Der binäre diagnostische Test

Um zur Diagnose einer Krankheit zu gelangen, benutzt der Arzt einen diagnostischen Test. Dieser kann z.B. ein Labortest, ein bildgebendes Verfahren, aber auch eine klinische Untersuchung sein, bei der Fachkenntnisse und Erfahrungen einfließen. Bekanntlich sind Diagnosen mit einer gewissen Unsicherheit verbunden. Somit ist ein diagnostischer Test ein Verfahren, das Vorliegen einer Krankheit anhand diagnostischer Kriterien mit einer gewissen Wahrscheinlichkeit (WSK) zu bestimmen. Es klassifiziert im binären Fall Patienten anhand von Laborwerten, Symptomen, Röntgenbildern, klinischen Befunden ua. in "krank" oder "gesund". Hohe Laborwerte, Verschattungen im Röntgenbild, Verfärbungen der Haut und vieles andere können als diagnostische Kriterien dienen.
Mit guten Tests werden viele tatsächlich Kranke und tatsächlich Gesunde richtig erkannt. Diese Fähigkeit eines diagnostischen Tests zwischen Kranken und nicht Kranken zu unterscheiden, die Validität, wird bestimmt durch zwei bedingte WSK: Sensitivität (Sens) und Spezifität (Spez).

Sens = WSK(T+|K+) ist die WSK, dass der Test positiv reagiert bei Kranken,
Spez = WSK(T-|K-) ist die WSK, dass der Test negativ reagiert bei Gesunden.

Diese Gütekriterien eines diagnostischen Tests werden in Diagnosestudien mit Hilfe eines binären Goldstandards ermittelt und sind prävalenzunabhängig. Als Goldstandard dienen tatsächlich Kranke und tatsächlich Gesunde, deren Status durch andere Verfahren (z.B. feingewebliche Untersuchungen) zweifelsfrei ermittelt wurde. In der Praxis werden auch Expertenmeinungen als Goldstandard verwendet, falls dieser nur aufwändig oder risikoreich verfügbar ist.

Beispiel 1: Diagnostik der Zahnkaries (EbM-Splitter, DZZ 57, 2002)

Eine gegebene Zahl von Zahnärzten sollte klinisch (mit Spiegel und Sonde) beurteilen, ob die ihnen vorgelegten 5722 Kontaktflächen von Zähnen kariös oder gesund sind. Als Goldstandard diente hier ein Röntgenbild der Zähne, das von Experten begutachtet wurde (Expertenmeinung). Die Ergebnisse sind in folgender Vierfeldertafel zusammengefasst:

		Röntgenbild		
		kariös	gesund	∑
Klinische	kariös	172	5	177
Untersuchung	gesund	385	5160	5545
	∑	557	5165	5722

Von 557 tatsächlich kariösen Kontaktflächen wurden von den Zahnärzten 172 erkannt und von den 5165 gesunden Zähnen waren es 5160. Somit ergab sich für die Güte der klinischen Beurteilung:
Sens = 172/557 = 0,309 (**30,9%**) und **Spez** = 5160/5165 = 0,999 (**99,9%**) für die beteiligten Zahnärzte.
Falsch negativ beurteilt wurden 385 von 557 Kontaktflächen (69,1%) und falsch positiv 5 von 5165 (0,1%).
Aufgrund der großen Zahl gesunder Kontaktflächen, die fast alle richtig erkannt wurden, gibt es in diesem Fall einen hohen Anteil übereinstimmender Befunde, 172 + 5160 = 5332 und insgesamt 5332 / 5722 = 0,932 (93,2%). Allerdings wurden nur 30,9% der kariösen Kontaktflächen richtig erkannt, was auf die im Allgemeinen schwierige Beurteilung der Kontaktflächen zurückzuführen ist.

