Skript Evidenzbasierte Medizin I & II

Inhaltsverzeichnis

Dieses Skript wurde verfasst von Peter von Philipsborn, Jacob Burns, Caroline Jung-Sievers, Kerstin Sell, Jan Stratil und Eva Rehfuess. Es deckt die Inhalte der Lehrveranstaltungen EbM I & II ab. Für die Lehrveranstaltung EbM III gibt es ein eigenes Skript. Wir freuen uns über Rückmeldungen und Verbesserungsvorschläge, die an pphilipsborn@ibe.med.uni-muechen.de gesandt werden können.

Lernziele

EbM I: Grundlagen ärztlichen Handelns – was ist evidenzbasierte Medizin, und was bringt sie mir? AbsolventInnen der Veranstaltung „EbM I“ können …

- den Begriff der Evidenzbasierung und die ihm zugrundeliegenden Prinzipien von Systematik, Transparenz und Bewertung von Unsicherheit erläutern.
- Problemstellungen in präzise wissenschaftliche Fragestellungen übersetzen, die in Fach- und Literaturdatenbanken recherchierbar sind.
- wesentliche gesundheitsrelevante Datenbanken und Internetportale benennen.
- systematische Reviews, Health Technology Assessments und Leitlinien als wichtige Instrumente der Vermittlung von Evidenz in die Praxis beschreiben.

EbM II: Fehlerquellen finden und bewerten – welcher Studie kann ich trauen? AbsolventInnen der Veranstaltung „EbM II“ können …

- die Begriffe interne und externe Validität, Studienqualität, Confounding und Bias erläutern.
- einzelne Studien hinsichtlich Relevanz und Validität kritisch bewerten.

EbM Workshop „Wenn Bauchgefühl und Wikipedia nicht reichen – wie finde ich die Evidenz, die ich brauche?“ AbsolventInnen des EbM Workshops können …

- Suchstrategien für PubMed und die Cochrane Library entwickeln und praktisch durchführen.
- die für die präzise wissenschaftliche Fragestellung relevanten und verlässlichen Publikationen identifizieren.
- Vor- und Nachteile der Verwendung von UpToDate, der Cochrane Library und PubMed im Zusammenhang mit einer Entscheidung im medizinischen Alltag benennen.

1. Einführung

1.1. Definition der evidenzbasierten Medizin (EbM)

Der Begriff der evidenzbasierten Medizin (EbM) wurde in den 1990er Jahren unter anderem von dem Mediziner und Epidemiologen David Sackett an der McMaster Universität in Hamilton, Kanada, geprägt [1]. Nach einer weit verbreiteten Definition bezeichnet die EbM das Fällen von gesundheitsbezogenen Entscheidungen auf Grundlage der systematischen und bewussten Berücksichtigung von drei Aspekten: der jeweils besten verfügbaren wissenschaftlichen Erkenntnisse; der klinischen Erfahrung der Behandelnden; und den Werten und Präferenzen der Patientin oder des Patienten (siehe Abbildung 1) [2].

Abbildung 1: *Evidenzbasierte Medizin (EbM) als Integration der klinischen Erfahrung der Behandler, der bestverfügbaren wissenschaftlichen Erkenntnisse und der Werte und Präferenzen der PatientIn.*

Das Konzept der EbM hat sich seitdem rasch verbreitet und ist heute eines der zentralen Paradigmen der Medizin und Gesundheitsversorgung [3]. So schreibt zum Beispiel das deutsche Sozialgesetzbuch vor, dass alle Gesundheitsleistungen, die von der gesetzlichen Krankenversicherung übernommen werden, den Prinzipien der EbM entsprechen müssen [1]. Die Bundesrahmenempfehlungen der Nationalen Präventionskonferenz sehen vor, dass auch im Rahmen des Präventionsgesetzes umgesetzte Maßnahmen evidenzbasiert sein sollten [4]. Auch die Weltgesundheitsorganisation (WHO) und zahlreiche weitere internationale und nationale Gesundheitsorganisationen bekennen sich zu einem evidenzbasierten Vorgehen [5].

Im Verlauf wurde das Konzept der EbM auf weitere Bereiche übertragen; so wurden Begriffe wie *Evidenzbasierte Pflege* und *Evidenzbasierte Public Health* geprägt [5] [6]. Analog zur eingangs zitierten Definition der EbM lässt sich das Verfahren einer *evidenzbasierten Entscheidungsfindung* allgemein definieren (siehe Kasten 1).

Kasten 1: Definition evidenzbasierter Entscheidungsfindung

Bei einer ***evidenzbasierten Entscheidungsfindung*** werden Entscheidungen auf Grundlage einer systematischen und bewussten Integration der für die Frage relevanten besten verfügbaren

wissenschaftlichen Erkenntnisse, der praktischen Erfahrungen und der Expertise relevanter Fachleute sowie der Werte und Präferenzen der betroffenen Personen getroffen.

Diese Definition von Evidenzbasierung bezieht sich auf das Verfahren, mit dem Entscheidungen getroffen werden. Der Begriff Evidenzbasierung kann sich auch auf konkrete Maßnahmen beziehen. Zum Teil wird dann von einer evidenzbasierten Maßnahme gesprochen, wenn es verlässliche wissenschaftliche Belege gibt, dass die entsprechende Maßnahme dafür geeignet ist, ihre intendierten Wirkungen hervorzubringen [7]. Diese Definition stellt den wissenschaftlichen Wirksamkeitsnachweis in den Vordergrund. Alternativ kann dann von einer evidenzbasierten Maßnahme gesprochen werden, wenn die Entscheidung für die jeweilige Maßnahme auf Grundlage eines evidenzbasierten Verfahrens getroffen wurde – d.h. eines Verfahrens, das die in Kasten 1 genannten Kriterien erfüllt. Dem vorliegenden Skript liegt dieses zweite Verständnis zu Grunde.

1.2. Wozu Evidenz in der Medizin?

Weshalb sollten klinische Entscheidungen auf Grundlage dieser drei Aspekte getroffen werden? Kurz zusammengefasst lässt sich die Notwendigkeit der drei Aspekte wie folgt erklären:

- Wissenschaftliche Evidenz ist nötig für die zuverlässige Beurteilung der Wirksamkeit verschiedener Therapien, der Aussagekraft diagnostischer Marker, der Prognose von Krankheiten und zahlreicher weiterer Aspekte klinischen Handelns.
- Klinische Erfahrung ist nötig für das Beurteilen von Befunden, das Einordnen von Diagnosen, die Therapieplanung, die Kommunikation mit der PatientIn und die Interpretation wissenschaftlicher Evidenz in Bezug auf eine bestimmte PatientIn.
- Die Vorgabe, die Präferenzen und Werte der betreffenden PatientIn zu berücksichtigen, ergibt sich aus dem ersten Gebot der Medizinethik, wie es von Beauchamp und Childress formuliert wurde: dem Respekt für die Autonomie, d.h. dem Selbstbestimmungsrecht der PatientIn.

Unter Evidenz (engl. *evidence*) werden hier Erkenntnisse aus systematischen wissenschaftlichen Untersuchungen verstanden.

Lange Zeit beruhte die Medizin hauptsächlich auf dem angesammelten Erfahrungswissen von ÄrztInnen. Dass dieses zwar wichtig ist, aber systematische wissenschaftliche Untersuchungen nicht ersetzen kann, wird im Kasten 2 anhand des Beispiels des Aderlassens erläutert.

Kasten 2: Das Aderlassen und die Notwendigkeit systematischer wissenschaftlicher Untersuchungen in der Medizin

Die Praxis des Aderlassens zu therapeutischen Zwecken ist seit dem 5 Jhr. v. Chr. in Europa, der arabischen Welt und Indien belegt, und war bis ins späte 19. Jahrhundert weit verbreitet.
Aderlassen wurde von zahlreichen Koryphäen der Medizin empfohlen und praktiziert, darunter Hippokrates, Herophilus, Galen und William Osler. Ein berühmtes Opfer des Aderlassens war George Washington, der erste Präsident der USA: Er erkrankte 1799 an einer Laryngitis, und verstarb nach mehreren Aderlässen, durch die ihm insgesamt 3,75 l Blut entnommen wurden, am hypovolämen Schock.

Heute wissen wir, dass Aderlassen von wenigen Ausnahmen abgesehen schädlich ist. Wieso wurde es trotzdem für so lange Zeit praktiziert? Eine Reihe von Erklärungsansätzen werden hierfür angeführt. Als ein Faktor wird angesehen, dass auf Einzelbeobachtungen beruhendes Erfahrungswissen in der Medizin irreführend sein kann. Viele Erkrankungen sind von einem wellenförmigen Verlauf geprägt – wenn bei solchen Erkrankungen eine Maßnahme (wie z.B. das Aderlassen) immer dann vorgenommen wird, wenn es einer PatientIn besonders schlecht geht, kann dies den falschen Eindruck erwecken, dass die Intervention eine dauerhafte oder temporäre Verbesserung des Gesundheitszustandes zur Folge hat (siehe Abbildung 2).

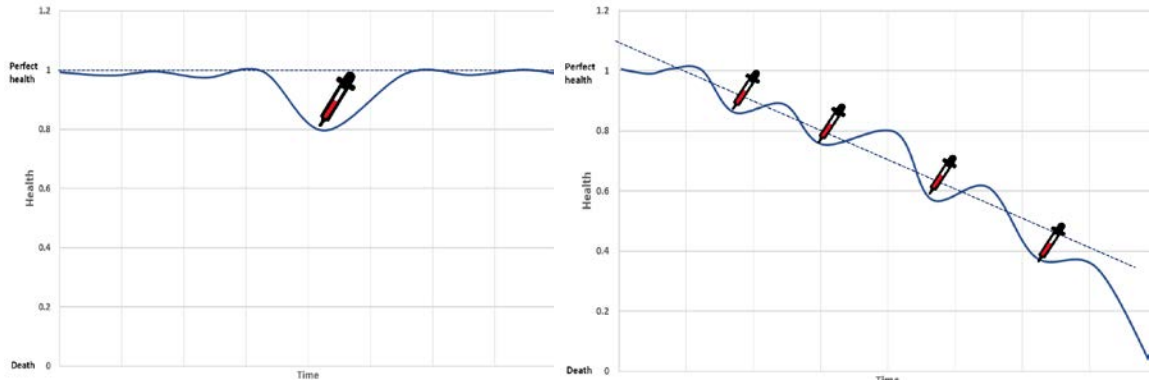


Abbildung 2*: Der wellenförmige Verlauf von Krankheiten kann die Illusion erzeugen, eine Maßnahme führe zur Verbesserung des Gesundheitszustandes (links: Verlauf einer akuten Erkrankung; rechts: Verlauf einer chronisch-progressiven Erkrankungen).*

Im 19. Jahrhundert verlor der Aderlass zunehmend an Popularität – mit hierzu trugen Untersuchungen des französischen Arztes Pierre Charles Louis bei, der die Anwendung der „numerischen Methode“ in der Medizin propagierte, einem Vorläufer der medizinischen Statistik und Epidemiologie. Unter anderen konnte er mit einer Untersuchung von 77 Patienten mit Lungenentzündung zeigen, dass PatientInnen, bei denen Aderlass praktiziert wurde, häufiger starben als solche, bei denen kein Aderlass praktiziert wurde [8].

2. Prinzipien der EbM

2.1. Systematik

Die Ergebnisse einzelner wissenschaftlicher Studien sind oft stark von Besonderheiten des lokalen Kontexts und methodischen Aspekten der jeweiligen Studie beeinflusst. Deshalb kann es irreführend sein, Entscheidungen auf einzelnen selektiv ausgewählten Studien zu basieren. Aus diesem Grund ist es ein Grundprinzip der EbM, wenn möglich stets den Gesamtkorpus an Evidenz (engl. *body of evidence*) zur jeweiligen Fragestellung zu betrachten [3]. Deshalb wird in der EbM großen Wert auf Systematik im Suchen, Auswählen und Auswerten wissenschaftlicher Evidenz gelegt. In vielen Fällen ist es empfehlenswert, nicht einzelne Studien, sondern – sofern vorhanden – *systematische Übersichtsarbeiten* (engl. *systematic reviews*) als Quelle zu nutzen [2]. In systematischen Übersichtsarbeiten wird der Evidenzkorpus zu definierten Fragestellungen systematisch und transparent erfasst und zusammengefasst (siehe Abschnitt 3.2).

In vielen Fällen ist es sinnvoll, auch weitere Schritte der Umsetzung von EbM möglichst systematisch zu gestalten. Dies gilt unter anderem für das Ableiten von Empfehlungen aus wissenschaftlicher Evidenz. In qualitativ hochwertigen Leitlinien – einem weiteren wichtigen Instrument der EbM – werden Empfehlungen zum Beispiel unter Nutzung formaler Verfahren der Konsensfindung entwickelt, um zu verhindern, dass einzelne besonders meinungsstarke Fachleute den Prozess dominieren. Sogenannte *Evidence-to Decision-Frameworks* dienen dazu, komplexe Entscheidungsfindungsprozesse zu strukturieren und die Berücksichtigung relevanter Kriterien zu systematisieren [9].

2.2. Partizipation

Es ist ein Grundprinzip der EbM, dass PatientInnen in die Entscheidungsfindung einbezogen und ihre Präferenzen und Werte berücksichtigt werden [3]. Hierzu gehört das Konzept der informierten Einwilligung (engl. *informed consent*) – PatientInnen sollten ausreichend gründlich und umfassend über Vor- und Nachteile bestimmter diagnostischer oder therapeutischer Maßnahmen aufgeklärt werden, um sich auf dieser Grundlage für oder gegen die Maßnahme entscheiden zu können. Noch weitergehend ist das Konzept des gemeinsamen Entscheidens (engl. *shared decision-making*): Behandler und PatientIn wägen Vor- und Nachteile und weitere entscheidungsrelevante Faktoren gemeinsam ab, bevor sie zu einer gemeinsamen Entscheidung kommen.

Besonders wichtig ist ein gemeinsames Entscheiden bei Fragen, die weitreichende und langfristige Folgen für das Leben der PatientIn haben– wie zum Beispiel der Entscheidung für oder gegen eine große, elektive Operation oder eine nebenwirkungsreiche Chemotherapie bei fortgeschrittener Krebserkrankung. Bei weniger weitreichenden Entscheidungen, sowie bei Entscheidungen, bei denen Patientenwünsche eine weniger wichtige Rolle spielen – wie z.B. der Entscheidung über die Dosierung eines häufig gegebenen, gut verträglichen Medikaments – kann es ausreichend und angemessen sein, die PatientIn über die Entscheidung zu informieren und ihm oder ihr die Möglichkeit des Einspruchs zu geben, oder eine Stellvertreterentscheidung zu treffen. Ähnliches gilt für Entscheidungen, die unter großem Zeitdruck getroffen werden müssen. Das Thema des gemeinsamen Entscheidens wird in der Lehrveranstaltung EbM III vertieft behandelt.

2.3. Integration

Ein weiterer Grundgedanke der EbM ist, dass Entscheidungen auf Grundlage einer bewussten und überlegten Integration verschiedener Aspekte getroffen werden sollten [3]. Die eingangs aufgeführte Definition von EbM betont die Integration von wissenschaftlicher Evidenz, der Erfahrung der Behandelnden und der Perspektiven der PatientInnen [2]. Darüber hinaus erstreckt sich das Grundprinzip der Integration noch auf weitere Aspekte. So sollte zum Beispiel nicht nur Evidenz zu den erwünschten Wirkungen von Maßnahmen berücksichtigt werden, sondern auch zu möglichen unerwünschten Wirkungen. Daneben können auch Faktoren wie Komorbiditäten, die Komedikation, die erwartete Adhärenz der PatientIn, und ihre Lebenssituation für die Entscheidungsfindung relevant sein.

