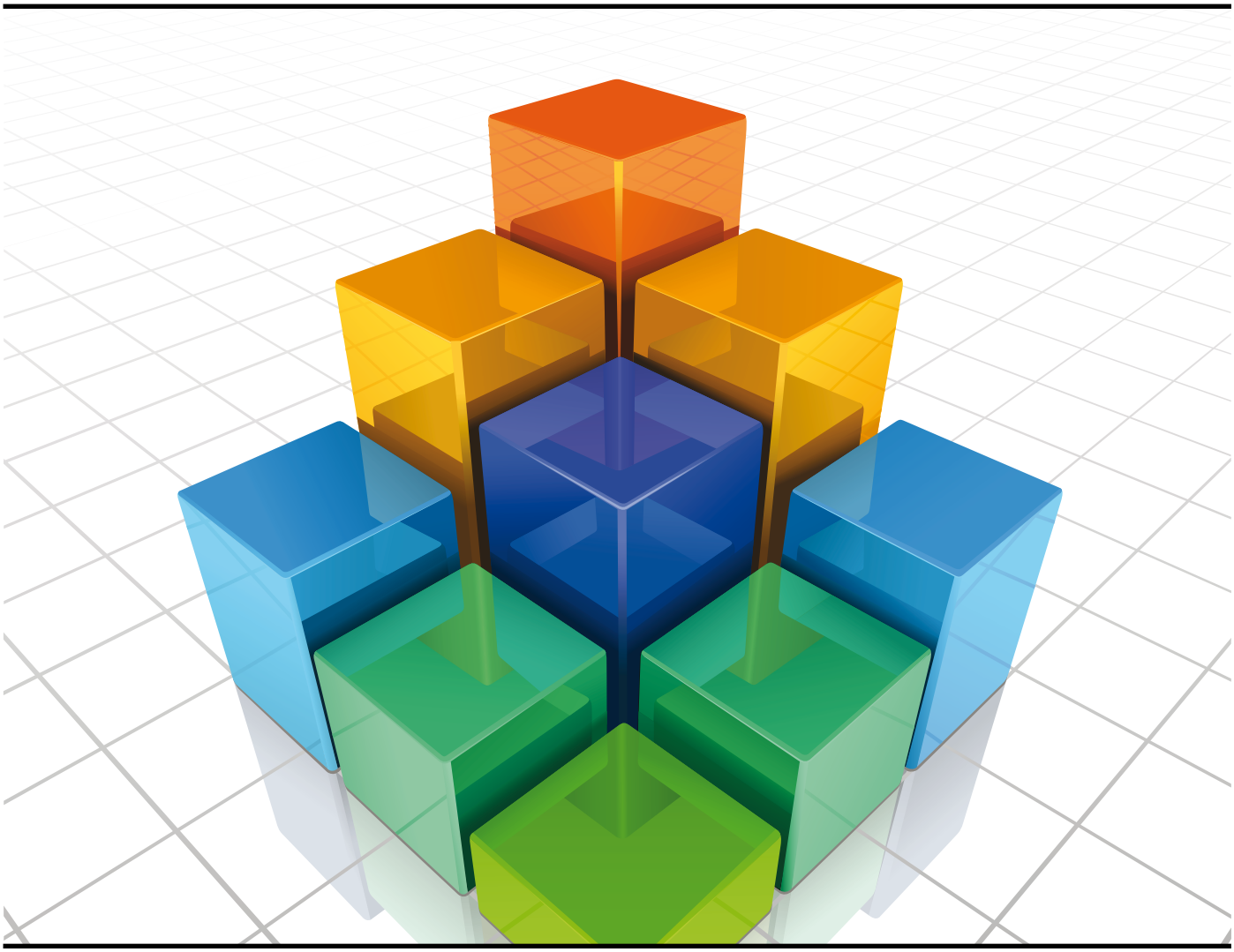
krebs:hilfe!

AUSTRIAN BREAST CANCER STUDY GROUP

CECOG Central European Cooperative Oncology Group

ÖSTERREICHISCHE KREBSHILFE

HEFT 2:2009



Grundlagen der Statistik

Zusammengestellt von Univ.-Prof. Dr. Martina Mittlböck und Univ.-Prof. Dr. Engelbert Hanzal

Liebe LeserInnen!

Zahlen lesen in der Medizin. Schlägt man heutzutage medizinische Fachjournale oder auch populärwissenschaftliche Zeitschriften auf, so findet man nur mehr selten einzelne Fallbeschreibungen, sondern meistens eine Beschreibung und Zusammenfassung von größeren Patientengruppen. Diese erfolgt durchwegs mit Grafiken und/oder statistischen Kennzahlen der erhobenen Daten. Viele Arbeiten zielen weiters darauf ab, Gruppen zu vergleichen und Erkenntnisse für verbesserte Prognosen und Therapien von Patienten zu gewinnen.

Die folgenden Artikel sollen einen kurzen Überblick über die gängigsten statistischen Methoden geben. Sie sollen die LeserInnen in grundlegende Definitionen einführen, um verwendete Begriffe besser verstehen zu können und selbstständig die Aussagen und Sinnhaftigkeit von statistischen Beschreibungen, Kennzahlen und Tests zu beurteilen und auch zu hinterfragen. Trotz der in den meisten wissenschaftlichen Journalen angewandten Peer-Review-Verfahren kann man nicht ausschließen, dass unentdeckte Mängel in der Publikation vorhanden sind. Nach wie vor obliegt es auch den Lesern, die Qualität von Publikationen zu beurteilen und zu fragen, ob und wieweit die publizierten Ergebnisse für ihre Patienten anwendbar sind. Die drei wichtigsten Fragen dazu sind:

1. Können die publizierten Ergebnisse stimmen?
2. Was sind die Ergebnisse und wie sind sie zu interpretieren?
3. Können mir die Ergebnisse in der Behandlung meiner Patienten helfen?

Das Ziel jeder Publikation sollte die Gewinnung eines Nutzens bzw. Mehrwerts für zukünftige PatientInnen sein. Selbstverständlich wird dadurch die Erfahrung jener Personen, die diesen Mehrwert an Wissen anwenden wollen, nicht obsolet. Im Gegenteil: Die Fähigkeit, Ergebnisse aus Studien mit den Vorstellungen der Betroffenen in einer spezifischen klinischen Situation zu integrieren, ist heute eine der wichtigsten und anspruchsvollsten Aufgaben in der Medizin.

Durch die rasant gewachsenen Möglichkeiten moderener Informationstechnologie und die stetig wachsende Anzahl an potenziell wichtigen wissenschaftlichen Daten wird dies einerseits wichtiger, andererseits schwieriger. Nicht nur für MedizinerInnen, auch für Pflegepersonen, PhysiotherapeutInnen, Angehörige des Gesundheitsmanagements und nicht zuletzt auch für PatientInnen und die Allgemeinheit wird es immer leichter, auf wissenschaftliche Daten zuzugreifen. Umso wichtiger erscheint es, sich angesichts dieser Situation mit Werkzeugen zu beschäftigen, die es ermöglichen, im Dickicht des Informationsdschungels die „Spreu vom Weizen" trennen zu können. Möge diese Sonderausgabe der krebs:hilfe! dazu einen kleinen Beitrag leisten!

Univ.-Prof. Dr. Martina Mittlböck, Institut für Medizinische Statistik und Informatik, Medizinische Universität Wien

Univ.-Prof. Dr. Engelbert Hanzal, Universitätsklinik für Frauenheilkunde, Wien

FOTOS: PRIVAT

Grafiken und Maßzahlen

VON MMAG. DI DDR. THOMAS BENESCH

Mit der deskriptiven (beschreibenden) Statistik wird versucht, die in einem Datensatz enthaltene Information durch bestimmte Kennzahlen und grafische Darstellungen zu veranschaulichen. Die deskriptive Statistik beschränkt sich ausschließlich auf die Beschreibung des vorliegenden Datenmaterials. Meistens liegt das Interesse an Maßzahlen (z.B. mittleres Alter von Personen). Resultate der deskriptiven Analyse dienen oft als Grundlage für die Planung weiterer wissenschaftlicher Studien.

Eine Beobachtungseinheit (Merkmalsträger) ist die kleinste Einheit, an der Eigenschaften direkt beobachtet werden. Jene Eigenschaften, die für die Beobachtungseinheiten erhoben werden, werden als Merkmale (Variable) bezeichnet. Merkmale sind durch ihre Merkmalsausprägungen (Ausprägungen) charakterisiert (siehe folgende Tabelle).

Merkmal und Merkmalausprägung

Merkmal	Mögliche Ausprägungen
Geschlecht	männlich und weiblich
Körpergröße	172cm, 184cm, ...
Körpergewicht	50kg, 75kg, 100kg, ...

Skalenniveaus

Welches statistische Verfahren in einem konkreten Fall am besten zur Beschreibung eines bestimmten Merkmals geeignet ist, hängt hauptsächlich von der Skala (Messniveau) ab, mit der die Messung der Ausprägungen erfolgt. Die vier wichtigsten Skalenniveaus werden nun in ihren unterschiedlichen Charakteristika betrachtet:

1. Nominalskala. Die Merkmalsausprägungen entsprechen begrifflichen Kategorien. Es ist nicht möglich, die Ausprägungen eines solchen Merkmals nach einer Größer-kleiner-Relation anzuordnen. Hat ein Merkmal nur zwei mögliche Ausprägungen, so wird von einem dichotomen Merkmal gesprochen; die speziellen Nominalskalen sind Alternativskalen.

2. Ordinalskala. Die Merkmalsausprägungen dieser Skala lassen eine Größer-kleiner-Relation zu. Die Abstände zwischen zwei Merkmalsausprägungen haben keine inhaltliche Bedeutung. Ordinalskalen sind z.B. bei der Charakterisierung von Gemeindegrößen durch Begriffe wie „Großstadt“ oder „Dorf“ anzutreffen. Hier wird klar differenziert, welcher Begriff eine größere Gemeinde beschreibt, die Unterschiede zwischen aufeinander folgenden Klassen sind jedoch nicht sinnvoll vergleichbar.

3. Intervallskala. Neben der Rangordnung einzelner Merkmalsausprägungen lassen sich die Abstände zwischen den einzelnen Merkmalsausprägungen interpretieren. Es existiert allerdings ein willkürlich gesetzter Nullpunkt. Aufgrund dieser Tatsache lassen sich keine sinnvollen Verhältniszahlen bilden.

4. Verhältnisskala. Die Merkmalsausprägungen haben einen natürlichen Nullpunkt, und es gelten die Eigenschaften der Intervallskala. Das Berechnen von Verhältniszahlen ist möglich und sinnvoll.

Intervall- und Verhältnisskala werden oft in dem Überbegriff „metrische Skala“ subsumiert. Nominal- und Ordinalskalen werden häufig als kategorielle Variable zusammengefasst, die festhält, zu welcher Kategorie oder Klasse jede Beobachtungseinheit bezüglich eines Merkmals gehört. Bei einer univariaten Auswertung eines Datensatzes wird jedes der einzelnen Merkmale gesondert analysiert.

Häufigkeit

Bei der Beschreibung nominaler Merkmale ist der Begriff der Häufigkeit von zentraler Bedeutung. Bei jeder Merkmalsausprägung wird ausgezählt, wie viele Beobachtungseinheiten diese spezielle Ausprägung annimmt. Diese Anzahl wird als absolute Häufigkeit bezeichnet. Wird die absolute Häufigkeit durch die Gesamtheit der Beobachtungseinheiten (**Stichprobenumfang n**) dividiert, ist das Ergebnis die relative Häufigkeit. Ergebnisse von Auszählungen dieser Art werden in Häufigkeitstabellen und Diagrammen dargestellt. Die (angeordneten) Merkmalsausprägungen und die zugehörigen (absoluten oder relativen) Häufigkeiten ergeben die Verteilung (oder Häufigkeitsverteilung) des betreffenden Merkmals.

Auch für ordinal skalierte Merkmale können Häufigkeitstabellen erstellt bzw. Stabdiagramme gezeichnet werden. In den Tabellen und grafischen Darstellungen für ordinale Merkmale wird allerdings die Tatsache berücksichtigt, dass die Merkmalsausprägungen einer Ordnungsrelation unterliegen. Beim Stabdiagramm sollte die Anordnung der Merkmalsausprägungen auf der x-Achse der vorliegenden Ordnungsrelation entsprechen.

Anhand der sortierten Werte (Rangliste) können für vorliegende Merkmalsausprägungen Quantile bestimmt werden. Ein p-Quantil ist jener Beobachtungswert, der größer oder gleich als mindestens 100 p Prozent der Werte und zugleich kleiner oder gleich als 100 (1-p) Prozent der Werte ist. Das 0,5-Quantil ist somit der Beobachtungswert, der größer oder gleich 50 Prozent der beobachteten Werte und zugleich kleiner oder gleich als 50 Prozent der Werte ist. Das 0,5-Quantil wird auch **Median $\tilde{x}$** oder zweites Quartil genannt, das 0,25-Quantil wird erstes Quartil oder unteres Quartil (Q1) und das 0,75-Quantil wird drittes Quartil oder oberes Quartil (Q3) genannt.

Lagemaße

Zur grafischen Beschreibung eines metrischen Merkmals werden die beobachteten Messwerte oft zu Gruppen zusammengefasst (klassifiziert). Um eine Klassifikation vorzunehmen, müssen zunächst sinnvolle Klassengrenzen bzw. Wertebereiche für die einzelnen Klassen festgelegt werden. Als Faustregel gilt, dass die Anzahl der zu bildenden Klassen ungefähr so groß sein soll wie die Quadratwurzel aus der Anzahl der Beobachtungen. Die Klassenbreiten sollten – wenn möglich – gleich lang gewählt werden. Die wichtigste grafische Darstellungsform für metrische Variablen ist das Histogramm.

Ein weiteres Lagemaß ist das **arithmetische Mittel** $\bar{x}$. Trotz seines breiten Anwendungsbereichs gibt das arithmetische Mittel bei bestimmten Merkmalen aus sachlogischen Gründen nicht den richtigen Durchschnitt an. Dies ist dann der Fall, wenn relative Änderungen als Merkmalsausprägung von Interesse sind, wie dies bei zeitabhängigen Messzahlen der Fall ist. Zeitabhängige Messzahlen sind zu erhalten, indem zwei Beobachtungen mit unterschiedlichem Zeitbezug, aber für dasselbe Merkmal, ins Verhältnis gesetzt werden. In diesem Fall ist das **geometrische Mittel** $\bar{x}_g$ das bessere Maß (die Tabelle fasst die Lagemaße und ihre Eigenschaften zusammen.)

