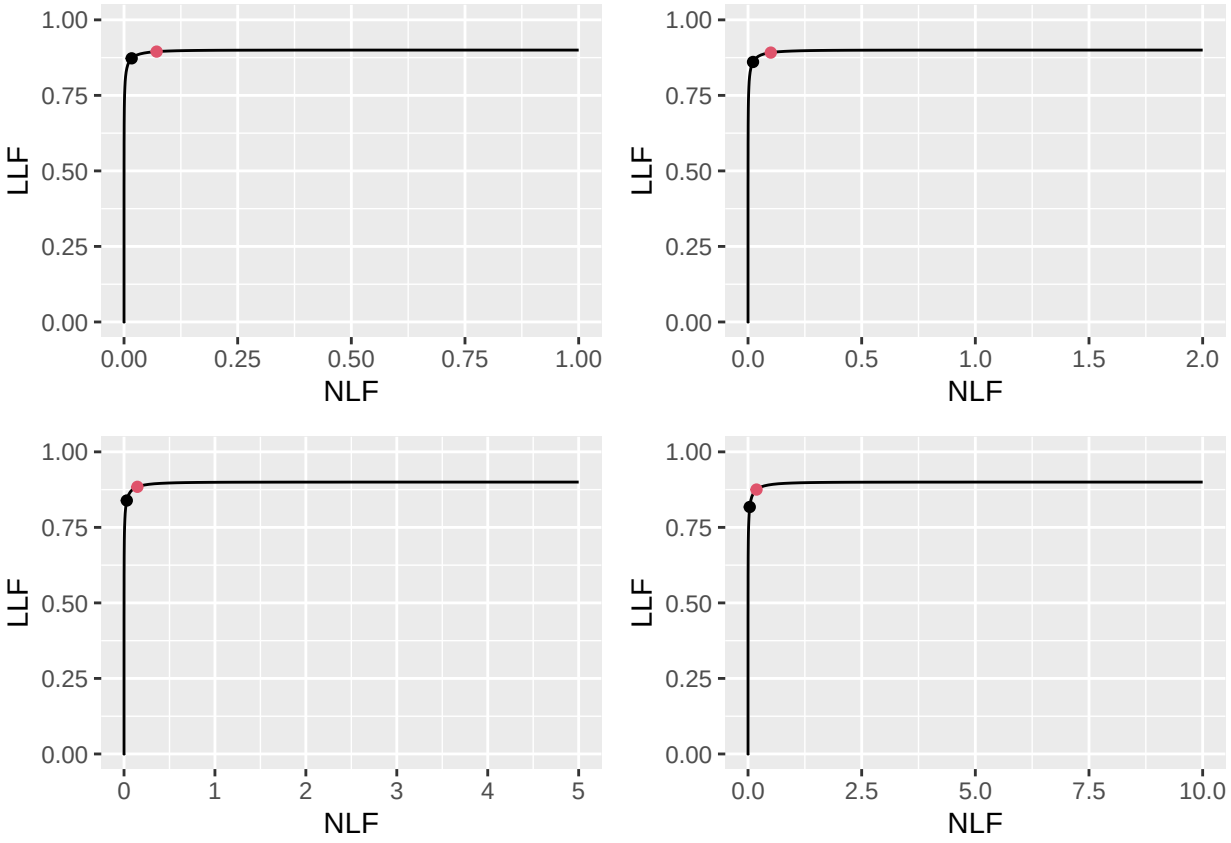


Figure 11.20: High performance varying λ FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

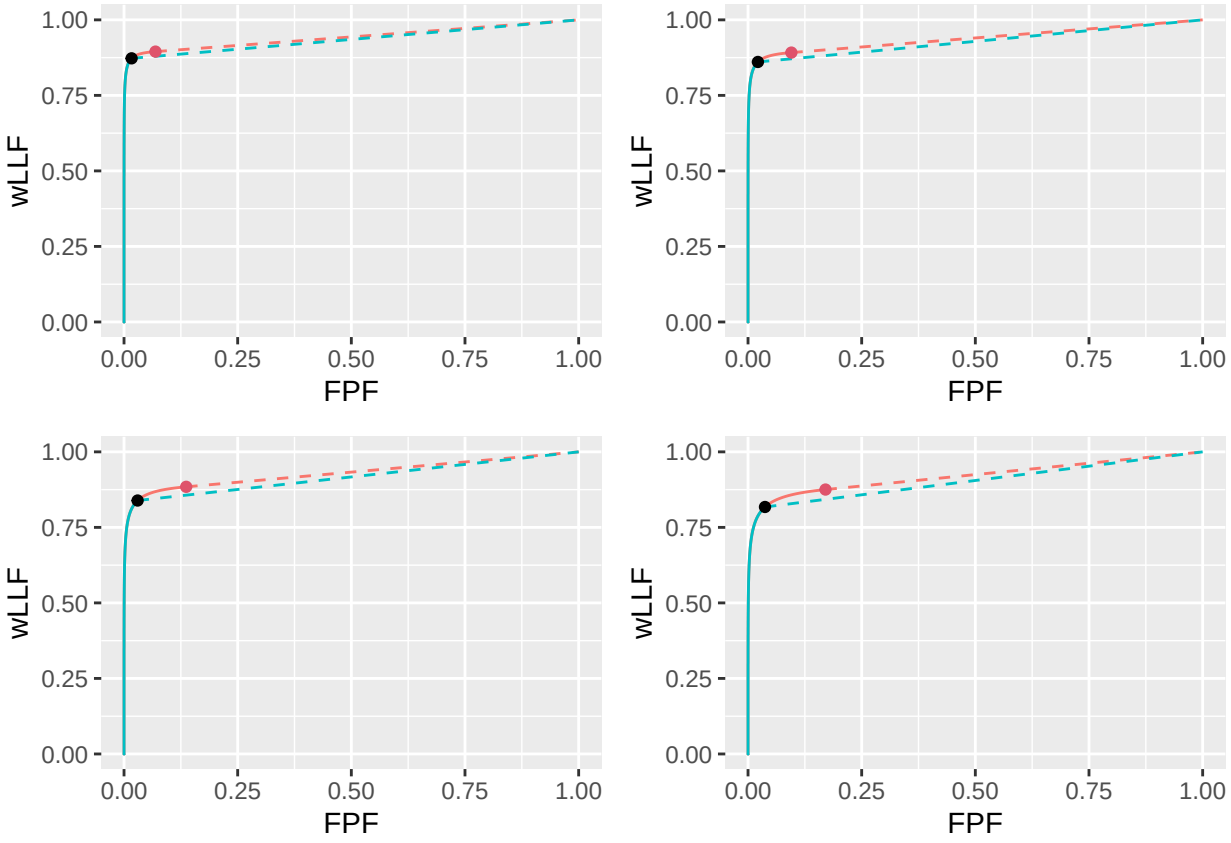


Figure 11.21: High performance varying λ wAFROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

