JOURNAL FÜR MENOPAUSE

LAURITZEN C

"Evidence based medicine" am Beispiel der präventiven Langzeitsubstitution mit Estrogenen-Gestagenen in der Postmenopause - Teil 2: Statistik

Journal für Menopause 2002; 9 (2) (Ausgabe für Schweiz), 58-64
Journal für Menopause 2002; 9 (2) (Ausgabe für Deutschland) 59-66
Journal für Menopause 2002; 9 (2) (Ausgabe für Österreich) 52-58

ZEITSCHRIFT FÜR DIAGNOSTISCHE, THERAPEUTISCHE UND PROPHYLAKTISCHE ASPEKTE IM KLIMAKTERIUM

Indexed in EMBASE/
Excerpta Medica

Krause & Pachernegg GmbH · VERLAG für MEDIZIN und WIRTSCHAFT · A-3003 Gablitz

C. Lauritzen

„Evidence based medicine“ am Beispiel der präventiven Langzeitsubstitution mit Estrogenen-Gestagenen in der Postmenopause – Teil 2: Statistik

Kriterien für die Zuverlässigkeit von Studien

Die Grundsätze der durch unwidersprochen beweiskräftige Daten abgesicherten Studien können heute allgemeinverbindlich formuliert werden [1]. Zuverlässige Studien müssen gekennzeichnet sein durch überzeugende praxisnahe Vorplanung. Dazu gehört die fehlerfreie, möglichst zufallsbedingte Rekrutierung von Fällen und Kontrollfällen in einer Zahl, die genügende statistische Aussagekraft (power) besitzt. Ferner müssen die praxisnahe Gestaltung der Intervention und die erforderliche Dauer der Untersuchung gewährleistet sein. Zu fordern ist weiter eine eindeutig überprüfbare, korrekte Ausführung und Überwachung der therapeutischen oder präventiven Intervention. Schließlich muß am Ende eine statistisch einwandfreie Berechnung der Daten einschließlich der Erfassung der Fehlerbreite und der Erörterung der möglichen Fehlerquellen stehen. Entscheidend ist eine kritische, vorurteilsfreie Interpretation dessen, was die Ergebnisse wirklich hergeben, wobei, noch einmal, Korrelation nicht mit Kausalität zu verwechseln ist. Bei der Übertragung von Ergebnissen auf Hypothesen, Theorien oder gar Anweisungen für die Praxis werden die meisten und schwersten Fehler gemacht. Die Erfüllung dieser Forderungen wird mit Recht zum Maßstab der Akzeptanz von Ergebnissen wissenschaftlicher Untersuchungen gemacht. Übersichten müssen die zitierten Studien unter den obengenannten Gesichtspunkten unter kritischer Würdigung der Zuverlässigkeitskriterien darstellen und dem Leser präzise sagen, welche Schlußfolgerungen für die Praxis aufgrund ihrer Befunde möglich und erlaubt sind [2, 3].

Die Studienformen der Epidemiologie unterscheiden sich vor allem hinsichtlich des Vorgehens bei der Stichproben- und der Datengewinnung [4, 5]. Grundsätzlich sind folgende Arten von epidemiologischen Studien zur Feststellung von Krankheitsursachen möglich, die eine unterschiedliche Analysekraft und Aussagefähigkeit (power) besitzen.

Die prospektive Studie (Längsschnitt-, Longitudinal- oder Panelstudie)

Dabei wird ein bestimmter Teil der Population mit einem gleichen, als Krankheitsursache verdächtigen Merkmal (beispielweise Frauen in der Menopause, mit Estrogenen behandelte und nicht behandelte) rekrutiert und von Anfang an über einen adäquaten Zeitraum begleitet (Interventionsstudie) und mehrfach untersucht. Dadurch werden der Verlauf und der Endausgang von behandelten und unbehandelten Patienten mit Eintritt oder Nichteintritt einer Erkrankung, zum Beispiel Fraktur, Myokardinfarkt, Brustkrebs, festgehalten.

Am Ende wird das Endergebnis (outcome) ermittelt und mit der wissenschaftlichen Hypothese (am besten der Null-Hypothese = unwirksam im Hinblick auf die Fragestellung) konfrontiert. Diese Patientinnen werden mit den unbehandelten Frauen der gleichen Population verglichen.

Ist die untersuchte Gruppe durch ein gemeinsames Merkmal (z. B. postmenopausale Frauen) einer wohldefinierten Population gekennzeichnet, so spricht man von einer Kohortenstudie.

