CHAPTER IX

DIAGNOSTIC STRATEGY STUDIES

Muriel Rabilloud, René Ecochard, Gilles Landrivon

Clinical or paraclinical examinations are used:

1 - Either to find out if the patient is a carrier of a condition (for example, ultrasound for cholelithiasis), in order, if the test is positive, to consider treatment (here, surgery).

2 - Either to quantify the value of a parameter (for example, digitalinemia during digitalis treatment) in order to adapt therapy (here, the dosage of cardiac tonic).

3 - Either to visualize normal or pathological structures (for example, the vascular network before an intracranial surgical intervention, or a lithiasis of the common bile duct during a retrograde catheterization of the papilla).

4 - Or finally to determine the extent of an attack, in order to establish a prognosis without any therapeutic decision.

Most of this chapter concerns situation 1. Situations 2 and 3 are dealt with in chapter XIV on measuring instruments. Situation 4 is similar to situation 1 (notions of sensitivity and specificity, etc.) but differs from it by the absence of a decision at the end of the result.

I - THE DIAGNOSTIC TEST IN ITS CONTEXT

In this paragraph, we will set the scene, introducing the terms pre-test probability, sensitivity, specificity, post-test probability, treatment threshold and utility.

Any examination finds its place in a story, a story that begins with a symptom, or a screening examination carried out in the absence of warning signs.

Upstream of the diagnostic test there is a clinical context (age, sex, history, symptoms already present, possibly results of other examinations), which makes it possible to establish a probability of existence of the disease (before carrying out the diagnostic test studied). This probability is called the pre-test probability.

A positive result of the test will change your opinion on the patient's condition, the probability of the existence of the disease being higher in this case, lower in front of a negative result. The probability that the patient is a carrier of the disease knowing the test result, is called post-test probability.

Sometimes the probability of disease knowing that the test is positive is 100% (equal to 1). This is the case if the test has no false positives, for example histology with the presence of neoplastic tissue on a biopsy. This test has a specificity of 100% (probability that the test will be negative in the absence of cancer).

Sometimes, the probability of the disease knowing that the test is negative is zero (equal to 0). The test has completely eliminated the hypothesis of damage by the condition sought because there are no false negatives. This is the case, for example, of the abdominal scanner for the diagnosis of renal cystic mass, the latter examination having a sensitivity of 100% (probability that the test will be negative in the presence of the disease).

Most often, the post-test probability is different from 0 or 100%. It depends on the clinical context (pre-test probability) and the quality of the test. If the test is very specific and it comes back positive, the information provided by the test to affirm the presence of the disease is important and leads to a significant increase in the probability of having the disease.

The test, whether positive or negative, may not change the treatment decision. What would then be its usefulness in situation 1 defined in the introduction to the chapter (knowing whether the patient has a condition, in order to consider treatment if the test is positive)?

Indeed, for each invasive therapeutic or diagnostic action, the clinician has a probability threshold below which he abstains. A breast biopsy is not performed for ordinary mammography images in the absence of an abnormality on palpation. This threshold is most often referred to as the processing threshold. It depends on the overall benefit expected from the intervention. Each invasive therapeutic or diagnostic procedure has an overall benefit, the result of the balance between the potential improvement in the state of health and the possible side effects.

Anticoagulant therapy in the event of suspected pulmonary embolism is a decision involving the risk of bleeding and the benefit provided by the treatment. The treatment threshold is low in the case of pulmonary embolism. As soon as there is a 10 to 20% probability of pulmonary embolism (or even less), the decision is made to immobilize the patient and to treat him with anticoagulant, before the results of additional examinations. This low threshold is due on the one hand to the seriousness of the complications avoided by the treatment, and on the other hand to the relative safety of a well conducted anticoagulant treatment.

On the contrary, when the gesture is fraught with consequences, such as a total gastrectomy or an amputation, we expect a high level of probability of the existence of the disease (for example, gastritis of Ménétérier, malignant bone tumour, etc.) before planning surgery. If the uncertainty remains, we prefer, in fact, to postpone the gesture and resume the additional examinations.

If the probability of illness greatly exceeds the treatment threshold, the latter is undertaken without carrying out other paraclinical examinations, which have, in fact, no chance of changing the treatment indication. A negative result would be labeled false negative, the rest of the context being too accusatory for this examination to change the diagnosis. Faced with a red and painful subcutaneous cord, the superficial phlebitis is affirmed and the treatment is put in place. A normal phlebography would have the advantage of eliminating deep damage, but would not call into question the superficial damage which would therefore be treated. The

test concerned (here, phlebography) therefore has the advantage of showing the extension of the attack, not of changing the probability of superficial phlebitis. Indeed, negative or positive, it does not have the possibility of influencing the diagnosis enough to question the treatment. Its sensitivity is insufficient for a negative result to lower the post-test probability below the treatment threshold.

It is customary (and correct) to say that a diagnostic test should only be performed if it has a chance of moving the probability of disease to the "other side" of the treatment threshold, i.e. say to change the decision. Faced with a hormonal assessment suggestive of adrenal insufficiency, a water test is not carried out. This one has in fact no chance of changing the diagnosis because it has too many false negatives. Conversely, in the follow-up of a neurological patient, the percussion of the Achilles tendon will be used, the loss of the reflex at this level having sufficient specificity for its positivity to trigger an additional assessment.

