ÜBERSICHTSARBEIT

Interpretation der Ergebnisse von 2×2-Tafeln

Teil 9 der Serie zur Bewertung wissenschaftlicher Publikationen

Wilhelm Sauerbrei, Maria Blettner

ZUSAMMENFASSUNG

Hintergrund: Ergebnisse von epidemiologischen Studien, diagnostischen Testverfahren und Therapievergleichen werden häufig in einer 2×2-Tafel dargestellt. Die richtige Interpretation der 2×2-Tafel ist Voraussetzung für das Verständnis der Ergebnisse solcher Studien.

Methoden: Darstellung grundlegender statistischer Zusammenhänge für die Analyse nominaler Daten unter Bezugnahme von Standardwerken der Statistik.

Ergebnisse: Relatives Risiko und Odds Ratio werden als Maßzahlen für den Zusammenhang von zwei binären Größen (zum Beispiel Exposition ja/nein, Erkrankung ja/nein) definiert. Untersucht wird der Einfluss des Stichprobenumfangs auf die Breite des Konfidenzintervalls und den p-Wert sowie Verzerrungen, die durch Messfehler entstehen. Häufig wird eine Exposition in drei Stufen (keine, niedrig, hoch) gemessen. Als Erweiterung betrachten die Autoren die 2×3-Tabelle und diskutieren die Kategorisierung stetiger Größen. Bei der Entwicklung einer Erkrankung ist typischerweise mehr als ein Faktor beteiligt. Es wird erläutert, welchen Einfluss ein weiterer Faktor auf den beobachteten Zusammenhang zwischen der Exposition und der Erkrankung haben kann.

Schlussfolgerungen: Umfang der Stichprobe, Messfehler, Kategorisierung und das Vorliegen von Störgrößen beeinflussen auf vielfältige Weise die Aussagekraft einer 2×2-Tafel. Der Leser einer wissenschaftlichen Publikation sollte die Probleme bei der Interpretation einer einfachen 2×2-Tafel kennen und darauf achten, ob die Autoren diese bei der Analyse und Interpretation hinreichend berücksichtigt haben.

Schlüsselwörter: Publikation, klinische Forschung, Epidemiologie, Statistik, Studie

Zitierweise: Dtsch Arztebl Int 2009; 106(48): 795–800
DOI: 10.3238/arztebl.2009.0795

Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg: Prof. Dr. rer. nat. Sauerbrei

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universitätsklinikum Mainz: Prof. Dr. rer. nat. Blettner

Ergebnisse von epidemiologischen Studien, diagnostischen Testverfahren und Therapievergleichen stellt man häufig in einer 2×2-Tafel dar. Häufig werden auch die Begriffe Vierfeldertafel, Kontingenztafel oder Kreuztabelle verwendet. So wurde kürzlich im British Medical Journal eine Fall-Kontrollstudie vorgestellt (1), in der die Assoziation zwischen Teekonsum und Ösophaguskarzinom untersucht wurde. Von 300 erkrankten Personen („Fällen“) gaben 249 an, dass sie nie oder sehr selten grünen Tee trinken, 17 (6,4 %) trinken häufig grünen Tee. Bei den 571 nicht erkrankten Personen („Kontrollen“) gaben 356 Personen einen seltenen Konsum und 30 Personen einen häufigen Konsum an. Diese Daten wurden in einer 2×2-Tabelle dargestellt (1). Es fällt auf, dass ein Teil der Personen keine Angaben machte (fehlende Daten). Als weiteres Beispiel betrachten wir eine klinische Studie, in der bei Patientinnen mit metastasiertem Brustkrebs unter Anderem der Einfluss der vorhergehenden Therapie untersucht wurde (2). Alle Patientinnen hatten Taxane erhalten. Bei 10 (28,6 %) von 35 Patientinnen mit vorheriger Anthracyclingabe stellte man eine Progression fest. Eine höhere Rate an Progressionen (15 von 26; 57,7 %) gab es in der Gruppe ohne vorherige Gabe von Anthracyclinen. In der Fall-Kontrollstudie könnte eine andere Aufteilung des Teekonsums, zum Beispiel in drei Gruppen (nie, mäßig, häufig) gewählt werden. In der Therapiestudie wird das Ansprechen des Tumors auch häufig in den Stufen „Komplettremission“, „partielle Remission“, „keine Änderung und Progression“ gemessen.