Insbesondere bei übergreifenden Fragen der Gesundheitsversorgung und Gesundheitspolitik ist es zudem oft wichtig, auch Evidenz zu möglichen ethischen, sozialen, politischen, wirtschaftlichen und ökologischen Auswirkungen zu berücksichtigen [9]. Dies erfordert oft eine Zusammenarbeit über Fach-

und Berufsgrenzen hinweg. Ein Beispiel hierfür sind die Maßnahmen, die im Frühjahr 2020 zur Eindämmung der Corona-Epidemie getroffen wurden, wie zum Beispiel die Schließung von Schulen und Geschäften. Aber auch bei weniger außergewöhnlichen Fragen der Gesundheitsversorgung können z.B. wirtschaftliche und ethische Überlegungen eine wichtige Rolle spielen – so bei der Frage, ob ein neues, teures Medikament für eine seltene Erkrankung in den Leistungskatalog der gesetzlichen Krankenversicherung aufgenommen werden soll (siehe Abschnitt 3.5). Bei solchen Entscheidungen ist immer zu bedenken, dass Gelder, die für einen bestimmten Zweck verwendet werden, nicht für andere Zwecke verwendet werden können (Konzept der Opportunitätskosten) – was relevante ethische Implikationen haben kann. Wie der Aspekt von Integration in der klinischen Praxis gehandhabt werden kann wird in der Lehrveranstaltung EbM III im Detail behandelt.

2.4. Transparenz

Ein weiteres Grundprinzip der EbM ist die Transparenz. Dies betrifft den transparenten Umgang mit Unsicherheit, der Offenlegung der Grundlage von Entscheidungen (Entscheidungskriterien, entscheidungsrelevante Überlegungen, Evidenzgrundlage), sowie die nachvollziehbare Präsentation von Methoden und Datenquellen [3].

Es ist ein Grundgedanke der EbM dass es keine absolute, über jede Zweifel erhabene Gewissheit gibt. Der Grad der Gewissheit bzw. die Vertrauenswürdigkeit von Evidenz (engl. *certainty of evidence*) lässt sich mittels geeigneter Verfahren abschätzen, und sollte transparent dargestellt werden. Wichtige Unsicherheiten, insbesondere bezüglich der Wirksamkeit und der Sicherheit einer Therapie sollten der PatientIn klar kommuniziert werden.

Die nachvollziehbare Präsentation der verwendeten Methoden und Datenquellen dient einerseits der Selbstkontrolle und der Sicherung der Systematik und der methodischen Qualität, und erlaubt es andererseits unabhängigen Fachleuten und anderen Betroffenen, das jeweilige Verfahren und die Ergebnisse kritisch zu hinterfragen. Die Wissenschaft lebt ebenso wie die Medizin von der kontinuierlichen kritischen Prüfung, Verfeinerung und Weiterentwicklung überlieferter Erkenntnisse [3]. Transparenz ermöglicht diese Entwicklung. Dies gilt für wissenschaftliche Veröffentlichungen und klinische Leitlinien ebenso wie für wichtige klinische Einzelentscheidungen, wie z.B. Entscheidungen eines Tumorboards.

Bei wissenschaftlichen Veröffentlichungen (und auch für medizinische Doktorarbeiten), ist es wichtig, dass die verwendeten Methoden und Datenquellen systematisch, vollständig und ausreichend detailliert beschrieben werden. So genannte Reporting Guidelines enthalten Empfehlungen dazu, welche Aspekte in wissenschaftlichen Veröffentlichungen berichtet werden sollen. Viele Fachzeitschriften verlangen die Einhaltung dieser Reporting Guidelines. Es empfiehlt sich, schon in der Frühphase der Planung einer wissenschaftlichen Arbeit (z.B. der Doktorarbeit) die für das jeweilige Studiendesign relevante Reporting Guideline zu konsultieren. Eine Zusammenstellung der wichtigsten Reporting Guidelines findet sich auf der Webseite des Equator-Netzwerks: www.equator-network.org. Welche Reporting Guideline anwendbar ist, hängt vom Studiendesign ab – weitere Informationen hierzu finden sich in Abschnitt 3.1.

2.5. Reflektierter Umgang mit Interessenkonflikten

Fachleute vertreten nicht notwendiger Weise nur die Interessen der Betroffenen oder des Gemeinwohls, sondern auch ihre eigenen, die von diesen Interessen abweichen oder sogar mit diesen im Konflikt stehen können. In diesem Fall spricht man von *Interessenkonflikten*. So liegt zum Beispiel ein *finanzieller Interessenkonflikt* vor, wenn Fachleute, die an der Bewertung eines bestimmten Medikaments beteiligt sind, finanzielle Zuwendungen von dem Hersteller des entsprechenden Medikaments erhalten. Das Ziel, Interessenkonflikte unter den Beteiligten zu vermeiden, muss abgewogen werden gegenüber dem Ziel, eine möglichst breite Beteiligung relevanter Fachleute zu erreichen. Es ist ein Grundprinzip der EbM, dass Akteure, bei denen davon ausgegangen werden muss, dass sie nicht primär die Interessen der Betroffenen oder des Gemeinwohls vertreten sondern ausschließlich oder vor allem im eigenen Interesse handeln, nicht an Entscheidungen beteiligt werden sollten [10]. Dies gilt insbesondere für VertreterInnen von Unternehmen und Lobbygruppen der Industrie.

3. Instrumente der EbM

3.1. Einzelstudien und ihre Studiendesigns

Wissenschaftliche Erkenntnisse werden in der Regel in Form von Einzelstudien veröffentlicht und verbreitet. In Abhängigkeit von der verwendeten Methodik werden dabei verschiedene Studiendesigns unterschieden. Je nach Fragestellung sind jeweils verschiedene Studiendesigns besonders geeignet, zuverlässige Ergebnisse zu liefern (siehe Tabelle 1). Merkmale dieser Studiendesigns werden in der Lehrveranstaltung zu klinischen Studien in größerem Detail behandelt.

Tabelle 1: Fragestellungen und dafür geeignete Studiendesigns (Beispiele)

Fragestellung	**Geeignete Studiendesigns**
Wirksamkeit von Therapiemaßnahmen (z.B. Medikamente, OP-Verfahren)	• Randomisierte kontrollierte Studien (RCTs) einschließlich Cluster-randomisierter kontrollierter Studien (cRCTs) • falls keine randomisierten Studien möglich oder verfügbar sind: sonstige experimentelle und quasi-experimentelle Studiendesigns
Nebenwirkungen von Therapiemaßnahmen	• Häufige Nebenwirkungen: RCTs oder Kohortenstudien • Gelegentlich auftretende Nebenwirkungen: Kohortenstudien • Seltene Nebenwirkungen: Fall-Kontroll-Studien, Registerstudien
Häufigkeit von Krankheiten und Risikofaktoren zu einem gegebenen Zeitpunkt	• Querschnittstudien
Häufigkeit von Krankheiten und Risikofaktoren im Zeitverlauf	• Längsschnitt- bzw. Kohortenstudien

Zusammenhang zwischen Risiko- und Schutzfaktoren und möglichen Folgezuständen (z.B. Krankheiten)	• Kohortenstudien • Fall-Kontroll-Studien
Werte und Präferenzen, subjektives Erleben von Gesundheit und Krankheit, Akzeptabilität von Maßnahmen	• Qualitative Studien (z.B. Interview-Studien, Fokus-Gruppen-Diskussionen)
Praktische Umsetzbarkeit und Akzeptabilität von Maßnahmen	• Prozessevaluationen (i.d.R. mit qualitativen und quantitativen Elementen)
Kosten von Maßnahmen	• Ökonomische Evaluationen

Von besonderer Bedeutung ist häufig die Frage, welche Wirkungen von bestimmten Maßnahmen, wie z.B. einer Präventionsstrategie oder medikamentösen Therapie, ausgehen. Die zuverlässigste Evidenz zu Fragestellungen dieser Art liefern *randomisierte kontrollierte Studien* (engl. *randomized controlled trials*, RCTs). Bei einer RCT werden die TeilnehmerInnen zu Beginn der Studie nach dem Zufallsprinzip, also randomisiert, einer Interventions- und einer Kontrollgruppe zugeteilt. Dies dient dazu, dass die beiden Gruppen zu Beginn möglichst ähnlich sind. Durch die Randomisierung soll sichergestellt werden, dass sowohl Einflussfaktoren, die bekannt sind (z.B. der Anteil der Personen mit Vorerkrankungen), als auch aktuell noch nicht bekannte Einflussfaktoren (z.B. eine noch unbekannte genetische Prädisposition für die Erkrankung), gleichmäßigt verteilt sind. Wenn die Intervention und Kontrollgruppen zu klein sind, kann – trotz Randomisierung – eine gleichmäßige Verteilung der Risikofaktoren nicht sichergestellt werden. Die Interventionsgruppe erhält anschließend die zu untersuchende Maßnahme (z.B. eine bestimmte Therapie), die Kontrollgruppe hingegen nicht. Abhängig davon, was untersucht werden soll, kann die Kontrollgruppe keine Therapie, die aktuell empfohlene Therapie oder eine Placebo-Therapie erhalten. Sofern die Randomisierung erfolgreich war und die beiden Gruppen ausreichend groß und zu Beginn der Studie tatsächlich vergleichbar waren, so lassen sich im Verlauf auftretende Unterschiede zwischen den beiden Gruppen kausal auf die untersuchte Maßnahme zurückführen. In Kasten 3 wird das Prinzip einer RCT anhand eines Beispiels erläutert.

Das wesentliche Merkmal einer RCT ist die Randomisierung, d.h. die Einteilung der TeilnehmerInnen in eine Interventions- und eine Kontrollgruppe nach dem Zufallsprinzip. Viele RCTs sind zudem *verblindet*. Dies bedeutet, dass die PatientInnen, die Behandelnden und/oder die ForscherInnen nicht wissen, wer welche Therapie erhalten hat. Um eine erfolgreiche Verblindung sicherzustellen, kann es ggf. nötig sein, der Kontrollgruppe ein Placebo zu verabreichen – wie z.B. Tabletten, die keinen Wirkstoff enthalten. Die Verblindung soll vermeiden, dass unbewusste psychologische Effekte das Ergebnis beeinflussen, z.B. auf der Seite der PatientInnen (Placeboeffekt) oder auf der Seite der TharapeutInnen (z.B. der sog. Rosenthal-Effekt).

Kasten 3: Randomisierte kontrollierte Studien – das Beispiel des WOMAN-Trials

Ein Beispiel für eine große, internationale, multizentrische randomisiert-kontrollierte Studie ist das WOMAN-Trial, bei der die Wirksamkeit von Tranexamsäure im Vergleich zu einem Placebo bei postpartalen Blutungen (Nachgeburtsblutung) mit Blick auf die Endpunkte von u.a. Mortalität und Hysterektomie untersucht wurde [11].

In der Studie wurden in 193 Krankenhäusern in 21 Ländern systematisch Frauen ab dem Alter von 16 Jahren mit der klinischen Diagnose einer postpartalen Blutung nach einer Vaginalgeburt oder einem Kaiserschnitt eingeschlossen. Die Frauen wurden nach dem Zufallsprinzip in eine Interventionsgruppe und eine Placebogruppe eingeteilt, die jeweils zusätzlich zur üblichen Behandlung entweder Tranexamsäure oder ein gleich aussehendes Placebo-Präparat erhielten. Sowohl Teilnehmerinnen, das behandelnde klinische Personal als auch diejenigen, die das Behandlungsergebnis beurteilten, waren gegenüber der Zuteilung in Interventions- und Kontrollgruppe verblindet. Das genaue Vorgehen der Studie, inklusive der Endpunkte, welche untersucht werden sollten, wurde vorausschauend (a priori) in einem Protokoll festgeschrieben, welches vor Beginn der Studie veröffentlicht wurde. Insgesamt wurden in der Studie 20.060 Mütter randomisiert [11]. Es zeigte sich, dass die intravenöse Gabe von Tranexamsäure in den ersten drei Stunden nach der Geburt die blutungsbedingte Mortalität um 30 Prozent gegenüber Placebogabe reduzierte (relatives Risiko 0,69, 95% Konfidenzintervall 0,52 – 0,91, p=0,008) [11].

Bei einer klassischen RCT werden individuelle Teilnehmer randomisiert. Allerdings kann man auch Personengruppen, Einrichtungen wie Schulen oder Krankenhäuser, oder geographische Einheiten wie Stadtteile oder Distrikte randomisiert einer Interventions- und Kontrollgruppe zuteilen. In diesen Fall spricht man von Cluster-randomisierten kontrollierten Studien (cRCTs). Eine Beispiel für eine solche Studie wird in Exkurs-Kasten 1 vorgestellt.

Exkurs-Kasten 1: Randomisierte kontrollierte Studien über Medikamentenversuche hinaus: – das Beispiel einer Studie zu städtischen Grünflächen und mentaler Gesundheit

Randomisiert-kontrollierte Studie werden oft im Kontext von klinischen Medikamentenstudien diskutiert, jedoch haben Sie auch in anderen Bereichen einen großen Nutzen. Dies sollte das Beispiel einer Studie zu städtischen Grünflächen und mentaler Gesundheit zeigen:

Studien zeigen, dass Menschen, die in der Nähe städtischer Grünflächen leben, seltener unter Depressionen leiden als Menschen, die fernab von Parks und Naherholungsgebieten leben. Aber liegt dies an den Grünflächen, oder vielleicht daran, dass Menschen, die in der Nähe von Grünflächen leben, im Durchschnitt wohlhabender sind, seltener arbeitslos sind, und psychisch weniger belastende Berufe ausüben? WissenschaftlerInnen der University of Pennsylvania in Philadelphia, USA, wollten dies herausfinden [14]. Sie identifizierten im Stadtgebiet von Philadelphia 541 verlassene, verwahrloste Grundstücke, und teilten diese nach dem Zufallsprinzip in drei Gruppen ein: ein Drittel der Grundstücke wurde von Müll und Unrat befreit, ein weiteres Drittel wurde zusätzlich begrünt und mit Bäumen bepflanzt, und das letzte Drittel wurde unverändert belassen. Personen, die in der Nähe dieser Grundstücke wohnten, wurden zu Beginn der Studie und 18 Monate nach der Säuberung und Begrünung der Grundstücke zu ihrer mentalen Gesundheit befragt. Unter den StudienteilnehmerInnen, die in der Nähe der begrünten Grundstücke wohnten, sank der Anteil der Personen mit depressiven Symptomen von 15% auf 10%, während er in den beiden anderen Gruppen unverändert blieb. Die AutorInnen schlussfolgern, dass städtische Grünflächen tatsächlich dazu beitragen, Depression vorzubeugen – das bloße Säubern verwahrloster Grundstücke hingegen scheint keinen Effekt zu haben [14].