II – EVALUATION OF A DIAGNOSTIC TEST

A – Evaluation studies of a diagnostic test

1 – The different evaluation phases

As for the evaluation of the effectiveness of a new drug, it is possible to define 3 phases in the evaluation of a new diagnostic test.

The first phase, also called the **exploratory phase**, corresponds to the early phase of evaluation of a new test. The objective is to know if this test can have a diagnostic interest. This involves, for example, verifying that a new biomarker has an average higher value in patients than in non-patients, and that it does better than mere chance in predicting the existence of the disease. At this stage, the studies carried out must make it possible to obtain a rapid response to decide whether to continue the evaluation or move on to something else.

The second phase, also called the **challenge phase**, aims to measure the diagnostic performance of a new test in different subgroups of patients and non-patients. The diagnostic performance of a test is quantified by its sensitivity and specificity for tests with a dichotomous response, or by the sensitivities and specificities associated with the different thresholds of positivity for a test with a ordinal or continuous response. It is classically said that sensitivity and specificity are the intrinsic qualities of a test because they do not involve the prevalence of the disease. Sensitivity is estimated in patients and specificity in non-patients. On the other hand, they often depend on the characteristics of the sick or the non-sick. For example, the sensitivity of mammography for diagnosing breast cancer depends on the size of the tumour. It is lower in a population of screened women than in a population of women coming to a specialist consultation at a more advanced stage of the disease. During this evaluation phase, the new test can also be compared to other existing tests.

The third phase, also called the **clinical phase**, aims to measure the diagnostic performance of a new test and compare it to other tests in the target population. This implies that the study involves a representative sample of the population in which the test will be used. For tests requiring interpretation by a reader, such as medical imaging examinations, it is also necessary to carry out the study with a representative sample of the doctors who will be required to read the examination. It is also during this phase preceding the use of the test in

clinical practice that the positivity threshold and the impact of the choice of threshold on sensitivity and specificity are studied for tests with an ordinal or continuous response.

2 – The different types of study

The main types of study found in the field of diagnostic test evaluation are case-control studies, cohort studies and randomized clinical trials.

Case-control studies

These studies are called case-control studies because when subjects enter the study, their sick or non-sick status is known. They are based on the constitution of a sample of subjects who are known to have the disease and independently of a sample of subjects who are known not to have the disease. The test to be evaluated is then measured in the group of sick subjects and in the group of non-sick subjects. This type of study is used in the exploratory phase of the evaluation of a new test. The subjects included in the sample of patients are often at a fairly advanced stage of the disease, whereas the subjects included in the sample of non-patients are often healthy subjects who have no pathology that could mimic the disease we are trying to diagnose. This often leads to an overestimation of the diagnostic performance of the test to be evaluated. This type of study is also used in the challenge phase because it makes it possible to form groups of sick subjects at different stages of the disease and groups of non-sick subjects with different characteristics, for example in terms of age or comorbidities. .

Cohort studies

These studies are called cohort-type studies because when subjects are included in the study, their diseased or non-sick status is not known. A representative sample of the population in which the test will be used is drawn up. The subjects included in the study all have the test to be evaluated and their sick or non-sick status is determined independently of the test result. Determining sick or non-sick status requires having a perfect reference test called the *gold standard*. The CASS study [1] is an example of a cohort type study. In this study, a sample of 1465 men for whom there is a suspicion of coronary artery disease was constituted. The objective of the study was to evaluate the performance of the stress test and the chest pain sought during the interrogation to make the diagnosis of coronary artery disease. All the subjects included in the study had, in addition to the 2 tests to be evaluated, a coronary angiography allowing them to be classified in the group of subjects with coronary artery disease or in the group of subjects without coronary artery disease. This type of study is mainly used in the clinical phase of the evaluation of a test. At this stage of the evaluation, it is recommended to favor multicenter studies to increase the representativeness of the sample studied.

Randomized clinical trials

When there is no perfect gold standard, the evaluation of a new test can be done by a randomized clinical trial with an arm corresponding to the usual diagnostic and therapeutic strategy and an arm which integrates the new test into the diagnostic and therapeutic strategy. In this type of study, the outcome criterion is a clinical criterion. The test will be deemed effective if the clinical result is significantly better in the arm including the new test. This type of study also makes it possible to evaluate the impact of the introduction of the test in the diagnostic and therapeutic strategy.

B – Evaluation of the performance of a diagnostic test

1 – Sensitivity and specificity

The sensitivity and specificity of a test are conditional probabilities. Sensitivity is the probability that the test will be positive (in favor of the disease) knowing that the subject is ill. This is the ability of the test to identify patients. Specificity is the probability that the test will be negative (not in favor of the disease) knowing that the subject does not have the disease. This is the ability of the test to identify non-sufferers.

Their estimate can be obtained from the results of a case-control or cohort-type study presented in the form of a table 2x2 (Table 1). There are four possible outcomes depending on the test result and disease status. The test result is positive and the subject is ill, this is a true positive (TP). The test result is negative and the subject is ill, this is a false negative (FN). The test result is negative and the subject is not sick, it is a true negative (TN). The test result is positive and the subject is not sick, this is a false positive (FP).