Aus der einfachen 2×2-Tabelle lassen sich das Relative Risiko (RR) oder das Odds Ratio (OR) berechnen. Deshalb ist es wichtig, die zentralen Eigenschaften der 2×2-Tabelle zu verstehen und zu wissen, wie schon einfache Erweiterungen die Analyse und Interpretation verändern können. Wird dies nicht beachtet, so kann es zu falschen Schlussfolgerungen kommen, die für den Patienten unter Umständen eine Fehleinschätzung des Risikos, der Diagnose, der Prognose oder der geeigneten Therapie bedeuten.

Bei der Entwicklung einer Erkrankung ist typischerweise mehr als ein Faktor beteiligt. Daher muss man in der Analyse in den meisten Situationen auch mehr als einen potenziellen Einflussfaktor betrachten. Zum Beispiel sind neben der Art und der Temperatur des Tees auch Kaffee und Alkoholkonsum zu berücksichtigen. Einfache Kontingenztafeln sind dann für die Analysen

KASTEN

Zusammenhangsmaße in einer 2×2-Tafel

Schema und Notation einer grundlegenden 2×2-Tafel

		Krankheit vorhanden ja (D +)	nein (D –)	
exponiert	ja	a	b	a + b
	nein	c	d	c + d
		a + c	b + d	n

Definition

- Risiko beschreibt die Wahrscheinlichkeit zu erkranken
- P_0 = Wahrscheinlichkeit zu erkranken für eine Person, die nicht exponiert ist
- P_1 gilt entsprechend für die Gruppe der exponierten Personen
- $P_0 = c / (c + d)$
- $P_1 = a / (a + b)$

Risikodifferenz: $RD = P_1 – P_0$
Relatives Risiko, Risiko-Quotient: $RR = P_1 / P_0$

O_0 = Odds (Chancenverhältnis) für die Nichtexponierten: $O_0 = P_0 / (1 – P_0)$.
O_1 = Odds für die Exponierten: $O_1 = P_1 / (1 – P_1)$
OR = odds ratio = $O_1 / O_0 = (P_1 / [1 – P_1]) / (P_0 / [1 – P_0]) = (a \times d) / (b \times c)$
Falls (a + c)/n „klein“, gilt in etwa RR = OR.

und Darstellung der Ergebnisse nicht mehr ausreichend. Notwendig sind Auswertungen mit multivariablen Modellen, in denen mehrere Variablen gleichzeitig berücksichtigt werden.

Im Folgenden benutzen die Autoren die Notation der 2×2-Tafel *(Kasten)* und diskutieren die Ergebnisse einer hypothetischen Studie *(Tabelle 1)*. Sie definieren die Begriffe Risiko, Relatives Risiko und Odds Ratio und diskutieren den Einfluss des Stichprobenumfangs auf die Breite des Konfidenzintervalls und den p-Wert. Weiterhin erläutern sie die durch Messfehler entstehende Verzerrung des Ergebnisses. Als einfache Erweiterung betrachten sie dann eine 2×3-Tafel und erklären, welchen Einfluss ein weiterer Faktor auf den beobachteten Zusammenhang zwischen Exposition und Erkrankung haben kann. Dazu benutzen die Autoren Bezeichnungen aus der Epidemiologie, das heißt sie sprechen von Risikofaktoren. Die erläuterten Aspekte gelten in gleicher Weise für die entsprechenden Maße der Diagnose, Prognose und Therapie. Im Rahmen von Prognosestudien sind weitere Aspekte in Sauerbrei & Schumacher (1999) (3) diskutiert. Als weiterführende Literatur empfehlen die Verfasser Fletcher et al. (2005) (4), Altman (1991) (5), Campbell et al. (2007) (6) und Schumacher & Schulgen (2008) (7).

Definitionen

Die Autoren betrachten eine Gruppe von n Personen und interessieren sich für zwei Eigenschaften

- Ist die Person exponiert oder nicht exponiert?
- Ist die Person erkrankt oder nicht erkrankt?

Das Wort „exponiert“ wird hier stellvertretend für verschiedene Charakteristiken, zum Beispiel Personen, die einer bestimmten beruflichen Belastung ausgesetzt sind, Personen, die eine bestimmte genetische Konstellation haben oder die für bestimmte Laborparameter Werte außerhalb der Normbereiche aufweisen, verwendet. Im vorliegenden Beispiel sind es die Personen, die „häufig“ grünen Tee trinken. In Therapiestudien kann man „exponiert“ durch „Therapie A“, „nicht exponiert“ durch „Therapie B“, „erkrankt“ durch „kein Therapieerfolg“ und „nicht erkrankt“ durch „Therapieerfolg“ ersetzen.

