Zur Diskrepanz zwischen Theorie und Praxis in der Medizin

→ 500 Jahre nach Paracelsus

Dem eigenen Erkennen und Verstehen ebenso rechtgeben können.

Einsiedler Kurse

Zeitgemässe Beurteilung medizinischen Wissens

„Critical Appraisal“ und Methoden der klinischen Epidemiologie für die praxisbezogene Bewertung von Studien

Kurs-Skriptum

Version 2g
(Mai 1999)

(c) Johannes G. Schmidt, Heiner C. Bucher, Johann Steurer

schmidt@paracelsus-heute.ch bucher@ubaclu.unibas.ch johann.steurer@dim.usz.ch

Motto: Es geht nicht darum, wie belesen ein Teilnehmer in bezug auf die letzten Studienergebnisse ist, sondern nur darum, den Verstand kritisch und logisch zu verwenden.

PARACELSUS HEUTE – STIFTUNG ZEITGEMÄSSE PRAXIS UND KRITISCHE WISSENSCHAFT IN DER MEDIZIN

Praxiszentrum Meinradsberg, Ilgenweidstrasse 3, CH-8840 Einsiedeln/Schweiz
Tel. (+41) 055 418 81 95 Fax 055 418 81 96 e-mail: stiftung@paracelsus-heute.ch
Kto. 306020-0024, Kantonalbank Schwyz Einsiedeln (PC 60-2577-8 · BLZ KBSZ CH 228811)

Inhalt

1. Einführung - Tour d'Horizon: Was wissen wir über das maligne Melanom?

- *Arbeit im Plenum*
- *Vorgehensweise interaktiv*
- *Beispiel malignes Melanom mit dem im Raum vorhandenen Wissen so weit wie möglich anhand konkreter oder in etwa vermuteter Zahlen aufarbeiten*
- *Wichtig ist die prinzipielle Art des Fragestellens*

1.1. Unechte statistische Zusammenhänge in Krebsstudien

Viele, zunächst sinnvoll und plausibel erscheinende statistische Zusammenhänge "beweisen" anscheinend die Vorteile der Krebsfrüherkennung. Folgende Fehlermöglichkeiten gilt es aber in Betracht zu ziehen:

- Lead-time bias (Diagnose-Vorverlegung)

Die Diagnose wird vorverlegt: => Verlängerung der Krankheitsphase = Verkürzung der krankheitsfreien Lebenszeit. Dadurch wird "Überlebenszeit" länger.

- Length bias (Wachstumsgeschwindigkeit)

Langsam wachsende, klinisch gutartigere Karzinome können durch regelmässige Untersuchungen leichter erkannt werden, wenn sie noch in einem frühen Stadium sind, weil sie sich im Intervall zwischen den Untersuchungen wenig verändern. Schnellwachsende, aggressiv verlaufende Karzinome hingegen wachsen oft so stark, dass sie bereits im Intervall zwischen den Untersuchungen zu einem grösseren, klinisch manifesten Tumor heranwachsen; sie fallen durch das Netz einer Früherkennung. Damit bestimmt die relative klinische Gutartigkeit des Karzinoms, und nicht notwendigerweise die Früherkennung, eine bessere Prognose.

- Overdiagnosis bias (Falsche Krebsdiagnosen)

Ein Screening führt immer auch zu falsch positiven Krebsdiagnosen bei histologisch benignen Tumoren, da die Histologie nicht fehlerfrei ist (siehe 1.3.). Ein Screening entdeckt zudem Fälle histologisch maligner Karzinome, die klinisch gutartig sind und ohne Suche lebenslänglich stumm bleiben.

- Selection bias/healthy screenee bias

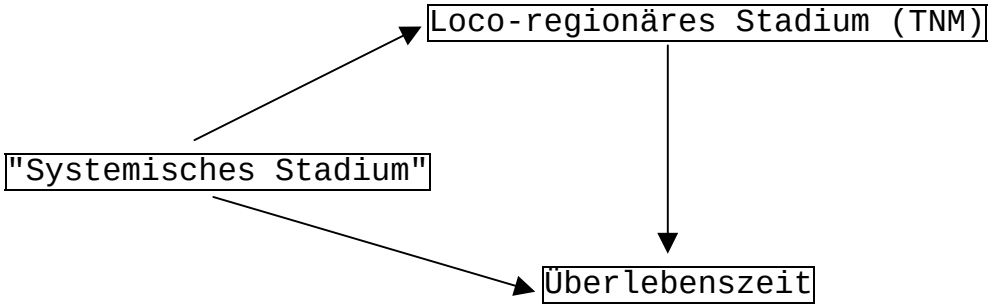
Teilnehmer an Screeninguntersuchungen sind regelmässig Personen, die gesünder sind als die "Verweigerer". Dadurch ist die bessere Prognose früh-entdeckter bzw. früh- oder präventiv behandelter Krankheiten eine Folge der Selektion von gesünderen (bzw. resistenteren) Patienten und nicht notwendigerweise das Resultat der Frühbehandlung.

Siehe Unterlage I:
Schmidt JG. Mammakarzinom: Früherkennungs-Glaube und Wirklichkeit. *Z. Allg. Med.* 1994; 70: 437-442

- Confounding effect/bias (Hintergrunds-Faktor)

Der präventive Nutzen bemisst sich nicht an "Stadien-Erfolgen", wie die oben erwähnten biases bereits klarmachen. Entscheidend ist, ob die letalen Verläufe durch die Früherfassung geheilt oder gebremst werden können.

Das Assoziations-Dreieck:



zeigt den möglichen "Stadienkosmetik"-Trugschluss (= ein confounding effect/bias) auf.
Es ist denkbar, dass der Zusammenhang zwischen frühem Krebsstadium (bzw. geringer Tumordicke beim Melanom) und der längeren Überlebenszeit nur indirekt besteht und auf dem direkten Zusammenhang zwischen dem klinisch nicht untersuchbaren systemischen Stadium und der Prognose beruht.
Das (hypothetische) systemische Stadium in dieser Erklärungs-Alternative bestimmt die Überlebenszeit möglicherweise direkt, das TNM-Stadium ist lediglich ein "Epi-phänomen", dessen chirurgische Entfernung die direkte Achse in diesem Modell unberührt lässt.

Siehe Anhang I
Abb. aus: Schmidt J. Die Brustkrebs-Vorsorgeuntersuchung: Die kritische und praktisch relevante Beurteilung wissenschaftlicher Daten. *Das Argument* (Berlin) 1988; AS 178: 100-122

1.2. Prävention, Stadienverteilung und Häufigkeit

Allzu oft wird ein vermeintlicher Präventionserfolg unüberlegt an der Stadienverteilung im Anschluss an präventive Massnahmen festgemacht.
1996, aus einem Schreiben eines Direktors einer dermatologischen Universitätsklinik: „Die zahlreichen sehr umfangreichen Studien ... belegen eindeutig, dass in den letzten Jahren insbesondere im Anschluss an Präventionskampagnen eine prozentuale Zunahme der dünnen Melanome zu verzeichnen ist. Andererseits ist der Zusammenhang zwischen Tumordicke und und Prognose unzweifelhaft. Wenn man also die prozentuale Zunahme der dünnen Melanome in der Folge der Aufklärungsaktionen sieht, bedeutet dies gleichzeitig eine eine Zunahme von Patienten mit verbesserter Prognose.“

Wieso ist es ein Trugschluss, "Stadienverteilungs-Kosmetik" mit präventiver Wirksamkeit gleichzusetzen? Früherkennungkampagnen führen dazu, dass die Inzidenz entdeckter Krebsfälle zunimmt, wobei vor allem mehr Frühformen diagnostiziert werden. Eine "günstigere" Stadienverteilung kann sich deshalb auch dann ergeben, wenn die Inzidenz fortgeschrittener Stadien (beim Melanom die Inzidenz von Tumoren mit fortgeschrittener Tumordicke) gleichbleibt. Dies wird in folgender Tabelle veranschaulicht:

	Verteilung von Tumordicke/-stadien				
	I	**II**	**III**	**IV**	Total
+ Kampagne	20	10	5	5	40
Prozentuale Verteilung	50%	25%	12,5%	12,5%	100%
- Kampagne	5	5	5	5	20
Prozentuale Verteilung	25%	25%	25%	25%	100%

z.B. Inzidenz auf 10'000

Die viel "günstigere" Stadienverteilung verbirgt, dass die absolute Anzahl "ungünstig" dicker Melanome mit und ohne Kampagne identisch ist.

Siehe Anhang II
Rees JL. The melanoma epidemic: reality and artefact. *BMJ* 1996; 312: 137-138

Merke: Eine Krebsfallzunahme (Inzidenz-Anstieg) kann zwei wichtige Ursachen haben:
1) Eine Zunahme der Krankheit an sich
2) Eine Zunahme der Zahl diagnostizierter Fälle in der Folge von Früherkennungskampagnen

Merke: Eine Kontrolle der genannten statistischen Artefakte ist nur möglich durch kontrollierte Studien mit gültigen Endpunkten (Mortalität und nicht nur Stadienverteilung etc.).

Zur Melanom-Früherkennung fehlen kontrollierte Studien, die eine Wirksamkeit über die statistischen Artefakte hinaus belegen könnten.

=> *kontrollierte Studien: Siehe 3.7.*

1.3. Hat die Histologie immer recht?

Wie spezifisch ist die Histologie?
Beim Melanom sind keine Untersuchungen bekannt.
Quervergleich Mammakarzinom: Bei 17% der Frühstadien und bei 6% aller in einem Früherkennungsprogramm entdeckten Karzinome war die Malignitätsdiagnose in einer späteren Nachbeurteilung strittig.

Wenn z.B. die Histologie 1 falsch positive Malignitätsdiagnose auf 1000 Präparate benigner Exzisate macht und die Malignitätswahrscheinlichkeit bei allen ziellos eingesandten Hautexzisaten 1:1000 beträgt, dann ist der positiv prädiktive Wert eines Malignitäts-Befundes 1 zu 1, d.h. 50%.
Eine Malignitätsdiagnose ist bei unselektioniertem Material deshalb unzuverlässig. Eine Früherkennung bringt es unvermeidbar mit sich, dass die Zahl falsch diagnostizierter Krebsfälle deutlich zunimmt (siehe overdiagnosis bias in Unterlage I).

=> *Spezifität, Sensitivität, PPV und NPV: Siehe 2.2.*

1.3. Ausmass eines Therapienutzens, Netto-Nutzen

Der Quervergleich Mammakarzinom zeigt, dass die (1) gesicherte präventive Wirkung der Brustkrebsfrüherkennung bei einem (2) viel häufigeren Karzinom als quantitativ unbedeutend betrachtet werden kann; 1 Todesfall in ca. 15'000 Frauenjahren kann verhütet werden. Die Zahl verhüteter Todesfälle des weit

selteneren Melanoms ist bei optimistischer Annahme einer tatsächlich vorhandenen präventiven Wirkung der Früherkennung verschwindend klein.

Hauptsächliche unerwünschte Wirkung ist die drastische Zunahme diagnostizierte Fälle, was bedeutet, dass die Leidenszeit durch eine frühere Diagnose verlängert wird und dass durch ein Programm zusätzlich Menschen zu verängstigten Krebspatienten gemacht werden, die ohne Programm bis zu ihrem Tod "unentdeckt" geblieben wären.

Die Frage ist, wie gross erwünschte Wirkungen im Vergleich zu unerwünschten Wirkungen sind. Es lässt sich grob abschätzen, dass "unnötige" Krebsdiagnosen um ein Vielfaches häufiger sind als die fraglich verhütbaren Melanom-Todesfälle.

Zur Beurteilung des Nettonutzens ergeben sich folgende Hauptfragen:

- Melanommortalitäts-Reduktion in absoluten Zahlen?

