Chapter 2: The binary paradigm

Table of contents

2.1: Introduction

In the previous chapter four observer performance paradigms were introduced: the receiver operating characteristic (ROC), the free-response ROC (FROC), the location ROC (LROC) and the region of interest (ROI). In the chapters comprising this section, i.e., **Chapter 02 - Chapter 07**, focus is on the ROC paradigm, where each case is rated for confidence in presence of disease. While a multiple point rating scale is generally used, *in this chapter it is assumed that the ratings are binary*, and the allowed values are "1" vs. "2". Equivalently, the ratings could be "non-diseased" vs. "diseased", "negative" vs. "positive", etc. In the literature this method of data acquisition is also termed the "yes/no" procedure[1,2]. The reason for restricting, for now, to the binary task is that the multiple rating task can be shown to be equivalent to a number of simultaneously conducted binary tasks. So understanding the simpler method is a good starting point.

Since the truth is also binary, this chapter could be named the *binary-truth binary-decision* task. The starting point is a 2 x 2 table summarizing the outcomes in such studies and useful fractions that can be defined from the counts in this table, the most important ones being true positive fraction (TPF) and false positive fraction (FPF). These are used to construct measures of performance, some of which are desirable from the researcher's point of view, but others are more relevant to radiologists. The concept of disease prevalence is introduced and used to formulate relations between the different types of measures. An **R** example of calculation of these quantities is given that is only slightly more complicated than the demonstration in the prior chapter.

2.2: Decision vs. truth: the fundamental 2x2 table of ROC analysis

> In this book, the term *case* is used for images obtained for diagnostic purposes, of a patient; often multiple images of a patient, sometimes from different modalities, are involved in an interpretation; all images of a single patient, that are used in the interpretation, are collectively referred to as a case. A familiar example is the 4-view presentation used in screening mammography, where two views of each breast are available for viewing.

Let *D* represent the radiologist's *decision*, with $D=1$ representing the decision "case is non-diseased" and $D=2$ representing the decision "case is diseased". Let *T* denote the *truth* with $T=1$ representing "case is actually non-diseased" and $T=2$ representing "case is actually diseased". It is assumed that, prior to the interpretation, the radiologist does not know the truth state of the case and the decision is based on information contained in the case. Each decision, one of two values, will be associated with one of two truth states, resulting in an entry in one of 4 cells arranged in a 2 x 2 layout, termed the *decision vs. truth table*, Table 2.1, *which is of fundamental importance in observer performance*. The cells are labeled as follows. The abbreviation *TN*, for true negative, represents a $D=1$ decision on a $T=1$ case. Likewise, *FN*, for false negative, represents a $D=1$ decision on a $T=2$ case (also termed a "miss"). Similarly, *FP*, for false positive, represents a $D=2$ decision on a $T=1$ case (a "false-alarm") and *TP*, for true positive, represents a $D=2$ decision on a $T=2$ case (a "hit").

Table 2.1: The decision vs. truth table: the fundamental 2x2 table of observer performance, showing the classification of decisions in the binary task.

Radiologist's decision D	**Case truth T**	
	Case is actually non-diseased; *T* =1	Case is actually diseased; *T* =2
"Case is diagnosed non-diseased" D=1	**TN**	**FN ("miss")**
"Case is diagnosed diseased" D=2	**FP ("false alarm")**	**TP ("hit")**

Table 2.2 shows the *numbers* (indicated by the hash symbol prefix) of decisions in each of the four categories defined in Table 2.1. Specifically, #TN is the number of true negative decisions, #FN is the number of false negative decisions, etc. The last row is the sum of the corresponding columns. The sum of the number of true negative decisions (#TN) and the number of false positive decisions (#FP) must equal the total number of non-diseased cases, denoted K_1. Likewise, the sum of the number of false negative decisions (#FN) and the number

of true positive decisions (#TP) must equal the total number of diseased cases, denoted K_2. The last column is the sum of the corresponding rows. The sum of the number of true negative (#TN) and false negative (#FN) decisions is the total number of negative decisions, denoted #N. Likewise, the sum of the number of false positive (#FP) and true positive (#TP) decisions is the total number of positive decisions, denoted #P. Since each case yields a decision, the bottom-right corner cell is #N + #P, which must also equal $K_1 + K_2$, the total number of cases *K*. These statements are summarized in Eqn. (2.1).