Lagemaße und ihre Eigenschaften

Lagemaß	Median	arithmetisches Mittel	geometrisches Mittel
Skala	Ordinalskala	Intervallskala	Verhältnisskala
Auswirkung von Extremwerten	geringe	große	mittlere

Streuungsmaße

Zur weiteren Beschreibung eines Merkmals dienen Streuungsmaße. Ist die zentrale Tendenz von Merkmalsausprägungen bekannt, so ist im Allgemeinen auch interessant, in welchem Ausmaß die einzelnen Werte um das Zentrum der Merkmalsausprägungen streuen. Streuungsmaße sind häufig Maßzahlen zur Risikoschätzung.

Als einfach zu bestimmende Streuungsmaße bietet sich der Interquartilsabstand oder auch die Spannweite an. Der **Interquartilsabstand IQR** ist die Differenz zwischen dem oberen Quartil Q3 und dem unteren Quartil Q1. Die **Spannweite R** wird durch die Differenz zwischen größtem und kleinstem Wert aller vorliegenden Beobachtungen errechnet. Die Spannweite gibt jenen Bereich an, in dem die gesamten Merkmalsausprägungen liegen; der Interquartilsabstand ist jener Bereich, in dem die zentralen 50 Prozent der Merkmalsausprägungen liegen. Die **Stichprobenvarianz** ist das Mittel (dividiert durch n-1) der quadratischen Abweichungen der Merkmalsausprägungen von ihrem arithmetischen Mittel. Die **Standardabweichung s** ist durch die Quadratwurzel der Stichprobenvarianz gegeben. Die Standardabweichung entspricht eher dem anschaulichen Begriff der „Streuung" als die Stichprobenvarianz. Die Standardabweichung hat die gleiche Dimension wie die Ursprungswerte; hat diese z.B. die Bezeichnung cm, so gilt dies ebenfalls für die Standardabweichung (die Tabelle fasst die wichtigsten Streuungsmaße und ihre Eigenschaften zusammen).

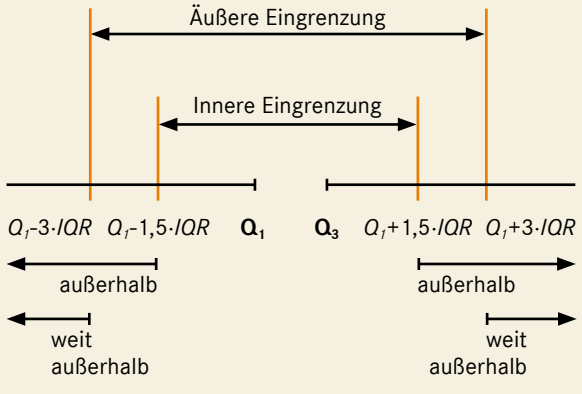
Streuungsmaße und ihre Eigenschaften

Streuungsmaß	Spannweite	Stichprobenvarianz	Standardabweichung
Skala	mindestens Intervallskala		
Auswirkung von Extremwerten	große	mittlere	mittlere

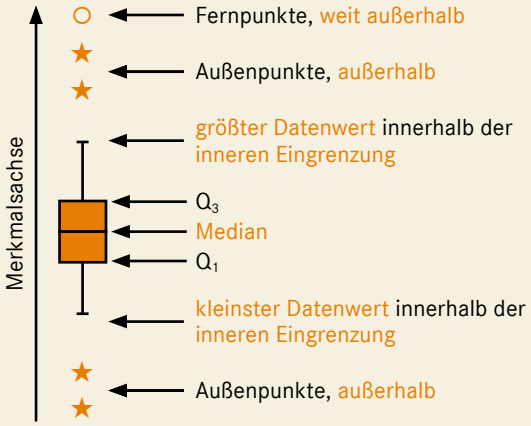
Mithilfe von Quartilen können Merkmale kurz und einfach beschrieben werden: z.B. durch die beiden Extremwerte „das Merkmal reicht von Minimum bis Maximum", den Median und das untere und obere Quartil „50 Prozent der Merkmalswerte liegen zwischen dem unteren und oberen Quartil". Diese informative Darstellung wird als das Fünf-Zahlen-Maß bezeichnet. Zur Veranschaulichung von Merkmalsausprägungen von metrisch skalierten Merkmalen kann der **Box-and-Whiskers-Plot, oder kurz Boxplot,** verwendet werden. Der untere (bzw. der obere) Rand des Kernstücks (der Box) wird vom unteren (bzw. vom oberen) Quartil gebildet. Somit liegen die mittleren 50 Prozent der Beobachtungswerte innerhalb der Box. Als Querlinie wird der Median eingezeichnet. Die Längen der Whiskers (Antennen) sind unterschiedlich definiert. Bei der einfachsten Variante enden sie bei der kleinsten und der größten Beobachtung. In der einfachsten Variante wird das Fünf-Zahlen-Maß direkt in einem Boxplot dargestellt.

Zum Bestimmen von extremen Beobachtungen wird der Interquartilsabstand IQR verwendet – dieser wird mit Eingrenzungen berechnet. Dabei werden zwei Arten von Eingrenzungen unterschieden: die inneren und die äußeren. Die folgende Abbildung veranschaulicht die Berechnung der Eingrenzungen.

Die Definition von Boxplot



Boxplot unter Berücksichtigung von Außenpunkten und Fernpunkten



Als Außenpunkte werden alle Merkmalsausprägungen zwischen innerer und äußerer Eingrenzung definiert. Fernpunkte sind alle Merkmalsausprägungen außerhalb der äußeren Eingrenzung. Die Längen der Whiskers enden bei den äußersten Punkten innerhalb der inneren Eingrenzung. Als Boxplot ergibt sich nach TUKEY die folgende Darstellung (Abbildung links):
Ein Boxplot liefert Indizien auf Werte, die eventuell als Ausreißer einzustufen sind. Große Werte können auf Verarbeitungsfehler zurückgehen und damit falsch sein; andererseits können sie auch korrekt und von wesentlicher Bedeutung sein. Als Beispiel: Bei der Untersuchung des Wasserstands darf ein Hochwasser, das als sehr großer Wert aufscheint, keinesfalls ignoriert werden.

MMag. DI DDr. Thomas Benesch
Besondere Einrichtung für Medizinische Statistik und Informatik, Institut für Medizinische Statistik, Medizinische Universität Wien

Einführung in das statistische Testen

VON UNIV.-PROF. DR. MARTINA MITTLBÖCK

Schlägt man heutzutage medizinische Fachjournale oder auch populärwissenschaftliche Zeitschriften auf, so findet man neben einer ausführlichen Beschreibung der erhobenen und gemessenen Daten zumeist auch die Ergebnisse von verschiedensten statistischen Tests. Diese sind entweder am angegebenen p-Wert erkennbar oder durch die Bezeichnungen signifikant bzw. nicht signifikant (n.sig.). Im Folgenden soll die Verwendung statistischer Tests motiviert werden. Es soll auf ihre Interpretationen eingegangen und die wichtigsten Annahmen bzw. Fallstricke sollen erläutert werden.

Das Testprinzip

Es gibt für (fast) jede Gelegenheit statistische Tests, die je nach Situation unterschiedlich heißen. Allen Tests liegt aber das gleiche „Testprinzip" zugrunde:

1. Formuliere eine einfache, aber präzise Frage, z.B.: „Gibt es Unterschiede im Body Mass Index (BMI) zwischen zwei Gruppen?", „Hat sich ein präoperativer Wert durch die Operation verändert?"

2. Definiere das Signifikanzniveau α, also die maximale Wahrscheinlichkeit, mit der man fälschlicherweise einen Gruppenunterschied behaupten möchte, wenn tatsächlich keiner besteht. In der Medizin wird das Signifikanzniveau durchwegs mit 0,05 (= fünf Prozent) angesetzt.

3. Berechne den p-Wert (meist mittels Computers), das ist die Wahrscheinlichkeit, dass die Gruppen nur zufällig diesen Unterschied (oder einen noch größeren Unterschied) aufweisen, wie er in der Stichprobe beobachtet wurde.

4. Testentscheidung: Wenn dieser p-Wert kleiner oder gleich dem festgelegten Signifikanzniveau ist, dann schließen wir den Zufall als Erklärungsmöglichkeit für die beobachteten Unterschiede aus. Wir gehen davon aus, dass tatsächlich ein Unterschied besteht.

Angenommen es ist p=0,043 angegeben, so bedeutet das, dass der Test signifikant (zu α=0,05) ist. Die Wahrscheinlichkeit, dass man nur durch Zufall den beobachteten Unterschied oder noch extremere Unterschiede erhält, ist nur 4,3 Prozent. Das ist üblicherweise unwahrscheinlich genug, sodass man mit gutem Gewissen von einer tatsächlichen Ungleichheit der Gruppen ausgehen kann.

Angenommen es ist p=0,132 angegeben, so würde der Test nicht signifikant (zum Signifikanzniveau von 0,05) sein. Häufig wird dann behauptet, dass kein Unterschied zwischen den Gruppen besteht. Kann man das wirklich behaupten? Nein – da man sich trotzdem irren kann. Man kann einen vorhandenen Unterschied übersehen (β-Fehler). Diesen β-Fehler kann man durch die Testprozedur nicht kontrollieren, sondern nur durch eine adäquate Stichprobenplanung. Ist die Zahl der Beobachtungen (Fallzahl) gering, so wird man vorhandene Unterschiede leicht übersehen,

FOTO: PRIVAT

d.h., bei einem nicht signifikanten Testergebnis kann man nicht automatisch davon ausgehen, dass kein Gruppenunterschied besteht. Für gut durchgeführte prospektive Studien ist eine Stichprobenplanung unerlässlich – nur so kann man bei nicht signifikanten Testergebnissen auch davon ausgehen, dass die Gruppen (nahezu) gleich sind (siehe folgende Tabelle).

Klassische Testsituation: Wahrheit vs. Testergebnis

		Wahrheit	
		Gruppen sind gleich	Gruppen sind unterschiedlich
Testergebnis	nicht signifikant	korrekterweise wird kein Gruppenunterschied gefunden	Gruppenunterschiede werden übersehen = β-Fehler
	signifikant	fälschlicherweise werden Gruppenunterschiede angenommen = α-Fehler	Gruppenunterschiede werden gefunden (Mächtigkeit, Power)

Die Wahl des Tests

Das „Testprinzip" ist im Wesentlichen immer gleich, aber je nach dem, welche Situation zugrunde liegt, wird man unterschiedliche Tests wählen. Einerseits unterscheidet man zwischen abhängigen und unabhängigen Beobachtungen. Abhängige Beobachtungen liegen z.B. vor, wenn an einer Person zwei oder mehrere Messungen durchgeführt werden. Das heißt, jeweils zusammengehörige Messungen haben gleiche Bedingungen, und dieser Umstand sollte beim Testen berücksichtigt werden. Unabhängige Beobachtungen liegen vor, wenn jede Messung unabhängig von allen anderen Messungen erfolgt.

Zweitens spielt das Skalenniveau der gemessenen Daten eine Rolle (siehe auch Artikel „Grafiken und Maßzahlen"). Wenn z.B. das Auftreten von Komplikationen geprüft werden soll, so werden die Daten mit Häufigkeiten und Prozentsätzen beschrieben. Sollte hingegen z.B. der Body Mass Index verglichen werden, so wird man ihn entweder mit Mittelwerten und Standardabweichungen (wenn symmetrisch/normalverteilt) oder mit Medianen, Minima und Maxima in den einzelnen Gruppen vergleichen. Entsprechend unterschiedlich werden auch die verwendeten Tests sein (eine Übersicht der gängigsten Tests beim Vergleich von zwei Gruppen ist in der Tabelle ganz unten zu finden).

Gruppenunterschied und Konfidenzintervall

Von einem angegebenen p-Wert kann man nicht auf die Größe des Gruppenunterschieds schließen. Dafür muss man die einzelnen Gruppen beschreiben, z.B. Differenzen der Prozentsätze bei kategoriellen Daten, Mittelwerte (und Standardabweichungen) bzw. Mittelwertsdifferenzen für normalverteilte Daten usw. Eine verbesserte Beschreibung der Gruppenunterschiede erhält man, indem man nicht nur Punktschätzungen, wie z.B. Prozentsätze und Mittelwertsdifferenzen, angibt, sondern auch ein Intervall, das mit hoher Wahrscheinlichkeit (meist mit 95 Prozent) den (unbekannten wahren) Gruppenunterschied überdeckt, das 95-%-Konfidenzintervall.

Retrospektive Untersuchungen

Bei prospektiven Analysen überlegt man sich vorher, welche primäre Frage gestellt wird und welcher Test dafür durchgeführt werden soll. Alle anderen Fragen sind sekundär, und es werden dafür meist nur beschreibende (explorative) Analysen durchgeführt. Bei retrospektiven Studien ist man meist nicht dazu gezwungen, sich eine primäre Forschungsfrage im vorhinein zu überlegen. Häufig wird solange gesucht, bis man Auffälligkeiten gefunden hat. In diesen Fällen sollten die Daten primär beschrieben werden, denn ist bei einem Test die Wahrscheinlichkeit, fälschlicherweise einen signifikanten Unterschied zu erhalten, fünf Prozent, so steigt die Wahrscheinlichkeit, bei zwei unabhängigen Tests fälschlicherweise zumindest einen signifikanten p-Wert zu erhalten, auf 9,75 Prozent, bei drei unabhängigen Tests auf 14,3 Prozent und bei zehn unabhängigen Tests auf 40,1 Prozent, wenn eigentlich keinerlei Gruppenunterschiede bestehen. Retrospektive Studien haben meist nur beschreibenden bzw. hypothesengenerierenden Charakter, d.h., die auf diese Weise gefundenen Auffälligkeiten müssen erst noch in weiteren unabhängigen Studien auf ihre tatsächliche Bedeutung geprüft werden.

