

W. Bautsch¹

Anforderungen und Bewertung der Ergebnisse von Laboruntersuchungen – Teil 5 der Serie zur Bewertung wissenschaftlicher Publikationen^{*}, ^{**}

Requirements and assessment of laboratory tests – part 5 of a series on evaluation of scientific publications

Hintergrund: Die hohen Sensitivitäten und Spezifitäten vorhandener Labortests verführen zusammen mit einem vergleichsweise niedrigen Preis der Einzeltestung zu einem breiten, weitgehend unkritischen Anforderungsverhalten.

Methoden: Erläuterung des Bayesschen Theorems und seine Anwendung auf laboratoriumsmedizinische Anforderungen in weitgehend nicht mathematischer Form unter Zuhilfenahme selektiver Literatur.

Ergebnisse und Schlussfolgerungen: Nach dem Bayesschen Theorem ist der positiv prädiktive Wert eines laboratoriumsmedizinischen Testergebnisses direkt von der Prävalenz der Erkrankung im Untersuchungskollektiv abhängig. Der klinische Aussagewert von laboratoriumsmedizinischen Testergebnissen ist daher von einer klaren Indikationsstellung abhängig. Ergebnisse von nicht indizierten Anforderungen sind ohne fundierte Datenbasis zur Prävalenz der Erkrankung klinisch unbrauchbar und sollten daher nicht angefordert werden. Das gilt unabhängig von ethischen oder ökonomischen Überlegungen.

(Dtsch Zahnärztl Z 2010, 65: 759–763)

Schlüsselwörter: Labordiagnostik, Blutanalyse, Diagnosestellung, PSA-Test, Borreliose

Background: Current laboratory tests exhibit high sensitivity and specificity combined with comparatively low costs thus favoring broad and uncritical ordering habits.

Methods: Introduction of Bayes' theorem and discussion of its implications for laboratory test results in a mostly non-technical form, accompanied by a selective literature review.

Results and conclusions: According to Bayes' theorem the positive predictive value of laboratory test results is directly dependent on the prevalence of the disease in a given patient cohort. Thus, the clinical value of a given test result is critically dependent on a precise indication. Ordering of tests that are not indicated in a given patient is clinically useless and undesirable, where detailed information on disease prevalence is missing. These considerations are valid irrespective of ethical or economic considerations.

Keywords: laboratory diagnostics, blood analysis, diagnosis, PSA test, borreliosis

^{*} Nachdruck aus: Dtsch Arztebl Int 2009; 106(24): 403–406; DOI: 10.3238/arztebl.2009.0403 © Deutscher Ärzte-Verlag GmbH Köln

^{**} Der folgende Artikel bezieht sich nicht unmittelbar auf die Bewertung medizinischer Publikationen. Er erscheint aber dennoch im Rahmen dieser Serie, weil er ein verwandtes Problem zum Inhalt hat: die statistische Bewertung einer klinischen Entscheidungssituation. Das im vorliegenden Text behandelte Thema des positiven prädiktiven Wertes taucht immer wieder in wissenschaftlichen Publikationen auf.

¹ Institut für Mikrobiologie, Immunologie und Krankenhaushygiene, Städtisches Klinikum Braunschweig gGmbH; Prof. Dr. med. habil. Dr. rer. nat. Bautsch

Peer reviewed article: eingereicht: 6.2.2007, revidierte Fassung angenommen: 19.10.2007

DOI: 10.3238/dzz.2010.0759

Einleitung

Das Anforderungsverhalten von Laboruntersuchungen ist in der Praxis häufig recht unkritisch. Ihre Kosten sind vergleichsweise niedrig – in Bezug zu etwa den bildgebenden Untersuchungsverfahren – bei gleichzeitig hoher Sensitivität und Spezifität der einzelnen Analyseergebnisse. Dadurch wird suggeriert, dass durch Anforderung vieler Laborparameter schnell, mit geringem Aufwand und vergleichsweise kostengünstig klinisch verwertbare Aussagen zu Erkrankungen erhalten werden können. Dies wird auch dann angenommen, wenn die getesteten Parameter mit dem vorliegenden Beschwerdebild des Patienten wenig oder nichts zu tun haben. Neben dem Anfordern von teilweise sehr umfangreichen Routineprofilen betrifft dies auch das Screening auf solche Erkrankungen, die vor Auftreten klinischer Symptome diagnostiziert werden sollten, etwa Tumorerkrankungen oder bestimmte, in Stadien verlaufende Infektionskrankheiten, wie etwa die Borreliose.