- "Epidemiologisches Discounting": Kann diese angesichts der kompetitiven Mortalität zum Tragen kommen ?

- Ausmass von Krebsfall-Zunahme und unnützer Krankheitsverlängerung?

Siehe Anhang III
Voraussetzungen für den sinnvollen Einsatz von Screeningtests aus: Bucher H. und Gutzwiller F. *Checkliste Gesundheitsberatung und Prävention*. Georg Thieme Verlag 1993

Siehe auch:
Schmidt JG. Früherkennung und Umgang mit Risikofaktoren. In: Kochen MM, ed. *Allgemein- und Familienmedizin (Lehrbuch Duale Reihe)*. Stuttgart: Hippokrates Verlag, 1998; 118-134

Anmerkung: Bei dieser Kurs-Übung ging es nicht um das Erarbeiten eines abschliessenden Urteils über Wirkung und Nutzen der Melanom-Prävention, sondern um eine Einführung in das Hinterfragen der vorliegenden Evidenz.
Eine aktuelle Durchsicht publizierter Studien durch den Kursleiter (anhand einer MEDLINE-Suche) lässt dennoch feststellen, dass Studien fehlen, die eine präventive Wirkung der Melanom-Früherkennung

belegen könnten, und dass sich durch Kampagnen regelmässig die Häufigkeit diagnostizierter Melanome verdoppelte, ohne dass dadurch eine Reduktion der Mortalität festgestellt werden konnte. Eine Australische Studie weist auf die Möglichkeit hin, dass mit der Früherkennung histologisch maligne Hauttumoren ans Tageslicht gezerrt werden könnten, die klinisch gutartig sind [Burton RC. Curr Opin Oncol 1995; 7: 170-174]. Die Kampagnen haben somit zusätzlich zahlreiche Menschen zu "Krebs"-Patienten gemacht, ohne dass eine präventive Wirkung feststeht. Eine unnötige Krebsdiagnose zu erhalten, ohne dass dadurch ein sicherer Nutzen entsteht, muss als gesundheitliche Beeinträchtigung gewertet werden.

2. Aussagewert medizinischer Tests

2.1. Fehlermöglichkeiten apparativer Untersuchungen

Tests sind nie hundert Prozent zuverlässig, ein Test kann entweder trotz vorliegender Krankheit falsch negativ sein (was die Sensitivität vermindert) oder trotz Abwesenheit von Krankheit falsch positiv (was die Spezifität vermindert).

Es gibt eine Reihe von **Fehlerquellen**:

1) "Technische" Fehler

- Manipulations-Fehler im Labor
  z.B. ungenaues Messen beim Verdünnen etc.

- Beobachter-Fehler
  z.B. Uneindeutige Grenzfälle beim Beurteilen eines Röntgenbildes oder Blutausstriches etc.

- Systematische Fehler
  z.B. zu hohe Temperaturen im Chemielabor oder fehlerhafte Tarierung; die fehlerhafte Abweichung geht dann systematisch in eine Richtung

2) Zufalls-Fehler

- Zufalls-Fehler
  Der Prozess der Messung oder Zählung stützt sich auf eine (möglichst repräsentative) Zufallsprobe (z.B. 4 Felder in der Leuko-

zyten-Zählkammer) und ist nicht kontrollierbaren "natürlichen Schwankungen" (z.B. kleine Temperaturschwankungen, die nach den Zufalls-Gesetzen in wenigen Momenten extrem sind) ausgesetzt.

Merke: Nicht kontrollierbare Zufallsfehler lassen sich in Ringversuchen nicht unterscheiden von Manipulations-Fehlern. Auch Resultate qualitativ bedenkenloser Labors sind unter Einbezug der klinischen Situation mit Vorsicht zu werten.

3) "Biologische" Fehler

- z.B. Antikörper, die sich trotz Infektion (noch) nicht gebildet haben oder Kreuzreaktionen
- Begleiterkrankungen, welche Test beeinflussen etc.
- Überlappung der Werte von Gesunden und Kranken; Grenzwerte sind of willkürlich als Mittelwert + 2 Standardabweichungen (=95% spezifisch) definiert.

2.2. Spezifität, PPV, Sensitivität, NPV, Pre- und Posttest-likelihood anhand der Vierfeldertafel

Sensitivität (Empfindlichkeit) heisst der prozentuale Anteil der Kranken, die vom Test erfasst, d.h. nicht fälschlich verpasst werden.

Spezifität heisst der prozentuale Anteil der Gesunden, die vom Test richtig erkannt werden, d.h. nicht fälschlich als krank bezeichnet werden.

Die Überlappung der Laborwerte von Gesunden und Kranken bedingt einen Grenzwert-Kompromiss, welcher einen Kompromiss zwischen Sensitivität (Empfindlichkeit) und Spezifität bedeutet. Wird der Grenzwert näher zu den Kranken verschoben, vermindert sich die Sensitivität, dafür erhöht sich die Spezifität - und umgekehrt.
Auch das Kombinieren mehrerer Tests unterliegt dieser Kompromiss-Logik. Wird krankhaft definiert als pathologische Werte in allen Tests, dann nehmen die falsch-positiven Fehler ab und damit die Spezi-

Vierfeldertafel ("Two-by-two table")

		Krankheit +	Krankheit -	
Test	+	richtig positive a	falsch positive b	Test-positive
	-	c falsch negative	d richtig negative	Test-negative
		Kranke	Gesunde	

$$\text{Sensitivität} = \frac{\text{richtig positive}}{\text{alle Kranke}} = \frac{a}{a+c}$$

$$\text{Spezifität} = \frac{\text{richtig negative}}{\text{alle Gesunde}} = \frac{d}{b+d}$$

$$\text{Vortest-Wahrscheinlichkeit} = \frac{\text{Kranke}}{\text{Alle}} = \frac{a+c}{a+b+c+d}$$

$$\text{PPV} = \frac{\text{richtig positive}}{\text{Testpositive}} = \frac{a}{a+b}$$

$$\text{NPV} = \frac{\text{richtig negative}}{\text{Testnegative}} = \frac{d}{c+d}$$

$$\text{Likelihood Ratio +} = \frac{\text{Sensitivität}}{100\% - \text{Spezifität}} = \frac{a}{a+c} : \frac{b}{b+d}$$

fität zu, dafür gehen diejenigen "durch die Latte", die in einzelnen Tests richtig positiv waren, aber in anderen Tests durch falsch negative Ergebnisse unbestätigt blieben; die Sensitivität nimmt dadurch ab. Die Fehler der verschiedenen Tests können mehr oder weniger abhängig von gemeinsamen Mechanismen sein, die Berechnung ist dementsprechend kompliziert (Theorem von Bayes) und übersteigt diesen Kurs. Wichtig ist die Kenntnis dieser prinzipiellen Abhängigkeit von Spezifität und Sensitivität. Wird also ein "sensitiver" Tests (z.B. ELISA bei HIV) mit einem "spezifischen" Test (Western Blot) kombiniert, dann ist die gesamte Testkaskade nicht einfach eine "Addition". Gilt nur ein bestätigtes Ergebnis als HIV-positiv, dann liegt die Sensitivität der Testkaskade unter der Sensitivität des "sensiblen" ELISA-Tests.

Merke: Die Spezifität hängt ab von den falsch positive Testfehlern bei tatsächlich Gesunden, die Sensitivität von den falsch negative Testfehlern bei tatsächlich Kranken.
Wird ein Test bzw. seine Kriterien so definiert, dass er spezifischer wird, dann verringert sich die Sensitivität - und umgekehrt.

PPV (positive predicive value) bedeutet Richtigkeit eines positiven Ergebnisses (Nachtest-Wahrscheinlichkeit des Krankheitsvorliegens)

NPV (negative predicive value) bedeutet Richtigkeit eines negativen Ergebnisses (Nachtest-Wahrscheinlichkeit der Krankheitsfreiheit)

Merke: Für die Praxis, d.h. die Testsituation entscheidend sind PPV und allenfalls NPV, denn die Testsituation ist nicht nur duch die Qualität des Testverfahrens, sondern im gleichen Ausmass durch die Wahrscheinlichkeit einer Erkrankung gegeben.

=> Diagnostischer Informationsgewinn = Unterschied positive Nachtest- minus Vortest-Wahrscheinlichkeit einer Krankheit

=> "Absicherungs"-Informationsgewinn = Unterschied negative Nachtest- minus Vortest-Wahrscheinlichkeit der Krankheitsfreiheit

PPV, NPV und Informationsgewinn in Abhängigkeit von von Vortest-Wahrscheinlichkeit (= Prävalenz) bei einer Spezifität von 95% und Sensitivität von 95%

Vortest-Wahrscheinlichkeit	PPV	NPV	Diagnostischer Informationsgewinn	Absicherungs-Informationsgewinn
99%	99,9%	16%	0,9% *	15% +
95%	99,7%	50%	5%	45%
90%	99%	68%	9%	58%
75%	98%	86%	23%	61%
50%	95%	95%	45%	45%
25%	86%	98%	61%	23%
10%	68%	99%	58%	9%
5%	50%	99,7%	45%	5%
1%	16%	99,9%	15%	0,9%
0,5%	9%	99,97%	8,5%	0,5%
0,1%	2%	99,99%	1,9%	0,09%

* PPV - Vortest-Wahrscheinlichkeit
\+ NPV - (100% - Vortest-Wahrscheinlichkeit)

Ein gutes Konzept zur Beurteilung von Testqualitäten sind auch Wahrscheinlichkeitsraten, sog. likelihood ratios, welche die durch Sensitivität und Spezifität gegebene Testleistung in einer Zahl zusammenfassen. Die Likelihood ratio für ein positives Testergebnis ist das Verhältnis der richtig positiven Testergebnisse (a/a+c [= Sensitivität]) zu den falsch positiven Testergebnissen (b/b+d [= 100%-Spezifität]). Die Likelihood ratio für ein negatives Testergebnis ist als Verhältnis der falsch negativen Testergebnisse (c/a+c) zu den richtig negativen Testergebnissen (d/b+d) definiert. Mit der likelihood ratio lässt sich mittels Lineal und speziellem Nomogramm aus der Vortest-Wahrscheinlichkeit unmittelbar die Nachtest-Wahrscheinlichkeit eines Testergebnisses berechnen. Je niedriger die Rate, umso höher ist die Wahrscheinlichkeit, dass es sich bei einem negativen Test um einen wahrhaft negativen Test handelt. Mit Hilfe des Nomogramms von Fagan können die Nachtest-Wahrscheinlichkeiten einfach abgelesen werden. Wir suchen die Vortest-Wahrscheinlichkeit auf der linken Skala auf (z.B. Ihre Schätzung von 30%), ziehen eine Linie zur mittleren Skala der 'likelihood ratio' (z.B. 1.6 für einen positiven Test und 0.12 für einen negativen Test) und lesen auf der rechten Skala die Nachtest-Wahrscheinlichkeit ab.

Siehe Anhang IV
Nomogramm von Fagan. Abb. aus: Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine - How to practice & teach EBM. Churchill Livingstone, New York/Edinburgh 1997

Merke: Ein Test hat (nur) einen Informationsnutzen mit therapeutischen Konsequenzen, wenn bereits einiges für das Vorliegen einer Erkrankung bzw. der gesuchten Erkrankung spricht und die Vortest-Wahrscheinlichkeit bereits beträchtlich ist (je nach Spezifität und Sensitivität auch tiefer). Bei sehr niedriger Wahrscheinlichkeit ist die Durchführung eines Tests wertlos, weil ein positives Resultat wegen der hohen Fehlerrate nicht verwertet werden kann. Ist andererseits der klinisch-anamnestische Krankheitsverdacht bereits sehr hoch, ist ein Test ebenfalls von zweifelhaftem Nutzen, weil ein negatives Resultat mit einiger Wahrscheinlichkeit als falsch negativ zu betrachten ist und eine (behandlungsbedürftige) Krankheit und eine Behandlung nicht ausschliesst.