$$\left.\begin{aligned}
&K_1 = \#TN + \#FP\\
&K_2 = \#FN + \#TN\\
&\#N = \#TN + \#FN\\
&\#P = \#TP + \#FP\\
&K = K_1 + K_2 = \#N + \#P
\end{aligned}\right\} \qquad . \tag{2.1}$$

Table 2.2: Decision vs. truth table, showing total counts in the different cells; the last row/column show the totals of the corresponding columns/rows [# denotes the number of counts of the corresponding cell].

Radiologist's decision: D	**Case truth: T**		Row totals
	Case is non-diseased: T=1	Case is diseased: T=2	
"Case is non-diseased": D=1	#TN	#FN	#N=#TN+#FN
"Case is diseased": D=2	#FP	#TP	#P=#FP+#TP
Column totals	$K_1 = \#TN + \#FP$	$K_2 = \#FN + \#TP$	$K_1 + K_2 =$ #N + #P

2.3: Sensitivity and specificity

The notation $P(D|T)$ indicates the *probability of diagnosis D given truth state T* (the vertical bar symbol is used to denote a *conditional probability*, i.e., what is to the left of the vertical bar depends on the *condition* appearing to the right of the vertical bar being true).

$$P(D|T) = P(\text{patient diagnosis is } D | \text{patient truth is } T) \qquad . \tag{2.2}$$

Therefore the probability that the radiologist will diagnose "case is diseased" when the case is actually diseased is $P(D=2|T=2)$, which is the probability of a true positive $P(TP)$.

$$P(TP) = P(D=2|T=2) \qquad . \tag{2.3}$$

Likewise, the probability that the radiologist will diagnose "case is non-diseased" when the case is actually diseased is $P(D=1|T=2)$, which is the probability of a false negative $P(FN)$.

$$P(FN)=P(D=1|T=2) \quad . \tag{2.4}$$

The corresponding probabilities for non-diseased cases, $P(TN)$ and $P(FP)$, are defined by:

$$\left.\begin{array}{l} P(TN)=P(D=1|T=1) \\ P(FP)=P(D=2|T=1) \end{array}\right\} \quad . \tag{2.5}$$

Since the diagnosis must be either $D=1$ or $D=2$, for each truth state the probabilities on non-diseased and diseased cases must sum to unity:

$$\left.\begin{array}{l} P(D=1|T=1)+P(D=2|T=1)=1 \\ P(D=1|T=2)+P(D=2|T=2)=1 \end{array}\right\} \quad . \tag{2.6}$$

Equivalently, these equations can be written:

$$\left.\begin{array}{l} P(TN)+P(FP)=1 \\ P(FN)+P(TP)=1 \end{array}\right\} \quad . \tag{2.7}$$

Comments:

1. An easy way to remember Eqn. (2.7) is to start by writing down the probability of one of the four probabilities, e.g., $P(TN)$, and “reversing” both terms inside the parentheses, i.e., T → F, and N → P. This yields the term $P(FP)$ which when added to the previous probability, $P(TN)$, yields unity, i.e., the 1st equation in Eqn. (2.7).
2. Because there are two equations in four unknowns, only two of the four probabilities, one per equation, are independent. By tradition these are chosen to be $P(D=1|T=1)$ and $P(D=2|T=2)$, i.e., $P(TN)$ and $P(TP)$, which happen to be the probabilities of correct decisions on non-diseased and diseased

cases, respectively. The two basic probabilities are so important that they have names: $P(D=2|T=2)$ $= P(TP)$ is termed *sensitivity* (*Se*) and $P(D=1|T=1) = P(TN)$ is termed *specificity* (*Sp*):

$$\left.\begin{aligned} Se = P(TP) = P(\mathrm{D}=2|\mathrm{T}=2) \\ Sp = P(TN) = P(\mathrm{D}=1|\mathrm{T}=1) \end{aligned}\right\} . \qquad \textbf{(2.8)}$$

