EVIDENCE-BASED OPHTHALMOLOGY: CRITICAL EVALUATION OF THE LITERATURE IN RELATION TO DIAGNOSTIC TESTS

OFTALMOLOGÍA BASADA EN PRUEBAS: EVALUACIÓN CRÍTICA DE LA LITERATURA SOBRE PRUEBAS DIAGNÓSTICAS

MESA JC¹, GARCÍA O², LILLO J², MASCARÓ F², ARRUGA J¹

ABSTRACT

Purpose: In general, articles on diagnostic tests have a very poor methodological quality. If we translate their conclusions to daily practice without a proper analysis it is easy to see a typical trend: physicians use new (and expensive) tests without increasing diagnostic capacity; they only increase the health budget.

Evidence-based medicine (EBM) consists of using the best evidence in decision-making. It deals with selected and relevant information, supported by data obtained through the most rigorous scientific method: epidemiology and statistics.

Critical evaluation of papers related to diagnostic tests is our aim. We provide with basic skills for evaluation and analysis of papers by means of simple topics on statistics, design of clinical assays and clinical epidemiology.

Methods: Review of the medical literature.

Results: To evaluate papers on diagnostic tests and to use a test correctly, it is necessary to know its diagnostic capacity, the level of certainty to start treatment, the probability of having the disease before using the test and the test capacity to change that probability.

RESUMEN

Objetivo: En general, los artículos sobre diagnóstico suelen adolecer de una calidad metodológica muy pobre. Si trasladamos sus conclusiones a nuestra práctica cotidiana sin un análisis detenido entenderemos fácilmente un fenómeno creciente: incorporamos acríticamente a nuestras estrategias diagnósticas las novísimas (y carísimas) pruebas sin que con ello aumente significativamente el rendimiento diagnóstico de tales estrategias; sólo se incrementa desorbitadamente el gasto.

La Medicina Basada en Evidencias es el uso de la mejor evidencia disponible en la toma de decisiones. Se trata de actuar utilizando información seleccionada y avalada por datos obtenidos a través del método científico más riguroso: la epidemiología y la estadística.

La evaluación crítica de artículos relacionados con pruebas diagnósticas constituirá nuestro objetivo. Suministrarremos las habilidades básicas para la evaluación y análisis de artículos científicos mediante el uso de una serie de conocimientos sensillos de estadística, diseño de investigaciones y epidemiología clínica.

Método: Revisión de la literatura.
**INTRODUCTION**

Evidence-Based Medicine (EBM) is the deliberate, explicit and sensible use of the best evidence available for taking decisions about individual patients (1). The goal is to utilize in the clinic selected and relevant information confirmed by data obtained through strictly scientific methods, i.e., epidemiology and statistics. EBM focuses on rigorous assessment of scientific evidence supplied by clinical research.

Traditional medical practice, which inherited classic paradigms which are still valid today in many environments, can be characterized by the belief that:

a) Observations derived from personnel clinical experience are a valid form of generating, validating and transmitting knowledge about prognosis of diseases, clinical performance of diagnostic tests and the efficiency of treatments.

b) An adequate knowledge of the underlying physiopathological theory, a combination of an ability for reasoning and logical speculation together with a fair amount of common sense allowing physicians to adequately interpret the signs of a disease and choose the most appropriate treatment; and finally

c) The updating of knowledge through textbooks and journals including traditional «revision articles» in which experts with a large amount of experience expound judicious statements about the solutions to clinical problems. This is why their arguments enjoy a high degree of credibility and the sections of introduction and discussion of original research papers are the focus of debate and set the action guidelines for «good practice».

In contrast with the traditional model, EBM is characterized by the belief that:

a) The information derived from clinical experience and intuition could lead to faulty conclusions when not firmly grounded in systematic observation;

b) The study and knowledge of the basic theoretical mechanisms of a disease is necessary but not sufficient to guide clinical practice; and

c) The physician needs to know certain rules for a precise assessment of the methodology utilized for obtaining scientific evidence on which to base his or her decisions.

In addition, the traditional system of updating knowledge by means of «Continuous Medical Training» is now obsolete because traditional textbooks are unable to convey the new scientific information produced in real-time. Thus, after 43 randomized clinical trials involving over 25,000 patients proving the efficacy of early thrombolysis in heart attack mortality rates, no medical textbook established that indication as a routine procedure. Even more: in 1990 and after 15 randomized clini-
cal trials and three meta-analyses, specialized texts continued to recommend the administration of lidocaine for preventing new infarcts, a completely inefficient measure (2).