Typische Beispiele für statistische Tests

Das Auftreten von Komplikationen (ja vs. nein) soll zwischen zwei Operationstechniken verglichen werden. Jeder Patient wird nur

Einfache Tests für den Vergleich von zwei Gruppen

Skalenniveau des zu untersuchenden Merkmals	unabhängige Beobachtungen	abhängige Beobachtungen
Binär (nur zwei Ausprägungen)	χ^2-Test Fishers exakter Test	McNemar-Test
Nominal (nicht ordenbare Kategorien)	χ^2-Test	marginaler Homogenitätstest Bowker-Test
Ordinal oder stetig, nicht normalverteilt	Wilcoxons Rangsummentest (=Mann-Whitney U Test)	Vorzeichentest Wilcoxons Vorzeichenrangtest
Stetig und normalverteilt	ungepaarter (Student's) t-Test	gepaarter t-Test

Wichtige Punkte

- Ein statistisch signifikantes Testergebnis bedeutet nur, dass die Wahrscheinlichkeit, fälschlicherweise zu behaupten, „es besteht ein Unterschied zwischen den Gruppen", klein ist, d.h. maximal dem Signifikanzniveau von fünf Prozent entspricht.
- Ein nicht signifikantes Testergebnis bedeutet nicht unbedingt, dass kein Unterschied zwischen den Gruppen besteht, sondern nur, dass wir keinen Unterschied nachweisen konnten. Es könnte auch sein, dass zu wenig Information (=zu kleine Stichprobe) vorhanden ist, um einen vorhandenen Unterschied nachzuweisen.
- Wiederholte statistische Tests erhöhen die Wahrscheinlichkeit für falsch positive signifikante Ergebnisse.

einmal operiert, daher besteht Unabhängigkeit. Die Zielgröße hat nur zwei Ausprägungen (binär) → Chi-Quadrat-Test (χ^2-Test) oder bei sehr kleinen Gruppengrößen auch Fishers exakter Test.
Es soll geprüft werden, ob der mittlere Body Mass Index zwischen zwei Therapiegruppen gleich ist. Jeder Patient hat nur einen BMI (Unabhängigkeit), und wenn die einzelnen Werte in jeder Gruppe annähernd symmetrisch sind → ungepaarter t-Test. Liegt keine symmetrische Verteilung vor oder gibt es Ausreißerwerte → Wilcoxons Rangsummentest.
Soll überprüft werden, ob der BMI sich nach der Operation gegenüber seinem Wert vor der Operation geändert hat, so liegt eine abhängige Situation vor → gepaarter t-Test, wenn die Differenzen annähernd symmetrisch verteilt sind, ansonsten → Vorzeichen-Test oder Wilcoxons Vorzeichenrangtest.

Univ.-Prof. Dr. Martina Mittlböck
Besondere Einrichtung für Medizinische Statistik und Informatik, Institut für Klinische Biometrie, Medizinische Universität Wien

Binäre Daten – Diagnosestudien

VON MMAG. DI DDR. THOMAS BENESCH

In der Epidemiologie (griechisch: Lehre über das Volk) werden Krankheiten und die damit in Zusammenhang stehenden Einflüsse untersucht. Zum Unterschied zur medizinischen Individualbetrachtung (hier wird die Krankheit eines einzelnen Patienten diagnostiziert und behandelt) beschäftigt sich die Epidemiologie mit dem Krankheitsbegriff innerhalb einer ganzen Bevölkerungsgruppe.

Erkrankungshäufigkeit

Spezielle epidemiologische Maßzahlen verhelfen zur deskriptiven Beschreibung der Erkrankungshäufigkeiten innerhalb einer Bevölkerung. Die Menge kranker Menschen zu einem bestimmten Zeitpunkt wird (**Punkt-**)**Prävalenz P** genannt; diese wird stets hinsichtlich einer bestimmten Bevölkerung angegeben.
Beispiel 1: *Die Prävalenz von Aids bei Österreichern/innen zwischen 15 und 49 Jahren im Jahre 2001 lag bei 0,2 Prozent; in Botswana (15- bis 49-Jährige Ende 2001) bei ca. 30 Prozent; bei Suchtgiftabhängigen (intravenous drug users) in Wien im Jahr 1990 bei ca. 27 Prozent.*
Sowohl eine genaue Beschreibung der betrachteten Bevölkerungsgruppe als auch der genaue Zeitpunkt sind zur Angabe der Prävalenz wesentlich.
Statt die Prävalenz P einer Erkrankung zu nennen, wird häufig die **Chance** (**Odds**) der Erkrankung angegeben: $\text{Odds} = \frac{P}{(1-P)}$
Umgekehrt kann die Prävalenz aus der Chance berechnet werden: $P = \frac{\text{Odds}}{(1+\text{Odds})}$.

Fortsetzung Beispiel 1: *Für einen Suchtgiftabhängigen in Wien ergab sich im Jahre 1990 Odds = 0,27/0,63 = 0,43. Das heißt, die Chance auf das Vorliegen von Aids war 0,43:1.*
Für eine/n Österreicher/in zwischen 15 und 49 Jahren im Jahre 2001 hingegen ergab sich Odds = 0,002/0,998 = 0,002004, das heißt, die Chance auf Aids war ungefähr 0,002:1. Hier ist die Prävalenz so gering, dass die Odds praktisch gleich der Prävalenz ist.
Die kumulative Inzidenz (**Risiko**) **CI** gibt die Wahrscheinlichkeit (in Prozent) an, dass eine zufällig ausgewählte gesunde Person innerhalb eines bestimmten Zeitraums (z.B. innerhalb eines Jahres) neu erkrankt.

$$CI = 100 \times \frac{\text{Anzahl der im Zeitraum neu Erkrankten}}{\text{Anzahl der Gesunden in der Population zu Beginn des Zeitraums}}$$

Fortsetzung Beispiel 1: *Im Jahr 2001 gab es innerhalb der gesamten österreichischen Bevölkerung 70 Neuerkrankungen an Aids. Bei einer geschätzten Anzahl von 8.075.000 gesunden Einwohnern ergibt sich somit eine kumulative Inzidenz von CI = 100 x 70/8.075.000 = 0,0009% (Quelle: UNAIDS/WHO Fact Sheet 2001).*

Zusammenhänge beschreiben

Anschließend werden epidemiologische Assoziationsmaße bestimmt, welche den Zusammenhang von Krankheitshäufigkeiten mit anderen Faktoren beschreiben. Zur übersichtlichen Darstellung der Zahlenwerte werden Kontingenztafeln verwendet.

FOTO: PRIVAT

Relatives Risiko. In der Epidemiologie werden Erkrankungswahrscheinlichkeiten als Risiko bezeichnet, daher wird von einem relativen Risiko gesprochen. Es gibt den multiplikativen Faktor an, um den sich die Erkrankungswahrscheinlichkeit bei einer definitiven Exposition erhöht. Das relative Risiko ist stets eine positive Zahl, die beliebig groß werden kann. Ist das relative Risiko gleich 1, dann sind beide Risiken identisch, und das bedeutet wiederum, dass die Exposition keinen Einfluss auf die Erkrankung hat. Das relative Risiko wird als Verhältnis der Erkrankungswahrscheinlichkeit von Exponierten (z.B. von Rauchern oder von Patienten mit bestimmten Symptomen oder Merkmalen) und der Erkrankungswahrscheinlichkeit von Nichtexponierten definiert:

$$\text{relatives Risiko} = \frac{\text{Erkrankungswahrscheinlichkeit bei Exposition}}{\text{Erkrankungswahrscheinlichkeit bei Nichtexposition}}$$

***Beispiel 2* (*Zusammenhang zwischen Rauchen und Tod durch Lungenkrebs*):** *Die folgende Vierfeldertafel beschreibt den Zusammenhang zwischen Rauchen und Tod durch Lungenkrebs.*
In diesem Beispiel wird nicht die Erkrankungswahrscheinlichkeit betrachtet, sondern die Sterbewahrscheinlichkeit durch Lungenkrebs. Es wird nun das relative Risiko für Raucher bestimmt: relatives Risiko = 60/140 / 40/160 = 0,4286/0,25 = 1,71. Das Sterberisiko für Raucher steigt somit um etwa das 1,7-Fache gegenüber Nichtrauchern.

→ Tod durch Lungenkrebs

Raucher	Ja	Nein	Zeilensumme
Ja	60	80	140
Nein	40	120	160
Spaltensumme	100	200	300

Odds Ratio (OR). Neben dem relativen Risiko gibt es ein weiteres Vergleichsmaß innerhalb der Epidemiologie, das nicht auf dem Risiko, sondern auf der Chance beruht. Die Odds Ratio ist als der Faktor zu interpretieren, um den die Chance bei Exposition steigt, und wird in vielen epidemiologischen Studien als Hauptzielparameter angesetzt. Analog zum relativen Risiko wird die Odds Ratio wie folgt definiert:

$$\text{Odds Ratio} = \frac{\text{Odds bei Exposition}}{\text{Odds bei Nichtexposition}}$$

Fortsetzung Beispiel 2: *Die Odds Ratio für Raucher, an Lungenkrebs zu sterben, berechnet sich zu OR = 60/80/40/120 = 0,75/0,33 = 2,25 Somit ist die Chance für Raucher, an Lungenkrebs zu sterben, um den Faktor 2,25 höher als für Nichtraucher.*

Diagnostische Tests

Nachfolgend werden zum besseren Verständnis einige wesentlich Begriffe im Zusammenhang mit diagnostischen Tests erklärt. Die ersten beiden Begriffe werden beim Klassifikationsergebnis eines Diagnoseverfahrens verwendet:
Die **Sensitivität** ist die Wahrscheinlichkeit für eine positive Diagnose, falls der Patient tatsächlich erkrankt ist (richtig positiv Rate).
Die **Spezifität** ist die Wahrscheinlichkeit für eine negative Diagnose, falls der Patient nicht erkrankt ist (richtig negativ Rate).
Es ist zu beachten, dass Sensitivität und Spezifität auf zwei verschiedenen Gesamtheiten beruhen: Die Sensitivität bezieht sich auf kranke Patienten, die Spezifität auf Gesunde.

***Beispiel 3* (*Klopftest für irreversible Pulpitis*):** *Um bei einem Patienten mit pulsierendem Schmerz besser entscheiden zu können, ob eine irreversible Pulpitis des betroffenen Zahns tatsächlich vorliegt, wird der so genannte Klopftest durchgeführt. Dabei wird auf den Zahn geklopft und überprüft, ob der Patient starke Schmerzen spürt. Falls er mit starken Schmerzen reagiert, handelt es sich um ein positives Testergebnis, das heißt, es wird eine Wurzelbehandlung durchgeführt.*
Bei einer Vierfeldertafel von 100 Patienten mit pulsierendem Schmerz ergeben sich folgende Häufigkeiten:

→ Irreversible Pulpitis

Klopftest	Ja	Nein	Zeilensumme
Positiv	(richtig positiv) 62	(falsch positiv) 1	63
Negativ	(falsch negativ) 3	(richtig negativ) 34	37
Spaltensumme	65	35	100

Die Sensitivität ist der Anteil der Patienten mit positivem Klopftest innerhalb der Patienten mit irreversibler Pulpitis, also Sensitivität = 62/65 = 0,954
Die Spezifität ist der Anteil der Patienten mit negativem Klopftest innerhalb der Patienten, ohne irreversibler Pulpitis, d.h. Spezifität = 34/35 = 0,971

Während Sensitivität und Spezifität dazu dienen, die Güte eines Diagnoseverfahrens zu beschreiben, sollen die folgenden beiden Begriffe die Vorhersagekraft eines diagnostischen Tests charakterisieren.
Der **positive Vorhersagewert** (**positive predictive value**) einer Diagnose ist die Wahrscheinlichkeit, dass die Erkrankung vorliegt, wenn die Diagnose positiv ist.
Der **negative Vorhersagewert** (**negative predictive value**) einer Diagnose ist die Wahrscheinlichkeit, dass die Erkrankung nicht vorliegt, wenn die Diagnose negativ ist.

Fortsetzung Beispiel 3: *Der positive Vorhersagewert des Klopftests ist der Anteil der Patienten, bei denen eine irreversible Pulpitis vorliegt, bezogen auf die Patienten mit positivem Klopftest, also positiver Vorhersagewert = 62/63 = 0,984. Der negative Vorhersagewert des Klopftests ist der Anteil der Patienten, bei denen eine irreversible Pulpitis nicht vorliegt, bezogen auf die Patienten mit negativem Klopftest, also negativer Vorhersagewert = 34/37 = 0,918.*

Das relative Risiko kann auch unter Verwendung des positiven und negativen Vorhersagewerts berechnet werden – relatives Risiko = positiver Vorhersagewert/1-negativer Vorhersagewert = 0,984/0,081 = 12,15. Die Odds Ratio ist schließlich – Odds Ratio = Chance der Exponierten/Chance der Nichtexponierten = 62/1 / 3/34 = 702,67.

MMag. DI DDr. Thomas Benesch
Besondere Einrichtung für Medizinische Statistik und Informatik, Institut für Medizinische Statistik, Medizinische Universität Wien

Kaplan-Meier-Kurven und die Hazard Ratio

VON UNIV.-PROF. DR. HARALD HEINZL

In vielen klinischen Bereichen, vor allem aber in der Onkologie, wird der Erfolg von Therapien anhand des Patientenüberlebens beurteilt. Neben dem Tod im engeren Sinn kann auch das Auftreten von Rezidiven, Metastasen u.ä. von Interesse sein. Wir sprechen daher von Überlebens- oder allgemeiner von Ereigniszeiten. Am Ende vieler Studien tritt häufig der erfreuliche Fall ein, dass Patienten noch leben bzw. rezidiv- und/oder metastasenfrei leben.

Zensierung, Abschneidung

Die ermittelten Zeitspannen sind dann Untergrenzen für die tatsächlichen Ereigniszeiten, wir sprechen von zensierten – genauer rechtszensierten – Beobachtungen. Obwohl hier nicht weiter behandelt, sei auch auf die Möglichkeit von links- und intervallzensierten Beobachtungen verwiesen, bei denen nur eine Obergrenze bzw. eine Unter- und Obergrenze für die tatsächliche Ereigniszeit beobachtet werden kann. So wird ein Lokalrezidiv irgendwann zwischen dem letzten und dem aktuellen Nachsorgetermin auftreten. Eine derartige Intervallzensierung wird gerne ignoriert und stattdessen der aktuelle Nachsorgetermin als Rezidivdatum verwendet. So ein Vorgehen erscheint nur bei einem entsprechend dichten und für alle Patienten einheitlichen Nachsorgeschema einigermaßen akzeptabel.

