35. Genç M, Mårdh P-A. A cost-effectiveness analysis of screening and treatment for *chlamdya trachomatis* infection in asymptomatic women. Ann Intern Med 1996; 124: 1-7.
36. Muir Gray JA (ed). Evidence-based healthcare. How to make health policy and management decisions. Churchill Livingstone, Edinburgh; 1997.
37. Haines A, Donald A. Getting research findings into practice. BMJ Publishing Group, 1998.
38. Price CP. Health technology assessment. RCPath Bull 2000; 110: 26-28.
39. Williams O. What is clinical audit? Ann R Coll Surg Engl 1996; 78: 406-411.
40. Lord J, Littlejohns P. Evaluating healthcare policies: the case of clinical audit. BMJ 1997; 315: 668-671.

Ned Tijdschr Klin Chem 2001; 26: 242-245

Pijlers van precisie: maten en getallen voor het onderbouwen van testgebruik

P. M.M. BOSSUYT

"*In God we trust, the others must provide sound data*". Dit is een tijd waarin weinigen nog op hun woord alleen worden geloofd. Alom klinkt de roep om onderbouwing van stellingen en de vraag naar rekenschap en verantwoording. Het liefst ziet men in de antwoorden maat en getal opduiken. Ook de professionals in de zorg ontkomen niet aan deze tendens. Bewust van alle vertekenende invloeden gaan zij op zoek naar gegevens die zich lenen voor het onderbouwen van besluiten over het te voeren beleid. Zie daar de basis voor 'Evidence based Medicine': een professioneel antwoord op een vraag naar onderbouwing, in een atmosfeer van toegenomen rekenschap en verantwoording (1). Worden patiënten hier beter van? Zo ja, in welke mate? Staat die verbetering in een redelijke verhouding tot wat die patiënten zelf, hun artsen en de maatschappij aan middelen moet investeren in die gezondheidswinst c.q. behoud van gezondheid? Dergelijke vragen zijn niet ongewoon bij nieuwe geneesmiddelen. Medische tests ontsnappen echter niet aan vergelijkbare verzoeken. Of het nu gaat om beeldvormend onderzoek of laboratoriumtesten, de vraag naar het waarom en waarvoor zal ook daar worden gesteld. In welke mate is voor een test een antwoord beschikbaar op deze vragen naar onderbouwing? Een verkenning hiervan kan enkel maar aanleiding geven tot gepaste bescheidenheid. Hieronder volgt een korte inleiding.

Sensitiviteit en specificiteit

De eerste vraag die wordt gesteld is: kan ik wel varen op de uitslagen van deze test? We gaan voor het gemak uit van de discussie dat evidente vragen over de veiligheid, de ijking en de betrouwbaarheid al naar tevredenheid zijn beantwoord. Spreekt de test de waarheid? Voor het antwoord hierop worden de uitslagen van de test die wordt geëvalueerd – laten we die de indextest noemen – vergeleken met die van een referentiestandaard. De mate van overeenkomst tussen de uitslagen kan op verschillende manieren worden uitgedrukt. Laten we als voorbeeld de evaluatie van D-dimer nemen, een test voor het aantonen c.q. uitsluiten van longembolie. Kline en collegae rapporteerden over deze test in JAMA (2). Ze hadden de test afgenomen bij 380 patiënten die met verdenking van longembolie op het "emergency department" van een van de deelnemende academisch ziekenhuizen waren gezien. De test was positief bij 164 van hen. Hiervan kon de diagnose bij 60 worden bevestigd: deze 60 hadden ook een positieve uitslag met de referentiestandaard. (Tabel 1). Omgekeerd hadden van de 216 patiënten met een negatieve uitslag op de D-dimer er 4 uiteindelijk toch een longembolie. De beste manier (op pathologie na) voor het aantonen c.q. uitsluiten van longembolie is longangiografie. Deze werd echter lang niet bij iedereen uitgevoerd. In het onderzoek werd een zogenaamde gemengde referentiestandaard toegepast. Dat betekent dat de diagnose longembolie kon worden geverifieerd door een "high-probability" V/Q scan, een afwijkende spiraal CT, een non-high V/Q scan of door overlijden tijdens de follow-up als niet kon worden uitgesloten dat dit het gevolg was van een veneuze tromboëmbolie. Een eerste blik op tabel 1 leert ons dat er een redelijke, maar verre van perfecte overeenkomst is tussen indextest en referentie standaard. Wie enkel zou varen op de uitslagen van deze indextest maakt een aantal fouten. Het is gebruikelijk om het percentage correcte uitslagen conditioneel op de geverifieerde ziektestatus uit te drukken.

