Klinička procjena medicinskih pretraga: još je dug put pred nama

Clinical evaluation of medical tests: still a long road to go

Patrick MM Bossuyt

Odsjek kliničke epidemiologije, biostatistike i bioinformatike, Klinički bolnički centar Sveučilišta u Amsterdamu

Dept. Clinical Epidemiology, Biostatistics & Bioinformatics, Academic Medical Center, University of Amsterdam

U doba medicine temeljene na dokazima i medicinske pretrage podliježu pažljivom ispitivanju. Prije uvođenja pretraga u praksu treba podrobno i kritički ispitati sve navode o vrijednosti podataka koje pružaju te o njihovoj korisnosti. Potrebna su vrlo kvalitetna istraživanja koje će nam dati dokaze o praktičnoj koristi dotične pretrage.

Glavna paradigma u kliničkoj procjeni pretraga je određivanje dijagnostičke točnosti pretrage. U istraživanjima dijagnostičke točnosti se rezultati jedne ili više pretraga uspoređuju sa zlatnim standardom. U svih bolesnika se izvodi dotična pretraga, kao i zlatni standard. Mjere dijagnostičke točnosti kazuju koliko dobro rezultati te pretrage odgovaraju "istini". Rezultati ovakve usporedbe mogu se izraziti na više načina, od kojih se najčešće u paru primjenjuju osjetljivost i specifičnost.

Nažalost, "istina" je neuhvatljiv koncept, a apsolutni zlatni standard rijetko postoji. Primjerice, kod venske trombembolije čak je i vrijednost autopsije kao zlatnog standarda upitna. Doista se možemo zapitati što je to "istina". Zbog tih razloga se pojam "istine" općenito zamjenjuje izrazima kao što su stanje od interesa, bolest te stadij ili tip bolesti. Temeljem dosadašnjih saznanja koncept "zlatnog standarda" zamjenjen je konceptom "referentnog standarda", odnosno najbolje raspoložive metode za utvrđivanje prisutnosti ili odsutnosti stanja od interesa.

Dugo je vremena prvenstvena namjena procjene dijagnostičke točnosti bilo izračunavanje osjetljivosti i specifičnosti, dok se malo pozornosti poklanjalo pitanjima ustroja i provođenja istraživanja. Meta-analize i sustavni pregledi vremenom su pokazali da su obilježja istraživanja od presudne važnosti te da loše provedena istraživanja mogu dovesti do precijenjenih procjena dijagnostičke točnosti (1,2,3). Najznačajniji utjecaj na rezultat istraživanja dijagnostičke točnosti ima izbor zdravih ispitanika za kontrolnu skupinu (4), kao i primjena dvaju ili više različitih referentnih standarda. U ispitivanjima koja su danas poznata kao istraživanja dijagnostike i liječenja (engl. *diagnostic management studies*) mogući su različiti načini potvrđivanja rezultata pretrage (engl. *differential verification*). Kod takvih strategija liječnici radi potvrde rezultata neke pret-

In this era of evidence-based medicine, medical tests cannot escape close scrutiny. Before tests are put into practice, claims about their information value and usefulness should be thoroughly questioned and critically tested. We need high quality studies to provide us with evidence of a test's practical use.

The dominant paradigm in the clinical evaluation of tests is determination of a test's diagnostic accuracy. In studies of diagnostic accuracy, the results of one or more tests are compared with the gold standard. All patients receive the index test and all receive the gold standard. Measures of diagnostic accuracy express how well the results of the index test correspond with the 'truth'. The results of such a comparison can be expressed in a number of ways, of which sensitivity and specificity pairs are most frequently used.

Unfortunately, 'truth' is an elusive concept and an untarnished gold standard seldom exists. In venous thromboembolism, for example, even autopsy has been distrusted and questioned as the gold standard. One can very well question what 'the truth' is. For these reasons, the notion of 'truth' has been generally replaced by 'target condition', the disease, disease stage or disease type, or condition for which the test is applied. Analogously, test research has replaced the concept of the 'gold standard' by that of the 'reference standard', i.e. the best method available for establishing the presence or absence of the target condition.

