The RJafroc Froc Book

Dev P. Chakraborty, PhD

2021-12-15

Contents

Preface

- It is intended as an online update to my print book (**?**).
- Since its publication in 2017 the `RJafroc` package, on which the `R` code examples in the book depend, has evolved considerably, causing many of the examples to "break".
- This also gives me the opportunity to improve on the book and include additional material.
- The online book is in 3 parts.
- This part is devoted to the FROC paradigm.

TBA How much finished

- HMF approximately 30%
- This book is currently (as of December 2021) in preparation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.

The pdf file of the book

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

The html version of the book

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

A note on the online distribution mechanism of the book

- In the hard-copy version of my book (**?**) the online distribution mechanism was `BitBucket`.

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

Structure of the book

The book is divided into parts as follows:

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

Contributing to this book

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

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

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

- CV
- Alternatives

Chapters needing heavy edits

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

Shelved vs. removed vs. parked folders needing heavy edits

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

Coding aids (for me)

- weird error with knitr not responding to changes in Rmd file: traced to upper case lower case confusion: 13A-froc-empirical1.Rmd which should be 13a-froc-empirical1.Rmd
- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:
- kbl(dfA, caption = "...", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning

message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

FROC paradigm

Chapter 1

The FROC paradigm and search

1.1 TBA How much finished

80%

1.2 Introduction

Until now the focus has been on the receiver operating characteristic (ROC) paradigm. For diagnostic tasks such as detecting diffuse interstitial lung disease[1], or diseases similar to it, where *disease location is implicit*, this is an appropriate paradigm in that essential information is not being lost by limiting the radiologist's response to a single rating categorizing the likelihood of presence of disease.

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

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

(or locations) associated with the suspected disease. Physicians term this *focal disease*, i.e., disease located at specific region(s) of the image.

For focal disease the ROC paradigm constrains the collected information to a single rating representing the confidence level that there is disease *somewhere* in the patient's imaged anatomy. The emphasis on "somewhere" is because it begs the question: if the radiologist believes the disease is somewhere, why not have them to point to it? In fact they do "point to it" in the sense that they record the location(s) of suspect regions in their clinical report, but the ROC paradigm cannot use this information. Clinicians have long recognized problems with ignoring location (**??**). From the observer performance measurement point of view the most important consideration is that neglect of location information leads to loss of statistical power. The basic reason for this is that additional noise is introduced in the measurement due to crediting the reader for correctly detecting the diseased condition but pointing to the wrong location - i.e., *being right for the wrong reason*. One can compensate for reduced statistical power by increasing the numbers of readers and cases, which increases the cost of the study and is also unethical because, by not using the optimal paradigm and analysis, one is subjecting more patients to imaging procedures (**?**).

1.2.1 Chapter outline

Four observer performance paradigms are compared as to the kinds of information collected and ignored. An essential characteristic of the FROC paradigm, namely *visual search*, is introduced. The FROC paradigm and its historical context is described. Key differences between FROC ratings and ROC ratings are noted. The FROC plot is introduced. A "solar" analogy is introduced – understanding this is key to obtaining a good intuitive feel for the FROC paradigm.

1.3 Location specific paradigms

Location-specific paradigms take into account, to varying degrees, information regarding the locations of perceived lesions, so they are sometimes referred to as lesion-specific (or lesion-level) paradigms: usage of these terms is discouraged. For example, all observer performance methods involve detecting the presence of true lesions; so ROC methodology is, in this sense, also lesion-specific. On the other hand *location* is a characteristic of true and perceived focal lesions, and methods that account for location are better termed *location-specific* than lesion-specific.

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

There are three location-specific paradigms:

- the free-response ROC (FROC) (**??**);
- the location ROC (LROC) (**??**);
- the region of interest (ROI) (**?**).

Fig. **??** shows a schematic mammogram interpreted according to current observer performance paradigms. The arrows point to two real lesions and the three light crosses indicate suspicious regions. If a suspicious region is marked it is indicated by a dark cross. Evidently the radiologist saw one of the lesions, missed the other lesion and mistook two normal structures for lesions.

In Fig. **??**, evidently the radiologist found one of the lesions (the light-shaded cross near the left most arrow), missed the other one (pointed to by the second arrow) and mistook two normal structures for lesions (the two light-shaded crosses that are relatively far from any true lesion).

- In the ROC paradigm, Fig. **??** (top-left), the radiologist assigns a single rating indicating the confidence level that there is at least one lesion somewhere in the image. Assuming a 1 – 5 positive directed integer rating scale, if the left-most light-shaded cross is a highly suspicious region then the ROC rating might be 5 (highest confidence for presence of disease).
- In the free-response (FROC) paradigm, Fig. **??** (top-right), the dark-shaded crosses indicate suspicious regions that were *marked* (or *reported* in the clinical report), and the adjacent numbers are the corresponding ratings, which apply to specific suspicious regions in the image, unlike the ROC paradigm, where the rating applies to the whole image. Assuming the allowed FROC ratings are 1 through 4, two marks are shown, one rated FROC-4, which is close to a true lesion, and the other rated FROC-1, which is not close to any true lesion. The third suspicious region, indicated by the light-shaded cross, was not marked, implying its confidence level did not exceed the lowest reporting threshold. The marked region rated FROC-4 (highest FROC confidence) is likely what caused the radiologist to assign the ROC-5 rating to this image in the top-left ROC paradigm figure.
- In the LROC paradigm, Fig. **??** (bottom-left), the radiologist rates the confidence that there is at least one lesion somewhere in the image (as in the ROC paradigm) and marks the most suspicious region in the image. In this example the rating might be LROC-5, the five rating being the same as in the ROC paradigm, and the mark may be the suspicious region rated FROC-4 in the FROC paradigm, and, since it is close to a true lesion, in LROC terminology it would be recorded as a *correct localization*. If the mark were not near a lesion it would be recorded as an *incorrect localization*. Only one mark is allowed in this paradigm, and in fact one