Zu vielen gesundheitsrelevanten Fragen liegen keine RCTs vor. Dies kann daran liegen, dass die Durchführung einer RCT für die jeweilige Frage praktisch schwierig oder gar unmöglich ist. Wollte man zum Beispiel die Auswirkungen eines nationalen oder regionalen Rauchverbots in Gaststätten auf den durchschnittlichen Tabakkonsum im Rahmen einer randomisierten Studie untersuchen, so müsste man eine größere Zahl an Ländern oder Regionen nach dem Zufallsprinzip in eine „Rauchverbots-Gruppe" und eine Kontrollgruppe einteilen – was politisch und praktisch kaum möglich ist. In anderen Fällen ist die Durchführung von RCTs zwar grundsätzlich denkbar, dennoch wurden noch keine solchen durchgeführt. In beiden Fällen muss auf alternative Studiendesigns zurückgegriffen werden. Für die Evaluation von Maßnahmen besonders relevant sind nicht-randomisierte Interventionsstudien (manchmal auch als *quasi-experimentelle Studien* oder als *natürliche Experimente-Studien* bezeichnet). Das einfachste nicht-randomisierte Studiendesign für die Evaluation einer Maßnahme ist ein einfacher Vorher-Nachher-Vergleich. So ließe sich z.B. der durchschnittliche Zigarettenkonsum vor und nach der Einführung eines Rauchverbots erheben. Ein Problem eines solchen Studiendesigns ist, dass die Ergebnisse durch langfristige Trends (wie z.B. einem allgemeinen Rückgang der Raucherzahlen aufgrund besserer Aufklärung oder stärkerer Besteuerung von Tabakprodukten) sowie Störfaktoren (wie z.B. einer zeitgleich durchgeführten Werbekampagne von Tabakunternehmen) verfälscht werden können. Es wurden daher diverse weitere Studiendesigns entwickelt, die ohne eine Randomisierung auskommen, aber dennoch versuchen, das Risiko solcher Verfälschungen zu reduzieren. Hierzu zählen unter anderem Unterbrochene-Zeitreihen-Studien (engl. *interrupted time series studies*), sowie kontrollierte Vorher-Nachher-Studien (engl. *controlled before-after studies* bzw. *difference-in-difference studies*). Die Entwicklung entsprechender Methoden hat in den letzten Jahren große Fortschritte gemacht, und schreitet weiter voran [12] [13].

Zu berücksichtigen ist zudem, dass RCTs nur dann indiziert sind, wenn tatsächlich Unklarheit darüber herrscht, ob eine Maßnahme Vorteile bietet, oder welche von mehreren Maßnahmen die beste ist (sog. klinische Equipoise). Denn wenn die Vorteile einer Maßnahme bereits belegt sind, dann kann es unethisch sein, einem Teil der TeilnehmerInnen diese vorzuenthalten, wie es bei einer RCT der Fall ist.

Während RCTs und andere Interventionsstudien helfen, die Auswirkungen von Maßnahmen abzuschätzen, können epidemiologische Beobachtungsstudien – insbesondere Quer- und Längsschnittstudien zur Prävalenz und Inzidenz – genutzt werden, um die Relevanz verschiedener Krankheiten und Risikofaktoren abzuschätzen. Qualitative Studien, z.B. Interviewstudien, können u.a. dafür genutzt werden, zu untersuchen, wie Gesundheitsfachkräfte, PatientInnen und andere Betroffene (z.B. Angehörige) bestimmte Maßnahmen wahrnehmen und erleben. Dies kann z.B. helfen die Gründe zu verstehen, warum Personen ein bestimmtes Risikoverhalten eingehen oder eine empfohlene Therapie nicht wahrnehmen. Die praktische Umsetzbarkeit und Akzeptabilität von Maßnahmen kann mit Prozessevaluationen untersucht werden, und ökonomische Evaluationen dienen u.a. dazu, die Kosten von Maßnahmen abzuschätzen. Allgemein gilt, dass das optimale Studiendesign von der Fragestellung abhängt.

3.2. Systematische Übersichtsarbeiten und die Evidenzpyramide

Wie eingangs dargestellt ist es ein Grundprinzip der EbM, Entscheidungen nicht auf Grundlage selektiv ausgewählter Studien zu fällen, sondern wenn möglich stets den Gesamtkorpus an Evidenz zur jeweiligen Fragestellung zu betrachten. Diesen zu erfassen und systematisch und transparent zusammenzufassen ist das Ziel *systematischer Übersichtsarbeiten* (engl. *systematic reviews*).

Die Durchführung qualitativ hochwertiger systematischer Übersichtsarbeiten umfasst die folgenden Schritte, die transparent und nachvollziehbar dokumentiert sein sollten:

- Formulierung einer klaren Fragestellung (z.B. „Wie wirkt sich die Behandlung von Personen mit Diabetes mellitus Typ 2 mit Metformin auf die kardiovaskuläre Mortalität aus?“).
- Festlegung von Kriterien für Studien, die zur Beantwortung der Frage geeignet sind, und die in die Übersichtsarbeit eingeschlossen werden sollen (z.B. Beschränkung auf RCTs, in denen die TeilnehmerInnen über mindestens zwei Jahre nachbeobachtet wurden).
- Durchführung einer umfassenden Literatursuche mit dem Ziel, alle Studien, welche die festgelegten Kriterien erfüllen, zu identifizieren.
- Standardisierte Betrachtung aller in die systematische Übersichtsarbeit eingeschlossenen Studien und systematische Zusammenführung der Ergebnisse.
- Kritische Prüfung und Bewertung der Zuverlässigkeit der Ergebnisse der einzelnen Studien sowie der Vertrauenswürdigkeit des Gesamtkorpus an Evidenz.

In manchen Fällen ist es möglich, die Ergebnisse der einzelnen Studien mit statistischen Methoden quantitativ zusammenzufassen (engl. *poolen*). In diesem Fall spricht man von einer *Meta-Analyse*. Eine Meta-Analyse ist jedoch nur dann sinnvoll, wenn die eingeschlossenen Studien hinreichend ähnlich sind.

Mittlerweile gibt es unzählige systematische Übersichtsarbeiten zu einer Vielzahl von Themen. Als systematische Übersichtsarbeiten der höchsten methodischen Qualität gelten sogenannte Cochrane Reviews. Cochrane ist ein internationales Netzwerk von Gesundheitsfachkräften und WissenschaftlerInnen, die sich zum Ziel gesetzt haben, zuverlässige Evidenz zu gesundheitsrelevanten Themen breit zugänglich und nutzbar zu machen (siehe weiterführende Ressourcen).

Aufgrund der besonderen Bedeutung von systematischen Übersichtsarbeiten für die EbM stehen diese in der sogenannten *Evidenzpyramide* (siehe Abbildung 3) über Einzelstudien. Die Evidenzpyramide gibt an, wie zuverlässig verschiedene Studiendesigns in der Beurteilung von Kausalzusammenhängen zwischen Maßnahmen und ihren Auswirkungen sind. Studiendesigns an der Spitze der Pyramide sind zuverlässiger als solche am Boden der Pyramide [15]. Zu beachten ist, dass sich die Evidenzpyramide nur auf die Evaluation der Wirksamkeit von Maßnahmen bezieht – für andere Fragestellungen sind oftmals andere Studiendesigns besser geeignet als RCTs bzw. systematische Übersichtsarbeiten von RCTs (siehe Tabelle 1). Der Grundsatz, dass Entscheidungen wenn möglich, nicht auf Einzelstudien, sondern auf methodisch gut durchgeführten systematischen Übersichtsarbeiten basiert werden sollten, gilt jedoch auch für andere Fragestellungen und methodische Ansätze.

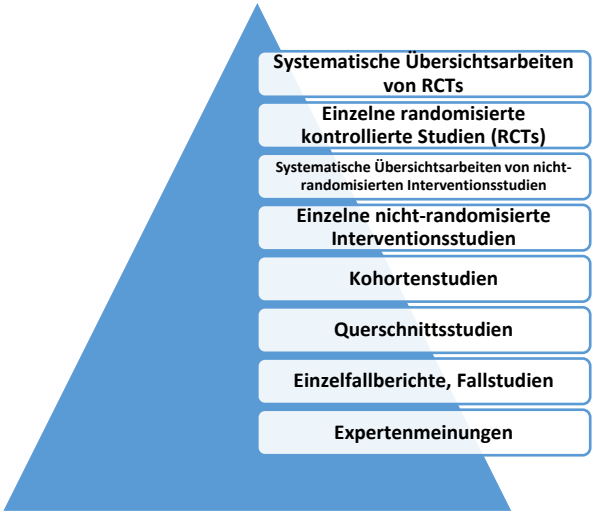


Abbildung 3: *Die Evidenzpyramide. Je weiter oben ein Studiendesign in der Pyramide steht, desto besser ist es dafür geeignet, zuverlässige Evidenz zum kausalen Zusammenhang zwischen Maßnahmen und ihren Auswirkungen (Wirksamkeit) zu liefern. Bildquelle: Eigene Darstellung.*

In Kasten 4 wird anhand der *Roll Back Malaria* Initiative beispielhaft dargestellt, wie Erkenntnisse aus systematischen Übersichtsarbeiten und Einzelstudien mit unterschiedlichen Studiendesigns zur Entwicklung einer erfolgreichen Strategie für die Bekämpfung der Malaria in Afrika beitrugen [5].

Kasten 4: EbM in der Praxis – das Beispiel der *Roll Back Malaria* Initiative

Im Jahr 1998 wurde von der damaligen Generaldirektorin der WHO, Gro Harlem Brundtland, die *Roll Back Malaria* Initiative ins Leben gerufen. Erklärtes Ziel war es, bis 2010 die Zahl der durch Malaria bedingten Todesfälle in Afrika zu halbieren [16]. Anfang der 2000er Jahre zeigten Zusammenfassungen von epidemiologischen Beobachtungsstudien, dass die Zahl der Malaria-bedingten Todesfälle in Afrika nicht zurückging, sondern vielmehr gegenüber 1998 angestiegen war. Diese ernüchternden Zahlen halfen, zusätzliche Geldmittel für den Kampf gegen die Malaria zu mobilisieren. Investiert wurde insbesondere in zwei Maßnahmen, deren Wirksamkeit durch systematische Übersichtsarbeiten mehrerer RCTs belegt war: das Bereitstellen von mit Pestiziden imprägnierten Moskitonetzen, und die Behandlung von Erkrankten mit einer Artesemin-Kombinations-Therapie. Der breite Einsatz dieser beiden Maßnahmen wird als wichtiger Erfolgsfaktor dafür angesehen, dass die WHO 2010 einen Rückgang der Malaria-bedingten Todesfälle in Afrika um 50% gegenüber dem Jahr 2000 melden konnte [5]. In Beobachtungsstudien konnte eine klare zeitliche und räumliche Korrelation zwischen dem Bereitstellen zusätzlicher Geldmittel, dem vermehrten Einsatz der beiden genannten Maßnahmen und dem Rückgang der Malaria-Sterblichkeit nachgewiesen werden [5].

Qualitativ hochwertige systematische Übersichtsarbeiten geben in der Regel an, wie vertrauenswürdig die Evidenz ist, die ihren Schlussfolgerungen oder Empfehlungen zu Grunde liegt. Ein weit verbreitetes Instrument für die Bewertung der Evidenzstärke, das unter anderem von der WHO und Cochrane verwendet wird, ist GRADE (Grading of Recommendations Assessment, Development and Evaluation) [18]. GRADE unterscheidet vier Stufen der Stärke bzw. Vertrauenswürdigkeit von Evidenz: Hohe, mittlere, niedrige und sehr niedrige Vertrauenswürdigkeit. Die Vertrauenswürdigkeit von Evidenz (engl. certainty of evidence) bezeichnet das Ausmaß der Gewissheit, dass ein bestimmter Effekt dem

realen Effekt entspricht. Die Evidenzgradierung mittels GRADE erfolgt anhand verschiedener Kriterien, welche die Vertrauenswürdigkeit verringern (z.B. inkonsistente Ergebnisse über Studien hinweg, ungenaue Ergebnisse mit breitem Konfidenzintervall, oder durch methodische Mängel verzerrte Ergebnisse) oder erhöhen (z.B. eine Dosis-Wirkung-Beziehung, sowie große Effektstärken) [18]. Die Bewertung der Evidenzstärke mit GRADE bezieht sich stets auf den Gesamtkorpus an Evidenz zu einer definierten Fragestellung (der sich aus allen Studien zusammensetzt, die zu der jeweiligen Fragestellung gefunden wurden), anders als die Bewertung des Verzerrungsrisikos und das Critical Appraisal (siehe Abschnitt 4.3), die sich auf einzelne Studien beziehen.

3.3. Leitlinien

Leitlinien (engl. *guidelines*) sind ein weiteres wichtiges Instrument der EbM und eine entscheidende Informationsquelle für ÄrztInnen. Leitlinien enthalten praxisorientierte Handlungsempfehlungen, und sollen mittels einer Zusammenfassung des aktuellen Erkenntnisstandes zu einem Gesundheitsproblem die Entscheidungsfindung von Gesundheitsfachkräften und PatientInnen unterstützen [17].

In Kasten 5 wird dargestellt, wie eine WHO-Leitlinie zur Behandlung postpartaler Blutungen basierend auf systematischen Übersichtsarbeiten und einer großen RCT entwickelt und überarbeitet wurde.

Kasten 5: Die WHO-Leitlinie zu intravenöser Tranexamsäure bei postpartalen Blutungen

Maternale Mortalität (Müttersterblichkeit) ist ein fortbestehendes Problem der globalen Gesundheit. Im Jahr 2017 sind weltweit 295.000 (80% KI: 279.000 bis 340.000) Frauen im Zusammenhang mit einer Geburt gestorben, entsprechend einer Sterblichkeit von 211 Müttern pro 100.000 Lebendgeburten (80% KI: 99 bis 243) [18]. Die Vereinten Nationen haben das Ziel ausgegeben, die Müttersterblichkeit bis 2030 auf unter 70 pro 100.000 Lebendgeburten zu senken (nachhaltiges Entwicklungsziel (engl. *sustainable development goal*) 3.1) [18]. Die Hauptlast maternaler Mortalität entfällt dabei aktuell auf Länder mit niedrigem und mittlerem Einkommen, während die Müttersterblichkeit in Europa etwa 10 pro 100.000 Lebendgeburten beträgt [18]. Die häufigste Ursache maternaler Mortalität in Ländern mit niedrigem Einkommen sind postpartale Blutungen, die für etwa ein Viertel der weltweiten Müttersterblichkeit verantwortlich sind [19].

Zur Behandlung postpartaler Blutungen veröffentlichte die Weltgesundheitsorganisation 2012 eine Leitlinie, basierend auf der besten vorliegenden Evidenz und der Bewertungen eines ExpertInnenpanels. Zu dem Zeitpunkt lagen zwei systematische Übersichtsarbeiten zur Wirksamkeit von Tranexamsäure, einem weit verbreiteten Anti-Fibrinolytikum, zur Behandlung postpartaler Blutungen vor. Diese Übersichtsarbeiten konnten jedoch nur kleine Studien von niedriger Qualität identifizieren, sodass nur eine schwache und bedingte Empfehlung für die Gabe von Tranexamsäure ausgesprochen werden konnte [19].

Im Verlauf mehrten sich jedoch Hinweise, dass eine rasche Tranexamsäuregabe in der Traumatologie effektiv zur Vermeidung von Blutverlusten beiträgt. Deshalb wurde die oben beschriebene WOMAN-Trial (siehe Kasten 3) durchgeführt, bei der die Wirkung des Medikaments gegenüber Placebo in einer großen RCT mit 20.060 Müttern in193 Krankenhäusern und 21 Ländern untersucht wurde. Hierbei zeigte sich, dass die intravenöse Gabe von Tranexamsäure in den ersten

drei Stunden nach der Geburt die blutungsbedingte Mortalität um 30 Prozent gegenüber Placebogabe reduzierte (relatives Risiko 0,69, 95% KI 0,52 – 0,91, p=0,008) [11].

Diese Studienergebnisse waren so aussagekräftig, dass ein rascher Revisionsprozess der WHO-Leitlinie initiiert wurde, und die WHO die Ergebnisse des WOMAN-Trial als Ergänzung zu der bestehenden Leitlinie publizierte [19]. Es bleibt nun zu beobachten, ob sich die intravenöse Tranexamsäuregabe in der Routineversorgung bei postpartalen Blutungen etabliert und ob durch eine orale Gabe von Tranexamsäure ähnliche Ergebnisse erzielt werden können, was den Einsatz in vielen Ländern mit niedrigem und mittleren Einkommen erleichtern würde. Mittel- und langfristig kann die Tranexamsäuregabe damit zur Reduktion maternaler Sterblichkeit beitragen. Dabei wird dieser Ansatz jedoch ohne eine deutliche Verbesserung des Zugangs zu Gesundheitsversorgung für Frauen nur unzureichend wirksam sein. Das Erreichen des nachhaltigen Entwicklungsziels für maternale Mortalität bis 2030 erscheint derzeit unwahrscheinlich – auf Basis der aktuellen Entwicklung wird geschätzt, dass auf Grund des nur langsamen Fortschritts bis 2030 rund eine Million Mütter zusätzlich im Zusammenhang mit einer Geburt sterben werden [18].