On the other hand, in recent decades we have witnessed an information explosion, an exponential growth of medical literature making it impossible to stay up-to-date if we want to be non-critical. In 1948 there were about 4,700 scientific journals, while in 1994 about 2 million articles were published in 20,000 such journals (3). Even though in our specialty this information explosion is easier to manage due to the lower (for now) volume of information to be consumed, it was estimated that to stay up to date a general physician would have read 19 articles per day every day of the year (4). If we combine this with the technological improvements for accessing information through the Internet, the result is a deluge of information which requires physicians who do not want their professional competencies to lose value with the passage of time to master abilities and systematic techniques reinforcing their critical sense to enable a selection of truly relevant information for everyday practice. What we need is a magnet that will help us find the needle in the haystack because, with the current pace of production of clinical trials and other research, it is no longer a matter of whether our decisions have a good scientific basis but of what percentage of available evidence is applied in daily practice (5).

EBM detaches the physician from intuition, unsystematic clinical experience and physiopathology as elements for taking clinical decisions and enhances the value of strict reviews of scientific evidence supplied by clinical research. To achieve this end, EBM requires an ability in the utilization of simple knowledge about statistics, research design and clinical epidemiology, which are added to the arsenal of knowledge and basic abilities of the medical profession.

Accordingly, physicians must acquire the responsibility of critically and independently evaluating the credibility of evidence as well as offered opinions. What is important is not the message but the method with which the data has been obtained. The sections «Material and Methods» and «Results» of articles now become the key part of medical research because they must be assessed in depth in order to evaluate the validity of the data supplied in the article. Therefore, in this scientific way of practicing medicine, the influence of established authorities, i.e., the «experts», is considerably reduced. For this reason Sackett (6), reluctantly turned into an expert himself, proposes the disappearance of this figure to facilitate the progress of science, in first place due to the tendency in the medical community to avoid contradicting them either out of respect or fear, and secondly because journal editors are tempted to accept or reject new ideas or evidence based on their agreement or not with the opinions of the «experts».

This should not be interpreted as a rejection of what we can learn from our teachers or peers. It only means that, if we are want to provide the best attention to our patients, «good practice» of modern medicine must necessarily be based on the comprehensive knowledge of the scientific evidence which supports each clinical practice.

The tools needed for assessing the quality of information supplied by articles discussing diagnostic tests are provided below.

SUBJECTS, MATERIAL AND METHOD

A bibliographic revision has been made of the available information discussing the topic of the paper. As evidence-based medicine is a relatively recent discipline (the term was coined in the eighties by a group of Canadian clinical epidemiologists of McMaster University and began to be extended in clinical practice as of 1992), its system of work is transmitted basically through the Internet and therefore all the information about EBM is freely available. Thus, in this revision almost all the references, spreadsheets and tools were obtained from multiple websites focused on this «new medical approach» which is progressively gaining ground in the medical community.

Having analyzed the information, an article on diagnostic tests published in a high impact Journal (Archives of Ophthalmology) was selected for a critical analysis by way of example.

RESULTS

Accordingly, decisions in the clinic are taken utilizing selected and relevant information supported
by data obtained through strict scientific methods such as epidemiology and statistics. This does not mean the physician must be an expert in epidemiology or statistics to apply EBM principles. It is feasible to acquire basic abilities enhancing critical judgment to obtain the best scientific evidence for each topic (7).

The quest for information

The sources for answering our clinical question are multiple. We can refer to traditional textbooks but, as discussed above, the information contained therein rapidly becomes obsolete making this source inadequate to respond 3-point questions.

Quality-filtered databases can be also utilized, such as Embase, the database of Excerpta Medica; or the popular Medline (www.ncbi.nlm.nih.gov), the database of Index Medicus produced by the National Library of Medicine which, in contrast to Embase, is freely accessible thanks to the Clinton administration. Medline includes a clinical question search engine (Clinical Queries) where users can introduce search terms for therapy, diagnostic, prognosis or etiology. This simplifies searches enormously.

Said databases would provide relevant articles, but these must be assessed to evaluate the evidence they provide. A third alternative is to search secondary journals in which the articles are assessed by physicians and therefore provide reviewed and classified information from the viewpoint of evidence on the basis of methodologically solid articles. At present we can find in the net different sources of providing this type of information. In the Spanish language, these can be accessed through websites such as www.fisterra.com, or www.infodoctor.org. The website of Washington University comprises links to different EBM sources, including the Cochrane initiative (http://healthlinks.washington.edu/ebp/ebpresources.html).

The diagnostic process

A critical reading of an article about a diagnostic test to enable an efficient use of a diagnostic test requires a prior understanding of several basic requirements:

1. A diagnostic test is useful from the clinical viewpoint only if it induces us to take the right therapeutic decisions.
2. Diagnosis is a process that the physician must undertake only in the context of uncertainty. In other words, the use of diagnostic tests makes sense only when anamnesis, physical exploration and other basic diagnostic tests have failed to provide sufficient certainty to carry out a therapeutic action. What is meant by «sufficient» certainty is because we cannot expect absolute certainty to implement a treatment.
3. It only makes sense to consider the use of new tests if we know that their results will reduce our uncertainty (fig. 1).