Von der Zensierung muss die Abschneidung (engl. truncation) unterschieden werden. Bei Zensierung ist das Auftreten des Ereignisses nicht genau bestimmbar. Bei Abschneidung hingegen bleibt die Existenz eines Patienten unbekannt, wenn sein Ereignis in ein gewisses Intervall fällt. Zum Beispiel werden in ein kardiologisches Zentrum nur transportfähige Herzinfarktpatienten mit einer gewissen Mindestlebensdauer eingeliefert, von den unmittelbar nach einem Herzinfarkt Verstorbenen erfährt das Zentrum nichts (die mögliche Gesamtstichprobe wird quasi „links abgeschnitten").

Wir wollen uns im Folgenden auf nicht informativ rechtszensierte Ereigniszeiten beschränken. Nicht informative Zensierung bedeutet, dass sich zensierte und nicht zensierte Patienten bezüglich ihrer tatsächlichen Ereigniszeiten nicht unterscheiden. Sie muss argumentativ begründet werden, denn zensierte Ereigniszeiten können definitionsgemäß nicht beobachtet werden. Sie ist bei randomisierten klinischen Studien am Ende der Follow-up-Periode wohl zumeist gegeben; denkbare Ausnahmen wären extrem lange Studien, wo sich die zugrundeliegenden Patientenpopulationen im Laufe der Zeit verändern. Problematischer verhält es sich bei Patienten, die während einer Studie ihre Teilnahme zurückziehen (withdrawal) oder verloren gehen (lost to follow-up), hier kann eine Änderung der Ereigniswahrscheinlichkeit nach Zensierung oft nur schwer ausgeschlossen werden.

→ Dos and Donts

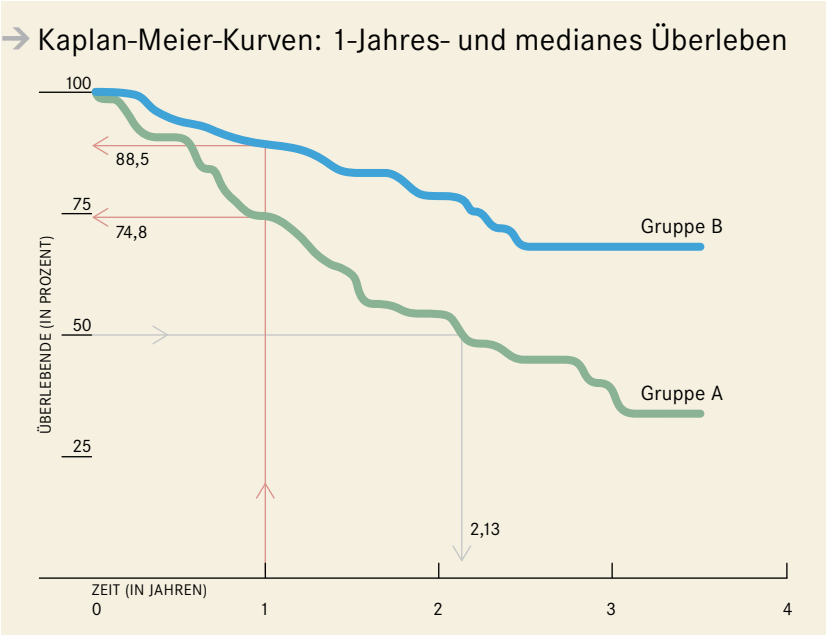
- Die elektronische Erfassung rechtszensierter Ereigniszeiten erfordert zwei Variablen: eine Zeit- und eine Statusvariable. Letztere gibt an, welche der gemessenen Zeitdauern tatsächlichen Ereignissen und welche Zensierungen entsprechen.
- Zensierte Beobachtungen sind keine fehlenden Werte, sie dürfen daher keinesfalls aus der Datei gelöscht werden.
- Die mediane Nachbeobachtungszeit (median follow-up time) kann mittels einer Kaplan-Meier-Kurve bestimmt werden, bei der die Bedeutung der Statusvariablen vertauscht wird (Zensierungen werden als Ereignisse gedeutet und vice versa).
- Kreuzende Kaplan-Meier-Kurven weisen auf nicht proportionale Hazardverläufe hin. In so einem Fall ist die Hazard Ratio keine Konstante mehr. Der Einsatz des Log-rank-Tests ist dann nicht möglich, und das Cox-Modell muss um zeitabhängige Hilfsvariablen erweitert werden, um den zeitabhängigen Verlauf der Hazard Ratio zu erfassen.

FOTOS: PRIVAT

Kaplan-Meier-Kurven

Das Auftreten von Zensierungen verhindert die Verwendung von graphischen Standardmethoden (z.B. Boxplot, Histogramm) zur Darstellung von Ereigniszeiten. Man weicht daher auf die kumulierte Verteilung aus. Unter Gültigkeit der nicht informativen Zensierung erhalten wir eine sogenannte Kaplan-Meier-Kurve (siehe Abbildung). Für jeden Zeitpunkt (x-Achse) können wir den Prozentsatz der noch lebenden Patienten ablesen (y-Achse). Zu Beginn leben natürlich noch alle, aber im Laufe der Zeit sinkt die Kurve immer mehr ab. Sie muss aber, aufgrund von zensierten Langzeitüberlebenden, nicht notwendigerweise null erreichen.

Gerne wird der genaue Überlebensprozentsatz für bevorzugte Zeitpunkte, wie das 1-Jahres-Überleben, angegeben. Man kann aber die Frage auch anders stellen und sich für die Zeitpunkte interessieren, bei denen noch ein bestimmter Prozentsatz der Patienten lebt. Üblich ist vor allem die Angabe der medianen Überlebenszeit, das ist der Zeitpunkt, zu dem 50 Prozent der Patienten noch leben. Die durchschnittliche Überlebenszeit ist bei zensierten Beobachtungen zumeist unbrauchbar, obwohl viele statistische Programme sie – leider – standardmäßig ausgeben.



Die Kaplan-Meier-Kurven zeigen die Überlebenswahrscheinlichkeiten über die Zeit für zwei Gruppen von Patienten. Gruppe A (grüne Linie) hat ein kontinuierlich höheres Ausfallsrisiko (Hazard) als Gruppe B (blaue Linie). Zur Bestimmung des 1-Jahres-Überlebens geht man an der 1-Jahres-Markierung senkrecht nach oben, bis man auf die Kaplan-Meier-Kurve trifft, und biegt dann orthogonal nach links zur Prozentskala ab (74,8 und 88,5 Prozent). Zur Bestimmung des medianen Überlebens geht man von der 50-Prozent-Markierung waagrecht nach rechts, bis man auf die Kaplan-Meier-Kurve trifft, und biegt dann orthogonal nach unten zur Zeitskala ab (2,13 Jahre bei Gruppe A). Beachte, für Gruppe B kann (noch) kein medianes Überleben bestimmt werden.

Hazard Ratio

Wenn eine Gruppe einen über die Zeit konstanten Überlebensvorteil (oder -nachteil) gegenüber einer Referenzgruppe aufweist, dann liegt eine sogenannte proportionale Hazard-Situation vor. In so einem Fall können Unterschiede zwischen Kaplan-Meier-Kurven mittels Log-rank-Test statistisch getestet werden. Für die Daten in der Abbildung ergibt sich ein p-Wert von 0,001. Er ist kleiner als das in der Medizin üblicherweise verwendete Signifikanzniveau von 0,05, womit wir den Zufall als Erklärung für die beobachteten Unterschiede verwerfen können.

Eine Erweiterung des Log-rank-Tests ist das proportionale Hazard-Modell von Cox, das hier eine Hazard Ratio von 2,5 (95% Konfidenzintervall: 1,4–4,3) ergibt. Die Hazard Ratio ist die Quantifizierung eines über die Zeit konstanten Überlebensnachteils bzw. -vorteils einer Gruppe zu einer Referenzgruppe. Das bedeutet konkret, dass Patienten in Gruppe A durchwegs ein 2,5-fach höheres Ausfallsrisiko als jene in Gruppe B haben. Andersherum ausgedrückt kann man auch von einem nur 0,40-fachen Ausfallsrisiko in Gruppe B im Vergleich zu Gruppe A sprechen (0,40=1/2,5). Übrigens, eine Hazard Ratio von eins würde bedeuten, dass kein Gruppenunterschied besteht.

Die Hazard Ratio ist auch für praktische Zwecke verwendbar. Die beobachtete Überlebenswahrscheinlichkeit zu einem Zeitpunkt in der Referenzgruppe hoch der geschätzten Hazard Ratio ergibt eine Schätzung für die Überlebenswahrscheinlichkeit in der interessierenden Gruppe. Auf das Ein-Jahres-Überleben von Gruppe B in der Abbildung angewandt, ergibt sich $0{,}885^{2{,}5}=0{,}737$ bzw. 73,7 Prozent als Schätzung für das Ein-Jahres-Überleben in Gruppe A. Der Unterschied zum beobachteten Wert von 74,8 Prozent ist gering und beruht auf zufälligen Stichprobenschwankungen.

Heutzutage gilt das proportionale Hazard-Modell von Cox als Standard zur Analyse von rechtszensierten Ereigniszeiten in der Medizin. Es ermöglicht die gemeinsame Untersuchung von mehreren, auch zeitabhängigen Einflussfaktoren (multivariable Analyse), um jene mit eigenständiger Bedeutung für das Patientenüberleben identifizieren zu können.

Für Interessierte sei abschließend noch auf die drei einführenden Arbeiten zur Überlebenszeitanalyse von Ziegler, Lange und Bender verwiesen, die 2007 in der „Deutschen Medizinischen Wochenschrift“ erschienen sind.

Univ.-Prof. Dr. Harald Heinzl
Besondere Einrichtung für Medizinische Statistik und Informatik, Institut für Klinische Biometrie, Medizinische Universität Wien

Epidemiologie und Demografie

VON UNIV.-PROF. DR. WILLIBALD STRONEGGER

Epidemiologie ist die Wissenschaft von den Krankheits- und Sterberisiken. Ausgangspunkt der Epidemiologie ist die Beobachtung, dass das Auftreten der meisten Krankheiten von Land zu Land oder von Bevölkerungsgruppe zu Bevölkerungsgruppe erstaunlich stark variiert. Dies ist auch bei vergleichbar ausgebauter medizinischer Versorgung der Fall. Die Lebensverhältnisse und physische wie gesellschaftliche Umweltfaktoren spielen folglich in der Entstehung und vor allem in der Verbreitung von Krankheiten eine große Rolle. Die Rolle dieser Faktoren zu erforschen und zu quantifizieren ist die Aufgabe der Epidemiologie.

Die Messung von Krankheitsrisiken

Was ist überhaupt ein Erkrankungsrisiko, und wie lässt es sich angeben und mitteilen? In der Praxis oft getätigte Aussagen wie „Ihr Risiko, an X zu erkranken, beträgt 20 Prozent" sind wertlos, sofern nicht explizit angegeben wird, auf welchen Zeitraum sich dieses Risiko bezieht. 20 Prozent mögen viel sein für ein halbes Jahr, aber wenig für die Lebenszeit. Nehmen wir nun an, das Risiko bezieht sich auf eine in der Risikokommunikation übliche Zeitspanne von fünf Jahren. Würde dann das Risiko, in den nächsten 25 Jahren zu erkranken, 100 Prozent betragen?
Risikoangaben dieser Art, sogenannte Inzidenzrisiken (auch „Inzidenzproportionen" oder „x-Jahres-Inzidenzen" genannt), eignen sich gut für die Kommunikation von Krankheits- oder Sterberisiken. Sie werden z.B. in Risk Charts für koronare Herzkrankheit angegeben. Inzidenzrisiken können aber nicht durch einfache Multiplikation auf andere Zeiträume übertragen werden, wie schon errechenbare Zahlenwerte von über 100 Prozent nahelegen!

a) Überschätzung von Risiken durch Verwendung des falschen Risikomaßes

Aus zwei Gründen wird das Risiko, an X zu erkranken, über einen längeren Zeitraum kleiner sein, als eine Multiplikation nahelegt. Erstens können Inzidenzrisiken 100 Prozent nicht überschreiten, nehmen also mit zunehmender Zeitspanne nicht linear zu; zweitens steigt mit zunehmender Zeitperiode die Wahrscheinlichkeit, dass eine Person gar nicht mehr an X erkranken kann, weil sie bereits vorher an einer anderen Krankheit Y verstorben ist.
Für beide Probleme wurden Gegenmittel entwickelt: Für das erste Problem gibt es die sogenannte Inzidenzrate (auch „Inzidenzdichte" oder „Hazard Rate" genannt), die für beliebige Zeiträume in Inzidenzrisiken (und in Überlebenswahrscheinlichkeiten) umrechenbar ist. Für das zweite Problem entwickelte man die „Competing Risk"-Analyse, welche die Verminderung des Risikos durch vorzeitiges Versterben berücksichtigt.

b) Inzidenzrate: Das Risiko, krank zu werden, pro Risikozeiteinheit

Die Inzidenzrate zählt jede Neuerkrankung (auch Wiederholungen in derselben Person!), und diese Anzahl wird auf die von den Personen bis zum Erkranken in „Gesundheit" verbrachte Zeit bezogen. Die Summe der Zeiten, in welchen eine Person noch nicht erkrankt ist, aber erkranken könnte, wird als Risikozeit bezeichnet und in „Personenjahren" (person years – py) angegeben:

$$\text{Inzidenzrate} = \frac{\text{Inzidenzfälle (Neuerkrankungen) in Zeitperiode}}{\text{Summe der Risikozeiten (in py)}}$$

Die Krebsinzidenzrate für Männer beträgt in Österreich etwa 3,3 Fälle auf 1000py. Für relativ kleine Krankheitsrisiken (und etwa zeitkonstante Inzidenz) erhält man das Inzidenzrisiko (IR) für eine Zeitperiode T näherungsweise einfach durch Multiplikation der Inzidenzrate I mit der Zeitspanne:
Inzidenzrisiko IR ≈ I x T = „kumulative Inzidenz"

Das Produkt aus Inzidenzrate mit Zeitperiode heißt kumulative Inzidenz CI (oder „kumulative Rate"). Die CI kann natürlich bei größeren Zeitperioden T-Werte über 1 (100 Prozent) annehmen, sie ist also kein Risiko. Fälschlicherweise wird dennoch häufig das Inzidenzrisiko IR als „kumulative Inzidenz" bezeichnet.
Für einen Mann beträgt das durchschnittliche Risiko, in zehn Jahren an Krebs zu erkranken, folglich ca. 3,3/100 = 3,3 Prozent. Hier wurde nicht berücksichtigt, dass die Krebsinzidenz mit dem Alter stark ansteigt. Die Krebsinzidenz bei Männern über 60 beträgt ca. 10/1000py. Das IR für 20 Jahre wäre nach der Näherungsformel I x T = 20 Prozent, nach der genauen Berechnung aber nur 18 Prozent. Bei nicht so selten (pro Zeiteinheit!) auftretenden Krankheiten wie Krebs kann das exakte IR erheblich kleiner sein als die kumulative Inzidenz I x T angibt!

c) Vernachlässigung der Competing Risks

Der zweite Grund, weshalb die tatsächlichen Inzidenzrisiken oft kleiner sind als die berechneten, liegt in der leider standardmäßig fehlenden Berücksichtigung der konkurrierenden Mortalität. Dies trifft besonders auf Risiken bei älteren Personen und für längere Zeiträume zu. Auch Lebenszeitrisiken werden oft nicht mit Competing-Risk-Analyse berechnet und mitunter stark überschätzt! Diese üblichen Berechnungen geben daher Inzidenzrisiken unter der Annahme an, dass die untersuchte Person das

Auftreten der betrachteten Krankheit jedenfalls erleben kann und nicht vorher verstirbt.