Es wurden hier Sens und Spez aus Stichproben geschätzt und daher sind beide mit einer gewissen Unsicherheit behaftet, deren Quantifizierung durch ein Konfidenzintervall (K.I.) erfolgt. Da es sich um Anteile handelt, findet die unter SRS_Teil2 genannte Formel (nach Wald) unter der Voraussetzung $n \cdot p \cdot (1-p) > 9$ Anwendung. Ist diese Voraussetzung nicht erfüllt, sollten exakte Berechnungen erfolgen.

$$P_{u,o} = \left(p \mp \frac{1}{2n} \right) \mp z_{1-\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n-1}} \cdot \sqrt{\left(1 - \frac{n}{N}\right)}$$

Unter Vernachlässigung von Endlichkeits- und Stetigkeitskorrektur erhält man für die **Sens** ein K.I. von
$\{0{,}309 - 1{,}96 \cdot (0{,}309 \cdot 0{,}691 / 556)^{1/2}$; $0{,}309 + 1{,}96 \cdot (0{,}309 \cdot 0{,}691 / 556)^{1/2}\} = (0{,}271$; $0{,}347)$
Die Voraussetzung $n \cdot p \cdot (1-p) = 557 \cdot 0{,}309 \cdot 0{,}691 = 119 > 9$ ist erfüllt.

Für die **Spez** findet man
$\{0{,}999 - 1{,}96 \cdot (0{,}999 \cdot 0{,}001 / 5164)^{1/2}\}$; $0{,}999 + 1{,}96 \cdot (0{,}999 \cdot 0{,}001 / 5164)^{1/2}\} = (0{,}998$; $1{,}000)$
Die Voraussetzung $n \cdot p \cdot (1-p) = 5165 \cdot 0{,}999 \cdot 0{,}001 = 5 < 9$ ist hier nicht erfüllt, so dass eine exakte Berechnung des K.I. heranzuziehen ist. Hier liefert z.B. die nach Clopper-Pearson auf drei Dezimalstellen gerundet allerdings das gleiche Ergebnis (0,998 ; 1,000).

Beispiel 2: Test auf HLA bei einer spezieller Gelenkerkrankung (Schlosstein et al 1973).
HLA = Humanes Leukozyten Antigen bei ankylosierender Spondylitis

Zur Bestimmung der Gütekriterien Sens und Spez wurde eine Gruppe nachweislich an der fraglichen Krankheit Erkrankter (40 Personen) und eine Gruppe nachweislich Gesunder (906 Personen) rekrutiert und an diesen Personen der Labortest auf HLA durchgeführt. Im Ergebnis wurden 35 von 40 Erkrankten als richtig positiv und 834 von 906 als richtig negativ klassifiziert. Somit ergab sich für die Güte des Tests:
Sens = 35 / 40 = 0,875 (**87,5%**) und **Spez** = 834 / 906 = 0,92053 (**92,1%**)
Wegen 40·0,875·0,125 = 4 < 9 wählen wir für die Sens das K.I. nach Clopper-Pearson zu (0,732 ; 0,958)

Wegen 906·0,921·0,079 = 66 > 9 wählen wir für die Spez das oben genannte K.I. nach Wald zu (0,903 ; 0,938)
Falsch negativ beurteilt wurden 5 von 40 (12,5%) und falsch positiv 72 von 906 (7,9%).

		Gelenkerkrankung ja	nein	∑
HLA - Test	positiv	35	72	107
	negativ	5	834	839
	∑	40	906	946

Das Programm "Open Epi" liefert weitere Berechnungsmöglichkeiten für das Konfidenzintervall (siehe Literatur und Software)

95% Confidence Limits for Proportion 35/40
Multiplier=100
Large population size or sample with replacement.

	Lower CL Per 100	Upper CL
Proportion as Percent	87.5	
Mid-P Exact	74.45	95.27
Fisher Exact(Clopper-Pearson)	73.2	95.81
Wald (Normal Approx.)	77.25	97.75
Modified Wald(Agresti-Coull)	73.42	95.01
Score(Wilson)*	73.89	94.54

95% Confidence Limits for Proportion 834/906
Multiplier=100
Large population size or sample with replacement.

