XVII. Evaluarea testelor diagnostice: criterii de validitate

Anormalitatea

Medicii petrec o bună parte a timpului distingînd normalul de anormal (Sunt aceste adenopatii patologice? Este consistenţa ficatului crescută? Este scăderea ponderală la acest individ determinată de stress sau are cancer?).
Când anomaliile sunt grosiere (adenopatii mari, ficat dur cu marginea ascuţită, pacient emaciat cu alte modificări biologice), este simplu de răspuns la aceste întrebări. Greutatea apare atunci când modificările sunt subtile, cum se întâmplă de obicei la debutul bolii – şi uneori diagnosticul este cu atât mai important. Atunci fie decidem că modificările reflectă o boală şi anunţăm pacientul, continuăm investigaţiile (cu unele mai invazive şi/sau mai scumpe) sau tratăm boala (operăm, administrăm tratamente care costă şi au efecte adverse), fie decidem că modificările nu ies din sfera normalului şi liniştim pacientul (dar ne asumăm riscul de a trece pe lângă o boală care totuşi există, iar când diagnosticul va fi clar, poate va fi prea târziu).
Testele diagnostice sunt de două feluri, calitative şi cantitative. Primele clasifică pacienţii ca bolnavi sau sănătoşi în funcţie de prezenţa sau absenţa unei modificări clinice sau de laborator (de exemplu, prezenţa sau absenţa unei pneumonii la examenul radiologic). Cele cantitative au ca rezultat o variabilă numerică continuă, iar clasificarea pacientului ca bolnav sau nu se face pe baza unei valori de prag (numite criteriu de pozitivitate), în funcţie de care testul este considerat pozitiv sau negativ (de exemplu TA, creatinina, transaminazele, VEMS etc.). Sunt mai multe metode de a hotărâ care este normalul:

1. **Metoda distributiei Gaussiene**, bazată pe presupunerea că valorile testului au o distribuţie normală (Gaussiană). Conform acestei metode, se încadrează în normal cei aflaţi în zona mediei ± 2 deviaţii standard, adică 95% din populaţie, extremele fiind anormale. Pentru a obţine aceste valori, se face un studiu pe o populaţie considerată "sănătoasă" (eventual separat pe grupe de vârstă, sex, rasă sau alte subgrupuri), se aplică testul diagnostic şi se calculează media şi deviaţia standard. Problema cu această metodă este aceea că rezultatele majorităţii testelor nu au o distribuţie normală (de exemplu media creatininemiei este în jurul valorii 0,9, în timp ce extremele se pot duce şi la 15 în sus, dar mult mai puţin în jos); de asemenea, conform acestei metode extremele de 2,5% ar reprezenta anormalul, deci toate bolile ar avea o prevalenţă fixă de 5%, ceea ce este absurd. Singurii pacienţi normali ar fi cei necăutaţi (dacă aplicăm 1 test, probabilitatea să iasă un rezultat normal ar fi 95%; dacă aceluiaşi pacient îi aplicăm 2 teste, probabilitatea să iasă normale scade la 90%; iar dacă îi aplicăm 20 de teste independente – nu ştiu câte teste are screeningul anual obligatoriu introdus recent de Ministerul Sanataţii – probabilitatea scade la 0.95^{20}=33%).
2. **Metoda procentuală**, care este aproape identică celei Gaussiene (se măsoară valorile testului într-o populaţie şi se defineşte normalul tot ca valorile în care se încadrează 95% din populaţie, iar patologicul fie 5% din valorile superioare, fie 2,5% din extremele de ambele părţi). Această metodă nu cere ca valorile testului să aibă o distribuţie normală, dar are aceleaşi dezavantaje ca şi metoda Gaussiană.
3. **Metoda "preferinţei culturale"**, care descrie ca normal ceea ce societatea consideră astfel (de exemplu greutatea la femei). În acest caz se creează o confuzie cu privire la rolul medicinei.