Die retrospektive Studie

Sie stellt den Versuch einer Abklärung an einem Teil der Population mit einem bestimmten Merkmal in der Vergangenheit dar (Anwendungsbeobachtungen). Beispiel: Patientinnen in der Postmenopause haben vor 10 Jahren über längere Zeit bis heute Estrogene erhalten (= Exposition). An dieser Gesamtheit wird festgestellt, wie oft eine bestimmte Erkrankung, etwa Myokardinfarkt oder Brustkrebs, aufgetreten ist. Dies geschieht durch Ermittlung aus Krankengeschichten oder Ambulanzkarteien, durch Erhebung mit Fragebögen und Telefon oder durch persönliche Interviews. Da Krankenunterlagen meist bezüglich Vorgeschichte und Befunden keineswegs lückenlos und vollzählig sind, ist eine zusätzliche Sicherung durch persönliche Befragung wünschenswert. Bei Interviews sind Angaben aus der Vergangenheit oft fehlerhaft oder ungenau (recall bias). Daher erhöht ein kontrollierender Rückgriff auf ärztliche Befunde die Zuverlässigkeit. Die beiden wichtigsten Formen der retrospektiven Studie sind die Querschnittstudie und die Fall-Kontroll-Studie. Bei der Querschnittstudie als einzeitiger, retrospektiver, bevölkerungsbezogener Methode wird jeder Fall einer Population nur einmal erfaßt. Bei der Längsschnittstudie wird jeder Fall über längere Zeit verfolgt und mehrfach untersucht. Beide Arten des Vorgehens werden unter dem Begriff der Beobachtungsstudie zusammengefaßt. Die Frage, welche Kontrollen geeignet sind, mit den Behandlungsfällen verglichen zu werden, ist für das Ergebnis und seine Zuverlässigkeit von allergrößter Bedeutung. Bei der prospektiven Studie erfolgt meist ein Vergleich mit Bevölkerungsstatistiken (externe Kontrolle) oder mit einer Krankenhaus- bzw. Ambulanzstatistik (innere Kontrolle). Ungleichheiten zwischen Fällen und Kontrollen können eventuell mit speziellen Verfahren herausgerechnet werden. Dies ist jedoch nicht immer zufriedenstellend möglich. Die deutsche Gesellschaft für medizinische Informatik, Biometrie und Epidemiologie hat zusammen mit dem Bundesamt für Arzneimittel Empfehlungen zu Anwendungsbeobachtungen herausgegeben.

Die Kontrollgruppe

Die Behandlungsgruppe muß mit einer Kontrollgruppe verglichen werden, um die Wirksamkeit einer Maßnahme oder eines Medikaments (Verum) nachzuweisen. Die Kontrolle kann gegen eine unbehandelte Gruppe mit gleichen Eigenschaften (z. B. Alter, Gewicht, soziale Herkunft, Ethnie, Ernährungsgewohnheiten, Alkoholkonsum, Medikamente) erfolgen,

wenn möglich, gegen Placebo. Beide Gruppen müssen also ein gleiches Basisrisiko aufweisen. Die Wahl der Vergleichsgruppe ist für die Zuverlässigkeit der Ergebnisse einer Studie entscheidend. Eine häufig vorkommende Fehlermöglichkeit ist die unerkannte Selektion der Behandelten oder der Kontrollen und eine falsche Erinnerung der Patientinnen (recall bias), vornehmlich bei Verwendung von Fragebögen, zumal wenn diese einen ungenügend hohen Rücklauf zeigen. Der Vergleich mit Klinikpatientinnen (sog. innere Kontrollen) ergibt aus verschiedenen Gründen öfter eine Unterbewertung der präventiven Wirkung der Estrogene. Ein Vergleich mit einer Population (sog. externe Kontrollen), aus der ja die Fälle stammen, vermeidet im allgemeinen die systemische Unterbewertung der Estrogenwirkung.

Matched Pairs

Bei der retrospektiven Studie werden die „Fälle" am besten mit sogenannten „matched pairs"-Kontrollen verglichen. Dies sind Patientinnen, die bezüglich der relevanten Eigenschaften (Risikofaktoren) mit den behandelten Fällen gleich sind. Bei mit Estrogenen langzeitig substituierten Frauen müßten die folgenden Faktoren der Kontrollen gleich sein: familiäre Belastung mit Krebs, Alter, Zeitpunkt der Menopause, Gewicht, Ethnie, soziale Herkunft, Ausbildung, Familienstand, Parität, Alkoholkonsum, Rauchen, frühere Einnahme von hormonalen Kontrazeptiva, Gewicht, Fehlen von Stoffwechselerkrankungen, gleiches Vorsorgeverhalten und sonstige Risikofaktoren, beispielweise für Brust-, Endometrium-, Eierstockkrebs, Thrombose und Gallenblasenerkrankungen [3].

Wenn beide Gruppen beispielsweise die Ambulanz zu regelmäßigen Vorsorgeuntersuchungen aufsuchen, dann unterscheiden sich Einnehmerinnen und Kontrollen praktisch so weit wie möglich nur durch die Estrogeneinnahme oder Nichteinnahme der Hormone. Bei Benutzung der „matched pairs"-Methodik wird gleichzeitig das Problem der „healthy users" beseitigt, das heißt, der Einwand wird widerlegt, daß Frauen, die Estrogene einnehmen, gesünder und gesundheitsbewußter seien als solche, die eine Einnahme von Hormonen nicht wünschen.

Die Estrogeneinnehmerinnen, so wird behauptet, seien nicht nur gesünder und gesundheitsbewußter, sie würden später auch seltener erkranken, worauf die scheinbar günstige Wirkung einer Prävention mit Estrogenen beruhen könne. Der Einwand der „healthy users" gilt auch nicht für randomisierte Studien und nicht für eine so homogene Population wie beispielsweise die der „Leisure World", einer medizinisch betreuten Altensiedlung in den USA, in der alle älteren Personen unter gleichen Bedingungen leben.

Die Interventionsstudie (synonym: Präventivstudie) ist gleichsam eine experimentelle Langzeitstudie, in der die durch eine Veränderung (z. B. Beginn einer Estrogensubstitution) oder das Ausschalten von als krankmachend geltenden Faktoren (z. B. überkalorische Ernährung) erzielten Effekte auf die Gesundheit von Bevölkerungsgruppen untersucht werden, also die Wirksamkeit von Präventivmaßnahmen. Zu diesem Typ gehört die Therapiestudie, die, anders verstanden, meist eine sekundäre oder tertiäre Prävention darstellt.

