Diagnostic Studies Made Easy

2

Fikri M. Abu-Zidan, Marco Ceresoli, and Saleh Abdel-Kader

2.1 Introduction

Diagnostic methods are one of the major pillars of our daily surgical practice. We routinely encounter a young lady who visits the clinic because she has noticed a breast mass and she is worried that it is malignant, or an elderly man who noticed a change in his bowel habit associated with bleeding per rectum and he is worried that he has colonic malignancy. Alternatively, we may admit a boy to the hospital with suspected appendicitis and we need to decide whether to operate on him or not. To properly solve these problems and to answer patients' concerns, we routinely use diagnostic studies to help us. Whether these methods are radiological, laboratory, endoscopic, or interventional, the main objective of these studies is to guide our clinical decision in finding whether the patient has that suspected disease or not, or occasionally to predict their clinical outcome. Understandably, the benefit of these diagnostic studies should overweigh their side effects especially for invasive procedures. The results of a diagnostic test can be dichotomous (either negative or positive, for example, a SARS-CoV2 PCR test), categorical (like the type of the breast tumor), ordinal (like staging), or continuous values (like the C-reactive protein level). These types of data are explained in more detail in Chap. 13. We aim to lay

F. M. Abu-Zidan (✉)
The Research Office, College of Medicine and Health Sciences, United Arab Emirates University, Al-Ain, United Arab Emirates

M. Ceresoli
General and Emergency Surgery Department, School of Medicine and Surgery, Milano-Bicocca University, Monza, Italy

S. Abdel-Kader
Department of Surgery, Ain Shams University, Cairo, Egypt

M. Ceresoli et al. (eds.), *Statistics and Research Methods for Acute Care and General Surgeons*, Hot Topics in Acute Care Surgery and Trauma,
https://doi.org/10.1007/978-3-031-13818-8_2

the principles of using these diagnostic tests in our clinical practice. This will help to critically appraise a diagnostic study, to design a diagnostic study, and to analyze its data.

Learning Objectives

- Understand the basic components of a diagnostic study.
- Recognize the criteria of a good diagnostic test.
- Define the predictor and outcome of a diagnostic study.
- Understand and be able to calculate the sensitivity, specificity, and predictive values of a test.
- Appreciate the importance of predictive values and likelihood ratios in clinical practice.
- Comprehend that the prior priority of a disease affects both the results and application of a diagnostic test.
- Highlight the most common mistakes encountered in submitted diagnostic study articles.

2.2 Nature of a Diagnostic Study

In principle, diagnostic studies are similar to the observational studies. Nevertheless, observational studies are usually designed to investigate the epidemiology of a disease, explore its etiology, or define its outcome. In contrast, diagnostic studies are commonly designed to answer the question whether the patient has a disease or not.

2.3 The Need for a Gold Standard

How can we reach the disease real status? It can only be reached by using a **gold standard** having a definitive outcome. Ideally the gold standard should be positive in almost all patients with the disease and negative in almost all patients without the disease. This may be an excisional biopsy of a breast mass or an appendectomy with proven histopathology for positive cases. Is this the same for negative cases? Definitely not. We will not operate on negative cases to prove that they were negative but we reach that conclusion mainly with follow-up of the patients. It sometimes gets a little tricky. Let us say that we want to study the diagnostic ability of ultrasound in detecting free intraperitoneal fluid in blunt abdominal trauma. We may decide to consider CT scan as our gold standard although it is not perfect. Occasionally, we may consider the gold standard as CT scan or laparotomy because laparotomy is more accurate than the CT scan. The gold standard is usually used to rule in the disease than to rule it out. The definition of the disease outcome based on a selected gold standard is the most important pillar of a successful diagnostic study.

2.4 Components of Diagnostic Studies

Let us consider the scenario of a male patient presenting with pain in the right iliac fossa. Once you examined his abdomen, you found that it was tender but soft. You are not sure clinically whether the patient has appendicitis or not, so you decided to perform an abdominal CT scan with intravenous contrast to help you in your surgical decision. The result of the CT scan (*test result* whether diagnostic of appendicitis or not) is the *predictor*, and your *outcome* is the disease (whether present or absent). You may decide to observe the patient or operate on him depending on the result. If you have already decided to operate before performing the study, then there is no value of performing the study. This actually may delay your management.