Die Messung von Risikounterschieden („Effekten")

Als Risikofaktor bezeichnet man einen Faktor, dem eine Person exponiert ist oder war (z.B. Rauchen, Stress, genetische Merkmale etc.) und welcher mit einem erhöhten oder verminderten Krankheitsrisiko einhergeht. In Zeitungsberichten über neu entdeckte Risikofaktoren finden sich so gut wie immer Aussagen der Form „Personen mit Faktor F weisen ein um 35 Prozent erhöhtes Risiko auf, an X zu erkranken". Um welche Zahl handelt es sich bei dieser Prozentangabe? Obwohl in Prozent angegeben, ist sie kein Inzidenzrisiko, sondern ein sogenanntes relatives Risiko (RR), das aus dem Vergleich der Krankheitsrisiken zwischen einer dem Faktor F exponierten Personengruppe mit einer Gruppe ohne Exposition entsteht.

$$\text{Relatives Risiko RR} = \frac{\text{Krankheitsrisiko bei exponierten Personen}}{\text{Krankheitsrisiko bei nicht exponierten Personen}}$$

Die Unterschiedlichkeit des Krankheitsrisikos in der Gruppe unter Exposition im Vergleich zu einer nicht exponierten Personengruppe wird als Effekt (eines Risikofaktors) bezeichnet. Dieser „Effekt" muss nicht notwendig kausal und auch der Faktor nicht praktisch beeinflussbar sein! Bei den meisten Erkrankungen ist das Alter der Person der größte bekannte Risikofaktor. Obwohl man das Alter nicht ändern kann, ist es für die epidemiologische Analyse und oft für die ärztliche Praxis wichtig zu wissen, in welcher Weise ein Krankheitsrisiko mit dem Alter zunimmt.

a) Relatives Risiko: Risikounterschiede relativ

Die meisten heute berichteten RR liegen im Bereich +20 Prozent bis +100 Prozent. Werte um +100 Prozent sehen schon beeindruckend aus, beim Vergleich mit bekannten Krebsrisiken wie z.B. dem Tabakkonsum verlieren sie schnell an Eindruckskraft. Ein Raucher erhöht sein Lungenkrebsrisiko um etwa +1000 Prozent bis +2000 Prozent oder mehr. Hier wird ersichtlich, dass relative Risiken, die sich auf unterschiedliche Krankheiten und Personengruppen beziehen, schwer zu vergleichen sind. Das RR eines Rauchers für einen Herzinfarkt beträgt (im Vergleich zu einem Nichtraucher) etwa +50 Prozent. Diese Zahlenangaben erzeugen bei den meisten Lesern die Auffassung, dass Rauchen vor allem zu einem hohen Krebsrisiko führt, weniger zu einem Herz-Kreislauf-Risiko. Tatsächlich steigt aber durch Rauchen das Risiko, von einem Infarkt betroffen zu sein, deutlich stärker an als jenes eines Lungenkarzinoms.

b) Attributables Risiko: Risikounterschiede absolut

Das Problem bei der Interpretation des RR besteht darin, dass es relativ ist, d.h. keine Aussage über die tatsächliche bzw. absolute Risikoänderung ermöglicht, solange das absolute Krankheitsrisiko ohne Exposition – das Grundrisiko I_O – nicht bekannt ist. Das Grundrisiko für ein Lungenkarzinom ist erheblich kleiner als jenes eines Herzinfarktes. Selbst das hohe RR für Lungenkarzinom erzeugt weniger zusätzliches absolutes Risiko als das vergleichsweise kleine RR von +50 Prozent für Herzinfarkt, das sich auf ein viel größeres Grundrisiko bezieht.

Die absolute Zu- oder Abnahme eines Krankheitsrisikos wird durch das attributable Risiko AR beschrieben:

Attributables Risiko AR = I_O x (RR-1) = Grundrisiko x relatives Risiko in Prozent

Das AR ermöglich die Quantifizierung der Wirkung eines Risikofaktors hinsichtlich der individuellen Risikosteigerung (-reduktion) als auch auf Fallzahlen in Bevölkerungen, was besonders in praxisnahen Bereichen von Interesse ist. Im Gegensatz dazu quantifiziert das RR mehr die kausale Bedeutung eines Faktors in der Entstehung oder Verbreitung einer Erkrankung. Daher kommt dem RR eine wichtige Rolle in der ätiologischen oder präventiven medizinischen Forschung zu. Heute oft berichtete RR von unter 100 Prozent Risikoerhöhung aus einzelnen Beobachtungsstudien haben in der Regel nur einen innerwissenschaftlichen Wert, aber keinen praktischen Aussagewert für Ärzte oder Patienten!

Die praktische Messung von Risiken und Effekten

Krankheitsrisiken sind an sich unbekannt! Man kennt sie nur, wenn man sie misst. Wie für jede Messung benötigt man geeignete Messinstrumente, und wie bei allen Messinstrumenten gibt es auch hier bessere und schlechtere, teurere und billigere.

a) Ideales Inzidenzmessinstrument der Epidemiologie: Kohortenstudie

Inzidenzen sind aufgrund ihres inhärenten Zeitbezugs oft nicht einfach zu messen. Das teuerste und aufwändigste, aber auch qualitativ beste Messinstrument für die Inzidenzmessung ist die Kohortenstudie. Sie ist eine (im Standardfall prospektive) Beobachtungsstudie, in der eine bestimmte Gruppe von gesunden Personen („Kohorte") hinsichtlich der Präsenz von Expositionen untersucht und über einen längeren Zeitraum auf das Auftreten einer bestimmten Krankheit (oder Tod) hin beobachtet wird.

b) Schnell und billig zum relativen Risiko: die Fall-Kontroll-Studie

Wäre es nicht denkbar, die aufwändigen und langwierigen Inzidenzmessungen zu umgehen und gleich „direkt" die relativen Risiken zu messen? Tatsächlich gelang es der Epidemiologie ein Messinstrument zu entwickeln, das ohne Inzidenzmessung eine Bestimmung der relativen Risiken erlaubt.

Fall-Kontroll-Studien beginnen mit der Sammlung von erkrankten Personen, die nachträglich auf das Vorliegen einer interessierenden Exposition untersucht werden. Dies wird in einem zweiten Schritt an einer Gruppe oder Stichprobe von Kontrollpersonen in analoger Weise durchgeführt. Aufgrund der nachträglichen Durchführung nennt man diese Studie oft „retrospektiv". Diese Bezeichnung sollte nicht generell auf die Fall-Kontroll-Studie angewendet werden, da es auch retrospektive Kohortenstudien und prospektive Fall-Kontroll-Studien (nested case-control study) gibt.

Es gehört zu den bemerkenswerten Resultaten der Epidemiologie, dass mit diesem Studientyp das relative Erkrankungsrisiko ohne Inzidenzmessung geschätzt werden kann, indem das Verhältnis der Expositions-Odds der Fälle zu jenem der Kontrollen berechnet wird. Für diese sogenannte „Odds Ratio", die die Größe des Zusammenhangs zwischen Exposition und Krankheitsrisiko quantifiziert, gilt:

RR = OR (nur in der Fall-Kontroll-Studie!)

Falls die Kontrollen – wie im klassischen Design von Cornfield (1951) – erst am Ende der Sammlung der Fälle aus einer fixen Kohorte (d.h. aus „survivors") ausgewählt werden, gilt dieser Zusammenhang nur bei Krankheiten mit kleinen Inzidenzen („rare disease assumption").

c) Überschätzung des RR in der Kohortenstudie durch das OR
Unabhängig vom Studientyp kann ein OR für den Vergleich von Krankheitsrisiken berechnet werden, wobei in Kohortenstudien (sowie in Querschnittstudien) der Wert des OR erheblich größer sein kann als jener des zugrundeliegenden RR, wenn die Krankheit oder Beschwerde häufig (>10 Prozent) auftritt! Bei einer Häufigkeit von z.B. 20 Prozent entspricht einem OR von 2,7 ein RR von zwei. Dies führt leicht zu einer Überschätzung der tatsächlichen Effekte einerseits durch Unkenntnis oder andererseits durch fälschliche Interpretation eines OR als RR.

Isolierte Risikofaktoren

Risikofaktoren treten im Alltagsleben vielfach gemeinsam auf. Möchte man z.B. den Gesundheitseffekt von Bewegungsmangel feststellen, indem eine Gruppe aktiver mit einer Gruppe weniger aktiver Personen verglichen wird, so finden sich in der zweiten Gruppe oft auch vermehrt Personen mit Fehlernährung. Dies führt zu einer Überschätzung des Effekts des Bewegungsmangels auf die untersuchte Krankheit. Eine solche Konstellation wird als Confounding bezeichnet, das durch die Störgröße (=Confounder) „Fehlernährung" entsteht. Confounding führt zu einer Über- oder Unterschätzung des Effekts, den man dem untersuchten Risikofaktor zuschreiben würde, wenn er „allein" wirksam wäre. Um den Gesundheitseffekt eines einzelnen Faktors beurteilen zu können, muss dieser also von der Wirkung der weiteren korrelierenden Faktoren „abgetrennt" bzw. isoliert werden.
Man kann sich leicht überlegen, dass in praktisch allen epidemiologischen Fragestellungen Störgrößen beteiligt sind. Fast ausnahmslos wirken das Alter und Geschlecht einer Person als Confounder, sodass in der Epidemiologie diese Größen immer zu berücksichtigen sind.
Beispiel: Aufgrund mangelhaft durchgeführter Beobachtungsstudien bestand die fälschliche Annahme, dass die Hormonersatztherapie bei postmenopausalen Frauen zu einer Reduktion des Herzinfarktrisikos führt. Die Einbeziehung des Sozialstatus der Frauen (= vergessener Confounder) zeigte, dass ein Confounding-Effekt vorlag, da offenbar vermehrt die gesünderen Frauen der oberen Sozialschichten therapiert wurden.

Wirksame und unwirksame Risikofaktoren

Eine Reihe großer Kohortenstudien zeigte starke positive Gesundheitseffekte der Vitaminaufnahme über die Ernährung. Bei der experimentellen Überprüfung mittels Verabreichung von Vitaminpräparaten in randomisierten Studien waren diese präventiven Effekte nicht mehr nachweisbar. Wie lässt sich der Widerspruch erklären?
Allein aus Beobachtungsstudien kann nicht auf die kausale Wirksamkeit eines Risikofaktors geschlossen werden! Das entscheidende Kriterium zum Nachweis der instrumentellen Kausalität ist die experimentelle Überprüfung, d.h. die Beobachtung des Krankheitsrisikos unter systematischer Beeinflussung des untersuchten Faktors. Da die Epidemiologie überwiegend gesundheitsschädigende Faktoren untersucht, ist anders als in der klinischen Forschung eine experimentelle Studie oftmals aus ethischen und praktischen Gründen undurchführbar. Man denke hier an typische Expositionen wie Tabakrauch oder Mobiltelefonie. Hier muss man sich beim Menschen auf Beobachtungsstudien beschränken, was einen direkten Nachweis einer kausalen Wirkung vereitelt.

Messfehler in der Risiko- und Effektmessung

Bisher gingen wir davon aus, dass gemessene Krankheitsrisiken und Effekte wirklich so sind, wie sie gemessen werden. Es weisen aber schon teure und geeichte physikalische oder chemische Messgeräte in der Praxis mitunter beträchtliche Messfehler auf, umso mehr darf man dies von den Messinstrumenten der Epidemiologie erwarten.
Jede Angabe eines Krankheitsrisikos oder eines relativen Risikos muss als Messergebnis verstanden werden, das einer Fehlerwahrscheinlichkeit unterliegt! Der Fehlerbereich in Form eines 95-Prozent-Konfidenzintervalls (oder eines p-Werts bei Existenzaussagen) sollte daher nie fehlen. Die Gültigkeit von p-Werten und Konfidenzintervallen ist allerdings nur unter der Voraussetzung gegeben, dass das Messinstrument richtig funktioniert, d.h. die Studie korrekt durchgeführt wurde! Ein systematischer Fehler in der Planung oder in der Durchführung einer Studie, ein sogenannter Bias, wird durch diese Fehlerabschätzungen nicht erfasst.

Univ.-Prof. Dr. Willibald Stronegger
Institut für Sozialmedizin und Epidemiologie, Medizinische Universität Graz

FOTO: PRIVAT

Das Lesen einer Studie – Qualitätsmerkmale analysieren

VON DR. CHRISTOPH GRIMM

Abstract

Sieht man sich mit einer Arbeit konfrontiert, der man bei all dem mathematischen Gewirr möglichst viel praktischen Wert entlocken will, bietet sich zunächst der Abstract an. Er gibt einen kurzen Überblick und sollte im Idealfall folgende Punkte abhandeln:

- Ein oder zwei Sätze, die die Fragestellung der Arbeit skizzieren und idealerweise eine klar formulierte primäre Fragestellung definieren.
- Die Methodik, mit der diese Frage im Rahmen der Studie zu beantworten versucht wurde, d.h. Informationen über Art der Studie, Anzahl der teilnehmenden Zentren, Fallzahl, Patientenkollektiv und statistischen Auswertung.
- Die Ergebnisse geordnet nach ihrer Wichtigkeit: Als Erstes sollte die primäre Fragestellung beantwortet werden, danach die sekundären Fragestellungen – nicht, wie in den meisten Fällen, das beste Ergebnis einer sekundären Fragestellung zuerst und das leider negative Ergebnis der primären Frage zuletzt oder gar nicht.
- Ein oder zwei abschließende Sätze, die die primäre Fragestellung klar beantworten und deren klinische Bedeutung diskutieren.

