

GRUNDLAGEN MEDIZINISCHER NACHBARGEBIETE

Editorial

Leider ist nach Abschluss der beiden Serien „Biometrie“ und „Klinische Pharmakologie“ in der Rubrik „Grundlagen medizinischer Nachbarggebiete“ eine Lücke eingetreten. Die Rubrik soll jedoch im nächsten Heft mit einer Reihe von Beiträgen zum Thema „Qualitätsmanagement“ wieder aufgenommen werden. Hierüber wird dann ausführlich berichtet.

In diesem Heft publizieren wir noch eine nachträglich eingegangene Arbeit zum Thema „Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung“ von R. Bender, Bielefeld, gewissermaßen als Ergänzung zu unserer Serie „Biometrie“. Für diejenigen Leser, die sich das Sonderheft mit den Beiträgen zur „Biometrie“ und zur „Klinischen Pharmakologie“ aufbewahrt haben, empfiehlt es sich, den Beitrag dieses Heftes als Ergänzung beizufügen.

J. Köbberling

Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung

Ralf Bender¹

ZUSAMMENFASSUNG

□ *Hintergrund:* Zur Beschreibung der Effizienz eines diagnostischen Tests sowie zur Darstellung der Effektivität einer Behandlung existiert eine Vielzahl von Maßen, die sich alle aus einer Vierfeldertafel absoluter Häufigkeiten ableiten lassen. Zum Verständnis dieser Maße ist eine Kenntnis ihrer Eigenschaften im Rahmen der Wahrscheinlichkeitsrechnung notwendig.

□ *Material und Methode:* Nach einer Einführung in grundlegende Begriffe, wie Wahrscheinlichkeit, Chance, gemeinsame und bedingte Wahrscheinlichkeit, werden die gebräuchlichen Maße Sensitivität, Spezifität, Likelihood Ratio, positiver und negativer prädiktiver Wert, relatives Risiko, Odds Ratio, relative Risikoreduktion, absolute Risikoreduktion und „Number Needed to Treat“ vorgestellt und erklärt. Anhand von Beispielen wird insbesondere auf die Bedeutung der Krankheitsprävalenz bzw. des Basisrisikos für die Interpretation dieser Maße eingegangen.

□ *Schlussfolgerung:* Wird die Höhe der Krankheitsprävalenz und des Basisrisikos nicht genügend berücksichtigt, so werden der Nutzen eines diagnostischen Tests und der Effekt einer Behandlung insbesondere im Rahmen von Screening und Prävention überschätzt.

Schlüsselwörter: Diagnostische Tests · Odds Ratio · Prädiktive Werte · Prävalenz · Wahrscheinlichkeit · Risiko · Sensitivität und Spezifität · Statistik

Med Klin 2001;96:116–21.

DOI 10.1007/s00063-001-1019-6

ABSTRACT

Interpretation of Efficacy Measures Derived from 2×2 Tables for the Evaluation of Diagnostic Tests and the Effect of Treatment

□ *Background:* To describe the efficacy of diagnostic tests and the effect of treatment a number of measures are used, which can be derived from 2×2 tables of frequencies. For the comprehension of these measures the knowledge of their properties in the framework of probability theory is necessary.

Im Rahmen von Screening und Prävention wird immer wieder versucht, den Nutzen eines Screening-Tests allein durch die Angabe von Effizienzmaßen wie Sensitivität und Spezifität und den Effekt einer Behandlung nur durch relative Maße wie relatives Risiko oder relative Risikoreduktion zu beschreiben. Die Bedeutung der Krankheitsprävalenz für den Vorhersagewert diagnostischer Maßnahmen und für den absoluten Effekt einer Behandlung wird häufig nicht genügend beachtet. Irrtümer und Trugschlüsse lassen sich nur vermeiden, wenn man die Eigenschaften der gebräuchlichen Effizienzmaße im Rahmen der Wahrscheinlichkeitsrechnung und ihre Vor- und Nachteile kennt. In dieser Arbeit werden in möglichst einfacher Weise die gebräuchlichen Effizienzmaße für Diagnostik und Behandlung aus einer Vierfeldertafel abgeleitet und ihre wesentlichen Eigenschaften dargestellt.

Vierfeldertafel und Wahrscheinlichkeiten

Eine Vierfeldertafel ist eine (2×2) -Anordnung absoluter Häufigkeiten, die sich aus einer Kreuzklassifikation von zwei binären Merkmalen X und Y ergeben (Tabelle 1). Ein binäres Merkmal ist eine Variable mit nur zwei Ausprägungen (zum Beispiel Krankheit ja/nein, Behandlung ja/nein, Erfolg ja/nein).

Um den Stellenwert einer gegebenen Vierfeldertafel beurteilen zu können,

¹ AG 3 – Epidemiologie und Medizinische Statistik, Fakultät für Gesundheitswissenschaften, Universität Bielefeld.