Für eine angemessene Nutzung von Leitlinien ist es wichtig, ihre Entstehung und Aussagekraft nachvollziehen zu können. In Deutschland wird die Entwicklung gesundheitsbezogener Leitlinien von der Arbeitsgemeinschaft wissenschaftlich-medizinischer Fachgesellschaften (AWMF) koordiniert. Die AWMF hat zwei Qualitätskriterien für Leitlinien definiert [20]:

- Evidenzbasierung: Die Leitlinie sollte auf einer systematischen Suche nach relevanter wissenschaftlicher Literatur beruhen.
- Konsensbasierung: Die Leitlinie sollte von einer möglichst repräsentativen Leitliniengruppe unter Nutzung formaler Verfahren der Konsensfindung entwickelt werden. Dies soll sicherstellen, dass Fachleute aller relevanten Fachbereiche einbezogen werden, und verhindern, dass einzelne besonders meinungsstarke Mitglieder der Leitliniengruppe den Prozess dominieren.

Auf Grundlage dieser beiden Kriterien werden vier Leitlinien-Entwicklungsstufen unterschieden: S1-, S2e-, S2k- und S3-Leitlinien (siehe Tabelle 2).

Tabelle 2: Die Leitlinien-Entwicklungsstufen der AWMF

S1	Einfachste Form der Leitlinie, weder evidenz- noch konsensbasiert: in einem informellen Verfahren entwickelte Handlungsempfehlungen von ExpertInnen
S2k	Konsens-, aber nicht evidenzbasierte Leitlinie: Repräsentatives Gremium, strukturierte Konsensfindung
S2e	Evidenz-, aber nicht konsensbasierte Leitlinie: Systematische Recherche, Auswahl und Bewertung der Literatur
S3	Evidenz- und konsensbasierte Leitlinie: Repräsentatives Gremium, systematische Recherche, Auswahl und Bewertung der Literatur, strukturierte Konsensfindung

In S3-Leitlinien wird in der Regel für einzelne Empfehlungen die sog. Konsensstärke, der Empfehlungsgrad sowie das Evidenzniveau (engl. *level of evidence*, abgekürzt *LoE*) angegeben (siehe Abbildung 4). Für die Angabe des Empfehlungsgrades, des Evidenzniveaus und der Konsensstärke gibt es jeweils festgelegte Abstufungen mit entsprechenden Kriterien (siehe Tabelle 3-5).

Empfehlungen	Empfehlungsgrad
5.21 Übergewichtige und adipöse Menschen sollen ermutigt werden, sich mehr körperlich zu bewegen. Körperliche Aktivität soll neben der Ernährungs- und Verhaltenstherapie ein Bestandteil der Maßnahmen zur Gewichtsreduktion sein. LoE 1++ bis 1+; starker Konsens Literatur: [2, 156, 209, 210, 241]	A
5.22 Es sollte sichergestellt werden, dass übergewichtige und adipöse Menschen keine Kontraindikationen für zusätzliche körperliche Aktivität aufweisen. LoE 4 (Expertenkonsens); starker Konsens	B
5.23 Für eine effektive Gewichtsabnahme sollte man sich > 150 Min./Woche mit einem Energieverbrauch von 1 200 bis 1 800 kcal/Woche bewegen. Krafttraining allein ist für die Gewichtsreduktion wenig effektiv. LoE 2++ bis 4; starker Konsens Literatur: [173]	B

Abbildung 4: *Drei Empfehlungen aus der S3-Leitlinie „Therapie und Prävention der Adipositas“ mit Empfehlungsgrad, Evidenzniveau (level of evidence, LoE) und Konsensstärke.*

Tabelle 3: Klassifikation der Konsensstärke in Leitlinien der AWMF	
Starker Konsens	Zustimmung von > 95 % der Mitglieder der Leitliniengruppe
Konsens	Zustimmung von > 75–95 % der Mitglieder der Leitliniengruppe
Mehrheitliche Zustimmung	Zustimmung von > 50–75 % der Mitglieder der Leitliniengruppe
Kein Konsens	Zustimmung von < 50 % der Mitglieder der Leitliniengruppe

Tabelle 4: Angabe des Evidenzniveaus in Leitlinien der AWMF	
Evidenzniveau	**Beschreibung**
1++	Qualitativ hochwertige Metaanalysen, systematische Übersichten von RCTs, oder RCTs mit sehr geringem Risiko systematischer Fehler (Bias)
1+	Gut durchgeführte Metaanalysen, systematische Übersichten von RCTs, oder RCTs mit geringem Risiko systematischer Fehler (Bias)
1-	Metaanalysen, systematische Übersichten von RCTs, oder RCTs mit hohem Risiko systematischer Fehler (Bias)
2++	Qualitativ hochwertige systematische Übersichten von Fall-Kontroll- oder Kohortenstudien oder qualitativ hochwertige Fall-Kontroll- oder Kohortenstudien mit sehr niedrigem Risiko systematischer Verzerrungen (Confounding, Bias, „Chance“) und hoher Wahrscheinlichkeit, dass die Beziehung ursächlich ist
2+	Gut durchgeführte Fall-Kontroll-Studien oder Kohortenstudien mit niedrigem Risiko systematischer Verzerrungen (Confounding, Bias, „Chance“) und moderater Wahrscheinlichkeit, dass die Beziehung ursächlich ist

2-	Fall-Kontroll-Studien oder Kohortenstudien mit einem hohen Risiko systematischerVerzerrungen (Confounding, Bias, „Chance“) und signifikantem Risiko, dass die Beziehung nicht ursächlich ist
3	Nicht-analytische Studien, z. B. Fallberichte, Fallserien
4	Expertenmeinung

Tabelle 5: Angabe des Empfehlungsgrades in Leitlinien der AWMF		
Empfehlungsgrad	**Beschreibung**	**Syntax**
A	Starke Empfehlung	soll/soll nicht
B	Empfehlung	sollte/sollte nicht
0	Empfehlung offen	kann erwogen werden/kann verzichtet werden

Für den Konsensgrad ist das Ausmaß der Einigkeit innerhalb der Leitliniengruppe ausschlaggebend. Das Evidenzniveau hängt von Art und Qualität der verfügbaren Studien ab. Bei der Festlegung des Empfehlungsgrades sollte die Leitliniengruppe das Evidenzniveau der zugrundeliegenden wissenschaftlichen Evidenz berücksichtigen. Darüber hinaus können noch eine Reihe weiterer Kriterien für die Festlegung des Empfehlungsgrades relevant sein, darunter [20]:

- Die Konsistenz der Studienergebnisse
- Die klinische Relevanz der Endpunkte und Effektstärken
- Das Nutzen-Risiko-Verhältnis
- Ethische, rechtliche und ökonomische Erwägungen
- Patientenpräferenzen
- Die Anwendbarkeit auf die Patientenzielgruppe und das deutsche Gesundheitssystem
- Die Umsetzbarkeit im Alltag und in verschiedenen Versorgungsbereichen

Viele Leitlinien liegen in einer Lang- und in einer Kurzfassung vor. Die Leitlinien-Kurzfassungen bieten praxisrelevante Informationen in knapper, übersichtlicher Form, oft in der Form von Behandlungsalgorithmen (siehe Abbildung 5).

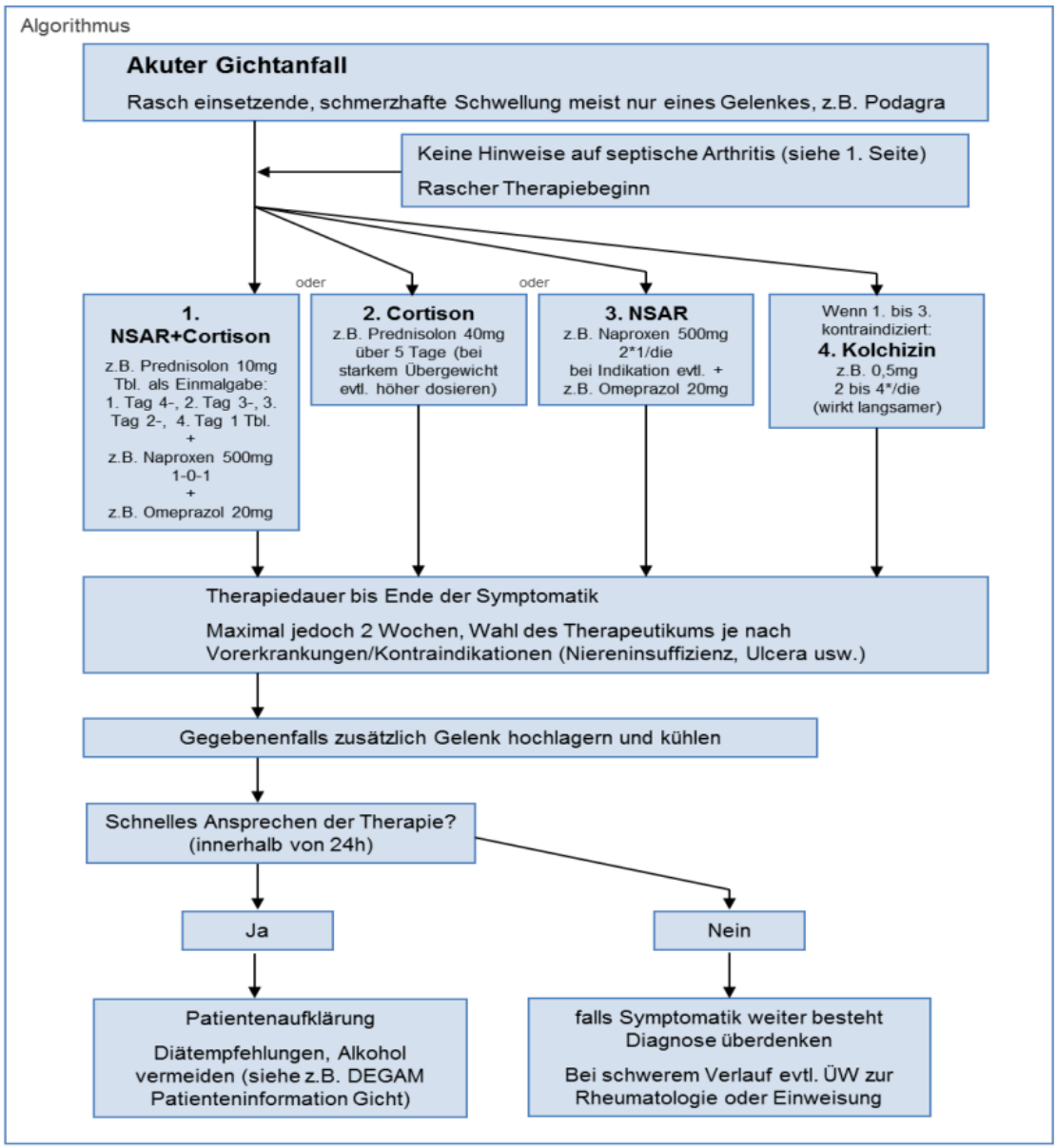


***Abbildung 5**: Behandlungsalgorithmus aus der Kurzfassung der S1-Leitlinie „Akute Gicht in der hausärztlichen Versorgung“.*

3.4. Klinische Referenzwerke

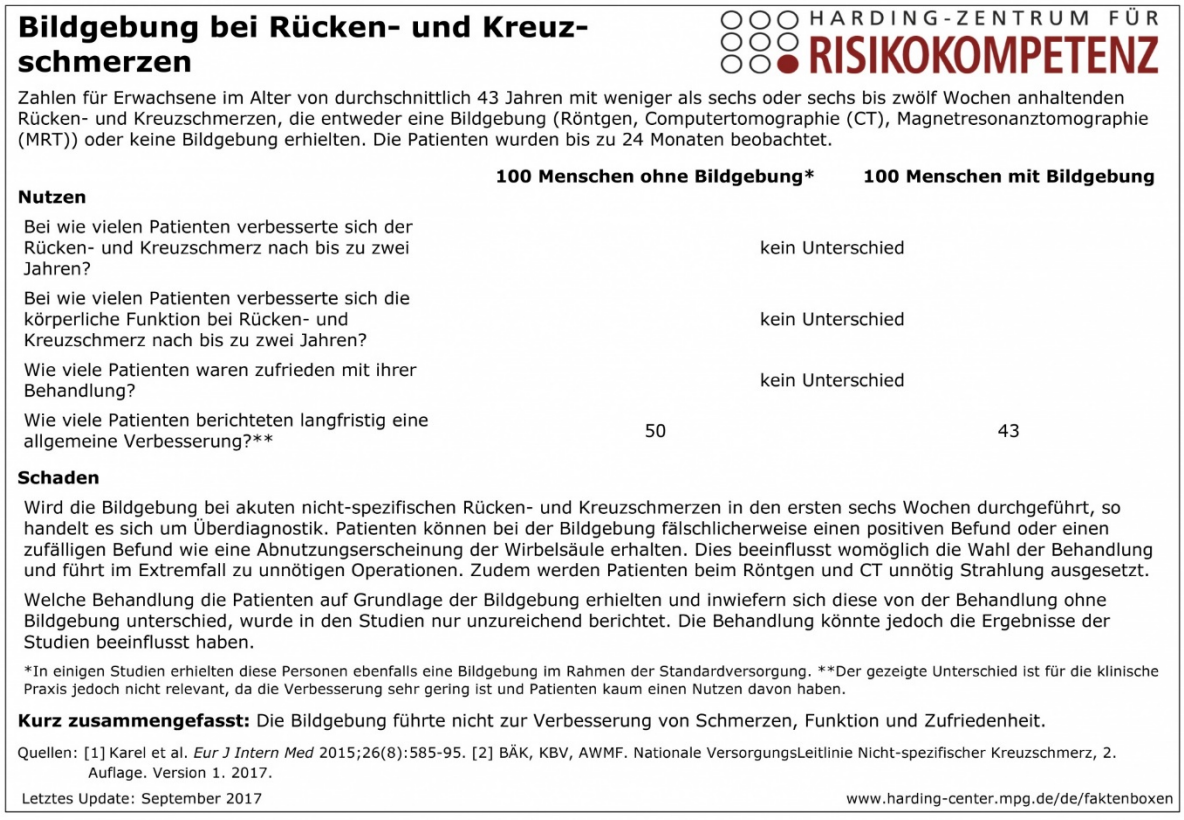
Ein in der klinischen Praxis wichtiges Instrument der EbM sind evidenzbasierte klinische Referenzwerke, die Erkenntnisse aus wissenschaftlichen Studien und Empfehlungen aus Leitlinien für die klinische Praxis aufbereiten und zugänglich machen. Das am weitesten verbreitete Referenzwerk dieser Art ist UpToDate, ein von dem medizinischen Fachverlag WoltersKluwer kuriertes Onlineportal. UpToDate beinhaltet knapp 12.000 Artikel, die auf Grundlage systematischer Literaturrecherchen von Fachautoren erstellt und regelmäßig aktualisiert werden. Die Behandlungsempfehlungen orientieren sich oft an den US-amerikanischen Leitlinien und können sich daher teilweise von deutschen Handlungsempfehlungen unterscheiden. UpToDate ist nicht frei zugänglich, allerdings besitzen viele Kliniken einen Zugang (www.uptodate.com).

Auch die kommerzielle Online-Lernplattform AMBOSS bietet mit dem sog. Arztmodus seit einiger Zeit zu einer zunehmenden Zahl an Themen vertiefte klinische Informationen, die auf Leitlinien, Lehrbüchern, UpToDate und anderen Quellen basieren (www.amboss.com/de/aerztinnen-aerzte).