In *Tabelle 1* betrachten die Autoren eine Kohortenstudie mit 450 Personen, von denen 36 erkrankt und 414 nicht erkrankt sind; zwei Drittel der Personen (300) sind exponiert, während ein Drittel nicht exponiert ist. Die Inzidenzrate beträgt 8 % (36 von 450) in der Gesamtgruppe, 10 % (30 von 300) bei den Personen, die exponiert sind und 4 % (6 von 150) bei den nicht exponierten Personen.

Aus der 2×2-Tabelle werden Inzidenzrate, Risiko, Relatives Risiko und Odds Ratio hergeleitet *(Kasten)*.

Ein Relatives Risiko von 1 (RR = 1) bedeutet, dass exponierte (Therapie A) und nicht exponierte (Therapie B) Personen das gleiche Risiko haben, zu erkranken („geheilt zu werden“). RR größer als 1 bedeutet, dass die exponierten Personen ein höheres Risiko haben als die anderen. RR = 1,5 bedeutet, dass das Risiko der exponierten Personen um 50 % höher ist als das Risiko der nicht exponierten Personen. RR = 2 bedeutet, dass sich das Risiko verdoppelt (andere Sprechweisen: das Risiko ist um 100 % erhöht; das Risiko ist auf 200 % angestiegen). RR = 0,5 bedeutet, dass Personen der exponierten Gruppe nur ein halb so großes Risiko haben wie Personen der nicht exponierten Gruppe. Man spricht dann auch von einem protektiven Faktor. Wichtig ist es, zu beachten, welche Gruppe als Basisgruppe zugrunde gelegt wird. Sei RR = 1,5 (zum Beispiel Raucher versus Nichtraucher) dann bedeutet dies, das Raucher ein um 50 % erhöhtes Risiko haben. Nimmt man die Raucher als Basisgruppe, so ergibt sich RR = 1/1,5 = 0,67. Im Vergleich zu Rauchern ist das Risiko von Nichtrauchern also um ein Drittel (1 – 0,67 = 0,33) reduziert.

Neben dem Relativen Risiko wird häufig das sogenannte Odds Ratio (OR, Chancenverhältnis oder Quotenverhältnis) als Maß für den Zusammenhang angegeben *(Kasten)*. Das Odds Ratio ist der Quotient der Chancen (Odds) einer Erkrankung (Heilung) der Personen ohne und der Personen mit Exposition (Therapie).

In Fall-Kontrollstudien kann das Relative Risiko nicht direkt berechnet werden. Das liegt daran, dass das Verhältnis von Fällen zu Kontrollen im Design festgelegt wird, also (a + c)/n vom Untersucher bestimmt wird. Daher ist weder a/(a + b) noch c/(c + d) eine sinn-

volle Kennzahl (also keine Inzidenzrate) und eine Berechnung des Relativen Risikos nicht möglich. Das Odds Ratio kann als eine Hilfskonstruktion für das Relative Risiko betrachtet werden. Odds Ratio und Relatives Risiko sind numerisch etwas gleich groß, wenn die Erkrankungswahrscheinlichkeiten (P_1 und P_0) beide klein sind. Ein Wert von etwa 1 % bis 5 % kann für diese Berechnungen noch als „klein" betrachtet werden. Zu beachten ist, dass Odds Ratio und Relatives Risiko nur in diesen Fällen etwa gleich groß sind, wobei bei einem Relativen Risiko größer 1 das Odds Ratio immer geringfügig größer als das Relative Risiko ist. In unserem Beispiel ist RR = 2,50 und OR = 2,67.

TABELLE 1

Darstellung der Egebnisse einer hypothetischen Studie in einer 2×2-Tafel

gesamt (% Spalte)	(% Zeile)	Krankheit vorhanden ja (D +)	nein (D –)	
exponiert	ja	30 (10,0 %) (83,3 %)	270 (90,0 %) (65,2 %)	300 (66,7 %)
	nein	6 (4,0 %) (16,7 %)	144 (96,0 %) (34,8 %)	150 (33,3 %)
		36 (8,0 %)	414 (92,0 %)	450