Kausalzusammenhänge zwischen vermuteten Ursachen und beobachteten Wirkungen können am besten durch Interventionsstudien wahrscheinlich gemacht oder nachgewiesen werden. Bei ihnen ist Praxisnähe in der Planung und Ausführung besonders wichtig. Leider fehlt es daran in nicht wenigen solcher Studien, z. B. der HERS-Studie.

Alle Studien können monozentrisch, also an einer einzigen, oder multizentrisch, also an mehreren Institutionen zur selben Zeit mit demselben Untersuchungsplan untersucht werden. Durch multizentrische Studien werden mögliche, das Ergebnis verzerrende Bedingungen oder Fehler (bias) einzelner Institute so weit wie möglich ausgeglichen.

Die Randomisierung

Bei der Rekrutierung der zu untersuchenden Stichprobe hat die Zuteilung durch ein Zufallssystem den Vorteil, jegliche durch Patienten oder den Arzt bedingte Fehler infolge Selektion (selection bias) der Teilnehmer auszuschließen. Von vornherein bestehende, grundlegende Unterschiede müssen daher auf Zufall beruhen. Die randomisierte Studie erlaubt die Evaluierung einer einzelnen Variablen. Man spricht von hypothetiko-deduktivem Vorgehen, da solche Untersuchungen eine eigene Hypothese eher falsifizieren als bestätigen. Der randomisierten, kontrollierten Studie wird eine größere Aussagekraft über die Wirksamkeit einer Intervention zugesprochen: Metaanalysen solcher Studien zeigen eine bessere Zuverlässigkeit als die von Beobachtungsstudien. Aufgrund ihrer einfacheren Konzeption sind Beobachtungsstudien jedoch leichter und mit weniger Material- und Kostenaufwand durchzuführen als randomisierte kontrollierte Studien.

Eigentlich haben statistische Berechnungen von Signifikanz oder Vertrauensgrenzen nur unter der Voraussetzung einer Randomisierung ihre Berechtigung und Bedeutung. Ein offenbarer Nachteil dieser Methode ist, daß vor der Randomisierung oft Frauen mit absoluten, aber oft auch solche mit relativen Kontraindikationen ausgesondert werden und daß die Indikationsstellung zur Intervention dann keinesfalls immer klar ist.

Randomisierte, kontrollierte Studien zeichnen sich bei richtiger Durchführung zwar durch eine hohe interne Validität aus. Es wird jedoch häufig der Fehler gemacht, nicht alle verfügbaren Patientinnen zu randomisieren, da die Patientinnen ja vorher nach ihrer Zustimmung gefragt werden müssen. Das Design ist daher nicht selten von Beginn an fehlerhaft, und die Ergebnisse sind oft nur einge-

schränkt generalisierbar und von den Gegebenheiten in der Praxis nicht selten ziemlich weit entfernt.

Generell muß hier noch die Tatsache angeführt werden, daß Epidemiologen und Statistiker ohne medizinische Erfahrung grundlegende Fehler im Design machen und daß große, randomisierte Studien wegen ihrer Aufwendigkeit öfter von Interessengruppen initiiert werden, die Planung, Verlauf und Ergebnisse der Studien beeinflussen.

Die Tabelle 1 gibt einen Anhalt für die Einstufung der Sicherheitsgrade der Evidenz aufgrund der unterschiedlichen Art der vorliegenden Untersuchungen.

Die Forderung nach kontrollierten, randomisierten Studien wird aber auch ernsthaft in Frage gestellt. Kürzlich haben Forscher der Universität Yale 136 in den Jahren 1985–1998 erschienene Publikationen ausgewertet, in denen jeweils vergleichbare Fragestellungen mittels randomisierter, kontrollierter Studien und andererseits mittels Beobachtungsstudien untersucht worden waren. In fast allen Fällen waren die Ergebnisse beider Verfahren, also von randomisierten, kontrollierten Studien und von einfachen Beobachtungsstudien, gleich ausgefallen [6]. Aufgrund dieser Ergebnisse muß die günstigere Beurteilung der randomisierten, kontrollierten Studien gegenüber reinen Beobachtungsstudien doch erneut kritisch diskutiert werden.

Placebo

Ein Placebo ist ein zu Studienzwecken eingesetztes Scheinmedikament, das keine pharmakologisch wirksame Substanz enthält. Das Placebo darf hinsichtlich Aussehen und Geschmack nicht vom Verum zu unterscheiden sein. In nicht wenigen Untersuchungen kommt die Placebowirkung dem Effekt des Verum in der subjektiven Beurteilung des Patienten sehr nahe. Dabei spielt die Erwartungshaltung des Patienten und des Arztes eine entscheidende Rolle. Im Crossover, dem Austausch der Medikamente in der zweiten Prüfphase, wird dann meist die Überlegenheit des Verum deutlich.

Die statistische Aufarbeitung

Für die statistische Aufarbeitung wird zum Vergleich zweier Stichproben bei Normalverteilung der Daten der Student-t-Test, bei unbekannter Verteilung der Wilcoxon oder der Mann-Whitney-Test verwendet, bei Vergleich mehrerer Stichproben die Varianzanalyse. Bei Abhängigkeit zweier Merkmale in einer Punktwolke wird der Korrelationskoeffizient, für Kontingenztafeln der Chi-Quadrat-Test auf Homogenität angewendet. Manche Befunde (zum Beispiel Lipidwerte und Blut-

Tabelle 1: Sicherheitsgrade der wissenschaftlichen Evidenz (nach Kurz et al., teilweise neu formuliert und ergänzt)