Vor der Durchführung/Verordnung einer Untersuchung sollte Klarheit über den möglichen Informationsgewinn bestehen; eine Untersuchung ist nur dann angezeigt, wenn aufgrund der Untersuchung (und ihrer rationalen Interpretation!) eine andere therapeutische Entscheidung resultiert.
Eine wichtige ärztliche Leistung ist heute der Mut, auf irrationale Absicherungs-Untersuchungen zu verzichten. Die oft bemühte Litigations-Angst scheint nicht mehr stichhaltig, wie auch erste juristische Gutachten festhalten:

Siehe Anhang V
Zusammenfassung von: Kuss E et al. Welcher Nutzen und welcher Schaden kann von Screening- und Routineuntersuchungen erwartet werden und von deren Unterlassung? *Geburtsh. Frauenheilk.* 1991; 51: 415-430

3. Trugschlüsse, Fehlermöglichkeiten und Aussagewert verschiedener Studientypen

Studien lassen sich einteilen in:

prospektiv <-> retrospektiv

Sind die Daten "vorausschauend" gesammelt worden, bevor Krankheitsereignisse oder Therapieergebnisse eingetroffen sind, oder sind z.B. Krankengeschichten "rückblickend" untersucht worden?

Verlaufsstudie <-> Querschnittsstudie

Sind Merkmale und Erkrankungen über einen längeren Zeitraum erfasst worden, sodass die zeitliche Abfolge der Ereignisse erkannt werden kann, oder sind Ereignisse und Merkmale mehr oder weniger gleichzeitig im "Querschnitt" erfasst worden?

Beobachtung <-> Intervention (experimentell)

Handelt es sich um eine Studie, die Vergleichsgruppen braucht, die sich sozusagen gewohnheitsmässig ergeben, oder werden Vergleichsgruppen "experimentell" beispielsweise durch das Los randomisiert gebildet? (Randomisierung = randomization bedeutet das Herstellen von Vergleichsgruppen durch das Los. Random samples sind Zufalls-Stichproben.)

kontrolliert <-> nicht kontrolliert

Hat die Studie eine Vergleichsgruppe mit keiner oder einer anderen Behandlung oder sind die Ergebnisse anhand von (Kohorten-)Untergruppen gewonnen worden, d.h. ohne vergleichbare Kontrollgruppe?

multizentrisch <-> "monozentrisch"

Handelt es sich um eine Studie, die in mehreren Institutionen oder Praxen durchgeführt wird oder um eine Studie, in der alle Erhebungen und Interventionen in einer Klinik durchgeführt werden?
Diese Unterscheidung kann bei allen Studientypen gemacht werden.

deskriptiv <-> analytisch

Ist die Studie nur beschreibend (z.B. Krankheitshäufigkeiten) oder macht sie beispielsweise Untergruppen-Analysen? Ein nicht sehr wichtiges Begriffspaar, das sich in der Literatur ab und zu findet.

Diese Merkmale können kombiniert auftreten.

Als Studientypen (study design) kennen wir:

3.1. Die Fallserie (case series)

Fallbeobachtungen lassen definitive Schlüsse über Wirkungszusammenhänge nicht zu, weil verschiedene Einflüsse oder Störfaktoren nicht kontrolliert werden können: Günstiger Spontanverlauf kann Therapieerfolg vortäuschen, "Droge Arzt" oder Placebo kann Erfolg ausmachen und nicht die spezifische Therapie, Zufall kann eine grosse Rolle spielen etc. Aussagekräftiger, aber auch nicht beweiskräftig, sind cross-over-Therapieversuche an einem Fall.
Sorgfältige Fallbeschreibungen können hingegen zu wertvollen Hypothesen führen, welche dann in kontrollierten Studien zu überprüfen sind.

3.2. Die n-of-1-Studie (n-of-1 trial)

ist eine kontrollierte Fallstudie und stellt in der Literatur einen neuen Studientyp dar. Sie ist durchführbar bei chronisch-stabilen Krankheitsverläufen, wobei zwei Therapien bzw. Therapie und Nicht-Therapie überkreuzt beim gleichen Patienten getestet werden. Die Wahl der Ersttherapie erfolgt durch das Los, und eine grössere Zahl solcher Therapieergebnisse kann mit Meta-Analyse-Technik statistisch ausgewertet werden.

Im Grunde genommen sind die altbekannten "Anwendungsbeobachtungen" mit neuen oder alternativen Therapien etwa bei therapieresistenten Patienten solche "cross-over"-Versuche und können als kontrollierte Studien betrachtet werden. Ein häufiges Problem publizierter "Anwendungsbeobachtungen" besteht aber darin, dass selektiv nur die günstigen Fälle berichtet werden, wobei der "zufällige" Einfluss des Spontanverlaufs ignoriert wird. Die randomisierte Zuordnung der Ersttherapie wäre ein kräftiges Mittel, um die Aussagekraft solcher Fallstudien zu verbessern.

3.3. Querschnittsstudien (cross-sectional study)

lassen Häufigkeiten von Erkrankungen ("Prävalenzstudien") oder manchmal Zusammenhänge zwischen Erkrankung und zeitgleich vorhandenen Merkmalen untersuchen und sind unter relativ kleinem Aufwand durchführbar. Einen Kausalitäts-Beweis lässt dieser Typ aber schon deshalb nie zu, weil der zeitliche Ablauf zwischen Krankheit und der vermuteten Erklärung unklar bleibt.
Wenn z.B. Rückenschmerz-Patienten gehäuft gewisse radiologische Zeichen degenerativer Veränderungen an der Wirbelsäule aufweisen, ist damit noch lange nicht gesagt, dass die degenerativen Veränderungen Ursache der Beschwerden sind; die Veränderungen am Skelett könnten auch Folge trophischer Störungen durch die die Rückenschmerzen begleitenden Muskelverspannungen sein, welche die regionäre Zirkulation kompromittieren. Die degenerativen Veränderungen wären dann Folge und nicht Ursache der Rückenschmerzen. Querschnittsbeobachtungen können diese Frage nicht klären.
In Querschnittstudien können Hypothesen eventuell näher eingegrenzt werden.

3.4. Kohortenstudien (cohort study)

sind prospektiv und in ihnen wird am Anfang eine Gruppe mit gewissen einheitlichen Kriterien (z.B. junge, Männer, gesunde etc.) einbezogen und über längere Zeit verfolgt, wobei das Auftreten bestimmter Erkrankungen in Abhängigkeit von am Anfang vorhandener Faktoren bzw. Merkmale analysiert wird. Eine der bekanntesten ist die Framingham-Studie. Der Einbezug der Probanden ("inception cohort") muss nach definierten Kriterien erfolgen.
Dieser Typ kann genaue Aussagen über die zeitliche Abfolge von Ursache bzw. Krankheitsvorläufer und Erkrankung machen. Wenn z.B. beobachtet worden ist, dass hospitalisierte Hirnschlag-Patienten einen durchschnittlich hohen Blutdruck aufweisen, so kann erst in der Kohortenstudie gezeigt werden, dass die Hypertonie einem Hirnschlag vorausgeht und der Blutdruckanstieg nicht Folge des akuten Hirnschlages ist.
Hauptächlicher Nachteil dieser Studie ist der im Vergleich zur Querschnitts- und retrospektiven Fallkontrollstudie grosse Aufwand.
Zur Interpretation: In Kohortenstudien wird die Wirkung einer Exposition (Risikofaktor oder Therapie) typischerweise mit dem Mass des *relativen*

Risikos angegeben (= Inzidenz unter Risiko/Therapie : Inzidenz ohne Risiko/Therapie; siehe 4.). Kohortenstudien zur Überprüfung einer Therapiewirkung sind aus Gründen eines immer möglichen Selektions-Bias (siehe 1.1.) praktisch immer unschlüssig. Nur wenn das Ergebnis sehr weit von 1 entfernt ist (z.B. 0.3 oder 3.0), kann u.U. angenommen werden, dass eine echte Wirkung vorhanden ist. Die Betrachtung der Gesamt-Mortalität und -Morbidität erlaubt in der Regel eine einfache Nachprüfung eines Selektions-Bias (siehe Übung III/IV). Ein (statistisch signifikanter) Zusammenhang zwischen Risikomerkmal und späterer Erkrankung muss unabhängig von anderen Merkmalen vorhanden sein, um als echter Risikofaktor zu gelten; ein confounding bias (siehe 1.1.) muss ausgeschlossen werden.

Siehe Anhang VI
Abstract von: Hulley SB et al. Epidemiology as a guide to clinical decisions: The association between triglycerides and coronary heart disease. *N Engl J Med* 1980: 302: 1383-1389
Der klare statistische Zusammenhang zwischen Triglyzeriden und Infarktrate lässt sich durch den Zusammenhang zwischen Plasma-Cholesterin und Infarktrate erklären; Triglyzeride sind ein abhängiges Risikomerkmal und damit ein unechter Risikofaktor.

Merke: Abhängigkeit und Unabhängigkeit von Risikofaktoren lassen sich nicht durch die statistische Analyse erfassen. Dass die Triglyzeride vom Plasma-Cholesterin abhängig sind und nicht umgekehrt, ist ein biologisch-klinischer Entscheid. Heute geht man davon aus, dass das Gesamtcholesterin ebenfalls ein unechter Risikofaktor und abhängig vom HDL/LDL-Quotienten ist.
Hinter jedem "unabhängigen" Risikofaktor kann sich ein bisher unbekannter tatsächlicher Faktor befinden. Die kausale Rolle kann auch aus diesem Grund erst durch kontrollierte Studien gesichert werden.

3.5. Fallkontrollstudien (case control studies)

gehen von (z.B. aus Krankengeschichten gesammelten) Krankheitsfällen aus und vergleichen diese mit ein bis mehreren "Kontrollen", d.h. anderen Patienten oder Personen, die in bekannten Charakteristika (Alter, Geschlecht, soziale Schicht etc.) gleich sind, aber die Krankheit nicht aufweisen.
Sie haben dadurch den Vorteil, dass Faktoren für seltene Erkrankungen untersucht werden können.
Obwohl dieser Typ auch etwas salopp Retrospektiv-Studie genannt wird, kann eine Fallkontrollstudie auch prospektiv ("matched pairs") durchgeführt werden und ist der nächstbessere Ersatz, wenn eine Randomisierung von der Mitmach-Bereitschaft her an Grenzen stösst (z.B. Hausgeburts-Studie).

Nachteile sind die fehlende Kontrollmöglichkeit gewisser Biases wie z.B. Selektion, Überdiagnose, Erinnerungslücken ("recall bias") etc. Das Ausmass einer Wirkung wird wegen dieser unkontrollierten Biases mit diesem Typ gerne überschätzt (z.B. "healthy screenee bias" bei präventiven Massnahmen; siehe etwa Fallkontrollstudien zum Mammographie-Screening).
Weil die Krankheits-Inzidenz in Fallkontrollstudien unbekannt bleibt, müssen an deren Stelle sogenannte "Odds" berechnet werden, das Ergebnis von Fallkontrollstudien wird rechnerisch entsprechend als Odds Ratio wiedergegeben (siehe 4.).