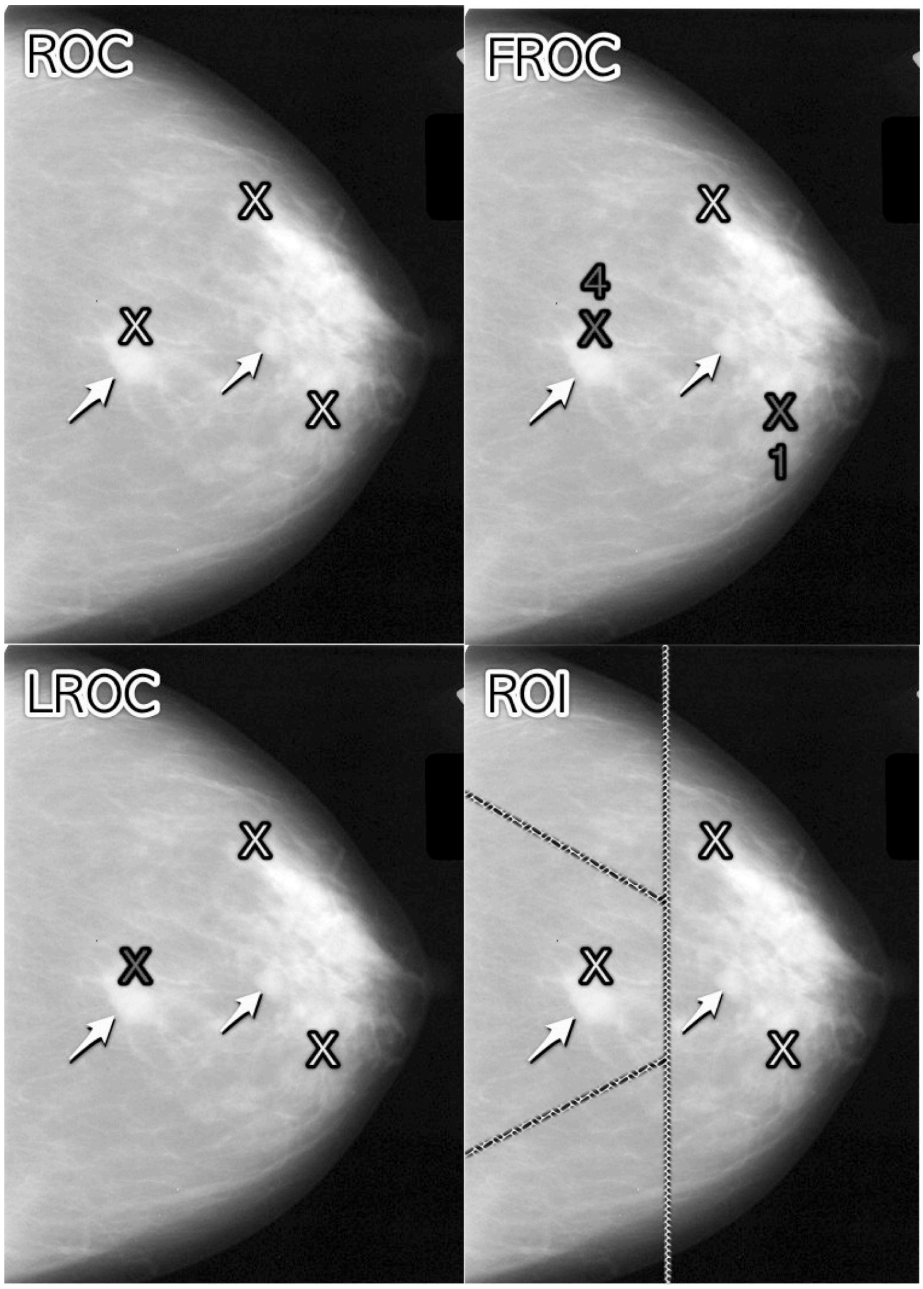


Figure 1.1: Upper Left: ROC, Upper Right: FROC, Lower Left: LROC, Lower Right: ROI

mark is *required* on every image, even if the observer does not find any suspicious regions to report.

- In the region of interest (ROI) paradigm, the researcher segments the image into a number of regions-of-interest (ROIs) and the radiologist rates each ROI for presence of at least one suspicious region somewhere in the ROI. The rating is similar to the ROC rating, except it applies to the ROI, not the whole image. Assuming a 1 – 5 positive directed integer rating scale in Fig. **??** (bottom-right) there are four ROIs. The ROI at ~9 o'clock might be rated ROI-5 as it contains the most suspicious light-shaded cross, the 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 3 (the unmarked lightshaded cross would tend to increase the confidence level) and the one at ~7 o'clock might be rated ROI-1. In the example shown in Fig. **??** (bottom-right), each case yields 4 ratings[2].

1.4 Visual search

The FROC paradigm in medical imaging is equivalent to a visual search task. Any search task has two components: (i) finding something and (ii) acting on it. Examples of a search tasks are looking for lost car-keys or a milk carton in the refrigerator. Success in a search task is finding the searched for object. Acting on it could be driving to work or drinking milk from the carton. There is expertise associated with any search task. Husbands are notoriously bad at finding the milk carton in the refrigerator (analogy due to Dr. Elizabeth Krupinski at an SPIE course taught jointly with the author). Like anything else, search expertise is honed by experience, i.e., lots of practice.

Likewise a medical imaging search task has two components (i) finding suspicious regions and (ii) acting on each finding. "Finding" is the actual term used by clinicians in their reports. Acting on a finding involves determining if the finding is sufficiently suspicious for cancer to warrant reporting. If a suspicious region is found and provided it is sufficiently suspicious the region is marked and rated for confidence that it is a true (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 a-priori that about 5 out of 1000 cases have cancers, so 99.5% of the time odds

[2]When different views of the same patient anatomy (perhaps in different modalities) are available, it is assumed that all images are segmented consistently, and the rating for each ROI takes into account all views of that ROI in the different views (or modalities). The segmentation shown in the figure is a schematic. In fact the ROIs could be clinically driven descriptors of location, such as "apex of lung" or "mediastinum", and the image does not have to have lines showing the ROIs (which would be distracting to the radiologist). The number of ROIs per image can be at the researcher's discretion and there is no requirement that every case have a fixed number of ROIs.

are that the case has no malignant lesions[3]. Considerably search expertise is needed for the radiologist to mark true 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, about 5 per 1000 cases in the US, and assuming expert-level radiologist interpretations, about 90% of breast cases do not generate any marks. About 10% of cases generate one or more marks and are recalled for further comprehensive imaging (termed diagnostic workup). Of marked cases about 90% generate one mark, about 10% generate 2 marks, and a rare case generates 3 or more marks (Dr. David Gur, private communication, ca. 2015).

