

Chapter 04: The ratings paradigm

Table of contents

1. Introduction
2. The ROC counts table
3. Operating points from counts table
4. Relation between ratings paradigm and the binary task
5. Ratings are not numerical values
6. A single "clinical" operating point from ratings data
7. The forced choice paradigm
8. Observer performance studies as laboratory simulations of clinical tasks
9. Discrete vs. continuous ratings: the Miller study
10. The BIRADS ratings scale and ROC studies
11. The controversy
12. Discussion
13. References

Online Supplementary Material

Online Appendix 4.A

4.1: Introduction

In **Chapter 02** the binary task and associated concepts of sensitivity, specificity, true positive fraction, false positive fraction, positive and negative predictive values were introduced. **Chapter 03** introduced the concepts of a random scalar decision variable, or z-sample for each case, which is compared, by the observer, to a fixed reporting threshold ζ , resulting in two types of decisions, “case is non-diseased” or “case is diseased” depending on whether the realized z-sample is less than, or greater than or equal to the reporting threshold. It described a statistical model, for the binary task, characterized by two unit-variance normal distributions separated by μ . The concept of an underlying receiver operating characteristic (ROC) curve with the reporting threshold defining an operating point on the curve was introduced and the advisability of using the area under the curve as a measure of performance, which is independent of reporting threshold, was stressed.

In this chapter the more commonly used *ratings* method will be described, which yields greater definition to the underlying ROC curve than just one operating point obtained in the binary task, and moreover, is more efficient. In this method, the observer assigns a rating to each case. Described first is a typical ROC counts table and how operating points (i.e., pairs of FPF and TPF values) are calculated from the counts data. A labeling convention for the operating points is introduced. Notation is introduced for the observed integers in the counts table and the rules for calculating operating points are expressed as formulae and implemented in **R**. The ratings method is contrasted to the binary method, in terms of efficiency and practicality. A theme occurring repeatedly in this book, that the *ratings are not numerical values but rather they are ordered labels* is illustrated with an example. A method of collecting ROC data on a 6-point scale is described that has the advantage of yielding an unambiguous single operating point. The forced choice paradigm is described. Two controversies are described: one on the utility of discrete (e.g., 1 to 6) vs. quasi-continuous (e.g., 0 to 100) ratings and the other on the applicability of a clinical screening mammography-reporting scale for ROC analyses. Both of these are important issues and it would be a disservice to the readers of the book if the author did not express his position on them.

4.2: The ROC counts table

In a positive-directed rating scale with five discrete levels, the ratings could be the ordered labels “1”: definitely non-diseased, “2”: probably non-diseased, “3”: could be non-diseased or diseased, “4”: probably diseased, “5”: definitely diseased. At the conclusion of the ROC study an *ROC counts table* is constructed. This is the generalization to rating studies of the 2 x 2 decision vs. truth table introduced in **Chapter 02**, Table 2.1. This type of data representation is sometimes called a *frequency table*, but frequency¹ means a *rate* of number of events per some unit, so the author prefers the clearer term “counts”.

Table 4.1 is a representative counts table for a 5-rating study that summarizes the collected data. It is the starting point for analysis. The top half of the table lists the number of counts in each ratings bin, listed separately for non-diseased and diseased cases, respectively. The data is from an actual clinical study¹.

¹ frequency |'frēkwənsē| noun (pl. frequencies)

The rate at which something occurs or is repeated over a particular period of time or in a given sample: shops have closed with increasing frequency during the period.

- the fact of being frequent or happening often.
- Statistics the ratio of the number of actual to possible occurrences of an event.
- Statistics the (relative) number of times something occurs in a given sample.

Table 4.1: A typical ROC counts table. Listed in the upper half of the table are the number of cases in specific ratings bins, listed separately for actually non-diseased and actually diseased cases. There are $K_1 = 60$ non-diseased cases and $K_2 = 50$ diseased cases in this dataset. The lower half of the table lists the corresponding FPF and TPF values, i.e., the abscissa and ordinate, respectively, of the operating points on the ROC plot.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	30	19	8	2	1
$K_2 = 50$	5	6	5	12	22
	Operating points				
	Ratings ≥ 5	Ratings ≥ 4	Ratings ≥ 3	Ratings ≥ 2	Ratings ≥ 1
FPF	0.017	0.050	0.183	0.500	1
TPF	0.440	0.680	0.780	0.900	1

In this example, there are $K_1 = 60$ non-diseased cases and $K_2 = 50$ diseased cases. Of the 60 non-diseased cases 30 were assigned the "1" rating, 19 were assigned the "2" rating, eight the "3" rating, two the "4" rating and one received the "5" rating. The distribution of counts is tilted towards the "1" rating end, but there is some spread and one actually non-diseased case appeared definitely diseased to the observer. In contrast, the distribution of the diseased cases is tilted towards the "5" rating end. Of the 50 diseased cases, 22 received the "5" rating, 12 the "4" rating, five the "3" rating, six the "2" rating and five the "1" rating. The spread appears to be more pronounced for the diseased cases, e.g., five of the 50 cases appeared to be definitely non-diseased to the observer. A little thought should convince you that the observed tilting of the counts, towards the "1" end for actually non-diseased cases, and towards the "5" end for actually diseased cases, is reasonable. However, one should be forewarned not to jump to conclusions about the spread of the data being larger for diseased than for non-diseased cases. While it turns out to be true, the ratings are *merely ordered labels*, and modeling is required, to be described in **Chapter 06**, that uses only the ordering information implicit in the labels, not the actual values, to reach quantitative conclusions.

4.3: Operating points from counts table

It is critical to understand the following example. The bottom half of Table 4.1 illustrates how ROC operating points are calculated from the cell counts. One starts with non-diseased cases that were rated five or more (in this example, since 5 is the highest allowed rating, the "or more" clause is superfluous) and divides by the total number of non-diseased cases, $K_1 = 60$. This yields the abscissa of the lowest non-trivial operating point, namely $FPF_{\geq 5} = 1/60 = 0.017$. The subscript on FPF is intended to make explicit which ratings are being cumulated. The corresponding ordinate is obtained by dividing the number of diseased cases rated "5" or more

and dividing by the total number of diseased cases, $K_2 = 50$, yielding $TPF_{\geq 5} = 22/50 = 0.440$. The coordinates of the lowest operating point are (0.017, 0.44). The abscissa of the next higher operating point is obtained by dividing the number of non-diseased cases that were rated "4" or more and dividing by the total number of non-diseased cases, i.e., $FPF_{\geq 4} = 3/60 = 0.05$. Similarly the ordinate of this operating point is obtained by dividing the number of diseased cases that were rated "4" or more and dividing by the total number of diseased cases, i.e., $TPF_{\geq 4} = 34/50 = 0.680$. The procedure, which at each stage *cumulates* the number of cases equal to or greater (in the sense of increased confidence level for disease presence) than a specified label, is repeated to yield the rest of the operating points listed in Table 4.1. Since they are computed *directly* from the data, without any assumption, they are called *empirical* or *observed* operating points. After done this once it would be nice to have a formula implementing the process, one use of which would be to code the procedure. First, one needs appropriate notation for the bin counts.

Let K_{1r} denote the number of non-diseased cases rated r , and K_{2r} denote the number of diseased cases rated r . For convenience, define dummy counts $K_{1(R+1)} = K_{2(R+1)} = 0$, where R is the number of ROC bins. This construct allows inclusion of the origin (0,0) in the formulae. *The range of r is $r = 1, 2, \dots, (R+1)$* . Within each truth-state, the individual bin counts sum to the total number of non-diseased and diseased cases, respectively. The following equations summarize all this:

$$\left. \begin{aligned} K_1 &= \sum_{r=1}^{R+1} K_{1r} \\ K_2 &= \sum_{r=1}^{R+1} K_{2r} \\ K_{1(R+1)} &= K_{2(R+1)} = 0 \\ r &= 1, 2, \dots, (R+1) \end{aligned} \right\} \quad . \quad (4.1)$$

To be clear, Table 4.1 is repeated to show the meaning of the counts notation, Table 4.2.