Dabei wird übersehen, dass der Aussagewert von Testergebnissen von einer klaren Indikationsstellung abhängt. Dieser Aspekt wird in den öffentlich geführten Diskussionen über den Wert von Screeninguntersuchungen zwar immer wieder genannt [1], spielt aber auch in der täglichen ärztlichen Praxis eine wichtige Rolle. Empfohlene Screeninguntersuchungen, etwa im Rahmen des Neugeborenen-Screenings, sind natürlich nicht gemeint, da die hier genannten Überlegungen bei solchen Empfehlungen explizit berücksichtigt wurden.

Diese Arbeit skizziert die zugrunde liegende mathematische Beziehung in weitgehend nicht mathematischer Form und erläutert die Konsequenzen für das diagnostische Anforderungsverhalten in der täglichen ärztlichen Praxis. Dieses Problem hat mit Statistik zu tun und gerade auf diesem Gebiet sind intuitive Vorstellungen häufig irreführend. Das grundlegende Problem wird in nachfolgender Multiple-Choice-Frage verdeutlicht.

A laboratory test (for example, for borreliosis) has a diagnostic specificity of 98 %. How probable is it that a patient who gives a positive test result does in fact have this disease?

- a) You have to know the sensitivity too to be able to answer this question.
- b) 98 %
- c) $(1 - \text{specificity}) \times 100 (\%) = 2 \%$
- d) None of these answers is correct.

Leser, die diese Frage richtig beantworten konnten, brauchen eigentlich nicht weiter zu lesen (die Lösung findet sich am Schluss des Artikels). Dieser Artikel kann für die praktische ärztliche Tätigkeit sehr hilfreich sein, denn das zugrunde liegende Problem taucht in vielen Variationen immer wieder auf.

Die meisten Menschen, denen dieses Problem gestellt wird, versuchen es, allein mithilfe der Spezifität zu lösen: Die Spezifität gibt an, bei wie vielen gesunden Probanden auch korrekterweise ein negatives Testergebnis erhalten wurde, umgekehrt gibt $1 - \text{Spezifität}$ an, bei wie vielen gesunden Probanden fälschlicherweise ein positives Testergebnis erhalten wurde (Falsch-positiv-Rate). Intuitiv ist man geneigt anzunehmen, dass jetzt alle erforderlichen Informationen vorliegen: Die Wahrscheinlichkeit sollte 98 % betragen – das ist aber falsch. Für die korrekte Lösung dieses Problems sind zusätzlich zwei weitere Angaben erforderlich, nämlich die Sensitivität des Tests und die Prävalenz der Erkrankung im Untersuchungskollektiv. Letztere ist der Anteil an erkrankten Patienten bezogen auf alle Personen, bei denen der Arzt eine entsprechende Untersuchung veranlasst. Der Grund hierfür wird in Kasten 1 erläutert.