Thus, the «rational» use of a test requires the clinician to:

a) Be aware of the probability that the patient may exhibit the disease before carrying out the test (prior probability or pre-test);

b) Be aware of the ability of a test to modify that probability (subsequent probability or post-test), and

c) Establish the level of certainty needed to take a therapeutic decision (action threshold).

The degree of uncertainty or certainty we may have about the occurrence of an event (for instance, that our patient does have a specific disease) can be expressed in two equivalent ways (fig. 2): by means of probability or by means of odds. A probability (a risk) is a quantity between zero and one which matches the frequency of emergence of the event expressed as the number of favorable cases divided by the total number of cases. The odds concept, widely used due to its advantages for calculations, is the same idea expressed in a less intuitive form, i.e., with an amount ranging between zero and 1 calculated with the number of favorable cases divided by the number of unfavorable cases (probability divided by its complement).

Fig. 1: Decision threshold.
In order to determine the capacity of the diagnostic test we selected to resolve our uncertainty we utilize sensitivity, specificity and predictive values. Sensitivity and specificity are considered to be the parameters which best evaluate diagnostic performance (internal validity) of a test, while sensitivity (S) represents the capacity of the test to detect cases, and specificity (E) is the ability of the test to detect healthy individuals (no cases). Mathematically, both are conditional probabilities which would be expressed in the following formula: Sensitivity = \( p(+/E) \), i.e., the probability that the test is positive because the subject is ill; Specificity = \( p(-/noE) \), the probability that the test is negative because the subject is healthy (not ill).

Both values are obtained after applying the test to populations whose status of disease is known with certainty (fig. 3). And extremely sensitive test (S) is utilized for discarding the presence of a disease and is defined as «SnOUT», an acronym which means that a negative result (n) excludes or leaves out (OUT) the presence of the disease. On the contrary, an extremely specific test (ES) is utilized to confirm the presence of a disease and is defined as «ESpIN», which means that a positive result (p) confirms (IN) the disease. Although diagnostic tests should have a high sensitivity and specificity, in general both parameters maintain an inverse relationship represented by the so-called ROC curve («receiver-operating characteristic») (8).

However, care must be exercised because the main difficulty is that, in daily clinical practice, these parameters are not needed. The uncertainty is that we do not know the state of health of our patient but what we are certain about is the result of the tests. So the question is if the positive or negative result of the test is right or wrong. The reply to this question is a further set of conditional probabilities, i.e., the predictive values of the test. The positive predictive value (PPV) is \( p(E+/) \), the probability that the patient is ill because the test is positive (pathognomonic sign will have a PPV of 100%), and the negative (NPV) is \( p(noE/-) \), i.e., the probability that the patient is healthy because the test is negative. Please note how the expression \( p(E+/) \) (positive predictive value) is absolutely different to \( p(+/E) \) (sensitivity). It is not a trifle but the well-known «conditional transposition fallacy». The global predictive value is defined as the probability a test has of being right (fig. 3).

What is the problem raised by the use of predictive values? That its calculation is not direct from the sensitivity and specificity but of the prevalence of the disease: The probability of a previous disease to make the test. Conditional probabilities are ruled by Bayes Theorem (fig. 4), and if we express the predictive value according to it, we shall see that the end result depends on the prevalence. This explains why the results of diagnostic tests vary when applied in different regions With different disease prevalence rates.

Lastly, in order to make efficient use of a diagnostic test we need to have the level of certainty we desire in order to carry out a therapeutic action. Having clarified the above concepts, we can now focus on critical reading. The requirements for a critical assessment of articles on diagnostic tests are listed in table I. These are the questions that the article must answer: are the results valid? Which are they? Will they be useful? (9).

1. Are the results valid?

   a) Independent (and blind) comparison with a reference test

We shall utilize the article titled «Use of the Progressive glaucomatous change of the optic disc as a standard reference in the assessment of glaucoma diagnostic tests», published in 2005 in the American Journal of Ophthalmology (10).

All diagnostic tests must be compared with the «truth», a reference test as objective as possible that allows a clear diagnosis of the disease, the so-called «golden standard» or reference test (for example, pathological anatomy for a tumor or a coronary vein graph for a coronary disease). If the selected reference test is not entirely valid, the results of the
study may be questioned. When the reference test is invasive, in case of doubt the usual decision is to follow up the patient for a reasonable period of time to confirm or discard the disease because, in the absence of signs of the disease, an invasive test with added risks is not justified.