Tabelle 1. Vierfeldertafel.

	X		Summe
	1	0	
Y	1 0	a b c d	a+b c+d
Summe	a+c	b+d	n = a+b+c+d

muss das Studiendesign berücksichtigt werden, mit dem die Daten erhoben worden sind [1]. In Abhängigkeit vom Design lassen sich aus einer Vierfeldertafel verschiedene Wahrscheinlichkeiten durch relative Häufigkeiten schätzen. Das allgemeine Prinzip wird zunächst erläutert unter der Annahme, dass die n Individuen aus Tabelle 1 eine Zufallsstichprobe aus einer interessierenden Grundgesamtheit darstellen. Aus der Vierfeldertafel lassen sich dann drei verschiedene Typen von Wahrscheinlichkeiten schätzen, auf denen die in den folgenden Kapiteln vorgestellten Effizienzmaße basieren. Einzelwahrscheinlichkeiten beziehen sich auf die Eintrittshäufigkeiten von Ausprägungen eines Merkmals, gemeinsame Wahrscheinlichkeiten auf das gemeinsame Auftreten von Ausprägungen zweier Merkmale und bedingte Wahrscheinlichkeiten auf die Eintrittshäufigkeiten von Ausprägungen eines Merkmals unter der Bedingung, dass die Ausprägung eines zweiten Merkmals vorliegt. Mit den Bezeichnungen aus Tabelle 1 gelten die folgenden Formeln (siehe Kasten unten).

Einzelwahrscheinlichkeiten:	
$P(X=1) = \frac{a+c}{n}$	$P(X=0) = 1 - P(X=1) = \frac{b+d}{n}$
$P(Y=1) = \frac{a+b}{n}$	$P(Y=0) = 1 - P(Y=1) = \frac{c+d}{n}$
Gemeinsame Wahrscheinlichkeiten:	
$P(X=1, Y=1) = \frac{a}{n}$	$P(X=0, Y=0) = \frac{d}{n}$
$P(X=1, Y=0) = \frac{c}{n}$	$P(X=0, Y=1) = \frac{b}{n}$
Bedingte Wahrscheinlichkeiten:	
$P(Y=1 X=1) = \frac{a}{a+c}$	$P(Y=0 X=1) = 1 - P(Y=1 X=1) = \frac{c}{a+c}$
$P(Y=1 X=0) = \frac{b}{b+d}$	$P(Y=0 X=0) = 1 - P(Y=1 X=0) = \frac{d}{b+d}$
$P(X=1 Y=1) = \frac{a}{a+b}$	$P(X=0 Y=1) = 1 - P(X=1 Y=1) = \frac{b}{a+b}$
$P(X=1 Y=0) = \frac{c}{c+d}$	$P(X=0 Y=0) = 1 - P(X=1 Y=0) = \frac{d}{c+d}$

□ **Material and Method:** After an introduction of basic terms such as probability, odds, joint and conditional probability the usual measures sensitivity, specificity, likelihood ratio, positive and negative predictive value, relative risk, odds ratio, relative risk reduction, absolute risk reduction, and number needed to treat are presented and explained. In particular, the importance of disease prevalence and baseline risk for the interpretation of these measures is pointed out by means of examples.

□ **Conclusion:** If the disease prevalence or the baseline risk is not appropriately taken into account, the efficacy of a diagnostic test and the effect of a treatment are overestimated, especially in screening and prevention trials.

Key Words: Diagnostic tests · Odds ratio · Predictive value of tests · Prevalence · Probability · Risk · Sensitivity and specificity · Statistics and numerical data

Med Klin 2001;96:116–21.
DOI 10.1007/s00063-001-1019-6

In endlichen Stichproben sind die angegebenen Formeln als Schätzungen für theoretische Wahrscheinlichkeiten anzusehen. In der Praxis sollte die Präzision der Schätzungen mit Hilfe von Konfidenzintervallen für Proportionen angegeben werden [17].

Chancen

Eine Wahrscheinlichkeit ist eine Quantifizierung der Eintrittshäufigkeit eines Ereignisses unter Verwendung des Zahlenintervalls (0,1). Eine weitere Möglichkeit, Eintrittshäufigkeiten zu beschreiben, ist die so genannte Chance (engl.: Odds). Bezeichnet p eine Wahrscheinlichkeit, dann ist die Chance definiert durch

$$\text{Odds} = \frac{p}{1-p}$$

Daraus ergibt sich, dass eine Chance immer im Zahlenbereich (0,∞) liegt. Im allgemeinen Sprachgebrauch wird Wahrscheinlichkeit mit Chance häufig gleichgesetzt. Das ist insofern richtig, als dass man eine Eintrittshäufigkeit sowohl durch eine Wahrscheinlichkeit als auch äquivalent hierzu durch eine Chance angeben kann. Da gilt