Weitere Parameter einbeziehen

Welche Konsequenzen hat das nun auf das eingangs beispielhaft angesprochene Problem einer Borreliose-Testung? Die Prävalenz einer aktiven Borreliose in der Bevölkerung ist nicht ausreichend bekannt. Schätzungen gehen von 10 bis 237 Fällen je 100.000 Einwohner aus [5] mit großen regionalen Schwankungen [6]. Das Robert Koch-Institut hat 2003 für Deutschland einen Wert von 25/100.000 publiziert [4], der auch im Folgenden verwendet werden soll, um die Rechnungen etwas zu vereinfachen. Für die modernen serologischen Enzym-Immuntests auf Borreliose zusammen mit dem empfohlenen Immunoblot werden Spezifitäten von über 98 % angenommen [2], die aber nicht genau be-

kannt sind und je nach Testsystem etwas verschieden sein dürften. Hier wird eine Spezifität von 98 % angenommen. Dann gilt, dass man bei 100.000 Untersuchungen in der Bevölkerung 25 tatsächlich positive Ergebnisse erhält. Es soll vernachlässigt werden, dass die Sensitivität der Tests $< 100 \%$ ist. Allerdings gibt es zwei weitere grundlegende Probleme im Zusammenhang mit der Interpretation von serologischen Testergebnissen bei Borreliose, die selbst dann existieren, wenn alle Untersuchungen streng indiziert angefordert werden.

- Ein negatives Testergebnis schließt eine aktive Borreliose – gerade in den Frühstadien – nicht sicher aus, weil die Tests keine 100-prozentige Sensitivität haben.
- Die verfügbaren serologischen Tests können nicht sicher zwischen einer aktiven Borreliose und einem Titer nach abgeheilter Borreliose unterscheiden, sodass auch eindeutig positive serologische Befunde per se noch keine Behandlungsindikation darstellen.

Neben den 25 tatsächlich positiven Ergebnissen wird es auch 2.000 falsch positive Testergebnisse geben, weil $1 - \text{Spezifität} = 2 \%$ beträgt. Man wird insgesamt also 2.025 positive Testergebnisse erhalten, von denen 25 durch eine aktive Borreliose verursacht sind. Das entspricht einer Wahrscheinlichkeit von etwa 1,25 %, dass ein Proband mit einem positiven Testergebnis auch tatsächlich an einer aktiven Borreliose leidet. Dieser Test eignet sich sicherlich nicht für ein Bevölkerungsscreening, weil ein positives Testergebnis mit fast 99-prozentiger Sicherheit falsch ist.

Der Arzt kann die Prävalenz einer Erkrankung beeinflussen. Gemeint ist nämlich immer die Prävalenz der Erkrankung im Untersuchungskollektiv, also bei den Patienten, bei denen er eine Untersuchung veranlasst. Wenn also eine Borreliose-Testung bei jedem Patienten angefordert wird – vollkommen unabhängig von den zugrunde liegenden Beschwerden – nähert sich der Aussagewert des Einzelergebnisses dem eines Bevölkerungsscreenings. Schließlich geht irgendwann jeder einmal zum Arzt. Und der Aussagewert eines positiven Testergebnisses geht dann gegen null.

Anders ist die Situation, wenn die Testung auch indiziert angefordert wird, zum Beispiel weil sich ein Patient mit ei-

Kasten 1

Sensitivität, Spezifität und positiv prädiktiver Wert

Sensitivität und Spezifität sind statistische Größen, die für die meisten Tests gut bekannt sind, weil sie im Prinzip leicht zu ermitteln sind. Hierzu testet man eine definierte Probenanzahl von Patienten, die erwiesenermaßen gesund sind oder an der gesuchten Erkrankung leiden. Jede Patientenprobe kann entweder ein positives oder negatives Testergebnis aufweisen. Die Ergebnisse lassen sich dann in einer 2 × 2 Matrix (4-Felder-Tafel) darstellen. Statt definierter Proben kann man auch einen Feldversuch durchführen. Hierbei muss man durch Referenzverfahren nachträglich klären, ob das untersuchte Material von einer erkrankten oder gesunden Person stammt. In der Tabelle 1 ist das Ergebnis eines solchen Feldversuchs durch die Zahlen in Klammern dargestellt.