In addition to an adequate reference test, it must be applied independently of the result of the test being studied. In the case that the gold standard is applied only when the test detects the disease, we would have what is known as the work-up detection bias. One example of this bias would be requesting a coronary scan only when the effort test is positive. In prospective studies, this bias can be resolved only with an adequate follow up of patients having negative results, while in retrospective studies the bias cannot be eliminated.

Ideally, the test interpretation should be blind, i.e., the gold standard must be applied without knowing the results of the test and vice versa. However this is not always possible. Thus, in studies on imaging tests and glaucoma the ophthalmologist may have clinical data about the patient and make a decision knowing the results of laser polarimetry (GDx). This introduces a bias in the estimation of the performance of the test. In addition, for imaging tests we must consider the possible observer bias according to which the expectations of the physician influenced the results of the measurements. If the ophthalmologist has clinical information about the patient, that information will influence the interpretation of the images. For this reason it is important to

Table I. General questions for an article on diagnostic tests

<table>
<thead>
<tr>
<th>Are the results of the study valid?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is there an independent comparison with an adequate reference test in all cases?</td>
</tr>
<tr>
<td>Is the range of patients of the sample adequate? (similar to the patients on whom the diagnostic assessment will be applied)</td>
</tr>
<tr>
<td>The methods described with sufficient detail to make them repeatable?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Which are the results?</th>
</tr>
</thead>
<tbody>
<tr>
<td>The data of the study allow for the calculation of probability quotients (likelihood ratios)?</td>
</tr>
<tr>
<td>What is the precision of the results?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Applicability of the results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the test and its interpretation reproducible in actual practice?</td>
</tr>
<tr>
<td>If yes, is it acceptable? (consider availability, costs, risk/benefit ratio)</td>
</tr>
<tr>
<td>With the test result modify the decision about how to proceed?</td>
</tr>
</tbody>
</table>

---

Fig. 3: Determination of sensitivity, specificity and predictive values on the basis of a 2x2 table.

<table>
<thead>
<tr>
<th>Ill</th>
<th>Not ill</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test +</strong></td>
<td>a</td>
</tr>
<tr>
<td><strong>Test -</strong></td>
<td>c</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

| Sensitivity = $p(+/E)=a/a+c$ |
| Specificity = $p(-/nE)=d/b+d$ |

Positive predictive value = $p(E+/)=a/a+b$
Negative predictive value = $p(nE/-)=c/c+d$
ensure a blind interpretation. In the instant article, the data were collected independently from the ophthalmologist assessment and before obtaining imaging tests.

Systematic or bias errors produced in the design or execution of the study can cause the sample results to be different from the population of origin. They do not correlate with the size of the sample and when not controlled tend to invalidate the conditions of the study, leading to faulty conclusions. In articles on diagnostic tests, the diagnostic or Berkson is irrelevant. To determine what is happening in the population, a sample of the population is selected and the risk factor under study is associated to a high probability of hospitalization. We also engage in that bias when we choose a control group made up of patients with a disease which is also associated to the exposure factor of the study.

b) The adequate range of patients

The patients included in the study must constitute a population very similar to that which would need the study in daily practice. A mistake in the selection of patients equals comparing clearly unhealthy subjects in which there is no diagnostic doubt with healthy subjects, giving rise to deceiving results when the test is applied in actual clinical practice. A similar problem arises if the proportion of patients with advanced processes is higher than usual because this increases the sensitivity. In most of the articles about diagnostic tests related to glaucoma, the diagnosis is based on the presence of the relevant structure abnormalities or alterations in the visual field of automated standard perimetry. However, the utilization of these parameters as a reference standard for diagnosing glaucoma has a significant limitation: The structural alteration of the papilla or anomalies in the visual fields of a glaucoma patient can be proved after a considerable number of nerve fibers has been lost. Accordingly, the patients included in these studies will probably be in a more advanced stage of the disease, which can facilitate their identification with imaging tests. It is very likely that in highly evolved cases the performance will increase. However, not many articles specify the percentage of incipient and evolved cases.

Another issue to be considered is that of losses. It could be that in some subjects the gold standard or the test in question cannot be applied or that, if applied, the result is not conclusive. In the article under analysis, eyes with important refractive errors were excluded from the study because the imaging tests tend to be less reliable in these patients.