$$p = \frac{\text{Odds}}{1 + \text{Odds}}$$

kann man Wahrscheinlichkeiten in Chancen umrechnen und umgekehrt. Eine Gleichsetzung von Wahrscheinlichkeit und Chance ist aber insofern falsch, als beide Maße unterschiedliche numerische Zahlen liefern. Nur bei Wahrscheinlichkeiten, die kleiner sind als 0,1, ergeben beide Maße ungefähr gleiche numerische Werte. Da die Chance ein Quotient von Wahrscheinlichkeiten ist, ist die Wahrscheinlichkeit leichter interpretierbar als die Chance. Die Darstellung von Eintrittshäufigkeiten als Chance hat aber bei gewissen Anwendungen Vorteile (zum Beispiel bei der Berechnung von Vorhersagewerten diagnostischer Tests). In Tabelle 2 werden beispielhaft einige äquivalente Wahrscheinlichkeiten und Chancen gegenübergestellt.

Effizienz eines diagnostischen Tests

Eine gute Übersicht über Methoden der Diagnoseevaluierung geben Richter et al. [15] sowie Jaeschke et al. [12, 13]. Ich beschränke mich hier auf die wesentlichsten Eigenschaften der üblichen Effizienzmaße diagnostischer Tests. Diese las-

GRUNDLAGEN MEDIZINISCHER NACHBARGEBIETE

p	Odds = p/(1-p)			
0,9000	0,9 / 0,1	=	9,0000	= 9:1
0,8000	0,8 / 0,2	=	4,0000	= 4:1
0,7500	0,75 / 0,25	=	3,0000	= 3:1
0,7000	0,7 / 0,3	=	2,3333	= 7:3
0,6667	0,667 / 0,333	=	2,0000	= 2:1
0,6000	0,6 / 0,4	=	1,5000	= 3:2
0,5000	0,5 / 0,5	=	1,0000	= 1:1
0,4000	0,4 / 0,6	=	0,6667	= 2:3
0,3000	0,3 / 0,7	=	0,4286	= 3:7
0,2000	0,2 / 0,8	=	0,2500	= 1:4
0,1000	0,1 / 0,9	=	0,1111	= 1:9
0,0500	0,05 / 0,95	=	0,0526	= 1:19
0,0100	0,01 / 0,99	=	0,0101	= 1:99
0,0010	0,001 / 0,999	=	0,0010	= 1:999
...	...	=

Tabelle 2. Beispiele für Wahrscheinlichkeit und Chance.

dar. Das Likelihood Ratio beschreibt die Macht eines Tests zur Veränderung der Prätest-Wahrscheinlichkeit (siehe nächster Abschnitt). Die Eignung eines Tests für bestimmte Zwecke ist situationsabhängig und muss von Fall zu Fall beurteilt werden, zum Beispiel weil in manchen Situationen mehr Wert auf hohe Sensitivität und in anderen mehr Wert auf hohe Spezifität gelegt werden muss.

Vorhersagewert eines diagnostischen Tests

Sensitivität, Spezifität und Likelihood Ratio beschreiben hinreichend die allgemeine Güte eines diagnostischen Tests. In der klinischen Anwendung beantworten diese Maße aber nicht die Frage nach der Wahrscheinlichkeit für das Vorliegen der Krankheit nach Durchführung des Tests. Für die diagnostische Situation in der klinischen Praxis sind daher die Vorhersagewerte

$$\text{Positiver prädiktiver Wert (PPW)} = \frac{P(K+ | T+)}{P(K+ | T+)} = \frac{P(T+ | K+) \cdot P(K+)}{P(T+ | K+) \cdot P(K+) + P(T+ | K-) \cdot P(K-)}$$

$$\text{Negativer prädiktiver Wert (NPW)} = \frac{P(K- | T-)}{P(K- | T-)} = \frac{P(T- | K-) \cdot P(K-)}{P(T- | K-) \cdot P(K-) + P(T- | K+) \cdot P(K+)}$$

wichtiger [4]. Wie die Krankheitsprävalenz lassen sich diese bedingten Wahrscheinlichkeiten nicht aus einer Vierfeldertafel schätzen, die auf zwei Stichproben – eine aus den Gesunden und eine aus den Kranken – basiert. Die Vorhersagewerte lassen sich aber mit Hilfe der Bayes'schen Formel [6] in Abhängigkeit von Sensitivität, Spezifität und Prävalenz darstellen. Es gilt

$$\text{PPW} = \frac{\text{Sensitivität} \times \text{Prävalenz}}{\text{Sensitivität} \times \text{Prävalenz} + (1 - \text{Spezifität}) \times (1 - \text{Prävalenz})}$$

$$\text{NPW} = \frac{\text{Spezifität} \times (1 - \text{Prävalenz})}{\text{Spezifität} \times (1 - \text{Prävalenz}) + (1 - \text{Sensitivität}) \times \text{Prävalenz}}$$