Studie	Evidenz-Typ	Beurteilung	Voraussetzungen, Bemerkungen
Ia	Evidenz aufgrund von Metaanalysen randomisierter kontrollierter Studien	Hohe Zuverlässigkeit; bei Praxisnähe sichere Grundlage ärztlicher Entscheidungen	Voraussetzungen: Praxisnähe; nicht zu viele Ablehnungen bei Randomisierung, wenig Ausschlüsse von der Behandlung, Qualität. Beurteilung durch unabhängige Gutachter. Ausschluß Bias und Confounding. Nachteile: fehlende Individualisierung der Therapie, Praxisferne
IIa	Evidenz aufgrund mindestens einer gut geplanten und ausgeführten randomisierten, kontrollierten Studie	Akzeptable Zuverlässigkeit; bei Praxisnähe annehmbare Grundlage ärztlicher Entscheidungen	Voraussetzungen und Nachteile wie Ia. Hohe Probandenzahl, statist. Power
IIb	Evidenz aufgrund einer gut geplanten und ausgeführten Studie von experimentellem Charakter	Vorläufig akzeptabler Grad der Zuverlässigkeit; bei Praxisnähe provisorische Grundlage ärztlicher Entscheidungen	Höherer Sicherheitsgrad bei Matching, Dosisabhängigkeit, niedrigem CI*, hohem relativem Risiko; Trendnachweis und Risikominderung nach Absetzen der Verum
III	Evidenz aufgrund gut geplanter und ausgeführter nichtexperimenteller Untersuchungen: Vergleichs-, Korrelations- und Fall-Kontroll-Studien	Eingeschränkte Zuverlässigkeit; bei Praxisnähe und Plausibilität vorläufige Grundlage ärztlicher Entscheidungen; Irrtum möglich; Sicherung nach Ia erforderlich	Wie IIa. Besser bei mehreren gleichlautenden und multizentrischen Ergebnissen und Rechenschaftsberichten, gleiche individualisierende Therapie bei bekannten, selbst befragten und betreuten Patienten
IV	Evidenz aufgrund von Fallberichten, autoritativen Meinungsäußerungen von Experten, Gremien, Konsensuskonferenzen und veröffentlichter Erfahrung	Fragliche wissenschaftliche Zuverlässigkeit; Änderung der Meinung aufgrund neuer Befunde häufig nötig; Sicherung nach Ia erforderlich	Abhängig von der Qualität der Begutachter und deren kritischer Formulierung

*CI = Konfidenzintervall, Vertrauensgrenze

druck) sind altersabhängig und manche Ergebnisse von der Dauer der Behandlung bestimmt. Eine solche zeitliche Korrelation wird als Trend bezeichnet. Der Nachweis eines Trends kann manchmal als Zusatzbeweis für das Vorliegen einer gesicherten Korrelation gelten. Einzelheiten sind in den Lehrbüchern der Statistik nachzulesen.

Intention to Treat

Um zu verhindern, daß ein durch Randomisierung hergestelltes Gleichgewicht zwischen zwei Gruppen durch Ausscheider (drop outs) oder Gruppenwechsler gestört wird, werden Patientinnen, die einer Behandlungsgruppe zugeteilt wurden, dort weitergeführt, gleich ob sie tatsächlich die vorgesehene (intendierte) Behandlung erhalten haben oder nicht. Nur durch dieses Vorgehen bleibt der Effekt der Randomisierung erhalten.

In Studien, die diesen Faktor nicht beachten, dürfte die Risikosenkung durch Estrogene eher unterbewertet werden. Steht die Frage im Vordergrund, ob die Behandlung bei denen wirkt, die sie ohne Nebenwirkungen vertragen, so werden für die Analyse auch nur Patienten herangezogen, welche die Studie bis zum Ende durchgeführt haben. Man spricht dann von einer explanativen oder Valid cases-Analyse. Durch diese Selektion erzielen die nebenwirkungsfreien Patienten ein fälschlich überdurchschnittlich gutes Ergebnis.

Number needed to treat

Das ist die Zahl der Patientinnen, die über einen bestimmten Zeitraum behandelt werden müssen, um ein zusätzliches unerwünschtes Ereignis, beispielsweise das Rezidiv eines Krebses oder eines Herzinfarkts, zu vermeiden. Die Zahl wird errechnet durch 1/absolute Risikoreduktion. Eine NNTT von 0,12 würde z. B. bedeuten, daß man 8 Patientinnen behandeln müßte, um einmal ein Rezidiv zu verhüten. Die Zahl der benötigten Patienten ist abhängig von der Höhe des Ausgangsrisikos. Die NNTT gibt an, ob es sinnvoll ist, eine präventive Behandlung zu beginnen und welches Verfahren vergleichsweise am günstigsten ist.

Die „number needed to harm" gibt, in gleicher Weise errechnet, diejenige Zahl der Patienten an, die man behandeln muß, bis unerwünschte Nebenwirkungen oder Schäden auftreten, z. B. eine Thrombose.

