Der Intuition auf der Spur? Das Bayes'sche Theorem und die Diagnostik in der Grundversorgung

Teil I

D. Pewsner[a], J. P. Bleuer[b], H. C. Bucher[c], M. Battaglia[d], P. Jüni[d,e], M. Egger[d,f]

Einleitung

Unsicherheit ist ein ständiger Begleiter der ärztlichen Tätigkeit: Im medizinischen Alltag müssen ständig Entscheidungen ohne hinreichende Information gefällt werden [1]. Erfahrene Ärzte und Ärztinnen zeichnen sich dadurch aus, dass sie anhand von Anamnese, Symptomen und klinischen Zeichen die Wahrscheinlichkeit einer bestimmten Erkrankung vielfach intuitiv richtig einschätzen. Nachteile des intuitiven Vorgehens sind die ungenügende Reproduzierbarkeit, die Schwierigkeit der didaktischen Vermittlung sowie vor allem die Störungsanfälligkeit unseres Gedächtnisses [1]. So erinnern wir kürzlich Erlebtes besser als länger Zurückliegendes. Auch bilden starke Emotionen für Erinnerungen einen zähen Klebstoff. Eine verpasste oligosymptomatische, offene Lungentuberkulose kann uns über Jahre zu einer Überschätzung der Prävalenz dieser Erkrankung verführen, was sich in einer unverhältnismässigen Verordnungshäufigkeit von Thoraxröntgenbildern widerspiegeln kann.

Im folgenden versuchen wir darzulegen, wie die konsequente Anwendung eines einfachen mathematischen Konzepts eine rationalere Diagnostik ermöglicht. Zwei Voraussetzungen spielen bei der diagnostischen Treffsicherheit eine wichtige Rolle: Die Fähigkeit, die Wahrscheinlichkeit der vermuteten Krankheit (Prävalenz oder Vortestwahrscheinlichkeit) abzuschätzen und die Kenntnis der Leistungsfähigkeit (Sensitivität und Spezifität) der angewendeten Tests. Das Bayes'sche Theorem [1, 2] verdeutlicht als formaler mathematischer Ansatz diesen Zusammenhang in reproduzierbarer Weise: Die Wahrscheinlichkeit der Diagnose wird berechenbar, wenn die Vortestwahrscheinlichkeit der fraglichen Diagnose und die Testcharakteristika bekannt sind. Ähnlich wie die Cochrane Datenbanken [3, 4] eine wichtige Grundlage rationaler Therapie darstellen, so benötigen wir auch in der Diagnostik Zugang zu einer Datenbank, welche die Eigenschaften diagnostischer Tests und die relevanten Vortestwahrscheinlichkeiten übersichtlich zusammenstellt. Dies umso mehr, als in der ärztlichen Praxis die Diagnostik häufig einen grösseren Raum als die Therapie einnimmt. Im Bereich der Therapie können wir mittlerweile auf Tausende von systematischen Übersichten, Meta-Analysen und klinischen Studien zugreifen, während in der Diagnostik entsprechende Bemühungen noch in den Kinderschuhen stecken. Das Potential einer diagnostischen Datenbank soll hier am Beispiel des Verdachts auf Appendizitis illustriert werden.

Ein Landwirt hat Bauchschmerzen

Ein 25jähriger Landwirt kommt mit mässig starken, bisher unbekannten diffusen Bauchschmerzen in die Sprechstunde. Die Schmerzen sind vor 48 Stunden neu aufgetreten und jetzt allmählich in den rechten Unterbauch gewandert. Der Patient klagt über leichte Nausea, der letzte Stuhlgang vor vier Stunden war normal. Die Appendix ist noch vorhanden. Sie messen rektal 38,2 °C Fieber und stellen einen leicht verminderten Allgemeinzustand fest. Bei der Untersuchung finden Sie ein weiches Abdomen ohne Loslassschmerz. Die Rektaluntersuchung ist unauffällig, die Darmgeräusche sind normal. Das Labor zeigt ein C-reaktives Protein von 25 mg/l, die Leukozyten betragen 13×10^9/l. Der Patient, der mit seiner Frau allein auf einem Hof lebt, möchte auf keinen Fall ins Spital.