Another point: If the design of the study is of the case/control type, the prevalence of disease in the sample will not be the same as that of the general population because the author selects 1,2 or x number of healthy subjects for each patient. Consequently, the predictive values of the test will not be applicable to the usual sample unless the selection of controls was made to maintain the same prevalence to be found subsequently in daily practice. How was the sample of patients in the instant study composed? It is an observational case/control study. The patients were also included in another study, the DIGS (Diagnostic Innovations

<table>
<thead>
<tr>
<th>Bayes theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ P(A/B) = p(A) \times p(B/A) / p(A) \times p(B/A) + p(A) \times p(B/nA) ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bayes theorem applied to the positive predictive value (PPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ P(E/+)=p(E)\times p(+/E)/p(E)\times p(+/E)+p(nE)\times p(+/nE) ]</td>
</tr>
</tbody>
</table>

Where
- \( p(E) \): the probability of not being ill = 1 - prevalence
- \( p(+/E) \): probability of positive test if ill = sensitivity
- \( p(nE) \): probability of not being ill = 1 - prevalence
- \( p(+/nE) \): probability of a positive test is not ill (False positive) = 1 specificity

Accordingly
\[ PPV = \text{prevalence} \times \text{sensitivity} / (\text{prevalence} \times \text{sensitivity} + (1 - \text{prevalence}) \times (1 - \text{specificity}) \]

<table>
<thead>
<tr>
<th>Bayes theory applied to the negative predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ P(\text{noE/-})=p(\text{noE})\times p(-/E)/p(E)\times p(-/E) - p(\text{noE})\times p(-/\text{noE}) ]</td>
</tr>
</tbody>
</table>

Where
- \( p(\text{noE}) \): probability of being healthy = 1 - prevalence
- \( p(-/\text{noE}) \): probability of negative test is healthy = specificity
- \( p(E) \): probability of being ill = prevalence
- \( p(-/E) \): probability of negative test if ill (False negative) = 1 - sensitivity

Accordingly
\[ NPV = (1 - \text{prevalence}) \times \text{specificity} / (\text{prevalence} \times (1 - \text{sensitivity}) + (1 - \text{prevalence}) \times \text{specificity} \]

Fig. 4: Bayes Theorem and its application to the determination of a predictive value.
in Glaucoma Study) designed for assessing the structure of the optic nerve and the visual function in glaucoma. The patients selected had to fulfill certain inclusion criteria such as the verification of structural changes and having a progressive glaucomatous change documented by means of stereoscopic photographs based on the thinning of the neuro-retinal ring, the increase of excavation and of the defect of the nervous fiber layer. The differences in the color of the neuro-retinal ring, the presence of hemorrhage in the papilla or papillary atrophy were not enough to diagnose glaucomatous progression. Initially and with these data, the population may not be similar to what we can find in our environment.

c) Description of the methods used

The description of the test must have sufficient detail for it to be applicable in patients and it must include the preparation of the patient as well as the execution of the test and the interpretation of its results. The article refers to GDx with variable corneal compensation (GDx-VCC).

The study should also include an additional information which is generally not provided, such as data on the reproducibility of the test, above all of imaging tests or «subjective» assessments. If the results of the test changes according to the observers, the test cannot be reliable. At this point, we will only add that in the case of categorical variables, coincidences between observers is usually measured with the Kappa index and continuous variables by means of graphic methods such as the Bland-Altman (11). The article we are reviewing does not provide any information of this type and, even though the assessment of the GDx parameters should not cause problems, the clinical assessment of the papilla involves a certain degree of subjectivity. The progressive change in the papilla was assessed on the basis of the thinning of the neuroretinal ring, the increase of excavation and of the nervous of fiber layer defects.

However, it has been suggested that the structural characteristics of the papilla should not be utilized as an inclusion criteria in the studies which assess the diagnostic precision of imaging tests. As these instruments evaluate the structural findings of the papilla or the nervous fiber layer, the inclusion of subject based on anomalies in these structures would lead to an over estimation of the sensitivity of the instrument.

Another form of critical interpretation is utilizing the tables of the STARD agreement (STAndards for the Reporting of Diagnostic accuracy studies), which has the aim of improving the reliability and precision of articles about diagnostics to allow readers to estimate the existence of any bias (internal validity) and determine if the conclusions can be extrapolated (external validity). Said agreement is a checklist with 25 points which the reader must verify in each section of the article (title, keyword, methods, results and discussion). Said table is freely downloadable at the following web site: www.stard-agreement.org (12).

2. Which are the results?

The article must include the likelihood ratios (LR) or at least provide the data necessary for calculating them. This term may be less familiar than sensitivity, specificity or predictive values but, as we shall see, it has more advantages. The article provides results in the form of ROC curves, generally difficult to interpret at the clinical level. The LR calculation provides easy to understand information.