Speziell für den hausärztlichen Bereich wurde die Online-Plattform Deximed entwickelt, die für Studierende ein kostenloses Abonnement anbietet (https://deximed.de). An der Weiterentwicklung der Plattform und ihrer Inhalte ist die Deutsche Gesellschaft für Allgemeinmedizin (DEGAM) beteiligt, was sicherstellen soll, dass aktuelle Studienevidenz und DEGAM-Leitlinien berücksichtigt werden. Die Plattform enthält ca. 3.800 Artikel.

3.5. Faktenboxen

Faktenboxen sind ein u.a. vom Max-Planck-Institut für Bildungsforschung in Berlin entwickeltes Format für die übersichtliche und verständliche Darstellung quantitativer Daten zu gesundheitsrelevanten Fragestellungen. Sie richten sich an Gesundheitsfachkräfte und PatientInnen, und beruhen auf Forschung dazu, in welcher Form Menschen Zahlen zu Häufigkeiten und Risiken am besten verstehen und bei Entscheidungen berücksichtigen. So können z.B. die meisten Menschen Angaben zu Risiken besser verstehen, wenn sie in ganzen Zahlen ausgedrückt werden (z.B. „5 von 100 PatientInnen“) und nicht als Prozentangaben (z.B. „5%“). Abbildungen 6 und 7 zeigen zwei Beispiele für Faktenboxen. In Deutschland werden Faktenboxen vom Harding-Zentrum für Risikokompetenz in Zusammenarbeit mit der AOK und der Bertelsmann Stiftung erstellt (siehe www.hardingcenter.de/de/projekte-und-kooperationen/faktenboxen).



Bildgebung bei Rücken- und Kreuzschmerzen

HARDING-ZENTRUM FÜR RISIKOKOMPETENZ

Zahlen für Erwachsene im Alter von durchschnittlich 43 Jahren mit weniger als sechs oder sechs bis zwölf Wochen anhaltenden Rücken- und Kreuzschmerzen, die entweder eine Bildgebung (Röntgen, Computertomographie (CT), Magnetresonanztomographie (MRT)) oder keine Bildgebung erhielten. Die Patienten wurden bis zu 24 Monaten beobachtet.

	100 Menschen ohne Bildgebung*	100 Menschen mit Bildgebung
Nutzen		
Bei wie vielen Patienten verbesserte sich der Rücken- und Kreuzschmerz nach bis zu zwei Jahren?	kein Unterschied	
Bei wie vielen Patienten verbesserte sich die körperliche Funktion bei Rücken- und Kreuzschmerz nach bis zu zwei Jahren?	kein Unterschied	
Wie viele Patienten waren zufrieden mit ihrer Behandlung?	kein Unterschied	
Wie viele Patienten berichteten langfristig eine allgemeine Verbesserung?**	50	43

Schaden

Wird die Bildgebung bei akuten nicht-spezifischen Rücken- und Kreuzschmerzen in den ersten sechs Wochen durchgeführt, so handelt es sich um Überdiagnostik. Patienten können bei der Bildgebung fälschlicherweise einen positiven Befund oder einen zufälligen Befund wie eine Abnutzungserscheinung der Wirbelsäule erhalten. Dies beeinflusst womöglich die Wahl der Behandlung und führt im Extremfall zu unnötigen Operationen. Zudem werden Patienten beim Röntgen und CT unnötig Strahlung ausgesetzt.

Welche Behandlung die Patienten auf Grundlage der Bildgebung erhielten und inwiefern sich diese von der Behandlung ohne Bildgebung unterschied, wurde in den Studien nur unzureichend berichtet. Die Behandlung könnte jedoch die Ergebnisse der Studien beeinflusst haben.

*In einigen Studien erhielten diese Personen ebenfalls eine Bildgebung im Rahmen der Standardversorgung. **Der gezeigte Unterschied ist für die klinische Praxis jedoch nicht relevant, da die Verbesserung sehr gering ist und Patienten kaum einen Nutzen davon haben.

Kurz zusammengefasst: Die Bildgebung führte nicht zur Verbesserung von Schmerzen, Funktion und Zufriedenheit.

Quellen: [1] Karel et al. *Eur J Intern Med* 2015;26(8):585-95. [2] BÄK, KBV, AWMF. Nationale VersorgungsLeitlinie Nicht-spezifischer Kreuzschmerz, 2. Auflage. Version 1. 2017.

Letztes Update: September 2017 www.harding-center.mpg.de/de/faktenboxen

Abbildung 6: *Faktenbox zur Bildgebung bei unkomplizierten Rückenschmerzen. Quelle: Harding-Zentrum für Risikokompetenz.*

Behandlungsstrategien bei hohem Risiko für einen und bei einem Herzinfarkt
(Angina pectoris und Nicht-ST-Hebungsinfarkt)

HARDING-ZENTRUM FÜR RISIKOKOMPETENZ

Zahlen für Erwachsene bis 75 Jahre mit einer instabilen Angina pectoris oder einem Nicht-ST-Hebungsinfarkt, die über 6 bis 12 Monate beobachtet wurden und entweder mit vorerst medikamentöser Behandlungsstrategie oder direkt invasiver Behandlungsstrategie versorgt wurden.

	100 Menschen mit vorerst medikamentöser Behandlungsstrategie	100 Menschen mit direkt invasiver Behandlungsstrategie
Nutzen		
Wie viele Menschen erlitten einen Herzinfarkt?	8	6
Wie viele Menschen litten unter anhaltenden Anginaschmerzen (refraktäre Angina)*?	33	21
Wie viele Menschen starben insgesamt?	4	4
Schaden		
Wie viele Menschen litten unter Blutungen als Komplikationen eines Eingriffes?	4	7
Wie viele Menschen litten unter einem durch einen Eingriff bedingten Herzinfarkt?	3	6
Wie viele Menschen mussten aufgrund eines akuten Koronarsyndroms (z.B. Brustschmerz) ins Krankenhaus wiederaufgenommen werden?	29	22

*Anhaltende Anginaschmerzen nach medikamentöser, Katheter- oder chirurgischer Behandlung, die durch weitere Behandlungen nicht verbessert werden können.

Kurz zusammengefasst: Bei der direkt invasiven Behandlungsstrategie erlitten weniger Menschen einen Herzinfarkt und weniger Menschen litten unter anhaltenden Anginaschmerzen. Die Behandlung hatte allerdings keinen Einfluss auf die Anzahl an Menschen, die insgesamt starben. Die direkt invasive Behandlungsstrategie war mit mehr Blutungskomplikationen und durch den Eingriff bedingte Herzinfarkte verbunden, aber führte zu weniger Krankenhauswiederaufnahmen.

Quelle: Fanning et al. *Cochrane Database Syst Rev* 2016(5):CD004815.

Letztes Update: März 2017 www.harding-center.mpg.de/de/faktenboxen

Abbildung 7: *Faktenbox zu Behandlungsstrategien bei instabiler Angina Pectoris. Quelle: Harding-Zentrum für Risikokompetenz.*

Die für den hausärztlichen Bereich entwickelte Software Arriba ermöglicht es, Faktenboxen anhand von Laborparametern und klinischen Befunden für einzelne PatientInnen zu individualisieren, die zugrundeliegenden Daten anschaulich und allgemeinverständlich zu visualisieren und direkt in die gemeinsame Entscheidungsfindung mit PatientInnen zu integrieren. Für Studierende ist das Programm kostenlos (https://arriba-hausarzt.de/).

3.6. Health Technology Assessments (HTAs)

Health Technology Assessments (HTAs) haben das Ziel, wissenschaftlich fundierte, interdisziplinäre Entscheidungshilfen für gesundheitspolitisch relevante Fragestellungen zu liefern. HTAs dienen der systematischen Bewertung der Auswirkungen von Technologien der medizinischen Versorgung (z.B. Medikamente, Medizinprodukte, Hilfsmittel, OP-Verfahren und andere Prozeduren) auf die Gesellschaft, unter Berücksichtigung gesundheitlicher, sozialer, wirtschaftlicher, rechtlicher und ethischer Aspekte.

Anders als klinische Leitlinien sollen HTAs keine Hilfestellungen bei Entscheidungen zur Behandlung einzelner PatientInnen liefern; sie sollen vielmehr die Entscheidungsgrundlage für übergeordnete Fragen der Gesundheitsversorgung sein. Im Zentrum steht dabei der Mehrwert, den eine Gesundheitstechnologie im Vergleich zu anderen neuen oder zu bestehenden Gesundheitstechnologien bietet. Dabei werden Wirksamkeit, Sicherheit, Kosten sowie ethische und soziale Aspekte berücksichtigt.

In Deutschland führen primär zwei Institutionen HTAs durch: das Deutsche Institut für Medizinische Dokumentation und Information (DIMDI) und das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Im Deutschen Gesundheitssystem spielen HTAs eine zentrale Rolle bei der Entscheidung zur Übernahme von Maßnahmen in den Leistungskatalog der Gesetzlichen Krankenversicherung durch den Gemeinsamen Bundesausschuss (G-BA).

4. EbM für die Praxis – die fünf Schritte der EbM

In der klinischen Praxis stellt sich immer wieder die Frage, welche Therapie für eine PatientIn optimal ist, was über die Wirkungen und Nebenwirkungen verschiedener Therapien bekannt ist, oder wie die Aussagekraft bestimmter diagnostischer Parameter zu beurteilen ist. Um in diesen Fällen wissenschaftliche Evidenz systematisch zu nutzen, bietet es sich an, in fünf Schritten vorzugehen [2]:

1) Die Formulierung einer klaren Fragestellung.
2) Die Suche nach der besten verfügbaren wissenschaftlichen Evidenz.
3) Die kritische Prüfung der wissenschaftlichen Erkenntnisse hinsichtlich ihrer Verlässlichkeit und Relevanz.
4) Die Anwendung der wissenschaftlichen Erkenntnisse.
5) Die Bewertung der Umsetzung, ob durch eine kritische Reflektion oder eine begleitende formale Evaluation.

Anmerkung: Als Hilfsmittel für die Praxis hat das MeCuM-EbM Team einen *EbM Pocket Guide entwickelt, welcher als kurzes Nachschlagewerk und Hilfsmittel für die Praxis dienen soll. An entsprechenden Teilen dieses Skripts finden sich Verweise auf den Pocket Guide. Beispielsweise findet sich in diesem auch eine kurze Erinnerungshilfe zu den fünf Schritten der EbM.*

4.1. Formulierung einer klaren Fragestellung

Zur Formulierung einer klaren Fragestellung kann in vielen Fällen das sogenannten PICO-Modell (***P**opulation, **I**ntervention, **C**omparison, **O**utcomes*) genutzt werden (siehe Tabelle 6). Es ist besonders nützlich, wenn die Wirksamkeit einer Maßnahme beurteilt werden soll. Das PICO-Modell kann auch über die EbM hinaus bei der Formulierung wissenschaftlicher Fragestellungen angewandt werden, und kann z.B. bei der Formulierung einer Fragestellung für eine medizinische Dissertation hilfreich sein. In dem in Tabelle 6 dargestellten Beispiel würde die Fragestellung lauten „Wie wirkt sich bei PatientInnen mit unkomplizierten Rückenschmerzen weiterführende Diagnostik per Bildgebung im Vergleich zur Standardtherapie ohne bildgebende Diagnostik auf die Intensität der Schmerzen nach 1-3 Wochen, Einschränkungen im Alltag und die Patientenzufriedenheit aus?“

Tabelle 6: Das PICO-Modell für die Konkretisierung einer Fragestellung	
PICO-Element	**Frage zur Konkretisierung der Fragestellung**
Population	Um was für Personen geht es? (z.B. PatientInnen mit unkomplizierten Rückenschmerzen)
Intervention	Um welche Maßnahmen geht es? (z.B. weiterführende Diagnostik per Bildgebung)

Comparison	Mit was soll diese Maßnahme verglichen werden? (z.B. Standard-Therapie ohne bildgebende Diagnostik)
Outcome	Anhand von welchen Endpunkten oder Kriterien soll die Maßnahme bewertet werden? (z.B. Intensität der Schmerzen nach 1-3 Wochen, Einschränkungen im Alltag, Zufriedenheit der PatientInnen)

4.2. Evidenzsuche

Im klinischen Alltag stellt sich Studierenden und ÄrztInnen aller Erfahrungsstufen regelmäßig die Aufgabe, zu einer definierten Fragestellung die bestverfügbare Evidenz zu identifizieren. Bei klinischen Fragestellungen bietet es sich an, hierfür zunächst klinische Referenzwerke zu konsultieren und nach relevanten, aktuellen Leitlinien zu suchen. Liefert dies nicht die notwendigen Informationen, empfiehlt es sich, nach systematischen Übersichtsarbeiten zu suchen, und Einzelstudien nur dann unterstützend heranzuziehen, wenn keine passenden systematischen Übersichtsarbeiten verfügbar sind. Für Leitlinien, systematische Übersichtsarbeiten und Einzelstudien gibt es jeweils spezifische Datenbanken, in denen nach relevanten Dokumenten gesucht werden kann (siehe weiterführende Ressourcen).

Bei Fragen, die nicht die unmittelbare PatientInnenversorgung betreffen (z.B. wissenschaftliche Fragen, die im Rahmen der Doktorarbeit auftreten), helfen klinische Referenzwerke und Leitlinien oft nicht weiter. In diesen Fällen kann es sich anbieten, direkt nach systematischen Übersichtsarbeiten sowie Primärstudien zu suchen.

Die Suche in wissenschaftlichen Datenbanken basiert darauf, dass bestimmte Schlagwörter in der Datenbank gesucht werden. Die Schlagworte können dabei mit bestimmten Operatoren miteinander verknüpft werden. Zentrale Operatoren sind über verschiedene wissenschaftliche Datenbanken hinweg gleich. Die wichtigsten hiervon sind die Operatoren AND, OR und NOT. Wenn man zwei Suchbegriffe z.B. „Diabetes“ und „Metformin“ mit OR verknüpft, erhält man alle Studien, welche “Diabets“ oder „Metformin“ (oder beide Begriffe) im entsprechenden Feld (z.B. dem Titel der Studie) aufweisen. Bei einer AND-Verknüpfung erhält man nur Studien, die sowohl „Diabetes“ als auch „Metformin“ aufweisen. Durch eine Verknüpfung mit NOT kann man Begriffe ausschließen, beispielsweise nur Studien mit „Diabetes“ im Titel, welche nicht das Schlagwort Metformin aufweisen (Siehe Abbildung 8)

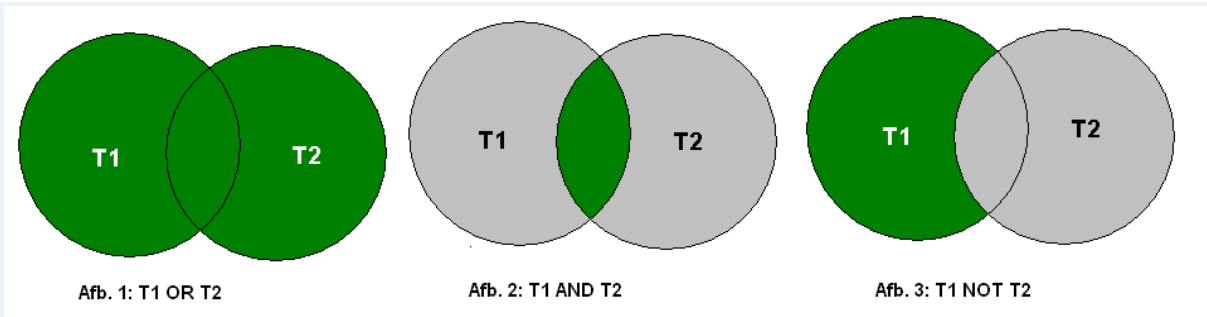


Abbildung 8: Wichtige Operatoren für die Suche in wissenschaftlichen Datenbanken

Praktisch alle wissenschaftlichen Datenbanken erlauben, dass man die Suche der Begriffe auf bestimmte Felder des Datenbankeintrags beschränkt. Dies kann mit dem entsprechenden „Builder“ der Webseite durchgeführt werden, oder durch bestimmte Operatoren, welche sich jedoch für die einzelnen Datenbanken unterscheiden. Beispielsweise liefert die Datenbank PubMed bei Suche von “Diabetes[Title/Abstract]“ nur Studien, welche im Titel oder Abstract der Studie das Wort Diabetes aufweisen.
Beim Bauen einer guten Suchstrategie gilt es darauf zu achten, dass diese sensitiv genug ist, um keine relevante Studie zu verpassen, jedoch auch gleichzeitig spezifisch genug, dass das Verhältnis von passenden zu unpassenden Studien in einem für den verfügbaren Zeitraumen in einem sinnvollen Rahmen bleibt. Für eine systematische Übersichtsarbeit, wo ein Schwerpunkt auf Sensitivität gelegt wird, kann sich das Prüfen der Einschlussfähigkeit (das s.g. Screenen) je nach Thema auf mehrere tausend Studien erstrecken.