Table 4.2: Explanation of the notation for cell counts for $R = 5$. The upper half of the table, reproduced from Table 4.1, illustrates the notation while the lower half shows the values or r , which in conjunction with Eqn. (4.2), determine the operating points.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	$K_{11} = 30$	$K_{12} = 19$	$K_{13} = 8$	$K_{14} = 2$	$K_{15} = 1$
$K_2 = 50$	$K_{21} = 5$	$K_{22} = 6$	$K_{23} = 5$	$K_{24} = 12$	$K_{25} = 22$

	Operating points				
	$r = R = 5$	$r = R - 1 = 4$	$r = R - 2 = 3$	$r = R - 3 = 2$	$r = R - 4 = 1$
	≥ 5	≥ 4	≥ 3	≥ 2	≥ 1
FPF	0.017	0.050	0.183	0.500	1
TPF	0.440	0.680	0.780	0.900	1

The operating points are defined by:

$$\begin{aligned} FPF_r &= \frac{1}{K_1} \sum_{s=r}^{R+1} K_{1s} \\ TPF_r &= \frac{1}{K_2} \sum_{s=r}^{R+1} K_{2s} \end{aligned} \quad . \quad (4.2)$$

The labeling of the points follows the following convention: $r = 1$ corresponds to the upper right corner (1,1) of the ROC plot, a *trivial operating point since it is common to all datasets*. Next, $r = 2$ is the next lower operating point, etc., and $r = R$ is the lowest non-trivial operating point and finally $r = R + 1$ is the origin (0,0) of the ROC plot, which is also a trivial operating point, because it is common to all datasets. In other words, the operating points are numbered starting with the upper right corner, labeled 1, and working down the curve, each time increasing the label by one.

If $r = 1$ one gets the uppermost "trivial" operating point (1,1):

$$\begin{aligned} FPF_1 &= \frac{1}{K_1} \sum_{s=1}^{R+1} K_{1s} = \frac{60}{60} = 1 \\ TPF_1 &= \frac{1}{K_2} \sum_{s=1}^{R+1} K_{2s} = \frac{50}{50} = 1 \end{aligned} \quad . \quad (4.3)$$

The uppermost non-trivial operating point is obtained for $r = 2$, when:

$$\begin{aligned} FPF_2 &= \frac{1}{K_1} \sum_{s=2}^{R+1} K_{1s} = \frac{30}{60} = 0.500 \\ TPF_2 &= \frac{1}{K_2} \sum_{s=2}^{R+1} K_{2s} = \frac{45}{50} = 0.900 \end{aligned} \quad . \quad (4.4)$$

The next lower operating point is obtained for $r = 3$:

$$\begin{aligned}
FPF_3 &= \frac{1}{K_1} \sum_{s=3}^{R+1} K_{1s} = \frac{11}{60} = 0.183 \\
TPF_3 &= \frac{1}{K_2} \sum_{s=3}^{R+1} K_{2s} = \frac{39}{50} = 0.780
\end{aligned}
\tag{4.5}$$

The next lower operating point is obtained for $r = 4$:

$$\begin{aligned}
FPF_4 &= \frac{1}{K_1} \sum_{s=4}^{R+1} K_{1s} = \frac{3}{60} = 0.050 \\
TPF_4 &= \frac{1}{K_2} \sum_{s=4}^{R+1} K_{2s} = \frac{34}{50} = 0.680
\end{aligned}
\tag{4.6}$$

The lowest non-trivial operating point is obtained for $r = 5$:

$$\begin{aligned}
FPF_5 &= \frac{1}{K_1} \sum_{s=5}^{R+1} K_{1s} = \frac{1}{60} = 0.017 \\
TPF_5 &= \frac{1}{K_2} \sum_{s=5}^{R+1} K_{2s} = \frac{22}{50} = 0.440
\end{aligned}
\tag{4.7}$$

The next value $r = 6$ yields the trivial operating point (0,0):

$$\begin{aligned}
FPF_6 &= \frac{1}{K_1} \sum_{s=6}^{R+1} K_{1s} = \frac{0}{60} = 0 \\
TPF_6 &= \frac{1}{K_2} \sum_{s=6}^{R+1} K_{2s} = \frac{0}{50} = 0
\end{aligned}
\tag{4.8}$$

This exercise shows explicitly that an R-rating ROC study can yield at most $R - 1$ distinct non-trivial operating points; i.e., those corresponding to $r = 2, 3, \dots, R$.

The modifier “at most” is needed, because if *both counts* (i.e., non-diseased and diseased) for bin r' are zeroes, then that operating point merges with the one immediately below-left of it:

$$\begin{aligned}
FPF_{r'} &= \frac{1}{K_1} \sum_{r=r'}^{R+1} K_{1r} = \frac{1}{K_1} \sum_{r=r'+1}^{R+1} K_{1r} = FPF_{r'+1} \\
TPF_{r'} &= \frac{1}{K_2} \sum_{r=r'}^{R+1} K_{2r} = \frac{1}{K_2} \sum_{r=r'+1}^{R+1} K_{2r} = TPF_{r'+1}
\end{aligned}
\tag{4.9}$$

Since bin r' is unpopulated, one can re-label the bins to exclude the unpopulated bin, and now the total number of bins is effectively $R-1$.

Since one is cumulating counts, which can never be negative, the highest non-trivial operating point resulting from cumulating the 2 through 5 ratings has to be to the upper-right of the next adjacent operating point resulting from cumulating the 3 through 5 ratings. This in turn has to be to the upper-right of the operating point resulting from cumulating the 4 through 5 ratings. This in turn has to be to the upper right of the operating point resulting from the 5 ratings. In other words, as one cumulates ratings bins, the operating point must move monotonically up and to the right, or more accurately, the point cannot move down or to the left. If a particular bin has *zero* counts for non-diseased cases, and *non-zero* counts for diseased cases, the operating point moves vertically *up* when this bin is cumulated; if it has zero counts for diseased cases, and non-zero counts for non-diseased cases, the operating point moves horizontally to the *right* when this bin is cumulated.

It is useful to replace the preceding detailed explanation with a simple algorithm that incorporates all the logic. Online Appendix 4.A describes the **R** code in **MainOpPtsFromCountsTable.R** for calculating operating points from counts for the data in Table 4.1. **Source** the file to get §4.3.1 and Fig. 4.1 (A - B).

4.3.1: Code output

```

> source('~\book2\02 A ROC analysis\A4 RatingsParadigm\software\MainOpPtsFromCountsTable.R')
FPF =
0.01667 0.05 0.1833 0.5
TPF =
0.44 0.68 0.78 0.9
uppermost point based estimate of mu = 1.282
corresponding estimate of Az = 0.8176
showing observed operating points and eq. var. fitted ROC curve
binormal estimate of Az = 0.8696
showing observed operating points and uneq. var. fitted ROC curve

```

The code lists the values of the arrays FPF and TPF, which correspond to those listed in Table 4.1.

It was shown in **Chapter 03** that in the equal variance binormal model, an operating point determines the parameters μ , Eqn. (3.21), or equivalently $A_{z;\sigma=1}$, Eqn. (3.30). The last three lines of §4.3.1 illustrate the application of these formulae using the coordinates (0.5, 0.9) of the uppermost non-trivial operating point. It should come as no surprise that the uppermost operating point in Fig. 4.1 (A) is *exactly* on the predicted curve: after all, this point was used to calculate $\mu = 1.282$. The corresponding value of ζ can be calculated from Eqn. (3.17), namely:

$$\Phi^{-1}(Sp) = \zeta \quad . \quad (4.10)$$

Alternatively, using Eqn. (3.18):

$$\mu - \zeta = \Phi^{-1}(Se) \Rightarrow \zeta = \mu - \Phi^{-1}(Se) \quad . \quad (4.11)$$

These are coded in §4.3.2.

4.3.2: Code snippet

```
> qnorm(1-0.5)
[1] 0
> mu-qnorm(0.9)
[1] 0
```

Either way, one gets the same result: $\zeta = 0$. It should be clear that $\zeta = 0$ makes sense: FPF = 0.5 is consistent with half of the (symmetrical) unit-normal non-diseased distribution being above $\zeta = 0$. The transformed value ζ is a genuine numerical value. *To reiterate, ratings cannot be treated as numerical values, but thresholds, estimated from an appropriate model, can.*

The ROC curve in Fig. 4.1 (A), as determined by the uppermost operating point, passes *exactly* through this point but misses the others. If a different operating point were used to estimate μ and $A_{z;\sigma=1}$, the estimated values would have been different and the new curve would pass exactly through the selected point. No choice of μ yields a satisfactory visual fit to the experimental data points. The reader should confirm these statements with appropriate modifications to the code. This is the reason one needs a modified model, with an extra parameter, namely the unequal variance binormal model, to fit radiologist data (the extra parameter is the ratio of the standard deviations of the two distributions).