4. **Metoda factorului de risc**: valorile devin patologice peste pragul la care este demonstrat că factorul respectiv devine factor de risc. Desigur, nu ne ajută cu nimic diagnosticarea unui factor de risc asupra căruia nu putem interveni.
5. **Metoda diagnostică sau a valorii predictive**: de acestă metodă ne vom ocupa în continuare, şi pe aceasta o aplicăm de fapt în clinică zi de zi.
6. **Metoda terapeutică**: valorile devin patologice de la nivelul la care s-a demonstrat că tratându-le se obţine mai mult beneficiu decât efecte adverse. Exemple pentru metoda aceasta şi cea a factorului de risc pot fi TA şi colesterolul, la care vedem că valorile de prag se modifică continuu. Pentru stabilirea valorilor de prag sunt necesare mari studii de cohortă (metoda factorului de risc) şi terapeutice (metoda terapeutică). Metoda terapeutică este cea mai pragmatică, derivând din metoda factorului de risc.

Performanţa măsurătorilor: validitatea şi reproductibilitatea

Validitatea (sau **acurateţea**) este corectitudinea măsurătorii – ne arată cât de aproape este valoarea unei măsurători de valoarea reală (cât de aproape este o potasemie măsurată de potasemia reală).
Reproductibilitatea (sau **precizia**) arată gradul în care o serie de măsurători fluctuează în jurul unei valori centrale, valoare care poate fi mai mult sau mai puţin apropiată de valoarea reală (în funcţie de **acurateţea** testului). De exemplu în determinarea unei potasemii (valoare reală de 4 mEq/l), metoda noastră de măsurare este validă dacă repetând măsurătoarea vom avea valori foarte apropiate de 4 mEq/l. Pentru ca metoda să fie precisă, măsurătorile trebuie să dea valori foarte apropiate între ele, chiar dacă diferite de cea reală (de exemplu dacă toate măsurătorile dau valoarea 7,49 mEq, înseamnă că metoda este foarte precisă, dar deloc validă).
Acurateţea se poate evalua determinând diferenţa dintre media rezultatelor măsurării şi valoarea reală, în timp ce precizia se evaluează observând distribuţia frecvenţelor măsurătorilor şi calculând deviaţia standard a acestora. O metodă de măsurare este bună atunci când este în acelaşi timp validă şi precisă.
Atunci când rezultatele testelor sunt binare (pozitiv/negativ), reproductibilitatea interobservator se calculează raportând numărul de rezultate convergente (în care investigatorii au dat acelaşi verdict) la numărul de rezultate divergente (în care părerile investigatorilor au fost diferite) - aceasta este reproductibilitatea calculată prin metoda simplă (Tabelul XVII.1 - doi clinicieni au consultat 100 de pacienţi cu dispnee: ambii sunt de acord cu prezenţa zgomotului 3 la 5 pacienţi şi cu absenţa lui la 75 de pacienţi – aşadar, “acordul simplu” este de (5+75)/100 = 0,80, sau acordul a fost prezent în proporţie de 80%).

Tabelul XVII.1 Calculul reproductiilităţii interobservator – metoda simplă. Acordul dintre doi clinicieni asupra prezenţei galopului protodiastolic la pacienţi cu dispnee.