For a long time, the primary purpose of accuracy tests was the calculation of sensitivity and specificity, with little attention paid to the issues of study design and execution. It gradually became clear through meta-analytical studies and systematic reviews that design features were crucially important and that suboptimal studies could lead to inflated estimates of diagnostic accuracy (1,2,3). Of particular concern were the inclusion of healthy controls (4) and the use of two or more reference standards. Differential verification is often used in what has become known as diagnostic management studies. In such strategies physicians base their decisions on further investigarage donose svoje odluke o daljnjim dijagnostičkim postupcima na temelju rezultata te pretrage ili pak na kliničkoj procjeni. Bolesnici s pozitivnim rezultatom pretrage upućuju se na daljnje pretrage, dok se oni s negativnim rezultatom šalju kući uz daljnje praćenje. Potom se pozitivni rezultati potvrđuju, primjerice, nalazima pretraga slikovnog prikazivanja koji ukazuju na stanje od interesa, dok se negativni rezultati potvrđuju normalnim nalazima tijekom praćenja.

U mnogim je radovima vrlo teško steći pravi uvid u ustroj istraživanja. Ostaje još puno prostora za unaprjeđenje radova o dijagnostičkoj točnosti (5). S tim u vezi međunarodna je grupa urednika, autora i drugih stručnjaka izradila Standarde za ustroj, provođenje i način prikazivanja rezultata istraživanja dijagnostičke točnosti. STARD smjernice su od 2003. godine objavili mnogi časopisi (6,7), od kojih je većina smjernice uklopila u svoje *Upute za autore* i kao takve postale su obvezne za radove o dijagnostičkoj točnosti. Nekoliko godina nakon objavljivanja grupa je autora procijenila učinke uvođenja STARD smjernica. Autori te procjene zabilježili su poboljšanje kvalitete radova o dijagnostičkoj točnosti, no još uvijek daleko od optimalnog (8).

Vrlo je teško procijeniti dijagnostičku točnost neke pretrage, ukoliko referentni standard ne postoji ili je nedostatan. Postoje načini kako je moguće ispraviti rezultate istraživanja u kojem je korišten nesavršeni referentni standard. Neki od njih su tzv. *latent class analysis*, *panel-based* metode te metode koje umjesto potvrde (engl. *verification*) koriste procjenu (engl. *validation*) rezultata pretrage.

Prevladavajući je stav da je točnost ključna za kliničku procjenu pretrage, no treba znati kako ona ipak nije dostatna. Testovi mogu biti točni, ali mi uz to želimo znati kakva je točnost testa u usporedbi s onom drugih testova, jer je to pokazatelj koji određuje buduću ulogu tog testa (9). Još je važnija tzv. *dodana vrijednost* pretrage, odnosno u kojoj mjeri dotična pretraga može smanjiti preostalu nesigurnost. Ne postoje općeprihvaćene metode za izražavanje ove dodane vrijednosti. Jedan je način osloniti se na dojam kliničara, tj. osobnu procjenu kliničke značajnosti neke pretrage. Nažalost, taj je način vrlo subjektivan i bitno ovisi o prethodnom očekivanju vrijednosti pretrage. U dobro ustrojenom istraživanju ne bi trebalo uzimati u obzir takve subjektivne procjene.

Isto tako nije prihvatljiva odluka liječnika o promjeni terapijskog pristupa bolesniku ukoliko se temelji na subjektivnoj procjeni kliničke značajnosti pretrage. Takvu procjenu liječnici često rabe, primjerice, u procjeni funkcijske MRI pri postavljanju dijagnoze konvulzivnih bolesti (10).

Na kraju, klinička značajnost primjene nove dijagnostičke pretrage uvelike ovisi o mogućnosti te pretrage da poboljša stanje bolesnika u odnosu na prijašnju dijagnostičku pretragu. Istraživanja dijagnostičke točnosti testa ne moraju uvijek dati dovoljno informacija da bi imale kliničku

tions to verify test results on the test results themselves, or on clinical assessment. Patients with positive tests results are referred for further testing, whereas those with negative results are sent home and contacted for follow up. Positive results are then verified by, say, imaging findings that point to the target condition, whereas negative results are verified by an uneventful follow up.

In many study reports, it is hard to find out what the design of the study actually was. Reporting leaves much to be desired (5). To remedy this, an international group of editors, authors and others have developed the Standards for Reporting Diagnostic Accuracy studies. This STARD initiative was published in 2003 in a large number of journals (6,7). Since then, others have published it as well and many more have made it a requirement for authors that want to submit manuscripts about diagnostic accuracy studies. The introduction of the STARD statement was evaluated a few years after its publication. The authors of that evaluation noticed an improvement, but reporting of diagnostic accuracy studies is still far from optimal (8).

The accuracy of a test may be difficult to estimate if there is no or only a deficient reference standard. Several methods exist to correct the results of an imperfect reference standard. These include latent class analysis, panel-based methods and methods aimed at validation, instead of verification, of test results.