Diese Formeln lassen sich durch Darstellung der Vorhersagewerte als Chancen anstelle von Wahrscheinlichkeiten wesentlich vereinfachen. Es gilt

$$\frac{\text{PPW}}{1 - \text{PPW}} = \frac{\text{Prävalenz}}{1 - \text{Prävalenz}} \times \text{PLR}$$

$$\frac{\text{NPW}}{1 - \text{NPW}} = \frac{\text{Prävalenz}}{1 - \text{Prävalenz}} \times \text{NLR}$$

sen sich als bedingte Wahrscheinlichkeiten aus einer Vierfeldertafel mit den beiden Merkmalen Krankheit ja/nein und Test positiv/negativ ableiten (Tabelle 3).

Zu beachten ist hierbei, dass in Studien zur Evaluierung diagnostischer Tests häufig die n Individuen keine Zufallsstichprobe aus der interessierenden Population darstellen, sondern dass zwei Stichproben vorliegen, eine aus der Population der Gesunden und eine aus der Population der Kranken. Die Stichprobenumfänge werden hierbei vom Untersucher vorgegeben, sodass nicht alle o. g. Wahrscheinlichkeiten aus einer solchen Vierfeldertafel mit teilweise vorgegebenen Randhäufigkeiten sinnvoll geschätzt werden können. Insbesondere die Krankheitsprävalenz ist nicht schätzbar und muss aus anderen Quellen abgeleitet werden.

Die grundlegenden Effizienzmaße eines diagnostischen Tests sind Sensitivität und Spezifität [3]. Die Sensitivität ist definiert als der Anteil positiver Tests unter den Kranken und die Spezifität als der Anteil der negativen Tests unter den Gesunden, das heißt

$$\text{Sensitivität} = P(T+ | K+) = \frac{a}{a+c}$$

$$\text{Spezifität} = P(T- | K-) = \frac{d}{b+d}$$

Ein Maß, das sowohl die Sensitivität als auch die Spezifität eines Tests berücksichtigt, ist das Likelihood Ratio [13]. Das Likelihood Ratio (LR) ist definiert als das Verhältnis der Anteile eines Testresultats unter den Kranken und unter den Gesun-

den. Bei binären Tests gibt es somit zwei Likelihood Ratios, und zwar

$$\text{Positives Likelihood Ratio (PLR)} = \frac{P(T+ | K+)}{P(T+ | K-)} = \frac{\text{Sensitivität}}{1 - \text{Spezifität}}$$

$$\text{Negatives Likelihood Ratio (NLR)} = \frac{P(T- | K+)}{P(T- | K-)} = \frac{1 - \text{Sensitivität}}{\text{Spezifität}}$$

Anhand der Likelihood Ratios lässt sich die Güte eines diagnostische Tests, wie in Tabelle 4 angegeben, grob klassifizieren.

Diese Klassifizierung der Güte eines Tests stellt nur eine grobe Faustregel

Tabelle 3. Vierfeldertafel zur Evaluation eines diagnostischen Tests.

	Krankheit		Summe
	ja (+)	nein (-)	
Test +	a	b	a+b
Test -	c	d	c+d
Summe	a+c	b+d	n= a+b+c+d

Tabelle 4. Grobe Einteilung des Effizienz eines diagnostischen Tests.

Positives Likelihood Ratio	Negatives Likelihood Ratio	Test-effizienz
> 10	< 0,10	sehr gut
5 – 10	0,1 – 0,2	gut
2 – 5	0,2 – 0,5	mäßig
1 – 2	0,5 – 1,0	schlecht

In Worten lässt sich somit die Berechnung von Vorhersagewerten ausdrücken durch

$$\text{Posttest-Chance} = \frac{\text{Prätest-Chance} \times \text{Likelihood Ratio}}{\dots}$$

Wichtig ist hier für konkrete Berechnungen die Unterscheidung von Chance und Wahrscheinlichkeit (zumindest für Wahrscheinlichkeiten, die nicht viel kleiner sind als 0,1). Man sieht an dieser Darstellung die Bedeutung sowohl der Effizienz des diagnostischen Tests (dargestellt durch das Likelihood Ratio) als auch der Krankheitsprävalenz für die Vorhersagekraft eines angewendeten Tests. Nur Tests mit sehr guter Effizienz ($\text{PLR} > 10$, $\text{NLR} < 0,1$) haben die Macht, eine gegebene Prätest-Wahrscheinlichkeit deutlich zu verändern. Weiterhin lässt sich ableiten: Geht die Prävalenz der betrachteten Krankheit gegen 0, so geht – auch bei sehr guter Testeffizienz – der positive prädiktive Wert auch gegen 0 (Tabelle 5).