		Clinician II (galop)	
		DA	NU
Clinician I (galop)	DA	5	5
	NU	15	75

Acest calcul are avantajul de a fi simplu de efectuat, dar nu ţine cont de faptul că cei doi observatori pot ajunge la acelaşi rezultat, măcar într-un număr de cazuri, din întâmplare (să ne închipuim că medicii, în loc să asculte zgomotele cardiace, diagnostichează prezenţa zgomotului 3 dând cu banul: capul = zgomot 3 prezent, pajura = zgomot 3 absent; este clar că, de câte ori cei doi vor ajunge la acelaşi rezultat prin această "metodă diagnostică", acest lucru va fi rezultatul întâmplării).
Pentru remedierea acestui neajuns, deci pentru a vedea care este acordul dintre doi investigatori în realitate, scăzând contribuţia întâmplării, se foloseşte **coeficientul de concordanţă κ**, care ia valori de la 0 la 1 (pentru valorile de la 0 la 0,2 concordanţa este uşoară, de la 0,2 la 0,4 este acceptabilă, de la 0,4 la 0,6 moderată, de la 0,6 la 0,8 substanţială, iar de la 0,8 la 1 aproape perfectă[1]). Pentru exemplul din Tabelul XVII.1, coeficientul de concordanţă κ este 0,23 (mult mai mic decât acordul simplu de 80%, considerându-se că acordul datorat întâmplării a fost în acest caz 74% - pentru calculul κ, vezi [1, 2]).
În Tabelele XVII.2 şi 3 putem vedea câteva exemple de acord interobservator (pentru o colecţie mult mai mare, vezi[2] şi mai ales[1]). Se observă că testele diagnostice pe care le folosim (de la semne clinice la teste paraclinice) sunt departe de a fi perfecte în privinţa reproductibilităţii. Chiar şi în privinţa analizelor de laborator, în care clinicianul nu are decât de interpretat un număr, dezacordul dintre investigatori este încă posibil (de exemplu, într-un studiu în care trei endocrinologi au văzut aceleaşi rezultate privind funcţia tiroidiană şi alte date clinice privind 55 de pacienţi consecutivi la care se suspecta boală tiroidiană, ei au fost în dezacord cu privire la diagnosticul final în 40% din cazuri[3]. Mai mult, nici măcar analiza computerizată a rezultatelor nu are o reproductibilitate mai bună: într-un studiu asupra unor perechi de electrocardiograme făcute la interval de un minut la 92 de pacienţi, interpretarea computerului a fost semnificativ diferită în 40% din cazuri, chiar dacă traseele nu difereau[4].

Tabelul XVII.2. Acordul interobservator – câteva exemple în cazul semnelor fizice (din[1]).

Semn	Coeficient κ
Piele	
Paloare (pacientul pare anemic)[5, 6]	0,23-0,48
Paloare conjunctivală[7]	0,54-0,75
Cianoză[5,8]	0,36-0,70
Icter[9]	0,65
Semne vitale	
Hipotensiune (TA sistolică < 90mmHg)	0,90
Febră (evaluată prin palparea pielii)[5]	0,09-0,23
Tahipnee[8]	0,25
Retinopatie diabetică[10, 11]	
Exudate	0,56-0,67
Hemoragii intraretiniene	0,89
Neovascularizaţie	0,1-0,48

Stadializare	0,65
Matitate la percuţia pulmonară[8,12, 13]	0,16-0,52
Raluri alveolare[8, 12, 14]	0,21-0,63
Raluri sibilante[8, 12, 14, 15]	0,43-0,93

Tabelul XVII.3. Acordul interobservator – câteva exemple în cazul testelor diagnostice (din[1]).

Semn (la testul diagostic)	Coeficient κ
Radiografie toracică	
Cardiomegalie[16]	0,48
Redistribuţia circulaţiei pulmonare[16]	0,50
Fibroza pulmonară (pe o scală cu 4 grade)[17]	0,45
Venografie cu substanţă de contrast	
Tromboză venoasă profundă[18]	0,53
Angiografie cu subtracţie digitală	
Stenoză de arteră renală[19]	0,65
Coronarografie	
Clasificarea leziunilor arterelor coronare[20]	0,33
Tomografia computerizată cerebrală[21]	
Normală sau anormală, la pacienţi cu accident vascular cerebral	0,60
Leziune pe partea dreaptă sau stângă, la pacienţi cu accident vascular cerebral	0,65
Efect de masă, prezent sau absent	0,52
Tomografa computerizată toracică	
Stadializarea cancerului pulmonar[22]	0,40-0,60
Rezonanţă magnetică nucleară crebrală	
Compatibilă cu scleroză multiplă[23]	0,57-0,87
Rezonanţă magnetică nucleară a coloanei vertebrale	
Modificări ale discului interbertebral sau normal[24]	0,59
Ecografie	
Tromboză venoasă profundă, prezentă sau absentă[25]	0,69
Nodul tiroidian, prezent sau absent[26, 27]	0,57-0,66
Examenul histopatologic al biopsiei hepatice[28]	
Colestază	0,40
Boală alcoolică a ficatului	0,49
Ciroză	0,59

Prin efectuarea unui test, clinicianul se străduieşte să clasifice starea reală dar necunoscută a unui subiect de observaţie cu ajutorul unui instrument imperfect.
Sursele de nesigutanţă/variabilitate sunt:

1. Instrumentul de măsură: imprecizia analitică (acelaşi test aplicat aceluiaşi pacient să dea acelaşi rezultat).
2. Variabilitatea subiectului: intraindividuală (regresia către medie – fluctuaţii fiziologice)/ interindividuală (cu cât acestă variabilitate este mai mare, avem nevoie de un eşantion mai mare pentru studiul nostru).
3. Variabilitatea interpretării: interindividuală (doi radiologi sau ecografişti văd lucruri diferite la acelaşi pacient) / intraindividuală (acelaşi radiolog/anatomopatolog vede lucruri diferite citind aceeaşi radiografie/lamă în momente diferite; studiile au arătat că aceste diferenţe sunt incredibil de mari). Această variabilitate este cuantificată prin mărimea coeficientului de concordanţă κ.
4. Validitatea intrinsecă a testului (sensibilitatea, specificitatea, raportul de probabilitate=*likelihood ratio*).
5. Prevalenţa bolii = probabilitatea pretest, care influenţează valorile predictive ale testului conform teoremei lui Bayes.

Ca la orice tip de studiu, şi la cele terapeutice ne interesează două lucruri: validitatea (calitatea metodologică) şi rezultatele. Este necesar ca în primul rând studiul să fie corect efectuat, pentru a ne putea baza pe rezultate.
Evaluarea validităţii studiului se face prin verificarea următoarelor criterii:

1. **A fost testul comparat cu un *gold standard* adevărat? A fost comparaţia "oarbă"?**

Gold standard-ul este testul etalon, cu care comparăm orice test nou şi în funcţie de care îl evaluăm pe acesta. Se presupune că *gold standard*-ul este testul perfect, care identifică toţi indivizii care au boala şi nu dă rezultate fals pozitive sau negative – de obicei este examenul histopatologic (în coronaropatii, este coronarografia). Există boli pentru care nu avem *gold standard* – de exemplu bolile pentru care avem criterii de diagnostic. În aceste cazuri trebuie căutat un alt tip de *gold standard,* cum ar fi evoluţia clinică (pentru artrita reumatoidă, de exemplu, pentru care nu avem un test perfect). Deloc surprinzător, de cele mai multe ori tocmai pentru astfel de boli fără *gold standard* se caută un test diagnostic bun. Acolo unde avem *gold standard* evaluăm teste noi fie pentru că sunt mai ieftine, fie mai puţin invazive, fie mai uşor de efectuat. Atunci când testul etalon este imperfect, şi evaluăm un test nou în funcţie de acesta, dacă testul nou este mai bun, prin natura studiilor diagnostice acest fapt nu va putea fi evidenţiat, iar testul nou întotdeauna apare ca fiind mai slab decât *gold standard*-ul.
Testul de evaluat trebuie să fie independent de *gold standard* – de exemplu, nu putem evalua anticorpii antiADN dublu catenar ca test diagnostic în lupus, având ca *gold standard* criteriile ACR pentru lupus, deoarece printe criterii se află şi aceşti autoanticorpi. Iar la evaluarea criteriilor de diagnostic pentru arterita cu celule gigante[29], *gold standard*-ul folosit a fost… părerea experţilor. Oare pe ce s-o fi bazat părerea acestora, dacă nu (măcar parţial) tot pe simptomele şi testele cuprinse în aceste criterii! Aşadar, atenţie la ce *gold standard* a fost folosit, şi cât de valid este acesta, pentru a aprecia validitatea studiului!
De asemenea, comparaţia dintre *gold standard* şi testul de evaluat trebuie să fie oarbă – adică cei care au efectuat şi interpretat testul de evaluat nu trebuie să ştie rezultatul *gold*

standard-ului şi invers. După ce medicii află de existenţa unui nodul pulmonar apărut la CT, îl vor vedea şi pe radiografie, iar după ecocardiografia care a arătat o regurgitare aortică, normal că vor auzi şi suflul! Astfel se introduce o eroare sistematică în favoarea testului de evaluat. Bineînţeles că, la fel ca la studiile terapeutice, cu cât testul este mai subiectiv (examen fizic, anamneză, radiologie, ecografie, chiar examen histologic – am vazut cât de mare este variabilitatea inter/intraobservator), cu atât mai mult este nevoie de precauţii care să asigure orbirea. Pentru testele biochimice – care nu pot fi influenţate de cel care le interpretează – lipsa orbirii nu prea poate introduce erori sistematice în favoarea acurateţei testului[30] - deşi nu este imposibilă înterpretarea subiectivă a rezultatelor biochimice[3]!