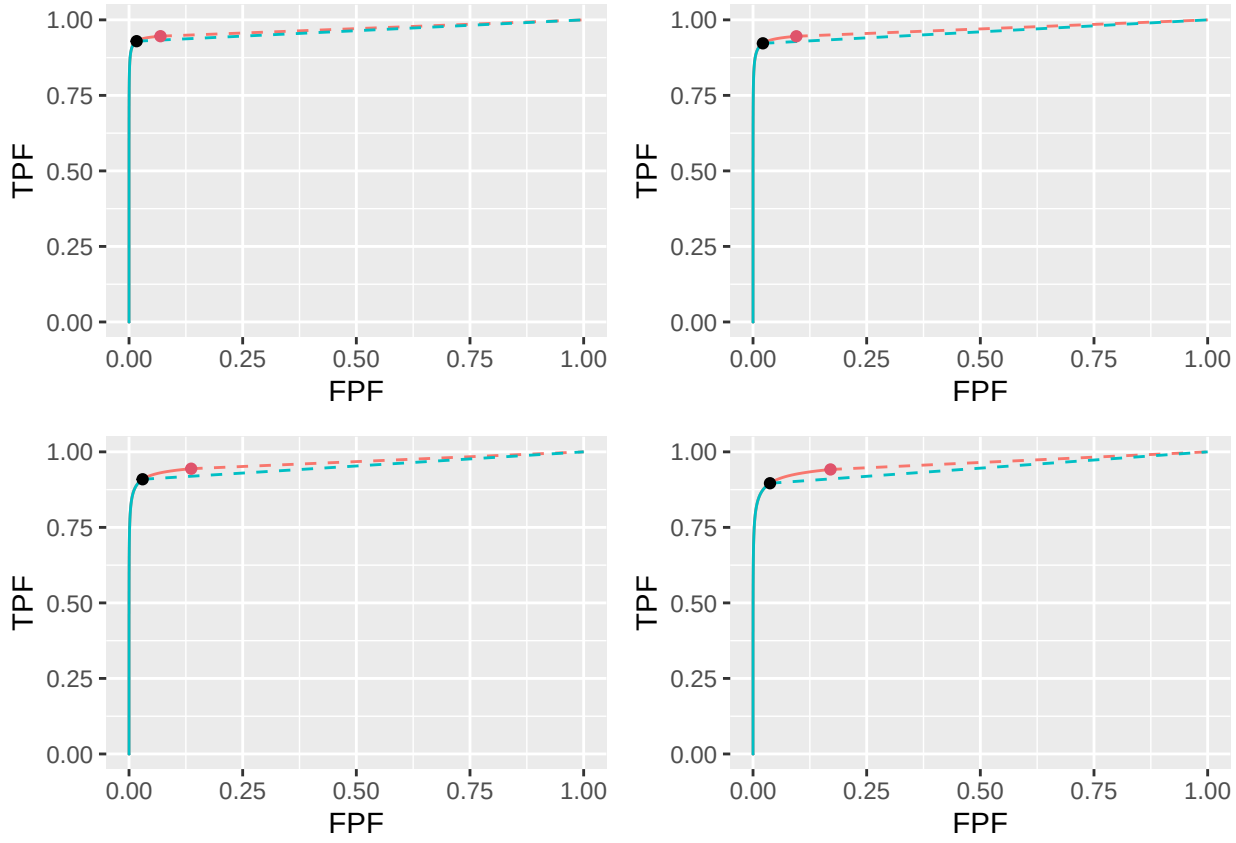


Figure 11.22: High performance varying λ ROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