All the information we need about the ability of a diagnostic test to diagnose or discard a disease independently of the prevalence rates is found in the probability ratio, veracity ratio (RV) or likelihood ratio (LR). These terms can be found in many texts in Spanish pairs probability quotient or ratio, but strictly speaking the mathematical properties of a probability function are very different to those of a veracity ratio (13). Even so, this definition became widespread since it was chosen by the Spanish group in the EBM. The veracity quotient of the result of a test is defined as the probability of that result in patients divided by the probability of the same result in healthy individuals. Thus defined, the LR constitutes the evidence provided by each result of tests for (or against) the disease. It indicates the increase or reduction of the probability of a disease. It operates as a relative risk (RR) because it indicates how many times that result is more likely in patients with the disease vis-à-vis those who do not have the disease (14). An LR ratio of 1 would indicate that the result is equally probable in both cases, and
therefore it does not differentiate healthy individuals from unhealthy ones. Similarly, a RR of 1 means that the risk of death is the same in treated patients as in untreated patients. Although the definition seems complicated, it is easy to calculate on the basis of sensitivity and specificity (fig. 5).

How can the LR help me in my practice? After all, knowing that the result of a test is for example twice more frequent in patients than in a healthy individual is not as useful as knowing the probability a subject has of having the disease, which is precisely what the predictive value is about (for instance, 85%). We shall see how by using the LR we can also obtain that probability. In our practice, when we suspect a disease we request a diagnostic test to confirm or discard our suspicion. The results of the test modifies the previous probability we entertained for that patient, also known as pre-test probability. For example, what is the probability of suffering glaucoma? Generally, this risk in the course of a lifetime is in the area of 0.5%, the equivalent of its emergence; the prevalence is of 1% (15).

Said probability increases in patients that are assessed due to increase of intra-ocular pressure, alterations in the papilla and abnormalities in the visual field. In other words, each data of the physical assessment and a supplementary explorations modifies the probability of a disease.

Accordingly, in practice the LR actually modifies the probability we had estimated (the pre-test probability), providing a new value which is the post-test probability. In fact, this change of probabilities cannot be done directly because the LR does not modify probability but odds. Therefore, we would have to transform the pre-test probability into odds, multiply that by the LR and transform once again the post-test odds into a probability. If this conversion from probability to odds seems complicated, we can always go for Fagan’s nomogram (16) (fig. 6). As we can see, the first column shows the possible pre-test probabilities, the possible RV values and the third the post-test probabilities. In order to determine the probability our patient has of contracting a specific disease, we only need to draw a line joining the first two columns according to the appropriate values and then extend the line towards value in the third column which, very simply, is the post-test probability.

The more extreme the LR value, the greater the importance of the diagnostic yield of the test. In clinical practice, we must utilize tests with a very high LR for positives and/or a very low LR for negatives. Table II shows the interpretation on the basis of value. The precision of results must have a confidence interval of 95% of the LR of each test result.

The authors of the selected article only provide sensitivity, areas under the ROC curves, variance analysis and LR for a positive result (we must remember that the objective of the study was to detect cases with a high probability of glaucoma). The study included 284 subjects: 61 of them healthy (true negatives) and 71 with the disease (true positives). On the basis of this data it is possible to calculate the rest of parameters with the above formulae, but an easier option is to introduce the data provided by the study into the CASPe network calculator found in its website (16), which is nothing more and nothing less than a spreadsheet. Table III shows the data introduced in table 2x2 provided by the calculator, while Table IV shows the end result. We must first check that the spreadsheet provides all the values that we need and also that the result match those published by authors.

What can we say about the test results? The LR for a positive result is of 1.63 or, in different terms, virtually useless for diagnosing glaucoma (as an initial diagnostic test and within our context). And the LR for negative tests is of 0.83, equally useless for discarding glaucoma under the same conditions as above.

In what concerns the precision of the results, we need not worry because the spreadsheet provides a confidence interval of 95%.

**DISCUSSION**

1) Will the test results modify the decision about how to proceed?

From the clinical viewpoint, a diagnostic test is useful only if it induces us to take the adequate
determine the action threshold; second, we must determine – in the strict conditions of each patient and scenario – the probability of the disease before the test, and finally calculate the post-test probability of the disease giving the test results. If the result of the test is able to take the probability of the disease beyond the threshold value, the use of the test is clinically justified (again, in the conditions of the scenario).

2) Determine the action threshold (AT)

Let us recall the scenario proposed by the article under analysis: it aims at establishing the value of an image test (GDx) in glaucoma.

Let us go over now the way in which our tools provide us a reasonable solution. In the situation of the scenario, the first thing to do is to determine the AT: what probability is sufficient to prescribe treatment. In other words, which is the probability of having glaucoma after which we would indicate treatment? How to calculate this? We have two options.

One is to return to the CASPe calculator where, on the basis of sensitivity, specificity, diagnostic test risk, estimated disease probability and usefulness of the various options available in actual clinical practice (treating a patient with a disease, treating a healthy person, not treating the person with the disease and not treating a healthy person) we obtained this value. Usefulness is a number between zero and one (the CASPe calculator requests an input between zero and 100) with which we measure the impact of the situation. It is up to the practitioner to decide how to measure it: Survival rates, costs, absence of complications, arbitrary units, etc. The calculator requests the risk of the test because it also calculates the diagnostic threshold (the probability of disease after which the test should be requested) (17).