Tabel 1. De uitslagen van een onderzoek naar de diagnostische waarde van D-dimer bij het uitsluiten van longembolie

	Referentiestandaard		
Indextest	Positief	Negatief	Totaal
Positief	60	104	164
Negatief	4	212	216
	64	316	380

Data uit Kline et al. JAMA 2001; 285: 761-768.

Afdeling Klinische Epidemiologie en Biostatistiek, Academisch Medisch Centrum, Universiteit van Amsterdam

Correspondentie: Prof. Dr. P.M.M. Bossuyt, Afdeling Klinische Epidemiologie en Biostatistiek, Academisch Medisch Centrum, Postbus 22660, 1100 DD Amsterdam
e-mail: p.m.bossuyt@amc.uva.nl

De 60 gedetecteerde longembolieën (de terecht positieven) worden dan vergeleken met het totaal van 64 gevonden longembolieën. Die verhouding wordt de sensitiviteit genoemd: 60/64 levert 94% op. Omgekeerd worden de 212 terecht als negatief bestempelde patiënten vergeleken met het totale aantal patiënten zonder longembolie: 212 van de 316 betekent 67%. Dit percentage wordt de specificiteit genoemd. Sensitiviteit en specificiteit zijn dus het complement van de respectievelijke conditionele foutkansen. Het complement van de sensitiviteit geeft de conditionele kans op een fout-negatief resultaat (100-94 = 6%) en het complement van de specificiteit levert ons de kans op een fout-positief resultaat (100-67 = 33%). Het gebruik van sensitiviteit en specificiteit betekent dat we altijd twee termen nodig hebben om de test te karakteriseren. De lezer kan zich afvragen waarom niet kan worden volstaan met een enkel getal, bijvoorbeeld met het totale percentage fouten of het complement daarvan. In dit geval zou dit inhouden: 104 plus 4 fout geclassificeerde patiënten, een foutpercentage van 108 van de 380 ofwel 28%. Dat betekent dat 72% van de bestudeerde patiënten door de indextest, de D-dimer, goed is ingedeeld. Het totaal aantal foute classificaties is echter om meerdere redenen een minder gelukkige maat voor het karakteriseren van een test. We noemen er twee. Een eerste reden is dat het percentage fouten zal afhangen van de mate waarin de te detecteren toestand in de populatie voorkomt. In dit onderzoek hadden 64 van de 380 patiënten een longembolie, een percentage van 17%. In Nederlands onderzoek wordt de drempel voor het bepalen van een D-dimer vaak hoger gelegd. Als het onderzoek in Nederland was uitgevoerd, met een percentage longembolieën van 30% en precies dezelfde sensitiviteit en specificiteit, dan was het totale percentage fouten neergekomen op 0,30 * 0,06 plus (1-0,30)*0,33, ofwel 25% fout geclassificeerde patiënten. Naarmate de prevalentie van longembolie toeneemt in de onderzoeksgroep, daalt het totale percentage fouten, omdat de sensitiviteit veel hoger is dan de specificiteit. Een tweede reden om conditionele foutpercentages of hun complementen te gebruiken is het verschil in consequenties van het maken van fouten. Het missen van een longembolie is een fout, het onterecht besluiten tot een longembolie en starten van behandeling met een heparine-product is eveneens een fout. De eerste fout kan echter veel ernstiger gevolgen hebben dan de tweede: het niet tijdig herkennen van een longembolie heeft een veel grotere kans op een dodelijke afloop. Om dit verschil in consequenties weer te geven, is kennis van de conditionele foutkansen belangrijk. Er bestaan ook enkelvoudige maten om de mate van overeenkomst tussen indextest en referentietest te typeren die niet afhankelijk zijn van die prevalentie. Eén ervan is de onder epidemiologen vrij populaire odds ratio. De diagnostische toepassing van de odds ratio (verderop tot DOR afgekort) is gelijk aan de (conditionele) odds op een positief testresultaat bij ziekte gedeeld door de odds op een positief testresultaat in afwezigheid van ziekte. Geschat in deze onderzoeksgroep is dat dus:

$$\text{DOR} = \frac{\dfrac{P(+\mid D)}{1-P(+\mid D)}}{\dfrac{P(+\mid \textit{no}D)}{1-P(+\mid \textit{no}D)}} \qquad \text{DOR} = \frac{\dfrac{60}{4}}{\dfrac{104}{212}} = 30.6$$