Fehlermöglichkeiten (bias), Healthy Users

Die wichtigste Fehlermöglichkeit ist die oben bereits erwähnte durch Selektion der Patientinnen (selection bias). Hierzu gehört die häufig geäußerte Annahme, daß Einnehmerinnen von Estrogenen von vornherein gesünder und gesundheitsbewußter seien als solche Frauen, die keine Hormone nehmen wollen. Hiervon gibt es zweifellos nicht wenige Ausnahmen, denn gerade die Verweigerer einer Hormonbehandlung sind oft besonders gesundheitsbewußt und bestehen auf natürlichen Behandlungsverfahren und pflanzlichen oder homöopathischen Arzneimitteln. Sie tragen also auch kein Behandlungsrisiko. Man spricht von einem „healthy user bias". Solche Annahmen gelten sicher nicht für eine relativ homogene Population wie die der Altensiedlungen (Leisure World) oder die Krankenschwesternpopulation der Colditz-Studie. In der Praxis werden die Patientinnen ja tatsächlich selektiert. Deshalb stehen eigentlich Untersuchungen an nach üblichen Verfahren durch Indikation und Kontraindikation selektierten Patienten der Praxis viel näher, jedenfalls sofern sie nicht systematische Selektionsfehler enthalten. Diese Fehlermöglichkeit der Selektion in der Praxis wird durch die Methode der matched pairs beseitigt, wobei in allen relevanten Eigenschaften gleiche Patienten miteinander verglichen werden. Speziell adjustierte Studien zu dieser Fragestellung haben wesentliche Unterschiede in Risikofaktoren zwischen Estrogeneinnehmerinnen und Nichteinnehmerinnen nicht bestätigt. Die adjustierten RR waren gleich, beispielsweise in der Lipid Research-Studie. Immerhin mag der Einwand der Selektion für einige Studien Gültigkeit besitzen. Entsprechende Berechnungen beziffern den selection bias für kardiovaskuläre Erkrankungen auf etwa 20 %. Das würde eine RR von 0,6 in den Metaanalysen nur mäßig schmälern. Ein anderer Einwand bezüglich eines bias befaßt sich mit der Einnahmezuverlässigkeit oder Behandlungstreue (Compliance). Sie soll besser sein bei Frauen, die freiwillig eine Estrogenbehandlung wählen. Bei zuverlässiger Einnahme liegen allerdings Morbidität und Mortalität für KHK deutlich niedriger. Bei Studien gegen Placebo ist daher die Kontrolle der Compliance besonders wichtig, und die regelmäßige Einnahme wird daher meist streng überwacht.

In einigen Studien werden frühere und jetzige Estrogenbenutzerinnen (past and current users) zusammengefaßt. Auch bei solchem Vorgehen wird die günstige Wirkung der Estrogene unterbewertet, und zwar entsprechend der Menge der past users, da bei diesen meist die günstige Wirkung bereits abgeklungen ist. Das gleiche gilt für die wenig aussagefähige Zusammenfassung „jemals oder niemals (ever/never) Estrogene", da hierbei meist Präparate, Dosen und Dauer der Behandlung nicht erfaßt werden. Oft werden unter die Benutzer sehr kurze Behandlungen und bezüglich der Compliance unkontrollierte Fälle eingeschlossen.

Störgrössen (confounding)

Von „confounding" (Verfälschung) spricht man, wenn ein Faktor, der

nicht unmittelbar Gegenstand der Untersuchung ist, mit der Exposition der Patientinnen oder mit der Zielgröße assoziiert gefunden wird und dadurch das Endergebnis verfälscht. Auf diese Weise kann es zur irrtümlichen Annahme eines Zusammenhangs zwischen zwei Faktoren kommen. Häufige Störgrößen sind Alter, Geschlecht, Gewicht, Rauchen, Alkohol. Eine Kontrolle der Störgrößen ist durch Randomisierung oder „matching" möglich, ferner durch Anwendung einer Stratifizierung der Daten und in begrenztem Umfang durch die Anwendung einer Multivarianzanalyse.

Surrogatmarker

Als solche werden Laborwerte bezeichnet, die vermutlich einen günstigen Einfluß auf bestimmte Erkrankungen anzeigen, beispielsweise Cholesterin-, LDL und HDL-Werte bei Herz-Kreislauf-Erkrankungen. Der Surrogatmarker wird zum Marker, wenn er diesen Einfluß tatsächlich anzeigt. Die wirklich harten, zuverlässigen Parameter wären beispielsweise Erkrankungshäufigkeit und Sterblichkeit.

Metaanalysen

versuchen die vorliegenden Studien unter Gewichtung ihrer Zuverlässigkeit auszuwerten und damit zur Gesamtbeurteilung einer bisher ungeklärten Fragestellung zu kommen. Mit der Metaanalyse ist die Hoffnung verbunden, die manchen Fall-Kontroll-Studien innewohnenden, zum Teil erheblichen Fehlerquellen durch Berechnung eines Mittel- oder Medianwertes auszugleichen [7–13]. Metaanalysen beruhen auf einem vorher festgelegten Protokoll, das die besonders relevanten und fehlerträchtigen Fragen berücksichtigt, nämlich die Form der Rekrutierung, die Art der Identifizierung der Daten, Einschluß- und Ausschlußkriterien, der Kombinationsfaktoren und die Verteilung der Risikofaktoren und Bias-Möglichkeiten in den Gruppen. Diese für die Zuverlässigkeit von Studien wichtigen Kriterien werden am besten von zwei unabhängigen Beurteilern anhand des Protokolls getrennt ermittelt und später verglichen. Die so ermittelte Zuverlässigkeit kann quantitativ in einem Zuverlässigkeitsscore bemessen und in der Gesamtauswertung gewichtet werden.