Problem 1: Stichprobengröße, Konfidenzintervall und p-Wert

Neben dem Relativen Risiko werden in vielen Publikationen Konfidenzintervall und p-Werte als zusammenfassende Beschreibung der Assoziation von zwei Faktoren angegeben. Ein p-Wert, der kleiner ist als eine "magische" Grenze, häufig 5 %, wird als statistisch signifikant bezeichnet. Konfidenzintervall und p-Wert sind (bei festem RR) vom Stichprobenumfang abhängig *(Tabelle 2)*. Im Autorenbeispiel ist das geschätzte RR = 2,5 mit 95-%-Konfidenzintervall (1,06–5,87). Der Test, ob es einen Zusammenhang zwischen den beiden Faktoren gibt (Chi-Quadrat-Test für Unabhängigkeit), ergibt einen p-Wert von p = 0,027. Verdopplung (n = 900) beziehungsweise Halbierung des Stichprobenumfangs (n = 225) verändert nicht den Schätzer, aber den p-Wert und das Konfidenzintervall. Der p-Wert wird kleiner (p = 0,002) beziehungsweise größer (p = 0,1118), das Konfidenzintervall schmäler beziehungsweise breiter. Bei Halbierung des Stichprobenumfanges ist der Wert „1,00" im Konfidenzintervall enthalten, der Effekt der Exposition wird folglich als statistisch nicht signifikant bezeichnet. Bei der Interpretation des p-Wertes sind der Schätzer des Relativen Risikos, die Stichprobengröße und das Konfidenzintervall gemeinsam zu betrachten.

Problem 2: Auswirkung des Messfehlers auf das relative Risiko

Es soll verdeutlicht werden, wie ein Fehler bei der Klassifikation von exponierten und nicht exponierten Personen das Ergebnis der 2×2-Tabelle beeinflussen kann *(Tabelle 3)*. Geht man davon aus, dass (nur) 10 % aller Personen falsch klassifiziert werden, so werden im Durchschnitt 30 exponierte Fälle fälschlicherweise als nicht exponierte Personen eingeordnet. Weiterhin werden etwa 15 nicht exponierte Fälle fälschlicherweise der exponierten Gruppe zugeordnet. Wir gehen davon aus, dass dieser Fehler nicht von dem Krankheitsstatus abhängt (nicht differenzielle Missklassifikation). In diesem Fall werden drei exponierte Fälle (10 % der 30 Fehlklassifikationen) und ein nicht exponierter Fall (4 % von 15 Fällen, ergibt 0,6 Fälle, gerundet 1 Fall) jeweils falsch klassifiziert. Wegen dieser Messfehler erhalten wir die Daten von *Tabelle 3* mit RR = 2,03 (95-%-KI = 0,95–4,34, p-Wert = 0,061). Das Ergebnis ist also „nicht signifikant". Eine Missklassifikation, die für Fälle und Kontrollen gleich groß ist (nicht differenzielle Missklassifikation) führt dazu, dass – wie man zeigen kann – das RR unterschätzt wird. Die Annahme einer nicht differenziellen Missklassifikation ist jedoch selten gerechtfertigt. Daher ist in jedem Fall eine detaillierte Untersuchung der Auswirkung potenzieller Messfehler notwendig.

Problem 3: Exposition in mehr als zwei Stufen

In der oben genannten Fall-Kontrollstudie wurde die Häufigkeit des Teekonsums in drei Gruppen (nie, mäßig, häufig) gemessen (1). Die Autoren erweitern unsere Kontingenztafel zu einer 2×3-Tafel mit der Exposition in drei Kategorien *(Tabelle 4)*. Das Relative Risiko der „hoch" exponierten gegenüber den nicht exponierten Personen beträgt 2,8, das der niedrig exponierten gegenüber den nicht exponierten Personen dagegen RR = 2,0. Der Einfluss der Exposition auf die Erkrankung ist ausgeprägter als in *Tabelle 1*. Wendet man aber (fälschlicherweise) einen 2-χ^2-Test für eine 2×3-Tafel

TABELLE 2

Einfluss des Stichprobenumfangs N auf die Breite des Konfidenzintervalls und den p-Wert

N	RR	95- %-KI	p-Wert	Interpretation
225	2,5	0,75–8,37	0,118	nicht signifikant
450	2,5	1,06–5,87	0,027	signifikant
900	2,5	1,37–4,57	0,002	hoch signifikant

KI, Konfidenzintervall

TABELLE 3

Einfluss eines 10 %-igen Klassifikationsfehlers (nicht differenzielle Missklassifikation) auf die 2×2-Tafel aus Tabelle 1