Definition der primären Fragestellung und der Endpunkte

Der Abstract gibt nur erste Einblicke in eine Arbeit, daher sollte sich der nächste Blick nicht auf die Einleitung, sondern auf den methodischen Teil der Arbeit richten. Ganz zentral ist die klare Definition der primären Fragestellung und gegebenenfalls der sekundären Fragestellungen. Sämtliche Fallzahlberechnungen, die Patientenrekrutierung und Planungen des Studiendesigns beziehen sich auf die primäre Fragestellung mit dem primären Endpunkt. **Es ist ganz entscheidend wie der primäre Endpunkt definiert ist;** einer der am besten definierten Endpunkte ist der Tod. Wesentlich schwieriger zu definieren ist z.B. der Zeitpunkt des Rezidivs einer Tumorerkrankung (Datum der histologischen Verifizierung, Datum des Tumormarkeranstiegs, Datum der ersten Symptome, …). Hierbei bedient man sich oft „weicher Endpunkte", sogenannter Surrogatparameter wie z.B. des Tumormarkeranstiegs als Hinweis für ein Tumorrezidiv. Je unklarer der Endpunkt, umso schwieriger die statistische Auswertung. Sekundäre Endpunkte können definiert sein, um Nebenfragestellungen zu beantworten. **Als Faustregel sollten allerdings nicht mehr als fünf sekundäre Fragestellungen untersucht werden,** da dies die statistische Aussagekraft der Ergebnisse beeinflussen könnte.

Definition des Patientenkollektivs

Eine Gratwanderung stellt üblicherweise die Definition des Patientenkollektivs dar. Neben der Auswahl des korrekten Kollektivs (bei Männern fällt die Evaluierung einer Hormonersatztherapie, die für Frauen gedacht ist, schwer), gilt es hier insbesondere auf die Argumentation der Ein- bzw. Ausschlusskriterien zu achten. Untersucht eine Studie beispielsweise den Einfluss einer ganz exakten Genmutation auf das Tumorwachstum des Prostatakarzinoms, sollte versucht werden, ein möglichst homogenes Kollektiv aus Patienten mit sehr ähnlichem Risikoprofil zu inkludieren (nur 60- bis 70-jährige Männer, Nichtraucher, negative Familienanamnese, keine Zweitmalignome, …). Dasselbe Kollektiv wäre natürlich nicht repräsentativ für die Evaluierung der Wirksamkeit eines Kopfschmerzmittels, das sowohl bei Jüngeren als auch Älteren sowie Frauen und Männer wirksam sein sollte. **Die Auswahl der Studienteilnehmer ist ganz entscheidend für die Aussagekraft einer Studie und muss daher klar definiert und argumentiert sein.**

Die erste Tabelle

Die erste Tabelle stellt üblicherweise die Fälle den Kontrollen gegenüber und zeigt, ob diese zwei Gruppen überhaupt miteinander vergleichbar sind oder ob die statistischen Auswertungen aufgrund einer ungleichen Verteilung von Risikofaktoren (Alter, Geschlecht, Rauchen, Familienanamnese, sozioökonomischer Status, …) verzerrt sind. Fallen hier Ungleichheiten auf, die nicht in die statistischen Analysen und bei der Interpretation der Ergebnisse einbezogen werden, kann bereits auf das Weiterlesen der Arbeit verzichtet werden.

Kreative Analyse, kreative Präsentation

Bei der Beurteilung der durchgeführten statistischen Analysen sollte man anfänglich überprüfen, um welchen Datentyp es sich bei den untersuchten Werten handelt, da sich danach die adäquaten Tests richten. Man unterscheidet

- metrische (Alter, Körpergewicht, Körpergröße),
- binäre (ja/nein, schwarz/weiß, männlich/weiblich) und
- kategoriale (Schulnoten, Tumorstadium) Daten.

Einerseits scheint es eine unüberschaubare Anzahl an statistischen Tests zu geben, andererseits kommen die meisten Arbeiten mit ungefähr einem Dutzend dieser Tests aus. **Aufpassen sollte man bei Daten, die auf gewöhnliche Weise gesammelt, aber außergewöhnlichen statistischen Tests unterzogen wurden.** Hier lohnt es sich, genau auf die Argumentation der Verfasser für

Klassische Formulierungsfallen im Rahmen der Methodik einer Arbeit

Getätigte Formulierung	Korrekte Formulierung	Falle
„Wir untersuchten, wie oft Frauenärzte nach familiärer Brustkrebsbelastung fragten“	„Wir durchsuchten Krankengeschichten und zählten, in wie vielen die Familienanamnese bezüglich Brustkrebs dokumentiert wurde“	Annahme, die Patientenaufzeichnungen wären 100-prozentig korrekt
„Wir untersuchten, wie Frauenärzte chronische Unterbauchschmerzen behandeln“	„Wir untersuchten die Angaben von Ärzten, wie sie chronische Unterbauchschmerzen therapieren“	Annahme: Arzt tut immer das, was er angibt zu tun
„Wir verglichen 20 Frauen und 20 Kontrollen“	„Folgende Parameter ... wurden in 20 kaukasischen postmenopausalen Frauen und 20 asiatischen postmenopausalen Frauen mittels t-Test miteinander verglichen“	Mangelnde Angaben über die Studienobjekte
„Wir verglichen eine antihormonelle Therapie mit Plazebo“	„Patientinnen in der Verumgruppe erhielten ... Tablette 1x1 20mg p.o. täglich über 20 Tage. Patientinnen in der Kontrollgruppe erhielten eine optisch idente Tablette (nur die Trägerstoffe enthaltend) 1x1 20mg p.o. täglich über 20 Tage“	Mangelnde Angaben über die Therapie bzw. Intervention

die Auswahl des statistischen Tests und den angegebenen Literaturhinweis zu achten.

Der schärfste Blick sollte darauf gerichtet sein, ob die Daten dem ursprünglichen Studienprotokoll bzw. -design entsprechend analysiert und diskutiert werden. **Die primäre Fragestellung sollte dabei im zentralen Fokus der Arbeit stehen** und nicht „interessante Ergebnisse“ aus sogenannten Subgruppenanalysen (Analysen sekundärer Fragestellungen), die überproportional diskutiert werden. Oft besteht hierbei nämlich das Problem des „multiple testing“ – führt man eine hohe Anzahl an Analysen durch, steigt die Wahrscheinlichkeit, ein zufälliges positives Ergebnis zu entdecken.

Ausreißer

Ein ganz zentrales Qualitätsmerkmal einer Arbeit stellt der Umgang mit statistischen Ausreißern bzw. unerwarteten Ergebnissen dar. Unerwartete Ergebnisse können durch den Patienten (z.B. abnormaler Stoffwechsel), Messfehler (fehlerhaftes Messinstrument, Untersucherfehler), Interpretationsfehler (Ablesefehler) oder Analysefehler (fehlerhafte Positionierung der Dezimalstelle) verursacht werden. Einerseits sollten Ausreißer mittels adäquater statistischer Analysen evaluiert werden. Andererseits sollten **unerwartete oder unlogisch erscheinende Ergebnisse adäquat diskutiert und nicht nur kommentarlos präsentiert werden.**

Die Einleitung

Hat man sich von der Qualität der Arbeit vergewissert, die adäquate Fragestellung, das korrekte Design, die richtig gewählten statistischen Tests, das korrekte Präsentieren der Ergebnisse und das adäquate Analysieren und Diskutieren von Ausreißern überprüft, kann man sich nun mit voller Konzentration der Einleitung der Arbeit widmen.

Dr. Christoph Grimm
Abteilung für allgemeine Gynäkologie und gynäkologische Onkologie, Universitätsklinik für Frauenheilkunde, Wien

FOTO: PRIVAT

Sensitivität und Spezifität

VON UNIV.-PROF. DR. ANTON STIFT

Die beiden Begriffe Sensitivität und Spezifität finden sich vor allem in Arbeiten, bei denen diagnostische Methoden zum Einsatz kommen, wobei deren Treffsicherheit beschrieben werden soll. Ziel eines jeden diagnostischen Verfahrens ist es festzustellen, ob tatsächlich eine Erkrankung vorliegt oder nicht. Yerushalmy hat die Begriffe der Sensitivität und Spezifität bereits im Jahre 1947 eingeführt, um ein Maß für die Aussagekraft eines Testverfahrens zu etablieren.

Sensitivität

Die Sensitivität eines Testverfahrens hinsichtlich einer vorliegenden Erkrankung ist der Anteil der positiv getesteten Personen von allen getesteten Personen, die die betreffende Erkrankung haben. Die Sensitivität beschreibt daher die bedingte Wahrscheinlichkeit eines positiven Testergebnisses, falls die Krankheit vorliegt. Addiert man die falsch negative Rate zum Wert der Sensitivität, erhält man 100 Prozent. Die Sensitivität wird deshalb auch Trefferquote genannt.

Spezifität

Die Spezifität eines Testverfahrens hinsichtlich einer Erkrankung ist der Anteil der negativ getesteten Personen von allen getesteten Personen, die die betreffende Erkrankung nicht haben. Die Spezifität beschreibt daher die bedingte Wahrscheinlichkeit eines negativen Testergebnisses, wenn die Krankheit nicht vorliegt. Addiert man die falsch positive Rate zum Wert der Spezifität, erhält man 100 Prozent.

Ein Beispiel aus der Praxis

In einer im September 2008 publizierten Arbeit untersuchten Wissenschafter die Genauigkeit der CT-Kolonografie hinsichtlich der Detektion von großen Adenomen bzw. Karzinomen im Dickdarm (NEJM 2008; 359 (12): 1207–17). Die Aussagekraft (Validität) eines derartigen diagnostischen Tests bezieht sich dabei auf dessen Fähigkeit, eine Läsion bei Patienten, die tatsächlich einen Tumor haben, zu erkennen und eben keine Läsion zu entdecken bei Personen, die tatsächlich keinen Tumor haben. Ein perfektes Testverfahren sollte beide Ergebnisse entsprechend abbilden, also bei Patienten mit Tumor diesen auch erkennen und bei jenen ohne Tumor auch keinen auszuweisen.

Leider stehen uns derartig perfekte Testverfahren nur selten zu Verfügung. Um festzustellen, wie gut eine Methode eine Pathologie feststellen kann, brauchen wir ein Verfahren, um der Wahrheit – ob bei diesem Patienten eine gewisse Erkrankung vorliegt oder nicht – möglichst nahezukommen. Um die Methode der CT-Kolonografie auf deren Treffsicherheit hinsichtlich Tumordetektion zu überprüfen, ist die Gewissheit nötig, dass bei einer Gruppe von Patienten tatsächlich ein Tumor vorliegt. In der oben genannten Studie wurde dies mittels Kolonoskopie überprüft. Diese sichere Information kann nun mit der „neuen" Methode verglichen werden. Die Methode, mit der das Vorhandensein einer Erkrankung sicher nachgewiesen werden kann, wird auch als Goldstandard bezeichnet und sollte den derzeit besten verfügbaren Untersuchungsgang darstellen.

Um die Sensitivität und Spezifität genau berechnen zu können, ist es erforderlich, das tatsächliche Vorhandensein bzw. Fehlen einer bestimmten Erkrankung zu kennen.

Berechnung von Sensitivität und Spezifität

Die Sensitivität ist definiert als der Anteil all jener mit der Erkrankung, die ein auch positives Testergebnis hatten, und ist daher ein Maß dafür, wie gut sich ein Testverfahren eignet, um eine bestimmte Erkrankung zu detektieren, wenn sie tatsächlich vorhanden ist.

Die Berechnung lässt sich daher wie folgt durchführen: Sensitivität in Prozent = WP/WP + FN x 100.

Die Spezifität ist definiert als der Anteil jener, die die Erkrankung nicht haben und auch ein negatives Testergebnis ausweisen. Die Spezifität ist daher ein Maß, wie exakt die Methode negative Ergebnisse ergibt, wenn die Erkrankung tatsächlich nicht vorliegt.

Die Berechnung lässt sich daher wie folgt durchführen:

Spezifität in Prozent = $\frac{WN}{FP+WN}$ x 100.

→ Goldstandard vs. neue Methode

Testergebnis neue Methode	Goldstandard: Krankheit vorhanden: ja	Goldstandard: Krankheit vorhanden: nein	
Positiv	WP	FP	WP+FP
Negativ	FN	WN	FN+WN
	WP+FN	FP+WN	

WP = richtig Positive, WN = falsch Positive, FP = falsch Positive, FN = falsch Negative

In der oben angeführten Screening-Studie ergab die CT-Kolonografie eine Sensitivität von 90 Prozent bei asymptomatischen Patienten, bei denen Läsionen mit einem Durchmesser von über 10mm detektiert werden konnten.

Zusammenfassung

Sensitivität und Spezifität sind Maßstäbe, die die Aussagekraft einer bestimmten Untersuchung darstellen. Je höher der Wert

für die Sensitivität ist, desto niedriger ist die falsch negative Rate, d.h. umso niedriger ist die Chance, eine vorhandene Erkrankung nicht zu erfassen. Je höher der Wert für die Spezifität ist, desto niedriger sind falsch positive Ergebnisse und damit die Gefahr, eine Erkrankung zu erfassen, die gar nicht vorhanden ist.