Dieses Gesetz ist insbesondere für Screening-Maßnahmen von Bedeutung, da hier häufig die Krankheitsprävalenz sehr gering ist.

□ Beispiel 1

Anhand der Daten der Göttinger Risiko-, Inzidenz- und Prävalenz-Studie (GRIPS) wurde die Cholesterin-Screening-Strategie der Deutschen Nationalen Cholesterin-Initiative (Nat Cl) zur Prävention des Herzinfarkts untersucht [14]. Bei einer Kohorte von $n = 6029$ Männern zwischen 40 und 60 Jahren wurden unter anderem der Cholesterinspiegel und die Inzidenz des Herzinfarkts innerhalb von fünf Jahren untersucht. Zur Prognose des Herzin-

farkts hatte die Screening-Strategie mit einer Sensitivität von 78,5% (95%-Konfidenzintervall: 69,3 bis 85,6%) und einer Spezifität von 67,0% (65,7 bis 68,3%) eine mäßige Testeffizienz ($\text{PLR} = 2,4$). Die Prävalenz des Herzinfarkts nach fünf Jahren betrug 1,96%. Damit hat die Cholesterin-Screening-Strategie in dieser Population einen positiven prädiktiven Wert von 4,5% (95%-Konfidenzintervall: 3,6 bis 5,6%). Das bedeutet, dass ungefähr nur 5% der Screening-Positiven nach fünf Jahren tatsächlich einen Herzinfarkt erlitten haben.

Maße zur Beschreibung eines Behandlungseffekts

Eine gute Übersicht zur Beurteilung von Studien über Therapie und Prävention geben Guyatt et al. [10, 11]. Ich beschränke mich hier auf die wesentlichsten Eigenschaften von Risiko- und Risikoverhältnismaßen. Die Grundlage dieser Maße zur Beschreibung eines Behandlungseffekts bildet die Vierfeldertafel von Tabelle 6.

Von Interesse sind die bedingten Wahrscheinlichkeiten

$$R_C = \text{Risiko der Unbehandelten („Control“)} = P(\text{Ereignis} \mid \text{Plazebo}) = \frac{b}{b+d}$$

$$R_T = \text{Risiko der Behandelten („Treatment“)} = P(\text{Ereignis} \mid \text{Medikament}) = \frac{a}{a+c}$$

Aus diesen Risiken werden die beiden Maße

$$\text{RR} = \text{relatives Risiko} = \text{Risk Ratio} = \frac{R_C}{R_T}$$

und

Tabelle 6. Vierfeldertafel zur Untersuchung eines Behandlungseffekts.

	Behandlung		Summe	
	Medikament	Plazebo		
Ereignis	ja	a	b	a+b
	nein	c	d	c+d
Summe	a+c	b+d	n	

OR = Chancenverhältnis = Odds Ratio =

$$\frac{\frac{R_C}{1-R_C}}{\frac{R_T}{1-R_T}} = \frac{R_C(1-R_T)}{R_T(1-R_C)}$$

abgeleitet. Ähnlich wie Wahrscheinlichkeit und Chance sind RR und OR äquivalent in dem Sinne, dass beide Maße eine Risikoerhöhung in Form eines multiplikativen Faktors darstellen. Dem Maß RR liegt allerdings die Skala von Wahrscheinlichkeiten, dem OR die Skala von Chancen zugrunde. Daher sind die absoluten Zahlen dieser beiden Maße verschieden, außer wenn die betrachteten Wahrscheinlichkeiten sehr klein ($< 0,1$) oder nahezu identisch sind. Einige Beispiele findet man in Tabelle 7.

Bei höheren Wahrscheinlichkeiten ist zum Beispiel bei einem Studienergebnis von $\text{OR} = 2$ eine Aussage der Art: „Das Risiko der Unbehandelten ist doppelt so hoch wie das der Behandelten“, falsch. Richtig ist in diesem Fall die Aussage: „Die Chance der Unbehandelten ist doppelt so hoch wie die der Behandelten“. Möchte man das Verhältnis der Risiken wissen, so muss man RR berechnen. Um aus einem gegebenen OR das entsprechende RR herzuleiten, benötigt man zusätzlich die Information entweder über R_C oder über R_T . Es gilt

$$\text{RR} = \text{OR} - (\text{OR} - 1) \times R_C$$

An dieser Darstellung sieht man, dass bei einem risikovermindernden Effekt der Behandlung ($\text{RR} > 1$) und bei hohem Basisrisiko ($R_C > 0,1$) RR deutlich kleiner ist als OR.

Das Maß RR ist in der Regel leichter interpretierbar als das OR. Das OR hat aber den Vorteil, dass es auch in Studien schätzbar ist, in denen das RR nicht sinnvoll berechnet werden kann (zum Beispiel retrospektive Fallkontrollstudien). Bei kleinen ($< 0,1$) und bei nahe-

Tabelle 5. Vorhersagewert diagnostischer Tests in Abhängigkeit von der Prävalenz.