	D +	D –	
E +	28 (9,8 %)	257 (90,2 %)	285 (63,3 %)
E –	8 (4,8 %)	157 (95,2 %)	165 (36,7 %)
	36 (8,0 %)	414 (92,0 %)	450

TABELLE 4

Erhebung der Exposition in drei Kategorien

a) ordinaler Faktor und resultierende 2×3-Tafel

	D +	D –	
E + hoch	22 (11,0 %)	178 (89,0 %)	200 (44,4 %)
E + niedrig	8 (8,0 %)	92 (92,0 %)	100 (22,2 %)
E –	6 (4,0 %)	144 (96,0 %)	150 (33,3 %)
	36 (8,0 %)	414 (92,0 %)	450

b) Ergebnisse verschiedener Analysen

Zusammenfassung von E +	RR	95- %-KI	p-Wert
ja (siehe Tabelle 1 und 2)	2,5	1,06–5,87	0,027
nein			0,058[a] 0,020[b]
E + niedrig vs. E – E + hoch vs. E –	2,0 2,8	0,72–5,59 1,14–6,61	
E + hoch vs. E + niedrig	1,4	0,64–2,98	

[a] χ^2 Unabhängigkeitstest in 2×3-Tabelle
[b] Test für Trend mit Score-Werten (0 – E –, 1 – E + niedrig, 2 – E + hoch)
KI, Konfidenzintervall

an, so ist der Zusammenhang jetzt nicht mehr signifikant. In der 2×2-Tafel hat der $2\chi^2$-Unabhängigkeitstest nur einen Freiheitsgrad, in der 2×3-Tafel dagegen zwei Freiheitsgrade. Für ein gegebenes Signifikanzniveau ist der kritische Wert in der 2×3-Tafel größer. Geht man so vor, wird aber nicht beachtet, dass die drei Expositionsstufen (nicht vorhanden, niedrig, hoch) geordnet sind. Ein geeigneter Trend-Test sollte aber die Rangfolge berücksichtigen. Wichtig ist, dass die Aufteilung der Kategorien vor der Auswertung mit inhaltlichen und biometrischen Argumenten festgelegt werden sollte. Ein nachträgliches „Suchen" nach geeigneter Kategorisierung, um einen kleineren p-Wert (also ein „signifikantes" Ergebnis) zu erhalten, ist wegen der Problematik multipler Tests strikt abzulehnen (8).

Problem 4: Kategorisierung stetiger Variablen

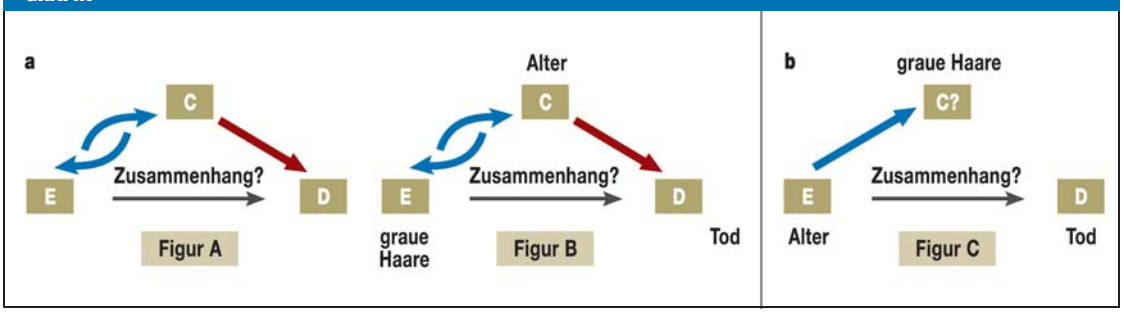
Obwohl die Exposition häufig als stetige Variable gemessen wird, also eine Variable mit vielen möglichen Ausprägungen (zum Beispiel Blutdruck), basiert die Auswertung oft auf kategorisierten Daten (hoch, mittel, niedrig). Die Problematik der Kategorisierung einer stetigen Variablen durch die Festlegung von Klassengrenzen hat viele Nachteile. Zunächst wird ein Teil der ursprünglich erhobenen detaillierten Information nicht verwendet. Bei einer geringen Anzahl von Kategorien (zum Beispiel ein Grenzwert für die Einstufung „hoch" beziehungsweise „niedrig") ist dieser Verlust am größten. Außerdem sind die Anzahl geeigneter Kategorien und die Grenzen dieser Kategorien festzulegen. Falls Kategorien mit zu kleiner Fallzahl gewählt werden, ist der Schätzer für den Effekt der entsprechenden Kategorien instabil. Die Berücksichtigung der Zielgröße bei der Festlegung der Grenzen führt zu einer starken Überschätzung des Effektes und zu falschen p-Werten. Für den häufig benutzen „optimalen" Cutpoint-Ansatz, bei dem datenabhängig alle möglichen Grenzwerte zur Kategorisierung in „hoch" beziehungsweise „niedrig" untersucht werden, verdeutlichen Altman et al. (1994) (9) vielfältige Probleme. Die Betrachtung vieler möglicher Cutpoints führt zu einem stark erhöhten Fehler erster Art, das heißt es kommt zu einem signifikanten Ergebnis, obwohl in der Realität kein Einfluss des untersuchten Faktors auf die Zielgröße vorhanden ist. Anstatt einer angenommenen Irrtumswahrscheinlichkeit von 5 % führt die mehrfache Anwendung von Tests zu einer Irrtumswahrscheinlichkeit von nahezu 50 % (9).