Univ.-Prof. Dr. Anton Stift
Abteilung für Allgemeinchirurgie, Universitätsklinik für Chirurgie, Wien

Kritische Bewertung von Studienergebnissen: Evidenzlevel

VON UNIV.-PROF. DR. LUKAS HEFLER UND UNIV.-PROF. DR. ALEXANDER REINTHALLER

Im Zeitalter der „evidence based medicine" (EBM) ist die richtige Interpretation von Studienergebnissen unerlässlich, um daraus die richtigen Schlussfolgerungen für die Diagnose und Behandlung von PatientInnen ziehen zu können.
Evidenzlevel sind in der aktuellen medizinischen Literatur oft gebrauchte Schlagworte: Keine Leitlinie kommt ohne die Erwähnung derselben aus. In medizinischen Fachdiskussionen werden neuen Publikationen sofort die entsprechenden Evidenzlevel zugeteilt. Die Vorteile der Verwendung von Evidenzlevel sind offensichtlich: Neue wissenschaftliche Daten sollen nach ihrer Qualität bewertet werden, um diesen die richtige „Bedeutung" beimessen zu können. Davon abgeleitet können nach Kriterien der EBM Empfehlungen für die klinische Praxis erstellt werden, die neuerlich ähnlich einem Schulnotensystem bewertet werden.
Man könnte den Eindruck haben, die Zuteilung eines Evidenzlevels zu einer Studie ist einfach. Betrachtet man die Thematik der Evidenzlevel jedoch genauer, merkt man sehr bald, dass deren Gebrauch auch einige Nachteile hat. Weltweit werden höchst uneinheitliche Definitionen verwendet. Ein und dieselbe Studie wird von unterschiedlichen Experten oft anders bewertet. Für viele Fragestellungen können keine bzw. werden nie Daten mit einem ausreichend hohen Evidenzlevel vorliegen, sei es aufgrund der Seltenheit einer Erkrankung oder der (ethischen oder rechtlichen) Unmöglichkeit, geeignete Studien durchzuführen.

Evidenzlevel in den USA und Großbritannien

Im Rahmen des vorliegenden Artikels sollen einige Definitionen von Evidenzlevel vorgestellt werden. Eine der renommiertesten „Rating-Agenturen" ist die U.S. Preventives Services Task Force (USPSTF). Diese Rating-Agentur verwendet ein dreistufiges Schema, um die Qualität der Evidenz von wissenschaftlichen Studien zu bewerten (gut, mittel, schwach): Um durch eine Übersetzung weder Definitionen noch Inhalte zu verfälschen, erlauben wir uns, die Definitionen im Originalwortlaut auf Englisch wiederzugeben.

USPSTF grading scheme for quality of evidence
Good: Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes.
Fair: Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, or consistency of the individual studies, generalizability to routine practice, or indirect nature of the evidence on health outcomes.
Poor: Evidence is insufficient to assess the effects on health outcomes because of limited number or power of studies, important flaws in their design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.
Basierend auf diesem Evidenzlevel hat die USPSTF ein Schema für die „Stärke" bzw. die Validität der von ihr erarbeiteten Empfehlungen vorgestellt. Damit kann der „Netto-Vorteil" einer Intervention für die PatientInnen besser eingestuft werden:

USPSTF grading scheme for recommendations
A. The USPSTF strongly recommends that clinicians provide [the service] to eligible patients. The USPSTF found good evidence that [the service] improves important health outcomes and concludes that benefits substantially outweigh harms.
B. The USPSTF recommends that clinicians provide [this service] to eligible patients. The USPSTF found at least fair evidence that [the service] improves important health outcomes and concludes that benefits outweigh harms.
C. The USPSTF makes no recommendation for or against routine provision of [the service]. The USPSTF found at least fair evidence that [the service] can improve health outcomes but concludes that the balance of benefits and harms is too close to justify a general recommendation.
D. The USPSTF recommends against routinely providing [the service] to asymptomatic patients. The USPSTF found at least fair evidence that [the service] is ineffective or that harms outweigh benefits.

I. The USPSTF concludes that the evidence is insufficient to recommend for or against routinely providing [the service]. Evidence that the [service] is effective is lacking, of poor quality, or conflicting and the balance of benefits and harms cannot be determined.
Quelle: U.S. Preventive Services Task Force (USPSTF)

Eine der angesehensten Zentren auf dem Gebiert der EBM in Europa ist sicher das Oxford Centre for Evidence-based Medicine. Dieses Zentrum hat kürzlich (März 2009) Richtlinien für die Beurteilung der Evidenz für unterschiedliche Fragestellungen (Therapie/Prävention/Ätiologie, Prognose, Diagnose, Differenzialdiagnose und ökonomische Entscheidungen) erarbeitet.
Diesen sehr umfangreichen Richtlinienkatalog, der die Dimensionen des Artikels sprengen würde, sowie die Klassifizierung klinischer Empfehlungen finden Sie unter
http://www.cebm.net/index.aspx?o=1025

Evidenzlevel in Kanada und dem deutschsprachigen Raum

Eine andere viel verwendete Klassifikation der Evidenzlevel ist die Einteilung der Canadian Task Force (CTF) on Preventive Health Care. Die CTF sieht sich als Verbindungsglied zwischen Forschung und der angewandten klinischen präventiven Medizin. Die erarbeiteten Empfehlungen sollen ÄrztInnen in der klinischen Praxis helfen, wirkungsvolle von unwirksamen Maßnahmen unter bestmöglichem Ausschluss von „bias" unterscheiden zu können. Neuerlich wird zwischen „Level of Evidence" und Stärke der Empfehlungen unterschieden.

Die kanadischen Evidenzlevel werden vielfach auch in deutschsprachigen Ländern verwendet und liegen daher auch auf Deutsch vor:
Grad I: Die Evidenz ist aufgrund randomisierter kontrollierter Studien (oder Metaanalysen) von genügendem Umfang derart, dass die Gefahr, dass sie falsch positive oder falsch negative Resultate beinhalten, gering ist.
Grad II: Die Evidenz basiert auf randomisierten, kontrollierten Studien, welche jedoch zu klein sind, um ihnen Grad I zuzusprechen; sie können positive Trends, welche jedoch statistisch nicht signifikant sind, oder gar keine Trends zeigen. Sie sind mit einem hohen Risiko falsch negativer Resultate verbunden.
Grad III: Die Evidenz basiert auf nicht randomisierten Kontroll- oder Kohortenstudien, Fallserien, Fallkontrollstudien oder Querschnittstudien.
Grad IV: Die Evidenz basiert auf der Meinung angesehener Experten oder Expertengremien, wie sie in publizierten Konsensus-Konferenzen oder in Guidelines angegeben werden.
Grad V: Die Evidenz basiert auf der Meinung derjenigen Personen, welche diese Guidelines geschrieben oder aktualisiert haben, beruhend auf ihrer Erfahrung, ihrer Kenntnis der einschlägigen Literatur und der Diskussion mit ihren Fachkollegen.

Die nachfolgende Tabelle stellt die Grade der Empfehlungen durch die CTF dar:
Recommendations Grades for Specific Clinical Preventive Actions
A. The CTF concludes that there is good evidence to recommend the clinical preventive action.
B. The CTF concludes that there is fair evidence to recommend the clinical preventive action.
C. The CTF concludes that the existing evidence is conflicting and does not allow making a recommendation for or against use of the clinical preventive action, however other factors may influence decision-making.
D. The CTF concludes that there is fair evidence to recommend against the clinical preventive action.
E. The CTF concludes that there is good evidence to recommend against the clinical preventive action.
I. The CTF concludes that there is insufficient evidence (in quantity and/or quality) to make a recommendation, however other factors may influence decision-making.
Quelle: Canadian Task Force on Preventive Health Care

Die „natürliche Grenze" der EBM

Eine Vielzahl von Interventionen sind in der täglichen Praxis integriert, ohne dass sie durch adäquate wissenschaftliche Evidenz untermauert wurden. Ein gutes Beispiel dafür ist sicherlich der zytologische Abstrich von der Portio uteri im Rahmen des Zervixkarzinom-Screenings. Der Nutzen dieser Intervention ist niemals durch eine prospektiv-randomisierte Studie nachgewiesen worden bzw. wird aus ethischen Gründen eine solche Studie auch niemals durchgeführt werden. Trotzdem ist der Nutzen dieser Intervention unbestritten. Hier stößt die Verwendung der EBM bzw. von Evidenzlevel scheinbar auf eine „natürliche Grenze". EBM soll jedoch nur die klinisch tätige ÄrztIn unterstützen und erhebt nicht den Anspruch, für sich alleine zu stehen. Vielmehr bedeutet EBM die Integration individueller klinischer Expertise mit der best verfügbaren externen Evidenz aus systematischer Forschung.

Zusammenfassung

- Die Bewertung von wissenschaftlichen Studien nach Evidenzlevel ist schwierig und erfordert Erfahrung und Übung.
- Da die Definitionen der Evidenzlevel sehr heterogen sind, sollten diese bei jeder Bewertung angegeben werden.
- Einmal festgelegt, sind Evidenzlevel bei richtiger Verwendung hilfreich in der Beurteilung, Interpretation und klinischen Umsetzung von neuen Erkenntnissen.

Univ.-Prof. Dr. Lukas Hefler

Univ.-Prof. Dr. Alexander Reinthaller
Abteilung für allgemeine Gynäkologie und gynäkologische Onkologie, Universitätsklinik für Frauenheilkunde, Wien

FOTOS: BARBARA KROBATH

Glossar

Absolute Risikoreduktion (**ARR**) beschreibt die absolute Differenz der Rate an ungünstigen Ereignissen in der experimentellen Gruppe (E) im Vergleich zur Kontrollgruppe (K), wenn die experimentelle Behandlung wirksam ist (ARR=K–E). Der Kehrwert der ARR ergibt die Number Needed to Treat (1/ARR=NNT).

Absolute Risikozunahme (**ARI, absolute risk increase**) beschreibt die absolute Differenz der Rate an ungünstigen Ereignissen in der experimentellen Gruppe (E) im Vergleich zur Kontrollgruppe (K), wenn die experimentelle Behandlung schlechter ist (ARI = |K–E|). Der Kehrwert der ARI ergibt die Number Needed to Harm (1/ARI=NNH).

Attrition. Der Verlust von Teilnehmern während der Studiendurchführung. Auch loss to follow up oder drop out genannt.

Bias (**systematischer Fehler**) ist die Tendenz der Studienergebnisse, systematisch von den „wahren" Ergebnissen abzuweichen. Bias führt entweder zu einer Über- oder Unterschätzung der wahren Wirkung einer Maßnahme oder Exposition. Die Ursachen dafür liegen vor allem im Design und der Durchführung der Studie und führen zu systematischen Unterschieden zwischen den Vergleichsgruppen, z.B. bei der Auswahl der Teilnehmer (Selektionsbias), der Erhebung der Endpunkte (measurement bias oder Messungsbias) oder dem Verlust von Teilnehmern in der Studie (attrition bias oder Verschleiß-Bias). Ergebnisse aus Studien mit geringem Risiko für Bias werden als valide angesehen.

Blindversuch. Wenn der Therapieerfolg oder die Beurteilung des Therapieerfolgs von der Erwartungshaltung des Patienten oder des Arztes abhängig ist, werden zur Ausschaltung eines systematischen Fehlers in einem einfachen Blindversuch der Patient oder in einem Doppelblindversuch Patient und Arzt über die im Einzelfall angewandte Therapie im Unklaren gelassen. Blindversuche sind ein wesentliches Instrument, um Beobachtungsgleichheit und Behandlungsgleichheit zu erreichen.

Confounding liegt vor, wenn ein Faktor, der nicht direkt Gegenstand der Untersuchung ist, sowohl mit der Intervention/Exposition als auch mit der Zielgröße assoziiert ist und dadurch „Verwirrung" stiftet. Häufige Confounder sind z.B. Alter, Geschlecht oder Nikotingenuss. Confounding lässt sich durch ein entsprechendes Studiendesign (z.B. Randomisierung oder Matching) oder durch die Anwendung bestimmter statistischer Verfahren bei der Analyse (Stratifizierung, multivariate Analyse) kontrollieren.

Control event rate (**CER; Ereignisrate in der Kontrollgruppe**). Anteil der Teilnehmer in der Kontrollgruppe, die in einem definierten Zeitraum ein Ereignis oder einen Endpunkt erleiden. Die Ereignisrate in der Kontrollgruppe wird zur Berechnung der absoluten Risikoreduktion und relativen Risikoreduktion benötigt.

Cross-over-Design. Studiendesign, in dem die zu vergleichenden Interventionen in den Vergleichsgruppen in zeitlicher Folge angewandt werden. Dabei erhält z.B. die eine Gruppe zunächst Therapie A, dann Therapie B, die andere Gruppe zuerst Therapie B und dann Therapie A.

Diskrete und dichotome Variable. Diskrete Variablen weisen im Gegensatz zu kontinuierlichen Variablen nur eine begrenzte Zahl eindeutig voneinander abgrenzbarer Zuständen auf (z.B. Augenfarbe: blau, grau, braun, grün). Eine Sonderform sind dichotome Variablen, die lediglich zwei Alternativen aufweisen, z.B. Raucher/Nichtraucher, lebend oder tot, Test-positiv oder -negativ.

Effektmaß. Maßzahl, um die Stärke eines Effekts zu quantifizieren. Gebräuchliche Effektmaße für dichotome Endpunkte sind das relative Risiko (RR) oder die Odds Ratio (OR), gebräuchliche Effektmaße für kontinuierliche Endpunkte sind in Einzelstudien die standardisierte mittlere Differenz (SMD) und in Metaanalysen die gewichtete mittlere Differenz (weighted mean difference, WMD).

Effectiveness (**Wirksamkeit unter Alltagsbedingungen**). Im Gegensatz zur Efficacy untersuchen Effectiveness-Studien die Frage: Wirkt die Maßnahme unter den Bedingungen der Routineversorgung?

Efficacy (**Wirksamkeit unter Idealbedingungen**). Efficacy-Studien zeichnen sich durch hohe innere Validität aus, die Ergebnisse sind jedoch möglicherweise nur bedingt auf die Routineversorgung übertragbar.

Erwartungstreue. Eine Schätzung eines unbekannten Parameters heißt erwartungstreu, wenn für alle Stichprobenumfänge und alle Werte des Parameters ihr Erwartungswert gleich dem Parameter ist. Die Abweichung des Erwartungswerts der Schätzung vom gesuchten Parameterwert heißt Bias (Verzerrung). Die Erwartungstreue ist eine wünschenswerte Eigenschaft einer optimalen Schätzung.

Evidenz (**evidence**). Der Begriff bezieht sich auf die Informationen aus klinischen Studien, die einen Sachverhalt erhärten oder widerlegen.

Experimental event rate (**EER; Ereignisrate in der experimentellen Gruppe**). Anteil der Teilnehmer in der experimentel-

len Gruppe einer klinischen Studie, die in einem definierten Zeitraum ein Ereignis oder einen Endpunkt erleiden. Die Ereignisrate (Risiko) in der experimentellen Gruppe wird zur Berechnung der absoluten Risikoreduktion und relativen Risikoreduktion benötigt.