Accuracy, although being the dominant paradigm, is by no means sufficient for clinical evaluation of tests. Tests may be accurate, but we also want to know how the accuracy of a test compares to that of other tests, a comparison that will be guided by the future role of the test (9). What matters even more is their added value: to what extent are these tests able to reduce the remaining uncertainty? There are no accepted methods for expressing this added value of tests. Some have relied on clinicians' reports: personal expressions of the extent to which test results were perceived to be useful. Unfortunately, such expressions are subjective beyond repair, very much prone to influences and expectations about the value of the test. For sound research, such expressions are not to be used.

Similar objections apply to the intended changes in management, another way of eliciting from clinicians the effect tests have on clinical practice. Such methods have been used, for example, in the evaluation of functional MRI for the diagnostic evaluation of seizure disorders (10).

In the end, the clinical value of using a new diagnostic test will depend to a large extent on its ability to improve patient outcomes beyond the outcomes achieved using an old diagnostic test. Diagnostic test accuracy studies may not always provide sufficient information to infer clinical value. Recently, Lord et al. have argued that accuracy studies suffice if a new diagnostic test is safer or more specific than, but of similar sensitivity to, an old test (11).

značajnost. Nedavno su Lord i sur. objavili rad u kojem tvrde da su istraživanja dijagnostičke točnosti dostatna ako je nova dijagnostička pretraga sigurnija ili specifičnija, ali podjednake osjetljivosti kao stara pretraga (11). Ako je nova pretraga osjetljivija od stare, to će dovesti do otkrivanja dodatnih slučajeva bolesti. U tom slučaju rezultati iz literaturnih navoda o uspješnosti liječenja bolesnika otkrivenih starom pretragom ne moraju vrijediti, osim ako se može dokazati da nova pretraga otkriva isti spektar i podtipove bolesti kao stara pretraga ili da je odgovor na liječenje sličan u čitavom spektru bolesti.

Odluku o tome treba li neku pretragu uvesti u kliničku praksu najbolje je temeljiti na rezultatima randomiziranih kliničkih dijagnostičkih istraživanja. Kao što se bolesnici nasumce razvrstavaju u skupine koje će primiti lijek ili placebo, kako je to slučaj u ispitivanjima lijekova, u dijagnostičkim se istraživanjima bolesnici nasumce razvrstavaju u dvije skupine od kojih jedna sadrži bolesnike kojima se ispitivana pretraga izvodi, dok se u drugoj skupini bolesnicima ta pretraga ne izvodi. Moguć je i takav ustroj u kojemu se bolesnici također nasumce razvrstavaju u dvije skupine u kojima se bolesnicima izvode dvije različite pretrage. Potom se bolesnici prate i uspoređuju se primarne mjere ishoda u dotičnim skupinama.

Nažalost, mnogo je otvorenih pitanja kod takvih dijagnostičkih ispitivanja (12). Prvo, ovakva ispitivanja ne omogućuju stvarnu procjenu pretrage, nego procjenjuju zajednički učinak pretrage i liječenja. Bolesnicima se stanje ustvari ne mijenja zbog izvedene pretrage već zahvaljujući odlukama o liječenju temeljenim na rezultatima te pretrage. To znači da protokol za daljnje donošenje odluka na osnovi rezultata pretrage treba unaprijed utvrditi. Ako se to ne učini, ispitivanje se ne može valjano procijeniti.

Nadalje, istraživanja zajedničkog učinka pretrage i liječenja obično zahtijevaju veliki broj ispitanika. Za razliku od ispitivanja lijekova, gdje svi bolesnici dobivaju lijek ili placebo, u dijagnostičkom ispitivanju će se skupine razlikovati zbog nesukladnih rezultata pretraga, što se najčešće odnosi na manji broj ispitanika.

Neki autori predlažu stupnjevitu procjenu učinkovitosti pretraga. Prvo valja ispitati dijagnostičku točnost, potom dodanu vrijednost te implikaciju za liječenje. Naposlijetku treba ispitati kako rezultat pretrage utječe na ishod bolesnika te načiniti ekonomsku procjenu. Fryback i Thornbury su izradili jednu od najpoznatijih shema za takvu stupnjevitu procjenu (13). U literaturi se može naći barem još dvadesetak drugih prijedloga.

Ove stupnjevite procjene, osmišljene po uzoru na ispitivanja lijekova u četiri faze, izgledaju primamljivo, ali im nedostaje bitno obilježje, ono koje dijagnostiku razlikuje od svih drugih domena zdravstvene skrbi. Pretrage daju informacije, a te se informacije – same po sebi ili u kontekstu rezultata drugih pretraga, bolesnikove anamneze i fizikalnog pregleda, vrednuju izolirano, a ne prema njiho-

If a new test is more sensitive than an old test, it leads to the detection of extra cases of disease. In these cases, results from treatment trials that enrolled only patients detected by the old test may not apply to these extra cases, unless one can show that the new test detects the same spectrum and subtype of disease as the old test or that treatment response is similar across the spectrum of disease.