Conceptually, a mammography report consists of the locations of regions that exceed the threshold and the corresponding levels of suspicion, reported as a Breast Imaging Reporting and Data System (BIRADS) rating (the BIRADS rating is actually assigned after the diagnostic workup following a 0-screening rating; the screening rating itself is binary: 0 for recall or 1 for normal).

1.4.1 Proximity criterion and scoring the data

In the first two clinical applications of the FROC paradigm (**??**) the marks and ratings were indicated by a grease pencil on an acrylic overlay aligned, in a reproducible way, to the CRT displayed chest image. Credit for a correct detection and localization, termed a lesion-localization or LL-event[4], was given only if a mark was sufficiently close (as per proximity criterion, see below) to an actual diseased region; otherwise, the observer's mark-rating pair was scored as a non-lesion localization or NL-event.

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

Definitions:

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

What is meant by sufficiently close? One adopts an acceptance radius (for spherical lesions) or *proximity criterion* (the more general case). What constitutes "close enough" is a clinical decision the answer to which depends on the application. It is not necessary for two radiologists to point to the same pixel in

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

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

order for them to agree that they are seeing the same suspicious region. Likewise, two physicians – e.g., the radiologist finding the lesion on an x-ray and the surgeon responsible for resecting it – do not have to agree on the exact center of a lesion in order to appropriately assess and treat it. More often than not, "clinical common sense" can be used to determine if a mark actually localized the real lesion. When in doubt, the researcher should ask an independent radiologist (i.e., not one used in the observer study) how to score ambiguous marks. A rigid definition of the proximity criterion should not be used.

For roughly spherical nodules a simple rule can be used. If a circular lesion is 10 mm in diameter, one can use the "touching-coins" analogy to determine the criterion for a mark to be classified as lesion localization. Each coin is 10 mm in diameter, so if they touch, their centers are separated by 10 mm, and the rule is to classify any mark within 10 mm of an actual lesion center as a LL mark, and if the separation is greater, the mark is classified as a NL mark. A recent paper (**?**) using FROC analysis gives more details on appropriate proximity criteria in the clinical context. [5]

1.4.2 Multiple marks in the same vicinity

Multiple marks near the same vicinity are rarely encountered with radiologists, especially if the perceived lesion is mass-like. [6] However, algorithmic readers, such as computer aided detection (CAD) algorithms, tend to find multiple regions in the same area. Algorithm designers generally incorporate a clustering step to reduce overlapping regions to a single region and assign the highest rating to it (i.e., the rating of the highest rated mark, not the rating of the closest mark. [7]

1.4.3 Historical context

The term "free-response" was coined by (**?**) to describe a task involving the detection of brief audio tone(s) against a background of white-noise (white-noise is what one hears if an FM tuner is set to an unused frequency). The

[5] Generally the proximity criterion is more stringent for smaller lesions than for larger one. However, for very small lesions allowance is made so that the criterion does not penalize the radiologist for normal marking "jitter". For 3D images the proximity criteria is different in the x-y plane vs. the slice thickness axis.

[6] The exception would be if the perceived lesions were speck-like objects in a mammogram, and even here radiologists tend to broadly outline the region containing perceived specks – they do not mark individual specks with great precision.

[7] The reason for using the highest rating is that this gives full and deserved credit for the localization. Other marks in the same vicinity with lower ratings need to be discarded from the analysis; specifically, they should not be classified as NLs, because each mark has successfully located the true lesion to within the clinically acceptable criterion, i.e., any one of them is a good decision because it would result in a patient recall and point further diagnostics to the true location.

tone(s) could occur at any instant within an active listening interval, defined by an indicator light bulb that is turned on. The listener's task was to respond by pressing a button at the specific instant(s) when a tone(s) was perceived (heard). The listener was uncertain how many true tones could occur in an active listening interval and when they might occur. Therefore, the number of responses (button presses) per active interval was a priori unpredictable: it could be zero, one or more. The Egan et al study did not require the listener to rate each button press, but apart from this difference and with two-dimensional images replacing the listening intervals, the acoustic signal detection study is similar to medical imaging search tasks.

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

The free-response receiver operating characteristic (FROC) plot was introduced (**?**) 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. If the points are connected by straight lines the resulting "curve" is termed the *empirical FROC curve*.
- The upper-right most operating point is termed the *observed end-point* and its coordinated are denoted $(\mathrm{NLF}_{\max}, \mathrm{LLF}_{\max})$.
- Unlike the ROC plot which is completely contained in the unit square, the FROC plot is not.

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

If *integer ratings* are used for each recorded mark then in a four-rating FROC study at most 4 FROC operating points will result: one corresponding to marks rated 4s; another corresponding to marks rated 4s or 3s; another to the 4s, 3s,

or 2s; and finally the 4s, 3s, 2s, or 1s. An R-rating study yields at most R operating points [8].

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

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

1.6 The "solar" analogy

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

FROC curve implications of this analogy are:

- Each 24-hour day corresponds to two "trials" in the (**?**) 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.

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

- Most important, $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$.

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

The conclusion is that in a task where a target is known to be present in the field of view and its location is known, the observer will always reach $\text{LLF}_{\text{max}} = 1$ and $\text{NLF}_{\text{max}} = 0$. Why are LLF and NLF subscripted "max"? By randomly choosing to not mark the position of the sun even though it is visible, for example, using a coin toss to decide whether or not to mark the sun, the observer can "walk down" the y-axis of the FROC plot, eventually reaching $LLF = 0$ and $NLF = 0$. The reason for allowing the observer to "walk down" the vertical is simply to demonstrate that a continuous FROC curve from the origin to (0,1) can, in fact, be realized.

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

The implication of this is that a fundamentally different mechanism from that considered in conventional observer performance methodology, namely *search*, is at work.

