Evaluarea unui test diagnostic

The evaluation of a diagnostic test

V. Bacârea[1*], Anca Bacârea[2], T. Călinici[3], M. Măruşteri[4], Ch. Petitot[5]

7. UMF Târgu Mureş, Disciplina de Epidemiologie
8. UMF Târgu Mureş, Disciplina de Laborator, biochimie clinică
9. UMF „Iuliu Haţieganu", Cluj Napoca, Disciplina de Informatică şi biostatistică medicală
10. UMF Târgu Mureş, Disciplina de Informatică medicală
11. OMS Geneve, Departement de recherche clinique

Rezumat

Un test diagnostic reprezintă posibilitatea de a determina prezenţa sau absenţa unei boli la un pacient provenind dintr-o populaţie de studiu. Parametrii care definesc un test diagnostic sunt sensibilitatea şi specificitatea. În limbaj medical o exprimare corectă ar fi sensibilitatea şi specificitatea medicală. Pentru calcularea acestor parametri se dezvoltă un protocol de comparare cu un "standard de aur". Dacă acesta lipseşte, este obligatorie construirea unuia nou, definit prin criterii subiective sau obiective. Dacă acestă construcţie nu este posibilă, se poate calcula concordanţa, ca indicator al testului diagnostic evaluat. Evaluarea unui test diagnostic sau screening poate deveni o sarcină dificilă dacă rigorile metodologice nu sunt respectate.

Cuvinte cheie: *test diagnostic, "standard de aur", concordanţă*

Abstract

A diagnostic test is a measurement used to indicate the presence or absence of a specific disease as condition in a patient from a specific patient population. The parameters that could be calculated are Sensitivity and Specificity. More exactly, in medicine we use clinical sensitivity and specificity. In order to calculate those parameters for a new diagnostic test it is necessary to develop a comparative procedure also called „golden standard". When this „golden standard" is not available, the construction of a new one becomes appropriate. Furthermore, if that construction is not possible we can calculate a level of agreement. Evaluating a diagnostic or even screening test could be a difficult task, if the methodological issues are not followed.

Keywords: *diagnostic test, "golden standard", agreement.*

[*] **Adresa de corespondenţă:** Str. Gh. Marinescu 38, UMF Târgu Mureş, Disciplina de Epidemiologie.
Telefon: 0265215551 int. 258; 0744645744; E -mail: bacarea@gmail.com

Introducere

Un test diagnostic este o metodă folosită pentru a indica prezența sau absența unei boli sau a unei condiții morbide la un pacient. În cel mai simplu mod de abordare, un pacient poate sau nu poate avea o boală sau o afecțiune. Vom folosi termenul de *bolnav* sau *fără boală* ca fiind cele mai corecte pentru a deosebi clar cele două mari grupuri de pacienți. Un test diagnostic indică prezența sau absența bolii la persoanele investigate, care sunt cu sau fără boală. Pentru acest test unic ca aplicare se descriu două moduri cantitative de a aprecia calitatea testului: *sensibilitatea* şi *specificitatea*.

Cea mai simplă definire:
Sensibilitatea – test pozitiv la un bolnav;
Specificitatea – test negativ la o persoană fără boală.

Există mai multe posibilități pentru evaluarea unui nou test diagnostic.

O posibilitate de a evalua un nou test diagnostic implică testarea unui eşantion reprezentativ dintr-o populație țintă, şi compararea rezultatelor noului test cu statusul clinic al pacienților sau cu rezultate obținute printr-o altă metodă. Când procedeul diagnostic este general acceptat (în lumea specialiştilor din lumea medicală) ca fiind un test cert, el devine aşa numitul „standard de aur", față de care se poate compara noul test. Acest „standard de aur" trebuie să fie foarte concis (pozitiv / negativ, prezent / absent, bolnav / fără boală), fără a permite apariția unor rezultate incerte sau intermediare[1,2].

Când un nou test este comparat cu un „standard de aur" sau un status clinic, *sensibilitatea* se va defini ca proporția dintre bolnavii depistați de noul test ca pozitivi față de toți bolnavii.

În mod similar, *specificitatea* se va putea calcula ca fiind proporția de persoane fără boală depistate corect față de toate persoanele fără boală.

Aceste valori vor fi doar nişte estimatori ale testului din cauza lotului de pacienți ales; dacă lotul este *reprezentativ* atunci aceşti estimatori devin corecți populațional, numindu-i statistic „nebiasați".