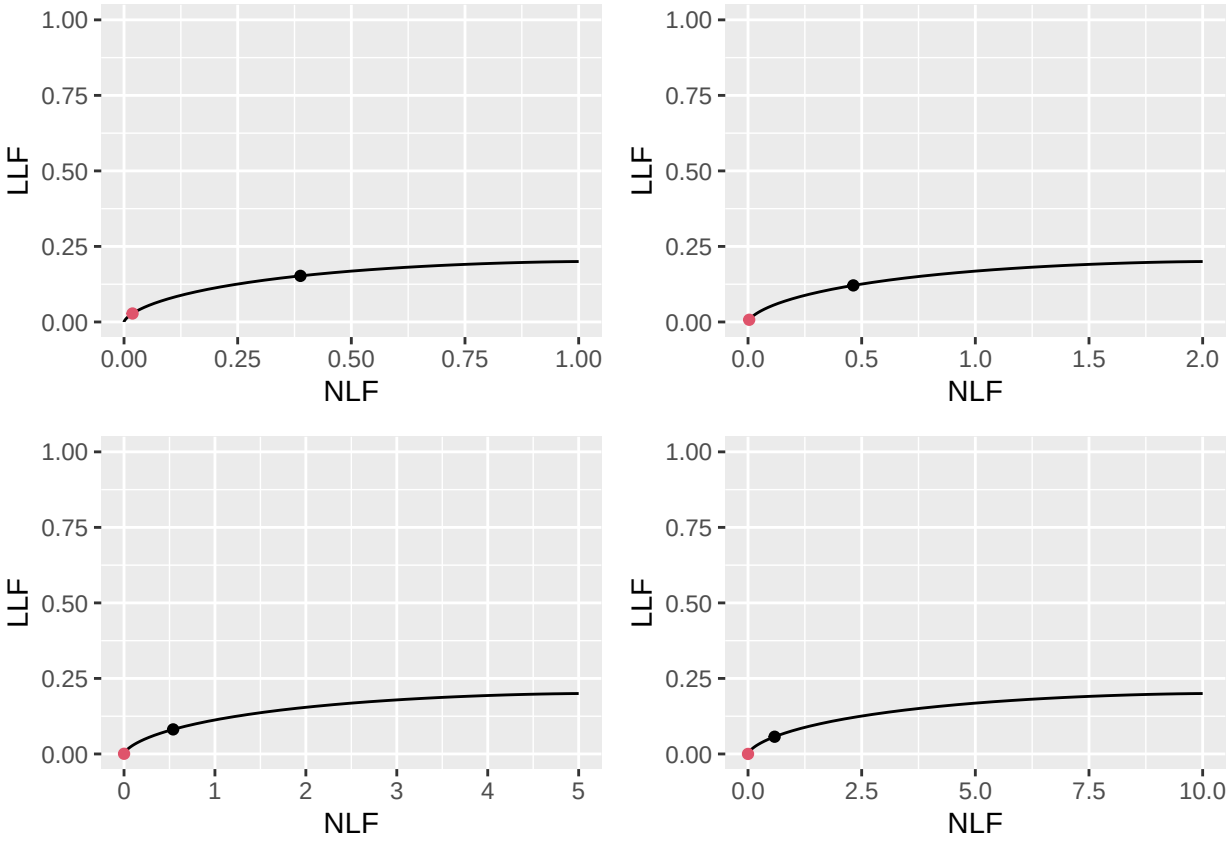


Figure 11.23: Low performance varying λ FROC plots with superimposed operating points. The red dot corresponds to wAFROC$_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

11.10.3.4.2 FROC

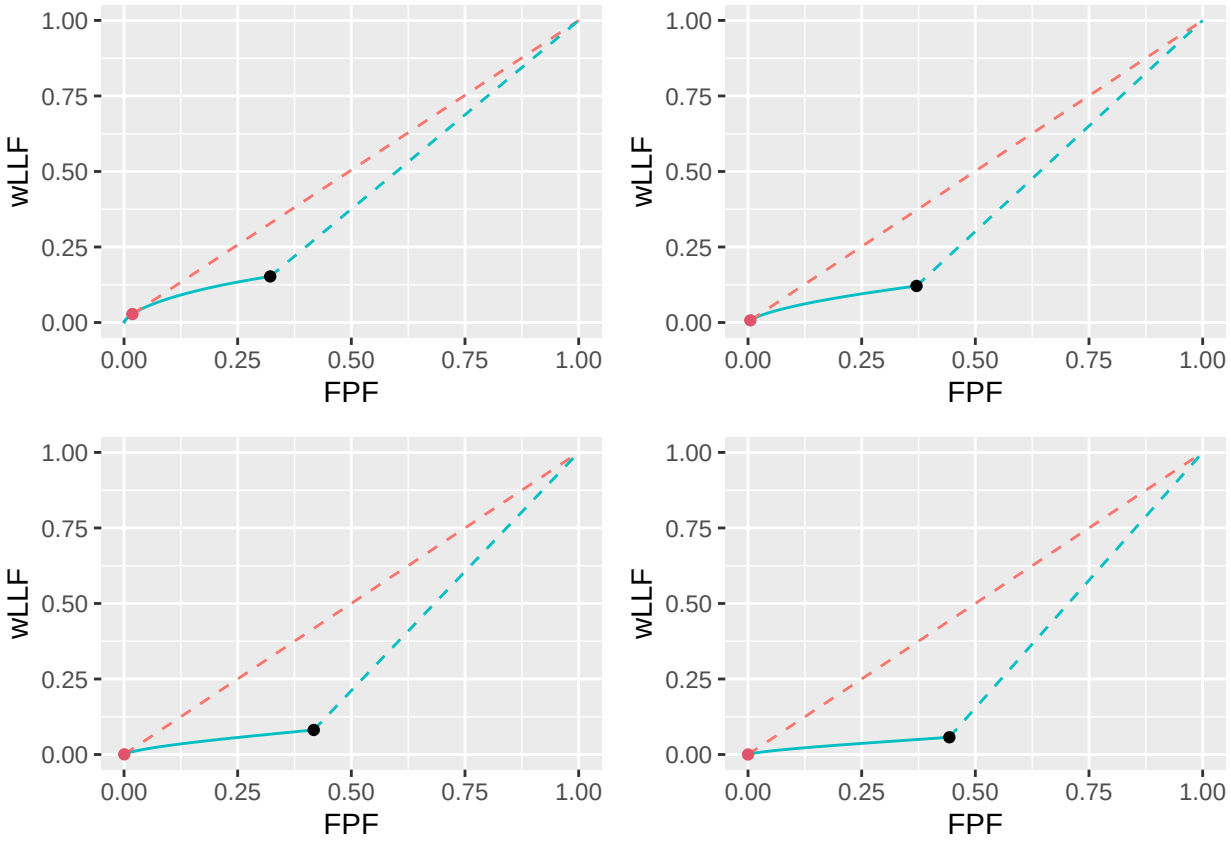


Figure 11.24: Low performance varying λ wAFROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