Let us say that the CT scan result showed acute appendicitis (positive result) and then you operated on the patient, removed the appendix, and sent it for histopathology. The appendix can be inflamed (true positive) or normal (false positive). Conversely, the CT scan was normal and you decided to observe the patient. The patient may improve so the result of the CT scan is true negative. In comparison, the patient may develop a frank picture of localized peritonitis and once you operate on the patient, he had an acute perforated appendicitis. Then the result of the CT scan is false negative. This is demonstrated in Fig. 2.1.

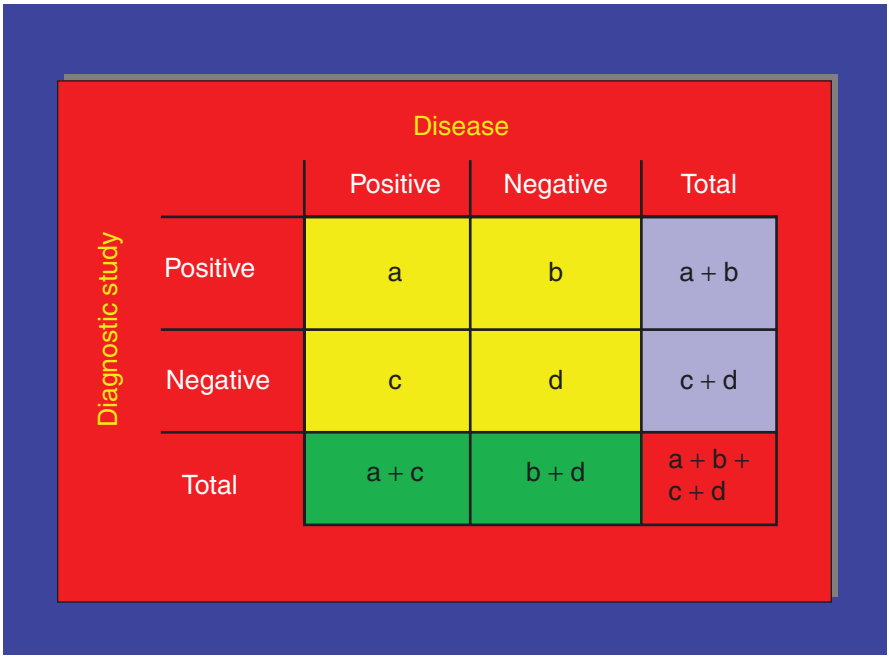

Fig. 2.1 A diagram showing the four cells stemming from the possibilities of the diagnostic study results depending on the disease status of the patient. *a* = true positive result (TP), *b* = false positive result (FP), *c* = false negative result (FN), and *d* = true negative result (TN)

Let us look at Fig. 2.1 carefully and take some time to digest it. That is the key for understanding, designing, and analyzing a diagnostic study. The real disease status is presented by the columns whether it is positive or negative. The results of the test are presented by the rows. Again, it is important to have this mental picture, status of the disease is in the vertical columns, while the results of the diagnostic study are in the horizontal rows. Just to simplify the idea, we will use the term normal for those who do not have the specific disease (although they may have another pathology). Accordingly, we will have four cells: (1) a cell for the positive tests in the diseased patients (a) which are the true positive (TP) results; (2) a cell for the positive tests in the normal patients (b) which are the false positive (FP) results; (3) a cell for the negative tests in the diseased patients (c) which are the false negative (FN) results; and (4) a cell for the negative tests in the normal patients (a) which are the true negative (TN) results.

The next step is to add the cells of each column and each row to have their total. This will give the number of real diseased patients ($a + c$), the number of normal patients ($b + d$); the total number of positive studies ($a + b$), the total number of negative studies ($c + d$), and the total number (n) of study population ($a + b + c + d$).

The third step is to pause, think, and look into the table again. This table can give us two important sides of the diagnostic study: the test and the patient. There are two important criteria that are related to the test which are sensitivity and specificity. Test **sensitivity** measures the ability of a test to detect presence of the disease. It can be calculated from the first column. It is the percentage of the true positive results in those having the disease. This can be calculated by $a/a + c$, in other words TP/TP + FN. In contrast, test **specificity** measures the ability of a test to detect the absence of the disease. This specificity can be calculated from the second column. The specificity is the percentage of the true negative results in those patients not having the disease which is $d/b + d$, in other words TN/TN + FP. It is very common that clinicians concentrate mainly on the sensitivity and specificity of a test. A good test should have high sensitivity and specificity (almost always positive in persons with the disease and negative in persons without the disease, preferably above 90%) but these are only two criteria of other important criteria of an ideal diagnostic test which are shown in Table 2.1.