Bei stetigem Risikofaktor sollte man statt der Kategorisierung eine Dosis-Wirkungs-Beziehung schätzen (10).

Problem 5: Einfluss eines dritten Faktors

Confounding und Simpson's Paradoxon – In vielen Studien wird der Einfluss der Exposition auf die Erkrankung von einem weiteren Faktor beeinflusst *(Grafik)*. Wir nehmen an, die in *Tabelle 1* angegebenen Ergebnisse wurden in einer Gruppe (Nichtraucher) erhoben. Es liegen aber auch Daten einer zweiten Gruppe (Raucher) vor *(Tabelle 5)*. In beiden Gruppen ist das

GRAFIK



Darstellung des Einflusses eines potenziellen Confounders; untersucht man den Zusammenhang von grauem Haar und Tod, stellt sich das Alter als logischer Confounder dar; wäre ein anderer Forscher an dem Zusammenhang zwischen Alter und Tod interessiert, würde er sicher über den Einfluss von grauem Haar auf diesen Zusammenhang staunen

RR größer als 1, bei Vorliegen der Exposition tritt die Erkrankung also häufiger auf.

Werden diese Gruppen bei der Auswertung nicht berücksichtigt, das heißt die Zahlen der beiden Gruppen addiert, so erhält man einen Schätzer für das Relative Risiko, der kleiner als 1 ist. Dieses Phänomen ist als Simpson's Paradoxon bekannt. Das liegt daran, dass die Verteilung der Exposition in den beiden Gruppen und das Erkrankungsrisiko für Nichtraucher und Raucher unterschiedlich sind. Eine Addition der Tabellen ist daher nicht gerechtfertigt. Bei der Auswertung von Studien mit mehreren potenziellen Einflussfaktoren ist jedoch oft nicht ersichtlich, welche Faktoren bei der Analyse berücksichtigt werden müssen. Dabei spielt die Assoziation (oder die Korrelation) zwischen den verschiedenen Einflussfaktoren eine Rolle. Als Confounder werden Variablen bezeichnet, die sowohl mit der Exposition als auch mit der Erkrankung korreliert sind. Die klassische Art mit kategoriellen Confoundern umzugehen, ist der sogenannte Mantel-Haenszel-Schätzer, der auf einer Stratifizierung der Daten nach der Confoundervariablen beruht. Der Mantel-Haenszel-Schätzer ist ein gewichteter Mittelwert aus den Odds Ratios der einzelnen Kategorien, wobei die Gewichte von deren Größe abhängen. Es ist offensichtlich, dass beim Vorliegen mehrerer Confounder diese Vorgehensweise sehr komplex werden kann. Insbesondere kann es dazu führen, dass einige Kategorien nur mit wenigen Fällen und Kontrollen besetzt sind. Dann ist eine Modellierung im Rahmen von Regressionsmodellen erforderlich. Für binäre Zielgrößen hat sich in der Medizin das logistische Regressionsmodell als Standard bei der simultanen Untersuchung mehrerer Faktoren etabliert.