Fig. 4.1 (B) shows the predicted ROC curve by the unequal variance binormal model, to be introduced in **Chapter 06**. Notice the improved visual quality of the fit. Each observed point is "not engraved in stone", rather both FPF and TPF are subject to sampling variability. Estimation of confidence intervals for FPF and TPF was addressed in §3.10. [A detail: the estimated confidence interval in the preceding chapter was for a *single* operating point; since the multiple operating points are correlated – some of the counts used to calculate them are common to two or more operating points – the method tends to overestimate the confidence interval. A modeling approach is possible to estimate confidence intervals that accounts for data correlation and this yields tighter confidence intervals.]

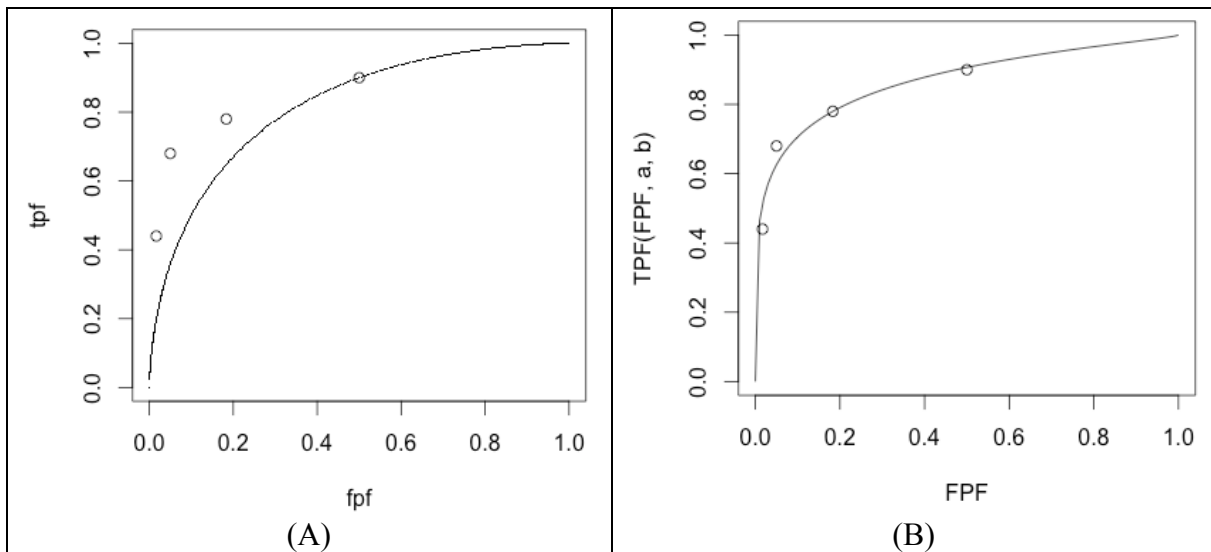


Fig. 4.1 (A): Predicted ROC curve for $\mu = 1.282$ superposed on the operating points obtained from the data in Table 4.1. (B): same data fitted with a two-parameter model $\mu = 2.17, \sigma = 1.65$ described in **Chapter 06**, in which μ is the separation of the normal distributions and σ is the standard deviation of the diseased distribution; the non-diseased distribution has unit standard deviation.

Consider what happens if the observer does not fully utilize the rating scale. For example, if the observer chooses to ignore the gradations implied by the intermediate ratings 2, 3 and 4, essentially lumping them into the 5 rating, then the observed ROC counts table would be as shown in Table 4.3. Essentially the observer responds with ones and fives, so R is effectively equal to two. In this example, the number of *distinct* non-trivial operating points is one, just as in the binary-decision task of **Chapter 02** (note the repeating operating points in the lower half of the table).

Table 4.3: An atypical ROC counts table where the observer ignores the gradations implied by the intermediate ratings 2, 3 and 4, essentially lumping them into the 5 rating. Otherwise, the structure of this table is identical to Table 4.1.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	30	0	0	0	30
$K_2 = 50$	5	0	0	0	45
	Operating points				
	≥ 5	≥ 4	≥ 3	≥ 2	≥ 1
FPF	0.500	0.500	0.500	0.500	1
TPF	0.900	0.900	0.900	0.900	1

Even though the operating point (0.5, 0.9) repeats in 4 columns, there is only one unique value.

What if the observer were to use only the 1 and 2 ratings bins, effectively ignoring the gradations implied by the upper levels of the rating scale? Or stated equivalently, the observer lumps the 3, 4 and 5 rating into the 2 rating, Table 4.4.

Table 4.4: Another atypical ROC counts table where the observer ignores the gradations implied by the intermediate ratings 3, 4 and 5, essentially lumping them into the 2 rating.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	30	30	0	0	0
$K_2 = 50$	5	45	0	0	0
	Operating points				
	≥ 5	≥ 4	≥ 3	≥ 2	≥ 1
FPF	0	0	0	0.500	1
TPF	0	0	0	0.900	1

Once again, the number of distinct non-trivial operating points is one, and the coordinates of the operating point (0.500, 0.900) are identical to that in Table 4.3, where the observer used the 1's and 5's only. This illustrates the intrinsic nature of the ratings as ordered labels. Given binary ratings 1s and 5s vs. 1s and 2s, one may *not* conclude that just because the actual difference between five and one is four times larger than that between two and one, that the discriminability between the non-diseased and diseased cases is 4 times larger. As long as the single operating point is unaltered, as between Table 4.3 and Table 4.4, there is no change in performance as quantified by μ or equivalently, by the area $A_{z, \sigma=1}$ under the ROC curve.

4.4: Relation between ratings paradigm and the binary paradigm

Table 4.2 corresponds to $R = 5$. In **Chapter 02** it was shown that the binary task requires a *single* fixed threshold parameter ζ and a decision rule, namely, to give the case a diseased rating of 2 if $Z \geq \zeta$ and a rating of 1 otherwise.

The R-rating task can be viewed as $(R - 1)$ *simultaneously* conducted binary tasks each with its own fixed threshold $\zeta_r, r = 1, 2, \dots, R - 1$. It is efficient compared to $(R - 1)$ *sequentially* conducted binary tasks; however, the onus is on the observer to maintain *fixed*-multiple thresholds through the duration of the study.

The rating method is a more efficient way of collecting the data compared to running the study repeatedly with appropriate instructions to cause the observer to adopt different fixed thresholds specific to each replication. In the clinical context such repeated studies would be impractical because it would introduce memory effects, wherein the diagnosis of a case would depend on how many times the case had been seen, along with other cases, in previous sessions. A second reason is that it is difficult for a radiologist to change the operating threshold in response to instructions. To the author's knowledge, repeated use of the binary paradigm has not been used in any clinical ROC study.

How does one model the binning? For convenience one defines dummy thresholds $\zeta_0 = -\infty$ and $\zeta_R = +\infty$, in which case the thresholds satisfy the ordering requirement $\zeta_{r-1} < \zeta_r, r = 1, 2, \dots, R$. The *rating rule* is:

$$\text{if } \zeta_{r-1} \leq z < \zeta_r \Rightarrow \text{rating} = r \quad . \quad (4.12)$$

For the dataset in Table 4.1 the *empirical* thresholds (as opposed to *modeled* thresholds via Eqn. (4.10) or Eqn. (4.11); note that empirical thresholds are not true numerical values) are as follows (the superscript E is for empirical):

$$\left. \begin{array}{l} \zeta_r^E = r \\ r = 1, \dots, R - 1; \\ \zeta_0^E = -\infty; \zeta_R^E = +\infty \end{array} \right\} \quad . \quad (4.13)$$

In Table 4.1 the number of bins is $R = 5$. The "simultaneously conducted binary tasks" nature of the rating task can be appreciated from the following examples. Suppose one selects the threshold for the first binary task to be $\zeta_4^E = 4$. Therefore a case rated 5 satisfies $\zeta_4^E \leq 5 < \zeta_5^E$, consistent with Eqn. (4.12). The operating point corresponding to $\zeta_r^E = 4$, obtained by cumulating all cases rated five, yields (0.017, 0.440). In the second binary-task, one selects as threshold $\zeta_3^E = 3$. Therefore, a case rated four satisfies the inequality $\zeta_3^E \leq 4 < \zeta_4^E$, again consistent with Eqn. (4.12). The operating point corresponding to $\zeta_3^E = 3$, obtained by cumulating all cases rated four or five, yields (0.05, 0.680). Similarly, for $\zeta_2^E = 2$, $\zeta_1^E = 1$ and $\zeta_0^E = 0$, which, according to the binning rule Eqn. (4.12), yield counts in bins 3, 2 and 1, respectively. The last is a trivial operating point. The non-trivial operating points are generated by thresholds ζ_r^E where $r = 1, 2, 3$ and 4. A five-rating study has four associated thresholds and a corresponding number of equivalent binary studies. In general, an R rating study has $R-1$ associated thresholds.