Table 2.1 Criteria of an ideal diagnostic test

Criteria
Accurate
Simple
Safe
Non-expensive
Non-invasive
Fast
Painless
Has point-of-care option
Reliable
Easy to learn
Generalizable

2.5 Predictive Values

Kindly note that we have looked only at one side of a diagnostic study which is the test. But that is not the way we clinically practice surgery. Figure 2.2 shows the normal process of using a diagnostic test in our practice. Once we meet a patient with a specific complaint, we listen to him/her, examine the patient, decide whether we need a diagnostic test, ask for one if deemed necessary, wait for the results, and finally get the results. The result can be conclusive being positive or negative or may not even give an answer (non-conclusive). In that case, we may need to select another test which can give the answer.

The clinical reality is that a clinician gets a test result (positive or negative) and ponders how accurate this result is in predicting the real disease status of the patient. These are actually the predictive value of a positive test and the predictive value of a negative test. These can be calculated from the horizontal rows (Fig. 2.1).

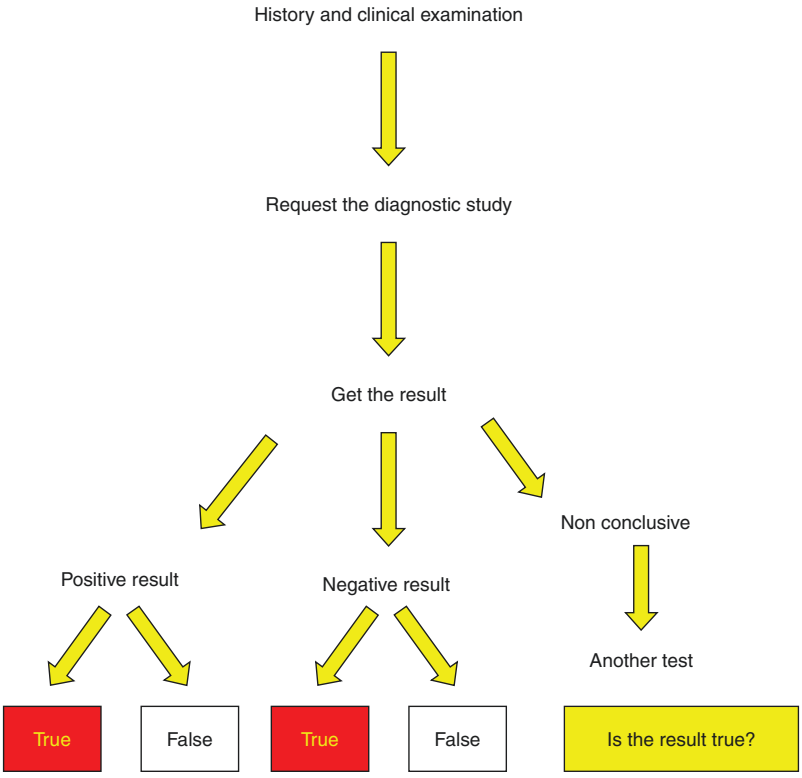

Fig. 2.2 A diagram demonstrating the natural process of the encounter between the doctor and the patient and the usual process for requesting a diagnostic study

The predictive value of a positive test in a study is the probability that a patient with a positive result actually has the disease. This can be calculated from the first row of Fig. 2.1 which is $a/a + b$, in other words TP/TP + FP. **The predictive value of a negative test** in a study is the probability that a patient with a negative result actually does not have the disease. This can be calculated from the second row of Fig. 2.1 which is $d/c + d$, in other words TN/TN + FN. Just to remember, if you evaluate the sensitivity or specificity of a test, calculate vertically. If you evaluate the predictive value of a positive or a negative result, calculate horizontally. Remember that we read horizontally not vertically, and attach that mentally to the clinical importance of the predictive values which is more important than the sensitivity and specificity.