The radiologist can be regarded as a diagnostic "test" yielding a binary decision under the binary truth condition. More generally, any test (e.g., a blood test for HIV) yielding a binary result (positive or negative) under a binary truth condition is said to be *sensitive* if it correctly detects the diseased condition most of the time. The test is said to be *specific* if it correctly detects the non-diseased condition most of the time. Sensitivity is how correct the test is at detecting a diseased condition, and specificity is how correct the test is at detecting a non-diseased condition.

2.4: Reasons for the names sensitivity and specificity

It is important to understand the reason for these names and an analogy may be helpful. Most of us are *sensitive* to temperature, especially if the choice is between ice-cold vs. steaming hot. The sense of touch is said to be *sensitive* to temperature. One can imagine some neurological condition rendering a person hypersensitive to temperature, such that the person responds "hot" no matter what is being touched. For such a person the sense of touch is not very *specific*, as it is unable to distinguish between the two temperatures. This person would be characterized by unit sensitivity (since the response is “hot” to all steaming hot objects) and zero specificity (since the response is never “cold” to ice-cold objects). Likewise, a different neurological condition could render a person hypersensitive to cold, and the response is "cold" no matter what is being touched. Such a person would have zero sensitivity (since the response is never “hot” when touching steaming hot) and unit specificity (since the response is “cold” when touching ice-cold). Already one suspects that there is an inverse relation between sensitivity and specificity.

2.5: Estimating sensitivity and specificity

Sensitivity and specificity are the probabilities of correct decisions, over diseased and non-diseased cases, respectively. The *true* values of these probabilities would require interpreting all diseased and non-diseased cases in the *entire population* of cases. In reality, one has a *finite sample* of cases and the corresponding quantities, calculated from this finite sample, are termed *estimates*. Population values are fixed, and in general unknown, while estimates are random variables. Intuitively, an estimate calculated over a larger number of cases is expected to be closer to the true or population value than an estimate calculated over a smaller number of cases.

Estimates of sensitivity and specificity follow from counting the numbers of TP and TN decisions in Table 2.2 and dividing by the appropriate denominators. For sensitivity, the appropriate denominator is the number of actually diseased cases, namely K_2, and for specificity the appropriate denominator is the number of actually non-diseased cases, namely K_1. The estimation equations for sensitivity specificity are (estimates are denoted by the "hat" or circumflex symbol ^):

$$\left.\begin{aligned}\widehat{Se} &= \widehat{P(TP)} = \frac{\#TP}{K_2} \\ \widehat{Sp} &= \widehat{P(TN)} = \frac{\#TN}{K_1}\end{aligned}\right\} . \qquad (2.9)$$

The ratio of the number of TP decisions to the number of actually diseased cases is termed *true positive fraction* $\widehat{TPF}$, which is an estimate of sensitivity, or equivalently, an estimate of $P(TP)$. Likewise, the ratio of the number of TN decisions to the number of actually non-diseased cases is termed *true negative fraction* $\widehat{TNF}$, which is an estimate of specificity, or equivalently, an estimate of $P(TN)$. The complements of $\widehat{TPF}$ and $\widehat{TNF}$ are termed *false negative fraction* $\widehat{FNF}$ and *false positive fraction* $\widehat{FPF}$, respectively, Table 2.3.

Table 2.3: This table shows estimates of two selected probabilities, sensitivity and specificity, in the binary decision task. The two other probabilities are the complements of these values. The probabilities follow from dividing the numbers of counts from Table 2.2 by the appropriate denominators.