4.5: Ratings are not numerical values

The ratings are to be thought of as *ordered labels*, not as numeric values. Arithmetic operations that are allowed on numeric values, such as averaging, are not allowed on ratings. One could have relabeled the ratings in Table 4.2 as A, B, C, D and E, where $A < B$ etc. As long as the counts in the body of the table are unaltered, such relabeling would have no effect on the observed operating points and the fitted curve. Of course one cannot average the labels A, B, etc. of different cases. The issue with numeric labels is not fundamentally different. At the root is that the difference in thresholds corresponding to the different operating points are not in relation to the difference between their numeric values. There is a way to estimate the underlying thresholds, if one assumes a specific model, for example the unequal-variance binormal model to be described in **Chapter 06**. The thresholds so obtained are genuine numeric values and can be averaged. [Not to hold the reader in suspense, the four thresholds corresponding to the data in Table 4.1 are $\zeta_1 = 0.007676989$, $\zeta_2 = 0.8962713$, $\zeta_3 = 1.515645$ and $\zeta_4 = 2.396711$; see §6.4.1; these values would be unchanged if, for example, the labels were doubled, with allowed values 2, 4, 6, 8 and 10, or any of an infinite number of rearrangements that preserves their ordering.]

The temptation to regard confidence levels / ratings as numeric values can be particularly strong when one uses a large number of bins to collect the data. One could use of quasi-continuous ratings scale, implemented for

example, by having a slider-bar user interface for selecting the rating. The slider bar typically extends from 0 to 100, and the rating could be recorded as a floating-point number, e.g., 63.45. Here too one cannot assume that the difference between a zero-rated case and a 10 rated case is a tenth of the difference between a zero-rated case and a 100 rated case. So averaging the ratings is not allowed. Additionally, one cannot assume that different observers use the labels in the same way. One observer's 4-rating is not equivalent to another observers 4-rating. *Working directly with the ratings is a bad idea: valid analytical methods use the rankings of the ratings, not their actual values.* The reason for the emphasis is that there are serious misconceptions about ratings. The author is aware of a publication stating, to the effect, that a modality resulted in an increase in average confidence level for diseased cases. Another publication used a specific numerical value of a rating to calculate the operating point for each observer – this assumes all observers use the rating scale in the same way.

4.6: A single "clinical" operating point from ratings data

The reason for the quotes in the title to this section is that a single operating point on a laboratory ROC plot, no matter how obtained, has little relevance to how radiologists operate in the clinic. However, some consider it useful to quote an operating point from an ROC study. For a 5-rating ROC study, Table 4.1, it is not possible to unambiguously calculate the operating point of the observer in the binary task of discriminating between non-diseased and diseased cases. One possibility would be to use the three and above ratings to define the operating point, but one might have chosen two and above. A second possibility is to instruct the radiologist that a four or higher rating, for example, implies the case would be reported “clinically” as diseased. However, the radiologist can only pretend so far that this study, which has no clinical consequences, is somehow a “clinical” study. If a single laboratory study based operating point is desired², the best strategy², in the author's opinion, is to obtain the rating via two questions. This method is also illustrated in a book on detection theory, Ref. 3, Table 3.1. The first question is "is the case diseased?" The binary (Yes/No) response to this question allows unambiguous calculation of the operating point, as in **Chapter 02**. The second question is: "what is your confidence in your previous decision?" and allow three responses, namely Low, Medium and High. The dual-question approach is equivalent to a 6-point rating scale, Fig. 4.2.

² The author owes this insight to Prof. Harold Kundel.

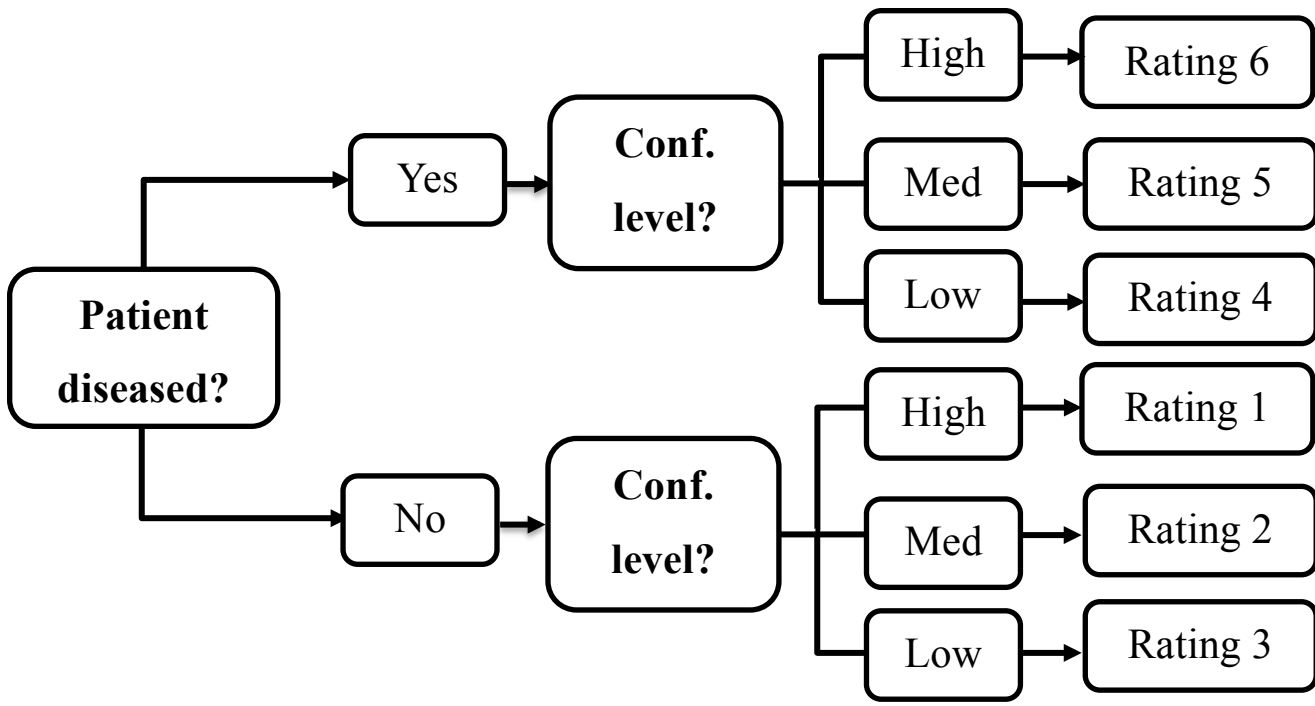


Fig. 4.2: This figure describes a method for acquiring ROC data on an effectively 6-point scale that also yields an unambiguous single operating point for declaring patients diseased. The data collection consists of two questions. The answer to the first question, is the patient diseased, allows unambiguous construction of a single operating point for disease presence. The answer to the second question, what is your confidence level in that decision, yields multiple operating points. Note the reversal of the final ratings in the last "column" in the lower half of the figure.

The ordering of the ratings can be understood as follows. The four, five and six ratings are as expected. If the radiologist states the patient is diseased and the confidence level is high that is clearly the highest end of the scale, i.e., six, and the lower confidence levels, five and four, follow, as shown. If, on the other hand, the radiologist states the patient is non-diseased, and the confidence level is high, then that must be the lowest end of the scale, i.e., "1". The lower confidence levels in a negative decision must be higher than "1", namely "2" and "3", as shown. As expected, the low confidence ratings, namely "3" (non-diseased, low confidence) and "4" (diseased, low confidence) are adjacent to each other. With this method of data-collection, there is no confusion as to what rating defines the single desired operating point as this is determined by the binary response to the first question. The 6-point rating scale is also sufficiently fine to not smooth out the ability of the radiologist to maintain distinct different levels. In the author's experience, using this scale one expects rating noise of about $\pm \frac{1}{2}$ a rating bin, i.e., the same difficult case, shown on different occasions to the same radiologist (with sufficient time lapse or other intervening cases to minimize memory effects) is expected to elicit a "3" or "4", with roughly equal probability.