2. A fost testul evaluat la pacienţii potriviţi?

Pacienţii pe care efectuăm studiul diagnostic trebuie să fie asemănători acelora la care l-am folosi în practică. Un test diagnostic este util atunci când face diferenţa între afecţiuni asemănătoare, între care fără el nu am prea putea face diagnosticul diferenţial. Oricine îşi dă seama că un pacient în vârstă, cu istoric de infarct anterior, ortopnee, subcrepitante, edeme, jugulare turgescente, cardiomegalie şi galop protodiastolic are insuficienţă cardiacă, iar o tânără cu crize de wheezing are altă cauză a dispneei, nu este nevoie pentru asta să dozăm peptidul natriuretic. La fel, un test pentru artrita reumatoidă ne este necesar la pacienţi care au poliartrită de scurt timp, şi ne întrebăm dacă este vorba despre o artrită reumatoidă sau o altă boală inflamatorie articulară. Dacă, pentru a evalua un test diagnostic se folosesc, ca bolnavi, pacienţi cu aspect clar al bolii respective, iar ca indivizi neavând boala oameni sănătoşi, testul va apărea cu o putere discriminativă mult mai mare decât în realitate. Aşadar, trebuie ca testul să fie evaluat la pacienţi consecutivi la care boala respectivă este suspectată.

3. A fost efectuat *gold standard*-ul la toti pacienţii, indiferent de rezultatul testului evaluat?

După cum am mai spus, presupunând că testul etalon este perfect, avem totuşi nevoie să dezvoltăm alte teste, mai ieftine, mai puţin laborioase sau mai puţin invazive decât acesta. Nu putem face coronarografie la toţi pacienţii pentru că este o investigaţie invazivă şi disponibilă în puţine centre (la noi în ţară), care necesită personal specializat, de asemenea preferăm să avem un alt test, imagistic prin care să diferenţiem formaţiunile intrahepatice decât examenul histopatologic al fragmentului obţinut prin biopsie ţintită. Din acest motiv căutăm să dezvoltăm alte teste, mai simple, cum ar fi computerul tomograf pentru bolile coronariene, leziunile colonice sau trombembolismul pulmonar, RMN pentru afecţiunile coledocului - în loc de coronarografie, colonoscopie, scintigrafie/arteriografie, colangiografie retrogradă endoscopică etc. - testele etalon. În studiile care consacrează aceste noi investigaţii, pentru a putea face comparaţia testului nou cu *gold standard*-ul, trebuie ca tuturor pacienţilor să li se aplice concomitent şi testul de evaluat, şi *gold standard*-ul şi nu, de exemplu, dacă pacientului nu-i iese nimic la colangio-RMN, să nu-i mai facem colangiografia retrogradă endoscopică, sau dacă tomografia cu emisie de pozitroni nu arată că leziunea este malignă, să nu-i mai facem biopsie din tumoră prin puncţie sau prin operaţie, ori dacă anticorpii antipeptid citrulinat (antiCCP) sunt negativi, să ne spunem că oricum nu părea o artrită reumatoidă şi să nu mai urmărim pacientul un an (evoluţia clinică fiind, înacest caz, *gold standard*-ul), deci să considerăm că pacientul nu are boala, fără să-i mai aplicăm *gold standard*-ul. Ori în toate aceste cazuri am omite cazurile fals negative şi testul de evaluat va părea, din nou, mai bun, cu o sensibilitate mult mai mare decât în realitate. Invers, dacă pacienţii la care