Radiologist's decision: D	**Case truth: T**	
	T = 1	T = 2
D = 1	$\widehat{Sp} = \widehat{TNF} = \widehat{P(TN)} = \frac{\#TN}{K_1}$	$1-\widehat{Se} = \widehat{FNF} = \widehat{P(FN)} = \frac{\#FN}{K_2}$
D = 2	$1-\widehat{Sp} = \widehat{FPF} = \widehat{P(FP)} = \frac{\#FP}{K_1}$	$\widehat{Se} = \widehat{TPF} = \widehat{P(TP)} = \frac{\#TP}{K_2}$

2.6: Disease prevalence

Disease prevalence, often abbreviated to *prevalence*, is defined as the *actual* or true probability that a randomly sampled case is of a diseased patient, i.e., the fraction of the entire population that is diseased. It is denoted, $P\left(D\middle|pop\right)$ when patients are randomly sampled from the population ("*pop*") and otherwise it is denoted $P\left(D\middle|lab\right)$, where the condition "*lab*" stands for a laboratory study, where cases may be artificially enriched, and thus not representative of the population value:

$$\left.\begin{array}{l} P\left(D\middle|pop\right)=P\left(T=2\middle|pop\right) \\ P\left(D\middle|lab\right)=P\left(T=2\middle|lab\right) \end{array}\right\} . \qquad \mathbf{(2.10)}$$

Since the patients must be either diseased on non-diseased, it follows with either sampling method, that:

$$\left.\begin{array}{l} P\left(T=1\middle|pop\right)+P\left(T=2\middle|pop\right)=1 \\ P\left(T=1\middle|lab\right)+P\left(T=2\middle|lab\right)=1 \end{array}\right\} . \qquad \mathbf{(2.11)}$$

If a finite number of patients are sampled randomly from the population (not true in most laboratory studies), then the fraction of diseased patients in the sample is an estimate of *true* disease prevalence.

$$\widehat{P\left(D\middle|pop\right)}=\left.\frac{K_2}{K_1+K_2}\right|_{pop} . \qquad \mathbf{(2.12)}$$

It is important to appreciate the distinction between *true* (population) prevalence and *laboratory* prevalence. As an example, true disease prevalence for breast cancer is about five per 1000 patients in the US, but most mammography studies are conducted with comparable numbers of non-diseased and diseased cases:

$$\left.\begin{array}{l} \widehat{P\left(D\middle|pop\right)}\sim 0.005 \\ \widehat{P\left(D\middle|lab\right)}\sim 0.5 >> \widehat{P\left(D\middle|pop\right)} \end{array}\right\} . \qquad \mathbf{(2.13)}$$

2.7: Accuracy

Accuracy is defined as the fraction of all decisions that are in fact correct. Denoting it by Ac one has for the corresponding estimate:

$$\widehat{Ac} = \frac{\#TN + \#TP}{\#TN + \#TP + \#FP + \#FN} \quad . \tag{2.14}$$

The numerator is the total number of correct decisions and the denominator is the total number of decisions. An equivalent expression is:

$$\widehat{Ac} = \widehat{Sp}\ \widehat{P(!D)} + \widehat{Se} \times \widehat{P(D)} \quad . \tag{2.15}$$

The exclamation mark symbol is used to denote the "*not*" or *negation* operator. For example, $P(!D)$ means the probability that the patient is not diseased. Eqn. (2.15) applies equally to laboratory or population studies, *provided sensitivity and specificity are estimated consistently*. In other words, one cannot combine a population estimate of prevalence with a laboratory measurement of sensitivity and / or specificity.