![Fagan's Nomogram](image)

Fig. 6: Fagan’s Nomogram.

(therapeutic) decisions in the context of uncertainty.

Basically, we must carry out three successive steps after studying the scenario: First we must

Table II. Clinical interpretation of the likelihood ratio

| LR+ >10: large increase: Excellent test |
| LR+ between 5 and 10: moderate increase: Good test |
| LR+ between 5 and 2: small increase: poor test |
| LR+ <2: Insignificant increase: useless test |

| LR- between 0.5 and 1: insignificant reduction: Useless test |
| LR- between 0.2 and 0.5: small reduction: poor test |
| LR- between 0.1 and 0.2: moderate reduction: good test |
| LR- <0.1: Large reduction: excellent test |

<table>
<thead>
<tr>
<th>Characteristic under assessment</th>
<th>Present (+ reference test)</th>
<th>Absent (− reference test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Diagnostic test</td>
<td>71</td>
<td>18</td>
</tr>
<tr>
<td>Negative Diagnostic Test</td>
<td>130</td>
<td>65</td>
</tr>
<tr>
<td>Total</td>
<td>201</td>
<td>83</td>
</tr>
</tbody>
</table>
A second and easier option excludes the diagnostic threshold and calculates only the therapeutic threshold on the basis of the damage that the treatment may cause vs. the expected benefits. The AT is established as a risk/benefit ratio (fig. 7) and therefore its value ranges between zero and one. To calculate it, we must first estimate the usefulness or «impact» of both options (treatment or no treatment) in the same scale which utilizes values between 0 and 1. In our scenario, the damage caused by the treatment lies in the side effects caused by anti-hypertension eye drops, the frequency of which has been estimated at 4-5% (14). As regards the expected benefit of the treatment, it is generally considered that avoiding the development of glaucomatous neuropathy involves a reduction of 40% in post-op complications (15). The end result for our scenario is that a probability of glaucoma equal to or above 12.5% would be sufficient for the ophthalmologist to consider prescribing treatment.

3) Determining the probability of disease before making the test

With the data provided by the article we can easily calculate the pre-test probability: 0.70 (201/284); there is a total of 201 patients with a positive reference test out of 284 patients included in the study (10).

We must take into account that these conversions between pre- and post-test probabilities are made automatically on the basis of the pre-test probability and the LR using Fagan’s nomogram or, even easier, in any personal computer with the CASPe calculator in the Internet (16,17).

4) Calculation of the post-test probability

Let us consider the probabilities obtained in the previous steps. If the probability of glaucoma for a patient referred for that reason is of 70% (pre-test probability) and if the TA for prescribing treatment is of 12.5%, we should prescribe treatment in all cases unless we have a diagnostic test capable of modifying that 70% pre-test probability to a number below the TA. Is GDx such a test? It doesn’t seem so because if the test was positive treatment would have to be prescribed in all cases, and if it were negative too because, since the LR of DGx is not sufficient to reduce the probability of glaucoma below 12.5%.

The discussion of the article analyzed states that, even though GDx provides good discrimination between patients with progressive changes in the papilla and healthy individuals, its precision is not adequate enough to suggest its use as an isolated test for diagnosing glaucoma. It ends with a conclusion which is not surprising, i.e., that GDx detects anomalies in patients with a previously confirmed diagnostic of glaucoma (10).

Table IV. Result provided by the CASPe calculator on the basis of table III

<table>
<thead>
<tr>
<th></th>
<th>IC 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>35.3%</td>
</tr>
<tr>
<td>Specificity</td>
<td>78.3%</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>79.8%</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>33.3%</td>
</tr>
<tr>
<td>Proportion of false positives</td>
<td>21.7%</td>
</tr>
<tr>
<td>Proportion of false negatives</td>
<td>64.7%</td>
</tr>
<tr>
<td>Precision</td>
<td>47.9%</td>
</tr>
<tr>
<td>Diagnostic odds ratio</td>
<td>1.97</td>
</tr>
<tr>
<td>J for Youden index</td>
<td>0.1</td>
</tr>
<tr>
<td>CPP or LR (+)</td>
<td>1.63</td>
</tr>
<tr>
<td>CPN or LR (–)</td>
<td>0.83</td>
</tr>
<tr>
<td>Pre-test probability</td>
<td>70.8%</td>
</tr>
</tbody>
</table>

Fig. 7: Calculation of the threshold action.