Een bekende eigenschap van de diagnostische odds ratio is dat deze ook omgekeerd kan worden geïnterpreteerd en berekend, namelijk als de odds op ziekte bij een positief testresultaat ten opzichte van de odds op ziekte bij een negatief testresultaat.

$$\text{DOR} = \frac{\dfrac{P(D\mid +)}{1-P(D\mid +)}}{\dfrac{P(D\mid -)}{1-P(D\mid -)}} \qquad \text{DOR} = \frac{\dfrac{60}{104}}{\dfrac{4}{212}} = 30.6$$

De diagnostische odds ratio neemt waarden aan van 0 tot oneindig. Een test zonder waarde heeft een DOR van 1. Hogere waarden geven betere tests aan.

Interpretatie van testresultaten

Sensitiviteit, specificiteit en diagnostische odds ratio zijn maten om de kwaliteiten van de toepassing van een test te karakteriseren. Er bestaan echter ook maten om de interpretatie van een testresultaat te vergemakkelijken. Bij klinische toepassing van de D-dimer wil een arts op het “emergency departement” waarschijnlijk graag weten wat de kans is op een longembolie nadat het testresultaat bekend is geworden. Voor de groep als geheel kan opnieuw naar Tabel 1 worden gekeken. Daaruit leren we dat van de 164 patiënten met een positieve test er 60 een longembolie hebben: een percentage van 58%. Omgekeerd hebben 212 van de 216 patiënten met een negatief testresultaat geen longembolie: een percentage van 98%. Deze waarden worden ook wel de voorspellende waarden van de test genoemd De percentages van de hele groep zijn echter maar beperkt bruikbaar voor nieuwe toepassingen van de test in individuele patiënten. Ze zijn immers eveneens afhankelijk van het percentage longembolieën in de hele groep. In het individuele geval kunnen ze ook nog een keer een vertekend beeld geven. De mate van verdenking bij individuele patiënten kan immers sterk variëren: van een sterke verdenking bij iemand met een voorgeschiedenis van veneuze tromboëmbolie, bekend met een stollingsstoornis, tot een lichte verdenking bij iemand zonder die voorgeschiedenis en vage klachten. Voor individuele toepassingen van de test moet gekeken worden naar de informatiewaarden van de resultaten van de test. Die hangt af van de mate waarin de resultaten van de test vaker voorkomen bij patiënten met dan wel zonder de ziekte. Deze verhoudingen worden “likelihood ratios” of aannemelijkheidswaarden genoemd. Een positief resultaat van de D-dimer komt voor bij 60 van de 64 patiënten met een longembolie (94%) tegenover 104 van de 316 patiënten zonder longembolie (33%) bijna drie keer vaker dus bij patiënten met een longembolie. Omgekeerd hadden slechts 4 van de 64 patiënten met een longembolie

een negatief resultaat op de D-dimer (6%) tegenover 212 van de 316 patiënten zonder longembolie (67%): bijna elf keer vaker dus bij patiënten zonder een longembolie. Het Theorema van Bayes geeft aan welk verband er dient te bestaan tussen de mate van verdenking vóór dat de test werd aangevraagd (uitgedrukt als waarschijnlijkheid in de voorafkans) en de mate van verdenking nadat het testresultaat bekend is geworden (achterafkans). Voor een positief testresultaat wordt de achterafkans verkregen uit de relatie:

$$\frac{P(D|+)}{1-P(D|+)} = \frac{P(D)}{1-P(D)} \cdot \frac{P(+|D)}{1-P(+|D)}$$

ofwel

$$\frac{P(D|+)}{1-P(D|+)} = \frac{P(D)}{1-P(D)} \cdot LR(+)$$

Neem een patiënt met een 60% voorafkans op een longembolie (3 tegen 2) en een positief testresultaat (aannemelijkheidsverhouding 2,9 tegen 1). De kansverhouding achteraf wordt dan (3 tegen 2) maal 2,9, of 8,7 tegen 2: een achterafkans van 8,7 / (8,7+2) = 81%.