4.7: The forced choice paradigm

In each of the four paradigms (ROC, FROC, LROC and ROI) described in **Chapter 01**, patient images are displayed one patient at a time. A fifth paradigm involves presentation of multiple images to the observer, where one image (or set of images from one patient, i.e., a case) is from a diseased patient, and the rest are from non-diseased patients. The observer's task is to pick the image, or the case, that is most likely to be from the diseased patient. If the observer is correct, the event is scored as a "one" and otherwise it is scored as a "zero". The process is repeated with other sets of independent patient images, each time satisfying the condition that one patient is diseased and the rest are non-diseased. The sum of the scores divided by the total number of scores is the probability of a correct choice, denoted $P(C)$. If the total number of cases presented at the same time is denoted n , then the task is termed n -alternative forced choice ($nAFC$)⁴. If only two cases are presented, one diseased and the other non-diseased, then $n = 2$ and the task is 2AFC. In Fig. 4.3, in the left image a Gaussian nodule is superposed on a square region extracted from a non-diseased mammogram. The right image is a region extracted from a different non-diseased mammogram (one should not use the same background in the two images – the analysis assumes that different, i.e., independent images, are shown). If the observer clicks on the left image, a correct choice is recorded. [In some 2AFC-studies, the backgrounds are simulated non-diseased images. They resemble mammograms; the resemblance depends on the expertise of the observer: expert radiologists can tell that they are not true mammograms. They are actually created by filtering the random white noise with a $1/f^3$ spatial filter⁵⁻¹².]

The 2AFC paradigm is popular, because its analysis is straightforward, and there exists a theorem⁴ that $P(C)$, the probability of a correct choice in the 2AFC task, equals, to within sampling variability, the *true* area under the true (not fitted, not empirical) ROC. Another reason for its popularity is possibly the speed at which data can be collected, sometimes only limited by the speed at which disk stored images can be displayed on the monitor. While useful for studies into human visual perception on relatively simple images, and the model observer community has performed many studies using this paradigm¹³⁻²⁰, the author cannot recommend it for clinical studies because it does not resemble any clinical task. In the clinic, radiologists never have to choose the diseased patient out of a pair consisting of one diseased and one non-diseased. Additionally, the forced-choice paradigm is wasteful of known-truth images, often a difficult/expensive resource to come by, because better statistics²¹ (tighter confidence intervals) are obtained by the ratings ROC method or by utilizing location specific extensions of the ROC paradigm. [The author is not aware of the 2AFC method being actually used to assess imaging systems using radiologists to perform real clinical tasks on real images.]

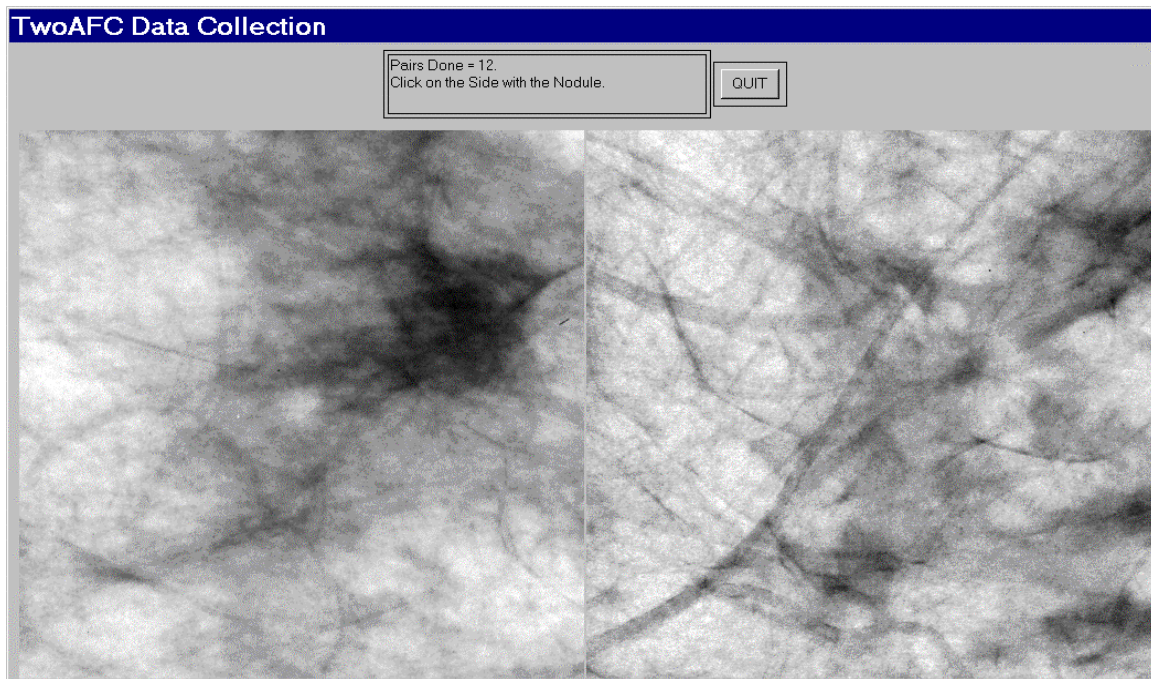


Fig. 4.3: Example of image presentation in a 2AFC study. The left image contains, at its center, a positive contrast Gaussian shape disk superposed on a non-diseased mammogram. The right image does not contain a lesion at its center and the background is from a different non-diseased patient. If the observer clicks on the left image it is recorded as a correct choice, otherwise it is recorded as an incorrect choice. The number of correct choices divided by the number of paired presentations is an estimate of the probability of a correct choice, which can be shown to be identical, apart from sampling variability, to the true area under the ROC curve. This is an example of a signal known exactly location known exactly (SKE-LKE) task widely used by the model observer community.

4.8: Observer performance studies as laboratory simulations of clinical tasks

Observer performance paradigms (ROC, FROC, LROC and ROI) should be regarded as experiments conducted in a laboratory (i.e., controlled) setting that are intended to be representative of the actual clinical task. They should not to be confused with performance in a real "live" clinical setting: there is a known "laboratory effect"²²⁻²⁴. For example, in one study radiologists performed better during live clinical interpretations than they did later, on the same cases, in a laboratory ROC study²². This is expected because there is more at stake during live interpretations: e.g., the patient's health and the radiologist's reputation, than during laboratory ROC studies. The claimed "laboratory effect" has caused some controversy. A paper²⁵ titled "*Screening mammography: test set data can reasonably describe actual clinical reporting*" argues against the laboratory effect.

Real clinical interpretations happen every day in radiology departments all over the world. In the laboratory, the radiologist is asked to interpret the images "as if in a clinical setting" and render a "diagnosis". The laboratory decisions have no clinical consequences, e.g., the radiologist will not be sued for mistakes and their ROC study

decisions have no impact on the clinical management of the patients. Usually laboratory ROC studies are conducted on retrospectively acquired images. Patients, whose images were used in an ROC study, have already been imaged in the clinic and decisions have already been made on how to manage them.

There is no guarantee that results of the laboratory study are directly applicable to clinical practice. Indeed there is an assumption that the laboratory study correlates with clinical performance. Strict equality is not required, simply that the performance in the laboratory is related monotonically to actual clinical performance.

Monotonicity assures preservation of performance orderings, e.g., a radiologist has greater performance than another does or one modality is superior to another, regardless of how they are measured, in the laboratory or in the clinic. The correlation is taken to be an axiomatic truth by researchers, when in fact it is an assumption. To the extent that the participating radiologist brings his/her full clinical expertise to bear on each laboratory image interpretation, this assumption is likely to be valid.

This section provoked a strong negative response from a collaborator. To paraphrase him, "*Dear Dev, my friend, I think it is a pity in this book chapter you argue that these studies are simulations. I mean, the reason people perform these studies is because they believe in the results*".

The author also believes in observer performance studies. Otherwise, he would not be writing this book. Distrust of the word "simulation" seems to be peculiar to this field. Simulations are widely used in "hard" sciences, e.g., they are used in astrophysics to determine conditions dating to 10^{-31} seconds after the big bang. Simulations are not to be taken lightly. Conducting clinical studies is very difficult as there are many factors not under the researcher's control. *Observer performance studies of the type described in this book are the closest that one can come to the "real thing" and the author is a firm believer in them.* These studies include key elements of the actual clinical task: the entire imaging system, radiologists (*assuming the radiologist take these studies seriously in the sense of bringing their full clinical expertise to bear on each image interpretation*) and real clinical images and as such are expected to correlate with real "live" interpretations. Proving this correlation is going to be difficult as there are many factors that complicated real interpretations. It is not clear to the author that proving or disproving this correlation is ever going to be a settled issue.