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

Think of the eye as two cameras: a low-resolution camera (peripheral vision) with a wide field-of-view plus a high-resolution camera (foveal vision) with a narrow field-of-view If one were limited to viewing with the high-resolution camera one would spend so much time steering the high-resolution narrow field-of-view camera from spot-to-spot that one would have a hard time finding the desired stellar object. Having a single high-resolution narrow field of view vision would also have negative evolutionary consequences as one would spend so much time scanning and processing the surroundings with the narrow field of view vision that one would miss dangers or opportunities. Nature has equipped us with

essentially two cameras; the first low-resolution camera is able to "digest" large areas of the surround and process it rapidly so that if danger (or opportunity) is sensed, then the eye-brain system rapidly steers the second high-resolution camera to the location of the danger (or opportunity). This is Nature's way of optimally using the eye-brain system. For a similar reason astronomical telescopes come with a wide field of view lower magnification "spotter scope".

When cloud cover completely blocks the fictitious random-position sun there is no stimulus to trigger the peripheral vision system to guide the fovea to the correct location. Lacking any stimulus, the observer is reduced to guessing and is led to different conclusions depending upon the benefits and costs involved. If, for example, the guessing observer earns a dollar for each LL and is fined a dollar for each NL, then the observer will likely not make any marks as the chance of winning a dollar is much smaller than losing many dollars. For this observer $\mathrm{LLF}_{\max} = 0$ and $\mathrm{NLF}_{\max} = 0$, and the operating point is "stuck" at the origin. If, on the other hand, the observer is told every LL is worth a dollar and there is no penalty to NLs, then with no risk of losing the observer will "fill up" the sky with false marks.

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

False sightings of objects in astronomy do occur. Finding a new astronomical object is a search task, where, as always, one can have two outcomes, correct localization (LL) or incorrect localizations (NLs). At the time of writing there is a hunt for a new planet, possibly a gas giant , that is much further than even the newly demoted Pluto.

1.7 Discussion

This chapter has served as a general introduction to the location specific paradigms. The FROC paradigm, in particular, is directly related to visual search. The terms lesion-localization (LL) and non-lesion localization (NL) were introduced, to mean locating a true lesion and a false region, respectively. A qualitative definition of search performance was given, namely, the ability to find lesions while not finding non-lesions. This will be quantified in a later chapter. A widely used operating characteristic associated with it, namely the FROC curve, was introduced. A "solar" analogy was given that brings out important characteristics of the FROC - in particular its dependence on lesion contrast.

In my experience the FROC paradigm is much misunderstood. Some of this has to do with loose terminology and some to misconceptions regarding the paradigm and the FROC curve. These are summarized below:

- Loose terminology: using the term "lesion-specific" to describe location-specific paradigms.
- Loose terminology: using the term "lesion" when one means a "suspicious region" that may or may not be a true lesion.
- 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 localization, that apply to localized regions of the image.
- Loose terminology: using the FROC-1 rating to mean in effect "I see no signs of disease in this image", when in fact it should be used as the lowest level of a reportable suspicious region. The former usage amounts to wasting a confidence level.
- Misconception: showing FROC curves as reaching the unit ordinate, as this is the exception rather than the rule.
- Misconception: believing that FROC curves extend to very large values along the abscissa and all the observer has to do to access this region is to lower their reporting threshold.
- Misconception: blaming the FROC paradigm for alleged arbitrariness of the proximity criterion and multiple marks in the same region. These are not clinically important.

The FROC plot is the first proposed way of visually summarizing FROC data. The next chapter deals with all empirical operating characteristics that can be defined from an FROC dataset and associated scalar measures of performance, termed *figures of merit* (FOMs).

1.8 References

Chapter 2

Empirical plots from FROC data

2.1 TBA How much finished

80%

2.2 TBA Introduction

FROC data is defined as consisting of mark-rating pairs. A distinction between latent and actual marks is made followed by a summary of FROC notation applicable to a single modality single reader dataset. This is a key table, which will be used in later chapters.

Section **??** defined the empirical FROC plot. This chapter introduces mathematical expressions for empirical operating characteristics for this and other plots possible with FROC data.

Operating characteristics are visual depicters of performance. Scalar quantities, typically area measures, derived from operating characteristics can serve as quantitative measures of performance, termed *figures of merit* (FOMs). This chapter defines the AUC associated with each introduced operating characteristic.

The observed end-point of an operating characteristic is defined. For the FROC plot it is demonstrated that the observed FROC curve is not contained in the unit square unlike the other operating characteristics which are contained in the unit square. In contrast the other introduced operating characteristics are each contained within the unit square.

An FROC dataset is used to illustrate the different operating characteristics. The correlation between the AUCs vs. ROC AUC is qualitatively examined via plots.

2.3 Mark rating pairs

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

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

2.3.1 Latent vs. actual marks

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

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

2.3.2 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

$r = 0, 1, ..., R_{ROC} - 1$ and R_{ROC} is the number of ROC study bins [1] and a binning rule that if $\zeta_r \leq z < \zeta_{r+1}$ the case is rated $r + 1$. Dummy cutoffs are defined as $\zeta_0 = -\infty$ and $\zeta_{R_{ROC}} = \infty$. The z-sample applies to the whole case. To summarize:

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

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

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

2.4 Notation

Clear notation is vital to understanding this paradigm. The notation needs to account for case and location dependencies of ratings and the distinction between case-level and location-level ground truth. The notation also has to account for cases with no marks.

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

2.4.1 Comments on Table ??

- 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

[1] The subscript is used to make explicit the paradigm used.

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

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

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 and 2 for diseased.
- 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 $k_t t$.
- 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 $z_{k_t t l_1 1}$. Latent NL marks are possible on non-diseased and diseased cases (both values of t are allowed). The range of a z-sample is $-\infty < z_{k_t t l_1 1} < \infty$, provided $l_1 \neq \emptyset$; otherwise, it is an unobservable event.
- Row 9: The z-sample of a latent LL is $z_{k_2 2 l_2 2}$. Unmarked lesions are observable events and are therefore assigned negative infinity ratings (the null-set notation is unnecessary for them).
- Row 10: R_{FROC} is the number of bins in the FROC study.