Problemen met de accuratesse van tests

Het berekenen van sensitiviteit en specificiteit is de standaardoplossing geworden voor het uitdrukken van de waarde van een test, al moet tegelijk worden gezegd dat die ‘sens & spec’ van heel wat vormen van diagnostiek nauwelijks bekend is. Jammer genoeg geeft het paradigma op zichzelf ook aanleiding tot veel verwarring. Wel begrijpelijk, maar grotendeels ook onnodig. In de best bekende vorm is de hiervoor genoemde referentietest de ‘gouden standaard’: een manier om onomstotelijk aan- of afwezigheid van de ziekte aan te tonen en op die manier de resultaten van de te onderzoeken test te verifiëren. In de praktijk van alledag blinkt er veel, maar het is lang niet altijd goud. Weten Kline et al. nu echt zeker welke van de 380 patiënten een longembolie hadden? Het eerlijk antwoord moet negatief zijn: ze gebruikten een mengeling van manieren om te besluiten tot de aan- dan wel afwezigheid van een longembolie. Een dergelijk gebruik van zilveren, bronzen, koperen of blikken referentietests wordt vaak als kritiek aangedragen. Het is de vraag of dat terecht is. Vanuit een zuiver theoretisch standpunt wellicht wel. Vanuit het standpunt van de gezondheidswinst is dat vermoedelijk veel minder het geval. Voor de patiënt is niet zozeer de vraag belangrijk of er een embolus is, hoe klein ook, maar of die embolus groot genoeg is om de nadelen van behandeling met middelen die de stolling beïnvloeden te laten opwegen tegen de voordelen van een dergelijke behandeling. Vanuit dat perspectief zijn piepkleine, subsegmentele emboli waarschijnlijk onvoldoende reden om behandeling met ‘bloedverdunners’ te starten of door te zetten. Niet het goudgehalte van de standaard is van belang, maar de mate waarin aan- of afwezigheid van een bepaalde toestand behandelbare gevolgen heeft voor de gezondheidstoestand van de betrokken patiëntengroep. Eens die pragmatische consequentie getrokken, moet ook worden erkend dat testeigenschappen als sensitiviteit en specificiteit geen eigenschappen van de test zijn, maar karakteriseringen die gelden voor toepassing van de test in nader te bepalen klinische omstandigheden. Ze zijn in de regel ook geen constanten: sensitiviteit en specificiteit kunnen variëren al naar gelang de situatie waarin de test wordt toegepast. Een tweede probleem is dat onderzoek naar de eigenschappen van tests vaak afschuwelijk slecht wordt uitgevoerd en de rapportage bar en boos is. Ondertussen is overtuigend aangetoond dat onderzoek met tekortkomingen in de methodologie tot te optimistische schattingen van de waarde van een test kan leiden (3). Dat heeft een groep “editors”, onderzoekers en methodologen aangezet tot het ontwikkelen van STARD: “standards for the reporting of diagnostic accuracy”. In navolging van het succesvolle CONSORT-initiatief, gericht op het maken van afspraken van een minimale, heldere rapportage van wat er in “clinical trials” is gebeurd, wil de STARD-groep tot vergelijkbare afspraken komen voor het rapporteren van onderzoek naar overeenkomsten tussen een indextest en een referentietest (4). (http://www.consort-statement.org)

Is that all there is?