4.9: Discrete vs. continuous ratings: the Miller study

There is controversy about the merits of discrete vs. continuous ratings²⁶⁻²⁸. Since the late Prof. Charles E. Metz and the late Dr. Robert F. Wagner have both backed the latter (i.e., continuous or quasi-continuous ratings) new ROC study designs sometimes tend to follow their advice. The author's recommendation is to follow the 6-point rating scale as outlined in Fig. 4.2. This section provides the background for the recommendation.

A widely cited (22,909 citations at the time of writing) 1954 paper by Miller²⁹ titled "*The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*" is relevant. It is a readable paper, freely downloadable in several languages (www.musanim.com/miller1956/). In the author's judgment, this paper has not received the attention it should have in the ROC community, and for this reason portions from it are reproduced below. [George Armitage Miller, February 3, 1920 – July 22, 2012, was one of the founders of the field of cognitive psychology.]

Miller's first objective was to comment on *absolute judgments of unidimensional stimuli*. Since all (univariate, i.e., single decision per case) ROC models assume a unidimensional decision variable, Miller's work is highly relevant. He comments on two papers by Pollack^{30,31}. Pollack asked listeners to identify tones by assigning numerals to them, analogous to a rating task described above. The tones differed in frequency, covering the range 100 to 8000 Hz in equal logarithmic steps. A tone was sounded and the listener responded by giving a numeral (i.e., a rating, with higher values corresponding to higher frequencies). After the listener had made his response, he was told the correct identification of the tone. *When only two or three tones were used, the listeners never confused them. With four different tones, confusions were quite rare, but with five or more tones, confusions were frequent. With fourteen different tones, the listeners made many mistakes.* Since it is so succinct, the entire content of the first (1952) paper by Pollack is reproduced below:

"In contrast to the extremely acute sensitivity of a human listener to discriminate small differences in the frequency or intensity between two sounds is his relative inability to identify (and name) sounds presented individually. When the frequency of a single tone is varied in equal - logarithmic steps in the range between 100 cps and 8000 cps (and when the level of the tone is randomly adjusted to reduce loudness cues), the amount of information transferred is about 2.3 bits per stimulus presentation. This is equivalent to perfect identification among only 5 tones. The information transferred, under the conditions of measurement employed, is reasonably invariant under wide variations in stimulus conditions."

By “information” is meant (essentially) the number of levels, measured in *bits* (binary digits), thereby making it independent of the unit of measurement: 1 bit corresponds to a binary rating scale, 2 bits to a four-point rating scale and 2.3 bits to $2^{2.3} = 4.9$, i.e., about 5 ratings bins. Based on Pollack’s’ original unpublished data, Miller put an upper limit of 2.5 bits (corresponding to about 6 ratings bins) on the amount of information that is transmitted by listeners who make absolute judgments of auditory pitch. A second paper³¹ by Pollack was related to: (1) the frequency range of tones; (2) the utilization of objective reference tones presented with the unknown tone; and (3) the “dimensionality”—the number of independently varying stimulus aspects. Little additional gain in information transmission was associated with the first factor; a moderate gain was associated with the second; and a relatively substantial gain was associated with the third (we return to the dimensionality issue below).

As an interesting side-note, Miller states:

“Most people are surprised that the number is as small as six. Of course, there is evidence that a musically sophisticated person with absolute pitch can identify accurately any one of 50 or 60 different pitches. Fortunately, I do not have time to discuss these remarkable exceptions. I say it is fortunate because I do not know how to explain their superior performance. So I shall stick to the more pedestrian fact that most of us can identify about one out of only five or six pitches before we begin to get confused.

It is interesting to consider that psychologists have been using seven-point rating scales for a long time, on the intuitive basis that trying to rate into finer categories does not really add much to the usefulness of the ratings. Pollack's results indicate that, at least for pitches, this intuition is fairly sound.

Next you can ask how reproducible this result is. Does it depend on the spacing of the tones or the various conditions of judgment? Pollack varied these conditions in a number of ways. The range of frequencies can be changed by a factor of about 20 without changing the amount of information transmitted more than a small percentage. Different groupings of the pitches decreased the transmission, but the loss was small. For example, if you can discriminate five high-pitched tones in one series and five low-pitched tones in another series, it is reasonable to expect that you could combine all ten into a single series and still tell them all apart without error. When you try it, however, it does not work. The channel capacity for pitch seems to be about six and that is the best you can do.”

Miller also quotes work³² on channel capacities for absolute judgments of *loudness*³² (2.3 bits), sensation of *saltiness*³³ (1.9 bits) and judgments of *visual position*³⁴ (3.25 bits).

In contrast to the careful experiments conducted in the psychophysical context to elucidate this issue, the author was unable to find a single study of the number of discrete rating levels that an observer can support. Even lacking such a study, a recommendation has been made to acquire data on a quasi-continuous scale²⁷.

There is no question that for multidimensional data, as observed in the second study by Pollack³¹, the observer can support more than 7 ratings bins. To quote Miller:

“You may have noticed that I have been careful to say that this magical number seven applies to one-dimensional judgments. Everyday experience teaches us that we can identify accurately any one of several hundred faces, any one of several thousand words, any one of several thousand objects, etc. The story certainly would not be complete if we stopped at this point. We must have some understanding of why the one-dimensional variables we judge in the laboratory give results so far out of line with what we do constantly in our behavior outside the laboratory. A possible explanation lies in the number of independently variable attributes of the stimuli that are being judged. Objects, faces, words, and the like differ from one another in many ways, whereas the simple stimuli we have considered thus far differ from one another in only one respect.”

In the medical imaging context, a trivial way to increase the number of ratings would be to color-code the images: red, green and blue; now one can assign a red image rated 3, a green image rated 2, etc., which would be meaningless unless the color encoded relevant diagnostic information. Another ability, quoted in the publication²⁷ advocating continuous ratings is the ability to recognize faces, again a multidimensional categorization task, as noted by Miller. Also quoted as an argument for continuous ratings is the ability of computer aided detection schemes that calculate many features for each perceived lesion and combine them into a single probability of malignancy, which is on a highly precise floating point 0 to 1 scale. Radiologists are not computers. Other arguments for greater number of bins: *it cannot hurt and one should acquire the rating data at greater precision than the noise, especially if the radiologist is able to maintain the finer distinctions*. The author worries that radiologists who are willing to go along with greater precision are over-anxious to co-operate with the experimentalist. In the author's experience, expert radiologists will not modify their reading

style and one should be suspicious when overzealous radiologists accede to an investigators request to interpret images in a style that does not closely resemble the clinic. Radiologists, especially experts, do not like more than about four ratings. The author has worked with a famous chest radiologist (the late Dr. Robert Fraser) who refused to use more than four ratings.

Another reason given for using continuous ratings is it reduces instances of data degeneracy. Data is sometimes said to be degenerate if the curve-fitting algorithm, the binormal model and the proper binormal model, cannot fit it. This occurs, for example, if there are no interior points on the ROC plot. Modifying radiologist behavior to accommodate the limitations of analytical methods seems to be inherently dubious. One could simply randomly add or subtract half an integer from the observed ratings, thereby making the rating scale more granular and reduce instances of degeneracy (this is actually done in some ROC software to overcome degeneracy issues). Another possibility is to use the empirical (trapezoidal) area under the ROC curve. This quantity can always be calculated; there are no degeneracy problems with it. Actually, fitting methods now exist that are robust to data degeneracy, such as discussed in **Chapter 18** and **Chapter 20**, so this reason for acquiring continuous data no longer applies.

The rating task involves a unidimensional scale and the author sees no way of getting around the basic channel-limitation noted by Miller and for this reason the author recommends a 6 point scale, as in Fig. 4.2.

On the other side of the controversy it has been argued that given a large number of allowed ratings levels the observer essentially bins the data into a much smaller number of bins (e.g., 0, 20, 40, 60, 80, 100) and adds a zero-mean noise term to appear to be "spreading out the ratings"³⁵. The author agrees with this reasoning.

4.10: The BI-RADS ratings scale and ROC studies

It is desirable that the rating scale be relevant to the radiologists' daily practice. This assures greater consistency – the fitting algorithms assume that the thresholds are held constant for the duration of the ROC study.