Eqn. (2.15) can be understood from the following argument. $\widehat{Sp}$ is the fraction of correct (i.e., negative) decisions on non-diseased cases. Multiplying this by $\widehat{P(!D)}$ yields $\widehat{Sp}\ \widehat{P(!D)}$, the fraction of correct negative decisions on all cases. Similarly, $\widehat{Se} \times \widehat{P(D)}$ is the fraction of correct positive decisions on all cases. Therefore, their sum is the fraction of (all, i.e., negative and positive) correct decisions on *all* cases. A formal mathematical derivation follows. The terms on the right hand side of Eqn. (2.9) can be "turned around" yielding:

$$\begin{aligned} \#TP &= K_2\,\widehat{Se} \\ \#TN &= K_1\,\widehat{Sp} \end{aligned} \quad . \tag{2.16}$$

Therefore,

$$\widehat{Ac} = \frac{\#TN + \#TP}{K} = \frac{K_1\,\widehat{Sp} + K_2\,\widehat{Se}}{K} = \widehat{Sp}\,\widehat{P(!D)} + \widehat{Se}\,\widehat{P(D)} \quad . \tag{2.17}$$

■

2.8: Positive and negative predictive values

Sensitivity and specificity have desirable characteristics, insofar as they reward the observer for correct decisions on actually diseased and actually non-diseased cases, respectively, so these quantities are expected to be independent of disease prevalence. Stated simply, one is dividing by the relevant denominator, so increased numbers of non-diseased cases are balanced by a corresponding increased number of correct decisions on non-diseased cases, and likewise for diseased cases. However, radiologists interpret cases in a “mixed” situation where cases could be positive or negative for disease and disease prevalence plays a crucial role in their decision-making – this point will be clarified shortly. Therefore, a measure of performance that is desirable from the researcher's point of view is not necessarily desirable from the radiologist's point of view. It should be obvious that if most cases are non-diseased, i.e., disease prevalence is close to zero, specificity, being correct on non-diseased cases, is more important to the radiologist. Otherwise, the radiologist would figuratively be crying "wolf" most of the time. The radiologist who makes too many FPs would discover it from subsequent clinical audits or daily case conferences, which are held in most large imaging departments. There is a cost to unnecessary false positives – the cost of additional imaging and / or needle-biopsy to rule out cancer, not to mention the pain and emotional trauma inflicted on the patient. Conversely, if disease prevalence is high, then sensitivity, being correct on diseased cases, is more important to the radiologist. With intermediate disease prevalence a weighted average of sensitivity and specificity, where the weighting involves disease prevalence, is desirable from the radiologist's point of view.

The radiologist is less interested in the *normalized* probability of a correct decision on non-diseased cases. Rather interest is in the probability that a patient diagnosed as non-diseased is actually non-diseased. The reader should notice how the two probability definitions are "turned around" - more on this below. Likewise, the radiologist is less interested in the *normalized* probability of correct decisions on diseased cases; rather interest is in the probability that a patient diagnosed as diseased is actually diseased. These are termed *negative and positive predictive values*, respectively, and denoted *NPV* and *PPV*.

Let us start with *NPV*, defined as the probability, given a non-diseased diagnosis, that the patient is actually non-diseased:

$$NPV = P\left(T=1 \mid D=1\right) \quad . \qquad \textbf{(2.18)}$$

Note that this equation is "turned around" from the definition of specificity, Eqn. (2.8), repeated below for ease of comparison:

$$Sp = P\left(D = 1 \mid T = 1\right) \quad . \qquad \textbf{(2.19)}$$

To estimate *NPV*, one divides the number of correct negative decisions (#TN) by the total number of negative decisions (#N). The latter is the sum of the number of correct negative decisions (#TN) and the number of incorrect negative decisions (#FN). Therefore,

$$\widehat{NPV} = \frac{\# TN}{\# TN + \# FN} \quad . \qquad \textbf{(2.20)}$$

Dividing the numerator and denominator by the total number of cases K , one gets:

$$\widehat{NPV} = \frac{\widehat{P_K\left(TN\right)}}{\widehat{P_K\left(TN\right)} + \widehat{P_K\left(FN\right)}} \quad . \qquad \textbf{(2.21)}$$

The estimate $\widehat{P_K\left(TN\right)}$ of the probability of a TN *over all cases* (hence the subscript K) equals the estimate of true negative fraction $\left(1 - \widehat{FPF}\right)$ multiplied by the estimate that the patient is non-diseased, i.e., $\widehat{P\left(!D\right)}$:

$$\widehat{P_K\left(TN\right)} = \widehat{P\left(!D\right)}\left(1 - \widehat{FPF}\right) \quad . \qquad \textbf{(2.22)}$$

Explanation: A similar logic to that used earlier applies: $\left(1 - \widehat{FPF}\right)$ is the probability of being correct on non-diseased cases. Multiplying this by the estimate of probability of disease *absence* yields the estimate of $\widehat{P_K\left(TN\right)}$.