3.6. Kohorten-Vergleiche (Historische oder geographische Kontrollen)

Dabei handelt es sich um den Vergleich von Patienten- oder Bevölkerungsgruppen, die nicht experimentell und nicht gleichzeitig gebildet worden sind. Eine historische Kontrolle ist eine vergleichbare Gruppe, die früher in der Literatur beschrieben worden ist, eine geographische Kontrolle ist eine vergleichbare Gruppe aus einer anderen Region oder Stadt, die in der Regel gleichzeitig untersucht wird. Der Vorteil besteht im relativ geringen Aufwand, Nachteile bestehen in der mangelhaften Kontrolle von Gleichheit der Gruppen und der Intervention (manchmal kann die Veränderung des Behandlungs-Settings und nicht die spezifische Therapie einen Unterschied ausmachen; Therapien verändern sich über die Zeit).
Eine Möglichkeit, die Gruppenvergleichbarkeit zu überprüfen, ist unter Umständen die nichtspezifische Mortalität, die "Gesamt-Sterbeprognose". Ein Beispiel zeigen Übung III und IV (siehe 6.2.): Der Unterschied in der Gesamt-Sterbeprognose in der ungeeigneten Vergleichsgruppe, und nicht die Brust-Selbstuntersuchung, erklärt somit die angeblich tiefere Brustkrebs-Mortalität.

3.7. Die randomisiert kontrollierte Studie (randomised controlled trial)

Randomisierung heisst Herstellung der Vergleichsgruppen durch ein Losverfahren; Behandlungs- und Kontrollgruppe werden nach dem Zufallsprinzip gebildet.
Bei adäquater Gruppengrösse ergibt sich dadurch eine gleichmässige Verteilung von prognostisch bedeutsamen Patientenmerkmalen auf beide Gruppen. Das Zufalls-Verfahren ist einem "Matching" (vgl. Fallkontroll-Studie) in der Aussagekraft klar überlegen, weil auch unbekannte, nicht erfassbare Einflussfaktoren nach dem Zufall gleichmässig verteilt werden. Unbekannte Einflussfaktoren sind oft in ihrem Ausmass wichtiger als die bekannten prognostischen Faktoren.

Entscheidend für den Wert einer Studie ist jedoch nicht nur das prinzipielle Design, sondern Fragen wie die Praxisrelevanz der geprüften Therapien, eine unvoreingenommene Beurteilung durch Beurteiler-Verblindung und eine sinnvolle Wahl der "Endpunkte".
Eine doppelte Verblindung (Patient und Arzt weiss nicht, ob Medikament ein Verum oder ein Placebo ist) ist nicht immer sinnvoll, weil damit mehr eine akademische als eine praktisch relevante Frage beantwortet wird; die "Placebowirkung" ist jeder ärztlichen Handlung inhärent und sollte je nach praktischer Fragestellung gar nicht ausgeklammert werden. Werden Therapien nämlich zu weitgehend standardisiert und damit künstlich und praxisfremd, kann dies die Aussagekraft einer Studie vernichten. Wichtiger ist oft ein Vergleich einer neuen Therapie mit einem bekannten Standard.
Als Kriterium der Wirksamkeit gilt es auch die vom Patienten wahrgenommene Linderung zu berücksichtigen, und auch eine sogenannte Placebowirkung kann wirksam und für den Patienten leidensmindernd sein und darf entsprechend Teil einer ärztlichen Behandlung ausmachen. "Placebo" ist nicht einfach eine inerte Pille, sondern kann aus einem ganzen therapeutischen Setting entstehen. Die entscheidenden Fragen sind die nach der relativen Wirksamkeit und Unschädlichkeit, Zweckmässigkeit und Wirtschaftlichkeit. Dies ist nicht etwa eine blosse Frage komplementärmedizinischer Verfahren, sondern eine Frage der Allgemeinmedizin schlechthin. In der Allgemeinmedizin hat sich dabei der Begriff "komprehensiv" oder "umfassend" und in der Komplementärmedizin der Begriff "ganzheitlich" für das gleiche Anliegen eingebürgert.

Es können auch (individualisierte) Therapiesysteme randomsierte miteinander verglichen werden, und auf diese Weise werden voraussichtlich in Zukunft viele entscheidende Studien durchgeführt werden und wichtige therapeutische Fragen geklärt werden.

3.7.b. Überkreuzte Therapievergleiche (Cross-over)

Wie in 3.2. (n-of-1 trial) schon angesprochen, können Therapievergleiche überkreuzt durchgeführt werden, d.h. nach einen ersten Phase werden die Behandlungen getauscht, die Kontrollgruppe enthält nun die Studienbehandlung (Verum), die Studiengruppe die Kontrollbehandlung (Placebo). Dabei muss die Dauer der Behandlungsphasen klinisch sinnvoll sein, d.h. auf Natur der Erkrankung und Wirkungsweise der Therapie abgestimmt sein (dies gilt natürlich für alle Studientypen). Ein grosser Vorteil dieses Verfahrens ist die statistisch grosse Aussagekraft (power) bei geringeren Probandenzahlen. Voraussetzung, dass dieses Verfahren angewandt werden kann, ist allerdings eine Erkrankung, die einen stabilen chronischen Verlauf aufweist. Dieser

Studientyp ist ein Sonderfall der randomisiert kontrollierten Studie.

3.8. Hierarchie der Evidenz

In der (klinisch adäquat durchgeführten!) randomisiert kontrollierten Studie können Störfaktoren (v.a. Selektions-Bias!) ausgeschaltet werden, die etwa bei Kohorten- und Fallkontroll-Studien immer wieder die Ergebnisse in Frage stellen. Deshalb ist die randomsiert kontrollierte Studie anderen Studien in der Aussagekraft überlegen und befindet sich in der Hierarchie der Evidenz an oberster Stelle.
Siehe auch Anhang VIb

3.9. Externe und interne Validität

Diese Frage bezieht sich auf alle Studientypen. Interne Validität bedeutet, dass die Studie derart aussagekräftig angelegt und fehlerfrei durchgeführt wird, dass das Ergebnis für die untersuchte Gruppe und Intervention gültig ist. Eine genügende interne Validität haben oft nur randomisiert kontrollierte Studien, welche einen immer zu vermutenden Selektions-Bias ausschliessen. Die externe Validität bezieht sich auf die Generalisierbarkeit der Ergebnisse auf entsprechende Patientengruppen und Interventionen; mit anderen Worten: Sind Studiengruppe, Art der Therapie und das Therapieziel überhaupt relevant für meine Patienten, die ich in der Praxis sehe? Eine entscheidende Rolle spielt dabei immer das basale Krankheits-Risiko der Studienpatienten im Vergleich zum Patienten, für den man eine therapeutische Entscheidung trifft (vgl. 4.: Absolutes Risiko bzw. "Number needed to treat").

3.10. Zur Komplementärmedizin

Im Prinzip gelten für komplementärmedizinische Therapieschulen und Behandlungen die gleichen Regeln.
Die Frage der Wissenschaftsmethodik und des Wirksamkeitsnachweises muss auf grundsätzlich zwei Ebenen aufgeteilt werden. In der Wissenschaft geht es einerseits um die (spekulative) Ebene der Wirkungsmechanismen, d.h. um das Generieren von möglichst widerspruchsfreien Hypothesen über Therapiewirkungen. Andererseits geht es um die ganz andere Frage der empirischen Dokumentation eines Therapienutzens unter Vermeidung von Störfaktoren und statistischen Trugschlüssen.
Für die erste Ebene der Hypothesengenerierung werden je nach Therapierichtung ganz verschiedene Wahrnehmungs- und Untersuchungsinstrumente verwendet, für die Ebene des Wirksamkeitsnachweises ist dies jedoch bedeutungslos. In Sachen Wirksamkeitsnachweis brauchen wir somit keine alternative Forschungsmethodik für die Komplementärmedizin.
Die häufig vorhandene Vorstellung, es gelte die Wirksamkeit der "Schulmedizin" oder "Komplementärmedizin" als solche zu belegen, ist falsch und überholt.

4. Wirkungsgrössen: Relative Risiken – Absolute Risiken/"Number needed to treat"

Das absolute Risiko steht dem relativen Risiko gegenüber. Die Berechnung beider Grössen ist einfach.

=> Das **relative Risiko** ist der Quotient der Inzidenz unter Behandlung/Exposition und der Inzidenz ohne Behandlung/Exposition.

=> Die Reduktion in Form des **absoluten Risikos** ist die Differenz zwischen diesen Inzidenzen.

		Krankheit +	Krankheit –	
Exposition/ Therapie	+	a	b	E
	–	c	d	NE
		D	ND	N

E: "Exposed" = a+b NE: "Non-exposed" = c+d
D: "Diseased" = a+c ND: "Non-diseased" = b+d
N: "Total number" = a+b+c+d

Inzidenz Studiengruppe: $\frac{a}{a + b}$

Inzidenz Kontrollgruppe: $\frac{c}{c + d}$

Relatives Risiko: $\frac{a}{a + b} : \frac{c}{c + d}$

Absolute Risikoreduktion: $\frac{c}{c + d} - \frac{a}{a + b}$

Number needed to treat: $1 : \left(\frac{c}{c + d} - \frac{a}{a + b} \right)$

Wir können diese Berechnung an Beispielen üben:

Beispiel 1: Helsinki Ultrasound Trial

Bei 4691 Schwangerschaften mit Routine-Ultraschall zwischen der 16. und 20. Woche traten 20 perinatale Todesfälle auf, nach 4619 Schwangerschaften ohne Routine-Ultraschall 39 Todesfälle.

Berechnung:

$$\text{Relatives Risiko} = \frac{20 : 4691}{39 : 4619} = 0{,}51 \quad (= 49\% \text{ Senkung})$$

$$\text{Absolutes Risiko} = \frac{39}{4619} - \frac{20}{4691} = 4{,}2 \text{ pro } 1000$$

(Die Reduktion der perinatalen Mortalität kam zustande durch einen Schwangerschaftsabbruch bei im Ultraschall entdeckten Missbildungen; in der Kontrollgruppe führten ausgetragene Missbildungen gehäuft zu Todgeburten. Die Rate gesund geborener Kinder war in beiden Gruppen identisch. Siehe Übungsbeispiel unten.)

Beispiel 2: MRC-Studie zur milden Hypertonie

Die aktive Behandlung einer Hypertonie (diast. 90-109 mmHg) bei 8700 Frauen und Männern zwischen 35 und 64 Jahren ergab in fünf Jahren 286 kardio-vaskuläre Krankheits-Ereignisse, in der Placebo-gruppe (n=8654) traten 352 Fälle auf.

Relatives Risiko = 0,81 (= 19 % Senkung)
Absolutes Risiko = 7,8 pro 1000

(Der Unterschied kam fast ausschliesslich durch die Reduktion zerebraler Insulte zustande. Für Herzinfarkte ergaben sich keine Unterschiede durch die Behandlung).

Beispiel 3: Zervixkarzinom in British Columbia

Die Zervixkarzinom-Mortalität bei Frauen im Alter von 30-64 in einer kanadischen Provinz mit einer hohen Zervixabstrich- Teilnahme betrug 8,3 gegen-über 16,8/100'000 jährlich in einer anderen Provinz mit praktisch fehlender Vorsorge-Aktivität.

Relatives Risiko = 0,49 (= 51% Senkung)
Absolutes Risiko = 8,5 pro 100'000 Frauenjahre

(Randomisierte Studien zur präventiven Wirksamkeit zytologischer Zervix-Abstriche fehlen.)

Wie an den Beispielen erkenntlich wird, sind nur Angaben des Therapienutzens in Form des absoluten Risikos bzw. der *Number needed to treat* aussagekräftig. Diese wird mathematisch von der Erkrankungsinzidenz (bzw. vom basalen Risiko) im gleichen Ausmass bestimmt wie durch die relative Risikoreduktion.