Table 1 - The four possible situations according to the result of the diagnostic test and the sick or not sick status

	Disease present	**Disease absent**	
Positive test	True Positive (TP)	False Positive (FP)	TP+FP
Negative test	False Negative (FN)	True Negative (TN)	FN+TN
	TP+FN	FP+TN	N

Sensitivity is estimated in patients by the proportion of positive tests: $\frac{TP}{TP + FN}$

Specificity is estimated in non-patients by the proportion of negative tests: $\frac{TN}{TN + FP}$

These are the most likely values for the sensitivity and specificity of the test given the observed data (maximum likelihood estimates). They are obtained by a vertical reading of table 2x2.

The results of the CASS study made it possible to estimate the sensitivity and specificity of chest pain for the diagnosis of coronary artery disease in a population of subjects at risk (Table 2).

Table 2 – Existence of chest pain according to the presence or absence of coronary artery disease in subjects at risk (CASS study)

	Coronary artery disease present	**Coronary artery disease absent**	
Chest pain	969	245	1214
No chest pain	54	197	251
	1023	442	1465

Chest pain sensitivity was estimated at: 969/1023 = 94,7%

Chest pain was present (positive) in approximately 95% of patients with coronary artery disease.

The specificity of chest pain was estimated at: 197/442 = 44,6%

Chest pain was absent (negative) in approximately 45% of subjects without coronary artery disease.

2 – The ROC curve and the area under the ROC curve

Three types of response can be distinguished for diagnostic tests. The response may be **dichotomous** as for chest pain. The response can be **ordinal** or **continuous quantitative**. An example of a test with an ordinal response is the BIRADS score developed by the American College of Radiology. This is a 5-level score that classifies mammograms according to the degree of suspicion of cancer. Biological markers such as PSA for the diagnosis of prostate cancer are examples of tests with a continuous quantitative response.

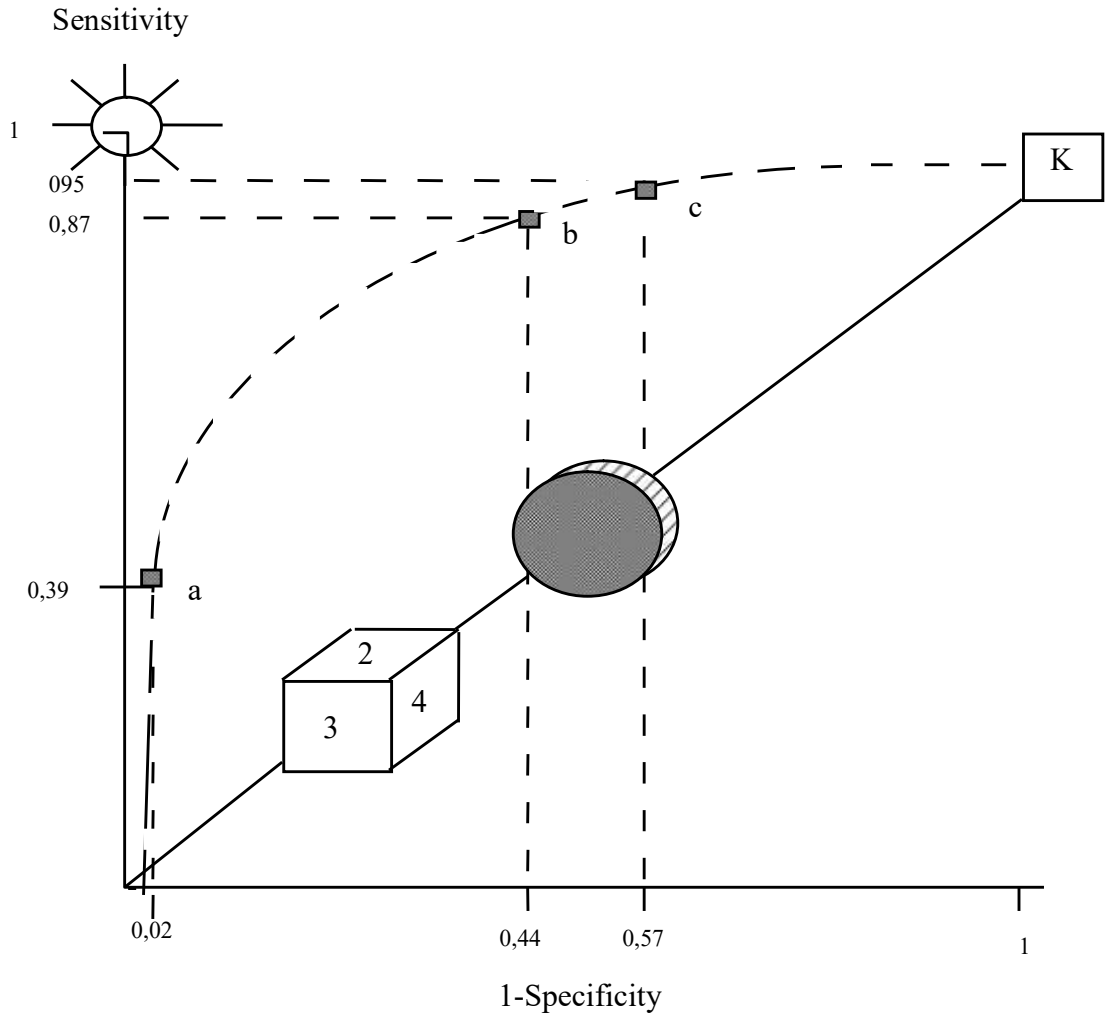
In the early phase of evaluating tests with a dichotomous response, performance is measured by sensitivity and specificity. With a continuous ordinal or quantitative response, it is not possible to summarize the diagnostic performance by estimating sensitivity and specificity. There are as many sensitivities and specificities as there are possible positivity thresholds. The ROC (Receiver Operator Characteristic) curve is used to represent the relationship between the probability that the test will be positive in patients (sensitivity) and the probability that the test will be positive in non-patients (1-specificity).