Dacă procedura de comparare este imperfectă, *sensibilitatea* şi *specificitatea* sunt aproape întotdeauna „biasate" (în mod *sistematic* cu valori prea mari sau prea mici). Chiar şi mai rău: nu se poate determina direcția biasului; singurul lucru care poate fi statuat este că aceşti estimatori sunt eronați. Deci, pentru a se obține valori corecte în ceea ce priveşte aceşti estimatori, testul nou trebuie comparat cu un status clinic foarte corect sau cu un standard perfect[4,6].

În alte cazuri compararea unui nou test cu un status clinic perfect sau un „standard de aur" este imposibilă, nepractică sau extrem de costisitoare. În acest caz noul test va fi comparat deseori cu un standard imperfect. În această situație *sensibilitatea* şi *specificitatea* nu sunt termenii corecți pentru descrierea rezultatelor comparației. Întrebarea care se pune, totuşi, este cum să raportăm rezultatele obținute din studierea unui test nou, atunci când procedura comparativă este imperfectă.

Vom prezenta în continuare cele mai corecte practici statistice de a prezenta rezultate din diverse studii care doresc să evalueze un nou test diagnostic.

Ghid statistic general pentru evaluarea unui test diagnostic

Cel mai important lucru este planificarea corectă a studiului înainte de începerea colectării primelor date. Acesta include şi ideea de a comunica sau nu *sensibilitatea* şi *specificitatea*. Dacă se doreşte comunicarea acestor estimatori este absolut necesară conturarea unui criteriu clinic cert sau stabilirea unui „standard de aur"[3].

Un alt pas cheie în planificarea studiului este stabilirea unei colaborări cu un specialist în biostatistică, pentru a se pune la punct protocolul de analiză statistică corect, precum şi metodele statistice care urmează a fi aplicate.

Tabelul I. Introducerea datelor în tabel de contingență

	B+	B-	Total
T+	44	1	**45**
T-	7	168	**175**
Total	**51**	**169**	**220**

După planificarea corectă a studiului se pot descrie patru eventuale posibilități:

1. Dacă există un „standard de aur" atunci trebuie folosit. Se pot estima corect *sensibilitatea* şi *specificitatea.*
2. Dacă „standardul de aur" există dar este nepractic de folosit, el trebuie folosit cât mai mult posibil. Se pot calcula valori ajustate de *sensibilitate* şi *specificitate.*
3. Dacă „standardul de aur" nu există, trebuie încercată definirea lui. Se pot calcula valori ale *sensibilității* şi *specificității* raportate la standardul definit.
4. Dacă „standardul de aur" nu există sau nu poate fi definit atunci compararea se face cu alt test care nu este „standard de aur", evaluându-se în ce măsură cele două teste dau rezultate similare.

1. Dacă există un „standard de aur" atunci trebuie folosit.

Din punct de vedere statistic este cea mai bună variantă pe care am putea să o folosim. Compararea unui test diagnostic cu un „standard de aur" sau cu un status clinic cert permite calcularea estimatorilor chiar în scopul de ai putea aplica în populația generală.

Exemplu:

Presupunem că evaluăm un nou test diagnostic. Folosim 220 pacienți. 51 dintre ei sunt bolnavi iar 169 sunt fără boală. Testul nou indică că 7 dintre bolnavi au rezultat negativ şi 1 dintre cei fără boală are test pozitiv. Atunci calculul parametrilor se face astfel (*Tabelul I*):

Sensibilitatea = 44/51 x 100 = 86,3%.

Specificitatea = 168/169 x 100 = 99,4%.

Intervalul de confidență 95% calculat pe baza unei distribuții binomiale a variabilelor se poate calcula folosind programul GraphPad, aplicându-se testul Chi Pătrat cu corecție Yates. Obținem valorile: (73,7%, 96,8%) pentru sensibilitate şi (96,8%, 100,0%) pentru specificitate.

Se poate calcula, de asemenea, şi valoarea predictivă a testului (*Figura 1*).

2. Dacă „standardul de aur" există dar este nepractic de folosit, el trebuie folosit cât mai mult posibil. Se pot calcula valori ajustate de *sensibilitate* şi *specificitate.*

Dacă utilizarea unui „standard de aur" este considerată impracticabilă sau nefezabilă se pot totuşi calcula estimatorii prin folosirea unui standard imperfect, folosind „standardul de aur" numai pentru un număr redus de cazuri.