- Rows 11, 12 and 13: The cutoffs in the FROC study. The lowest threshold is ζ_1. The other non-dummy thresholds are ζ_r where $r = 2, 3, ..., R_{FROC}$. The dummy thresholds are $\zeta_0 = -\infty$ and $\zeta_{R_{FROC}+1} = \infty$.
- Row 14: $W_{k_2 l_2}$ is the weight (i.e., clinical importance) of lesion $l_2 2$ in diseased case $k_2 2$. The weights of lesions in a case sum to unity: $\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1$.
- Row 15: L_{max} is the maximum number of lesions per case in the dataset.
- Row 16: L_T is the total number of lesions in the dataset.

2.4.2 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 [2]. In fact the first model of FROC data (**?**) predicts that, as the reporting threshold is lowered to $\zeta_1 = -\infty$, the number of NL marks per case approaches ∞ as does NLF_{max}. However, observed FROC curves end at a finite value of NLF_{max}. This is one reason I introduced the radiological search model (RSM) (**?**). I will have much more to say about this in Chapter **??**, 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.

With this background, let us return to the conceptual issue: why does the observer not keep "filling-up" the image with NL marks? The answer is that *the observer can only mark regions that have a non-zero chance of being a lesion.* For example, if the actual number of latent NLs on a particular case is 2, then, as the reporting threshold is lowered, the observer will make at most two NL marks. Having exhausted these two regions the observer will not mark any more regions because there are no more regions to be marked - *all other regions in the image have, in the perception of the observer, zero chance of being a lesion.*

The notational issue is how to handle images with no latent NL marks. Basically it involves restricting summations over cases $k_t t$ to those cases which have at least one latent NL mark, i.e., $N_{k_t t} \neq 0$, 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, and otherwise $N_{k_t t} = 0$ and $l_1 = \emptyset$, the null

[2] 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.

set. The possible values of l_1 are $l_1 = \{\emptyset\} \oplus \left\{1, 2, ...N_{k_t t}\right\}$. The null set applies when the case has no latent NL marks and $\oplus$ is the "exclusive-or" symbol ("exclusive-or" is used in the English sense: "one or the other, but not neither nor both"). In other words, l_1 can *either* be the null set or take on values $1, 2, ...N_{k_t t}$.

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

2.5 The empirical FROC

The FROC, Chapter **??**, is the plot of LLF (along the ordinate) vs. NLF (along the abscissa).

Using the notation of Table **??** and assuming binned data[3], then, corresponding to the operating point determined by threshold ζ_r, the FROC abscissa is $\text{NLF}_r \equiv \text{NLF}(\zeta_r)$, the total number of NLs rated $\geq$ threshold ζ_r divided by the total number of cases, and the corresponding ordinate is $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$, the total number of LLs rated $\geq$ threshold ζ_r divided by the total number of lesions:

$$\text{NLF}_r = \frac{n(\text{NLs rated} \geq \zeta_r)}{n(\text{cases})} \tag{2.3}$$

and

$$\text{LLF}_r = \frac{n(\text{LLs rated} \geq \zeta_r)}{n(\text{lesions})} \tag{2.4}$$

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

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

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

Equations (**??**) and (**??**) is are equivalent to:

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

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

and

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

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

In Eqn. (**??**) $\mathbb{I}\left(N_{k_t t} \neq 0\right)$ ensures that *only cases with at least one latent NL* are counted. Recall that $N_{k_t t}$ is the total number of latent NLs in case $k_t t$. The term $\mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right)$ counts over all NL marks with ratings $\geq \zeta_r$. The three summations yield the total number of NLs in the dataset with z-samples $\geq \zeta_r$ and dividing by the total number of cases yields NLF_r. This equation also shows explicitly that NLs on both non-diseased ($t = 1$) and diseased ($t = 2$) cases contribute to NLF.

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

2.5.1 Definition empirical plot and AUC

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}.

2.5.2 The origin, a trivial point

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

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

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

$$\mathrm{NLF}_1 = \frac{1}{K_1+K_2}\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \tag{2.8}$$

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

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

$$\left.\begin{aligned}\mathrm{LLF}_1 =& \frac{\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_1\right)}{L_T} \\ \leq& 1\end{aligned}\right\} \tag{2.9}$$

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

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

2.5.4 Futility of extrapolation outside the observed end-point

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

$$\mathrm{NLF}_0 = \frac{1}{K_1+K_2}\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq -\infty\right) \tag{2.10}$$

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

$$\left.\begin{aligned}\text{2nd term} =& \left(\frac{1}{K_1+K_2}\right)\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(-\infty \leq z_{k_t t l_1 1} < \zeta_1\right) \\ =& \frac{\text{unknown number}}{K_1+K_2}\end{aligned}\right\} \tag{2.11}$$

It represents the contribution of unmarked NLs, i.e., latent NLs whose z-samples were below ζ_1. It determines how much further to the right the observer's NLF

would have moved, relative to NLF_1, if one could get the observer to lower the reporting criterion to $-\infty$. *Since the observer may not oblige, this term cannot, in general, be evaluated.* Therefore NLF_0 cannot be evaluated. The basic problem is that *unmarked latent NLs represent unobservable events.*

Turning our attention to LLF_0:

$$\left.\begin{aligned} \mathrm{LLF}_0 =& \frac{\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq -\infty\right)}{L_T} \\ =& 1 \end{aligned}\right\} \tag{2.12}$$