Depending on the clinical task, a natural rating scale may already exist. For example, in 1992 the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) to standardize mammography reporting³⁶. There are six assessment categories: category 0 indicates need for additional imaging; category 1 is a negative (clearly non-diseased) interpretation; category 2 is a benign finding; category 3 is probably benign, with short-interval follow-up suggested; category 4 is a suspicious abnormality for which biopsy should be considered; category 5 is highly suggestive of malignancy and appropriate action should be

taken. The 4th edition of the BI-RADS manual³⁷ divides category 4 into three subcategories 4A, 4B and 4C and adds category 6 for a proven malignancy. The 3-category may be further subdivided into “probably benign with a recommendation for normal or short-term follow-up” and a 3+ category, “probably benign with a recommendation for immediate follow-up”. Apart from categories 0 and 2, the categories form an ordered set with higher categories representing greater confidence in presence of cancer. How to handle the 0s and the 2s is the subject of some controversy, described next.

4.11: The controversy

Two large clinical studies have been reported in which BI-RADS category data were acquired for > 400,00 screening mammograms interpreted by many (124 in the 1st study) radiologists^{38,39}. The purpose of the 1st study was to relate radiologist characteristics to actual performance (e.g., does performance depend on reading volume – the number of cases interpreted per year), so it could be regarded as a more elaborate version of the Beam et al. study⁴⁰, described in **Chapter 03**. The purpose of the second study was to determine the effectiveness of computer-aided detection (CAD) in screening mammography.

The reported ROC analyses used the BIRADS assessments labels ordered as follows: $1 < 2 < 3 < 3+ < 0 < 4 < 5$. The last column of Table 4.5 shows that with this ordering the numbers of cancer per 1000 patients increases monotonically. The CAD study is discussed later, for now the focus is on the adopted BIRADS scale ordering that is common to both studies and which has raised controversy.

Table 4.5: The ordering of the BI-RADS ratings shown in the first column correlates with the cancer-rate shown in the last column. Data from Barlow et al. Ref. 38.

BI-RADS assessment	Total number of mammograms	Mammograms in women without breast cancer (%)	Mammograms in women with breast cancer (%)	Cancers per 1,000 screening mammograms
1: Normal	356,030	355,734 (76.2)	296 (12.3)	0.83
2: Benign finding	56,614	56,533 (12.1)	81 (3.4)	1.43
3: Probably benign with a recommendation for Normal or Short-term follow-up	8692	8,627 (1.8)	65 (2.7)	7.48
3+: Probably benign with a recommendation for immediate work-up	3094	3,049 (0.7)	45 (1.9)	14.54
0: Need additional imaging evaluation	42,823	41,442 (8.9)	1,381 (57.5)	32.25
4: Suspicious abnormality, biopsy should be considered	2022	1,687 (0.4)	335 (13.9)	165.68
5: Highly suggestive of malignancy	237	38 (0.0)	199 (8.3)	839.66

The use of the BI-RADS ratings shown in Table 4.5 has been criticized⁴¹ in an editorial titled “*BI-RADS Data Should Not Be Used to Estimate ROC Curves*”. Since BI-RADS is a clinical rating scheme widely used in mammography, the editorial, if correct, implies that ROC analysis of clinical mammography data is not possible. Since the BI-RADS scale was arrived at after considerable deliberation, inability to perform ROC analysis with it would strike at the root of clinical utility of the ROC method. The purpose of this section is to express the reasons why the author has a different take on this controversy.

It is claimed in the editorial⁴¹ that the Barlow et al. method confuses cancer yield with confidence level and that BI-RADS categories 1 and 2 should not be separate entries of the confidence scale, because both indicate no suspicion for cancer.

The author agrees with the Barlow et al. suggested ordering of the "2s" as more likely to have cancer than the "1s". A category-2 means the radiologist found *something to report*, and the *location* of the finding is part of the clinical report. Even if the radiologist believes the finding is definitely benign, there is a *finite* probability that a category-2 finding is cancer, as evident in the last column of Table 4.5 ($1.43 > 0.83$). In contrast, there are no findings associated with a category-1 report. [Independent of the Barlow et al. study, a paper⁴² titled "*Benign*

breast disease and the risk of breast cancer" should convince any doubters that benign lesions do have a finite chance of cancer]

The problem with “where to put the 0s” arises only when one tries to analyze clinical BI-RADS data. In a laboratory study, the radiologist would not be given the category-0 option. In analyzing a clinical study it is incumbent on the study designer to justify the choice of the rating scale adopted. Showing that the proposed ordering agrees with the probability of cancer is justification – and in the author’s opinion, given the very large sample size this was accomplished convincingly in the Barlow et al. study. Moreover, the last column of Table 4.5 suggests that any other ordering would violate an important principle, namely, optimal ordering is achieved when each case is rated according to its *likelihood ratios*: defined as the probability of the case being diseased divided by the probability of the case being non-diseased. The likelihood ratio is the "betting odds" of the case being diseased, which is expected to be monotonic with the empirical probability of the case being diseased, i.e., the last column of Table 4.5. Therefore, the ordering adopted in Table 4.5 is equivalent to adopting a likelihood ratio scale and any other ordering would not be monotonic with likelihood ratio.

The likelihood ratio is described in more detail in the **Chapter 20**, which describes ROC fitting methods that yield "proper" ROC curves, i.e., ones that have monotonically decreasing slope as the operating point moves up the curve from (0,0) to (1,1) and therefore do not (inappropriately) cross the chance diagonal. Key to these fitting methods is adoption of a likelihood ratio scale to rank-order cases, instead of the ratings assumed by the unequal variance binormal model. The proper ROC fitting algorithm implemented in PROPROC software *reorders* confidence levels assumed by the binormal model, **Chapter 20**, paragraph following Fig. 20.4. This is analogous to the reordering of the clinical ratings based on cancer rates assumed in Table 4.5. It is illogical to allow reordering of ratings in "blind" software but question the same when done in a principled way by a researcher. As expected, the modeled ROC curves in the Barlow publication, their Fig. 4, show no evidence of improper behavior. This is in contrast to a clinical study (about fifty thousands patients spread over 33 hospitals with each mammogram interpreted by two radiologists) using a non-BIRADS 7-point rating scale which yielded markedly improper ROC curves⁴³ for the film modality when using ROC ratings (not BIRADS). This suggests that use of a non-clinical ratings scale for clinical studies, without independent confirmation of the ordering implied by the scale, may be problematical.

The reader might be interested as to reason for the 0-ratings being more predictive of cancer than a 3+ rating, Table 4.5. In the clinic the zero rating implies, in effect, "*defer decision, incomplete information, additional imaging necessary*". A zero rating could be due to technical problems with the images: e.g., improper

positioning (e.g., missing breast tissue close to the chest wall) or incorrect imaging technique (improper selection of kilovoltage and/or tube charge), making it impossible to properly interpret the images. Since the images are part of the permanent patient record, there are both healthcare and legal reasons why the images need to be optimal. Incorrect technical factors are expected to occur randomly and *not* be predictive of cancer. However, if there is a suspicious finding and the image quality is sub-optimal, the radiologist may be unable to commit to a decision, they may seek additional imaging, perhaps better compression or a slightly different view angle to resolve the ambiguity. Such zero ratings are expected with suspicious findings, and therefore are expected to be more predictive of cancer than pure technical reason zero ratings.

[As an aside, the 2nd paper³⁹ using the ordering shown in Table 4.5 questioned the utility of CAD for breast cancer screening (this was ca. 2007). This paper was met with flurry of correspondence⁴⁴⁻⁵¹ disputing the methodology (summarized above). The finding regarding utility of CAD has been validated by more recent studies, again with very large case and reader samples, showing that usage of CAD can actually be detrimental to patient outcome⁵² and a call¹⁹ for ending insurance reimbursement for CAD.]

4.12: Discussion

In this chapter the widely used ratings paradigm was described and illustrated with a sample dataset, Table 4.1. The calculation of ROC operating points from this table was detailed. A formal notation was introduced to describe the counts in this table and the construction of operating points and an **R** example was given. The author does not wish to leave the impression that the ratings paradigm is used only in medical imaging. In fact the historical reference³ to the two-question six-point scale in Fig. 4.2, namely Table 3.1 in the book by MacMillan and Creelman, was for a rating study on performance in recognizing odors. The early users of the ROC ratings paradigm were mostly experimental psychologists and psychophysicists interested in studying perception of signals, some in the auditory domain, and some in other sensory domains.

