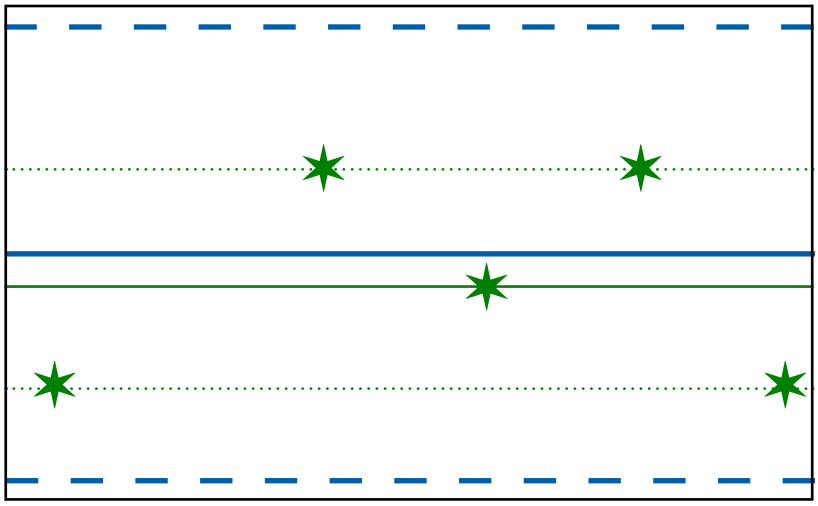
Eine Einführung zur Wahrheit und Aussagekraft diagnostischer Messwerte in der medizinischen Statistik

Moderne Medizin beruht vielfach auf statistischen Erhebungen – Erhebungen zur Häufigkeit von Krankheiten, Erhebungen zur Erfolgswahrscheinlichkeit von Therapien, Erhebungen, welche Laborwerte »normal« sind, und viele andere. Dies hat Konsequenzen für die Aussagen, die auf deren Basis gemacht werden. Auch die Aussagekraft von Laborwerten bzw. Testergebnissen leitet sich aus den statistischen Gegebenheiten ab, die den Test beschreiben oder mit denen die zur Entscheidungsfindung herangezogenen Grenzen definiert wurden. Das Verständnis dieser statistischen Gegebenheiten ist damit ein wichtiges Werkzeug bei der Interpretation von Laborergebnissen. Was bedeutet es z. B. für den Patienten, wenn ein bestimmter Wert gemessen wird?

Eine diagnostische Messung erfolgt theoretisch in der Absicht, Kenntnis zu erlangen über Gegebenheiten im Patienten – Konzentrationen oder ihre Veränderung, Reaktionszeiten, örtliche Position von Teilen der Anatomie, Morphologie von Gewebe etc. In vielen Fällen wird dazu eine Probe genommen, z. B. eine Blut- oder Urinprobe. Aus diesen Proben wird wiederum ein Teil entnommen und dem jeweiligen Test zugeführt. Bedenkt man nun, dass die bestimmte Komponente, zum Beispiel die Konzentration an Leukozyten, nicht überall gleichförmig ist, sondern zufällig verteilt, ergibt sich daraus, dass ein derartiges Messerergebnis immer nur eine Näherung des tatsächlichen Patientenstatus ist. Befinden sich im Urin des Patienten durchschnittlich 10 Leukozyten/µL, so kann lokal – zum Beispiel am Ort der Probenentnahme – durchaus auch eine Konzentration von 12 Leukozyten/µL vorliegen, die an anderer Stelle durch eine entsprechend niedrigere Konzentration ausgeglichen wird. Es ergibt sich also rein statistisch, aufgrund der Probenentnahme, eine mögliche Schwankungsbreite. Diese ist bestimmt durch die Konzentration selbst. Bei geringen Konzentrationen des Analyts bringen schon kleine absolute Schwankungen eine prozentual drastische Veränderung. Im obigen Fall resultiert die statistische Schwankung in einem Fehler von 20 %! Bei einer Messung von Leukozyten in einer normalen Blutprobe dagegen macht eine absolute Schwankung selbst in zehnfacher Höhe keinen signifikanten Unterschied.