Prävalenz	Sehr gute Testeffizienz: Sensitivität = Spezifität = 0,95 (PLR = 19)		Mäßige Testeffizienz: Sensitivität = Spezifität = 0,7 (PLR = 2,3)	
	PPW	NPW	PPW	NPW
90 %	99,4 %	67,9 %	95,5 %	21 %
50 %	95,0 %	95,0 %	70 %	70 %
10 %	67,9 %	99,4 %	21 %	95 %
1 %	16,1 %	> 99,9 %	2,3 %	99,6 %
0,1 %	1,9 %	≈ 100 %	0,2 %	> 99,9 %
0,01 %	< 0,2 %	≈ 100 %	< 0,1 %	≈ 100 %

GRUNDLAGEN MEDIZINISCHER NACHBARGEBIETE

zu identischen Risiken von Kontrolle und Behandlung liefern RR und OR nahezu gleiche Werte, und das OR ist auch als guter Schätzer für RR verwendbar.

Ein weiteres beliebtes Maß ist

$$RRR = \text{relative Risikoreduktion} = \frac{R_C - R_T}{R_C}$$

Das Maß RRR gibt den Anteil der Risikodifferenz bezogen auf das Risiko der Unbehandelten an (häufig in Prozent und ist damit eine Prozentzahl von einer Prozentzahl). Häufig werden in Studien, die einen Behandlungseffekt zeigen wollen, nur RR oder RRR angegeben. Mit diesen relativen Maßen, insbesondere RRR, lassen sich oft eindrucksvolle Zahlen erzeugen, obwohl der absolu-

te Effekt der Behandlung gering ist. RR und RRR sind relative Maße und können demzufolge nicht zwischen absolut hohen und geringen Effekten unterscheiden. Dies wird in Tabelle 8 gezeigt, wo der absolute Effekt durch Absinken des Basisrisikos immer geringer wird, RR und RRR aber konstant hoch bleiben.

Um den absoluten Effekt einer Behandlung zu beschreiben, benötigt man absolute Maße. Das einfachste ist die absolute Differenz zwischen R_C und R_T

$$ARR = \text{absolute Risikoreduktion} = R_C - R_T$$

In der Epidemiologie wird ARR auch als Exzessrisiko oder irrtümlicherweise auch als attributables Risiko bezeichnet.

Ein weiteres absolutes Maß ist die Zahl „Number Needed to Treat“ (NNT), die definiert ist als Kehrwert von ARR [7]

$$NNT = \text{Number Needed to Treat} = \frac{1}{ARR}$$

Das Maß „Number Needed to Treat“ beinhaltet im Prinzip die gleiche Information wie ARR, ist aber besser interpretierbar. NNT ist die geschätzte mittlere Anzahl von Patienten, die behandelt werden muss, um ein zusätzliches (unerwünschtes) Ereignis zu verhindern. Hierbei wird die Höhe des Basisrisikos mit berücksichtigt. Je geringer die Wahrscheinlichkeit ist, dass ein Ereignis eintritt, desto höher ist die Zahl der Patienten, die behandelt werden

Tabelle 7. RR und OR für verschiedene Risikokombinationen. Es wird für jede Kombination von R_C und R_T zuerst RR (oben) und dann OR (unten) angegeben. Beispiel: $R_C = 0,9$, $R_T = 0,5$ ergibt $RR = 1,8$ und $OR = 9$. Nur im grauen Bereich sind RR und OR nahezu gleich (Verhältnis OR/RR zwischen 0,9 und 1,1).