While it is possible to use the equal variance binormal model to obtain a measure of performance, the results depend upon the choice of operating point, and evidence was presented for the generally observed fact that most ROC ratings datasets are inconsistent with the equal variance binormal model. This indicates the need for an extended model, to be discussed in **Chapter 06**.

The rating paradigm is a more efficient way of collecting the data compared to repeating the binary paradigm with instructions to cause the observer to adopt different fixed thresholds specific to each repetition. The rating paradigm is also more efficient 2AFC paradigm; more importantly, it is more clinically realistic.

Two controversial but important issues were addressed: the reason for the author's recommendation for adopting a discrete 6-point rating scale, and correct usage of clinical BIRADS ratings in ROC studies. When a clinical scale exists, the empirical disease occurrence rate associated with each rating should be used to order the ratings. Ignoring an existing clinical scale would be a disservice to the radiology community.

The next step is to describe a model for ratings data. Before doing that, it is necessary to introduce an empirical performance measure, namely the area under the empirical or trapezoidal ROC, which does not require any modeling.

4.13: References

-
-
1. Barnes G, Sabbagh E, Chakraborty D, et al. A comparison of dual-energy digital radiography and screen-film imaging in the detection of subtle interstitial pulmonary disease. *Investigative radiology*. 1989;24(8):585-591.
 2. Nishikawa R. Estimating sensitivity and specificity in an ROC experiment. *Breast Imaging*. 2012:690-696.
 3. Macmillan NA, Creelman CD. *Detection Theory: A User's Guide*. New York: Cambridge University Press; 1991.
 4. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons; 1966.
 5. Burgess AE. Visual Perception Studies and Observer Models in Medical Imaging. *Seminars in Nuclear Medicine*. 2011;41(6):419-436.
 6. Burgess AE. On the noise variance of a digital mammography system. *Medical Physics*. 2004;31(7):1987-1995.
 7. Burgess AE, Judy PF. Detection in power-law noise: spectrum exponents and CD diagram slopes. *Proc SPIE*. 2003;5034:57-62.
 8. Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. *Med Phys*. 2001;28(4):419-437.

9. Burgess AE. Evaluation of detection model performance in power-law noise. *Proc SPIE*. 2001;4324:419-437.
10. Burgess AE, Jacobson FL, Judy PF. On the detection of lesions in mammographic structure. *SPIE*. 1999;3663(Medical Imaging):304-315.
11. Burgess AE, Chakraborty S. Producing lesions for hybrid mammograms: Extracted tumours and simulated microcalcifications. *Proc of SPIE*. 1999;3663:316-321.
12. Chakraborty DP, Kundel HL. Anomalous nodule visibility effects in mammographic images. Paper presented at: Proc. SPIE Medical Imaging 2001: Image Perception and Performance 2001.
13. Burgess AE, Li X, Abbey CK. Visual signal detectability with two noise components: anomalous masking effects. *Journal Opt Soc Am A*. 1997;14(9):2420-2442.
14. Bochud FO, Abbey CK, Eckstein MP. Visual Signal Detection in structured backgrounds IV, Calculation of Figures of Merit for Model Observers in Non-Stationary Backgrounds. *Journal of the Optical Society of America, A, Optics, Image Science, & Vision*. 1999;17(2):206-217.
15. Bochud FO, Abbey CK, Eckstein MP. Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically non-stationary backgrounds. *J Opt Soc Am A*. 2000;17(2):193-216.
16. Eckstein M, FO B, Abbey C. Visual signal detection in structured backgrounds IV. Figures of merit for model observers in multiple alternative forced choice with response correlations. *Journal of the Optical Society of America A*. 2000;17:206-217.
17. Eckstein MP, Abbey CK, Bochud FO. A Practical Guide to Model Observers for Visual Detection in Synthetic and Natural Noisy Images. In: Kundel H, Beutel J, Van-Metter R, eds. *Handbook of Medical Imaging*. Bellingham, Washington: SPIE; 2000:593-628.
18. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J Optical Soc Am A*. 2001;18(3):473-488.
19. Abbey CK, Eckstein MP. Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision*. 2002;2(1):66-78.
20. Bochud FO, Abbey CK, Eckstein MP. Search for lesions in mammograms: Statistical characterization of observer responses. *Med Phys*. 2004;31(1):24-36.
21. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med Phys*. 1995;22(5):643-655.

22. Gur D, Bandos AI, Cohen CS, et al. The "Laboratory" Effect: Comparing Radiologists' Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations. *Radiology*. 2008;249(1):47-53.
23. Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE. The Prevalence Effect in a Laboratory Environment: Changing the Confidence Ratings. *Acad Radiol*. 2007;14:49–53.
24. Gur D, Rockette HE, Armfield DR, et al. Prevalence Effect in a Laboratory Environment. *Radiology*. 2003;228:10-14.
25. Soh BP, Lee W, McEntee MF, et al. Screening mammography: test set data can reasonably describe actual clinical reporting. *Radiology*. 2013;268(1):46-53.
26. Rockette HE, Gur D, Metz CE. The Use of Continuous and Discrete Confidence Judgments in Receiver Operating Characteristic Studies of Diagnostic Imaging Techniques. *Investigative Radiology*. 1992;27:169-172.
27. Wagner RF, Beiden SV, Metz CE. Continuous versus Categorical Data for ROC Analysis: Some Quantitative Considerations. *Academic Radiology*. 2001;8(4):328-334.
28. Metz CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*. 1998;17(9):1033-1053.
29. Miller GA. The Magical Number Seven, Plus or Minus Two: Some limits on our capacity for processing information. *The Psychological Review*. 1956;63(2):81-97.
30. Pollack I. The information of elementary auditory displays. *The Journal of the Acoustical Society of America*. 1952;24(6):745-749.
31. Pollack I. The information of elementary auditory displays. II. *The Journal of the Acoustical Society of America*. 1953;25(4):765-769.
32. Garner W. An informational analysis of absolute judgments of loudness. *Journal of experimental psychology*. 1953;46(5):373.
33. Beebe-Center JG, Rogers M, O'connell D. Transmission of information about sucrose and saline solutions through the sense of taste. *The Journal of Psychology*. 1955;39(1):157-160.
34. Hake HW, Garner W. The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal of experimental psychology*. 1951;42(5):358.
35. Berbaum KS, Dorfman DD, Franken EA, Caldwell RT. An Empirical Comparison of Discrete Ratings and Subjective Probability Ratings. *Academic Radiology*. 2002;9(7):756-763.

36. D'Orsi CJ, Bassett LW, Feig SA, et al. *Illustrated Breast Imaging Reporting and Data System*. Reston, Va: American College of Radiology; 1998.
37. D'Orsi CJ, Bassett LW, Berg WA. *ACR BI-RADS-Mammography*. 4th ed. Reston, Va: American College of Radiology; 2003.
38. Barlow WE, Chi C, Carney PA, et al. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. *Journal of the National Cancer Institute*. 2004;96(24):1840-1850.
39. Fenton JJ, Taplin SH, Carney PA, et al. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N Engl J Med*. 2007;356(14):1399-1409.
40. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine*. 1996;156(2):209-213.
41. Jiang Y, Metz CE. BI-RADS Data Should Not Be Used to Estimate ROC Curves. *Radiology*. 2010;256(1):29-31.
42. Hartmann LC, Sellers TA, Frost MH, et al. Benign breast disease and the risk of breast cancer. *New England Journal of Medicine*. 2005;353(3):229-237.
43. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*. 2005;353(17):1-11.
44. Ciatto S, Houssami N. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
45. Feig SA, Birdwell RL, Linver MN. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
46. Fenton J, J, Barlow WE, Elmore JG. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
47. Gur D. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
48. Nishikawa RM, Schmidt RA, Metz CE. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
49. Ruiz JF. Computer-Aided Screening Mammography. *New England Journal of Medicine*. 2007;357(1):83-85.
50. Berry DA. Computer-Assisted Detection and Screening Mammography: Where's the Beef? *Journal of the National Cancer Institute*. 2011.

51. Nishikawa RM, Giger ML, Jiang Y, Metz CE. Re: Effectiveness of Computer-Aided Detection in Community Mammography Practice. *Journal of the National Cancer Institute*. 2012;104(1):77.
52. Philpotts LE. Can Computer-aided Detection Be Detrimental to Mammographic Interpretation? *Radiology*. 2009;253(1):17-22.