Wie ein Behandlungsnutzen kommuniziert wird und mit dem Patienten besprochen wird, kann individuell verschieden sein. Gemäss Beispiel 2 erfolgt durch eine antihypertensive Behandlung einer milden Hypertonie (bis diast. 110 mmHg) eine Abnahme kardiovaskulärer Erkrankungen um etwa 20%. Man kann nun sagen: "Ihr Blutdruck ist an der Grenze. Wenn wir 130 gleiche Patienten 5 Jahre lang behandeln, kann bei 1 eine Krankheit verhütet werden; das ist der Nutzen, den Sie erwarten können." Oder: Ihr Krankheitsrisiko ist 4%, eine Behandlung kann es auf 3,3% senken." Oder vielleicht versteht es der Patient am besten so: "Ihr Blutdruck ist an der Grenze. Ihre Chance, die nächsten 5 Jahre gesund zu bleiben, beträgt 96%. Durch eine Behandlung können wir diese Chance auf 96,7% erhöhen. Sollen wir diese Behandlung machen?" Es ist klar, dass die Einschätzung von Nebenwirkungen und Inkonvenienzen den Entscheid nun stark mitbestimmt. Es wird auch klar, dass eine solche Behandlung nur eine mögliche Option ist, die nicht unbedingt sein muss.

Siehe Anhang VII
Abb. aus: Bucher H. und Gutzwiller F. *Ckeckliste Gesundheitsberatung und Prävention*. Georg Thieme Verlag 1993
Die NNT (Anzahl zu Behandelnder) hängt entscheidend vom basalen Risiko ab.

Gleichzeitig ist auch ein Vergleich von Nutzen und Risiken nur durch eine Darstellung in Form absoluter Risiken möglich, d.h. in Form der Anzahl Ereignisse pro 1000 oder 100'000 Behandlungsjahre. Erst dann kann nämlich verglichen werden, wieviele unerwünschte Ereignisse einem erwünschten Ereignis gegenüberstehen (Siehe 1.4.).
Die Beispiele lassen auch erkennen, dass eine Vorsorge-Massnahme bei einer Krankheit, deren Folgen ein Individuum nur mit einer geringen Wahrscheinlichkeit treffen, von vornherein nur einen geringen Nutzen haben kann.

Das relative Risiko seinerseits ist ein gutes Mass für die Strenge eines statistischen Zusammenhangs und findet deshalb bei der theoretischen Beurteilung der möglichen kausalen oder ätiologischen Rolle eines Faktors Anwendung (auch ein hohes

relatives Risiko lässt jedoch die Frage offen, ob ein Faktor ursächlich oder nur als "Risikoindikator" bzw. "Risikomarker" an der Krankheitsentstehung beteiligt ist).
Die Verwendung des relativen Risikos für die praktische Bestimmung einer Therapie-Notwendigkeit bzw. eines Therapie-Nutzens - obschon in der Literatur üblich - muss als inadäquat bezeichnet werden.

Die manchmal erwähnte Odds Ratio ist dem relativen Risiko in etwa gleichzusetzen und weist bei Inzidenzen <10% praktisch gleiche Werte auf. Bei höheren Inzidenzen wird der Wert im Sinne des prozentualen Unterschieds aufgebläht (etwa 1.7 statt 1.5 bzw. etwa 0.3 statt 0.5). Die Odds Ratio findet vor allem bei Fallkontrollstudien Verwendung, weil dort keine Inzidenzen berechnet werden können. Mathematisch entspricht sie: (a*d/b*c).

=> *Das Kurs-Computerprogramm berechnet relative und absolute Risiken sowie die Differenz pro 1000 Behandlungsjahre*

5. p-Wert und Vertrauensintervalle Statistische Fehler I. und II. Ordnung

Diese Grössen sind Parameter der Zufalls-Mathematik und der entsprechenden Signifikanz-Testung. Wir müssen bei einer Studie abschätzen, wieweit der Zufall ein Ergebnis versursacht haben könnte.

Der mathematische Zufall "stört" die Ergebnisse von Studien durch unbekannte Störfaktoren ("confounder") so wie Störgeräusche eine Musik. Nur wenn der Geräuschpegel auf genügend niedrigem Niveau gehalten wird, wird der Klang der Melodie erkennbar. Das Mittel, den Zufalls-Pegel zu kontrollieren, ist eine genügend grosse Anzahl von Probanden, die in eine Studie einbezogen werden.

5.1. p-Wert = statistischer Fehler I. Ordnung

Der bekannte p-Wert drückt aus, wie gross die Wahrscheinlichkeit ist, dass der Zufall einen Unterschied zwischen Therapie- und Kontroll-Gruppe verursacht hat (falsch positives Studienergebnis). Ein p-Wert von 0,15 bzw. von 0,01 bedeutet eine Wahrscheinlichkeit von 15% bzw. von 1%, dass der

Zufall zu einem Unterschied geführt hat, der eigentlich nicht vorhanden ist.

Als Konvention gilt, dass ein "Zufalls-Fehler" von höchstens 5% (p = 0,05) in Kauf genommen wird, um ein Studienergebnis als "signifikant" anzuerkennen. Ein p-Wert von 0,06 gegenüber einem p-Wert von 0,04 bedeutet, dass die Wahrscheinlichkeit eines falsch positiven Studienergebnisses 6% statt 4% beträgt.

Merke: Wenn nach der Konvention ein Ergebnis mit einem p-Wert von 0,04 als "signifikant" und damit als echt akzeptiert wird, ein solches mit einem p-Wert von 0,06 aber als "nicht signifikant" verworfen wird, muss die Willkür dabei erkannt werden. "Statistisch signifikant" ist kein sakrosanktes Verdikt und kann weit weniger bedeutend sein als die praktische oder klinische Signifikanz eines Studienergebnisses.

5.2. Vertrauensintervalle

Ein p = 0,05 bedeutet ein Vertrauen, es mit 95% Wahrscheinlichkeit nicht mit einem Zufalls-Ergebnis zu tun zu haben.

So lassen sich für relative und absolute Risiken 95%-Vertrauensintervalle berechnen. Man darf von einer Studie eigentlich nie sagen, sie hätte 20% weniger Infarkte unter einer Behandlung gezeigt, denn diese 20% sind nur das zufällige Ergebnis; die wahre Senkung ist vielleicht 40% oder nur 5%. Das 95%-Vertrauensintervall bezeichnet den Bereich, in welchem der wahre Wert mit 95% Sicherheit irgendwo liegt.
Ein weites Intervall bedeutet eine grosse Zufalls-Störung (entsprechend einem hohen p-Wert), ein enges Intervall bedeutet ein relativ stabiles Ergebnis ohne grosse Zufalls-Schwankungen (entsprechend einem niedrigen p-Wert).

Merke: Vertrauensintervalle und p-Werte sagen praktisch das Gleiche aus. Vertrauensintervalle lassen sich jedoch im Gegensatz zu p-Werten graphisch gut darstellen, was insbesondere bei der Ergebnis-Darstellung aus mehreren Studien angenehm ist (vgl. Meta-Analyse in Unterlage VI). Die Breite des Intervalls widerspiegelt zudem auch den statistischen Power einer Studie, sodass im Vertrauensintervall mehr Information enthalten ist als im p-Wert.

=> Das Kurs-Computerprogramm berechnet p-Wert und Vertrauensintervalle für absolute und relative Risiken

5.3. Fehler II. Ordnung

Der p-Wert drückt die Wahrscheinlichkeit aus, mit der der Zufall einen Unterschied zwischen Therapie- und Kontroll-Gruppe verursacht hat (falsch positives Studienergebnis = Fehler I. Ordnung).

Der Fehler II. Ordnung bezeichnet die Wahrscheinlichkeit falsch negativer Studienergebnisse. Wenn eine Studie zu klein ist, können tatsächliche Behandlungsunterschiede unentdeckt bleiben bzw. zu klein sein, um "statistisch signifikant" zu erscheinen.

Beide Fehler verhalten sich zueinander wie Spezifität und Sensitivität (siehe 2.2.). Beide Fehler können aber gleichzeitig verringert werden durch eine Vergrösserung der Studienpopulation.

Merke: Meta-Analysen erlauben eine bessere Kontrolle des Zufalls-"Geräuschpegels", weil durch ein Zusammenfassen mehrerer (gleichartiger) Studien die verwertbare Probandenzahl erhöht werden kann. Deutliche, aber in kleineren Einzelstudien statistisch nicht signifikante Therapieunterschiede können so "statistisch signifikant" werden, und falsch positive Einzelstudien können als Zufallsergebnisse entlarvt werden. Der statistische Prozess der Meta-Analyse darf aber dann nicht überbewertet werden, wenn Zweifel an der Gleichartigkeit der verwendeten Studien besteht.

Zahlreiche Studien mit Cholesterinsenkern zeigten übereinstimmend eine Tendenz zur Zunahme nichtkardialer Todesfälle unter der Therapie. Durch Meta-Analyse konnte berechnet werden, dass eine solche Mortalitätszunahme hochsignifikant ist, d.h. weit über den Zufall hinaus vorhanden ist.

Siehe Unterlage V:
Schmidt JG. Cholesterol lowering treatment and mortality. *BMJ* 1992; 305: 1226-1227

=> *Diese Meta-Analyse ist mit dem Kurs-Computerprogramm durchgeführt worden.*

6. Kritische Beurteilung publizierter Studien

6.1. Vorgehen bei der Beurteilung einer Interventions-studie

Eselsbrücke: **SPION**

S

1) Studien-Design

Ist dieses überhaupt erkennbar? Querschnitt oder Längsverlauf? Prospektiv oder retrospektiv? Intervention oder Observation? Falls Intervention: Randomisierte oder andere Art der Kontrollgruppe?

P

2) Probanden-/Patientengruppe

Ist die Intervention oder die Frage nach Risikofaktoren in dieser Gruppe sinnvoll, sind die richtigen Patienten oder Probanden ausgewählt worden?

I

3) Einfluss-Faktor (Intervention, Risikomerkmal)

Sind Interventions- oder Risiko-Faktoren gut beschrieben und nachvollziehbar?

O

4) Was ist die Zielvariable bzw. die Zielerkrankung

Handelt es sich um eine klinisch "echte" Erkrankung oder um ein apparatives Surrogat (Laborkosmetik)? Sind die Therapieziele sinnvoll festgelegt worden? Sind die Endpunkte klinisch relevant?

N

5) Resultate/NNT

Sind die Resultate dargestellt in RR oder AR, sind Vertrauensintervalle angegeben? Ist das Ausmass der Wirkung bzw. der Unterschied zwischen Studien- und Kontrollgruppe klinisch wirklich relevant?

6) Aussagekraft der Studie - Vergleich mit Schlussfolgerungen der Autoren

Lässt der Studientyp die Schlussfolgerungen der Autoren überhaupt zu? Auf was für Patienten beziehen sich die Ergebnisse?

Die Reihenfolge dieser Fragen kann unter Umständen vom Vorwissen abhängen. Ist zum Beispiel das Studiendesign (Punkt S) in der Hierarchie der Evidenz gegenüber bereits publizierten Studien zur gleichen Therapie minderwertig (z.B. Fallkontroll-studie gegenüber randomisert kontrollierter

Studie), kann die Studie übergangen werden, und damit erübrigt sich auch die Beantwortung der anderen Punkte.

6.2. Übungsbeispiele

Übung I

Unterlage VI:
Graboys TB et al. Long-term survival of patients with malignant ventricular arrhythmia treated with antiarrhythmic drugs. *Am J Cardiol* 1982; 50: 437-443

1) Arrhythmiekomplikationen (plötzlicher Herztod)

2) Medikamentöse Arrhythmie-Kontrolle

3) Patienten-Einschluss: Arrhythmie-Patienten mit st.n. kardialer Reanimation. In der Regel Postinfarkt-Patienten, Homogenität der Kohorte ein Problem

4) Prospektive Kohortenstudie mit Möglichkeit einer Reihe von Biases und Confoundern:
Selektionsbias: Gute "Therapiekontrolle" wegen Compliance und leichter Erkrankungsform

5) Patienten mit erfolgreicher Arrythmie-Suppression zeigen weit bessere Prognose als Patienten mit schlecht kontrollierter Arrhythmie

6) Die Behauptung der (renomierten) Autoren, eine medikamentöse Arrhythmie-Suppression könne den plötzlichen Herztod verhindern, geht weit über das hinaus, was der Studientyp erlaubt.