RR OR	R_T													
	0,001	0,01	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,99
R_C	0,001	0,01	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,99
0,001	1 1	0,100 0,099	0,020 0,019	0,010 0,009	0,005 0,004	0,003 0,002	0,003 0,002	0,002 0,001	0,002 7×10^{-4}	0,001 4×10^{-4}	0,001 3×10^{-4}	0,001 1×10^{-4}	0,001 5×10^{-5}	0,001 1×10^{-5}
0,01	10 10,090	1 1	0,2 0,192	0,1 0,091	0,05 0,04	0,033 0,024	0,025 0,015	0,02 0,01	0,017 0,007	0,014 0,004	0,013 0,003	0,011 0,001	0,011 5×10^{-4}	0,01 1×10^{-4}
0,05	50 52,58	5 5,211	1 1	0,5 0,474	0,25 0,211	0,167 0,123	0,125 0,079	0,1 0,053	0,083 0,035	0,071 0,023	0,063 0,013	0,056 0,006	0,053 0,003	0,051 5×10^{-4}
0,1	100 111	10 11	2 2,111	1 1	0,5 0,444	0,333 0,259	0,25 0,167	0,2 0,111	0,167 0,074	0,143 0,048	0,125 0,028	0,111 0,012	0,105 0,006	0,101 0,001
0,2	200 249,8	20 24,75	4 4,75	2 2,25	1 1	0,667 0,583	0,5 0,375	0,4 0,25	0,333 0,167	0,286 0,107	0,25 0,063	0,222 0,028	0,211 0,013	0,202 0,003
0,3	300 428,1	30 42,43	6 8,14	3 3,86	1,5 1,714	1 1	0,75 0,643	0,6 0,429	0,5 0,286	0,429 0,184	0,375 0,107	0,333 0,048	0,316 0,023	0,303 0,004
0,4	400 666	40 66	8 12,7	4 6	2 2,667	1,333 1,556	1 1	0,8 0,667	0,667 0,444	0,571 0,286	0,5 0,167	0,444 0,074	0,421 0,035	0,404 0,007
0,5	500 999	50 99	10 19	5 9	2,5 4	1,667 2,333	1,25 1,5	1 1	0,833 0,667	0,714 0,429	0,625 0,25	0,556 0,111	0,526 0,053	0,505 0,01
0,6	600 1499	12 148,5	60 28,5	6 13,5	3 6	2 3,5	1,5 2,25	1,2 1,5	1 1	0,857 0,643	0,75 0,375	0,667 0,167	0,632 0,079	0,606 0,015
0,7	700 2331	70 231	14 44,33	7 21	3,5 9,333	2,333 5,444	1,75 3,5	1,4 2,333	1,167 1,556	1 1	0,875 0,583	0,778 0,259	0,737 0,123	0,707 0,024
0,8	800 3996	80 396	16 76	8 36	4 16	2,667 9,333	2 6	1,6 4	1,333 2,667	1,143 1,714	1 1	0,889 0,444	0,842 0,211	0,808 0,04
0,9	900 8991	90 891	18 171	9 81	4,5 36	3 21	2,25 13,5	1,8 9	1,5 6	1,286 3,857	1,125 2,25	1 1	0,947 0,474	0,909 0,091
0,95	950 18981	95 1881	19 361	9,5 171	4,75 76	3,167 44,33	2,375 28,5	1,9 19	1,583 12,67	1,375 8,143	1,188 4,75	1,056 2,111	1 1	0,96 0,192
0,99	990 98901	99 9801	19,8 1881	9,9 891	4,95 396	3,3 231	2,475 148,5	1,98 99	1,65 66	1,414 42,43	1,238 24,75	1,1 11	1,042 5,211	1 1

Tabelle 8. RR, RRR, ARR und NNT bei sinkendem Basisrisiko.

R_C	R_T	RR	RRR	ARR	NNT
70 %	20 %	3,5	71,4 %	50 %	2
7 %	2 %	3,5	71,4 %	5 %	20
0,7 %	0,2 %	3,5	71,4 %	0,5 %	200
0,07 %	0,02 %	3,5	71,4 %	0,05 %	2000

muss, um ein Ereignis zu verhindern (Tabelle 8). Da eine Behandlung auch einen schädlichen Effekt haben kann, schlägt Altman [2] zur Angabe der Richtung des Effekts die Bezeichnungen „number needed to treat for one patient to benefit (NNTB) or be harmed (NNTH)“ vor. Methoden zur Berechnung und Präsentation von Konfidenzintervallen für NNTs beschreiben Altman [1] und Bender [7].

□ Beispiel 2a

Im Diabetes Control and Complications Trial (DCCT) wurde unter anderem die gewöhnliche Insulintherapie (Control) mit der intensivierten Insulintherapie (Treatment) bezüglich der Inzidenz von Neuropathie in fünf Jahren verglichen [16]. Es ergaben sich die Risiken $R_C = 16,9\%$ bei gewöhnlicher und $R_T = 6,7\%$ bei intensivierter Insulintherapie für die Entwicklung einer Neuropathie in fünf Jahren. Die relativen Maße belegen einen hohen (relativen) Effekt ($RR = 2,5$, $RRR = 60,4\%$). Der absolute Effekt lässt sich beschreiben durch $ARR = 10,2\%$ und $NNT = 9,8 \approx 10$ (95%-Konfidenzintervall: 6 bis 19). Das bedeutet, dass zehn Diabetiker fünf Jahre mit intensivierter anstelle von gewöhnlicher Insulintherapie behandelt werden müssen, um eine Neuropathie zu verhindern.