\[
\text{Threshold action} = \frac{\text{damage expected with the treatment}}{\text{improvement expected with the treatment}}
\]

\[
\text{Damage: frequency of adverse effects (AE) produced by the treatment} \times \text{impact of AE (in a scale from 0 to 1)}
\]

\[
\text{Improvement: frequency of AE prevented by the treatment} \times \text{impact of AE (in the same scale from 0 to 1)}
\]

Calculation in our scenario
- Adverse effect of the treatment
- Adverse effect prevented by avoiding the development of glaucoma
- We consider that the impact of the treatment AE (value=1) is equivalent to the impact of the adverse effects prevented by the treatment (value=1)
- Frequency of adverse effects of the treatment: 5% (8)
- Frequency of adverse effects prevented by the treatment: 40% (9)

\[
\text{Action threshold} = \frac{\text{Expected damage with the treatment/expected improvement with the treatment}}{\text{prevalence}}
\]

\[
0.05 \times 1/0.4 \times 1 = 0.125 \times 12.5\%
\]
5) Resolution of the clinical scenario

The application of this simple diagnostic test has modified our therapeutic attitude for the specific patient of the scenario.

What would have happened had we requested the GDx in spite of a negative result? Because, even with a positive GDx, we would have considered it a false positive because, assuming a LR+ of 5 for GDx, the probability of glaucoma would rise from 2.5% to 11.4%. Even with these values treatment would not be prescribed (threshold of 12.5%). This is a good example of the inadequate use of the diagnostic tests. If a positive result is unable to take us beyond the action threshold, then why request the test?

Even if we strike a point of balance, in orthodox practice the steps to be taken in case of a positive GDx for glaucoma would not be exactly the same. Let us imagine that the patient had a NFI (Nerve Fiber Indicator) indicator of 36 (the NFI index assigns a value of 0-100: with higher absolute values, higher probability of glaucoma). In that case we prescribe treatment but, should we ask for a previous visual field? As stated above, each new diagnostic test modifies the probability of a disease, and now the post-test probability has become the pre-test probability of the perimetry.

Now the ophthalmologist is facing a new scenario with a second level of decision in which the above approach again provides a rational solution. First the new TA must be determined in the new scenario. The therapeutic approach is whether treatment should be prescribed and not before requesting a visual field. To prescribe treatment, obviously the AT should be well above 12.5%. the ophthalmologist knows that other considerations are involved in the AT such as the percentage of non-glaucoma patients with which he is willing to pursue treatment, or a number of economic considerations.

Finally, the ophthalmologist must know the capacity of diagnostic tests available in the scenario in order to change the probability. When the patient arrives, a thorough clinical exploration and anamnesis must be carried out. If at the end he is convinced that the probability of glaucoma is of 90% (a value above his AT), he must prescribe treatment. If, on the contrary, the exploration only raised the probability of glaucoma to 60%, an image test must be requested, the one having the highest LR+ in order to take the glaucoma probability beyond the threshold.

Meaning of the story: in practice, the diagnostic performance of a test mainly depends on the pre-test probability. This increases the importance of learning to estimate our level of certainty at each point of the diagnostic process. Only in this way we will be able to utilize sensibly the diagnostic tests.

Table V below shows how the changes are very small with extreme pre-test probabilities. The diagnostic performance (change in probabilities) is in the situation of maximum uncertainty. The maximum benefit occurred when we were unable to decide for or against the diagnostic, i.e., when the pre-test probability is of 50%. What can we do when we opt for seeing a patient with a suspect papilla after six months and request a visual field? We allow time to pass before assessing again to see whether the pre-test probability has increased. Maybe then we will be above the action threshold and then it will no longer be necessary to request new tests.

In this revision we have discussed the requirements a study on diagnostic tests must have for its results to be acceptable. However, the fact that a diagnostic test has high sensitivity and specificity does not mean that its utilization would perform adequately in all cases. The determination of the action threshold, previous probability of disease in a specific patient and the capacity of the test to modify this probability (LR+ and LR-) are the basic requirements for rational use. We must remember that however good the test may be, we will obtain maximum performance in a situation of maximum uncertainty or a probability of disease of 50% estimated at that specific stage of the diagnostic

<table>
<thead>
<tr>
<th>Pretest probability (%)</th>
<th>Post test probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.6</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>59.1</td>
</tr>
<tr>
<td>25</td>
<td>81.3</td>
</tr>
<tr>
<td>50</td>
<td>92.9</td>
</tr>
<tr>
<td>75</td>
<td>97.5</td>
</tr>
<tr>
<td>90</td>
<td>99.2</td>
</tr>
</tbody>
</table>
process. Tip: use of the tests when the pre-test probability has increased to the maximum after an exhaustive anamnesis and physical exploration.

BIBLIOGRAFÍA