11.10.3.4.3 wAFROC

11.10.3.4.4 ROC

11.10.3.5 High performance varying ν

```
muArr <- c(4)
lambdaArr <- c(1)
nuArr <- c(0.6, 0.7, 0.8, 0.9)
```

11.10.3.5.1 Summary table

11.10.3.5.2 FROC

11.10.3.5.3 wAFROC

11.10.3.5.4 ROC

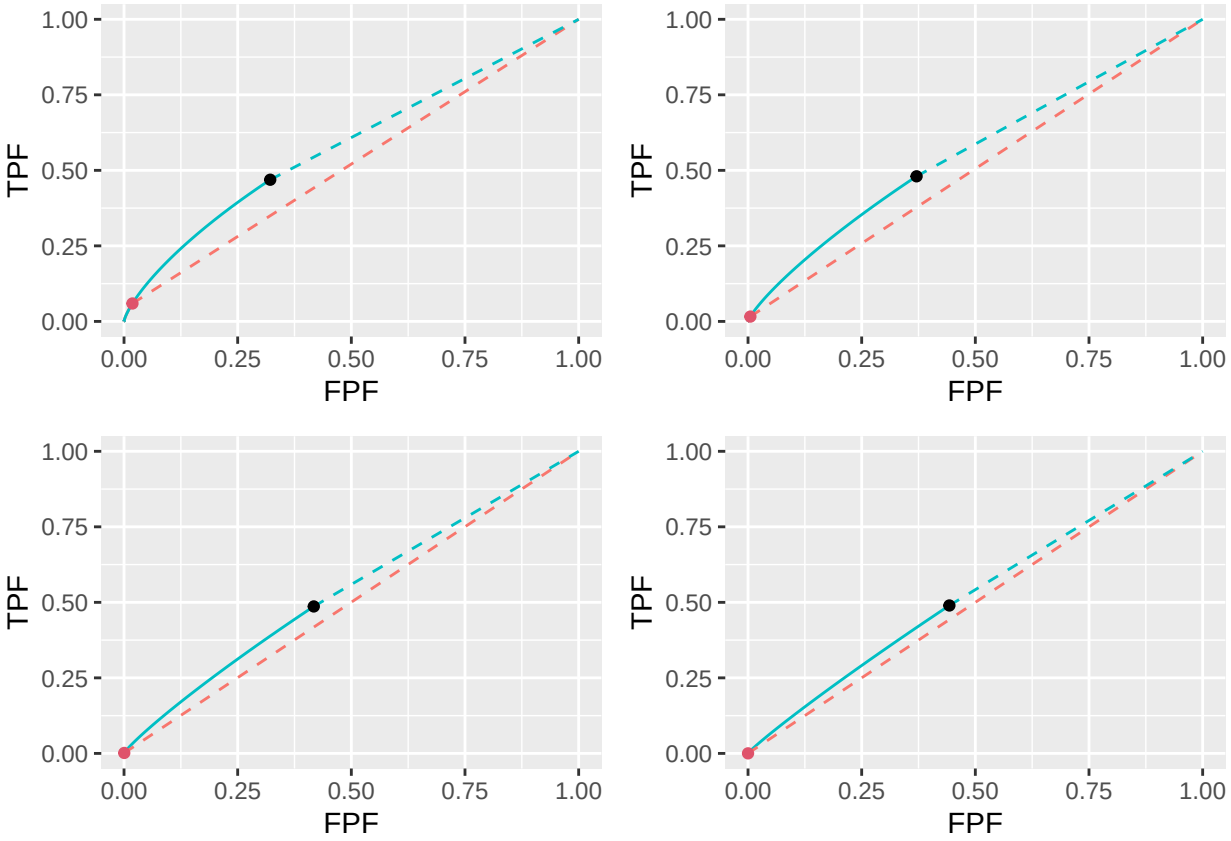


Figure 11.25: Low performance varying λ ROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of λ are: top-left $\lambda = 1$, top-right $\lambda = 2$, bottom-left $\lambda = 5$ and bottom-right $\lambda = 10$.

Table 11.9: Results for $\mu = 4$, $\lambda = 1$ and varying ν. Column labeling as in previous tables.

FOM	ν	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

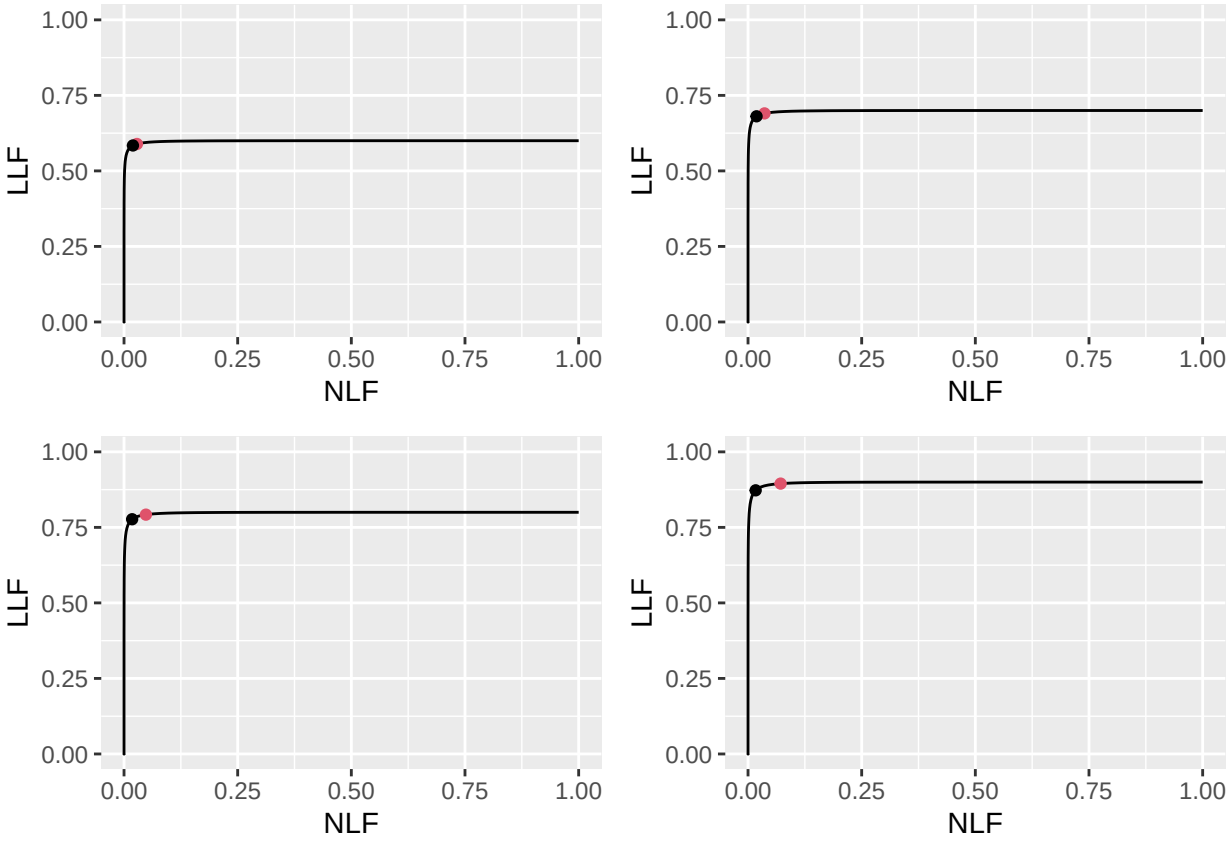


Figure 11.26: High performance varying ν FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$.