The study by Hall FM et al. [2] involved 400 women who had a breast biopsy for suspicion of cancer on mammography and normal palpation. Of these women, 119 had breast cancer. The authors reread the mammograms without knowing the result of the biopsy and classified them according to the degree of suspicion of cancer *(Table 3*). Depending on the positivity threshold chosen to classify mammograms as positive, sensitivity and specificity evolve in opposite directions. If only mammograms with a high degree of cancer suspicion are classified as positive, the sensitivity is low because there are many false negatives. On the contrary, the specificity is high because there are few false positives. The lower the chosen positivity threshold, the better the sensitivity and the worse the specificity. From the observed data *(Figure 1)*, it is possible to estimate the sensitivity (Se) and specificity (Sp) of mammography for each threshold in a population of women with suspected cancer.

Table 3 – Classification of 119 women with breast cancer and 281 women without breast cancer according to the degree of suspicion of cancer on mammography

Mammography result Degree of suspicion of cancer	**Breast cancer**	**No breast cancer**
High	47	6
Mean	57	117
Low	9	37
Minimal	6	121
Total	119	281

Figure 1 – ROC curve of mammography for the diagnosis of breast cancer in women with a positive biopsy (study by Hall FM et al)



- Mammography considered positive for high suspicions of cancer and negative for medium, mild and minimal suspicions:

(a) **Se** = 47/119 = **0,39** **1-Sp** = 1- ((281-6)/281) = 1- (275/281) = 1- 0,979 = **0,02**

- Mammography considered positive for high or medium suspicions

(b) **Se** = 104/119 = **0,87** **1-Sp** = 1- ((281- (6+117)/281) = 1- (158/281) = 1- 0,56 = **0,44**

- Mammography considered positive for high, medium or slight suspicions

(c) **Se** = 113/119 = **0,95** **1-Sp** = 1- ((281- (6+117+37)/281) = 1- (121/281) = 1- 0,43 = **0,57**

If all mammograms are considered positive, all women with cancer are well classified (Se=1), but all women without cancer are falsely classified positive (Sp=0). At the other extreme, if all mammograms are considered negative, all women without cancer are well classified (Sp=1), but all women with cancer are falsely negative (Se=0). From each couple (sensitivity, 1- specificity) estimated for the different thresholds observed, it is possible by connecting the points to construct the empirical ROC curve (Figure 1) making it possible to represent the overall diagnostic performance of mammography. If we consider that the measure of the degree of suspicion of cancer is a continuum between normality and certain cancer, the dotted curve represents the ROC curve of the continuous quantitative latent measure from which the observed ordinal response is derived.

The closer the ROC curve is to the upper left corner corresponding to a sensitivity of 1 and a specificity of 1, the better the overall performance of the test. At most, a quantitative test whose ROC curve passes through the point of sensitivity 1 and specificity 1, is a perfect gold standard. In this case, the distributions of the values in the patients and the non-patients do not overlap and all the subjects are well classified. This rarely achieved ideal is symbolized by the sun in Figure 1.

A diagnostic test whose ROC curve is on the diagonal, is a test for which the probability of having a positive response in patients is equal to the probability of having a positive response in non-patients, whatever the threshold of positivity. It doesn't do better than chance. The coin symbolizes the situation one would have by tossing a coin and deciding that the test is positive each time one lands heads and negative each time one lands tails (Se=0.5 and 1-Sp=0.5).

K symbolizes the diagnostic skills of Doctor Knock dear to Jules Romain. The doctor affirming that "everyone in good health is a patient who does not know it" has a perfect sensitivity but unnecessarily worries all the healthy. Its specificity is equal to 0.

The overall diagnostic performance of the test is all the better as the ROC curve moves away from the diagonal. It is quantified by estimating the area under the curve. A test whose ROC curve is on the diagonal and which is therefore of no diagnostic interest, has an area under the curve of 0.5. It can be interpreted as the probability that a sick subject has a higher test value than a non-sick subject, when a high test value is in favor of the disease. The test is all the better at discriminating sick from non-sick as its area under the curve approaches 1.

A nonparametric method of estimating the area under the curve consists in calculating for all the pairs (sick, not sick), the proportion of pairs for which the value of the test in the sick subject is higher than the value of the test in the non-ill subject, when a high value of the test is in favor of the disease. This is the Mann and Whitney statistic. The area under the mammography ROC curve to diagnose breast cancer in the study by Hall FM et al. is estimated at 0.81 with a 95% confidence interval between 0.76 and 0.85. Mammography does significantly better than chance because the lower limit of the confidence interval is greater than 0.5.