	Lower CL Per 100	Upper CL
Proportion as Percent	92.053	
Mid-P Exact	90.15	93.68
Fisher Exact(Clopper-Pearson)	90.1	93.73
Wald (Normal Approx.)	90.29	93.81
Modified Wald(Agresti-Coull)	90.1	93.65
Score(Wilson)*	90.11	93.64

Die Prädiktiven Werte PW+ und PW-
Patienten möchten normalerweise nicht wissen, wie gut der Test bei Gesunden und Kranken die richtige Auskunft gibt. Sie möchten wissen, ob sie erkrankt sind bei positivem Test oder nicht erkrankt bei negativem Test. Eine Antwort geben die prävalenzabhängigen prädiktiven Werte PW+ und PW- . Dabei ist

PW+ = WSK(K+|T+) die WSK krank zu sein bei positivem Test,
PW- = WSK(K-|T-) die WSK gesund zu sein bei negativem Test.

Beide Werte sind **prävalenzabhängig**. Man erhält sie durch Anwendung der Bayes - Formeln:

$$PW+ = \frac{P \cdot SENS}{(P \cdot SENS) + (1-P) \cdot (1-SPEZ)} \quad \text{und} \quad PW- = \frac{(1-P) \cdot SPEZ}{(1-P) \cdot SPEZ + P \cdot (1-SENS)}$$

Dabei ist P die Prävalenz in der Population (nicht in der Test-Studie), die vor Anwendung des Tests geschätzt werden muss, und SENS und SPEZ die Gütekriterien des Tests.

Berechnung der prädiktiven Werte beim Test auf HLA in Abhängigkeit von der Prävalenz
Die **Prävalenz der Gelenkerkrankung** in der deutschen **Allgemeinbevölkerung** sei etwa **1,9%** = 0,019 .
Bei einem Screening von 1000 Personen in der Allgemeinbevölkerung erhält man für die prädiktiven Werte mit Hilfe der **Bayes - Formeln**:

PW+ = (0,019·0,875) / {(0,019·0,875) + (0,981·0,07947)} = 0,17577 = 17,6%
PW- = (0,981·0,92053) / {(0,981·0,92053) + (0,019·0,125)} = 0,9974 = 99,7%

Eine **zweite Möglichkeit** zur Berechnung der prädiktiven Werte bietet eine konstruierte 2x2 - Tafel, bei der die Prävalenz dem Wert (a+c)/n entspricht mit den gegebenen Werten für p = (a+c)/n = 0,019 ,
Sens = a/(a+c) = 0,875 , Spez = d/(b+d) = 0,92053 und n = 1000. Man rechnet:

(a+c) = 1000·0,019 = 19 ; a = 19·0,875 = 16,625 ; (b+d) = n - ((a+c) = 1000 - 19 = 981
d = 981·0,92053 = 903,04 ; b = 981 - 903 = 78 ; c = 19 - 17 = 2

Damit erhält man folgende Vierfeldertafel (Werte gerundet):

	Krankheit		
	ja	nein	
T+	a	b	a+b
T-	c	d	c+d
	a+c	b+d	n

	Krankheit		
	ja	nein	
T+	17	78	95
T-	2	903	905
	19	981	1000

Ein approximatives K.I. bekommt man jetzt durch Berechnung der Konfidenzgrenzen für den Anteil der tatsächlich Kranken (Gesunden) an den Testpositiven (Testnegativen), den man direkt aus der 2x2 - Tafel entnehmen kann:

Konfidenzintervall für PW+ = W(K+|T+) = Anteil der Kranken unter den Testpositiven
Zahl der Kranken unter den Testpositiven: 1000·0,019·Sens = 16,625 ≈ 17 = a
Zahl der Testpositiven: 16,625 / 0,17557 = 94,6916 ≈ 95 = (a+b); **Anteil** (gerundet): 17/95 = a/(a+b)
K.I. (10,2% ; 25,6%)