Wie wahrscheinlich ist das Vorliegen einer Appendizitis? Dürfen Sie mit der Hospitalisation zuwarten?

Diagnostische Tests

Teil jeder Diagnostik ist die Differentialdiagnose, welche in Frage kommende Diagnosen hierarchisch nach abnehmender Wahrscheinlichkeit gliedert (Tab. 1). Da diagnostische Tests meist nicht perfekt sind, gelingt es selten, den wahren Zustand des Patienten mit absoluter

[a] Praxis für Innere Medizin FMH, Bern
[b] Mediscope AG PF, 3000 Bern 23
[c] Medizinische Universitäts-Poliklinik, Kantonsspital, Basel
[d] Institut für Sozial- und Präventivmedizin, Universität Bern
[e] Rheumatologische Universitätsklinik, Inselspital, Bern
[f] Division of Health Services Research, Department of Social Medicine, University of Bristol

Korrespondenz:
Dr. Daniel Pewsner
Postfach
CH-3000 Bern 26

daniel.pewsner@bluewin.ch

Tabelle 1.
Regeln der Informationsgewinnung bei der Anamneseerhebung.

Differentialdiagnostische Abklärung erfolgt weitgehend hypothesengesteuert [1, 6]. So werden Daten schon bei Beginn der Befragung nicht einfach systematisch gesammelt. Hypothesen werden aufgrund der Befragung ständig erhoben und dann im Rahmen eines dreistufigen zyklischen Prozesses falsifiziert oder verifiziert:

Zyklischer Prozess der Differentialdiagnose
I Zuhören → Hypothesen generieren
II Hypothese prüfen
III Beurteilung – Entscheidung – Therapie

Beim Generieren der Hypothesen spielt gemäss dem Bayes'schen Theorem die Häufigkeit (Prävalenz) einer möglichen Erkrankung eine herausragende Rolle. Auch müssen Hypothesen von selteneren Erkrankungen überprüft werden, die «nicht verpasst werden dürfen», da sie einer sofortigen Behandlung bedürfen. Weiter gilt, dass bei vorgängig gesunden Patienten eine einzige Diagnose als Erklärung einer gegebenen Symptomatik wahrscheinlicher als eine Kombination von zwei oder mehr Diagnosen ist. Bei sehr unspezifischen Krankheitssymptomen bzw. Befunden, welche die Bildung von Hypothesen erschweren, gelangen systemanamnestische Fragen, die sich durch eine hohe Sensitivität bei meist schlechter Spezifität auszeichnen, zur Anwendung («SnNout-Tests», Tab. 2). Die Überprüfung einer Hypothese erfolgt nach einem System: In der ersten Phase erlauben Fragen hoher Sensitivität bei negativem Befund den Ausschluss der Hypothese («SnNout»). Kann eine Hypothese damit nicht verworfen werden, so folgen Fragen hoher Spezifität, die bei positivem Befund den Einschluss der Hypothese: («SpPin», Tab. 2) ermöglichen. Diese Prinzipien der Generation und Überprüfung von Hypothesen gelten auch für klinische und apparative Untersuchungen.

Sicherheit zu ergründen. Die Diagnose soll jedoch für praktische Zwecke genügend sicher sein, sich also in ihrer Wahrscheinlichkeit genügend von den weiteren Differentialdiagnosen unterscheiden. Mit diagnostischen Tests sind nicht nur apparative Untersuchungen gemeint: Auch anamnestische Angaben und klinische Untersuchungsbefunde stellen Tests dar und lassen wesentliche Schlussfolgerungen zu. So erhöht die Aussage eines Patienten, dass er einen *neuartigen* Bauchschmerz *bisher unbekannter Qualität* verspüre, die Wahrscheinlichkeit einer Appendizitis. Die gezielte Frage nach der Schmerzqualität darf deshalb durchaus als Test bezeichnet werden.