De exemplu, se aplică compararea cu „standardul de aur" pentru toți indivizii al căror rezultat a fost diferit de standardul imperfect şi a unui lot redus, ales la întâmplare, din cei unde rezultatul a fost în concordanță cu standardul imperfect. Folosind aceste date, putem calcula estimatorii ajustați. Totuşi este nevoie de o retestare pe un număr extins de cazuri pentru a obține o precizie acceptabilă.

Rareori este posibilă calcularea estimatorilor (*sensibilitate* şi *specificitate*), fără utilizarea unui „standard de aur". Acest lucru este posibil cînd *sensibilitatea* şi *specificitatea* standardului imperfect au fost corect stabilite prin comparări anterioare cu un „standard de aur" utilizând aceeşi populație țintă.

Calculul propriu-zis are aceeaşi desfăşurare ca în cazul 1.

3. Dacă „standardul de aur" nu există, trebuie încercată definirea lui. Se pot calcula valori ale *sensibilității* şi *specificității* raportate la standardului definit.

Probabil un expert poate să definească un set de criterii care poate servi ca „standard de aur". Deşi partea inițială a studiului va dura mai mult, rezultatele vor fi corecte. Este însă important ca „standardul de aur" construit să fie prezentat în protocolul de studiu.

Calculul propriu-zis are aceeaşi desfăşurare ca în cazul 1.

Exemplu "standard de aur"

Chi-square Test

The two-sided P value is < 0.0001, considered extremely significant.
The row/column association is statistically significant.

Note: With such small values, the chi-square P value
is not accurate. Use Fisher's exact test instead.

Calculation details:
Chi-square statistic (with Yates correction) = 171.55
Degrees of freedom = 1

Sensitivity and specificity

Variable	Value	95% Confidence Interval
Sensitivity	0.8627	0.7376 to 0.9430
Specificity	0.9941	0.9675 to 0.9999
Positive Predictive Value	0.9778	0.8823 to 0.9994
Negative Predictive Value	0.9600	0.9194 to 0.9838
Likelihood Ratio	145.80	

Data analyzed

	B+	B-	Total
T+	44	1	45
	(20%)	(0%)	(20%)
T-	7	168	175
	(3%)	(76%)	(80%)
Total	51	169	220
	(23%)	(77%)	(100%)

Figura 1. Calcularea parametrilor unui test diagnostic folosint utilitarul GraphPad

4. Dacă „standardul de aur” nu există sau nu poate fi definit atunci compararea se face cu alt test care nu este „standard de aur”, evaluându-se în ce măsură cele două teste dau rezultate similare.

În acest caz, tot ceea ce avem la dispoziție este standardul imperfect. Când un test nou este comparat cu un standard imperfect nu se pot calcula direct şi fără erori estimatorii doriți (*sensibilitate* şi *specificitate*). Deci termenii *sensibilitate* şi *specificitate* nu sunt corect folosiți în a exprima rezultatul comparării. Totuşi rezultatele calculului rezultat din tabelul 2 x 2 pentru compararea celor două teste trebuie comunicat. Se descrie în mod obligatoriu standardul imperfect şi apoi se calculează *concordanța* testului studiat cu standardul imperfect. Se expune de asemenea şi intervalul de confidență pentru *concordanță*[5,7].

Se poate vorbi despre două inconveniente majore în ceea ce priveşte calculul *concordanței*. În primul rând, acest parametru nu este o măsură a corectitudunii pentru că cele două teste pot fi concordante în „greşirea diagnosticului”. De fapt, ambele pot concorda foarte bine, însă ambele pot să aibă *sensibilitate* şi *specificitate* reduse. În al doilea rând *concordanța* variază în funcție de prevalența bolii studiate[6].

Ca exemplu ipotetic, putem presupune

Tabelul II. Tabel de contingență pentru calcularea concordanței

	Standard imperfect	
	+	-
T+	a	b
T-	c	d
Total	**a+c**	**b+d**

că testul studiat şi standardul imperfect ales concordă perfect când diagnosticul este negativ şi nu concordă pentru diagnostic pozitiv. *Concordanța* va fi mare pentru boli cu prevalență redusă şi mică pentru boli cu prevalență mare. Performanța noului test față de standardul imperfect nu se schimbă, dar *concordanța* dintre ele variază datorită prevalenței diferite[1].

Exemplu:

Când comparăm un test cu un standard imperfect calculele uzuale din tabelul 2 x 2 estimează eronat (biasat) *sensibilitatea* şi *specificitatea* din cauză că standardul imperfect nu este întotdeauna corect. Totuşi, a afla cât de des rezultatele noului test concordă cu standardul imperfect poate fi de folos. Pentru a calcula *concordanța,* unui eşantion de indivizi i se aplică ambele teste, testul nou şi standardul imperfect. Rezultatele sunt introduse în tabelul 2 x 2 (*Tabelul II*).