Die am häufigsten verwendete Methode der Datenkombination ist die Mantel-Haenszel-Technik. Dieses Verfahren beruht auf der (manchmal nicht zutreffenden Annahme), daß die Behandlungserfolge in die gleiche Richtung gehen und etwa die gleiche Größe aufweisen. Unter Benutzung von Odds-Ratios (Odds = Chance, Aussicht) werden die Ergebnisse der individuellen Studien – zum Teil nach Reanalysen – kombiniert, um einen Summenwert der Wirkung mit den jeweiligen Vertrauensgrenzen aus der Varianz aller Studien einschließlich des Schwankungsbereichs zu errechnen. Dabei müssen aber oft Arbeiten unterschiedlicher Methodik summiert werden. Die unterschiedliche Qualität der Studien versucht man durch eine Gewichtung zu mildern. Dennoch bleibt durch die mehr oder weniger willkürliche Auswahl der Arbeiten ein inhärentes Fehlerrisiko bestehen. Größere Unterschiede zwischen individuellen Studien infolge Heterogenität können demnach das Ergebnis und die Schlußfolgerungen einer Metaanalyse zweifelhaft machen. Wenn solche Heterogenität nicht durch Zufall entstanden ist, sondern durch unterschiedliche Populationen oder unterschiedliche Interventionen, und wenn sie daher aus unterschiedlichen Ergebnissen stammt, so kann die Kombination der Studien sehr fehlleitend sein. Leider haben die Tests für Heterogenität nur begrenzte Aussagefähigkeit. Es empfiehlt sich daher, mittlere Befunde und „Ausreißer" in eine graphische Darstellung einzutragen, so daß der Leser das Ausmaß der Heterogenität selbst beurteilen kann. Beruht eine Metaanalyse auf einer Reihe gleichartig verzerrter Studien, so wird die kumulative Evidenz diesen Fehler nicht ausgleichen können. Wird eine einzelne Studie als besser eingeordnet als sie ist, so wird das Ergebnis eine Pseudoevidenz sein. Evidenzfallen sind außerdem Studien, in denen unbewußt oder bewußt und gezielt Evidenz produziert wird, beispielsweise für bestimmte neu einzuführende Medikamente. Auch die Metaanalyse hat also ihre Grenzen. Letzten Endes ist auch ihr Ergebnis fast immer angreifbar.

Einer der entscheidenden Nachteile ist, daß in den meisten Studien und insbesondere bei Metaanalysen die Qualität der Behandlung und Gesamtbetreuung aus den Ergebnissen in keiner Weise ablesbar ist: Empathie, verbale und nichtverbale Kommunikation beeinflussen ein Behandlungsergebnis. Insbesondere eine mit keinem Verfahren meßbare, individuell maßgeschneiderte Therapie ist in hohem Maße geeignet, die Häufigkeit von Nebenwirkungen zu reduzieren. Die in diesem Absatz genannten wichtigen Faktoren sind aber in den meisten Arbeiten und Metaanalysen nicht erkennbar und sind sehr oft nicht gegeben.

Das relative Risiko

Folgende Begriffe werden zur quantitativen Beurteilung eines relativen Risikos verwendet:

Inzidenz = Zahl der Neuerkrankungen an einer bestimmten Erkrankung durch Gesamtzahl der Bevölkerung.

Prävalenz = Zahl der Erkrankten an einer bestimmten Erkrankung durch Gesamtzahl der Bevölkerung. Die Wahrscheinlichkeit, mit der eine bestimmte Erkrankung verhütet oder gefördert wird, gibt man als Odds

Ratio (OR) oder als relatives Risiko (RR) an. Das relative Risiko soll angeben, um welchen Faktor unter oder über 1,0 (normales Risiko) exponierte Personen stärker oder weniger gefährdet sind als nicht exponierte Personen. Der Begriff „relatives Risiko" gibt keine Information über die Häufigkeit einer Erkrankung, sondern erfaßt lediglich das Verhältnis der Inzidenzen zwischen zwei Vergleichsgruppen. Er läßt keinen Schluß auf die klinische Bedeutung einer Risikoerhöhung zu. Beispielweise kann die 10fache Erhöhung einer sehr seltenen Erkrankung eine weitaus geringere Bedeutung haben als die Steigerung des RR einer sehr häufigen Erkrankung beispielsweise um den Faktor 2 [14]. Risikofaktoren sind jene pathogenen Bedingungen, die in Bevölkerungsstudien bei der Untersuchung der Entstehungsbedingungen bestimmter Erkrankungen als korrelierend oder vermutlich als ursächlich bestimmend (kausal) mit adäquaten statistischen Verfahren mit einem bestimmten Sicherheitsgrad wahrscheinlich gemacht wurden. Der Begriff „Risikofaktor" stellt ebenfalls nicht notwendigerweise einen kausalen Bezug zwischen dem gegebenen Faktor und dem Krankheitsbild her, sondern beschreibt die gemeinsamen Eigenschaften einer Gruppe von Faktoren, bei denen die Inzidenz einer Erkrankung erhöht ist. Für diesen Zweck wäre die Bezeichnung Risikoindikator oder Risikomarker sinnvoller. Unter attributivem Risiko versteht man ein Risiko, das einer bestimmten Ursache zuzuordnen ist. Zu den errechneten Mittelwerten oder dem errechneten relativen Risiko müssen die Vertrauensgrenzen (Konfidenzintervalle = CI) kalkuliert werden. Das CI bezeichnet das Intervall, das den Erwartungswert mit einer bestimmten Wahrscheinlichkeit einschließt, beziehungsweise den Bereich, in dem der errechnete Wert mit einer bestimmten Wahrscheinlichkeit gelegen sein wird. Meist wird das 95%ige Intervall angegeben. Je enger die Vertrauensgrenzen liegen, desto näher liegen die Werte zusammen, desto geringer ist also die Streuung und desto zuverlässiger ist wahrscheinlich das Ergebnis. Der p-Wert gibt die Wahrscheinlichkeit (p = probability) oder die Höhe der Irrtumsmöglichkeit an, mit welcher das Ereignis reell, also wirklich prädiktiv ist oder mit welcher ein Ereignis voraussichtlich wirklich eintreten wird; p = 0,05 bedeutet 95 % Wahrscheinlichkeit, 0,01 = 99 % Wahrscheinlichkeit, 0,001 = 1 Promille. Ein Ergebnis von 1 ‰ wird als hochsignifikant, von 1 % als signifikant, von 5 % als schwach signifikant bezeichnet. Der Begriff „signifikant" wird oft verwendet, um eine Untersuchung mit dem Stempel der Wissenschaftlichkeit zu versehen. Signifikanz bedeutet aber im wissenschaftlichen Sinne nur, daß das betreffende Ergebnis vermutlich nicht durch Zufall (durch die Null-Hypothese) erklärbar ist, allerdings unter dem Vorbehalt der Höhe des p-Wertes. Der Begriff „signifikant" macht jedoch eine Aussage darüber, ob die Versuchsplanung, die Erhebung oder die Auswertung korrekt waren, oder ob das gefundene Ergebnis vermutlich relevant ist. Der Vorhersagewert (predictive value) einer Untersuchung ist die Wahrscheinlichkeit, mit der eine Vorhersage vermutlich eintreffen wird. Sensitivität, Empfindlichkeit einer Untersuchung = Zahl der zutreffend als positiv (z. B. als krank) beurteilten Fälle geteilt durch die Gesamtzahl der positiven (z. B. kranken) Fälle. Spezifität = Zahl der zutreffend als negativ (z. B. als gesund) erkannten Fälle/die Gesamtzahl der positiven (erkrankten) Fälle.