Parameter der Testleistung: Sensitivität, Spezifität

Die Leistungsfähigkeit eines Tests kann auf verschiedene Art gemessen werden. Allgemein bekannt sind die Grössen *Sensitivität* und *Spezifität* [1, 5]. Vorausgesetzt, dass ein positiver Testbefund «krank» bedeutet, gilt:

Sensitivität:
Wahrscheinlichkeit, dass ein kranker Proband einen positiven Test aufweist.
Wird als Richtig Positive Rate (True Positive Rate; TPR) bezeichnet.
Spezifität:
Wahrscheinlichkeit, dass ein gesunder Proband einen negativen Test aufweist.
Wird als Richtig Negative Rate (True Negative Rate; TNR) bezeichnet.

Sensitivität und Spezifität sind Parameter, welche die Leistungsfähigkeit eines diagnostischen Tests objektiv beschreiben (Tab. 2). Sie werden empirisch durch Vergleich mit einem Referenztest, dem sogenannten «Goldstandard» bestimmt. Der Goldstandard ist diejenige Untersuchung, die den Zustand des Patienten so wahrheitsgetreu wie möglich erfasst. Sensitivität und Spezifität sind im Hinblick auf die Praxis allerdings nur bedingt hilfreich: Mehr als die Wahrscheinlichkeit eines positiven Befunds bei Krankheit interessiert die Wahrscheinlichkeit des Vorhandenseins der Krankheit bei einem positiven bzw. einem negativen Resultat. Diese in der Praxis eigentlich relevanteren Aussagen über die sogenannte «Nachtestwahrscheinlichkeit» werden auch «Positiver» bzw. «Negativer Vorhersagewert» genannt (im Englischen «positive» und «negative predictive value»; PPV und NPV).

Wahrscheinlichkeit einer Diagnose und Krankheitsprävalenz: Das Bayes'sche Theorem

Gute Diagnostik beruht auf der Erkenntnis, dass das Seltene selten und das Häufige häufig ist: «If you hear hoofbeats, think of horses, not zebras» [1]. So lautet bei einem Patienten mit Kopfweh, Gliederschmerzen und hohem Fieber aus Zürich die differentialdiagnostische Rangfolge von Grippe und Malaria anders als bei jemandem, der soeben aus Abidjan zurückge-

Abbildung 1.
Thomas Bayes (1702–1761) wirkte als Priester in der Presbyterianischen Kirche von Tunbridge Wells, 50 km südlich von London. Obwohl er schon zu Lebzeiten als Mathematiker bekannt war, erschien sein Hauptwerk «Essay towards Solving a Problem in the Doctrine of Chances» erst 1764 posthum in den «Philosophical Transactions of the Roxal Society of London».

kehrt ist (Tab. 1). Die Wahrscheinlichkeit, dass eine bestimmte Krankheit vorliegt, leitet sich somit auch von der Häufigkeit (Prävalenz) ab, mit der diese in einer Gruppe vergleichbarer Individuen vorkommt: Angaben über die Prävalenz, welche der Vortestwahrscheinlichkeit von Erkrankungen entspricht, finden sich in der Literatur. Allerdings sind relevante Artikel nicht immer einfach zu finden. Die Indexierung dieser Artikel in den bibliographischen Datenbanken (zum Beispiel in MEDLINE) ist uneinheitlich. Die Vortestwahrscheinlichkeiten sind vom untersuchten Kollektiv abhängig, und auch mit ausgedehnten Suchstrategien finden sich oft nur wenige Arbeiten, die sich auf für Grundversorger repräsentative Patientenkollektive beziehen.