Diese Schwankungen sind rein statistisch und methodenunabhängig. Bedingt sind sie durch die Entnahme einer Probe aus einer Gesamtheit (z. B. dem gesamten Blut in einem Patienten bzw. dem gesamten Blut in einem Röhrchen gegenüber dem aus letzterem entnommenen und vermessenen Volumen) und der zufälligen Verteilung der gemessenen Eigenschaft über die Gesamtheit. Sie bedingen einen statistischen Variationskoeffizienten (VK), der eine Schwankungsbreite der Messung um den »wahren« Wert angibt und rein konzentrationsabhängig ist.

Er findet sich z. B. in der Rümke-Tabelle wieder, die zeigt, wieviele Zellen gemessen werden müssen, um eine bestimmte Zuverlässigkeit des Messwertes zu erreichen. Zeigen Messungen im Labor einen vermeintlich kleineren Variationskoeffizienten, so ist zu beachten, dass es sich dabei um einen Variationskoeffizienten um den Mittelwert der Messungen, nicht aber um den »wahren« Wert herum, handelt.



Grün: mittlerer Messwert mit Schwankungsbreite,
✶: einzelne Messergebnisse
Blau: wahre Konzentration im Patienten mit statistischer Schwankungsbreite

***Abb. 1** Wahrer Wert (blau, fett) mit statistischem Variationskoeffizienten (blau, gestrichelt). Darin experimenteller Mittelwert (grün) mit Schwankungsbreite der Messwerte (grün, gestrichelt). Der experimentelle VK scheint hier geringer zu sein als statistisch möglich, tatsächlich aber liegen nur die gefundenen Messwerte zufällig in einem engeren Bereich um den gefundenen Mittelwert. Sie schöpfen zumindest in eine Richtung den Variationskoeffizienten fast vollständig aus. Weitere Messungen würden den statistischen VK bestätigen und den Mittelwert an den wahren Wert weiter annähern.*

Zu diesem statistischen Variationskoeffizienten kommen weitere Quellen der Schwankungsbreite hinzu, dazu gehören schwankendes Pipettiervolumen, subjektive Interpretation nach nichteinheitlichen Standards (ggf. durch wechselndes Personal) etc. Die Messung eines Parameters erfolgt daher immer im Widerstreit zwischen Präzision und Praktikabilität. Entscheidend werden dabei letztendlich auch Fragen der klinischen Relevanz – wie genau muss ich einen Parameter bestimmen, um verantwortungsvoll klinische Entscheidungen zu treffen?

Variationskoeffizient (VK):
Der Variationskoeffizient zeigt an, wie unsicher der Messwert bei einer bestimmten Konzentration im Verhältnis zu dieser Konzentration ist. Er ist ein Maß für die Impräzision einer Methode und wird berechnet, indem die Standardabweichung durch den Mittelwert geteilt wird. Das Ergebnis wird in Prozent angegeben. Im Gegensatz zur Standardabweichung, die einen Absolutwert für die Schwankungsbreite angibt, lässt sich aus dem VK ersehen, wie drastisch sich diese Unsicherheit bei der jeweiligen Konzentration auswirkt.