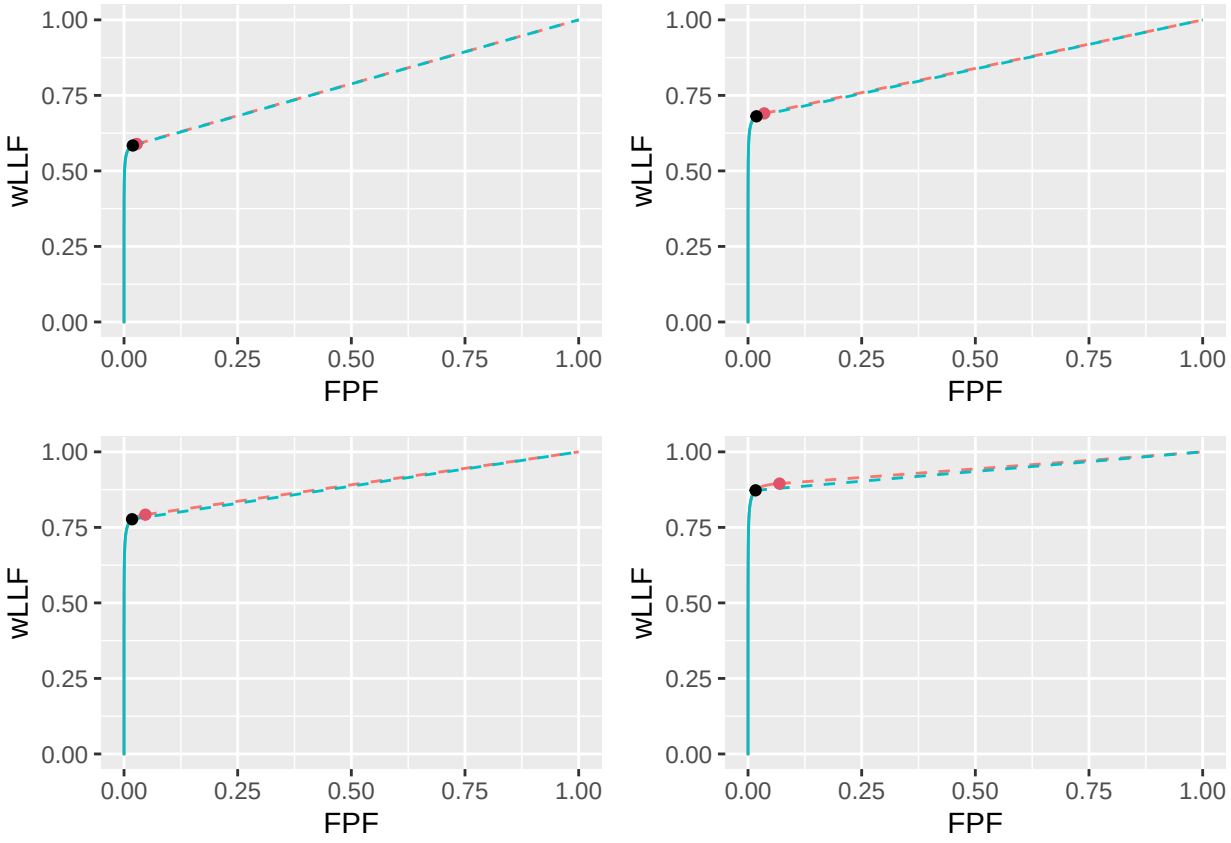


Figure 11.27: High performance varying ν wAFROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$.