Das vom englischen Priester Thomas Bayes (Abb. 1) im 18. Jahrhundert entwickelte Theorem besagt, dass die Aussagekraft eines Tests bei gegebener Testgüte (Sensitivität und Spezifität) davon abhängt, wie gross die Wahrscheinlichkeit der gesuchten Diagnose vor dem Test ist. Das Prinzip wird leichter verständlich, wenn wir uns bewusst sind, dass ein bestimmtes Testresultat eine Diagnose nicht mit Sicherheit, sondern lediglich mit einer bestimmten Wahrscheinlichkeit nachweist oder ausschliesst. Die Wahrscheinlichkeit nach dem Test (Nachtestwahrscheinlichkeit) wird naturgemäss umso höher, je grösser die Wahrscheinlichkeit bereits vor dem Test war. Anders ausgedrückt kann ein Test als ein «*Wahrscheinlichkeitsumwandler*» bezeichnet werden (Abb. 2), der die Vortestwahrscheinlichkeit in eine Nachtestwahrscheinlichkeit umwandelt. Letztere entspricht dem positiven oder negativen Vorhersagewert (PPV bzw. NPV). Bei kleiner Vortestwahrscheinlichkeit muss ein Test überragende Testqualitäten aufweisen, um eine hohe Nachtestwahrscheinlichkeit zu generieren. Anders ausgedrückt können nur wenige Tests «aus Wasser Wein machen».

Tabelle 2.
«SpPins» und «SnNouts».

Aus der Definition der Sensitivität kann die Faustregel abgeleitet werden, dass ein negativer Test mit hoher Sensitivität (tiefe falsch negative Rate) den Ausschluss einer gesuchten Erkrankung (z.B. D-Dimere [ELISA] bei der Frage nach Lungenembolie) erlaubt, bei positivem Resultat hingegen keinen «Einschluss». Anschaulicher ist die Vorstellung, dass ein Fischer ein sehr engmaschiges Netz (= hohe Sensitivität) verwendet, um auch kleine Fische nicht zu verpassen. Dafür muss er viel Unrat im Netz in Kauf nehmen (ein schweres Netz darf nicht zum Schluss verleiten, dass ein Fisch gefangen worden ist). Diese Regel wird in der angelsächsischen Literatur häufig als «SnNout» bezeichnet [2, 7].

«SnNout»: «Sensitivity Negative rule the target disorder out»: Ein negatives Testresultat bei einem hochsensitiven Test erlaubt Ausschluss einer Erkrankung.

Umgekehrt gilt die Faustregel, dass ein positiver Test mit hoher Spezifität (tiefe falsch positive Rate) den «Einschluss» einer gesuchten Erkrankung (z.B. positive Tbc-Sputumkultur) gestattet, nicht jedoch deren Ausschluss bei negativem Resultat. Derartige Befunde werden auch als «pathognomonisch» bezeichnet. Ein weitmaschiges Netz oder viel mehr noch eine Reuse lässt zwar viele Fische durchschlüpfen, die Wahrscheinlichkeit ist jedoch hoch, dass bei einem seltenen Fang tatsächlich ein gesuchter Fisch gefangen worden ist. Es wird vom «SpPin» gesprochen.

«SpPin»: «Specificity Positive rule the target disorder in»: Ein positives Testresultat bei einem hochspezifischen Test erlaubt Einschluss einer Erkrankung.

Einschränkend sei angefügt, dass die Aussagekraft eines Tests – sei sie bezüglich Ausschluss oder Einschluss – von Sensitivität und Spezifität gemeinsam abhängt. Die Faustregeln «SnNout» und «SpPin» gelten deshalb nur, wenn die dazugehörige Spezifität bzw. Sensitivität nicht weniger als 40–50% beträgt; eine Bedingung, welche die meisten gebräuchlichen Tests allerdings erfüllen.