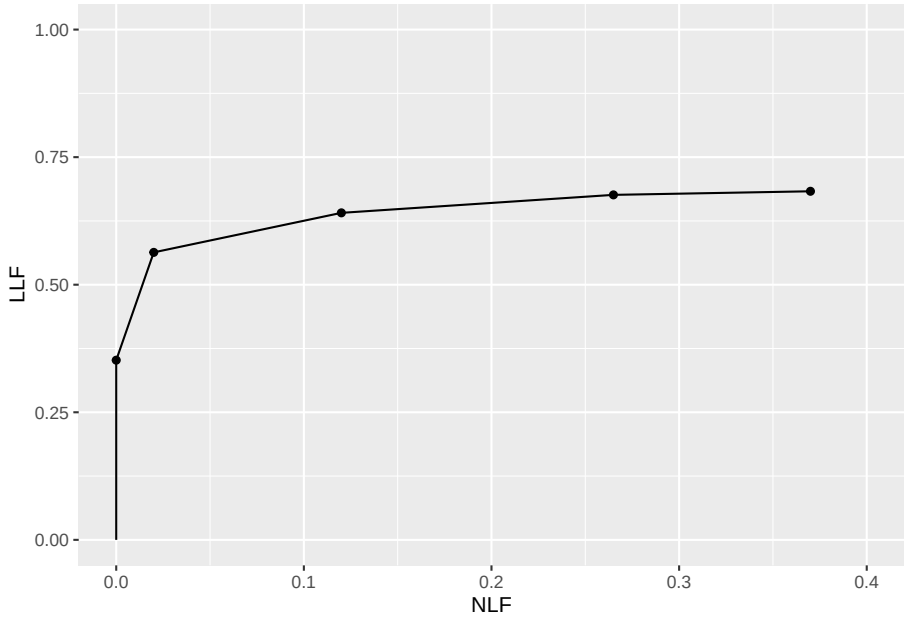
Unlike unmarked latent NLs, **unmarked lesions can safely be assigned the $-\infty$ rating, because an unmarked lesion is an observable event*. The right hand side of Eqn. (**??**) evaluates to unity. However, since the corresponding abscissa NLF_0 is undefined, one cannot plot this point. It follows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

The formalism should not obscure the fact that the futility of extrapolation outside the observed end-point of the FROC is a fairly obvious property: one does not know how far to the right the abscissa of the observed end-point might extend if one could get the observer to report every latent NL.

2.5.5 Illustration with a dataset

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

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



The study in question was a 5 rating FROC study. The lowest non-trivial point corresponds to the marks rating 5, the next higher one corresponds to marks rated 4 or 5, etc. The FROC plots vary widely but share the common characteristic that the operating points cannot move downward-left as one cumulates lower confidence level marks.

Shown next is calculation of the figure of merit for this dataset. All 20 values are shown. The value for `trt1` and `rdr1` is the area under the FROC plot shown above.

```
UtilFigureOfMerit(dataset04, FOM = "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
```

2.6 The inferred ROC plot

By adopting a rational rule for converting the zero or more mark-rating data per case to a single rating per case, and commonly the highest rating rule is

used [4], it is possible to infer ROC data from FROC mark-rating data.

2.6.1 Inferred-ROC rating

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

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

Definitions:

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

Definition of ROC plot:

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

The mathematical definition of the ROC follows.

2.6.2 Inferred FPF

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

$$\left.\begin{aligned} FP_{k_1 1} &= \max_{l1}\left(z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset\right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{2.13}$$

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

[4] The highest rating method was used in early FROC modeling in (**?**) and in (**?**), the latter in the context of LROC paradigm modeling.

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:

$$\text{FPF}_r \equiv \text{FPF}\left(\zeta_r\right) = \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_r\right) \tag{2.14}$$

2.6.3 Inferred TPF

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

$$TP_{k_2 2} = \max_{l_1 l_2}\left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2} \mid l_1 \neq \emptyset\right) \tag{2.15}$$

or

$$TP_{k_2 2} = \max_{l_2}\left(z_{k_2 2 l_2 2} \mid \left(l_1 = \emptyset \wedge \left(\max_{l_2}\left(z_{k_2 2 l_2 2}\right) \neq -\infty\right)\right)\right) \tag{2.16}$$

or

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

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

The inferred true positive fraction TPF_r is defined by:

$$\text{TPF}_r \equiv \text{TPF}(\zeta_r) = \frac{1}{K_2}\sum_{k_2=1}^{K_2} \mathbb{I}\left(TP_{k_2 2} \geq \zeta_r\right) \tag{2.18}$$

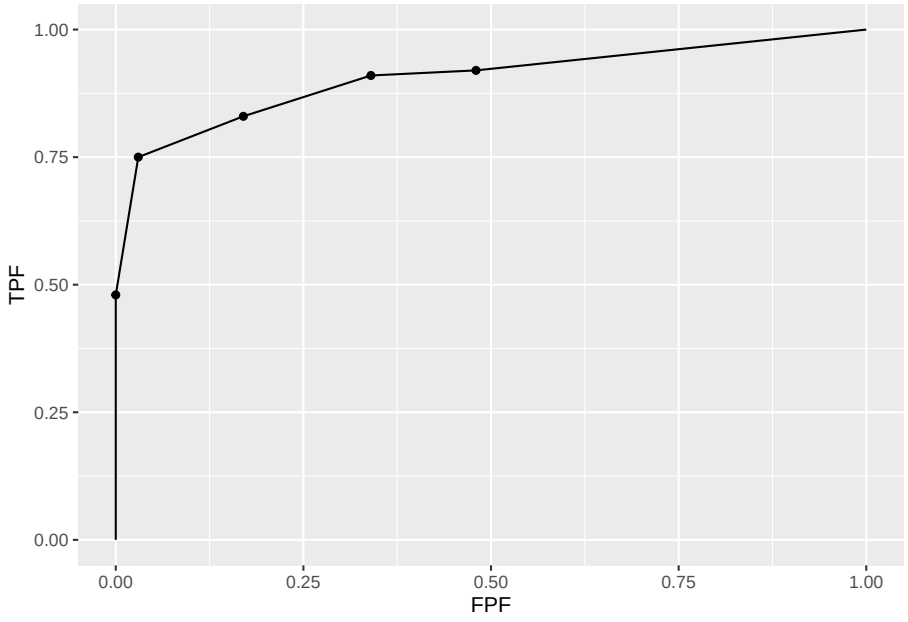
2.6.4 Definition empirical plot and AUC

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

2.6.5 Illustration with a dataset

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

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

2.7 The alternative FROC (AFROC) plot

- Fig. 4 in (**?**) anticipated another way of visualizing FROC data. I subsequently termed[5] this the *alternative FROC (AFROC)* plot (**?**).
- The empirical AFROC is defined as the plot of $\text{LLF}(\zeta_r)$ along the ordinate vs. $\text{FPF}(\zeta_r)$ along the abscissa.
- $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$ was defined in Eqn. (**??**).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$ was defined in Eqn. (**??**).