2.6 Prior Probability of the Disease (Prevalence)

There is a need to define the prior probability of the disease in the studied population because the predictive values and the clinical implications of the test when using the likelihood ratios in decision-making will differ depending on the prior probability of the disease. The predictive value of a positive test (PPV) will increase with the increased prior probability. The prior probability of the disease (prevalence) is defined as the percentage of patients who have the disease out of those tested for the disease. In other words, TP + FN/total number of patients (n).

2.7 The Likelihood Ratio (LR)

It is the likelihood that a patient having the disease would have a certain test result divided by the likelihood that a patient without the disease would have the same result. In other words, it is the ratio of the true positive rate to the false positive rate. Sensitivity is the true positive rate, while 1 − specificity is the false positive rate. Accordingly, LR can be calculated as sensitivity/(1 − specificity).

The likelihood ratio is very useful in clinical practice when it is high because of its discriminating power. Figure 2.3 shows the Fagan nomogram. It is a graph which is used to estimate the extent of change in the probability that a patient has a disease depending on the likelihood ratio. The figure gives a theoretical comparison between two diagnostic tests (A and B) that were used to diagnose the disease in the same population having a prevalence of the disease (pre-test probability) of 50%. The diagram enabled us to define the post-test probability when the test was positive. In the A diagnostic test, having LR of 5, the post-test probability of the disease increased to 82% while for the B diagnostic test having LR of 1 the post-test probability of the disease stayed the same at 50%.

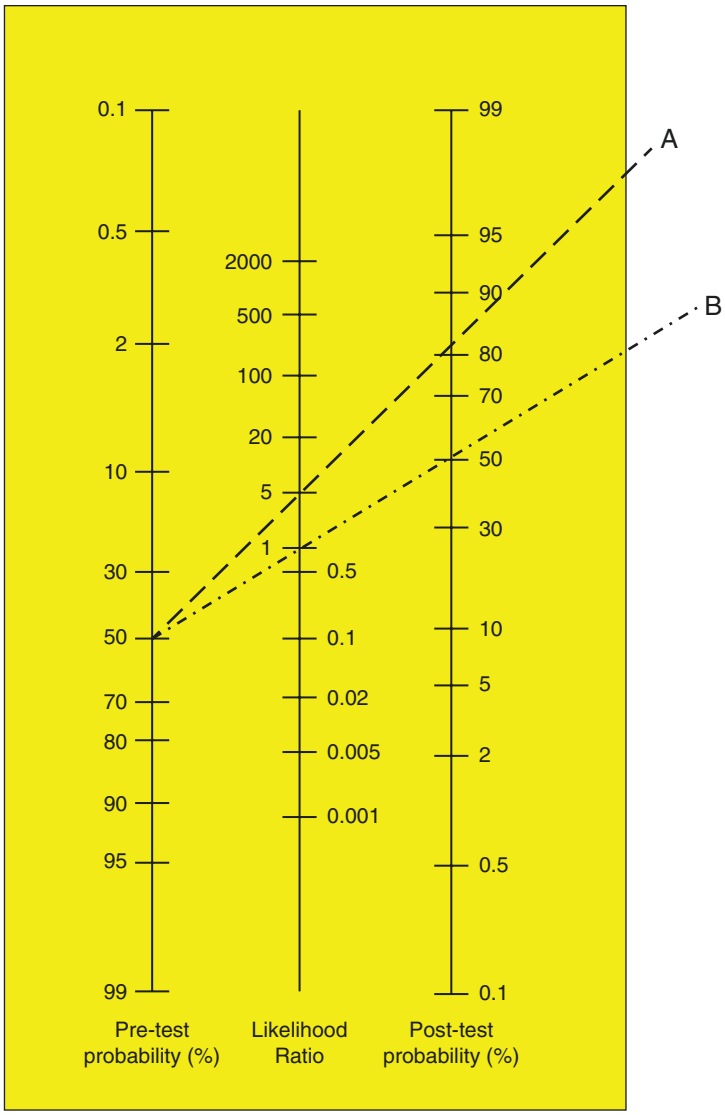

Fig. 2.3 A Fagan nomogram is used to estimate the extent of change of the probability that a patient has a disease depending on the likelihood ratio. The figure compares two diagnostic tests (A and B) that were used to diagnose the disease in the same population having a prevalence of the disease of 50%. For the A diagnostic test, having LR of 5, the post-test probability of the disease increased to 82% while for the B diagnostic test having LR of 1 the post-test probability of the disease stayed the same at 50%

2.8 Receiver Operating Characteristics (ROC) Curves

The characteristics of a diagnostic test can be demonstrated graphically using the ROC curves. They were developed from the analysis of radar receivers during WWII from which they were called receiver operating characteristics curves. ROC curves and their analysis can compare diagnostic performances of different tests and evaluate the best cut-off value for a diagnostic test. They are the graphical representation of diagnostic characteristics of a test having an ordinal or continuous outcome at each possible cut-off point of the test result.