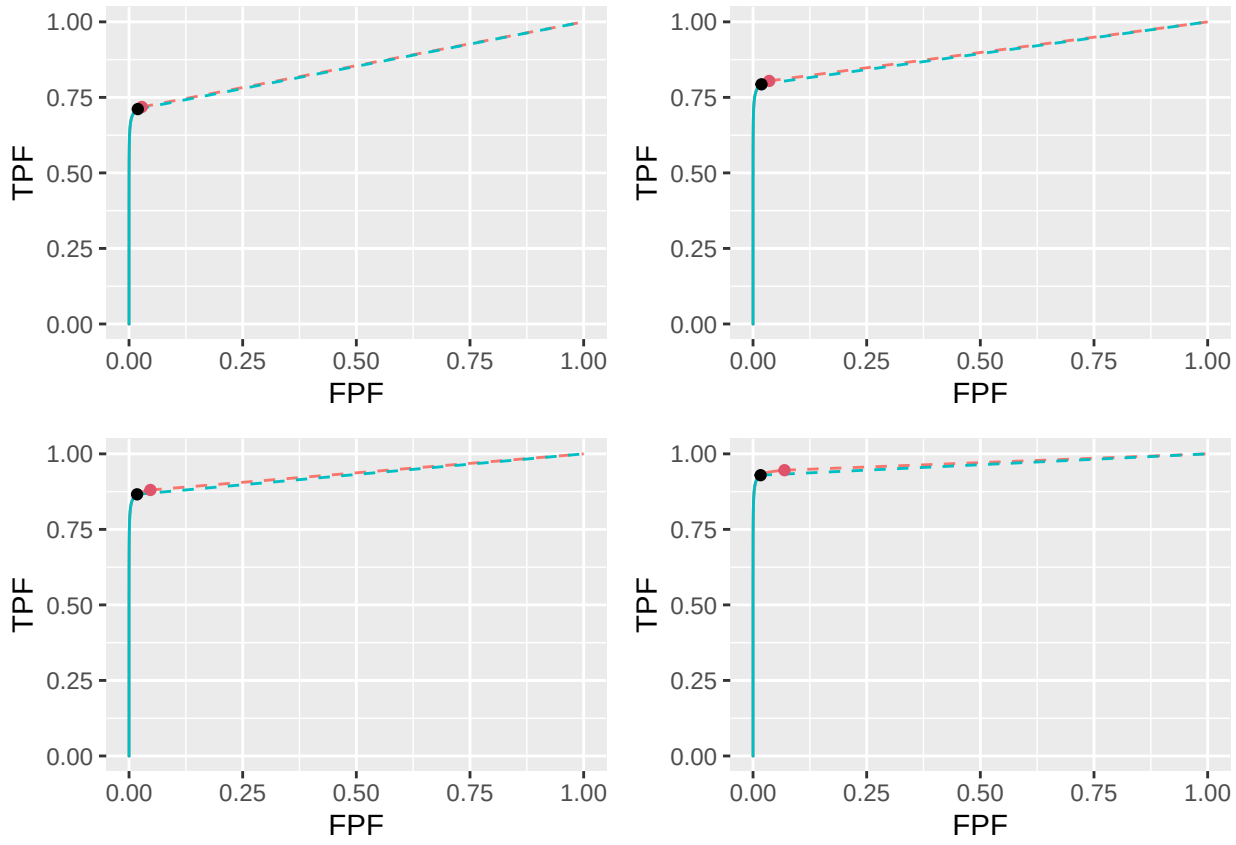


Figure 11.28: High performance varying ν ROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.6$, top-right $\nu = 0.7$, bottom-left $\nu = 0.8$ and bottom-right $\nu = 0.9$.

Table 11.10: Results for $\mu = 1$, $\lambda = 10$ and varying ν. Column labeling as in previous tables.

FOM	ν	ζ_1	$\text{wAFROC}_{\text{AUC}}$	ROC_{AUC}	(NLF, LLF)
$\text{wAFROC}_{\text{AUC}}$	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-index	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)

11.10.3.6 Low performance varying ν

```
muArr <- c(1)
lambdaArr <- c(10)
nuArr <- c(0.1, 0.2, 0.3, 0.4)
```

11.10.3.6.1 Summary table

11.10.3.6.2 FROC

11.10.3.6.3 wAFROC

11.10.3.6.4 ROC

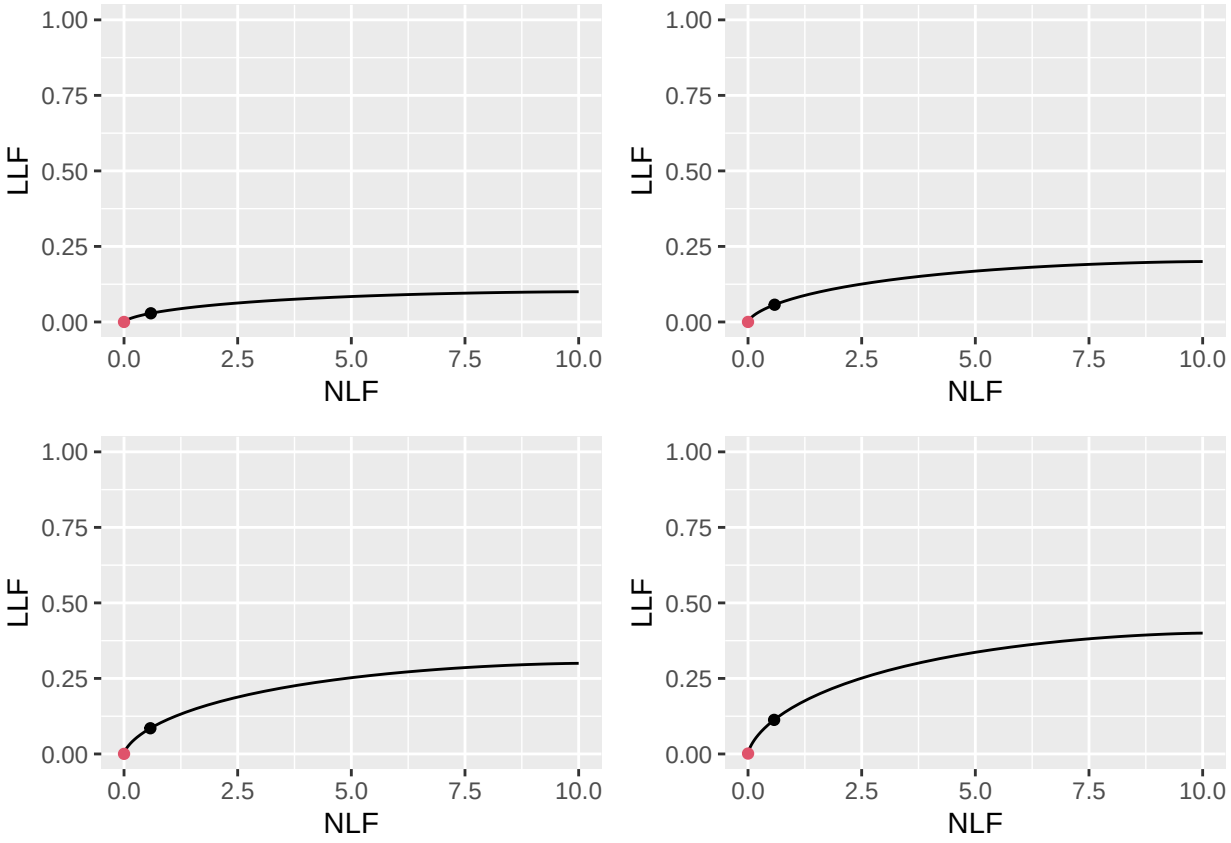


Figure 11.29: Low performance varying ν FROC plots with superimposed operating points. The red dot corresponds to $\text{wAFROC}_{\text{AUC}}$ optimization and the black dot to Youden-index optimization. The values of ν are: top-left $\nu = 0.1$, top-right $\nu = 0.2$, bottom-left $\nu = 0.3$ and bottom-right $\nu = 0.4$.