		„Wahrer Wert“	
		krank	gesund
Testergebnis	positiv	a (398)	b (12)
	negativ	c (22)	d (1.012)

Tabelle 1

Die Sensitivität eines Tests gibt an, wie viele der erkrankten Personen durch den Test als positiv erkannt werden. Aus der 4-Felder-Tafel kann man unmittelbar ableiten, dass die Sensitivität $a/(a+c)$ sein muss: (a+c) ist die Anzahl aller Erkrankten im untersuchten Kollektiv, von denen a Patienten ein positives Testergebnis aufwiesen. Die Sensitivität des Tests beträgt in diesem Zahlenbeispiel daher $398/(398+22) = 0,9476$ oder etwa 94,8 %.

Die Spezifität gibt dagegen an, wie viele der gesunden Patienten in diesem Test auch ein negatives Testergebnis aufweisen und ist durch $d/(b+d)$ gegeben: (b+d) ist die Anzahl aller gesunden Probanden im untersuchten Kollektiv, von denen d Patienten ein negatives Testergebnis haben. Die Spezifität des Tests beträgt in dem Zahlenbeispiel daher $1.012/(1.012+12) = 0,9883$ oder etwa 98,8 %.

Wie wahrscheinlich ist es, dass ein Patient mit einem positiven Testergebnis auch krank ist? Hauptsächlich dies interessiert den behandelnden Arzt in einer klinischen Situation. Diese Wahrscheinlichkeit bezeichnet man als positiv prädiktiven Wert (PPW). Aus der 2 × 2 Matrix ergibt sich: Die Anzahl der positiv getesteten Patienten beträgt (a+b) = 410. Davon sind a = 398 Probanden auch krank. Die Wahrscheinlichkeit, dass ein Proband mit einem positiven Testergebnis auch krank ist, beträgt in unserem Beispiel daher $PPW = a/(a+b) = 398/410 = 0,9707$ oder etwa 97,1 %.

Das ist etwas anderes als die Spezifität: Bei der Ermittlung der Spezifität musste die rechte Spalte „gesund“, bei der Ermittlung des PPW dagegen die obere Zeile (Testergebnis positiv) ausgewertet werden. Im ersten Fall wurden Zeilen, im zweiten Spalten ausgewertet.

Nun ist der Unterschied in dem gewählten Beispiel nicht besonders groß (Spezifität 98,8 % gegenüber PPW 97,1 %). Das liegt daran, dass der prozentuale Anteil an kranken Proban-

den (Prävalenz der Erkrankung) in diesem Feldversuch sehr hoch war, und zwar $(a+c)/(a+b+c+d)$ etwa 29,1 %. Die Zahl der gesunden Probanden könnte in dem Feldversuch zehnmal höher sein, also 96,1 %, entsprechend einer Prävalenz von 3,9 % (Tabelle 2). Dies ist häufig eine realistische Annahme. Hieraus ergeben sich folgende Werte:

Die Sensitivität beträgt weiterhin 94,8 % und auch die Spezifität hat sich nicht geändert: $d/(b+d) = 10.120/(10.120 + 120) = 0,9883$ oder etwa 98,8 %. Auf den PPW hat dieser Umstand dagegen entscheidenden Einfluss, denn er beträgt jetzt nur noch $PPW = a/(a+b) = 398/(398+120) = 0,768$ oder 76,8 %.

		„Wahrer Wert“	
		krank	gesund
Testergebnis	positiv	a (398)	b (120)
	negativ	c (22)	d (10.120)

Tabelle 2

Bei einer noch niedrigeren Prävalenz von zum Beispiel 0,41 % (entsprechend einer nochmals 10-fach höheren Anzahl gesunder Probanden) würde der PPW auf 24,9 % sinken. Ein positives Testergebnis würde dann mit über 75-prozentiger Wahrscheinlichkeit bedeuten, dass der Proband die Erkrankung nicht hat.