Diferența dintre tabelul de contingență cu „standard de aur" şi acesta este că aici pe coloane nu este realitatea absolută, ci o comparare între două teste.

Cel mai uşor mod de a calcula *concordanța absolută* este evaluarea proporției de subiecți la care cele două teste au dat rezultat similar.

Concordanța absolută =(a+d)/(a+b+c+d)x100

Această valoare indică concordanța globală, atât pentru bolnavi cât şi pentru cei fără boală. Este folositor calculul *concordanței pozitive* şi *negative*:

Concordanța pozitivă = a/(a+c)x100;

Concordanța negativă = d/(b+d)x100.

Cei 220 de pacienți sunt testați cu ambele teste (*Tabelul III*), rezultând:

Concordanța absolută = 95,9% cu IC (92,4%, 98,1%);

Concordanța pozitivă = 90,9%;

Concordanța negativă = 97,2%.

Se poate trage concluzia că cele două teste sunt concordante, în general, dar mai ales pentru depistarea lipsei bolii. Altfel spus putem spune că şi testul nou are „o specificitate" mare.

Comparând datele cu cele obținute la exemplul 1, unde s-a folosit „standard de aur", constatăm: standardul imperfect clasifică 44 indivizi ca fiind bolnavi în loc de 51 câți sunt de fapt. Deoarece standardul imperfect greşeşte uneori, nu se pot calcula fără bias *sensibilitatea* şi *specificitatea,* doar *concordanța.*

Calcularea *concordanței* are două dezavantaje majore. Unul este acela că „concordant" nu înseamnă „corect". Al doilea, că ea variază în funcție de prevalență.

Când două teste sunt concordante nu putem trage concluzia că sunt şi corecte. Pentru a demonstra corectitudinea, este nevoie de un tabel de contingență 3 x 3, unde să fie introduse rezultatele testului nou, ale standardului imperfect şi ale „standardului de aur" (*Tabelul IV*).

Se poate constata faptul că testul nou şi standardul imperfect dau rezultate similare la 40 + 171 = 211 persoane testate, dar greşesc ambele față de „standardul de aur" cu 4 + 5 = 9 persoane.

Tabel III. Aplicarea practică a calculului concordanței

	Standard imperfect		
	+	-	**Total**
T+	40	5	**45**
T-	4	171	**175**
Total	**44**	**176**	**220**

Tabel IV. Evaluarea practică a sensului de concordanță a două teste prin compararea lor cu „standard de aur”

Test nou	Standard imperfect	Total persoane testate	„Standard de aur”	
			+	-
+	+	40	39	1
+	-	5	5	0
-	+	4	1	3
-	-	171	6	165
Total		**220**	**51**	**169**

După raportarea față de „standardul de aur”, cercetătorul este cel care va putea trage concluzii în ceea ce privește calitățile testului diagnostic evaluat.

Bibliografie

1. Akobeng AK, Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice. Acta Paediatrica (Oslo, Norway: 1992), 2007 Apr; Vol. 96 (4), pp. 487-91;
2. Leeflang MM, Meta-analysis of diagnostic test accuracy. Clinical Infectious Diseases: An Official Publication Of The Infectious Diseases Society Of America, 2006 Nov 1; Vol. 43 (9), pp. 1220;
3. Lesaffre E., Bayes and diagnostic testing. (eng; includes abstract), Veterinary Parasitology [Vet Parasitol], 2007 Aug 19; Vol. 148 (1), pp. 58-61;
4. O'Connor A, Critically appraising studies reporting assessing diagnostic tests. The Veterinary Clinics Of North America. Small Animal Practice, 2007 May; Vol. 37 (3), pp. 487-97;
5. Orenstein EW, Methodologic issues regarding the use of three observational study designs to assess influenza vaccine effectiveness. International Journal Of Epidemiology 2007 Jun; Vol. 36 (3), pp. 623-31;
6. Pepe MS, Insights into latent class analysis of diagnostic test performance. Biostatistics (Oxford, England), 2007 Apr; Vol. 8 (2), pp. 474-84;
7. Stamey JD, Bayesian estimation of intervention effect with pre- and post-misclassified binomial data. Journal Of Biopharmaceutical Statistics, 2007; Vol. 17 (1), pp. 93-108;