Eine Standardabweichung von 2 bei einer Konzentration von 2 Zellen/µL bedeutet einen großen Variationskoeffizienten von 100 %. Bei 200 Zellen/ µL ist die gleiche Standardabweichung aber wesentlich unproblematischer: nur 1 %.

a	n = 100	n = 500	n = 1000
1	0 – 4	0 – 1	0 – 1
2	0 – 6	0 – 3	0 – 2
3	0 – 9	1 – 5	2 – 5
4	1 – 10	2 – 7	2 – 6
5	1 – 12	3 – 8	3 – 7
6	2 – 13	4 – 9	4 – 8
7	2 – 14	4 – 10	5 – 9
8	3 – 16	5 – 11	6 – 10
9	4 – 17	6 – 12	7 – 11
10	4 – 18	7 – 13	8 – 13

Tabelle 1 *Ausschnitt aus einer von Rümke veröffentlichten Tabelle zu den 95%-Konfidenzintervallen eines Prozentsatzes an Zellen bei gegebener Gesamtzahl gezählter Zellen. a = gefundener Prozentsatz einer Population, z.B. in der Differenzierung; n = ausgezählte Zellen; X-Y = 95%-Konfidenzintervall des Ergebnisses. Das heißt, dass das Ergebnis in 95% der Fälle in diesem Bereich liegen wird, in 2,5% liegt es darunter und in weiteren 2,5% der Fälle liegt es darüber. Als Beispiel bedeutet dies für die manuelle Differenzierung von 100 WBC, dass bei gefundenen 3% Eosinophilen der wahre Wert in 95% der Messungen irgendwo zwischen 0 und 9 liegen kann. Selbst bei einer Differenzierung von 1.000 Leukozyten ist nicht auszuschließen, dass tatsächlich nicht 3, sondern 5% Eosinophile vorliegen. Die Tabelle lässt sich entsprechend für höhere Prozentsätze (z.B. relevant für Neutrophile) fortsetzen, aber auch für größere Zahlen ausgezählter Zellen (wie im automatischen Analysensystem).*

Automatisierung kann hier daher auf zweierlei Weise helfen: Bei hohen Konzentrationen (z. B. RBC im Vollblut) wird der statistische VK durch die Zählung einer besonders hohen Anzahl Zellen klein gehalten. Gleichzeitig werden zusätzliche Fehlerquellen minimiert, da die Probenentnahme und -auswertung grundsätzlich auf die immer gleiche Art und Weise erfolgt.

Hat man schließlich einen Wert, so ist die Frage, welche Aussage er erlaubt. Um zu entscheiden, ob ein bestimmter Wert pathologisch ist, wird ein Referenzintervall herangezogen, das die Verteilung der Messwerte in einer gesunden Bevölkerung widerspiegelt. Um dieses zu ermitteln, wurden Messwerte bei gesunden Probanden erhoben (ggf. auch solchen, die eine Krankheit haben, die den Messwert nicht beeinflusst). Von diesen wird gewöhnlich das 95%-Konfidenzintervall, d. h. der Bereich, in dem 95% der Messwerte liegen, als Referenzintervall verwendet. Ein Wert außerhalb des Referenzbereiches bedeutet also keineswegs zwangsläufig, dass der Proband krank ist, vielmehr haben 5% der Bevölkerung Werte, die höher oder niedriger als dieses Intervall liegen. Die Frage ist: Gehört der Proband dazu? Umgekehrt bedeutet auch ein Messwert im Referenzbereich nicht notwendigerweise, dass der Patient gesund ist.

Die eigentliche diagnostische Fragestellung wird häufig reduziert auf Gegensätze des Typs »krank« bzw. »nicht krank« oder »Therapie wirkt« vs. »Therapie wirkt nicht«. Die Entscheidung wird dabei anhand bestimmter Grenzwerte, sogenannter »Cut-offs«, oder auch Entscheidungsgrenzen, gefällt. Zur Bestimmung der Güte eines Tests wird dieser dabei verglichen mit etablierten Tests – z. B. dem jeweiligen »Goldstandard«, d. h. der vermeintlich besten verfügbaren Methode – oder sogenannten »Referenzmethoden«, d.h. Methoden, die explizit als Standard für den spezifischen Parameter definiert wurden.

Man kann die unterschiedlichen Tests dann anhand einer sogenannten „»Vierfeldertafel« vergleichen. Bei einem reinen Methodenvergleich lässt sich so die sogenannte »Konkordanz«, die Übereinstimmung zweier Testergebnisse, bestimmen. Dabei ist zu berücksichtigen, dass das Ergebnis der alten Methode nicht notwendigerweise korrekt ist.