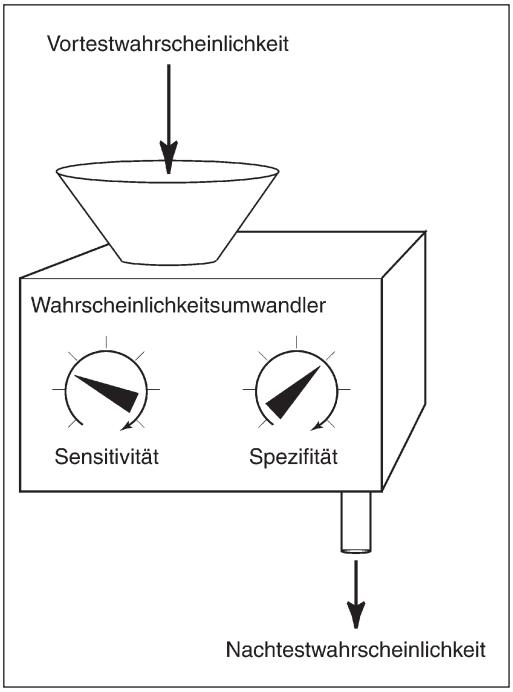

Abbildung 2.
Ein Test wandelt gemäss seiner Leistungsfähigkeit (Sensitivität bzw. Spezifität) eine gegebene Vortestwahrscheinlichkeit in eine Nachtestwahrscheinlichkeit um.

Die Likelihood-Ratio

Die *Likelihood-Ratio* [1, 2, 5] beschreibt, von der Vortestwahrscheinlichkeit unabhängig, die Leistung eines Tests. Sie fasst die Testqualitäten der Sensitivität und Spezifität in einer einzigen Zahl zusammen und ist somit ein objektiver Parameter der Testleistung. Sie ist definiert als Verhältnis der Wahrscheinlichkeiten eines bestimmten Testresultats bei Kranken und Gesunden. Die *Likelihood-Ratio für einen positiven Test* (LR+) besagt, wieviel mal wahrscheinlicher sich ein bestimmtes Testresultat (eine anamnestische Angabe, ein Symptom, ein Labor- oder anderes Testresultat) bei einem Kranken als bei einem Gesunden *findet*, sie ist eine Masszahl der «Einschlusskraft». Die *Likelihood-Ratio für ein negatives Resultat* (LR–) sagt aus, wieviel mal wahrscheinlicher ein bestimmtes Testresultat bei einem Kranken als bei einem Gesunden *fehlt*, sie ist eine Masszahl der «Ausschlusskraft».

LR+:	Rate der richtig Testpositiven	=	TPR	=	Sensitivität
	Rate der falsch Testpositiven		FPR		1-Spezifität
LR-:	Rate der falsch Testnegativen	=	FNR	=	1-Sensitivität
	Rate der richtig Testnegativen		TNR		Spezifität

Bei einem positiven Testresultat ist der Informationsgewinn umso höher, je grösser die LR+ des Tests ist. Umgekehrt sagt ein negatives Resultat um so mehr aus, je mehr sich die LR– Null annähert. Eine LR+ bzw. LR– von 1 ist ohne Aussage. Eine LR+ von über 10 lässt ausser bei sehr kleiner Vortestwahrscheinlichkeit eine gesuchte Krankheit mit grosser Wahrscheinlichkeit bestätigen. Umgekehrt kann eine LR– von unter 0,1 unter der Bedingung, dass die Vortestwahrscheinlichkeit nicht ausgesprochen hoch ist, eine gesuchte Krankheit mit grosser Wahrscheinlichkeit ausschliessen.