Ein positives Testergebnis kann also entweder von kranken oder von gesunden Probanden stammen. Im letzteren Fall handelt es sich um ein falsch positives, das heißt unspezifisches Ergebnis. Bei sinkender Prävalenz der Erkrankung, das heißt prozentual immer weniger kranken Personen im Untersuchungskollektiv, werden aber immer mehr gesunde Probanden untersucht. Deshalb ist ein positives Testergebnis auch mit immer höherer Wahrscheinlichkeit falsch positiv. Zusammenfassend bedeutet dies, dass der positiv prädiktive Wert eines Tests nicht nur von Sensitivität und Spezifität, sondern auch von der Prävalenz der Erkrankung im Untersuchungskollektiv abhängig ist: Je niedriger die Prävalenz, desto niedriger der PPW.

Analoge Überlegungen gelten für den negativ prädiktiven Wert (NPW), der die Wahrscheinlichkeit beschreibt, dass ein Patient bei einem negativen Testergebnis tatsächlich nicht erkrankt ist. Aus Tabelle 1 ergibt sich, dass der $NPW = d/(c+d) = 1.012/(1.012+22) = 97,7\%$ ist. Im Gegensatz zum PPW sinkt der negativ prädiktive Wert mit steigender Prävalenz der Erkrankung: Erhöht man die Anzahl der kranken Patienten im Untersuchungskollektiv um den Faktor 100, entsprechend einer Prävalenz der Erkrankung von 97,6 % im Untersuchungskollektiv, ergeben sich folgende Werte (Tabelle 3): Sensitivität: $a/(a+c) = 39.800/(39.800+2.200) = 94,8\%$ und Spezifität: $d/(d+c) = 1.012/(1.012+12) = 98,8\%$ bleiben unverändert. Der NPW beträgt dagegen jetzt nur noch $d/(c+d) = 1.012/(2.200+1.012) = 31,5\%$, das heißt, nur in 31,5 % der Fälle sind Patienten mit einem negativen Testergebnis auch tatsächlich gesund. So kann ein negatives Testergebnis entweder von gesunden oder kranken Probanden stammen. Im letzteren Fall ist es ein falsch negatives Ergebnis. Bei steigen-

der Prävalenz der Erkrankung werden aber auch immer mehr kranke Probanden untersucht. Entsprechend ist ein negatives Testergebnis auch mit immer höherer Wahrscheinlichkeit falsch negativ. Der hier dargestellte mathematische Zusammenhang lässt sich auch mit Wahrscheinlichkeitsrechnung darstellen (Kasten 2).

		„Wahrer Wert“	
		krank	gesund
Testergebnis	positiv	a (39.800)	b (12)
	negativ	c (2.200)	d (1.012)

Tabelle 3

Kasten 2

Bayessches Theorem

Sensitivität, Spezifität und positiv prädiktiver Wert lassen sich auch als bedingte Wahrscheinlichkeiten ausdrücken: $p(B/A)$ sei die Wahrscheinlichkeit, dass das Ereignis B eintritt unter der Voraussetzung, dass A vorliegt. Zum Beispiel ist die bedingte Wahrscheinlichkeit p (Testergebnis positiv/Proband krank) nichts anderes als die in Kasten 1 definierte Sensitivität und durch $a/(a+c)$ gegeben. Meistens interessiert aber die umgekehrte Wahrscheinlichkeit p (Proband krank/Testergebnis positiv), also der positiv prädiktive Wert. In gewisser Weise werden Ursache und Wirkung vertauscht. Meistens ist die

Wahrscheinlichkeit bekannt, dass die Ursache, also die Erkrankung, zu einem positiven Testergebnis führt. Man interessiert sich oft aber für einen anderen Aspekt, und zwar dafür, wie von einem positiven Testergebnis auf die Ursache (Erkrankung) geschlossen werden kann. Der korrekte mathematische Zusammenhang ist durch das von Thomas Bayes aufgestellte Bayessche Theorem gegeben: $p(A/B) = p(B/A) \times p(A) / p(B)$ beziehungsweise nach Umformung, weil $p(B)$ nicht unmittelbar bekannt ist: $p(A/B) = p(B/A) \times p(A) : [p(B/Ac) \times p(Ac) + p(B/A) \times p(A)]$ mit $Ac =$ Komplement von A, das heißt A liegt nicht vor.