Übung II

Unterlage VII:
Echt DE et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo - The Cardiac Arrhythmia Suppression Trial. *N Engl J Med* 1991; 324: 781-788

1) Ebenfalls Arrhythmiekomplikationen (plötzlicher Herztod)

2) Medikamentöse Arrhythmie-Kontrolle mit Flecainid und Encainid, welche bis anhin als überdurch-

schnittlich wirksam und gut verträglich galten

3) Postinfarkt-Patienten, Selektion adäquat nach praktisch-klinischen Gesichtspunkten (Therapie-responder)

4) Randomisiert kontrollierte Studie

5) Unter aktiver Therapie erfolgreiche Arrhytmie-Suppression, aber 4mal häufiger plötzlicher Herztod als in der Placebogruppe mit schlecht kontrollierten Arrhythmien

6) Der Schluss, dass die verwendeten Antiarrhythmika die Mortalität ungünstig beeinflussen, wird durch keine nennenswerten Studienfehler beeinträchtigt.
(Damit wird gleichtzeitig die krasse Unzulänglichkeit der vorhergehenden Studie dokumentiert, auch wenn die Patienten der ersten Studie eine Gruppe mit weit höherem Risiko bildeten.)

Übung III

Unterlage VIII:
Gästrin G. Preliminary results of primary screening for breast cancer with the Mama Program. *Soz Präventivmed* 1993; 38: 280-287

1) Brustkrebsmortalität und andere Nebenparameter wie Compliance etc.

2) Anleitung zur Brustselbstuntersuchung

3) Angeleitete Frauen, die auf Umfrage geantwortet haben

4) Kohortenstudie mit geographischer Kontrolle mit Möglichkeit von Biases: Selektionsbias
Ein in einem Nebensatz erwähnter Unterschied in der Gesamtmortalität von 30% (RR=0,7) beweist die Unvergleichbarkeit der durch die Compliance selektionierten Studienkohorte mit der geographischen Kontrollgruppe.

5) Frauen mit Selbstuntersuchung zeigen eine 30% geringere Brustkrebsmortalität.

6) Die Behauptung der Autorin, die Brustkrebssterblichkeit würde sich durch die Selbstuntersuchung um 30% reduzieren lassen, lässt das Studiendesign nicht zu. Vielmehr beweist der Hinweis auf die Gesamtmortalität, dass nur ein Selektions-, aber kein Interventions-Effekt zu dem Ergebnis geführt

hat (eine 30%-ige Reduktion der Brustkrebsmortalität kann die Gesamtmortalität theoretisch nur um 1% senken und damit nicht signifikant beeinflussen).

Übung IV

Unterlage IX:
Gastrin G, Miller AB, To T et al. Incidence and mortality from breast cancer in the Mama Program for breast screening in Finland, 1973-1986. *Cancer* 1994; 73: 2168-2174

Gleiche Studie wie vorher, nun in einer methodisch besser betreuten Zeitschrift publiziert.

5) In dieser Version wird die Gesamtmortalität "ordentlich" dargestellt.

6) Die Autoren halten nun fest, dass ein Selektionsbias die Ergebnisse relativiert. Allerdings wird die Hypothese einer möglicherweise echten Mortalitätssenkung aufrechterhalten und der Selektionsbias als eher zweitrangig abgehandelt, obwohl der Vergleich mit der Gesamtmortalität zeigt, dass die Ergebnisse durch den Selektionsbias allein zu erklären sind.
Diese insgesamt weit adäquatere Publikation kontrastiert mit den krassen Mängel der ersten Publikation; sie lässt unter anderem auch eigentliche Datenfälschungen in der ersten Publikation erkennen wie z.B. die dort mit selektiven Daten vorgetäuschte günstige Stadienverteilung.

Bemerkung: Eine grosse britische Studie zur gleichen Frage zeigte keine Wirkung der Selbstuntersuchung im Vergleich mit einer sorgfältigen geographischen Kontrolle.

Übung V

Unterlage X:
Luck CA. Value of routine ultrasound scanning at 19 weeks: a four year study of 8849 deliveries. *BMJ* 1992; 304: 1474-1478

1) Kindliche Geburtskomplikationen

2) Routine-Ultraschall in Frühschwangerschaft mit sehr interventionsfreudigem Eingriffen bei patholo-

gischen Befunden

3) Schwangere

4) Prospektive Kohortenstudie ohne Vergleichsgruppe, deskriptiv ohne Analyse von Risikomermalen etc.

5) Deskription der Befunde und Eingriffe mit der ungeprüften Annahme, Früherkennung und Frühein-griffe seien per se von Nutzen und präventiver Wirkung

6) Die Behauptung der Autorin, die Studie zeige einen Nutzen des Routine-Ultraschalls, geht über das hinaus, was der Studientyp erlaubt.

Übung VI

Unterlage XI:
Bucher HC, Schmidt JG. Does routine ultrasound scanning improve outcome in pregnancy? Meta-analysis of various outcome measures. *BMJ* 1993; 307: 13-17

1) Ebenfalls kindliche Geburtskomplikationen (Lebendgeburtsrate, perinatale Mortalität, Apgar)

2) Routine-Ultraschall in Frühschwangerschaft, Interruptio bei Missbildungen

3) Schwangere

4) Meta-Analyse von randomisiert kontrollierten Studien

5) Unter Routine-Ultraschall gleiches resultat wie bei selektivem, klinisch indiziertem Ultraschall in bezug auf Lebendgeburtsrate und Apgar; geringere perinatale Mortalität als wahrscheinlicher statistischer Artefakt.

6) Der Schluss der Nutzlosigkeit des Routine-Ultraschalls in der Frühschwangerschaft wird durch keine nennenswerten Studienfehler beeinträchtigt und durch die zusammenfassende Datenauswertung in einer Meta-Analyse erhärtet.
(Damit wird die Unzulänglichkeit der vorhergehenden Studien-Schlussfolgerung dokumentiert. Siehe Diskussion in Unterlage XII.)

Unterlage XII:
Bucher HC, Schmidt JG. Routine ultrasound scanning in pregnancy - authors' reply. *BMJ* 1993; 307: 560

7. Weiterführende Literatur

Bücher

Evidence-based medicine - How to practice & teach EBM. Sackett DL, Richardson WS et al. Churchill Livingstone, New York/Edinburgh 1997

How to read a paper. The basics of evidence based medicine. Greenhalgh Trisha. BMJ Publishing Group, London 1997

Clinical Epidemiology; A basic Science für Clinical Medicine. Sackett DL, Haynes RB, Tugwell P. Little, Brown and Company, Boston/Toronto 1991

Medical Decision Making. Sox HC, Blatt MA, Higgins MC, Marton KI. Butterworth-Heinemann, Boston 1988

„Placebo“ - Wertvoll wenn es dem Patienten nützt? Methodologie einer nutzensorientierten, pragmatischen klinischen Forschung. Schmidt J.G. (Hrsg.) Forschende Komplementärmedizin 1998, Supplement 1. S. Karger Verlag, Freiburg/D

Kritik der medizinischen Vernunft: Schritte zu einer zeitgemässen Praxis - Ein Lesebuch. Schmidt JG, Steele RE. Verlag Kirchheim, Mainz 1994

Können bei der STIFTUNG bezogen werden

Diagnostic Strategies for Common Medical Problems (2nd ed). Black ER, Bordley DR, Tape TG, Panzer RJ (eds). ACP, Philadelphia, Pennsylvania, 1999.ISBN 0-943126-74-6

Checkliste Gesundheitsberatung und Prävention. Bucher H, Gutzwiller F. Georg Thieme Verlag, Stuttgart 1993

Evidenz-basierte Medizin, Wissenschaft im Praxisalltag. M.Perleth, G.Antes. MMV Medien und Medizin Verlag, 1999

PDQ Epidemiology. Steiner LD, Norman GR. Mosby-Year Book Inc., St. Louis 1996

Epidemiology for the Uninitiated. Coggon D, Rose G, Barker DJP. BMJ Publishing Group, London 1997

Torheiten und Trugschlüsse in der Medizin. Skrabanek P, McCormick J. Verlag Kirchheim, Mainz 1991
(Original: Follies and fallacies in medicine, The Tarragon Press, Glasgow 1989)

Clinical Research Methodology for complementary Therapies. Lewith G, Aldridge D. Hodder & Stoughton, London/Sydney/Auckland 1993

A guide to effective care in pregnancy and childbirth. Enkin M et al. Oxford University Press 1995

Methodological Errors in Medical Research. Andersen B. Blackwell Scientific Publications, Oxford 1990

Studying a Study and Testing a Test: How to Read the Medical Literature. Riegelman RK, Hirsch R. Little, Brown and Company, Boston/Toronto 1989

Elektronische Medien

The Cochrane Library. Update Software

Best Evidence. Americal College of Physicians

Siehe auch Internet:

http://paracelsus-heute.ch

http://evimed.ch

Artikel zu Evidence-based medicine

[1] Guyatt GH, Rennie D. Users' guides to the medical literature [editorial]. JAMA 1993;270(17):2096-7.

[2] Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993;270(17):2093-5.

[3] Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1993;270(21):2598-601.

[4] Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. JAMA 1994;271(1):59-63.

[5] Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271(5):389-91.

[6] Jaeschke R, Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence- Based Medicine Working Group. JAMA 1994;271(9):703-7.

[7] Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. JAMA 1994;271(20):1615-9.

[8] Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. JAMA 1994;272(3):234-7.

[9] Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group [see comments]. JAMA 1994;272(17):1367-71.

[10] Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? Evidence Based Medicine Working Group. JAMA 1995;273(20):1610-3.

[11] Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1995;273(16):1292-5.

[12] Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? The Evidence-Based Medicine Working Group. JAMA 1995;274(7):570-4.

[13] Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt G. Users' guides to the Medical Literature. VIII. How to use clinical practice guidelines. B. what are the recommendations and will they help you in caring for your patients? The Evidence-Based Medicine Working Group. JAMA 1995;274(20):1630-2.

[14] Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group [published erratum appears in JAMA 1996;275(16):1232. JAMA 1995;274(22):1800-4.

[15] Naylor CD, Guyatt GH. Users' guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. The Evidence-Based Medicine Working Group. JAMA 1996;275(7):554-8.

[16] Naylor CD, Guyatt GH. Users' guides to the medical literature. XI. How to use an article about a clinical utilization review. Evidence Based Medicine Working Group. JAMA 1996;275(18):1435-9.

[17] Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. Evidence-Based Medicine Working Group. JAMA 1997;277(15):1232-7.

[18] Drummond MF, Richardson WS, OBrien BJ, Levine M, Heyland D. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1997;277(19):1552-7.

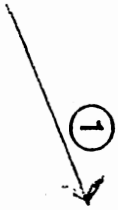
[19] O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group [published erratum appears in JAMA 1997 Oct 1;278(13):1064]. JAMA 1997;277(22):1802-6.

[20] Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. JAMA 1998;279(7):545-9.