Beurteilung der Zahlen des relativen Risikos

Der Begriff des relativen Risikos (RR) hat bei Laien und Ärzten teilweise zu Mißverständnissen Anlaß gegeben. Fälschlich wurde angenommen, daß beispielsweise bei einem Risiko von 1,32 zusätzlich 32 % über die normale Erkrankungshäufigkeit der Bevölkerung hinaus an Brustkrebs erkranken, etwa, wenn sie länger als 5 Jahre Estrogene einnehmen. Ein RR von 1,32 (z. B. bei Colditz et al. für Brustkrebs [15]) bedeutet aber tatsächlich, daß während (nicht infolge) Estrogeneinnahme bei 32 % der Untergruppe von den 10 % der (unbehandelten) weiblichen Gesamtpopulation, die in ihrem Leben ein Mammakarzinom entwickeln werden, vermutlich ein Brustkrebs zur Diagnose kommen wird, der möglicherweise sonst nicht oder wenigstens nicht zu diesem Zeitpunkt nachgewiesen worden wäre. Dies wären tatsächlich 2 Fälle auf 1000 postmenopausale Frauen bei Einnahme der Estrogene über mehr als 5 Jahre. Ein RR von 1,32 mit einem 95 %-Vertrauensbereich von 1,12–1,54 bedeutet, daß auch eine nur 12%ige Zunahme oder eine 54%ige Zunahme bei den 10 % der Erkrankten möglich wäre. Ein unterer Wert des Vertrauensbereichs unter 1,0 bedeutet, daß im allgemeinen nicht von einer Signifikanz der Befunde ausgegangen werden kann. Die oben genannten 2 Fälle auf 60/1000 Brustkrebse werden oft als zusätzliche Fälle (zu den 60 pro 1000 spontan auftretenden Fällen) bezeichnet. In Wahrheit sind sie in den 60 pro 1000 oder den 47.000 Brustkrebsfällen, die jährlich in Deutschland registriert werden, bereits enthalten. Das zusätzliche Auftreten ist eine Hypothese, die aus den Zahlen nicht unmittelbar hervorgeht. Sehr wahrscheinlich handelt es sich ja auch um Krebse, die undiagnostiziert schon vorher bestanden, die aber durch rascheres Wachstum kumulativ früher zur Diagnose gelangt sind.

Von welcher Höhe an zeigt das relative Risiko eine tatsächliche bestehende Risikosteigerung an? M. Angell (New Engl J Med) hat sich folgendermaßen geäußert: „Wir schauen nach einem relativen Risiko, das bei 3 oder höher liegt, ehe wir eine Arbeit zur Publikation annehmen, insbesondere dann, wenn das Ergebnis biologisch nicht plausibel ist, oder wenn es sich um ein ganz neuartiges Ergebnis handelt." R. Temple (Food and Drug

Administration) schreibt: „Meine Grundregel lautet: Wenn das relative Risiko nicht höher ist als 3 oder 4, so vergiß es." Die gleiche Meinung haben D. Trichopoulos (Harvard) und S. Shapiro (New York) geäußert. Aus diesen Stellungnahmen von Experten der statistischen Auswertung wird klar, wie unsinnig es ist, aus RR-Zahlen wie 1,32 oder 1,2 zu berechnen, mit welcher Größe das Mammakarzinom pro Jahr zunimmt (z. B. Beral). Das ist Pseudomathematik.

Weitere Begriffe

Mortalität = Zahl der Gestorbenen/Gesamtzahl der Bevölkerung. Der Begriff der Mortalität wird teilweise als Synonym für Inzidenz oder Prävalenz verwendet.

Letalität = Zahl der an einer bestimmten Erkrankung Gestorbenen/Zahl der Erkrankten. Die oben genannten relativen Häufigkeiten werden meist mit 1000, 10.000 oder 100.000 multipliziert, so daß die genannten Maßzahlen dann absolute Häufigkeiten ausdrücken = *standardisierte Rate.* Während die relative Häufigkeit ein Erfahrungswert ist, der aus einer größeren definierten Beobachtungsreihe gewonnen wurde, ist die Wahrscheinlichkeit ein abstrakter theoretischer Wert, der dazu dient, Voraussagen über zukünftige Beobachtungen zu machen.