3 - Choice of the positivity threshold of a test

During the clinical phase of the evaluation of an ordinal or continuous diagnostic test, the determination of a positivity threshold is necessary. The optimal positivity threshold is the

one that maximizes utility in the population in which the test is used. **Utility** is defined as a measure of health state or preference for a health state; this is, for example, life expectancy weighted by quality of life. The average utility in the population depends on the utility of each of the situations (sick and treated subject, sick and untreated subject, not sick and untreated subject, not sick and treated subject) and on the frequency of each of these situations.

The average utility for the threshold c, noted U(c) is written:

$U(c) = Se*p*U_{TP} + (1-Se)*p*U_{FN} + Sp*(1-p)*U_{TN} + (1-Sp)*(1-p)*U_{FP}$

Se = sensitivity of the test
Sp = test specificity
p = disease prevalence or pre-test probability

U_{TP}, U_{FN}, U_{TN}, U_{FP} are the utilities associated with the four situations: subject sick and treated (*true positive*), subject sick and not treated (*false negative*), subject not sick and untreated (*true negative*), subject not sick and treated (*false positive*).

The average utility, U(c), can be rewritten in terms of the net benefit in terms of utility of rightly treating a sick subject and the net cost in terms of utility of wrongly treating a non-sick subject.

Methods for estimating the threshold that maximizes average utility are beyond the scope of this book and are therefore not presented. The interested reader can find these methods in the references given at the end of the chapter.

C - Evolution of the probability of having the disease at the end of the test

1 – Predictive values and Bayes' theorem

The estimation of sensitivity and specificity makes it possible to evaluate the diagnostic performance of a test, but for the clinician who will use the test, what matters are the positive and negative predictive values.

The positive predictive value (PPV) is the probability that the subject has the disease knowing that he has a positive test. The negative predictive value (NPV) is the probability that the subject does not have the disease knowing that he has a negative test. These predictive values depend on the sensitivity and specificity of the test, ***but also on the prevalence of the disease or pre-test probability***. In a cohort-type study, it is possible to estimate the predictive values directly from table 2x2 by reading the table horizontally. Let us look again at the results of the CASS study presented in *Table 2*. The sample formed for the study is *a priori* representative of the population of subjects referred for coronary angiography because of suspected coronary artery disease. The prevalence of the disease in this population can be estimated from the study data at 70% (1023/1465).

The positive predictive value of chest pain is estimated at: 969/1224 = 80%

The negative predictive value of chest pain is estimated at: 197/251 = 78%

On the other hand, case-control studies do not make it possible to directly estimate the predictive values, because they are not based on the inclusion of a representative sample of a population, but on the independent inclusion of a sample of patients and a sample of non-patients whose numbers are set by the investigator. However, it is possible to estimate the values positive and negative predictive values of the test using Bayes' theorem, provided you have 1- an estimate of the sensitivity and specificity of the test from a case-control type study, and moreover 2- an estimate of the prevalence of the disease in the population of interest.

Bayes' theorem generally makes it possible to reverse the conditional probabilities and to pass, for example, from the probability that the test is positive knowing that the subject is sick (sensitivity) to the probability that the subject has the disease knowing that the test is positive (VPP).

$$\begin{aligned}\text{VPP} = \text{P}(\text{M/Test} +) &= \frac{\text{P}(\text{M et Test} +)}{\text{P}(\text{Test} +)} = \frac{\text{P}(\text{Test} + /\text{M}) \times \text{P}(\text{M})}{\text{P}(\text{Test} + \text{et M}) + \text{P}(\text{Test} + \text{et NM})} \\ &= \frac{\text{P}(\text{Test} + /\text{M}) \times \text{P}(\text{M})}{\text{P}(\text{Test} + /\text{M}) \times \text{P}(\text{M}) + \text{P}(\text{Test} + /\text{NM}) \times \text{P}(\text{NM})} \\ &= \frac{\text{Se} \times \text{Prévalence}}{\text{Se} \times \text{Prévalence} + (1 - \text{Sp}) \times (1 - \text{Prévalence})}\end{aligned}$$

Bayes' theorem also makes it possible to go from the probability that the test is negative knowing that the subject is not sick to the probability that the subject is not sick knowing that the test is negative.

$$\begin{aligned}\text{VPN} = \text{P}(\text{NM/Test} -) &= \frac{\text{P}(\text{NM et Test} -)}{\text{P}(\text{Test} -)} = \frac{\text{P}(\text{Test} - /\text{NM}) \times \text{P}(\text{NM})}{\text{P}(\text{Test} - \text{et NM}) + \text{P}(\text{Test} - \text{et M})} \\ &= \frac{\text{P}(\text{Test} - /\text{NM}) \times \text{P}(\text{NM})}{\text{P}(\text{Test} - /\text{NM}) \times \text{P}(\text{NM}) + \text{P}(\text{Test} - /\text{M}) \times \text{P}(\text{M})} \\ &= \frac{\text{Sp} \times (1 - \text{Prévalence})}{\text{Sp} \times (1 - \text{Prévalence}) + (1 - \text{Se}) \times \text{Pr}\textit{évalence}}\end{aligned}$$

Se = sensitivity, Sp = specificity, M = diseased, NM = not diseased

To illustrate the fact that the predictive values depend a lot on the prevalence, we will take the example of the use of mammography to make the diagnosis of breast cancer in a screening situation or in a specialized consultation. According to the results of the study by Hall FM et al, and considering as positive the mammograms for which there is a high suspicion of cancer, the sensitivity is estimated at 39% and the specificity at 98%.