Likewise, the estimate $\widehat{P_K(FN)}$ of the probability of a FN over all cases equals the estimate of false negative fraction, which is $\left(1-\widehat{TPF}\right)$, multiplied by the estimate of the probability that the patient is diseased, i.e., $\widehat{P(D)}$:

$$\widehat{P_K(FN)}=\widehat{P(D)}\left(1-\widehat{TPF}\right) \quad . \qquad \textbf{(2.23)}$$

Putting this all together, one has:

$$\widehat{NPV}=\frac{\widehat{P(!D)}\left(1-\widehat{FPF}\right)}{\widehat{P(!D)}\left(1-\widehat{FPF}\right)+\widehat{P(D)}\left(1-\widehat{TPF}\right)} \quad . \qquad \textbf{(2.21)}$$

For the population,

$$NPV=\frac{P(!D)(1-FPF)}{P(!D)(1-FPF)+P(D)(1-TPF)} \quad . \qquad \textbf{(2.22)}$$

Likewise, it can be shown that PPV is given by

$$PPV=\frac{P(D)\times TPF}{P(D)\times TPF+P(!D)\times FPF} \quad . \qquad \textbf{(2.23)}$$

In words,

$$negative\,predictive\,value=\frac{(1-prevalence)(specificity)}{(1-prevalence)(specificity)+(prevalence)(1-sensitivity)} \quad . \qquad \textbf{(2.24)}$$

$$positive\,predictive\,value=\frac{(prevalence)(sensitivity)}{(prevalence)(sensitivity)+(1-prevalence)(1-specificity)} \quad . \qquad \textbf{(2.25)}$$

The equations defining NPV and PPV are actually special cases of Bayes' theorem[3]. The general theorem is:

$$P\left(A\middle|B\right)=\frac{P\left(B\middle|A\right)P\left(A\right)}{P\left(B\right)}=\frac{P\left(A\right)P\left(B\middle|A\right)}{P\left(A\right)P\left(B\middle|A\right)+P\left(!A\right)P\left(B\middle|!A\right)} \quad . \qquad \textbf{(2.26)}$$

An easy way to remember Eqn. (2.26) is to start with the numerator, which is the "reversed" form of the desired probability on the left hand side, multiplied by an appropriate probability. For example, if the desired probability is $P\left(A\middle|B\right)$, one starts with the "reversed" form, i.e., $P\left(B\middle|A\right)$, multiplied by $P\left(A\right)$. This yields the numerator. The denominator is the sum of two probabilities: the probability of *B* given *A*, i.e., $P\left(B\middle|A\right)$, multiplied by $P\left(A\right)$ plus the probability of *B* given *!A*, i.e., $P\left(B\middle|!A\right)$, multiplied by $P\left(!A\right)$.

2.9: Example: Calculation of PPV, NPV and accuracy

Typical disease prevalence in the US in screening mammography is 0.005. A typical operating point, for an expert mammographer, is FPF = 0.1, TPF = 0.8. What are NPV and PPV? While this can be done using a hand calculator, since one has **R/RStudio**, why not use it. In the online **software** folder for this chapter, open the **RStudio** project file, always named **software.Rproj** in this book, and use the **Files** menu to open **mainNpvPpv.R**, a listing of which follows:

2.9.1: Code Listing

```
# mainNpvPpv.R
rm(list = ls())
prevalence <- 0.005 # disease prevalence in US screening mammography
FPF <- 0.1 # typical operating point
TPF <- 0.8 # do:
specificity <- 1-FPF
sensitivity <- TPF
NPV <- (1-prevalence)*(specificity)/((1-prevalence)*(specificity) + prevalence*(1-sensitivity))
PPV <- prevalence*sensitivity/(prevalence*sensitivity + (1-prevalence)*(1-specificity))
cat("NPV = ", NPV, "PPV = ", PPV, "\n")
accuracy <-(1-prevalence)*(specificity)+(prevalence)*(sensitivity)
cat("accuracy = ", accuracy, "\n")
```

Line 3 initializes the variable **prevalence**, the disease prevalence. In other words, **prevalence <- 0.005** causes the value 0.005 to be assigned to the variable **prevalence**. Do not use **prevalence = 0.005** as an assignment statement: it may work some of the time, but can cause problems when one least expects it; code that works some of the time is worse, in the author's opinion, than code that never works. In **R** one does not need to worry about the type of variable - integer, float, double, or declaring variables before using them; this can lead to "sloppy" programming constructs but for the most part **R** behaves reasonably. Line 4 assigns 0.1 to **FPF** and line 5 assigns 0.8 to **TPF**. Lines 6 and 7 initialize the variables **specificity** and **sensitivity**, respectively.

Line 8 calculates **NPV**, using Eqn. (2.24) and line 9 calculates **PPV**, using Eqn. (2.25). Line 10 prints the values of **NPV** and **PPV**, with a helpful message. The **cat()** function stands for *concatenate and print the comma-separated components of the argument*. The **cat()** function starts by printing the string variable **"NPV = "**, then it encounters a comma, then the variable name **NPV**, so it prints the value of the variable. Then it encounters another comma, and the string **"PPV = "**, which it prints. Then it encounters another comma and the variable name **PPV**, so it prints the value of this variable. Finally, it encounters the last comma, and the string **"\n"**, which stand for a *newline* character, which positions any subsequent output to the next line; without it any subsequent print statements would appear on the same line, which is usually not the intent. Line 11 calculates accuracy, Eqn. (2.17) and the next line prints it. Click on the **Source** button (in future this will be abbreviated to "**source** the code") on the top-right corner of the source-file window; one gets the following output in the **Console** window:

2.9.2: Code Output

```
> source('~/book2/A ROC analysis/A2 BinaryTask/software/mainNpvPpv.R')
NPV =  0.9988846 PPV =  0.03864734
accuracy =  0.8995
```

If a woman has a negative diagnosis, chances are very small that she has breast cancer: the probability that the radiologist is incorrect in the negative diagnosis is 1 - NPV = 0.00111. Even is she has a positive diagnosis, the probability that she actually has cancer is still only 0.039. That is why following a positive screening diagnosis the woman is recalled for further imaging, and if that reveals cause for reasonable suspicion, then additional imaging is performed, perhaps augmented with a needle-biopsy to confirm actual disease status. If the biopsy turns out positive, only then is the woman referred for cancer therapy. Overall, accuracy is 90%, i.e., the radiologist is accurate! The numbers in this illustration are for expert radiologists. In practice there is wide variability in radiologist performance[4].

Consider what happens if the radiologist simply call every case negative for disease. The radiologist will be correct on all of the actually non-diseased cases but will be incorrect on all of the diseased cases. Since there are 995 non-diseased cases and 5 diseased cases, accuracy will be 0.995, higher than that achieved by the expert in the previous example. *This tells us that accuracy is not a good measure of performance*. If the radiologist responds non-diseased to every case, then both FPF and TPF will be zero. Making these changes and sourcing the code one gets:

2.9.3: Code Output

```
> source('~/book2/A ROC analysis/A2 BinaryTask/software/mainNpvPpv.R')
NPV =  0.995 PPV =  NaN
accuracy =  0.995
```

This confirms our expectation for accuracy. The reason **PPV** is **NaN** (Not a Number) is because one has zero *correct* positive decisions out of a *total* of zero positive decisions, leading to a 0 divided by 0 situation, Eqn. (2.23).