Figure 2.4 shows the ROC of the WSES sepsis severity score in predicting mortality (Sartelli et al., World J Emergency surgery 2015). The *X* axis represents the 1 − Specificity value (false positive rate), while the *Y* axis represents the sensitivity value (true positive rate). Table 2.2 is the SPSS output of the coordinates from which this graph was drawn. Each point of the score (1 − 15) will dichotomize the data. The

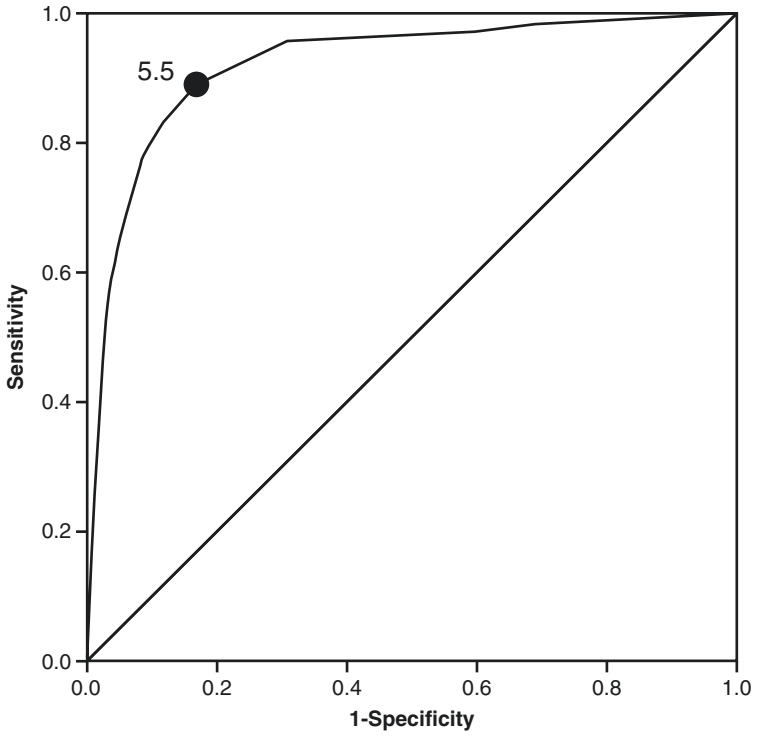

Fig. 2.4 Receiver operating characteristics (ROC) curve for the best WSES sepsis severity score that predicted mortality in patients having complicated intra-abdominal infection, global study of 132 centers ($n = 4553$). The best cut-off point for predicting mortality was 5.5. (Reproduced from the study of Sartelli M et al. Global validation of the WSES Sepsis Severity Score for patients with complicated intra-abdominal infections: a prospective multicenter study (WISS Study). World J Emerg Surg 2015; 10: 61 which is distributed under the terms of the Creative Commons Attribution 4.0 International License)

Table 2.2 SPSS outcome for the WSES sepsis severity score study with the coordinates of the data which were used to produce the ROC so as to define the best cut-off point of the score that predicts death

Positive if more than or equal to	Sensitivity	1 – Specificity
−1	1	1
0.5	0.986	0.766
1.5	0.986	0.725
2.5	0.978	0.653
3.5	0.964	0.395
4.5	0.942	0.323
5.5	**0.896**	**0.221**
6.5	0.802	0.101
7.5	0.757	0.081
8.5	0.598	0.043
9.5	0.436	0.019
10.5	0.335	0.013
11.5	0.159	0.004
12.5	0.101	0.001
13.5	0.024	0
15	0	0

Note that 5.5 had the best sensitivity and specificity

test will be considered true positive if death occurred at the WSES severity score which is greater than or equal to that point. It will be considered true negative if survival occurred if the score was less than that point. The test will be considered false positive if survival occurred at a score which is greater than or equal to that point and will be considered false negative if death occurred at a score less than that point. These dots draw a curve that describes the diagnostic accuracy of the WSES sepsis severity score in predicting mortality. A perfect test is the one which can vertically reach the left upper corner and then becomes horizontal. This would have a sensitivity of 100% and specificity of 100%. The diagonal line represents the reference line and is the result of a test that has 50% of specificity and 50% of sensitivity (like a coin tossing). The best cut-off point is usually where the curve turns with a corner, which was 5.5 in this case.