testul de evaluat iese pozitiv îi considerăm ca având boala, renunţând să le mai facem testul *gold standard*, omitem fals negativii, şi atunci testul va părea că are o specificitate mult mai mare decât în realitate.
Când pacienţii au un test diagnostic negativ, investigatorii sunt tentaţi să nu mai aplice *gold standard*-ul, iar când acesta este invaziv sau riscant (de exemplu angiografie) poate că nici nu merită să fie făcut la pacienţii la care testul de evaluat a fost negativ. Pentru a trece peste această deficienţă, investigatorii pot utiliza în aceste cazuri un nou standard de referinţă pentru a demonstra că pacienţii într-adevăr nu au avut boala – de exemplu evoluţia clinică fără probleme în absenţa tratamentului. O dovadă convingătoare că un pacient cu suspiciune clinică de tromboză venoasă profundă nu a avut-o include lipsa oricărei complicaţii pe durata unei urmăriri îndelungate fără tratament anticoagulant[31] - ceea ce, de fapt, nu constituie de loc o dovadă că pacientul nu a avut boala. În schimb, este o evaluare mult mai pragmatică a valorii unui test diagnostic – nici nu mă interesează să descopăr o tromboză care nu face rău, pentru că nu este nevoie să o tratez – aşa s-a procedat şi în celebrul studiu PIOPED[32].
De obicei aplicăm testele pe bolnavi, şi atunci, în lipsa unui studiu diagnostic bun, ne bizuim pe aceste rezultate deformate de lipsa cunoştinţelor privind rezultatele testului la pcienţii care nu au boala. Rezonanţă magnetică nucleară (RMN), de exemplu, se face pacienţilor cu dureri la nivelul coloanei lombare, şi se descoperă tot felul de modificări, care au determinat explozia intervenţiilor chirurgicale la acest nivel la începutul anilor 90. Întrebarea legitimă este: cum arată coloana la indivizi fără dureri, şi pentru a răspunde s-a efectuat RMN la 98 de voluntari care nu se plângeau de nimic[33]. Radiologii care au citit rezultatele nu ştiau asta (erau "orbi"), şi au descoperit la fel de des protruzii discale ca şi la pacienţii care făceau RMN pentru dureri (şi care fuseseră, probabil, operaţi de mult…).

4. **S-a demonstrat că testul este reproductibil inter/intraobservator?**

Dacă un medic efectuează un test de două ori asupra unui subiect a cărui condiţie nu s-a schimbat, într-o anumită proporţie de cazuri el va obţine rezultate diferite – acest lucru este valabil pentru toate testele şi toţi investigatorii, numai că gradul de concordanţă variază – una este să fie de 99%, alta 50%. Cu atât mai mult variază rezultatele când testul este efectuat de investigatori diferiţi. Aşadar, înainte de a începe evaluarea unui test, trebuie să vedem dacă reproductibilitatea lui este acceptabilă (nu este mai puţin adevărat că acelaşi lucru se întâmplă şi cu *gold standard*-ul – am mai spus, există variabilitate intra- şi interobservator şi la citirea aceleiaşi lame histologice; în aceste cazuri se poate apela la citirea testului de către doi sau mai mulţi investigatori, "orbi" unul faţă de celălalt).

5. **Au fost furnizate intervalele de încredere pentru sensibilitate, specificitate şi ceilalţi parametri ai testului?**

6. **Este furnizat raportul de probabilitate (Likelihood ratio) al testului, sau datele din care acesta poate fi calculat?**

Bibliografie

[1]. McGee S. Evidence-based physical diagnosis. WB Saunders, Philadelphia, 2001. p.33-50.

[2]. Sackett DL, Haynes B, Guyatt G, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd edition, Little, Brown. Toronto. 1991. p. 25-35.

[3]. Jarlov AE et al. Observer variation in the clinical and laboratory evaluation of patients with thyroid dysfunction and goiter. *Thyroid* 1998; 8:393-398.

[4]. Spodick DH, Bishop RL. Computer treason: ntraobserver variability of an electrocardiographic computr system. *Am J Cardiol* 1997; 80:102-103.

[5]. Gjorup T et al. Global assessment of patients – a bedside study. II. Interobserver variation and frequency of clunical findings. *J Intern Med* 1990; 228:147-150.

[6]. Gjorup T et al. A critical evaluation of the clinical diagnosis of anemia. *Am J Epidemiol* 1986; 124:657-665.

[7]. Sheth TN et al. The relation of conjunctival pallor to the presence or absence of anemia. *J Gen Intern Med* 1997; 12:102-106.

[8]. Spiteri MA, Cook DG, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet* 1988; 2:873-875.

[9]. Espinoza P et al. Interobserver agreement in the physical diagnosis of alcoholic liver disease. *Dig Dis Sci* 1987; 32:244-247.