Konfidenzintervall für PW- = W(K-|T-) = Anteil der Gesunden unter den Testnegativen
Zahl der Gesunden unter den Testnegativen: 1000·0,981·Spez = 903,04
Zahl der Testnegativen: 903,04 / 0,9974 = 905,394 ; **Anteil** (gerundet): 903/905 ; K.I. (99,5% ; 100%)

Ergebnis: Bei positivem Test in der Allgemeinbevölkerung mit einer Prävalenz von 1,9% beträgt die WSK, erkrankt zu sein nur 17,7%. Bei negativem Test kann man mit 99,7% recht sicher sein, die Erkrankung nicht zu haben.

Prävalenz der Gelenkerkrankung 5%

Betrachtet man nun eine Gruppe von **Personen mit unspezifischen Gelenkbeschwerden**, die deshalb ihren Hausarzt aufsuchen, der einen HLA-Test verordnet. In einer solchen Gruppe sei die Prävalenz der Erkrankung etwa 5% = 0,05. Man erhält dann:
PW+ = (0,05·0,875) / {(0,05·0,875) + (0,95·0,07947)} = 0,36689 = 36,7%
PW- = (0,95·0,92053) / {(0,95·0,92053) + (0,05·0,125)} = 0,9929 = 99,3%

Konfidenzintervall für PW+
Zahl der Kranken unter den Testpositiven: 1000·0,05·Sens = 43,75
Zahl der Testpositiven: 43,75 / 0,36689 = 119,24555 ; Anteil (gerundet): 44/119 ; K.I. (28,3% ; 45,6%)
Konfidenzintervall für PW-
Zahl der Gesunden unter den Testnegativen: 1000·0,95·Spez = 874,5035
Zahl der Testnegativen: 874,5035 / 0,9929 = 880,7569 ; Anteil (gerundet): 874/881 ; K.I. (98,6% ; 99,8%)

Ergebnis: Personen mit unspezifischen Gelenkbeschwerden, die einen positiven Test erhalten, haben jetzt bereits eine WSK erkrankt zu sein von 36,7%.
Ist der Test negativ kann man mit 99,3% wiederum recht sicher sein, die Erkrankung nicht zu haben.

Je größer die Prävalenz (A-priori-WSK, Vortest-WSK) in der untersuchten Gruppe, desto größer ist die WSK, bei einem positivem Test erkrankt zu sein (A-posteriori-WSK).

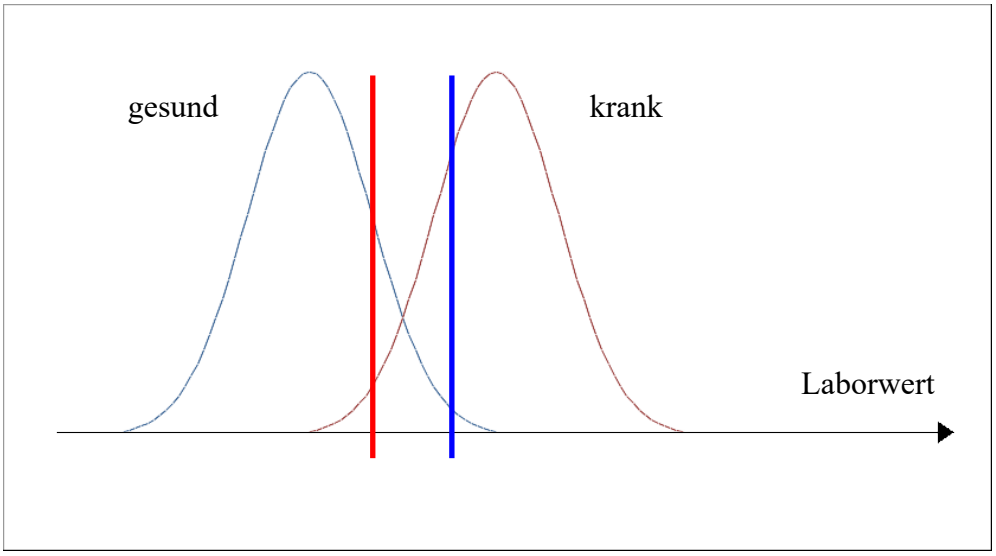
Anwendung sensitiver und spezifischer Tests