Die wichtigste Datenbank für Leitlinien im deutschen Kontext ist die Leitlinien-Plattform der Arbeitsgemeinschaft wissenschaftlich-medizinischer Fachgesellschaften (AWMF) (www.awmf.org/leitlinien/leitlinien-suche.html). Nach systematischen Übersichtsarbeiten kann über die Suchplattform Epistemonikos (www.epistemonikos.org/de) sowie in der Cochrane Library (www.cochranelibrary.com) gesucht werden. Die größte Datenbank für gesundheitsbezogene Veröffentlichungen, die Einzelstudien und systematische Übersichtsarbeiten enthält, ist PubMed (www.ncbi.nlm.nih.gov/pubmed). PubMed ist eine öffentliche Datenbank, die von den *National Institutes of Health*, einer staatlichen Gesundheitsbehörde in den USA, unterhalten wird.

Die Lehrveranstaltung EBM Workshop vertieft und erprobt die Erstellung von Suchstrategien und die Suche in unterschiedlichen Datenbanken. Im EbM Poket Guide findet sich auch eine kurze Hilfestellung zur Literaturrecherche.

Anwendungsbeispiel: Sie möchten eine rasche Literatursuche für die klinische Praxis durchführen um sich über den aktuellen Stand der Wirksamkeit des Medikaments Metformin bei PatientInnen mit Diabetes Mellitus Typ II mit Blick auf den Endpunkt Reduktion der Mortalität zu informieren. In der Datenbank PubMed führen Sie daher eine fokussierte Literatursuche mit folgenden Begriffen durch: ("metformin"[Title] AND "diabetes"[Title/Abstract]) AND ("mortality"[Title/Abstract] OR „survival"[Title/Abstract]). Sie beschränken die Suche mittels der Filterfunktion der Plattform auf systematische Übersichtsarbeiten und Publikationen der letzten 5 Jahre. So erhalten Sie 34 Treffer, welche praktisch alle von Relevanz für ihre Fragestellung sind.

4.3. Kritische Prüfung der Evidenz

Ein weiterer wichtiger Schritt bei der Umsetzung von EbM ist die kritische Prüfung (engl. *critical appraisal*) der Studien, die man bei der Literatursuche identifiziert hat [1]. Hierzu gehören:

- die Überprüfung der **externen Validität**, das heißt die Generalisierbarkeit bzw. Übertragbarkeit und Anwendbarkeit der Ergebnisse der Studie auf die vorliegende Fragestellung oder die Frage.
- die Überprüfung der **internen Validität**, d.h. der Glaubwürdigkeit der Ergebnisse der Studie, d.h. inwieweit wir uns darauf verlassen können, dass der beobachtete Effekt nicht durch systematische oder zufällige Fehler verzerrt wurde.

- die Überprüfung der **Relevanz**, d.h. der klinischen Bedeutung der Ergebnisse für den konkreten Fall.

Im EbM Pocket Guide finden sich hierzu einen kurze Zusammenfassung relevanter Begriffe, eine Anleitung zum Formulieren von PICO-Fragestellungen sowie für die schnelle kritische Bewertung der externen und internen Validität und der Relevanz von Interventionsstudien.

Fallbeispiel: Eine 80-jährige Patientin mit Diabetes Mellitus Typ II, Bluthochdruck und einer Herzinsuffizienz (NYHA II) weist einen HbA1c-Wert von 8.1 % auf (HbA1c-Zielkorridor 6,5 – 7,5 %). Sie überlegen, ein zusätzliches Medikament zur Blutzuckerkontrolle anzulegen. Ein – hypothetisches – neues Medikament wurde bisher in einer RCT auf die Wirksamkeit getestet.

4.3.1. Externe Validität

Die Überprüfung der externen Validität kann anhand des eingangs erwähnten PICO-Modells erfolgen (siehe Tabelle 7).

Tabelle 7: Die PICO-Elemente für die Prüfung der Anwendbarkeit wissenschaftlicher Evidenz auf eine konkrete Fragestellung	
PICO-Elemente	**Frage zur Überprüfung der Anwendbarkeit**
Population	Sind die TeilnehmerInnen der Studie mit den Personen vergleichbar, um die es geht?
Intervention	Ist die in der Studie untersuchte Maßnahme mit der Maßnahme vergleichbar, die bewertet werden soll?
Comparison	Wurde in der Studie die untersuchte Maßnahme mit derselben Alternative verglichen, die auch im vorliegenden Fall relevant ist?
Outcome	Wurden in der Studie die Kriterien untersucht, die auch im vorliegenden Fall von Interesse sind?

Fallbeispiel: Sie überprüfen die PICO-Elemente der Studie und stellen fest, dass die Studienpopulation größtenteils männlich (70%) war, deutlich jünger als Ihre Patientin (im Durchschnitt 55 Jahre mit nur 10% der StudienteilnehmerInnen über 70 Jahre) und das eine Herzinsuffizienz ein Ausschlusskriterium zur Teilnahme an der Studie darstellte.
Sie fragen sich, ob die externe Validität für ihren konkreten Fall gegeben ist, d.h. ob die Studienergebnisse auf die Situation Ihrer Patientin übertragbar sind.

4.3.2. Interne Validität

Die interne Validität bezeichnet das Ausmaß, in dem eine Studie tatsächlich das misst, was sie zu messen ausgibt. Die interne Validität einer Studie kann u.a. durch systematische Fehler (engl. *bias*), durch Zufallsfehler sowie sog. *Confounding* eingeschränkt werden. In Abschnitt 5 werden diese Konzepte genauer erklärt.

Fallbeispiel: Bei der Studie handelt es sich um eine große, doppelt-verblindete randomisiert-kontrollierte Studie. Sie finden keinen Hinweis auf systematische Verzerrungen. Die Effekte sind

statistisch signifikant (d.h. mit hoher Wahrscheinlichkeit nicht durch Zufall erklärbar). Sie nehmen daher an, dass die interne Validität der Studie gegeben ist und sie daher die Studienergebnisse für glaubwürdig halten können.

4.3.3. Klinische Relevanz

Neben der externen und internen Validität ist zudem die *klinische Relevanz* des jeweiligen Ergebnisses zu prüfen. Die klinische Relevanz bezieht sich darauf, ob ein beobachteter Effekt praktische Auswirkungen auf (Therapie-)Entscheidungen hat. Die klinische Relevanz ist insbesondere von der Effektstärke abhängig, kann aber auch von weiteren Faktoren (wie z.B. Kosten oder Nebenwirkungen) beeinflusst werden.

Zentrale Fragen bei dieser Bewertung der Relevanz können unter anderem sein:

- Was für eine Art von Endpunkt (Outcome) wurde betrachtet und für wie relevant wird dieser von dem/der PatientIn eingeschätzt? (z.B. handelt es sich um einen Laborwert (z.B. HbA1c) oder um Lebenszeit oder Einschränkung der Lebensqualität?)
- Wie stark ist die zu erwartende Veränderung und wie wahrscheinlich tritt diese für den/die PatientIn auf? (z.B. wie viele Personen müssen die Therapie (z.B. eine Darmspiegelung) erhalten, damit ein Endpunkt (z.B. ein Todesfall durch Darmkrebs) verhindert wird?)
- Wie schwer oder (un)erträglich ist der aktuelle Zustands? (z.B. könnten PatientInnen mit starker Symptombelastung und schlechter Prognose auch in geringen Erfolgsaussichten eine Relevanz sehen.)
- Welche Alternativen sind verfügbar? (z.B. sind ähnliche Verbesserungen durch andere Maßnahmen erreichbar (z.B. Sport oder Ernährungsänderung im Vergleich zu einer zusätzlichen Tablette)?)

Fallbeispiel: Die von Ihnen identifizierte Studie zeigte, dass bei Gabe des neuen Medikaments zusätzlich zu Metformin eine statistisch signifikante Reduktion des HbA1C von 0,2 Prozentpunkten zu erwarten ist. Daten zu klinischen Endpunkten oder zur Mortalität liegen nicht vor.
In diesem Fall könnte man sich ggf gegen die Gabe des Medikamentes entscheiden, da die klinische Relevanz fraglich erscheint.

4.3.4. Bewertungsinstrumente für die Praxis

Für eine erste, orientierende Prüfung der externen und internen Validität können Critical-Appraisal-Checklisten verwendet werden (siehe weiterführende Ressourcen) [21]. In systematischen Übersichtsarbeiten wird die interne Validität von Studien in der Regel mit *Instrumenten zur Beurteilung des Verzerrungsrisikos* (engl. *risk of bias assessment tools*) untersucht. Das Verzerrungsrisiko (engl. *risk of bias*) beschreibt die Wahrscheinlichkeit, mit der das Ergebnis einer Studie systematisch vom wahren Wert abweicht (was sowohl zu einer Über- als auch Unterschätzung des Effekts führen kann). Beispiele für Instrumente zur Beurteilung des Verzerrungsrisikos sind das *Cochrane Risk of Bias Tool* sowie das *Risk Of Bias In Non-Randomized Studies of Interventions tool* (ROBINS-I) (siehe weiterführende Ressourcen) [22].

Weitere Details zur kritischen Prüfung von Evidenz werden im Abschnitt 5 (Kriterien für Studienqualität) dargestellt.

4.4. Anwendung der Evidenz und Bewertung der Umsetzung

Bei der Anwendung der Evidenz auf den jeweiligen konkreten Fall kommt neben der wissenschaftlichen Evidenz den beiden weiteren Aspekten der EbM eine besonders wichtige Rolle zu: den Werten und Präferenzen der PatientIn, und der klinischen Erfahrung der Behandelnden.

Im Sinne der informierten Einwilligung und des gemeinsamen Entscheidens sollte die Patientin, soweit im jeweiligen Fall möglich und angemessen, über relevante Informationen zu den verschiedenen möglichen Handlungsoptionen aufgeklärt werden. Hierbei können auch Visualisierungen von Daten, z.B. Faktenboxen oder das Arriba-Programm (siehe Abschnitt 3.5), eingesetzt werden.

Bei der Anwendung und Umsetzung wissenschaftlicher Evidenz – wie z.B. von Empfehlungen aus Leitlinien – ist klinisch-praktische Erfahrung unabdingbar. Dies gilt zum Beispiel für die Frage, wann Empfehlungen aus Leitlinien anwendbar sind, und in welchen begründeten Fällen von diesen abgewichen werden sollte. Die Bewertung der Umsetzung kann durch begleitende Reflektion erfolgen, sowohl individuell als auch in Formaten wie Teambesprechungen und Fallkonferenzen.

5. Mögliche Fehlerquellen in wissenschaftlichen Studien

5.1. Überblick

Im Folgenden werden wichtige mögliche Fehlerquellen wissenschaftlicher Studien vorgestellt. Dies soll dazu dienen, die entsprechenden Konzepte in allgemeiner Form einzuführen. Tiefergehende mathematische und statistische Aspekte werden in späteren Lehrveranstaltungen des MeCuM-Science-Curriculums behandelt.

Mögliche Fehlerquellen in wissenschaftlichen Studien sind:

- Zufallsfehler
- Systematische Fehler (engl. *bias*)
- Störfaktoren (engl. *confounder*)
- Bewusste Manipulation

5.2. Zufallsfehler

5.2.1. Zufallsfehler, Präzision, und das Konfidenzintervall

Ein wichtiges Kriterium für die Aussagekraft des Ergebnisses einer Studie ist die *Präzision* bzw. das Ausmaß des *Zufallsfehlers*. Die Bedeutung dieser beiden Konzepte soll im Folgenden anhand des Beispiels einer (hypothetischen) Beobachtungsstudie verdeutlicht werden:

Nehmen wir an, unser Ziel sei es, die durchschnittliche Körpergröße der deutschen Bevölkerung zu erfassen. Die genaueste Methode zur Feststellung des Wertes wäre die Vermessung der gesamten deutschen Bevölkerung– in einer gewaltigen Querschnittstudie. Nehmen wir an, wir hätten alle 83 Mio.

Einwohner Deutschland vermessen und hätten (hypothetisch) dabei einen Wert von 173 cm für die Durchschnittsgröße der Bevölkerung erhalten. Dabei würden wir auch feststellen, dass 95% der Körpergrößen zwischen 161 cm und 185 cm liegen: nur 5% der Bevölkerung sind größer oder kleiner (siehe Abbildung 8).

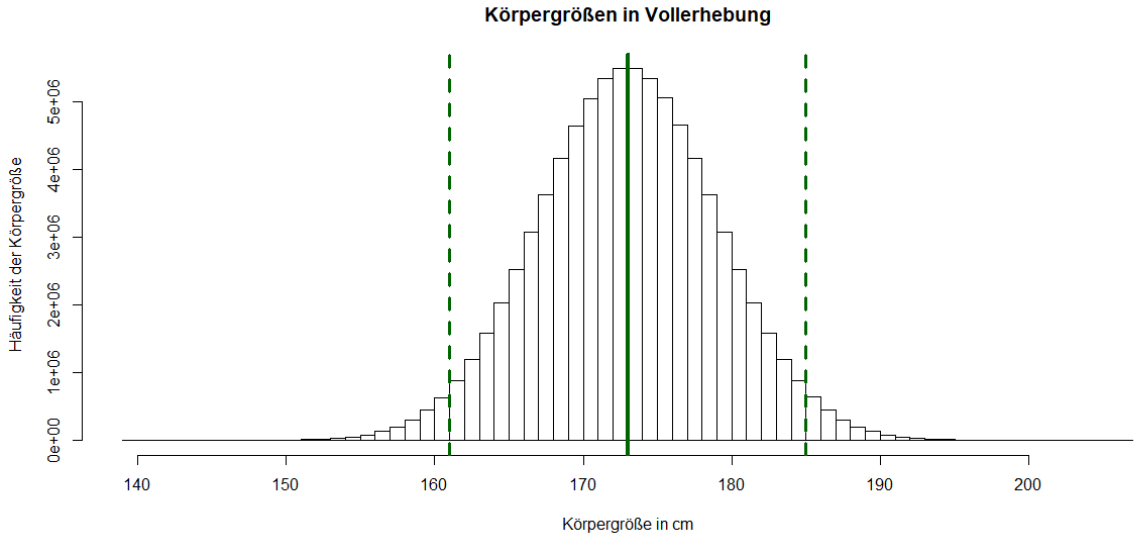


***Abbildung 9**: Größenverteilung der Gesamtpopulation. Es ergibt sich ein Durchschnittswert von 173 cm (durchgezogene grüne Linie). 95% der Bevölkerung hat eine Körpergröße zwischen 161 und 185 cm (gestrichelte grüne Linien).*

Da eine Vermessung der gesamten deutschen Bevölkerung nicht praktikabel ist, vermessen wir stattdessen eine Stichprobe (engl. *sample*): In einem zweiten Gedankenexperiment greifen wir 25 Personen zufällig aus der Gesamtpopulation heraus, und vermessen deren Körpergröße. Die Auswahl der Personen erfolgt rein zufällig – es handelt sich somit um eine *Zufallsstichprobe*. Dies bedeutet, dass jede Person in der Gesamtpopulation die gleiche Wahrscheinlichkeit hat, ausgewählt zu werden. Aufgrund der natürlichen Varianz der Körpergröße in der Bevölkerung, und der zufälligen Auswahl einer nur kleinen Stichprobe wird der so bestimmte Wert mit einer gewissen Wahrscheinlichkeit von dem wahren Wert abweichen. In unserem Beispiel (siehe Abbildung 10) ergibt sich in der Zufallsstichprobe ein Wert von 170.5 cm (rote Linie), der von dem wahren Wert von 173.0 cm (grüne Linie) abweicht.