	Alter Test positiv	Alter Test negativ
neuer Test positiv	A	B
neuer Test negativ	C	D

Tabelle 2 *Beispiel einer Vierfeldertafel für den allgemeinen Methodenvergleich*

Anders sieht die Situation aus, wenn mit einer Referenzmethode verglichen wird. Dann ist deren Ergebnis per definitionem korrekt (was Probleme aufwerfen kann, wenn die Referenzmethode eigentlich nicht sehr genau ist).

Referenzmethode:
Eine Referenzmethode ist eine analytische Methode, die z. B. von den relevanten Fachgesellschaften als zuverlässigste Methode anerkannt ist, so dass der mit ihr ermittelte Wert als »wahr« angesehen wird. Häufig wird der Begriff aber fälschlich für die verbreitetste Methode verwendet. Dies führt zu Problemen, denn auch verbreitete Methoden können falsche Ergebnisse liefern und daher nicht einfach als »wahr« angesehen werden. Dies hat Konsequenzen sowohl für die Aussagen des Methodenvergleiches als auch für die zu verwendenden Auswertungsmethoden. In solchen Fällen sollte besser ein Begriff wie »ausgewählte Vergleichsmethode« verwendet werden.

	Krankheit liegt vor	Krankheit liegt nicht vor	
Neuer Test positiv	A (z. B. 67)	B (z.B. 3) (falsch positives Ergebnis)	Wie oft hat ein positiver Test korrekt die Krankheit angezeigt? (70 Tests positiv, davon 67 korrekt)
Neuer Test negativ	C (z. B. 12) (falsch negatives Ergebnis)	D (z.B. 128)	Wie oft hat ein negativer Test korrekt die Krankheit ausgeschlossen? (140 Tests negativ, davon 128 korrekt)
	Wie oft wurde eine positive Probe auch als solche erkannt? (79 Proben positiv, 67 davon wurden korrekt erkannt)	Wie oft wurde eine negative Probe auch als solche erkannt? (131 Proben negativ, 128 davon wurden korrekt erkannt)	

Tabelle 3 *Beispiel einer Vierfeldertafel zur Güteabschätzung eines diagnostischen Tests. Mit den Beispieldaten ergeben sich eine Sensitivität von 85% (67 von 79 Proben), eine Spezifität von 98% (128 von 131 Proben), ein positiver Vorhersagewert von 96% (67 von 70 positiven Tests waren korrekt) und ein negativer Vorhersagewert von 91% (128 von 140 Tests waren korrekt).*

In diesem Fall sind die nicht übereinstimmenden Proben falsch positiv (B) oder falsch negativ (C). Mit diesen Daten können nun bei Vergleich mit einer Referenzmethode folgende Daten bestimmt werden:

- Wie häufig wird eine positive Probe auch als solche erkannt? (Sensitivität)
- Wie häufig wird eine negative Probe auch als solche erkannt? (Spezifität)

Sensitivität:

Die Sensitivität gibt an, wie häufig eine positive Probe von einem Test auch positiv erfasst wird. Bei geringer Sensitivität übersieht der Test viele positiven Proben.

Spezifität:

Die Spezifität gibt an, wie häufig eine negative Probe von einem Test auch als negativ erfasst wird. Eine geringe Spezifität führt zu vielen Fehlalarmen.