Wenn das Übersehen einer Krankheit bei einem diagnostischen Test schwerwiegende Folgen hat (z.B. bei HIV oder Krebs), ist ein sensitiver Test (hohe Sensitivität) sinnvoll, bei dem wenig falsch negative Ergebnisse auftreten. Ist in solchen Fällen ein Test negativ, kann die Krankheit mit hoher WSK ausgeschlossen werden.
Bei einem spezifischen Test gibt es sehr wenige falsch positive Ergebnisse. Mit hochspezifischen Tests kann eine Krankheit im Falle eines positiven Testergebnisses bestätigt werden.

Der diagnostische Test mit metrischen Daten

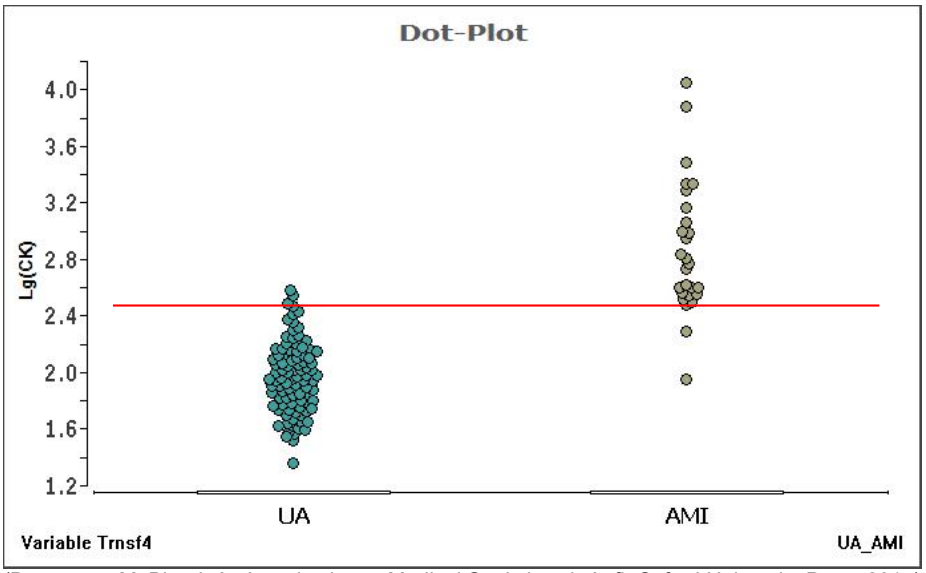
Zur Diagnose vieler Erkrankungen werden Laborwerte mit metrischer Skala herangezogen (z.B. Blutdruck, Cholesterin). Dabei ist es hilfreich, wenn sich die Verteilungen der Werte für Kranke und Gesunde unterscheiden. Die erste Grafik demonstriert ein fiktives Beispiel, bei dem ein hoher Laborwert als Hinweis auf eine Erkrankung, ein niedriger Laborwert für gesunde Personen spricht. Je mehr sich die Verteilungen überschneiden, desto ungenauer trennt der Test mit diesem Laborwert. Je weniger sich die Verteilungen überschneiden, desto genauer trennt der Test. Ziel ist es, einen optimalen Grenzwert (cut-point) für einen diagnostischen Test zu bestimmen, bei dem Personen mit höheren Laborwerten als "krank" und solche mit

Werten kleiner als der Grenzwert als "gesund" klassifiziert werden. In vielen Fällen, in denen Patienten noch nicht erkrankt sind sondern lediglich ein erhöhtes Risiko für eine Erkrankung aufweisen (KHK bei hohem Cholesterinwert), kann eine Entscheidung "behandlungsbedürftig" (z.B. Cholesterinsenker) oder "nicht behandlungsbedürftig" getroffen werden.