Another important element of the graph is the area depicted by the curve, called area under the curve (AUC): the higher this value, the higher is the diagnostic accuracy. The area under the reference line is 0.5 and represents a test with no diagnostic abilities. A test with good sensitivity and specificity will have a higher AUC: the maximum is 1. The AUC of the WSES sepsis severity score in predicting mortality was 0.92.

Let us take our trauma registry as another example. We want to evaluate the diagnostic performances of age and injury severity score (ISS) in predicting mortality of trauma patients. Figure 2.5 shows the diagnostic performances of age (blue line) and ISS (green line) in predicting trauma death. The AUC of age was 0.687 and the AUC of ISS was 0.92. ISS shows a better diagnostic performance in predicting mortality with a higher AUC compared with age.

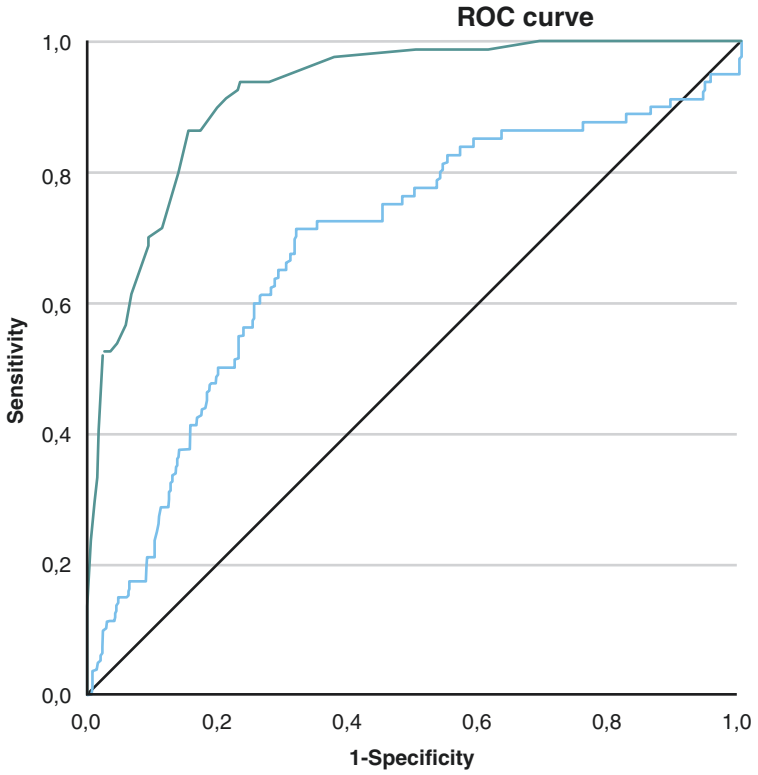

Fig. 2.5 Receiver operating characteristic (ROC) curve comparing the ability of age (blue line) and ISS (green line) in predicting trauma mortality. The area under the curve (AUC) of age was 0.687 and of ISS was 0.92 indicating that ISS had a much better predictor ability

2.8.1 Choosing a Cut-off Point: The Youden Index

What cut-off value can we choose to predict mortality in our clinical practice? According to the cut-off point chosen, the test will have different diagnostic abilities. A high cut-off point will produce a very specific test (low false positive rate because patients having an ISS above the chosen cut-off point will have a very high probability of death) but also a low sensitive test (high false negative rate because mortality may occur with ISS less than the chosen cut-off point).

On the contrary if we choose a low cut-off point we will have a very sensitive test (low false negative rate because death is unlikely if the ISS is less than the chosen cut-off point) but a poorly specific test (high false positive rate). Choosing the best cut-off point of a diagnostic test is not straightforward and should take into consideration the clinical contest. For example, some diseases require high sensitivity (screening tests) and others require high specificity. It is clear that the cut-off point plays a pivotal role in balancing diagnostic characteristic of a test (sensitivity and specificity).