[10]. Milton RC, Ganley JP, Lynk RH. Variability in grading diabetic retinopathy from stereo fundus photographs: comparison of physician and lay readers. *Br J Ophtalmol* 1977; 61:192-201.

[11]. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs: an extension of the modified Airlie House classification. ETDRS report number 10. *Ophtalmology* 1991; 98:786-806.

[12]. Mulrow CD et al. Observer variation in the pulmonary examination. *J Gen Intern Med* 1986; 94:188-196.

[13]. Gjorup T, Bugge PM, Jensen AM. Interobserver variation in assessment of respiratory signs: physicians' guesses as to interobserver variation. *Acta Med Scand* 1984; 216:61-66.

[14]. Holleman DR, Simmel DL, Goldberg JS. Diagnosis of obstrucive airways disease from the clinical examination. *J Gen Intern Med* 1993; 8:63-68.

[15]. Badgett DG et al. Can moderate chronic obstructive pulmonary disease be diagnosed by historical and physical findings alone? *Am J Med* 1993; 94:188-196.

[16]. Butman SM et al. Bedside cardiovascular examination in patients with severe chronic heart failure: Importance of rest or inducible jugular venous distension. *J Am Coll Cardiol* 1993; 22:968-974.

[17]. Baughman RP et al. Crackles in interstitial lung disease: comparison of sarcoidosis and fibrosing alveolitis. *Chest* 1991; 100:96-101.

[18]. Illescas FF et al. Interobserver variability in the interpretation of contrast venography, technetium-99m red blood cell venography and impedance plethysmography for deep venous thrombosis. *J Can Assoc Radiol* 1990; 41:264-269.

[19]. DeVries ar et al. Interobserver variability in assessing renal artery stenosis by digital subtraction angiography. *Diagn Imag Clin Med* 1984; 53:277-281.

[20]. Herman JPR et al. Inter- and intra-observer variability in the qualitative categorization of coronary angiograms. *Int J Card Imag* 1996; 12:21-30.

[21]. Shinar D et al. Interobserver reliability in the interpretation of computed tomographic scans of stroke patients. *Arch Neurol* 1987; 44:149-155.

[22]. Webb WR et al. Interobserver variability in CT and MR staging of lung cancer. *J Comput Assist Tomogr* 1993; 17:841-846.

[23]. Barkhof F et al. Interobserver agreement for diagnostic MRI criteria in suspected multiple sclerosis. *Neuroradiology* 1999; 41:347-350.

[24]. Jensen MC et al. Magnetic resonance imaging of the lumbar spine in eople without back pain. *N Engl J Med* 1994; 331:69-73.

[25]. Atri M et al. Accuracy of sonography in the evaluation of calf deep vein thrombosis in both postoperative surveillance and symptomatic patients. *Aqm J Radiol* 1996; 166:1361-1367.

[26]. Jarlov AE et al. Observer variation in ultrasound assessment of the thyroid gland. *Br J Radiol* 1993; 66:625-627.

[27]. Schneider AB et al. Thyroid nodules in the follow-up of irradiated individuals: Comparison of thyroid ultrasound with scanning and palpation. *J Clin Endocrinol Metab* 1997; 82:4020-4027.

[28]. Theodossi A et al. Observer variation in assessment of liver biopsies including analysis by kappa statistics. *Gastroenterology* 1980; 79:232-241.

[29]. Hunder GG, Bloch DA, Michel BA et al. The American College of Rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis Rheum.* 1990; 33:1122-8.

[30]. Meissner K, Distel H, Mitzdorf U. Evidence for placebo effects on physical but not on biochemical outcome parameters: a review of clinical trials. *BMC Med*. 2007; 5:3

[31]. Strauss SE, Richardson S, Glasziou P, Haynes B. Evidence-based medicine. How to practice and teach EBM. 3rd edition, Elsevier, London, 2005. p.73.

[32]. The PIOPED investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA* 1990; 263: 2753-59.

[33]. Jensen MC, Brant-Zawadzki MN, obuchowski N, Modic MT, Malkasian D, Ross JS. Magnetic resonance imaging of the lumbar spine in people withot back pain. *N Engl J Med* 1994; 331: 69-73.