Dabei stehen diese beiden Charakteristika im Widerstreit: Je empfindlicher ein Test ist, desto höher meist auch die Wahrscheinlichkeit, irrtümlich eine negative Probe als positiv zu bestimmen. Wichtig ist hier der Anwendungsbereich eines Tests. Man kann z. B. durch ein Verschieben des Cut-off-Wertes die Kennzahlen eines Tests optimieren. Je nach spezifischer Fragestellung kann eine hohe Sensitivität oder eine hohe Spezifität das wünschenswertere Kriterium sein – es ist keineswegs immer nötig, ein optimales Gleichgewicht zwischen den beiden Charakteristika zu finden. Sensitivität und Spezifität sagen aber wenig über einen konkreten Fall aus. Sie beschreiben mehr die gesellschaftliche Sicht oder die des Versorgers. Sie geben allgemeine Indizien über die Zuverlässigkeit eines Tests.

Ebenfalls abgelesen werden können aus diesen Daten die sogenannten »Vorhersagewerte« oder prädiktiven Werte. Wo Sensitivität und Spezifität fragen, wie häufig der Zustand des Patienten korrekt erfasst wird, zeigen der positive und negative prädiktive Wert an, wie häufig ein positiver bzw. negativer Befund tatsächlich korrekt ist. Ein hoher positiv-prädiktiver Wert gibt Sicherheit, dass eine positive Probe auch mit einer Erkrankung einhergeht. Das heißt aber nicht, dass ein nicht-positiver Befund »keine Erkrankung« bedeutet, es können also Krankheiten übersehen werden. Umgekehrt erlaubt ein hoher negativ-prädiktiver Wert, negative Proben sicher als unverdächtig zu klassifizieren.

Positiver Vorhersagewert:
Der positive Vorhersagewert gibt an, wie wahrscheinlich es ist, dass z.B. eine bestimmte Krankheit tatsächlich vorliegt, wenn ein Test darauf positiv war.

Negativer Vorhersagewert:
Der negative Vorhersagewert gibt an, wie wahrscheinlich es ist, dass man bei einem negativen Testergebnis auch tatsächlich nicht unter der gesuchten Krankheit leidet.

Wichtig dabei ist es, die Häufigkeit, mit der eine Krankheit auftritt (Prävalenz), im Auge zu behalten. Ist eine Krankheit sehr selten, wird bei Proben aus der Gesamtbevölkerung auch ein hoher Anteil an negativen Proben dabei sein. Hier ist es sehr schwer, einen guten positiv-prädiktiven Wert zu erzielen. Die prädiktiven Werte beschreiben gut die Sicht des Arztes oder Patienten. Diese wollen wissen, was ein positiver oder negativer Befund bedeutet: Wie wahrscheinlich ist es, dass ich trotz eines negativen Ergebnisses an der Krankheit leide? Wie beunruhigt muss ich durch ein positives Ergebnis sein? Manche Screening-Tests nehmen eine hohe falsch-positive Rate in Kauf, da sie schnell und kostengünstig durchzuführen sind und das Ergebnis gegebenenfalls mit einem weiteren Test bei einer dann reduzierten Anzahl Probanden mit höherer Prävalenz bestätigt werden kann.

Bei der Berücksichtigung der Prävalenz muss allerdings daran gedacht werden, mit was für Proben man es zu tun hat. In einem Krankenhauslabor kann selten die Häufigkeit in der Gesamtbevölkerung verwendet werden: Die meisten Patienten befinden sich aus gutem Grund im Krankenhaus; die Wahrscheinlichkeit, dass eine Krankheit bei ihnen vorliegt, ist also wesentlich höher als im Schnitt der Gesellschaft.

Als Häufigkeit gibt die Prävalenz auch die Wahrscheinlichkeit an, dass eine bestimmte Krankheit vorliegt, wenn ich sonst nichts über den Patienten weiß. Ein guter diagnostischer Test verändert diese Wahrscheinlichkeit möglichst deutlich: Bei einem positiven Test sollte ich deutlich sicherer sein, dass die Krankheit vorliegt. Ist dies nicht der Fall, trägt der Test nicht wirklich zum Erkenntnisgewinn bei.