2.7.1 Definition empirical 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 inferred AFROC AUC, denoted A_{AFROC}.

Key points:

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

2.7.2 The constrained observed end-point of the AFROC

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

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

$$\left.\begin{aligned} \text{FPF}_{R_{FROC}+1} =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \infty\right) \\ =& 0 \\ \text{LLF}_{R_{FROC}+1} =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq \infty\right) \\ =& 0 \end{aligned}\right\} \tag{2.19}$$

and

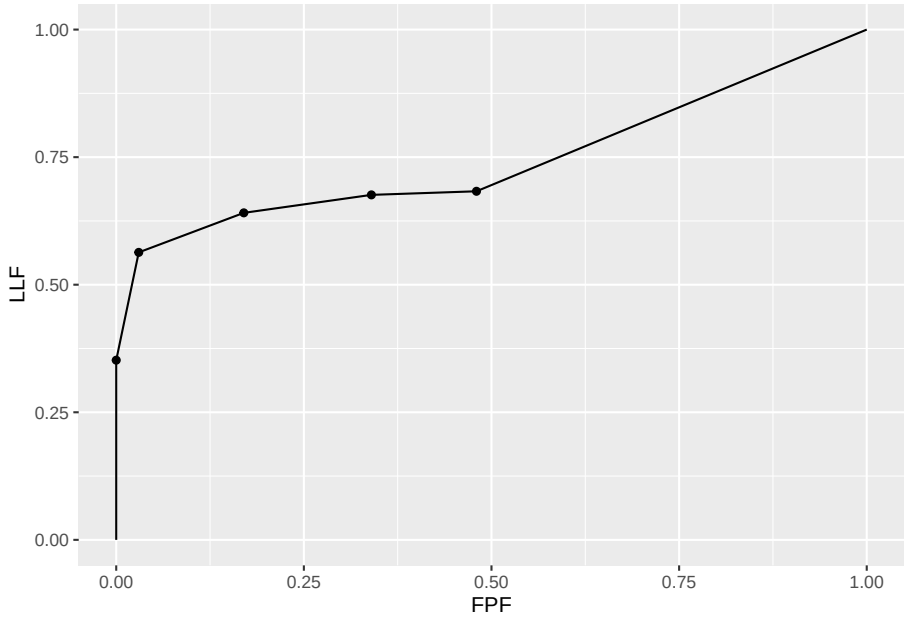
$$\left.\begin{aligned} \text{FPF}_0 =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq -\infty\right) \\ =& 1 \\ \text{LLF}_0 =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq -\infty\right) \\ =& 1 \end{aligned}\right\} \tag{2.20}$$

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

2.7.3 Illustration with a dataset

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

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



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

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

2.8 The weighted-AFROC (wAFROC) plot

The AFROC ordinate defined in Eqn. (**??**) gives equal importance to every lesion in a case. Therefore, a case with more lesions will have more influence on the AFROC (see TBA Chapter 14 for an explicit demonstration of this fact). This is undesirable since each case (i.e., patient) should get equal importance in the analysis – as with ROC analysis, one wishes to draw conclusions about the population of cases and each case is regarded as an equally valid sample from the population. In particular, one does not want the analysis to be skewed towards cases with greater than the average number of lesions. [6]

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

Another issue is that the AFROC assigns equal *clinical* importance to each lesion in a case. Lesion weights were introduced (**?**) to allow for the possibility that the clinical importance of finding a lesion might be lesion-dependent (**?**). 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 (**?**).

Let $W_{k_2 l_2} \geq 0$ denote the *weight* (i.e., short for clinical importance) of lesion l_2 in diseased case k_2 (since weights are only applicable to diseased cases one can, without ambiguity, drop the case-level and location-level truth subscripts, i.e., the notation $W_{k_2 2 l_2 2}$ would be superfluous). For each diseased case $k_2 2$ the weights are subject to the constraint:

$$\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1 \tag{2.21}$$

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

The weighted lesion localization fraction wLLF_r is defined by (**?**):

$$\text{wLLF}_r \equiv \text{wLLF}\left(\zeta_r\right) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 l_2 2} \geq \zeta_r\right) \tag{2.22}$$

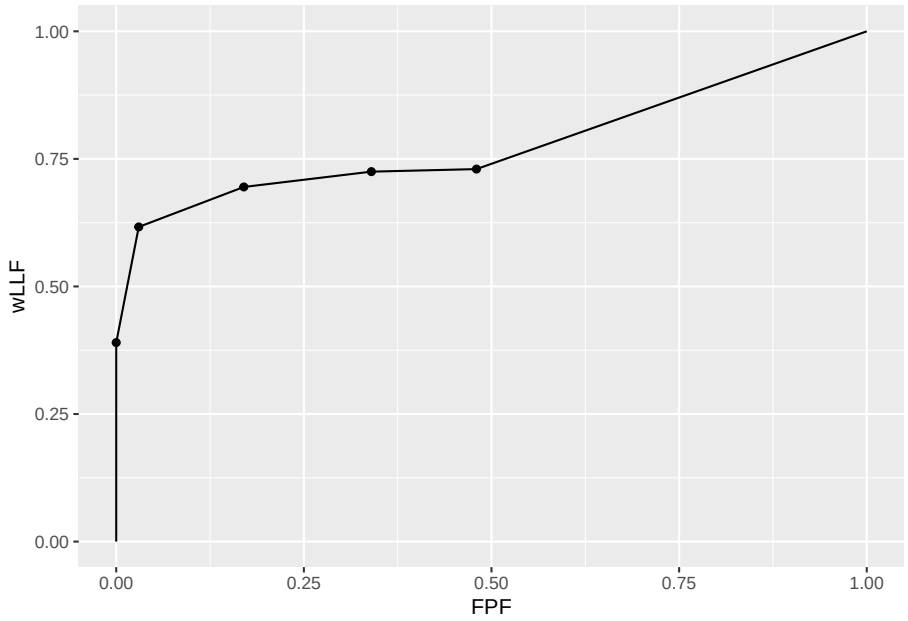
2.8.1 Definition empirical plot and AUC

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

2.8.2 Illustration with a dataset

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

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



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

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

2.9 The AFROC1 plot

Historically the AFROC originally used a different definition of FPF, which is retrospectively termed the AFROC1 plot. Since NLs can occur on diseased cases, it is possible to define an inferred "FP" rating on a *diseased case* as the maximum of all NL ratings on the case, or $-\infty$ if the case has no NLs. The quotes emphasize that this is non-standard usage of ROC terminology: in an ROC study, a FP can only occur on a *non-diseased case*. Since both case-level

truth states are allowed, the highest false positive (FP) z-sample for case $k_t t$ is [the "1" superscript below is necessary to distinguish it from Eqn. (**??**)]:

$$\left.\begin{aligned} FP^1_{k_t t} &= \max_{l_1}\left(z_{k_t t l_1 1} \mid l_1 \neq \emptyset\right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{2.23}$$