For a prevalence of 4 per thousand in the population of women screened between 50 and 65 years of age, the positive predictive value is:

$$\text{VPP} = \frac{0{,}39 \times 0{,}004}{0{,}39 \times 0{,}004 + (1 - 0{,}98) \times (1 - 0{,}004)} = 7{,}3\%$$

For a prevalence of 30% in the population coming for a specialist consultation, the predictive value is:

$$\text{VPP} = \frac{0{,}39 \times 0{,}3}{0{,}39 \times 0{,}3 + (1 - 0{,}98) \times (1 - 0{,}3)} = 89\%$$

The information provided by the test is the same in both cases, but the pre-test probability of the disease is very different. The positive predictive value is better in the population where the proportion of patients is greater. On the other hand, the negative predictive value is better in the population where the proportion of non-sick people is higher. It is estimated at 99.8% in the population of women screened and at 78.9% in the population of women who come for a specialist consultation.

2 – Pre and post-test probabilities and likelihood ratios

The information provided by the test depends on its sensitivity and specificity and can be quantified by the likelihood ratios. A distinction is made between the positive likelihood ratio, which corresponds to the information provided by the test when the test is positive, and the negative likelihood ratio, which corresponds to the information provided by the test when the test is negative.

The positive likelihood ratio of a test (LR+) is the ratio of the likelihood of a positive result in patients to the likelihood of a positive result in non-patients:

$$\text{RV+} = \frac{\text{P}(\text{Test} + /\text{M})}{\text{P}(\text{Test} + /\text{NM})} = \frac{Se}{1 - Sp}$$

A test that does no better than chance in discriminating sick from non-sick is a test for which the likelihood of a positive result in sick is equal to the likelihood of a positive result in non-sick. This situation corresponds to an LR+ equal to 1. The more the positive likelihood ratio is greater than 1, the more information provided by a positive test result is important.

The RV+ makes it possible to pass from the pre-test probability to the post-test probability when the test is positive. It multiplies the pre-test Odds of the disease. Let's take the example of mammography with a positive threshold corresponding to a high suspicion of cancer.

$$\text{RV} + = \frac{0{,}39}{1 - 0{,}98} = 19{,}5$$

The Odds of breast cancer in a screening situation is equal to:

$$\text{Odds pré test} = \frac{\text{prévalence}}{1 - \text{prévalence}} = \frac{0.004}{1 - 0.004} \approx 0{,}004$$

Post-test Odds when the mammogram is positive:

$$\text{Odds post test} = \text{Odds pré test} \times \text{RV} + = 0{,}004 \times 19{,}5 = 0{,}078$$

The post-test probability is equal to: $\dfrac{\text{Odds post test}}{1 + \text{Odds post test}} = 7{,}2\%$

We find the positive predictive value or probability of having the disease knowing that the test is positive. This is another way to apply Bayes' theorem.

The negative likelihood ratio (LR-) is the ratio of the likelihood of a negative result in patients to the likelihood of a negative result in non-patients:

$$\mathrm{RV}- = \frac{\mathrm{P}(\mathrm{Test}-/\mathrm{M})}{\mathrm{P}(\mathrm{Test}-/\mathrm{NM})} = \frac{1-Se}{Sp} \qquad \textit{(LR- is noted RV- in this formula)}$$

The closer the negative likelihood ratio is to 0, the more information provided by a negative test result.

The LR- (noted RV-) of the mammography is equal to: $\mathrm{RV}- = \frac{1-0{,}39}{0{,}98} = 0{,}62$

If the mammogram is negative, the Odds of the disease is divided by 1.6.

$$\text{Odds post test} = \text{Odds pré test} \times \mathrm{RV}- = 0{,}004 \times 0{,}62 \approx 0{,}0025$$

The post-test probability is equal to: $\frac{\text{Odds post test}}{1+\text{Odds post test}} \approx 2{,}5$ pour mille

The post-test probability corresponds to the probability of having the disease knowing that the test is negative. This is 1 minus the negative predictive value.

The LR+ depends mainly on the specificity of the test. The better the specificity of the test, the better the test is at confirming the presence of the disease when it is positive. The RV- depends above all on the sensitivity. The better the sensitivity, the better the test will rule out disease when negative. Let us take the example of 3 tests: gasometry in arterial blood to make the diagnosis of pulmonary embolism, culture of pleural fluid to make the diagnosis of tuberculosis and CT scan to make the diagnosis of cystic renal mass (table 4).

Table 4 – Sensitivity, specificity and positive and negative likelihood ratios of three different tests

Test	Sensitivity	Specificity	LV+	LV-
Gasometry for diagnosis of pulmonary embolism	0.95	0.5	1.9	1/10 = 0.1
Culture of pleural fluid for the diagnosis of tuberculosis	0.24	0.99	24	1/1.3 = 0.77
Scanner for the diagnosis of renal cystic mass	1	0.98	50	0

Gasometry is sensitive but not very specific for the diagnosis of pulmonary embolism. This test significantly reduces the probability of having the disease if it is negative by dividing the pre-test Odds by 10. On the other hand, it only multiplies the pre-test Odds by 2 when it is positive.