2.10: PPV and NPV are irrelevant to laboratory tasks

According to the hierarchy of assessment methods described in **Chapter 01**, Table 1.1, PPV and NPV are level-3 measurements, which are calculated from "live" interpretations. In the clinic, the radiologist adjusts the operating point to achieve a balance between sensitivity and specificity. *The balance depends critically on the known disease prevalence.* Based on geographical location and type of practice, the radiologist over time develops an idea of actual disease prevalence, or it can be found in various databases. For example, a breast-imaging clinic that specializes in imaging high-risk women will have higher disease prevalence than the general population and the radiologist is expected to err more on the side of reduced specificity because of the expected benefit of increased sensitivity. However, in the context of a laboratory study, where one uses enriched case sets, the concepts of NPV and PPV are meaningless. For example, it would be rather difficult to perform a laboratory study with 10,000 randomly sampled women, which would ensure about 50 actually diseased patients, which is large enough to get a reasonably precise estimate of sensitivity (estimating specificity is inherently more precise because most women are actually non-diseased). Rather, in a laboratory study one uses enriched data sets where the numbers of diseased-cases is much larger than in the general population, Eqn. (2.13). *The radiologist cannot interpret these cases pretending that the actual prevalence is very low*. Negative and positive predictive values, while they can be calculated from laboratory data, have very little, if any, clinical meanings, since they have no effect on radiologist thinking. As noted in **Chapter 01** the whole purpose of level-3 measurements is to determine the effect on radiologist thinking. There are no diagnostic decisions riding on laboratory ROC interpretations of retrospectively acquired patient images. However, PPV and NPV do have clinical meanings when calculated from very large population based "live" studies[5-7]. For example, the 2011 Fenton et al study sampled 684,956 women and used the results of "live" interpretations of their images. In contrast, laboratory ROC studies are typically conducted with 50-100 non-diseased and 50-100 diseased cases. A study using about 300 cases total would be considered a "large" ROC study.

2.11: Summary

This chapter introduced the terms sensitivity (identical to TPF), specificity (the complement of FPF), disease prevalence, and positive and negative predictive values and accuracy. It is shown that, due to its strong dependence on disease prevalence, accuracy is a relatively poor measure of performance. Radiologists generally have a good, almost visceral, understanding of positive and negative predictive values, as these terms are relevant in the clinical context, being in effect, their "batting averages". A caveat on the use of PPV and NPV calculated from laboratory studies is noted; these quantities only make sense in the context of "live" clinical interpretations.

2.12: References

1. Green DM, Swets JA. *Signal Detection Theory and Psychophysics.* New York: John Wiley & Sons; 1966.
2. Egan JP. *Signal Detection Theory and ROC Analysis.* first ed. New York: Academic Press, Inc.; 1975.
3. Larsen RJ, Marx ML. *An Introduction to Mathematical Statistics and Its Applications.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall Inc; 2001.
4. 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.
5. Barlow WE, Chi C, Carney PA, Taplin SH, D’Orsi C, Cutter G, Hendrick RE, Elmore JG. Accuracy of Screening Mammography Interpretation by Characteristics of Radiologists. *Journal of the National Cancer Institute.* 2004;96(24):1840-1850.
6. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, Berns EA, Cutter G, Hendrick RE, Barlow WE, Elmore JG. Influence of Computer-Aided Detection on Performance of Screening Mammography. *N Engl J Med.* 2007;356(14):1399-1409.
7. Fenton JJ, Abraham L, Taplin SH, Geller BM, Carney PA, D’Orsi C, Elmore JG, Barlow WE, Consortium BCS. Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer institute.* 2011;103(15):1152-1161.
8. Croft WB, Metzler D, Strohman T. *Search engines: Information retrieval in practice.* Vol 283: Addison-Wesley Reading; 2010.
9. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3:32-35.