Beeldt u zich een imaginaire wereld in. Stel dat van alle tests, in alle denkbare situaties, voor elke denkbare patiëntengroep, de sensitiviteit en de specificiteit bekend zijn. Is dan meteen ook de waarde bekend van die tests? Het antwoord op die vraag moet negatief zijn. Sensitiviteit en specificiteit zeggen wel wat over de (conditionele) kansen op een passend testresultaat, maar niet alles over de gezondheidswinst of de doelmatigheid (verhouding van ingezette middelen tot winst). Twee argumenten om die stelling te onderbouwen. Ten eerste zeggen sensitiviteit en specificiteit weinig over de toegevoegde waarde van een test. Men heeft niet zo veel aan een goed onderscheid als dat al eerder, met andere middelen kon worden gemaakt. Sensitiviteit en specificiteit zeggen dus weinig over de reductie van onzekerheid over de aan- dan wel afwezigheid van de vermoede ziekte. Ten tweede zeggen sensitiviteit en specificiteit niet zo veel over de gezondheidswinst die het gevolg is van het gemaakte onderscheid. Wat voor nut heeft het onderscheid als er geen gevolgen aan kunnen worden verbonden? Voor de purist is dat van weinig belang, voor de clinicus practicus wellicht des te meer. Als reactie op die laatste vraag wordt de roep gehoord naar diagnostische RCT’s. Als het gerandomiseerd, vergelijkend onderzoek zo goed functioneert als gouden toetssteen voor het evalueren van geneesmiddelen en andere vormen van therapie, waarom dan ook niet voor het beoordelen van tests? Ook hier is, driewerf helaas, het antwoord minder evident dan de vraag (5). Er is nog veel werk aan de winkel voor alle professionals die de handschoen hebben opgepakt en, zich bewust van alle mogelijke vormen van vertekening, op zoek gaan naar het antwoord op de terechte vragen naar rekenschap en verantwoording ‘Worden mensen hier echt beter van?’.

Literatuur

1. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. New York: Churchill Livingstone, 1997.
2. Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. JAMA 2001; 285: 761-768.
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282: 1061-1066.
4. Moher M, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomised trials. Jama 2001; 285: 1987-1991.
5. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000; 356: 1844-1847.

Ned Tijdschr Klin Chem 2001; 26: 245-248

Structured validation of laboratory test results

W.P. OOSTERHUIS[1] and H.J.L.M. ULENKATE[2]

Implementation of all kinds of quality controls did not eliminate errors from the clinical laboratories. Now and then, human handling in the preanalytical, analytical and postanalytical phase will introduce mistakes and blunders. In a hospital environment, pathological results will always exist. A validator or validation program should discriminate between erroneous and pathological results. Manual or automated validation processes inevitable result in non-validated test results. To improve decision rules on these non-validated test result more patient information is necessary. Patient information could be derived from the request form or in the future from the electronic patient file. Therefore, we suggest that clinical chemists should participate from the beginning in the development of the electronic patient files to organise that an automated laboratory validation program can use this medical information.

Key-words: validation; automation; verification; authorisation

Clinical laboratories are expected to fulfil high quality standards. Although a zero error level is the target, errors do inevitably occur. Not all of the approaches to quality control are equally effective. What are the sources of errors and what are the most effective ways to eliminate each type of error? Much effort has been devoted to quality control schemes. Shewhart at Bell Laboratories showed that variation in the production process could be described statistically to identify when a process was drifting out of control (1). In time adjustment of the process could prevent errors. This approach was very successful, but had its limits. A further reduction of process variation did not lead to the expected further reduction of the error rate, because most remaining errors did not result from process variation. They result from human mistakes and blunders (1). Because each type of mistake is a rare event, the frequency of mistakes cannot be predicted by sampling methods. Factors resulting in a systematic error such as errors at test order entry, drug interactions, improper collection and handling, patient condition and labelling of specimen, can affect the accuracy of the laboratory test result without necessarily affecting the analytical quality (2). When the analytical quality is unaffected, these types of errors are invisible to statistical quality control schemes. To control these types of errors in automobile industry, quality control methods were changed dramatically to a 100% inspection method. In this method all products (results) are inspected, not a sample representing a series of products (results). This can be accomplished by automating the inspection process (1). Inspection methods focus on detecting defect-causing conditions upstream of the process and correcting mistakes before they result in production errors. For example, a batch of wrong printed OMR (optical mark reading) request forms can cause many errors due to non-reading of requested tests. Correcting these errors downstream in the process is very inefficient. Inspection of a new batch of request forms will eliminate this type of error at an early stage. Incorrect samples will cause wrong test results or laborious correction. Methods aimed at false-proof guidance of the technician during the process of blood sampling, e.g. by using unique sample labels with sample type instructions, will reduce the error rate significantly.

St. Elisabeth Hospital, Department of Clinical Chemistry and Haematology[1], Tilburg, The Netherlands and Diagnostic Center SSDZ, Department of Clinical Chemistry[2], Delft, The Netherlands.

Address for correspondence: Dr. W.P. Oosterhuis, St. Elisabeth Hospital, Department of Clinical Chemistry and Haematology, PO Box 90151, 5000 LC Tilburg. The Netherlands
e-mail: oosthuis@knmg.nl.