Anhang I

aus: Schmidt J. Die Brustkrebs-Vorsorgeuntersuchung: Die kritische und praktisch relevante Beurteilung wissenschaftlicher Daten. Das Argument (Berlin) 1988; AS 178: 100-122

Abb. 1:
Modell eines disseminierten Wachstums des Mammakarzinoms



1: Konfundierende Assoziation; 2: Kausale (»unabhängige«) Assoziation, 3: Unechte (»abhängige«) Assoziation / Kausal nur in Fällen streng lokaler Determination des Verlaufs

Anhang II

Rees JL. The melanoma epidemic: reality and artefact. BMJ 1996; 312: 137-138

The melanoma epidemic: reality and artefact

Warrants a reappraisal of the relation between histology and clinical behaviour

Skin watchers have their own equivalent of the Heisenberg uncertainty principle. Just as you can't know exactly both the momentum and the position of a single electron, once you excise a pigmented lesion and know its histology you forfeit the chance of knowing what would have happened if you had left it in situ; progression, metastasis, or even involution are all possible. Removal of pigmented lesions is now one of the commonest surgical procedures and has radically changed the pattern of referrals to dermatologists. Epidemiological studies show a dramatic increase in histologically confirmed melanomas[1] and raise the important questions of whether the increase is real or a diagnostic artefact and whether we should reconsider the relation between sun exposure and melanoma.

The incidence of melanoma has increased by 3-7% per year from the mid-1950s to the early 1980s.[1-3] These changes have been seen in both sexes and in a large number of different caucasian communities in both the northern and southern hemispheres.[1-5] A smaller rise has also been seen in mortality, with a reduction in case fatality from an estimated level of over 85% in 1925[3] to under 20% today. Cohort analyses show, perhaps surprisingly, that mortality from melanoma rose from the 1890s to the 1950s and then started to decline, with forecasts that the overall mortality from malignant melanoma will peak early in the next century.[1-3 5] These cohort effects are not easily explained by changes in leisure activity. However, more recently (from the early 1980s onwards) several studies have shown steep rises in the incidence of melanoma, of 15-43% per year; in parts of New South Wales a doubling of histologically confirmed melanomas has been reported over a two year period.[1 4 6]

Armstrong suggests that these recent steep increases cannot be solely attributed to earlier diagnosis, changes in histopathological criteria for melanoma, or an increase in the proportion of excised lesions being referred for histological opinion.[1 4 6 7] In the case of New South Wales, he argues that since the increase in detection and removal of thin lesions has not been followed by a reduction in incidence of thicker lesions (which one might expect if the increase was due to earlier case ascertainment), this apparent epidemic of melanoma represents in part an increasing recognition of a pre-existing non-metastasising but invasive form of melanoma. The presence of such a lesion should not be surprising. In many, if not most, cancer systems the majority of dysplastic lesions do not progress to clinically relevant malignancy. Perhaps the clearest example is squamous cell cancer of the skin, where careful epidemiological studies have shown that, despite the presence of multiple genetic changes,[8] over 99·9% of actinic keratoses fail to progress to invasive tumours in any one year and as many as 25% may regress.[2]

Understanding the relation between melanocytes, naevi, and melanoma is considerably more confusing: some argue that melanomas are nearly always derived from pre-existing naevi,[9] while others argue that this is a rare event[10]; some believe that melanomas in the radial growth phase are incapable of metastasis (although what percentage of radial growth phase proceeds to vertical growth phase is unclear)[9]; and the never ending change in terminology for atypical or dysplastic naevi betrays the difficulty in relating histology to clinical behaviour.

Does questioning the relation between histopathological description and clinical behaviour have implications for those seeking to persuade people to change their attitude to sun exposure? The answer is surely "yes." The arguments relating melanoma to sun exposure are well rehearsed,[2] but the relation is not nearly as clear cut as it is between sun exposure and squamous cell malignancy. Most melanomas occur on skin that is only intermittently exposed; individuals with higher continuous sun exposure have lower rates than those exposed intermittently; and there seems an important interaction between skin type and incidence of melanoma. There is much else that is unclear[11]: we do not understand which part of the sun's spectrum is responsible for melanoma, nor the relative importance of ultraviolet radiation to mutagenesis, tumour promotion, or impairment of cutaneous immunity in the pathogenesis of melanoma.

The need for such understanding is underlined by models that predict that changes in the pattern of leisure exposure to the sun or that the use of sunscreens may actually increase rather than decrease melanoma risk.[12] There is after all no robust empirical evidence to defend most health promotion in this area. It has been suggested that the antithesis of science is not art but politics[13]; melanoma is perhaps an example of the two having become mistakenly intertwined. An amicable separation is required. The certainties of Health of the Nation and "slip-slap-slop" already look a little shaded: molecules care little for consensus.

JONATHAN L REES
Professor of dermatology

University Department of Dermatology,
Royal Victoria Infirmary,
Newcastle upon Tyne NE1 4LP

1 Armstrong BK, Kicker A. Cutaneous melanoma. *Cancer Surveys* 1994;19-20:219-40.
2 Marks R. An overview of skin cancers: Incidence and causation. *Cancer* 1995;75(suppl):607-12.
3 Elder DE. Skin cancer: melanoma and other specific nonmelanoma skin cancers. *Cancer* 1995;75(suppl):245-56.
4 Burton, RC, Coates MS, Hersey P, Roberts G, Chetty, MP, Chen S, *et al.* An analysis of a melanoma epidemic. *Int J Cancer* 1993;55:765-70.
5 Scotto J, Pitcher H, Lee JAH. Indications of future decreasing trends in skin-melanoma mortality among whites in the United States. *Int J Cancer* 1991;49:490-7.
6 Burton RC, Armstrong BK. Current melanoma epidemic: a nonmetastasizing form of melanoma? *World J Surg* 1995;19:330-3.
7 Burton RC, Armstrong BK. Recent incidence trends imply a nonmetastasizing form of invasive melanoma. *Melanoma Research* 1994;4:107-13.
8 Rehman I, Quinn AG, Healy E, Rees JL. High frequency of loss of heterozygosity in actinic keratoses, a usually benign disease. *Lancet* 1994;344:788-9.
9 Clark WH, Elder DE, Van Horn M. The biologic forms of malignant melanoma. *Human Pathology* 1986;17:443-50.
10 Ross PM. Apparent absence of a benign precursor lesion: implications for the pathogenesis of malignant melanoma. *J Am Acad Dermatol* 1989;21:529-38.
11 Diffey BL, Healy E, Thody AJ, Rees JL. Melanin, melanocytes, and melanoma. *Lancet* 1995;346:1713.
12 Armstrong BK. The epidemiology of melanoma: where do we go from here? In: Gallagher RP, Elwood JM, eds. *Epidemiological aspects of cutaneous malignant melanoma.* Boston: Kluwer, 1994:307-22.
13 Carey P, ed. *The Faber book of science.* London: Faber and Faber, 1994.

Anhang III

aus: Bucher H. und Gutzwiller F. Ckeckliste Gesundheitsberatung und Prävention. Georg Thieme Verlag 1993

Tabelle 2 Voraussetzungen für den sinnvollen Einsatz von Screeningtests

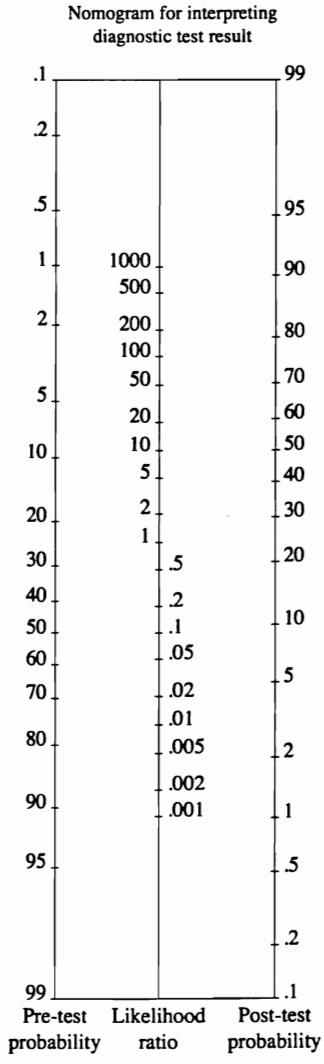
Bedingungen der zu screenenden Krankheit oder des Risikofaktors:	• Bedeutsame Auswirkungen auf Lebensqualität, Morbidität und Mortalität • Genügend hohe Prävalenz in der Bevölkerung • Natürlicher Verlauf der Krankheit hinreichend bekannt • Asymptomatische Krankheitsperiode, während welcher Frühdiagnose und Intervention eine bedeutsame Reduktion von Morbidität und Mortalität erwarten lassen
Bedingungen an den Screeningtest:	• Hohe Sensitivität, Spezifität und Prädiktivität • Hohe Reproduzierbarkeit mit geringer inter- und intraindividueller Variabilität von Meßresultaten • Hohe Testsicherheit. Da der Test an großen Kollektiven angewandt wird, sind seltene Komplikationen bedeutsam • Kostengünstig: Folgekosten durch nötige Zusatzabklärungen müssen berücksichtigt werden • Einfache Applizierbarkeit, vertretbarer Arbeits- und Materialaufwand • Hohe Akzeptanz bei Arzt und bei Patienten • Höhere Effektivität bei früher als bei später Intervention
Bedingungen des Interventionsprogrammes:	• Vorteile müssen gesundheitliche und andere Risiken überwiegen (an sich geringe Nebenwirkungen können durch hohe Verbreitung einer Maßnahme relevant werden) • Ressourcen für Follow-up-Diagnostik und therapeutische Interventionen müssen in ausreichender Form vorhanden sein • Hohe Akzeptanz bei Arzt und Patient

Anhang IV

aus: Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine - How to practice & teach EBM. Churchill Livingstone, New York/Edinburgh 1997

Card 3B The Fagan Likelihood Ratio nomogram

Adapted from Fagan T J 1975 Nomogram for Bayes's Theorem (c). New England Journal of Medicine 293:257



From Sackett, Richardson, Rosenberg & Haynes: Evidence-Based Medicine; How to Practice and Teach EBM. London: Churchill Livingstone, 1997

Anhang V

aus: Kuss E et al. Welcher Nutzen und welcher Schaden kann von Screening- und Routineuntersuchungen erwartet werden und von deren Unterlassung? Geburtsh. Frauenheilk. 1991; 51: 415-430

E. Kuss[1], M. Tryba[2], R. Kürzl[1], K. Ulsenheimer[3]

[1] I. Frauenklinik der Universität München. [2] Universitätsklinik für Anästhesiologie, Intensiv- und Schmerztherapie, Bochum. [3] Rechtsanwaltskanzlei, München

Zusammenfassung

Die Frage nach dem Nutzen und Schaden von Screening- und Routineuntersuchungen und von deren Unterlassung wird in vier Abschnitten beantwortet. Im ersten Abschnitt werden die Methoden beschrieben, mit denen der Erkenntniswert medizinischer Untersuchungen geschätzt werden kann. Es werden die Begriffe diagnostische Sensitivität, diagnostische Spezifität einer Untersuchung sowie Prätest- und Posttest-Wahrscheinlichkeit einer Diagnose definiert. Anschließend wird dargestellt, wie die Begriffsinhalte voneinander abhängen und wie deren numerische Werte errechnet werden können (»Bayes-Theorem«). Im zweiten und dritten Abschnitt wird unter Berücksichtigung der oben genannten Zusammenhänge der Wert präoperativer Routineuntersuchung aus anästhesiologischer Sicht erörtert, aus gynäkologischer Sicht außerdem auch der Wert anderer Screening- und auch Nachsorge-Untersuchungen. Präoperative Laboratoriumsuntersuchungen sind dann, aber auch nur dann notwendig, wenn sorgfältige Erhebung der Anamnese und sorgfältige körperliche Untersuchung Hinweise auf Organschäden und Risikofaktoren ergeben. Der Nutzen klinisch-chemischer Screening- und Nachsorge-Routineuntersuchungen, deren Durchführung Frauenärzten empfohlen wird, ist gering. Dies ist, wie belegt wird, zum einen die Folge des hohen Anteils „gesunder" Frauen unter den Patientinnen der Frauenärzte, zum anderen die Folge der Tatsache, daß im Rahmen der gynäkologischen Onkologie die Behandlung eines frühzeitig erkannten Rezidivs nicht nachweislich erfolgreicher ist als die eines später entdeckten. Im vierten Abschnitt wird schließlich ausgeführt, daß keine nachteiligen forensischen Folgen zu erwarten sind, wenn diagnostische Untersuchungen wegen nachweislich geringem Erkenntniswert unterlassen werden. Wird ein einschlägiges Rechtsverfahren eröffnet, dann muß der medizinische Sachverständige den Erkenntniswert der eingeklagten Untersuchung objektiv aus der Sicht ex-ante bestimmen; hierzu dienen die im ersten Abschnitt besprochenen Rechenverfahren. Unter Berufung auf die Judikatur von vornherein „aus Sicherheitsgründen" Diagnostik zu betreiben, aus der sich mit hoher Wahrscheinlichkeit keine therapeutischen Konsequenzen ergeben, ist nicht gerechtfertigt; die fehlende medizinische Indikation darf nicht durch eine „forensische Indikation" ersetzt werden. Diagnostische Untersuchungen wegen mangelnder Erfolgswahrscheinlichkeit zu unterlassen, bleibt nicht nur ohne nachteilige forensische Folge, sondern ist aus ärztlichen, wirtschaftlichen und ethischen Gründen geboten.