□ Beispiel 2b

Bei einer Übersicht über Studien bezüglich milder Hypertonie und Schlaganfall ergaben sich in fünf Jahren $R_C = 1,5\%$ Schlaganfälle unter Placebo und $R_T = 0,9\%$ Schlaganfälle bei antihypertensiver Behandlung [8]. Die relativen Maße belegen ähnlich wie in Beispiel 2a einen hohen relativen Effekt ($RR = 1,7$, $RRR = 40,0\%$). Der absolute Effekt ist allerdings aufgrund

des niedrigen Basisrisikos hier nur gering, denn es gilt: $ARR = 0,6\%$ und $NNT = 166,67 \approx 167$ (95%-Konfidenzintervall: 118 bis 282). Das bedeutet, dass 167 Patienten mit milder Hypertonie fünf Jahre lang antihypertensiv behandelt werden müssen, um einen Schlaganfall zu verhindern.

SCHLUSSFOLGERUNGEN

Zur Beurteilung von Maßnahmen im Rahmen von Screening und Prävention besitzt die Prävalenz der betrachteten Krankheit oder des betrachteten Ereignisses eine fundamentale Bedeutung. Ist die Prävalenz gering, so ist auch der positive prädiktive Wert – selbst bei sehr guter Testeffizienz – gering. Die Mehrheit der Screening-Positiven ist in Wahrheit gesund, wenn ein diagnostischer Test auf ganze Populationen mit geringer Krankheitsprävalenz angewendet wird.

Ebenso ist der absolute Effekt einer Behandlung gering, wenn das Basisrisiko des betrachteten Ereignisses gering ist. Die Angabe relativer Maße wie relatives Risiko oder relative Risikoreduktion ist unzureichend, um einen Behandlungseffekt zu quantifizieren. Erforderlich ist immer auch die Angabe absoluter Maße wie zum Beispiel der Zahl „Number Needed to Treat“, die interpretiert werden kann als die mittlere Anzahl von Patienten, die behandelt werden muss, um ein Ereignis zu vermeiden.

Bei hohem Basisrisiko erhält man für das Odds Ratio extremere Werte als für das Risk Ratio. Wird in solchen Fällen das Odds Ratio fälschlicherweise als Risk Ratio interpretiert, ergibt sich eine Überschätzung des Effekts.

Wird die Höhe der Krankheitsprävalenz bzw. des Basisrisikos nicht genügend berücksichtigt, so werden der Nutzen eines diagnostischen Tests und der Effekt einer Behandlung insbeson-

dere im Rahmen von Screening und Prävention überschätzt. Weitere Irrtümer und Trugschlüsse im Bereich der medizinischen Forschung beschreibt Andersen [5].

Literatur

- Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991.
- Altman DG. Confidence intervals for the number needed to treat. Br Med J 1998;3317:1309–12.
- Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. Br Med J 1994;308:1552.
- Altman DG, Bland JM. Diagnostic tests 2: predictive values. Br Med J 1994;309:102.
- Andersen B. Methodological errors in medical research. London: Blackwell, 1990.
- Bayes ET. An essay toward solving a problem in the doctrine of chance. Phil Trans R Soc 1763;53:370–418. Wiederabdruck in Biometrika 1958;45:213–315.
- Bender R. Calculating confidence intervals for the number needed to treat. Contr Clin Trials 2001;22:in press.
- Collins R, Peto R, MacMahon x, et al. Blood pressure, stroke and coronary heart disease. II. Short-term reductions in blood pressure: overview of randomised drug trials in their epidemiologic context. Lancet 1990; 335:827–38.
- Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. Br Med J 1995;310:452–54. Correction: Br Med J 1995;310:1056.
- Guyatt G, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users'guides to the medical literature II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA 1993;270:2598–601.
- Guyatt G, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users'guides to the medical literature II. How to use an article about therapy or prevention. B. What are the results and will they help me in caring for my patients? JAMA 1994;271:59–63.
- Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group. Users'guides to the medical literature III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–91.
- Jaeschke R, Guyatt G, Sackett DL, for the Evidence-Based Medicine Working Group. Users'guides to the medical literature III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA 1994;271:703–7.
- Muche R, Gefeller O, Nagel D, Cremer P. Cholesterol screening: an appropriate tool for prevention of CHD? In: König W, Hombach V, Bond MG, eds. Current aspects in atherosclerosis. Berlin: Blackwell, 1995:445–8.
- Richter K, Lange S. Methoden der Diagnoseevaluierung. Internist (Berlin) 1997;38:325–36.
- The DCCT Research Group. The effect of intensive diabetes therapy on the development and progression of neuropathy. Ann Intern Med 1995;122:561–8.
- Vollset SE. Confidence intervals for a binomial proportion. Stat Med 1993;12:809–24.

Korrespondenzanschrift

Dr. Ralf Bender
AG 3 – Epidemiologie und Medizinische Statistik
Fakultät für Gesundheitswissenschaften
Universität Bielefeld
Postfach 100131
D-33501 Bielefeld
Telefon (+49/521) 106-3803,
Fax -6465,
E-Mail: Ralf.Bender@uni-bielefeld.de