Conversely, culture of pleural fluid is very specific for the diagnosis of tuberculosis but very insensitive. This test makes it possible to significantly increase the probability of having

the disease if it is positive by multiplying the pre-test Odds by 24. On the other hand, it divides the pre-test Odds only by 1.3 if it is negative.

CT is a test that has both 100% sensitivity and high specificity for the diagnosis of kidney cyst. It is a test that has no false negatives. It eliminates the disease when it is negative. A positive test multiplies the pre-test Odds by 50.

III - SOME SPECIFIC ASPECTS OF DIAGNOSTIC TEST EVALUATION STUDIES

A – Some biases specific to studies evaluating diagnostic tests

1 – Verification bias

In studies evaluating diagnostic strategies, there is a risk of obtaining biased estimates whenever the sick/non-sick status is not measured independently of the test to be evaluated, or vice versa. For example, the incorporation bias occurs when the determination of sick, non-sick status is based at least in part on the result of the test to be assessed. This leads to an overestimation of sensitivity and specificity.

In this category of bias, we find the verification bias that arises when the probability of having the *gold standard* depends on the result of the test to be evaluated. This situation typically arises when the *gold standard* is invasive or expensive and cannot be achieved by everyone. In this case, it is more often performed in subjects who have a positive test than in those who have a negative test.

A study set up to assess the diagnostic performance of the stress electrocardiogram involved 414 subjects at risk of coronary artery disease. All subjects had a stress electrocardiogram. All subjects with a positive exercise electrocardiogram underwent coronary angiography. For subjects with a negative exercise electrocardiogram, only 40% taken at random had a coronary angiogram. The results are presented in Table 5. The estimated sensitivity and specificity in subjects who underwent coronary angiography are respectively:

$$Se = 92/(92+46) = 67\% \quad \text{and} \quad Sp = 72/(72+27) = 73\%$$

Since the probability of having a coronary angiogram is higher in subjects with a positive test than in those with a negative test, there is an overrepresentation of positive tests. The sensitivity of the test is overestimated and the specificity underestimated. Since the probability of having a coronary angiogram depends only on the test result, it is possible to obtain unbiased estimates of sensitivity and specificity using Bayes' theorem.

Table 5 – Exercise electrocardiogram (ECG) results in 414 subjects at risk for coronary artery disease

	Coronary angiography result			
	Coronary heart disease	No coronary heart disease	No coronary angiography	Total
Positive stress ECG	92	27	0	119
Negative stress ECG	46	72	177	295

From the results presented in Table 5, we have an estimate of:

- the probability that the test is positive in the population of subjects at risk of coronary artery disease:

$$119/(119+295) = 28,7\%$$

- the probability of having the disease knowing the positive test:

$$92/(92+27) = 73,7\%$$

- the probability of not having the disease knowing the negative test:

$$72/(72+46) = 61\%$$

Sensitivity estimate:

$$Se = P(Test+ / M) = \frac{P(M/Test+) \times P(Test+)}{P(M/Test+) \times P(Test+) + P(M/Test-) \times P(Test-)}$$

$$= \frac{0,773 \times 0,287}{0,773 \times 0,287 + (1-0,61) \times (1-0,287)} = 44\%$$

Specificity estimate:

$$Sp = P(Test- / NM) = \frac{P(NM/Test-) \times P(Test-)}{P(NM/Test-) \times P(Test-) + P(NM/Test+) \times P(Test+)}$$

$$= \frac{0,61 \times (1-0,287)}{0,61 \times (1-0,287) + (1-0,773) \times 0,287} = 87\%$$

When there is a *gold standard* but it cannot be used in all the subjects included in the study, it is possible to estimate the performance of the test to be evaluated using the gold standard on a sample of positive subjects and a sample of negative subjects taken at random.

2 – Bias related to the use of an imperfect gold standard

It is very common that the test used as a reference is not perfect. If the sensitivity and specificity of the test to be evaluated are estimated by acting as if the reference test were perfect, these estimates are biased. In particular, it is impossible to show the superiority of the new test compared to the reference test.

Let's take the example of a perfect new test, whose performance is evaluated against a reference test that has a sensitivity of 90% and a specificity of 90%. In a study involving 100 diseased subjects and 100 non-diseased subjects, 10 diseased subjects will be classified as negative by the reference test and 10 non-diseased subjects will be classified as positive by the reference test (Table 6). The sensitivity and specificity of the new test will be underestimated by 90%.

Table 6 – Results of a study to evaluate the performance of a new perfect test compared to a reference test which has a sensitivity of 90% and a specificity of 90%. The study covers a sample of 200 subjects including 100 patients and 100 non-patients

	Positive baseline test	Negative reference test
Test to evaluate positive	90 TP	10 FP
Test to evaluate negative	10 FN	90 TN
	100 = 90 VP + 10 FN	100 = 90 NV + 10 FP

In the case where the 2 tests are independent conditionally on the status vis-à-vis the disease, a lack of sensitivity of the reference test leads to an underestimation of the specificity of the new test. Conversely, a lack of specificity of the reference test leads to an underestimation of the sensitivity of the new test.

It is possible to estimate the diagnostic performance of a test in the situation where the reference test is not perfect. The sick, non-sick status of the subjects included in the study is not directly observed, it is a latent variable. The positive or negative results of the test to be evaluated and of the reference test provide information on the status of the subjects.