Let us have not more science,
but better science and truer science
in the medical curriculum (53).

Weltweit erarbeiten Wissenschaftler, Habilitanden, Doktoranden in Universitäten und Industrien neue Möglichkeiten, Bekanntes besser oder jedenfalls anders zu analysieren und bisher Unbekanntes oder Unmeßbares nachweisbar oder meßbar zu machen. Täglich erscheinen Journale und publizieren die Ergebnisse in der ersten oder einer weiteren Ausführung. Diese Aktivitäten könnten uns der Wahrheit stetig näher bringen, dem praktizierenden Arzt die diagnostische Aufgabe fortwährend erleichtern, das Vertrauen des Patienten zunehmend steigern und der Gesellschaft den Unterhalt ihres Gesundheitswesens stetig erleichtern, würden wir in einer idealen Welt leben. Arzt, Patient und Gesellschaft der realen Welt wissen oder fürchten, daß dem nicht so ist.

Komplexität von Gesundheit und Krankheit und die variablen Vorstellungen davon, Ignoranz und Oberflächlichkeit, Spezialistentum, Phobien, Geltungsbedürfnis, „publish or perish", Marketing-Erfordernisse, Verkaufsargumente, Journalismus, Gewinnstreben, Dogmatismus, Anspruchshaltung, Jurisdiktion und was darüber kolportiert wird, Vollständigkeitswahn, „universitärer Standard", Kostenindifferenz, Sicherheitssuggestion und andere Faktoren, mehr oder weniger gleichmäßig auf alle Beteiligten verteilt, induzieren diagnostischen Aktionismus und progrediente Unsicherheit trotz und wegen der Fülle der Ergebnisse. In Deutschland trägt der nicht sehr hoch eingeschätzte Stand der klinischen Forschung (88) das Seine dazu bei. Der folgende Beitrag soll dem praktizieren-

* Ergebnis der Arbeitsgemeinschaft Klinische Chemie und Biochemie in Frauenkliniken. Deutsche Gesellschaft für Gynäkologie und Geburtshilfe. 48. Kongreß, Hamburg 11.09.1990

Geburtsh. u. Frauenheilk. 51 (1991) 415–430

Anhang VI

aus: Hulley SB et al. Epidemiology as a guide to clinical decisions: The association between triglycerides and coronary heart disease. N Engl J Med 1980: 302: 1383-1389

SPECIAL ARTICLE

EPIDEMIOLOGY AS A GUIDE TO CLINICAL DECISIONS

The Association between Triglyceride and Coronary Heart Disease

STEPHEN B. HULLEY, M.D., M.P.H., RAY H. ROSENMAN, M.D., RICHARD D. BAWOL, PH.D., AND RICHARD J. BRAND, PH.D.

Abstract The hypothesis that triglyceride is a cause of coronary heart disease, although unconfirmed and never universally accepted, has nonetheless strongly influenced the practice of preventive medicine. We have examined the epidemiologic association between triglyceride and coronary heart disease to evaluate the validity of inferring that there is a causal relation between the two. Neither the evidence from published studies nor an analysis of data from the Western Collaborative Group Study provides strong support for the causal hypothesis. Information from other scientific disciplines is also meager, contrasting with the coherence of diverse evidence supporting the hypothesis that cholesterol is a cause of coronary heart disease.

These arguments fall short of disproving the belief that lowering triglyceride will prevent coronary heart disease, especially since triglyceride and cholesterol are inextricably associated through mutual lipoprotein carriers. But we propose that the ethics of preventive medicine place the burden of proof on the proponents of intervention. We therefore recommend that widespread screening and treatment of healthy persons for hypertriglyceridemia be abandoned until more persuasive evidence becomes available. (N Engl J Med. 1980; 302:1383-9.)

IN 1959, Albrink and Man found high serum levels of triglyceride in men with a history of myocardial infarction[1] and proposed this lipid as a cause of coronary heart disease. The important clinical implications of the hypothesis spawned many attempts to confirm it, but two decades later the controversy over whether serum triglyceride is an appropriate target for efforts to prevent coronary heart disease remains unresolved. The basic epidemiologic association has been confirmed by many studies,[2-33] but its biologic basis is uncertain: Is triglyceride a cause of coronary heart disease, or are both triglyceride and coronary heart disease consequences of some third, confounding variable?

Ideally, firm answers to these questions should have preceded clinical efforts directed at triglyceride. But such answers are often elusive, and the attractiveness of positive action has led many physicians to decide that triglyceride should be a target for clinical management. This judgment has evolved over the years. The initial enthusiasm for screening programs and for prescribing diet and drugs for otherwise healthy persons with hypertriglyceridemia[34,35] has given way to a reactionary phase, and the evidence that triglyceride is an independent causal risk factor is now considered unconvincing by many authorities.[14,37-39]

Despite this trend in the prevailing scientific judgment, strong recommendations for screening and treatment continue to appear. The 1980 edition of *Harrison's Principles of Internal Medicine* suggests clofibrate for asymptomatic persons with serum triglyceride levels exceeding 300 mg per deciliter,[40] and a recent review concludes that "whether or not drug therapy is used, a dietary program should be prescribed."[41] The policy extends even to children, as in the 1978 recommendation of the American Heart Association "that children with elevated cholesterol or triglyceride be placed on an appropriate diet."[42] Triglyceride is often lumped together with cholesterol in this fashion, which tends to discourage physicians from making separate judgments for the two hyperlipidemias. Interest in triglyceride is also sustained by the profit that its measurement and reduction represent for the pharmaceutical, laboratory, and health-care-delivery industries.[43,44]

The result is a problem for the practitioner, who must decide whether to identify and treat hypertriglyceridemia in otherwise healthy patients. In addition, there is the more general problem of the manner in which the medical profession should create and implement a coherent policy on the issue. The process of

From the Department of Epidemiology and International Health, University of California School of Medicine, San Francisco; the Harold Brunn Institute, Mount Zion Hospital and Medical Center, San Francisco; and the Department of Biomedical and Environmental Health Sciences, School of Public Health, University of California, Berkeley (address reprint requests to Dr. Hulley at the Department of Epidemiology and International Health, University of California, San Francisco, CA 94143).

Supported by a Research Grant (HL 03429) from the National Heart, Lung, and Blood Institute and by the Robert Wood Johnson Foundation Clinical Scholars Program.

Anhang VIb

Richtlinien zur Qualitätsbeurteilung der wissenschaftlichen Evidenz medizinischer Interventionen

(gemäss Canadian Task Force for the Periodic Health Examination)

I	Evidenz von mindestens einer adäquaten randomisiert kontrollierten Studie
II–1	Evidenz von mindestens einer kontrollierten, nicht randomisierten Studie
II–2	Evidenz von Kohorten oder Fall–Kontrollstudien, nach Möglichkeit unabhängig an mehreren Orten durchgeführt
II–3	Evidenz von Vergleichsstudien, die behandelte und unbehandelte Populationen in verschiedenen Zeitabschnitten oder an verschiedenen Orten vergleichen
III	Meinungen von respektierten Experten über ihre klinische Erfahrung oder Berichte von Expertengremien

---

Empfehlungen zu einer Therapie

A	gute Evidenz, um Therapie/Vorgehen zu empfehlen
B	einige Evidenz, um Therapie/Vorgehen zu empfehlen
C	wenig Evidenz, um Therapie/Vorgehen zu empfehlen
D	einige Evidenz, um Therapie/Vorgehen nicht zu empfehlen
E	gute Evidenz, um Therapie/Vorgehen nicht zu empfehlen

Anhang VII

aus: Bucher H. und Gutzwiller F. Ckeckliste Gesundheitsberatung und Prävention. Georg Thieme Verlag 1993

Tabelle 7 Auswirkung des basalen Risikos und relativer Risikoreduktion auf die Anzahl zu Behandelnder (aus: Sackett, D.L., et al.: Clinical Epidemiology. Little Brown & Company, Boston 1991)

	Relative Risikoreduktion unter Behandlung						
Basales Risiko (ohne Behandlung)	50%	40%	30%	25%	20%	15%	10%
0,9	2	3	4	4	6	7	11
0,6	3	4	6	7	8	11	17
0,3	7	8	11	13	17	22	33
0,2	10	13	17	20	25	33	50
0,1	20	25	33	40	50	67	100
0,05	40	50	67	80	100	133	200
0,01	200	250	333	400	500	667	1000
0,005	400	500	667	800	1000	1333	2000
0,001	2000	2500	3333	4000	5000	6667	10000

Anhang VIII Zum Computerprogramm

STATISTICAL TESTS FOR PROPORTIONS (Up-Date April 1996)

(c) Johannes G. Schmidt, CH-8840 Einsiedeln

The program includes:

A. CHI-SQUARE-TEST for independent groups

B. McNEMAR for paired groups

C. MANTEL-HAENSZEL for meta-analysis of several studies, including test for heterogeneity

The program calculates

(1) Relative Risks RR, (2) Risk Differences (absolute risks) AR and the Number needed to treat NNT, (3) Odds Ratios, and their Confidence Intervals

(use CapsLock in entire program !)

(the program can be interrupted by pressing Ctrl-C and be quit by entering 'System' after the 'Ok' appears)

Chi-Square Test of CARDIOVASC. in MRC Study

5 YEARS STUDY DURATION

STUDY GROUP	a: 286	b: 8414	E: 8700	3.287 %
CONTROL GROUP	c: 352	d: 8302	NE: 8654	4.067 %
	D: 638	ND: 16716	N: 17354	

==> X-SQUARE = 7.456 (1 DF) => P = .0063

CONFIDENCE INTERVALS:	95% CI	90% CI	Vertrauens intervalle
Risiko-Differenz → AR = 7.8 / 1000	(2.2 , 13.4)	(3.1 , 12.5)	
1.56	(.44 , 2.68)	(.62 , 2.5)	annually — pro Jahr
Number needed to treat → NNT = 128	(75 , 454)	(80 , 322)	
641	(373 , 2271)	(400 , 1612)	annually
Relative Risiko → RR = .808	(.693 , .942)	(.711 , .919)	
Risiko-Reduktion → 19.2 %	(30.7 , 5.8)	(28.9 , 8.09)	
Odds Ratio → OR = .802	(.684 , .94)	(.702 , .916)	

Print Screen ? (YES=<return> / NO='N') ?