Interaktion – Bisher sind die Autoren davon ausgegangen, dass der zusätzliche Faktor (hier: Rauchen) nicht von primärem Interesse ist. Er hat lediglich Einfluss auf die Beziehung zwischen der Exposition und der Erkrankung und muss bei der Auswertung berücksichtigt werden. Häufig ist man jedoch außer an dem Effekt der einzelnen Faktoren auch an ihrem gemeinsamen Einfluss auf die Krankheit interessiert. In der medizinischen Forschung geht man oft davon aus, dass die Faktoren multiplikativ wirken. Das heißt, beim gleichzeitigen Vorhandensein von zwei oder mehr Faktoren errechnet sich das Relative Risiko als Produkt der einzelnen Relativen Risiken. Weicht das Ergebnis der Untersuchung beträchtlich von dem errechneten Produkt ab, gibt es eine Interaktion (Wechselwirkung) zwischen den Faktoren. In realen Studien werden immer (kleinere) Abweichungen beobachtet. Ein Test auf Interaktion kann untersuchen, ob diese Abweichungen zufällig oder statistisch auffällig (signifikant) sind.

Im vorliegenden Beispiel sei die Variable „Rauchen" als zweiter Faktor ebenfalls von Interesse. Für Raucher erhält man RR = (60/150)/(36/450) = 5,0 *(Tabelle 6a)*, falls die Exposition E nicht berücksichtigt wird. Anderseits zeigt *Tabelle 5b*, dass das RR = 0,83 für die Exposition E beträgt, falls der Raucherstatus nicht berücksichtigt wird. Ein multiplikativer Effekt bedeutet, dass Personen, die sowohl rauchen als auch gegenüber E exponiert sind, im Vergleich zu Nichtrauchern ohne Exposition E ein um 0,83 × 5,0 = 4,15 erhöhtes Risiko haben.

TABELLE 5

Einfluss eines zusätzlichen Faktors

a) Einfluss von E in 2 Strata

Stratum I (Nichtraucher); RR = 2,5

	D +	D –	
E +	30 (10,0 %)	270 (90,0 %)	300 (66,7 %)
E –	6 (4,0 %)	144 (96,0 %)	150 (33,3 %)
	36 (8,0 %)	414 (92,0 %)	450

Stratum II (Raucher); RR = 1,38

	D +	D –	
E +	20 (50,0 %)	20 (50,0 %)	40 (26,7 %)
E –	40 (36,0 %)	70 (63,6 %)	110 (73,3 %)
	60 (40,0 %)	90 (66,0 %)	150

b) Einfluss von E

ohne Berücksichtigung der Strata; RR = 0,83

	D +	D –	
E +	50 (14,7 %)	290 (85,3 %)	340 (56,7 %)
E –	46 (17,7 %)	214 (82,3 %)	260 (43,3 %)
	96 (16,0 %)	504 (84,0 %)	600

Aus *Tabelle 6b* geht aber hervor, dass das Risiko tatsächlich um den Faktor (20/40)/(6/150) = 12,5 erhöht ist. Es handelt sich hier um eine Abweichung von der Annahme der Multiplikativität, also um eine Interaktion zwischen den beiden Faktoren. Weitere Publikationen erörtern eine weiterführende Diskussion und erläutern geeignete Methoden zur Untersuchung von Interaktionen (11 – 13).

TABELLE 6

Relatives Risiko für Raucher und Interaktionen zwischen E und Rauchen (hergeleitet aus Tabelle 5a)

a) Einfluss des Rauchens; keine Berücksichtigung der Exposition E; RR = 5,00

	D +	D –	
Nichtraucher	36 (8,0 %)	414 (92,0 %)	450 (75,0 %)
Raucher	60 (40,0 %)	90 (60,0 %)	150 (25,0 %)
	96 (16,0 %)	504 (84,0 %)	600

b) gemeinsamer Einfluss von Rauchen und der Exposition E; RR = 12,5; die Kombinationen (Nichtraucher, E +) und (Raucher, E –) sind nicht angegeben.

	D +	D –	
(Nichtraucher, E –)	6 (4,0 %)	144 (96,0 %)	150
(Raucher, E +)	20 (50,0 %)	20 (50,0 %)	40
	26 (13,7 %)	164 (86,3 %)	190

Diskussion

Jede Publikation einer klinischen oder epidemiologischen Studie sollte eine einfache deskriptive Darstellung der Ergebnisse enthalten (14). In vielen Fällen ist bereits die 2×2-Tafel eine übersichtliche Methode, die Hauptergebnisse darzustellen. Allerdings ist die Interpretation der scheinbar einfachen Tafel mit einigen Tücken verbunden. Der Leser einer wissenschaftlichen Publikation sollte diese kennen und darauf achten, ob die Autoren hinreichend auf mögliche Probleme hingewiesen haben.