Einer der grossen Vorteile der Likelihood-Ratios ist, dass sich durch deren Multiplikation mit der Vortestwahrscheinlichkeit die Nachtestwahrscheinlichkeit direkt errechnen lässt (Vortestwahrscheinlichkeit × Likelihood-Ratio- = Nachtestwahrscheinlichkeit). Dies setzt allerdings eine andere «Wahrscheinlichkeitswährung» voraus: Wir sind uns gewohnt, Wahrscheinlichkeiten in Prozenten zu definieren. Damit ist das Verhältnis von positiven Ereignissen zu *allen* (d.h. positiven *und* negativen) Ereignissen gemeint. Bei der Vortestwahrscheinlichkeit entspricht dies dem Verhältnis zwischen *Kranken* und *allen Untersuchten (Kranke/[Gesunde + Kranke])*. Die Statistiker geben der «Währung» Odds zur Bezeichnung der Wahrscheinlichkeit den Vorzug, weil sie mathematisch besser einsetzbar ist. Die Odds entsprechen dem Verhältnis zwischen positiven und negativen Ereignissen, d.h. zwischen *Kranken* und *Gesunden (Kranke/Gesunde)*. Wir kennen diese Art der Angabe der Wahrscheinlichkeit aus dem Pferderennsport, wo die Gewinnchance eines bestimmten Pferdes nicht mit 20% sondern mit 1 zu 4 (entspricht 20%/[100%–20%] = 20%/80%) angegeben wird [5]. Diese «Währung» ist bei uns wenig geläufig, während sie zum Beispiel in England, wo das Wetten auf Pferde populär ist, jedem Kind vertraut ist.

Es gilt also folgende Beziehung: Vortest-Odds × Likelihood-Ratio = Nachtest-Odds.

Ein weiterer Vorteil der Likelihood-Ratios besteht darin, dass bei der Durchführung mehrerer Tests die entsprechenden Ratios miteinander multipliziert werden können. Daraus ergibt sich die Gesamt-Likelihood-Ratio der Testkombination. Wird diese mit den Vortest-Odds multipliziert, so erhält man die Nachtest-Odds für die Gesamtheit der durchgeführten Tests, die Reihenfolge der Tests spielt dabei keine Rolle.

Vortest-Odds × LR_{Test1} × LR_{Test2} × LR_{Test3} = Nachtest-Odds

Ein Landwirt mit Bauchschmerzen: Wie wahrscheinlich ist eine Appendizitis?

In einer zweiten Folge dieses Artikels in der nächsten Ausgabe des «Swiss Medical Forum» soll aufgezeigt werden, wie aufgrund von Prävalenz und Testcharakteristika mit Hilfe des Bayes'schen Theorems im hier beschriebenen Fall die Nachtestwahrscheinlichkeit einer Appendizitis berechnet werden kann. Auch wird das Konzept der Entscheidungsschwellen erörtert, welche es bei einer bestimmten Erkrankung erlaubt, aufgrund der ermittelten Nachtestwahrscheinlichkeit angemessen zu handeln. Schliesslich wird auf das Potential einer diagnostischen Datenbank eingegangen, welche es Grundversorgern ermöglichen soll, die Aussagen des Bayes'schen Theorems im Praxisalltag anzuwenden.

Verdankung

Wir bedanken uns bei Drs. A. Jaggi und Ch. Junker, Bern, für die kritische Durchsicht des Manuskripts.

Literatur

1 Sox HC, Blatt MA, Hinggins MC, Marton KI. Medical Decision-Making. Boston: Butterworth-Heinemann; 1987.
2 Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based Medicine. New York: Churchill Livingstone; 1997.
3 Bero L, Rennie D. The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. JAMA 1995;274:1935–8.
4 Antes G, Oxman AD. The Cochrane Collaboration. In Egger M, Smith GD, Altman DG, eds. Systematic Reviews in Health Care: Meta-Analysis in Context, London: Br Med J Books; 2000.
5 Bucher HC, Schmidt JG, Steurer J. Kritische Beurteilung einer Arbeit zu einem diagnostischen Test. Schweiz Rundsch Med Prax 1998; 87:1096–102.
6 Kassirer JP, Gorry GA. Clinical problem solving: a behavioral analysis. Ann Intern Med 1978;89:245–55.
7 Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making 1993;13:313–21.