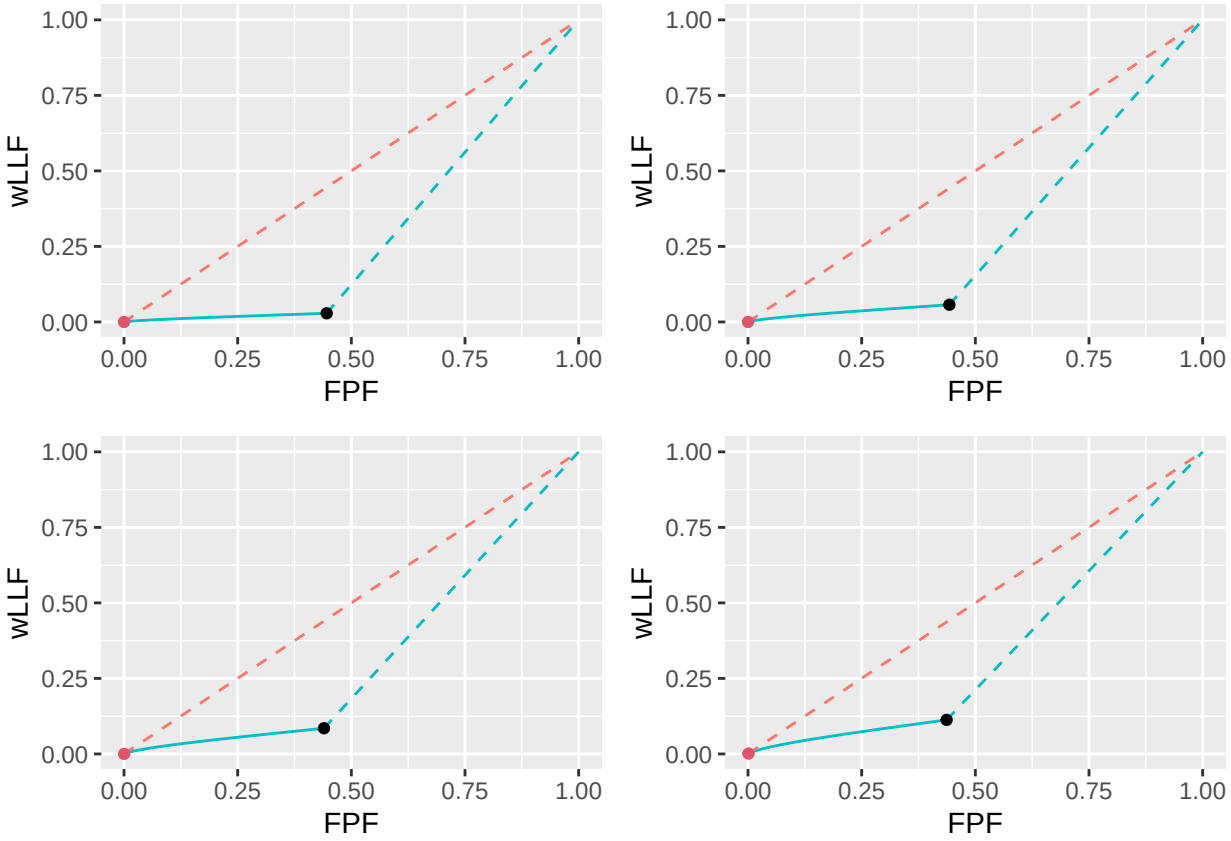


Figure 11.30: Low performance varying ν wAFROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.1$, top-right $\nu = 0.2$, bottom-left $\nu = 0.3$ and bottom-right $\nu = 0.4$.

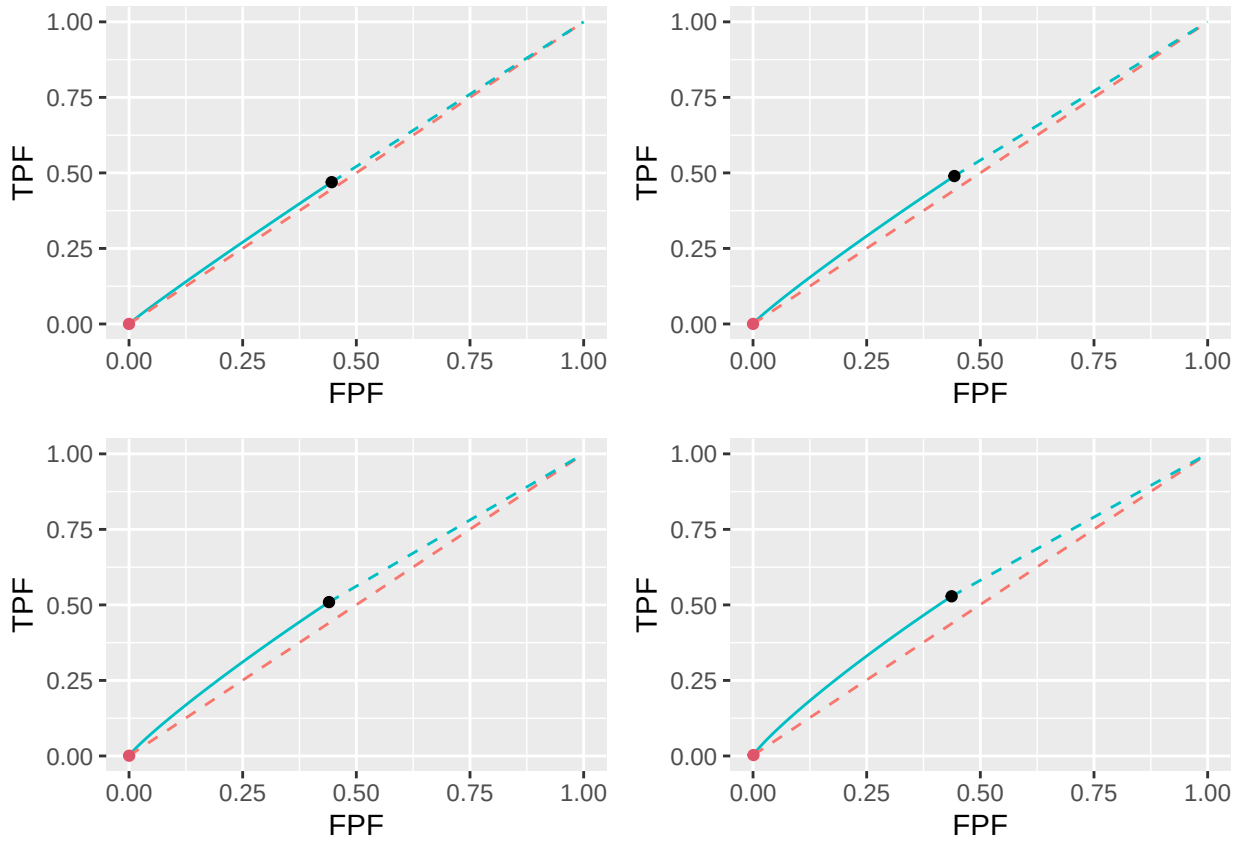


Figure 11.31: Low performance varying ν ROC plots for the two optimization methods with superimposed operating points. The color coding is as in previous figures. The values of ν are: top-left $\nu = 0.1$, top-right $\nu = 0.2$, bottom-left $\nu = 0.3$ and bottom-right $\nu = 0.4$.