ner akut aufgetretenen peripheren Fazialisparese vorstellt. Die Prävalenz der Borreliose bei Patienten mit akuter Fazialisparese ist nicht sehr gut untersucht, eine neuere norwegische Arbeit gibt etwa 10 % an [3], bei Kindern liegt dieser Wert sicherlich deutlich höher. In diesem Patientenkollektiv resultieren ganz andere Ergebnisse: Bei 1.000 Untersuchungen werden 18 falsch positiv sein (1 – Spezifität = 2 % von 900 negativen Patienten), aber es gibt 100 tatsächlich positive Befunde. Die Wahrscheinlichkeit, dass ein Patient mit einem positiven Testergebnis auch tatsächlich eine Borreliose hat, beträgt also $(100/100+18) \times 100 \sim 85 \%$, bei Kindern wird das Ergebnis entsprechend höher sein.

Zusammenfassung

Sensitivität und Spezifität sind testspezifische Eigenschaften, die der Arzt nicht aktiv beeinflussen kann. Dies gilt unter

der Voraussetzung korrekter Testdurchführung und -befundung, einschließlich Prä- und Postanalytik. Dagegen ist der Aussagewert eines positiven Testergebnisses, der positiv prädiktive Wert, kritisch von der Prävalenz der Erkrankung im Testkollektiv abhängig – und die kann der Arzt beeinflussen. Tests sollten grundsätzlich nur indiziert angefordert werden, denn nur dann ist ein Testergebnis auch klinisch verwertbar. Ergebnisse von nicht indizierten Anforderungen sind ohne fundierte Datenbasis zur Prävalenz der Erkrankung klinisch unbrauchbar und sollten daher nicht erbeten werden. Dies gilt unabhängig von ökonomischen und ethischen Argumenten.

Was hier am Beispiel des Borreliosetests gezeigt wurde, gilt für alle anderen Labortests. Die Überlegungen sind auch unabhängig davon, ob es sich um einen Labortest oder andere Untersuchungen handelt, etwa endoskopische, röntgenologische, sonografische, elektrokardiografische oder klinische Verfahren:

Nicht indizierte Untersuchungen führen zu einer Senkung des positiv prädiktiven Werts und somit zur Zunahme falsch positiver Testergebnisse.

Die richtige Antwort auf die eingangs gestellte Multiple-choice-Frage lautet: d. 

Interessenkonflikt: Der Autor erklärt, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Korrespondenzadresse

Prof. Dr. med. habil. Dr. rer. nat.
Wilfried Bautsch
Institut für Mikrobiologie,
Immunologie und Krankenhaushygiene
Städtisches Klinikum Braunschweig
gGmbH
Celler Straße 38, 38814 Braunschweig
E-Mail: w.bautsch@klinikum-braunschweig.de

Literatur

1. Bögermann C, Rübber H: Früherkennung des Prostatakarzinoms. Dtsch Arztebl 104, A 503–504 (2007)
2. O'Connell S, Granström M, Gray JS, Stanek G: Epidemiology of European Lyme borreliosis. Zentralbl Bakteriol 287, 229–240 (1998)
3. Talaska T: Borreliose-Epidemiologie. Brandenburgisches Ärzteblatt 11, 338–340 (2002); www.laekb.de/15/15Beitraege/95021TH0211.pdf
4. Mehnert WH, Krause G: Surveillance of lyme borreliosis in Germany, 2002 and 2003. Euro Surveill 10, 83–85 (2005)
5. Goettner G, Schulte-Spechtel U, Hillermann R, Liegl G, Wilske B, Fingerle V: Improvement of Lyme borreliosis serodiagnosis by a newly developed recombinant immunoglobulin G (IgG) and IgM line immunoblot assay and addition of VlsE and DbpA homologues. J Microbiol 43, 3602–3609 (2005)
6. Ljostad U, Okstad S, Topstad T, Mygland A, Monstad P: Acute peripheral facial palsy in adults. J Neurol 252, 672–676 (2005)