Regression und Korrelation

Der Korrelationskoeffizient macht eine Aussage darüber, wie eng der statistische Zusammenhang zwischen zwei Größen ist. Die Regressionsanalyse dient dazu, Schlüsse von einer Größe auf die andere zu ziehen, also z. B. auszusagen, welcher y-Wert bei einem bestimmten x-Wert zu erwarten ist. Die Regressionswerte, beispielsweise in einer Punktwolke auf einer diagonalen Geraden dargestellt, sind um so zuverlässiger korreliert, je enger die einzelnen Punkte an der Regressionsgeraden zusammenliegen. Der Korrelationskoeffizient ist umso niedriger, je größer die Streuung ist. Eine positive Korrelation kann durchaus zufallsbedingt sein. Kausalität ist daraus nicht zu schließen. Dies ist allenfalls möglich, wenn die Korrelation sehr stark ist, sich in gleichen Untersuchungen immer gleich darstellt, wenn es eine Dosis-Wirkungs-Beziehung gibt, wenn die Assoziation epidemiologisch und biologisch Sinn ergibt und spezifisch zu sein scheint.

Die primäre und sekundäre Prävention

Die primäre Prävention versucht, gesundheitliche Bedingungen zu schaffen oder zu beeinflussen, die dem Entstehen oder Auftreten von Risiken entgegengesetzt sind. Sie will also drohende Krankheiten verhüten. Die sekundäre oder tertiäre Prävention auf der Basis eines Risikokonzepts zielt ab auf die Reduktion von Morbidität oder vorzeitiger Mortalität bei bereits vorbestehender Erkrankung (z. B. Versuch einer präventiven Estrogenbehandlung nach Herzinfarkt oder Bypassoperation wie in der HERS-Studie). Es handelt sich häufig eigentlich um eine Therapiestudie. Die Intervention kann erfolgen durch Medikamente (z. B. Estrogene oder Statine), Diät oder durch die Beseitigung von Risikofaktoren.

Literatur:

1. Schulz KF, Chalmers I, Hayes R, Aleman DG. Empirical evidence of bias. Dimension of methodological quality, associated with estimates of treatment effects in controlled trials. New Engl J Med 1983; 309: 1358–61.

2. Begg CVB, Berlin JA. Publication bias: A problem in interpreting medical data. JR Statist Soc 1988; 151: 419–63.

3. Dickerson K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. Br Med J 1994; 309: 1286–91.

4. Rothman KJ, Greenland SA. Modern epidemiology. 2nd ed. Lippincott Raven, Philadelphia, 1998.

5. Sackett DCL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology, A Basic Science for Clinical Medicine. 2nd edition. Little Brown, Boston, 1991.

6. Benson K, Hartz AJ. A comparision of observational studies and randomized clinical trials. N Engl J Med 2000; 342: 1878–86.

7. Antman E, Lau J, Kapelnik B. A comparision of results of metaanalyses of randomized clinical trials and recommendations of clinical experts. Treatment for myocardial infarction. JAMA 1992; 268: 240–8.

8. Hughes EG. Beyond Publication Bias. Meta-analysis and medical literature. ORGANON Magazine on womens Health. Organorama 1993; 3: 12.

9. Hulley S, Grady D, Bush T et al. for the Heart and Estrogen/Progestin Replacement Study (HERS) Research Group. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. JAMA 1998; 280: 605–13.

10. Meyer R. Metaanalysen und ihre Grenzen. Deutsches Ärzteblatt 1999; 20: C 1464.

11. Moher D, Olkin I. Meta-analysis of randomized controlled trials: a concern for standards. JAMA 1995; 274: 1962–3.

12. Assmann G, Cullen P, Schultze H. Plasma risk factors and cardiovascular disease. Europ Menopause J 1996; 3 (Special Issue 3): 203–8.

13. Victor N. The challenge of meta-analysis: indications and contraindications for meta-analysis. J Clin Epidemiol 1995; 48: 5–7.

14. Walter SD. Determination of significant relative risks and optimal sampling procedure in prospective and retrospective studies of various sites. Am J Epidemiol 1977; 105: 387–97.

15. Colditz GA, Willett WC, Stampfer MJ et al. Menopause and the risk of coronary heart disease in woman. N Engl J Med 1987; 316: 1105–10.

Teil 3: „EBM in bezug auf Probleme in den Wechseljahren"
folgt im Journal für Menopause 3/2002

Mitteilungen aus der Redaktion

Besuchen Sie unsere Rubrik

☑ Medizintechnik-Produkte

Neues CRT-D Implantat
Intica 7 HF-T QP von Biotronik

Artis pheno
Siemens Healthcare Diagnostics GmbH

Philips Azurion:
Innovative Bildgebungslösung

Aspirator 3
Labotect GmbH

InControl 1050
Labotect GmbH

e-Journal-Abo

Beziehen Sie die elektronischen Ausgaben dieser Zeitschrift hier.

Die Lieferung umfasst 4–5 Ausgaben pro Jahr zzgl. allfälliger Sonderhefte.

Unsere e-Journale stehen als PDF-Datei zur Verfügung und sind auf den meisten der marktüblichen e-Book-Readern, Tablets sowie auf iPad funktionsfähig.

☑ Bestellung e-Journal-Abo