Fall-Kontroll-Studie. Darunter versteht man eine retrospektive Erhebung. Aus einer definierten Grundgesamtheit wird eine Stichprobe von Personen mit der interessierenden Erkrankung (Fälle) gezogen. Aus der gleichen Grundgesamtheit wird eine Stichprobe von Personen ohne diese Erkrankung (Kontrollen) gezogen. Die Exposition in der Vergangenheit gegenüber potenziellen Risikofaktoren wird ermittelt. Das wichtigste Risikomaß in Fall-Kontroll-Studien ist die Odds Ratio.

Heterogenität/Homogenität. In systematischen Reviews oder Metaanalysen bezeichnet Homogenität (Heterogenität), inwieweit die in den eingeschlossenen Studien gefundenen Effekte ähnlich (homogen) oder verschieden (heterogen) sind. Mit statistischen Heterogenitätstests kann festgestellt werden, ob die Unterschiede zwischen den Studien größer sind, als zufallsbedingt zu erwarten wäre. Als Ursachen für Heterogenität kommen Unterschiede in den Patientencharakteristika, Intervention oder Endpunkte zwischen den Studien infrage, was aus klinischer Sicht beurteilt werden muss. Die Durchführung einer Metaanalyse aus heterogenen Studien ist problematisch.

Inferenz. Der Vorgang des Schließens: Von einer Stichprobe ausgehend wird mittels statistischer Hypothesentests auf Eigenschaften der gesamten Grundgesamtheit geschlossen. Wesenszug der Inferenzstatistik ist also die Überprüfung von Hypothesen.

Intention-to-treat-Analyse. Analysetechnik, bei der die Patienten nach ihrer ursprünglichen Gruppenzuteilung analysiert werden, unabhängig davon, ob sie die zugeordnete (intendierte) Therapieform vollständig, partiell oder gar nicht erhalten haben.

Inzidenz beschreibt die in einem bestimmten Zeitraum neu aufgetretene Anzahl an Krankheitsfällen in einer definierten Population.

Klinische Studie. Unscharf definierter Begriff für eine Studie, in der eine Intervention an einer Gruppe von Patienten untersucht wird. Oberbegriff für unterschiedliche Studientypen, z.B. nicht kontrollierte, kontrollierte und randomisierte klinische Studien.

Kohortenstudie. Vergleichende Beobachtungsstudie, in der Personen (Kohorte) mit bzw. ohne eine Intervention/Exposition (zu der sie nicht von dem Studienarzt zugeteilt wurden) über einen definierten Zeitraum beobachtet werden, um Unterschiede im Auftreten der Zielerkrankung festzustellen. Kohortenstudien können prospektiv oder retrospektiv durchgeführt werden.

Konfidenzintervall (Vertrauensbereich, confidence interval – CI). Bereich, in dem der „wahre" Wert einer Messung (Effektgröße) mit einer bestimmten Wahrscheinlichkeit erwartet werden kann (üblicherweise 95-%-Konfidenzintervall). Das Konfidenzintervall beschreibt die Unsicherheit über die Zuverlässigkeit der Aussage zur Effektgröße. Die Breite des Konfidenzintervalls hängt u.a. von der Zahl der in die Studie eingeschlossenen Patienten ab und wird mit zunehmender Patientenzahl enger, d.h. die Effektgröße kann präziser geschätzt werden.

Korrelation misst den Zusammenhang zwischen zwei quantitativen Merkmalen. Eine Maßzahl für die Stärke der Korrelation ist der Korrelationskoeffizient.

Log-Rank-Test. Test zum Vergleich zweier Überlebenszeitkurven, die sich jedoch nicht überschneiden dürfen. Mathematisch gesehen ein gewöhnlicher „k x 2 Felder Chi Quadrat"-Test (zwei Kurven und jeweils k Messstellen). Man vergleicht zu den gegebenen Zeitpunkten (k) die erwarteten Häufigkeiten (die eine Kurve) und gemessenen Häufigkeiten (die andere Kurve).

Metaanalyse. Statistisches Verfahren, um die Ergebnisse mehrerer Studien, die die gleiche Frage bearbeiten, quantitativ zu einem Gesamtergebnis zusammenzufassen und dadurch die Aussagekraft (Genauigkeit der Effektschätzer) gegenüber Einzelstudien zu erhöhen. Metaanalysen werden mit zunehmender Häufigkeit in systematischen Reviews eingesetzt. Allerdings beruht nicht jede Metaanalyse auf einem systematischen Review.

Nullhypothese. Bei der Durchführung statistischer Signifikanztests wird üblicherweise die Hypothese aufgestellt, dass zwischen den verschiedenen Gruppen einer Studie kein Unterschied (Nullhypothese) besteht. Aus statistischer Sicht ist eine Maßnahme wirksam, wenn man durch den statistischen Test die Nullhypothese, dass es zwischen den Ergebnissen der experimentellen und der Kontrollgruppe keinen Unterschied gibt, verwerfen kann (s.a. statistische Signifikanz).

Number Needed to Treat (NNT). Klinisch intuitives Effektmaß für dichotome Endpunkte, um die Auswirkung einer Behandlung zu beschreiben. Sie gibt die Anzahl an Patienten wieder, die behandelt werden müssen, um ein zusätzliches ungünstiges Ereignis zu verhindern. Die NNT wird als 1/ARR (s. absolute Risikoreduktion) berechnet.

Odds. Das Verhältnis der Wahrscheinlichkeit p, dass ein Ereignis eintritt, zur Wahrscheinlichkeit 1-p, dass das Ereignis nicht eintritt, nennt man Odds.

Odds Ratio (OR, Chancenverhältnis). Effektmaß für dichotome Daten. Bezeichnet das Verhältnis (Ratio) der Odds, dass ein Ereignis oder Endpunkt in der experimentellen Gruppe eintritt, zu der Odds, dass das Ereignis in der Kontrollgruppe eintritt. Eine OR

von 1 bedeutet, dass zwischen den Vergleichsgruppen kein Unterschied besteht. Bei ungünstigen Endpunkten zeigt eine OR<1, dass die experimentelle Intervention wirksam ist, um die Odds für das Auftreten dieser ungünstigen Endpunkte zu senken

p-Werte (**p von probability**) beschreiben die Wahrscheinlichkeit, dass der beobachtete (oder ein noch extremerer) Effekt einer Studie aufgetreten sein könnte, wenn die Nullhypothese richtig und der Effekt auf das Spiel des Zufalls zurückzuführen ist. Je kleiner der Wert, desto deutlicher spricht das beobachtete Ergebnis gegen die Nullhypothese. Es ist eine Konvention, dass ein p-Wert gleich oder kleiner 0,05 als statistisch signifikant angesehen wird. Wenn die Signifikanz von Effekten interpretiert wird, sollten p-Werte immer im Zusammenhang mit Konfidenzintervallen verwendet werden.

Power (**statistische Trennschärfe**). Die Fähigkeit einer Studie, einen tatsächlich vorhandenen Unterschied statistisch signifikant nachzuweisen und die Nullhypothese zu verwerfen, wenn sie tatsächlich falsch ist. Der Nachweis bezieht sich auf a priori festgelegte Unterschiede in den Endpunkten („Outcomes") von Therapie- und Kontrollgruppe. Da die Power u.a. entscheidend vom Stichprobenumfang abhängt, kann der allgemein übliche Wert von 80 Prozent z.B. durch eine ausreichend große Stichprobe sichergestellt werden.

Prävalenz beschreibt den Anteil Erkrankter an der Gesamtzahl einer definierten Population zu einem bestimmten Zeitpunkt.

Prospektiv heißt eine Untersuchung, wenn die Datenerhebung begonnen wird, bevor die interessierenden Ereignisse eingetreten sind.

Publikationsbias (**publication bias**). Systematischer Fehler (Bias) aufgrund einer selektiven Publikationspraxis, bei der Studien mit positiven und signifikanten Ergebnissen eine größere Chance haben, publiziert zu werden, als Studien mit negativen und nicht signifikanten Resultaten. Ein systematischer Review oder eine Metaanalyse, die sich ausschließlich auf publizierte Studien stützen, laufen Gefahr, den Effekt der untersuchten Intervention zu überschätzen.

Quasi-Randomisierung. Methoden der Studienzuordnung, die zwar nicht randomisiert sind, jedoch mit der Absicht angewandt werden, bei der Teilnehmerzuordnung ähnliche Gruppen zu gewährleisten. Beispiele: Zuordnung nach Geburtsdatum oder Krankenhausidentifikationsnummer, alternierende Zuordnung.

Randomisierte kontrollierte Studie. Eine experimentelle Studie, bei der die Patienten nach einem Zufallsverfahren (mit verdeckter Zuordnung) auf die Therapie- bzw. die Kontrollgruppe verteilt (Randomisierung) und auf das Auftreten der festgelegten Endpunkte in den einzelnen Gruppen nachbeobachtet werden.

Relative Risikoreduktion (**RRR**). Effektmaß für dichotome Variablen. Die relative Senkung der Rate an ungünstigen Ereignissen in der experimentellen Gruppe (E) einer Studie im Vergleich zur Kontrollgruppe. Sie wird wie folgt berechnet. Beispiel: Das Risiko für eine gastrointestinale Blutung auf einer Intensivstation beträgt ohne Behandlung (Kontrollgruppe) zehn Prozent oder 0,10, bei Prophylaxe mit H2-Blockern (E) sieben Prozent oder 0,07: Die RRR beträgt = 0,3 oder 30 Prozent.

Relatives Risiko (**RR**) bezeichnet das Verhältnis zwischen dem Risiko in der experimentellen Gruppe und dem Risiko in der Kontrollgruppe. Ein relatives Risiko von 1 bedeutet, dass zwischen den Vergleichsgruppen kein Unterschied besteht. Bei ungünstigen Ereignissen zeigt ein RR<1, dass die experimentelle Intervention wirksam ist, um das Auftreten von ungünstigen Ereignissen zu senken.

Risiko (**Rate, Ereignisrate**). Der Anteil von Personen in einer Gruppe, bei denen ein bestimmter Endpunkt auftritt. Wenn z.B. in einer Gruppe von 100 Personen 30 einen bestimmten Endpunkt entwickeln (und bei 70 Personen das Ereignis nicht auftritt), ist das Risiko (oder die Ereignisrate) 0,3 oder 30% (s. Odds).

Sensitivitätsanalyse. Wiederholung der ursprünglichen Analyse unter anderen Annahmen, um zu überprüfen, inwieweit sich dies auf die Ergebnisse auswirkt. Beispiele sind Änderungen der Einschlusskriterien oder Annahmen für fehlende Werte.

Standardabweichung. Maß für die Streuung von Messwerten um den Durchschnittswert.

Statistische Signifikanz. Ein statistisch signifikantes Ergebnis einer Studie ist ein Ergebnis, das gegen die Nullhypothese spricht. Die Aussage basiert auf einem statistischen Test, der zur Prüfung einer vorab festgelegten Hypothese mit vorab festgelegter Irrtumswahrscheinlichkeit durchgeführt wird. Statistische Signifikanz darf nicht mit klinischer Relevanz gleichgesetzt werden.

Störgrößen sind Einflussgrößen, die im Versuchsplan nicht berücksichtigt und auch nicht erfasst werden (siehe auch Confounder).

Surrogatendpunkte (**intermediäre Endpunkte**). Endpunkte, die selbst nicht von unmittelbarer Bedeutung für die Patienten sind, aber stellvertretend für wichtige Endpunkte stehen können (z.B. Blutdruck als Risikofaktor für Schlaganfall). Surrogatendpunkte sind oft physiologische oder biochemische Marker, die relativ schnell und einfach gemessen werden können. Für viele Surrogatendpunkt ist eine zuverlässige Vorhersage eines späteren Ereignisses nicht nachgewiesen.

Systematischer Review. Sekundärforschung, bei der zu einer klar formulierten Frage alle verfügbaren Primärstudien systema-

tisch und nach expliziten Methoden identifiziert, ausgewählt und kritisch bewertet und die Ergebnisse extrahiert und deskriptiv oder mit statistischen Methoden quantitativ (Metaanalyse) zusammengefasst werden. Nicht jeder systematische Review führt zu einer Metaanalyse.

Überlebenszeitanalyse. Eine statistische Analyse, bei der die Zeit bis zum einem bestimmten Ereignis („time to event") zwischen zwei oder mehr Gruppen verglichen wird, um die Wirkung von prognostischen Faktoren, medizinischer Behandlung oder schädlichen Einflüssen zu schätzen. Das Ereignis kann dabei Tod sein, jedoch auch beliebige andere Endpunkte wie Heilung, Rezidiv oder Eintreten einer Komplikation. Beispiele für eine solche Analyse sind die Kaplan-Meier-Methode oder die Cox-Regression.

Übertragbarkeit beschreibt die Übertragbarkeit von Studienergebnissen auf die Patienten in der Routineversorgung, d.h. auf Patienten, die nicht an der Studie teilgenommen haben (s.a. Validität).

Validität (**innere Validität, Glaubwürdigkeit**) bezeichnet das Ausmaß, mit dem die Ergebnisse einer Studie die „wahren" Effekte einer Intervention/Exposition wiedergeben, d.h. frei von systematischen Fehlern (Bias) sind. Die innere Validität beruht auf der Integrität des Studiendesigns und ist Voraussetzung für die Anwendbarkeit der Studienergebnisse in der Routineversorgung.

Verblindung. Geheimhaltung der Gruppenzuordnung (Therapie oder Kontrolle) vor Patienten, Studienärzten, Pflegepersonal oder Auswertern, die an einer Studie teilnehmen. Damit soll verhindert werden, dass durch das Wissen um die Gruppenzugehörigkeit die Therapieantwort der Patienten, das Verhalten der Ärzte oder die Bewertung der Ergebnisse beeinflusst wird. In einfach blinden Studien wissen nur die Patienten nicht über ihre Zuordnung Bescheid. In doppelblinden Studien bleibt die Zuordnung Patient und behandelndem Arzt verborgen. Die Verblindung von Ärzten und Patienten ist nicht immer durchführbar (z.B. beim Vergleich von chirurgischen mit medikamentösen Verfahren), wobei eine Verblindung der Endpunkt-Auswerter in der Regel möglich ist.

Quellen:
http://www.cochrane.de/de/glossary.htm
http://www.reiter1.com/Download.htm
http://www.mh-hannover.de/institute/biometrie/JUMBO/bio/glossar.html#F