Diese Grundcharakteristika geben die allgemeine diagnostische Güte eines Tests an. Sie können aber auf verschiedenste Weise dargestellt werden. Statt der prädiktiven Werte können auch sogenannte »likelihood ratios« oder »odds ratios«, d.h. Wahrscheinlichkeitsverhältnisse, angegeben sein. Sensitivität und Spezifität können in Abhängigkeit von bestimmten Cut-off-Werten in einer sogenannten »receiver operating characteristic (ROC)«-Kurve dargestellt werden. Ggf. wird vereinfacht auch die Fläche unter der Kurve angegeben: Je näher diese bei 1 ist, desto besser können hohe Sensitivität und hohe Spezifität gleichzeitig erreicht werden.

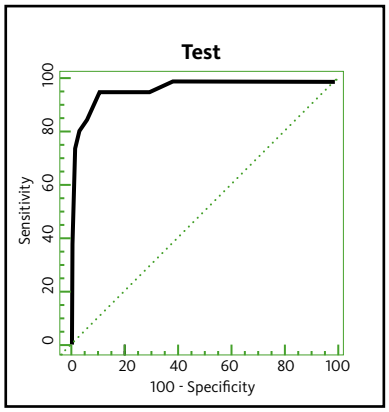


Abbildung 2a *Beispiel für eine ROC-Kurve eines Tests, der gute Sensitivität und Spezifität bietet und diese auch gut miteinander vereinbar macht. Die Fläche unter der Kurve ist hier sehr nahe an der Fläche des gesamten Diagramms, d.h. 1.*

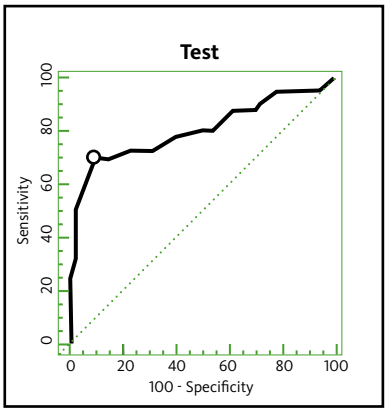


Abbildung 2b *Beispiel für einen Test, der hohe Sensitivität nur auf Kosten eines starken Abfalls der Spezifität erreichen kann. Markiert ist der Punkt, der die beste Kombination bietet. Die Fläche unter der Kurve ist deutlich geringer als bei A.*

Letztendlich liegen diesen Darstellungsformen aber die gleichen Daten zugrunde und sie können ggf. ineinander überführt werden. Auf eine tiefergehende Erörterung an dieser Stelle wird daher verzichtet. Angesichts der Vielfalt statistischer Aspekte, die sowohl bei der Interpretation von Laborwerten an sich, als auch bei der Bewertung und beim Vergleich verschiedener Tests eine Rolle spielen, werden zukünftige Beiträge zu diesem Thema diese Punkte gegebenenfalls aufgreifen.

ROC-Kurve:
»Receiver operating characteristic«-Kurve: Auftragung der Sensitivität gegen »1-Spezifität«. Erlaubt abzulesen, inwieweit Sensitivität und Spezifität optimiert werden können. Die Diagonale (Winkelhalbierende) der Achsen beschreibt eine 50 %- Chance des Vorliegens der Krankheit. Tests, die sich nahe an dieser Linie befinden, sind diagnostisch nicht viel wertvoller als ein Münzwurf. Die Fläche unter der Kurve (Area under curve, AUC) der Diagonalen ist 0,5. Tests, deren Kurve sich in die obere linke Ecke drängen und eine AUC nahe 1 haben, erlauben dagegen gleichzeitig eine hohe Sensitivität und Spezifität. Tatsächlich muss aber bei vielen Fragestellungen (z.B. beim Screening) gar nicht das Optimum der gesamten Kurve gefunden werden, da z. B. eine hohe Sensitivität ausreicht.

Sysmex Deutschland GmbH
Bornbarch 1, 22848 Norderstedt, Germany
Phone +49 (40) 534102-0 · Fax +49 (40) 5232302
www.sysmex.de