Diagnostic randomized clinical trials of tests may seem to be a perfect solution to all problematic attempts to collect evidence for sound decision making about the introduction or propagation of tests in clinical trials. Instead of random patient allocation to active treatment or placebo, as in drug trials, patients are randomly allocated to testing or no testing, or one type of testing versus a second test. One then follows patients and compares the primary outcome measures of the respective groups.

Unfortunately, there are a number of issues with such diagnostic trials (12). To start, such trials are not strict evaluation of a test, but of test-treatment combinations. Patients seldom improve from a test itself – they will get better from the management decisions that are based on this test. This implies that the protocol for subsequent decision making based on the test results should be specified in advance. If not, the trial cannot be properly evaluated.

Another consequence of the test-treatment combinations is that diagnostic trials will usually also need large sample sizes. Unlike during trials, where all patients receive either the active treatment or placebo, the difference between the groups in a diagnostic trial will be generated by the group of discordant test results, which will usually be only a minority of those tested.

Several authors have proposed a staged evaluation of the efficacy of tests, in which tests move from an evaluation of their accuracy to studies of the added value and implications for management, and ultimately to studies of patient outcomes and economic evaluation. One of the best known evaluation schemes is the one put forward by Fryback and Thornbury (13). At least two dozen other proposals can be found in the literature.

These staged evaluations, modeled after the four-phase approach to drug trials, seem attractive, but they seem to lack an essential feature of tests, one that distinguishes them from many other actions in health care. Tests generate information and that information, either by itself or in context with other test results and elements from patient history and physical examination, is valued in itself, not only for its implications for the management (14). When we feel ill, we usually would like to know why it has happened and when we will get better. If we have complaints, many of us would like to know what has caused them.

The desire for knowledge, especially about one's own condition, is an essential feature of human beings. The desire vim implikacijama za liječenje (14). Kad smo bolesni, obično želimo znati zašto smo se razboljeli i kada će nam biti bolje. Imamo li kakvih tegoba, najčešće želimo znati što ih je uzrokovalo.

Želja za saznanjima, poglavito o vlastitom stanju, bitno je obilježje ljudskih bića. Težnji za preživljenjem, za stabilnošću jednaka je želja za znanjem i shvaćanjem. I za postizanjem kontrole.

Još uvijek nema zadovoljavajućeg načina kako procijeniti pretragu i pri tome obuhvatiti sva ova otvorena pitanja. Još je dug put pred nama.

Adresa za dopisivanje:

Patrick MM Bossuyt
Dept. Clinical Epidemiology, Biostatistics & Bioinformatics
Academic Medical Center, University of Amsterdam
Room J1b-212; PO Box 22700; 1100 DE Amsterdam, The Netherlands
e-pošta: *p.m.bossuyt@amc.uva.nl*
tel: +31(20)566 3240 (voice)
Fax: +31(20)691 2683(fax)

for survival, for stability is matched by the desire to know, to understand. And to control.

The evaluation of tests has not yet been extended to cover these other issues – and all previous – in a satisfactory way. We still have a long road to go.

Corresponding author:

Patrick MM Bossuyt
Dept. Clinical Epidemiology, Biostatistics & Bioinformatics
Academic Medical Center, University of Amsterdam
Room J1b-212; PO Box 22700; 1100 DE Amsterdam, The Netherlands
e-mail: *p.m.bossuyt@amc.uva.nl*
Phone: +31(20)566 3240 (voice)
Fax: +31(20)691 2683(fax)

Literatura / References

1. *Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Meulen JHP van der, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.*
2. *Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.*
3. *Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-76.*
4. *Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem 2005;51:1335-41.*
5. *Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347-53.*
6. *Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Clin Chem 2003;49:1-6.*
7. *Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.*
8. *Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006;67:792-7.*
9. *Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.*
10. *Medina LS, Bernal B, Dunoyer C, Cervantes L, Rodriguez M, Pacheco E, Jayakar P, Morrison G, Ragheb J, Altman NR. Seizure disorders: functional MR imaging for diagnostic evaluation and surgical treatment – prospective study. Radiology 2005; 236:247-53.*
11. *Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850-5.*
12. *Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844-7.*
13. *Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.*
14. *Asch DA, Patton JP, Hershey JC. Knowing for the sake of knowing: the value of prognostic information. Med Decis Making 1990;10:47-57.*