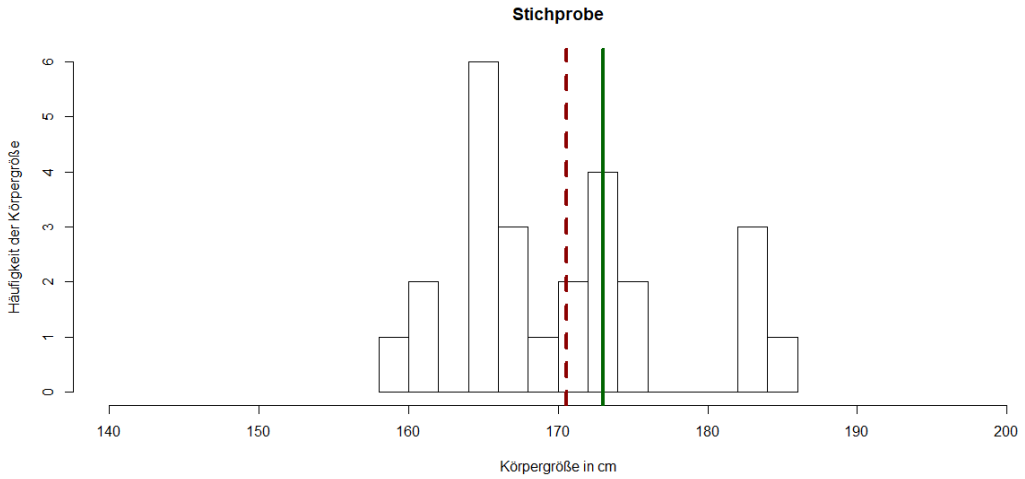


Abbildung 10: *Verteilung der Körpergrößen in einer Stichprobe von 25 Personen, die zufällig aus der Gesamtpopulation gezogen wurden*

Diese Abweichung zwischen dem Mittelwert der Stichprobe und dem Mittelwert der Gesamtpopulation (dem wahren Wert) bezeichnet man als ***Zufallsfehler***. Je kleiner der Zufallsfehler, desto größer ist die ***Präzision*** (engl. *precision*) des Ergebnisses. Er wird Zufallsfehler genannt, weil er sich aus der zufälligen Auswahl der Stichprobe ergibt, und nicht aus einem systematischen Fehler bei der Messung.

Eine Möglichkeit, die Größe des Zufallsfehlers zu quantifizieren sind sog. Konfidenzintervalle (abgekürzt *KI,* engl. *confidence interval*). Auch dieses Konzept sei anhand eines Beispiels illustriert: In diesem dritten Gedankenexperiment wiederholen wir das zweite Experiment – das Ziehen einer Zufallsstichprobe von 25 Personen und die Bestimmung des Mittelwerts der Körpergröße dieser 25 Personen – so oft, bis sich die Verteilung der resultierenden Mittelwerte durch weitere Wiederholungen des Experiments nicht mehr ändert. Die resultierenden Mittelwerte streuen alle um den wahren Wert von 173 cm, weichen aber zum Teil auch deutlich von diesem ab. Nehmen wir an, dass 95% der resultierenden Mittelwerte im Bereich von 171-175 cm (173 cm +/- 2 cm) liegen – dies ist das sog. ***95% Konfidenzintervall*** dieser Messung**.** Der Konfidenzintervall gilt für die gezogene Stichprobe und jede neuen Stichprobe ergibt einen anderen Konfidernzintervall. In unserem Beispiel ergäbe das Experiment also das folgende Ergebnis für die durchschnittliche Körpergröße der Bevölkerung: 173 cm (95% KI: 171 – 175 cm). Zu beachten ist, dass das 95% Konfidenzintervall *nicht* besagt, dass 95% der Bevölkerung eine Körpergröße zwischen 171 und 175 cm haben – es besagt vielmehr, dass bei einer hypothetisch unendlichen Wiederholung unseres Experiments (Ziehen einer Zufallsstichprobe von 25 Personen und Bestimmung des Mittelwerts der Körpergröße dieser 25 Personen) 95% der 95%-Konfidenzintervalle den wahren Wert enthalten (hier 173 cm). Würden wir die Stichprobe vergrößern (z.B. statt 25 jedes Mal 100 oder 1000 Personen vermessen) würde das Konfidenzintervall kleiner werden, obwohl sich die zugrundeliegende Verteilung der Körpergröße der Bevölkerung nicht ändert.

In der Realität können wir weder die gesamte Bevölkerung vermessen, noch unser Experiment beliebig oft wiederholen. In der Realität werden wir das Experiment – in unserem Fall das Vermessen einer Zufallsstichprobe von 25 Personen – nur ein einziges Mal durchführen. Wir werden daher einen Wert erhalten, von dem wir nicht wissen, wie weit er vom wahren Wert entfernt ist. Wir wissen nicht, ob wir mit unserem Experiment sehr nahe an dem wahren Wert liegen oder relativ weit davon entfernt (z.B. ob unsere Stichprobe eine von den zahlreichen mit Ergebnissen zwischen 172.0 und 174.0 cm ist, oder einer der Ausreißer mit 169,8 cm oder 176,8 cm). Es ist jedoch möglich, anhand der Verteilung der Körpergrößen innerhalb unserer Zufallsstichprobe mit einer mathematischen Näherungsformel das 95% Konfidenzintervall zu berechnen. Diese Formel lautet (bei normalverteilten Werten) wie folgt:

95% KI = Mittelwert +/- 2 x Standardfehler des Mittelwerts

Die Formel für den Standardfehler des Mittelwerts (engl. *standard error of the mean*, SEM) lautet:

$$SEM = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}$$

Die Details der Berechnung des Konfidenzintervalls werden in späteren Lehrveranstaltungen in MeCuM Science behandelt.

Die Größe des Konfidenzintervalls hängt vom Ausmaß der Varianz der zugrundeliegenden Werte ab, und von der Größe der Stichprobe. Je größer die Varianz, und je kleiner die Stichprobe, desto größer das Konfidenzintervall.

5.2.2. P-Wert und statistische Signifikanz

Ein weiteres zentrales Konzept ist die statistische Signifikanz eines Ergebnisses. Auch dieses soll anhand eines Beispiels erläutert werden. Nehmen wir an, unser Ziel sei es zu klären, ob sich Frauen und Männer in ihrer Körpergröße unterscheiden.

Wir formulieren hierzu zwei Hypothesen:

- **Nullhypothese H_0:** Es gibt *keinen* Unterschied in der durchschnittlichen Körpergröße von Frauen und Männern.
- **Alternativhypothese H_1**: Es gibt *einen* Unterschied in der durchschnittlichen Körpergröße von Frauen und Männern.

Wir ziehen nun zwei Zufallsstichproben von je 100 Frauen und Männern, und bestimmen und vergleichen die Mittelwerte. Nehmen wir an, dass wir eine durchschnittliche Größe von 180 cm bei Männern und 167 cm bei Frauen in unserer Stichprobe finden. Hieraus ergibt sich eine Mittelwertdifferenz von 13 cm (95% KI: 2-24 cm) sowie ein p-Wert von p = 0,04 ergibt.

Der p-Wert beschreibt die Glaubwürdigkeit der Nullhypothese bzw. wie wahrscheinlich es ist, dass es *keinen* Zusammenhang zwischen den beobachteten Phänomenen – in diesem Fall Körpergröße und Geschlecht – gibt.

In unserem Beispiel bedeutet der p-Wert von 0,04 folgendes: Wenn die Nullhypothese richtig wäre (es also keinen Unterschied in der durchschnittlichen Körpergröße von Männern und Frauen gäbe), dann würde sich – rein durch Zufall – bei einer sehr häufigen Wiederholung des Experiments nur in 4% der Fälle eine Mittelwertdifferenz von 13 cm oder größer ergeben.

Anders formuliert: Wenn die Nullhypothese richtig wäre (es also keinen Unterschied in der durchschnittlichen Körpergröße von Männern und Frauen gäbe), dann betrüge die Wahrscheinlichkeit, bei dem genannten Experiment als Ergebnis eine Mittelwertdifferenz von 13 cm oder größer zu messen, 4%. Da dies eine sehr kleine Wahrscheinlichkeit ist, kann die Nullhypothese verworfen werden, und die Richtigkeit der Alternativhypothese angenommen werden.

Wenn die Nullhypothese zugunsten der Alternativhypothese verworfen wird, wird das Resultat als ***statistisch signifikant*** bezeichnet. In verschiedenen wissenschaftlichen Disziplinen haben sich verschiedene Signifikanzniveau wie 5%, 1% oder 0,1% etabliert, wobei sich in den Gesundheitswissenschaften ein Signifikanzniveau von 5% etabliert hat. Ergebnisse gelten dann als statistisch signifikant, wenn der p-Wert kleiner als 0,05 ist.

Wie auch bei der Bestimmung des Konfidenzinterfalls werden auch für die Bestimmung des p-Wert in der Realität nicht Experimente beliebig oft wiederholt. Vielmehr wird der p-Wert mit mathematischen Methoden aus der Größe der Stichprobe und der Verteilung der Werte errechnet.

Die Größe des p-Werts macht keine Aussage über die Größe des wahren Effekts. Bei einer sehr großen Stichprobe können auch sehr kleine Unterschiede statistisch signifikant sein. Deshalb ist neben der statistischen Signifikanz immer auch die klinische Relevanz eines Ergebnisses zu beurteilen.

Des Weiteren ist zu beachten, dass der p-Wert nur die Wahrscheinlichkeit von zufälligen Fehlern erfasst. Er lässt keine Aussagen über das Vorliegen möglicher systematischer Verzerrungen zu (siehe Abschnitt 5.3).

5.2.3. p-Hacking: Manipulation des Zufallsfehlers

In der Medizin wird der p-Wert und das Signifikanzniveau von 0,05 als oft als zentraler Maßstab dafür genommen, ob ein beobachteter Effekt durch Zufallseffekte erklärbar ist (bei einem p-Wert größer als 0,05) oder ob er einem realen Zusammenhang entspricht (bei einem p-Wert kleiner als 0,05). Diese Interpretation ist insofern problematisch, da der p-Wert in Wirklichkeit ein Kontinuum darstellt: Je größer der p-Wert, desto unsicherer sind wir, ob der beobachtete Effekt nicht auch durch Zufall zustande gekommen sein könnte. Die Grenzziehung bei 0,05, und die daran festgemachte Unterscheidung in „statistisch signifikant“ und „statistisch nicht signifikant“ ist zwar allgemein üblich, aber dennoch willkürlich.

Eine Auswirkung der Grenzziehung bei 0,05 ist, dass oft ein starker Anreiz besteht, den p-Wert künstlich unter diese Grenze zu drücken. Dies wird als *p-Hacking* bezeichnet. Dies kann bewusst – in manipulativer Absicht – geschehen, es kann jedoch auch ohne bösartige Absichten durch schlechte Studiendurchführung auftreten. p-Hacking kann finanzielle Motive haben – wenn z.B. pharmazeutische Unternehmen einen „statistisch signifikanten Effekt“ für ein neues Medikament nachweisen möchten. Ein weiteres Motiv ist, dass es als einfacher gilt, Ergebnisse in renommierten Fachzeitschriften zu publizieren, wenn sie statistisch signifikant sind.

Potentiell manipulative Methoden zum Erreichen eines signifikanten p-Werts sind unter anderem:

- Sobald eine Studie statistisch signifikante Ergebnisse zeigt, wird die Datenerhebung (vorzeitig) beendet.
- Wenn die Daten keine statistische Signifikanz zeigen, werden weitere Daten gesammelt (Nacherhebungen).
- Es werden einige Daten (sog. Ausreißer) fallengelassen, die „abwegig“ erscheinen.
- Man fügt der Berechnung weitere Variablen hinzu, bis das Ergebnis signifikant wird.
- Man transformiert die Daten (z.B. logarithmiert) solange, bis die Ergebnisse signifikant werden.
- HARKing, was bedeutet: „Hypothesizing After the Results are Known“. D.h. man führt zahlreiche Versuche durch und stellt die „zu überprüfende Hypothese“ erst auf, nachdem man die Daten analysiert und gesehen hat, welche Ergebnisse signifikante Werte aufweisen. Eine verwandte Praktik ist das „Fischen nach Signifikanz“ (engl. *fishing for significance*): Man testet eine größere Zahl an Zusammenhängen, bis man einen statistisch signifikanten identifiziert.

„Fischen nach Signifikanz“ ist aus dem folgenden Grund irreführend und potentiell manipulativ: Prüft man 100 mögliche Zusammenhänge auf statistische Signifikanz, so ist zu erwarten, dass sich hierbei in

fünf Fällen *rein durch Zufall* ein statistisch signifikantes Ergebnis zeigt (wenn man ein Signifikanzniveau von 5% zugrunde legt). Werden in einer wissenschaftlichen Veröffentlichung nur diese fünf statistisch signifikanten Ergebnisse berichtet, und verschwiegen, dass auch 95 andere Zusammenhänge auf statistische Signifikanz getestet wurden, so erweckt dies den falschen Eindruck, dass die fünf Zusammenhänge, die berichtet werden, mit hoher Wahrscheinlichkeit nicht durch Zufall entstanden seien.

Die Webseite *Spurious Correlations* (www.tylervigen.com/spurious-correlations) illustriert, zu welchen absurden Ergebnissen das „Fischen nach Signifikanz" führen kann. Die Webseite zeigt zahlreiche statistisch hoch signifikante Zusammenhänge zwischen Phänomenen, zwischen denen ganz offensichtlich keinerlei Zusammenhang besteht (siehe Abbildungen 11 und 12). Die Webseite beruht auf dem folgenden Prinzip: Sie fischt aus einer potentiell fast unbegrenzt großen Zahl möglicher Zusammenhänge systematisch statistisch signifikante heraus, und stellt selektiv nur diese dar.