To evaluate the best cut-off point, we can adopt the Youden's J statistics of Youden's index. For each cut-off point we can calculate the Youden's index as *"Sensitivity + Specificity − 1."* This index could assume values between 0 and 1, where value 1 indicates the perfect diagnostic test. The cut-off value with the highest Youden's index indicates the value with the maximum available sensitivity and specificity.

2.9 Common Errors Encountered in Submitted Diagnostic Studies

We hope that by highlighting common errors of diagnostic studies, we will educate young researchers to avoid them when submitting their articles to journals. This will possibly reduce the chance of rejection of their papers. Other common errors encountered in research design are detailed in Chap. 3. Those that we have encountered when reviewing diagnostic studies submitted to acute care surgical journals include:

1. *No clear gold standard*: Using a gold standard is pivotal to assure the validity of the study. Missing the gold standard indicates that you cannot be sure of your results.
2. *Lack of definition of the test results:* The definition of each of the results (true positive, true negative, false positive, and false negative) should be clearly defined in the protocol and should be followed through the whole study.
3. *Not reporting the predictive values or likelihood ratio*: These important clinical values should be calculated and reported. Reporting only sensitivity and specificity is not enough.
4. *Ignoring the learning curve of the operator:* This is a common problem in diagnostic tests that need high technical skills. If the results of the study depend on the skill of the operator like laparoscopy or ultrasound, then the operator should have passed the learning curve stage so the poor results of the test are not attributed to the operator.
5. *Improper study population*: This can be a fatal mistake. The studied population should be that which will benefit from the study. An example for that is selecting a population that has a very high prior probability of the disease like studying the role of C-reactive protein in diagnosing acute appendicitis in those who were already operated. Those who were operated will have a prior probability of almost 90% of acute appendicitis which may be even higher than the sensitivity of the diagnostic test.
6. *Not addressing the generalizability:* Diagnostic studies should be useful in the real clinical situation for a particular setting. For example, the excellent results of diagnosing acute appendicitis by ultrasound experts may not be reproducible in other hospitals without proper training or expertise.
7. *Ignoring the non-conclusive findings:* The percentage of the non-conclusive results should be reported because it may affect the practical usefulness of the test. It is very interesting to note that many diagnostic studies ignore the

non-conclusive studies. These should be minimum to have a test which is useful. Let us assume that a study was done in a population and it was not conclusive in 50% of the patients, do you consider this as a good test!

Do and Don't

- Think about the components of the diagnostic studies you use.
- Use the predictive values and likelihood ratios in your clinical practice.
- Value the impact of prior priority of the disease on the results of diagnostic studies.
- Understand the structure of a diagnostic study. This will help to avoid common errors encountered in designing diagnostic studies.
- Do not concentrate only on the sensitivity and specificity of a study and know their limitations.

Take Home Messages

- There are two major components of a diagnostic study: The method and the population in which it was used.
- Calculate the sensitivity and specificity vertically and calculate the predictive values of a test horizontally.
- Overusing unnecessary diagnostic methods can be sometimes misleading.

Conflict of Interest None declared by the author.

Further Reading

Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? Med Decis Mak. 1994;14:175–9.

Browner WS, Newman TB, Cummings SR. Designing a new study: III. Diagnostic tests. In: Hulley SB, Cummings SR, editors. Designing clinical trials. Baltimore: Williams and Wilkins; 1988. p. 87–97.

Clarke JR. A scientific approach to surgical reasoning. II. Probability revision-odds ratios, likelihood ratios, and bayes theorem. Theor Surg. 1990;5:206–10.

Clarke JR, Hayward CZ. A scientific approach to surgical reasoning. I. Diagnostic accuracy-sensitivity, specificity, prevalence, and predictive value. Theor Surg. 1990;5:129–32.

Clarke JR, O'Donnell TF Jr. A scientific approach to surgical reasoning. III. What is abnormal? Test results with continuous values and receiver operating characteristic (ROC) curves. Theor Surg. 1990;6:45–51.

Goldin J, Sayre W. A guide to clinical epidemiology for radiologists: part II statistical analysis. Clin Radiol. 1996;51:317–24.

Peacock JL, Peacock JL. Diagnostic studies. In: Peacock JL, Peacock JL, editors. Oxford handbook of medical statistics. 1st ed. Oxford: Oxford University Press; 2011. p. 339–51.