Chapter 12

Analyzing a dataset with only diseased cases

12.1 TBA How much finished

0%

12.2 The problem

How to analyze $K_1 = 0$ datasets.

ROC-like plot of TPF vs. FPF1 is possible, see Section 3.12.1. Can create a ROC-like dataset with equal number of "non-diseased" and diseased cases (the ratings of the non-diseased cases are the FP ratings on diseased cases). Fit RSM to this dataset. Proceed as before. Key assumption being violated: the FP ratings on diseased cases are independent of the TP ratings on same cases. However, without this assumption one cannot estimate RSM parameters. Need `RJafroc` function to handle this special case: `FitRsmRoc1`? No! Just need function to create a "ROC" dataset from one that only has diseased cases. e.g., `DfNoNormalsDataset`?

12.2.1 Step 1: Create a test (diseased cases only) dataset

Save TONY dataset to `dsTony`. Create copy `dsNoNormals`. Remove all normal cases from it.

```
dsTony <- RJafroc::dataset01 # TONY dataset
K2 <- length(dsTony$lesions$perCase)
K1 <- length(dsTony$ratings$NL[1,1,,1]) - K2
dsNoNormals <- dsTony
# Remove all normal cases
dsNoNormals$ratings$NL <- dsNoNormals$ratings$NL[,,-(1:K1),]
# And fix truthTableStr
dsNoNormals$descriptions$truthTableStr <-
  dsNoNormals$descriptions$truthTableStr[,,-(1:K1),]
RJafroc::UtilFigureOfMerit(dsTony,FOM = "wAFROC")
#>            rdr1      rdr2      rdr3      rdr4      rdr5
#> trtBT 0.7602704 0.8406191 0.8171524 0.8153090 0.8278324
#> trtDM 0.6425854 0.7049977 0.7518434 0.7724426 0.6836962
#RJafroc::UtilFigureOfMerit(dsNoNormals,FOM = "wAFROC") #this will generate an error
RJafroc::UtilFigureOfMerit(dsTony,FOM = "wAFROC1")
#>            rdr1      rdr2      rdr3      rdr4      rdr5
```

```
#> trtBT 0.8079866 0.8696629 0.8747798 0.8517613 0.8563468
#> trtDM 0.7277103 0.7781506 0.8225630 0.7968418 0.7496963
RJafroc::UtilFigureOfMerit(dsNoNormals,FOM = "wAFROC1")
#>            rdr1      rdr2      rdr3      rdr4      rdr5
#> trtBT 0.8594559 0.9009910 0.9369398 0.8910807 0.8871039
#> trtDM 0.8195304 0.8570572 0.8988448 0.8231600 0.8208875
st <- St(dsTony,FOM = "wAFROC")
st1 <- St(dsNoNormals,FOM = "wAFROC1")
st$RRRC
#> $FTests
#>                 DF          MS    FStat           p
#> Treatment  1.00000 0.025564954 10.29883 0.003668578
#> Error     24.70276 0.002482317       NA          NA
#>
#> $ciDiffTrt
#>              Estimate     StdErr       DF        t       PrGTt    CILower
#> trtBT-trtDM 0.1011236 0.03151074 24.70276 3.209178 0.003668578 0.03618638
#>               CIUpper
#> trtBT-trtDM 0.1660608
#>
#> $ciAvgRdrEachTrt
#>        Estimate     StdErr       DF   CILower   CIUpper         Cov2
#> trtBT 0.8122367 0.02698434 59.28149 0.7582465 0.8662268 0.0005390098
#> trtDM 0.7111131 0.03391021 17.78930 0.6398098 0.7824163 0.0006046324
st1$RRRC
#> $FTests
#>                 DF           MS    FStat           p
#> Treatment   1.0000 0.0065582806 7.957961 0.005193632
#> Error     236.8821 0.0008241157       NA          NA
#>
#> $ciDiffTrt
#>               Estimate     StdErr       DF        t       PrGTt    CILower
#> trtBT-trtDM 0.05121828 0.01815616 236.8821 2.820986 0.005193632 0.01545011
#>                CIUpper
#> trtBT-trtDM 0.08698645
#>
#> $ciAvgRdrEachTrt
#>        Estimate     StdErr       DF   CILower   CIUpper         Cov2
#> trtBT 0.8951143 0.01974550 24.73302 0.8544254 0.9358031 0.0002330913
#> trtDM 0.8438960 0.02497063 27.62144 0.7927144 0.8950776 0.0003862498
```

- `dsNoNormals` is the dataset with no non-diseased cases.
- `st` contains the results of significance testing using the wAFROC-AUC figure of merit for the full dataset.
- `st1` contains the results of significance testing using the wAFROC1-AUC figure of merit for the dataset with no non-diseased cases.

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.

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. (2023). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.1.3.

Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response receiver operating characteristic (froc) data. *Medical physics*, 16(4):561–568.

Chakraborty, D. P. (1997). Computer analysis of mammography phantom images (campi): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics*, 24(8):1269–1277.

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. (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 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. 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. P., Greenberg, G. Z., and Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *The Journal of the Acoustical Society of America*, 33(8):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.

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., 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. 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.

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.

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.

Siddiqui, K. M., Johnson, J. P., Reiner, B. I., and Siegel, E. L. (2005). Discrete cosine transform jpeg compression vs. 2d jpeg2000 compression: Jndmetrix visual discrimination model image quality analysis. In *Medical Imaging 2005: PACS and Imaging Informatics*, volume 5748, pages 202–207. International Society for Optics and Photonics.

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.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

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.