$FP^1_{k_t t}$ 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 [the "1" superscript is necessary to distinguish it from Eqn. (**??**)]:

$$FPF^1_r \equiv FPF^1_r(\zeta_r) = \frac{1}{K_1 + K_2}\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(FP^1_{k_t t} \geq \zeta_r\right) \tag{2.24}$$

Note the subtle differences between Eqn. (**??**) and Eqn. (**??**). The latter counts "FPs" on non-diseased and diseased cases while Eqn. (**??**) counts FPs on non-diseased cases only, and for that reason the denominators in the two equations are different. The advisability of allowing a diseased case to be both a TP and a FP is questionable from both clinical and statistical considerations. However, this operating characteristic can be useful in applications where all or almost all cases are diseased.

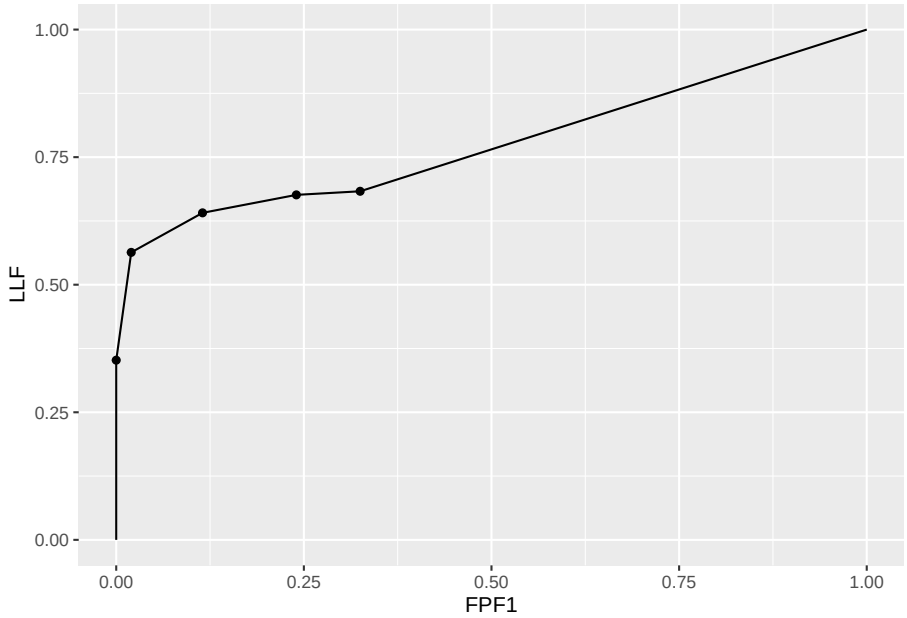
2.9.1 Definition empirical plot and AUC

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

2.9.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "AFROC1")$Plot` with different treatments and readers specified.

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



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

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

2.10 The weighted-AFROC1 (wAFROC1) plot

2.10.1 Definition empirical plot and AUC

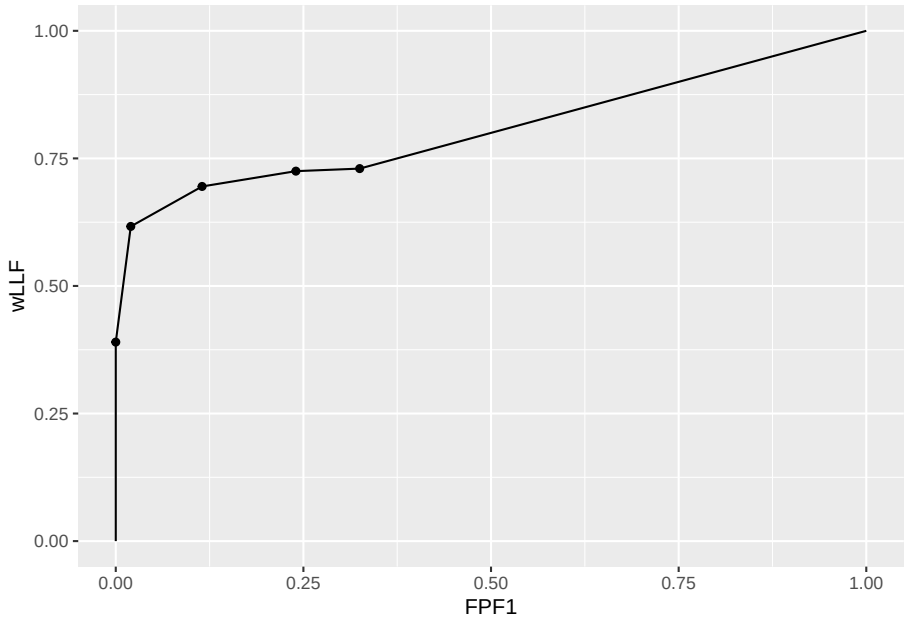
The empirical weighted-AFROC1 (wAFROC1) plot connects adjacent operating points (FPF_r^1, wLLF_r), including the origin (0,0) and (1,1), with straight lines. The only difference between it and the wAFROC plot is in the x-axis. The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC1}.

2.10.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical wAFROC plot1.
The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04,`

`trts = 1, rdrs = 1, opChType = wAFROC1")$Plot` with different treatments and readers specified.

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



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

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

2.11 Plots of FROC, AFROC and wAFROC AUC vs. ROC AUC

Plots of A_{FROC}, A_{AFROC} and A_{wAFROC} vs. A_{ROC} were generated for the dataset used in the previous illustrations.