In a study involving a sample of a population in which the subjects included had the test to be evaluated and the reference test, there are 5 parameters to be estimated: *the sensitivity and specificity of the new test*, the sensitivity and specificity of the reference test and *the prevalence of the disease*. Table 2x2 presenting the combined results of the 2 tests makes it possible to estimate 3 parameters. If the sensitivity and specificity of the reference test are known, then it is possible to estimate the sensitivity and the specificity of the new test and the disease prevalence. The observed data provide 3 degrees of freedom.

If none of the parameters is known with certainty, then it is necessary to increase the information provided by the data. Hui and Walter [3] proposed to use samples of subjects from 2 populations with very different disease prevalences. They took the example of the evaluation of a new skin test to diagnose tuberculosis, the Tine test. The reference test is the Mantoux skin test. They used data from 2 studies: one in which the 2 tests were applied to a sample from the population of a school district with a low prevalence of the disease, and the other in which the 2 tests were applied to a sample of a population of a sanatorium with a high prevalence of the disease.

Under the assumption of conditional independence of the 2 tests and identical diagnostic performance in the 2 populations, there are 6 parameters to be estimated: the sensitivity and specificity of each of the tests and the prevalence of the disease in each of the populations. Tables 2x2 presenting the cross-referenced results of the 2 tests in each of the populations each provide us with 3 degrees of freedom. The number of degrees of freedom is 6, the information provided by the data is therefore sufficient to estimate all the parameters. This approach can be generalized to more than 2 tests or more than 2 populations.

The presentation of estimation methods is beyond the scope of this book. The interested reader can find these methods in the references given at the end of the chapter.

B – The confidence interval and the calculation of the number of subjects required

1 – Confidence interval of sensitivity and specificity

Sensitivity and specificity are estimated by the proportion of positive results in patients and the proportion of negative results in non-patients, respectively. Their estimated variance and standard error is the variance and standard error of a proportion.

For sensitivity: $\text{Variance} = \frac{Se \times (1 - Se)}{M}$ $\text{Erreur standard} = \sqrt{\frac{Se \times (1 - Se)}{M}}$

M = number of patients

For specificity: $\text{Variance} = \frac{Sp \times (1 - Sp)}{NM}$ $\text{Erreur standard} = \sqrt{\frac{Sp \times (1 - Sp)}{NM}}$

NM = number of non-patients

If the numbers of patients and non-patients are large enough and if the sensitivity and specificity are not too close to 100%, the confidence interval can then be constructed using the method based on the approximation of the binomial distribution by a Gaussian distribution.

95% confidence interval of the estimated sensitivity or specificity:

$$P \pm 1{,}96 \times \sqrt{\frac{P \times (1 - P)}{N}}$$

P is the estimated sensitivity or specificity
N corresponds to the number of sick or non-sick people

This method of construction of the confidence interval based on the Gaussian approximation is generally applicable when NP>= 5 and N(1-P)>=5.

When the numbers are too small or the estimates too close to 100%, the exact confidence interval should be constructed based on the binomial distribution.

2 – Calculation of the number of subjects required

When the expected sensitivity and specificity are not too close to 100%, the method for calculating the number of subjects needed to estimate sensitivity and specificity is the same as for estimating a proportion of patients in a population (prevalence of the disease).

If the study set up is a case-control type study, the number of patients to include to estimate the sensitivity and the number of non-patients to include to estimate the specificity are determined separately. It is then necessary to set the expected sensitivity and specificity, the desired width of the confidence interval and its coverage probability, which is generally 95%.

If the study set up is a cohort type study, it is necessary to take into account the prevalence of the disease in the population from which the study sample will be drawn. In most cases, the prevalence of the disease is less than 50%. The strategy to follow is then the following. The number of patients to be included is calculated to estimate the sensitivity, then the number of subjects to be included is calculated to have the necessary number of patients, taking into account the prevalence.

$$N_{Total} = \frac{M}{Prévalence}$$

N_{Total} is the total number of subjects to be included in the study
M is the number of patients

If the prevalence of the disease is greater than 50%, the same strategy is applied, but the number of non-patients to include to estimate the specificity is calculated first.

If the expected sensitivity and specificity are close to 100%, an exact method of calculating the number of subjects based on the binomial distribution should be used.

IV- CONCLUSION

The objective of this chapter is to provide the reader with the methodological tools necessary for setting up a study aimed at estimating the diagnostic capacities of a new test. Emphasis was placed on studies aimed at estimating sensitivity and specificity, which corresponds to the early phase of the evaluation of a new test. The two books which are given as a reference will allow readers who so wish to go further [4;5].

References

1. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tritani F, Chaitman BR, Fisher LD. Correlations among history of angina, ST-segment response and prevalence of coronary artery disease in the coronary artery surgery study (CASS). N Engl J Med 1979; 301: 230-5.

2. Hall FM, Storella JM, Silverstone DZ, Wyshak G. Non palpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography. Radiology 1988; 167: 353-8.

3. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-71.

4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Ed Oxford University Press, New York, 2003.

5. Zhou XH, Obuchowsky NA, McClish DK. Statistical methods in diagnostic medicine. Ed John Wiley & Sons, New York, 2002.