Interessenkonflikt
Die Autoren erklären, dass kein Interessenkonflikt im Sinne der Richtlinien des International Committee of Medical Journal Editors besteht.

Manuskriptdaten
eingereicht: 25. 3. 2009, revidierte Fassung angenommen: 16. 6. 2009

LITERATUR

1. Islami F, Pourshams A, Nasrollahzadeh D, et al.: Tea drinking habits and oesophageal cancer in a high risk area in northern Iran: population based case-control study. BMJ 2009; 338: b929. Doi: 10.1136/bmj.b929
2. Andreetta C, Puppin, C, Minisini A, et al.: Thymidine phosphorylase expression and benefit from capecitabine in patients with advanced breast cancer. Annals of Oncology 2009; 20: 265–71.
3. Sauerbrei W, Schumacher M: Aspekte der statistischen Evaluation neuer Prognosefaktoren: Illustration bei Studien in der Onkologie. Geburtshilfe und Frauenheilkunde 1999; 59: 483–7.
4. Fletcher RH, Fletcher SW: Klinische Epidemiologie. 2. Aufl. Bern: Huber 2007.
5. Altman DG: Practical statistics for medical research. London: Chapman and Hall 1991.
6. Campbell MJ, Machin D, Walters SJ: Medical statistics—a textbook for the health sciences. 4. Aufl. Chichester: Wiley 2007.
7. Schumacher M, Schulgen G: Methodik klinischer Studien – Methodische Grundlagen der Planung, Durchführung und Auswertung. 3. Aufl. Berlin: Springer 2008.
8. Victor A, Elsäßer A, Hommel G, Blettner M: Wie bewertet man die p-Wert-Flut – Hinweise zum Umgang mit dem multiplen Testen. 2009 (Deutsches Ärzteblatt, in press)
9. Altman DG, Lausen B, Sauerbrei W, Schumacher M: Dangers of using „optimal" cutpoints in the evaluation of prognostic factors. J Net Cancer Inst 1994; 86: 829–35.
10. Royston P, Sauerbrei W: Multivariable model-building—a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: Wiley 2008.
11. Altman DG, Matthews JNS: Interaction 1: heterogeneity of effects. BMJ 1996; 313: 486.
12. Assmann SF, Pocock SJ, Enos LE, Kasten LE: Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064–9.
13. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM: Statistics in medicine—reporting of subgroup analyses in clinical trials. The New England Journal of Medicine 2007; 357: 2189–94.
14. Spriestersbach A, Gerhold-Ay A, du Prel JB, Röhrig B, Blettner M: Deskriptive Statistik. Dtsch Arztebl Int 2009; 106(36): 578–83.

Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie, Epidemiologie und Informatik (IMBEI)
Klinikum der Universität Mainz, 55101 Mainz
E-Mail: blettner@imbei.uni-mainz.de

SUMMARY

Interpreting Results in 2×2 Tables: Extensions and Problems—Part 9 of a Series on Evaluation of Scientific Publications

Background: The findings of epidemiological studies, diagnostic tests, and comparative therapeutic trials are often presented in 2×2 tables. These must be interpreted correctly for a proper understanding of the findings.

Methods: The authors present basic statistical concepts required for the analysis of nominal data, referring to standard works in statistics.

Results: The relative risk and odds ratio are defined to be indices for the relationship between two binary quantities (e.g., exposure – yes/no and disease – yes/no). The topics dealt with in this article include the effect of sample size on the width of the confidence interval and the p-value, and also inaccuracies caused by measuring error. Exposures are often expressed on a three-level scale (none, low, high). The authors also consider the 2×3 table as an extension of the 2×2 table and discuss the categorization of continuous quantities. Typically, more than one factor is involved in the development of a disease. The effect that a further factor can have on the observed relationship between the exposure and the disease is discussed.

Conclusions: Sample size, measurement error, categorization, and confounders influence the statistical interpretation of 2×2 tables in many ways. Readers of scientific publications should know the inherent problems in the interpretation of simple 2×2 tables and check that the authors have taken these into account adequately in analyzing and interpreting their data.

Key words: publications, clinical research, epidemiology, statistics, clinical trial

Zitierweise: Dtsch Arztebl Int 2009; 106(48): 795–800
DOI: 10.3238/arztebl.2009.0795

The English version of this article is available online:
www.aerzteblatt-international.de