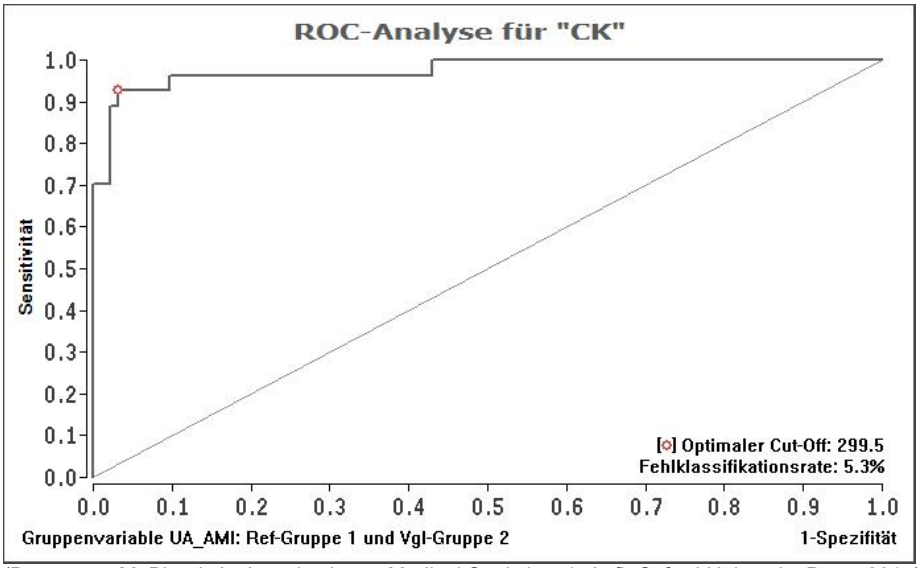
Legt man den Grenzwert auf einen niedrigen Laborwert (Rot), werden zwar die meisten Kranken als krank klassifiziert, viele Gesunde allerdings auch (viele falsch positive).
Legt man den Grenzwert auf einen höheren Laborwert (Blau), werden die meisten Gesunden als gesund klassifiziert, doch darunter fallen auch viele wirklich Kranke (viele falsch negative).

Die zweite Grafik zeigt ein reales Beispiel für die Verteilung des Laborwertes Kreatinkinase (Enzym in Muskelzellen) im Blut von Patienten zur Diagnose von unstabiler Angina (UA) oder akutem Herzinfarkt (AMI).



(Daten aus: M. Bland: An Introduction to Medical Statistics. 4. Aufl. Oxford University Press 2015)

Eine Möglichkeit, den Zusammenhang zwischen Sens und Spez grafisch darzustellen bietet die sog. ROC-Kurve. Hier werden für verschiedene Grenzwerte (cut-poits) die daraus resultierenden Werte für Sens auf der y-Achse und (1-Spez) auf der x-Achse aufgetragen. Die dritte Grafik zeigt eine ROC-Kurve für das Beispiel Kreatinkinase (Enzym in Muskelzellen). Die Software "BiAS", mit der die letzten zwei Grafiken erstellt wurden, ermittelt den optimalen Grenzwert, bei dem sowohl Sens, als auch Spez hoch sind, für diesen Labortest zu 299,5 U/l (SENS = 93%, SPEZ = 97%). Die Fläche unter der Kurve (area under the curve AUC) ist ein Maß für die Leistung des Testes (diagnostic accuracy) und gleich der WSK für eine Person mit der entsprechenden Krankheit, einen größeren Laborwert zu erhalten als eine Person ohne diese Krankheit (hier AUC = 0,975).



(Daten aus: M. Bland: An Introduction to Medical Statistics. 4. Aufl. Oxford University Press 2015)