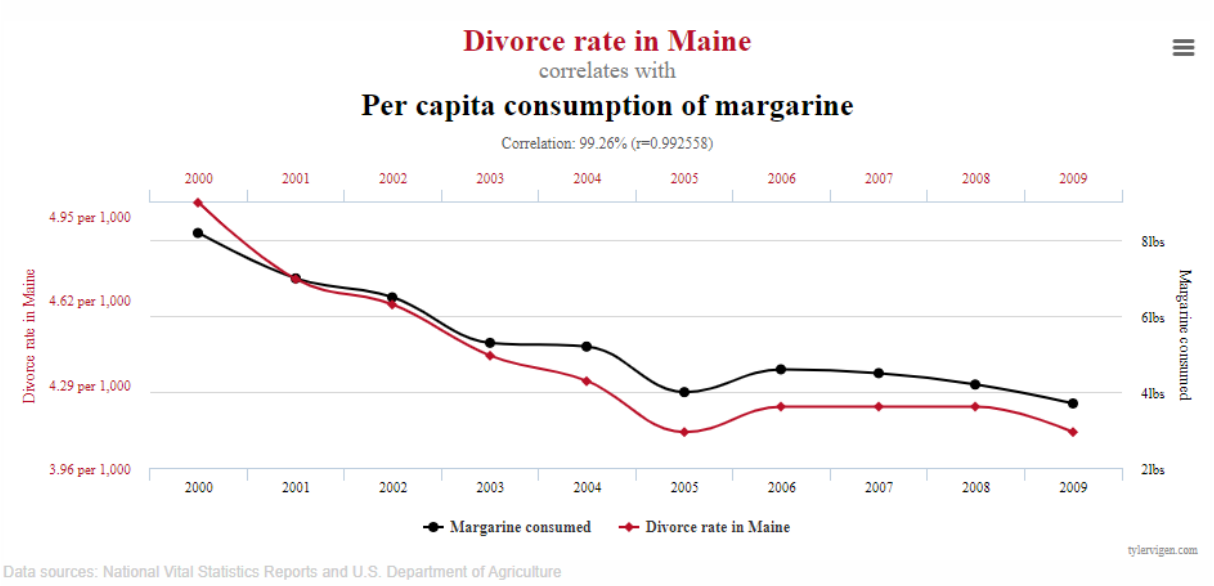


Abbildung 11: *Zusammenhang zwischen dem pro-Kopf-Konsum von Margarine und der Scheidungsrate im US-Bundesstaat Maine. Quelle: www.tylervigen.com/spurious-correlations.*

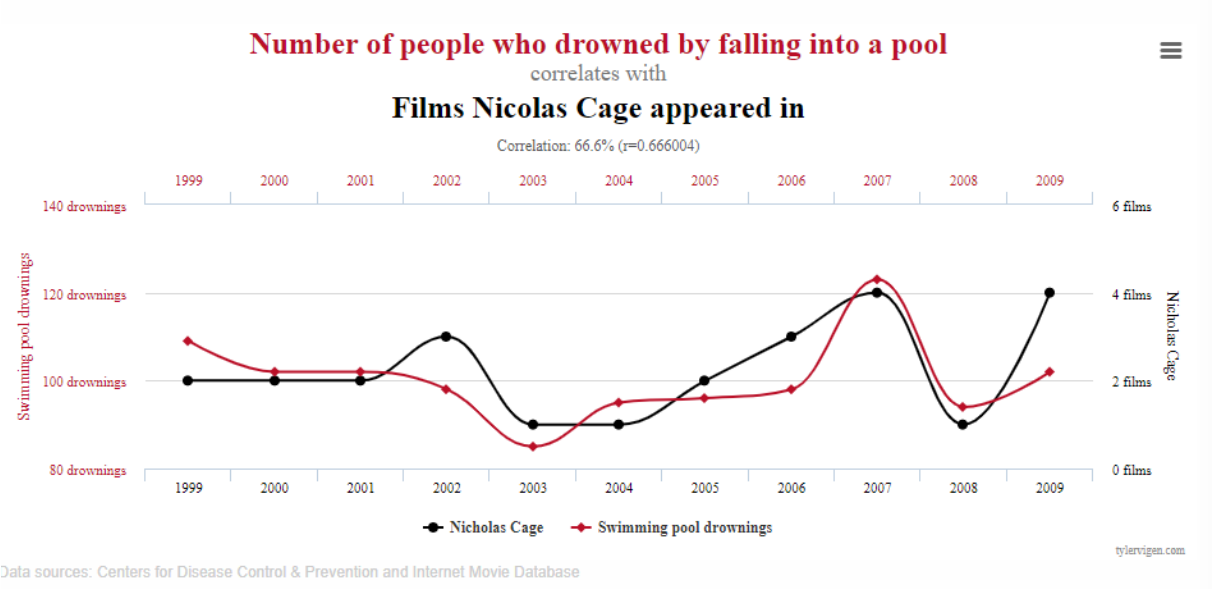


Abbildung 13: *Zusammenhang zwischen der Zahl an Todesfällen durch Ertrinken in Swimmingpools in den USA und der Zahl an neuen Filmen mit Nicolas Cage. Quelle: www.tylervigen.com/spurious-correlations.*

Dies illustriert das folgende Prinzip: Statistische Signifikanz besitzt nur dann Aussagekraft, wenn man vorab eine, oder eine kleine, klar begrenzte Zahl an Null- und Alternativhypothesen formuliert, nur diese auf statistische Signifikanz prüft, und anschließend die Ergebnisse für alle überprüften Zusammenhänge berichtet. Durch die Praxis des Fischens nach Signifikanz verliert das Konzept der statistischen Signifikanz hingegen jegliche Aussagekraft – wie die Spurious-Correlations-Webseite eindrucksvoll illustriert.

Die Beispiele auf der Spurious-Correlations-Webseite sind offensichtlich absurd. Leider kann „Fischen nach Signifikanz" auch weniger offensichtliche, und ernstere Formen annehmen. Hierzu ein Beispiel: Ein pharmazeutisches Unternehmen möchte die Wirksamkeit eines Medikaments in einer Placebo-kontrollierten Studie überprüfen. Als die Studie keine signifikante Wirksamkeit des Medikaments gegenüber der Kontrollgruppe zeigt, führt das Unternehmen zahlreiche Subgruppenuntersuchungen durch (z.B. Wirksamkeit bei Männern zwischen 20-30 Jahren, bei Frauen nach der Menopause etc.), bis sich eines der zahlreichen Ergebnisse als statistisch signifikant herausstellt (auch wenn es sich tatsächlich durch die zahlreichen Ziehungen ausschließlich um ein zufälliges Ergebnis handelt). Die Studie wird am Ende als wirksam für Männer zwischen 50 und 75 Jahren veröffentlicht.

Es gibt verschiedene Wege, um p-Hacking zu vermeiden oder in Studien zu erkennen:

- Vorregistrierung: Das Studienprotokoll, welches festschreibt was genau das Ziel der Untersuchung ist und welche Untersuchungen durchgeführt werden, wird vor Auswertung der Daten veröffentlicht und kann mit der finalen Studie abgeglichen werden.
- Transparenz: Alle erhobenen Variablen und alle durchgeführten Analysen und Versuchsbedingungen werden berichtet. Auch solche, welche negative Ergebnisse gezeigt haben.
- Es werden Konfidenzintervalle verwendet, und der p-Wert vollständig vermieden (einzelne Fachzeitschriften berichten mittlerweile keine p-Werte mehr).
- Wenn eine große Anzahl an Berechnungen durchgeführt wird, wird für multiples Testen statistisch korrigiert (z.B. mit der Bonferroni-Korrektur).
- Wenn eine Analyse eine Kovariable enthält, werden die statistischen Ergebnisse der Analyse mit und ohne die Kovariable(n) angegeben und eine Begründung für die Auswahl der Kovariable(n) geliefert.

5.3. Systematische Fehler (engl. *bias*)

Der im vorausgehenden Abschnitt dargestellte *Zufallsfehler* einer Messung ergibt sich aus der zufälligen Streuung der Messwerte in der jeweiligen Stichprobe. Neben diesem Zufallsfehler gibt es auch sog. *systematische Fehler* bzw. *systematische Verzerrungen* (engl. *bias*). Systematische Fehler verzerren das Ergebnis systematisch in eine bestimmte Richtung. Systematische Fehler können eine Vielzahl von Ursachen haben.

Unterschieden werden u.a. die folgenden Arten von systematischen Fehlern:

- **Stichprobenverzerrung (engl. *selection bias*):** Ein Fehler, der sich aus der Art der Stichprobenziehung ergibt. Wenn z.B. die Stichprobe für das oben genannte Experiment

(Bestimmung der durchschnittlichen Körpergröße der deutschen Bevölkerung) auf einem Basketballturnier gezogen würde, dürfte das Ergebnis systematisch nach oben verzerrt sein.

- **Systematische Messabweichung (engl. *measurement bias***): Ein Fehler, der sich aus der Art der Messung ergibt. Wenn beim genannten Experiment die Teilnehmer mit Schuhen vermessen werden, dürfte dies den Messwert ebenfalls systematisch nach oben verzerren.
- **Ausscheider-Bias (engl. *attrition bias*)**: Ein systematischer Fehler, der sich aus dem Verlust an StudienteilnehmerInnen ergibt. Wenn z.B. in einer Studie zur Wirksamkeit eines neuen Schmerzmittels TeilnehmerInnen, bei denen das Medikament nicht anschlägt, frühzeitig aus der Studie ausscheiden und bei der Analyse nicht berücksichtigt werden, kann sich hieraus eine systematische Überschätzung der Wirksamkeit des Medikaments ergeben.
- **Publikationsbias (engl**. ***publication bias*)**: Wenn Studien mit negativen oder statistisch nicht signifikanten Ergebnissen nicht veröffentlicht werden, kann dies ebenfalls zu einer Überschätzung der Wirksamkeit führen.
- **Soziale Erwünschtheit (engl. *social desirability bias*)**: Bei Untersuchungen, die auf subjektiven Angaben der TeilnehmerInnen beruhen (z.B. bei Befragungen zum Essverhalten oder Alkoholkonsum) können Ergebnisse dadurch verzerrt werden, dass TeilnehmerInnen eher erwünschte oder sozial respektierte Antworten geben. So dürfte ein einfache Umfrage zum Alkoholkonsum wahrscheinlich die Höhe des tatsächlichen Konsums unterschätzen.

Kasten 6: Beispiele für systematische Fehler (bias)

Stichprobenverzerrung (engl. *selection bias*)

Viele medizinische Studien schließen mehr männliche als weibliche TeilnehmerInnen ein. Dabei hat insbesondere die Unterrepräsentation von Frauen in der pharmakologischen Forschung zu einer Gefährdung von Patientinnen durch unerwartete Nebenwirkungen geführt [23, 24]. Eine Untersuchung zum Anteil von Frauen an klinischen Studien zur kardiovaskulären Prävention zeigte, dass dieser im Jahr 1970 bei 9% lag, und im Zeitverlauf bis zum Jahr 2006 auf erst 41% angestiegen war [25]. Möglicherweise hat dieser Fokus auf die Erkrankung bei Männern auch mit dazu beigetragen, dass Symptome, die eher bei Männern auftreten, häufiger als „typisch" klassifiziert werden als Symptome, die bei Frauen verbreiteter sind. Inzwischen wird das Problem dieses schwerwiegenden *sample selection bias* vermehrt adressiert, beispielsweise in Studien, die molekulare, psychologische und soziale Aspekte von biologischem Geschlecht (engl. *sex*) und sozialem Geschlecht (engl. *gender*) in Bezug auf kardiovaskuläre Erkrankungen, aber auch z.B. Krebs und Infektionserkrankungen untersuchen [26]. Studien aus den USA zeigen zudem, dass auch ethnische Minderheiten in klinischen Studien oft unterrepräsentiert sind [27]. Möglicherweise trägt dies mit zu der höheren kardiovaskulären Morbidität und Mortalität unter ethnischen Minderheiten in den USA bei, wobei trotzdem wahrscheinlich andere Faktoren (wie z.B. Exposition gegenüber Risikofaktoren, Zugang zu Gesundheitsversorgung, und Ungleichbehandlung in Gesundheitseinrichtungen) eine größere Rolle spielen [28]. Weitere Informationen zur Unter-Repräsentation von Frauen in der (medizinischen) Forschung und diesem eindrücklichen Selektionsbias finden sich in folgendem Buch: Criado Perez, C: Invisible Women. Random House UK Ltd. ISBN: 1784706280.

Systematische Messabweichung (engl. *measurement bias*)

In einer hypothetischen Studie soll untersucht werden, ob regelmäßige Kontrolluntersuchungen zur Optimierung der Therapie die Lebensqualität von PatientInnen mit chronisch obstruktiver Lungenerkrankung (COPD) verbessern. In einer RCT werden 1025 PatientInnen mit COPD nach dem Zufallsprinzip in eine von zwei Gruppen eingeteilt: eine Interventionsgruppe, die im festen Abstand von 3 Monaten Kontrolltermine bei ihrer HausärzIn erhielt, und eine Kontrollgruppe mit Standard-Versorgung (d.h. ohne feste Kontrolltermine). Die Studie hatte eine Dauer von zwei Jahren. Die Lebensqualität wurde mit einem standardisierten Fragebogen erhoben. In der Interventionsgruppe wurde der Fragebogen von den PatientInnen beim letzten Kontrolltermin zusammen mit einem medizinischen Fachangestellten (MFA) ausgefüllt, in der Kontrollgruppe wurde der Fragebogen den TeilnehmerInnen per Post zugeschickt.
In dieser Studie könnte es zu einer Ergebnis-Verzerrung durch systematische Messabweichung kommen: Auch wenn in beiden Gruppen derselbe Fragebogen verwendet wurde, ist es denkbar, dass das gemeinsame Ausfüllen mit einer MFA zu anderen Ergebnissen führt als das Ausfüllen alleine zu Hause: möglicherweise werden einzelne Fragen von den TeilnehmerInnen beim Ausfüllen alleine nicht richtig oder anders verstanden; möglicherweise wollen TeilnehmerInnen beim gemeinsamen Ausfüllen mit einer MFA nicht undankbar erscheinen, und geben deshalb positivere Antworten (dies wäre ein Beispiel für eine Verzerrung durch soziale Erwünschtheit, der in diesem Fall zu einer systematischen Messabweichung führen könnte).

Ausscheider-Bias (engl. *attrition bias*)
In einer hypothetischen Studie soll der Effekt einer zusätzlichen neoadjuvanten Chemotherapie bei fortgeschrittenem Ovarialkarzinom untersucht werden. In einer RCT werden 550 Patientinnen mit fortgeschrittenem Ovarialkarzinom nach dem Zufallsprinzip in zwei Gruppen eingeteilt: 272 PatientInnen erhielten vor ihrer OP und der adjuvanten Chemotherapie zusätzlich eine neoadjuvante Chemotherapie, und 278 Patientinnen erhielten die Standard-Therapie (OP plus adjuvante Chemotherapie). Der primäre Endpunkt war das Überleben nach drei Jahren. In die Analyse eingeschlossen wurden Daten von 209 Patientinnen aus der Interventionsgruppe und 265 Patientinnen aus der Kontrollgruppe.
In dieser Studie könnte es zu einer systematischen Verzerrung durch Ausscheider-Bias gekommen sein. In der Interventionsgruppe sind deutlich mehr Patientinnen vorzeitig aus der Studie ausgeschieden, und sind bei der Auswertung unberücksichtigt geblieben. Dies kann daran liegen, dass in der Interventionsgruppe mehr Patientinnen mit der Therapie unzufrieden waren oder verstorben sind ohne dass dies vom Studienteam registriert wurde, oder zu krank waren um die Studie fortzusetzen. Da die Daten von diesen Patientinnen bei der Auswertung nicht berücksichtigt wurden, kann sich hieraus eine Verzerrung der Ergebnisse ergeben haben.

Eine weitere mögliche Ursache systematischer Verzerrungen sind Interessenkonflikte (siehe Abschnitt 2.5). In einer Analyse von RCTs zur Wirksamkeit von Medikamenten, die in renommierten Fachzeitschriften erschienen sind, wurde untersucht, wie sich finanzielle Verbindungen zwischen der StudienleiterIn und dem Hersteller des untersuchten Medikaments auf die Ergebnisse der jeweiligen Studie auswirkte [29]. Es zeigte sich, dass in 68% aller Studien finanzielle Verbindungen zwischen der StudienleiterIn und dem Hersteller bestanden, und dass diese rund drei Mal häufiger positive Effekte berichteten als Studien, deren StudienleiterIn keine finanziellen Verbindungen zum Hersteller des Medikaments hatten (Odds Ratio 3,23, 95% KI: 1,7-6,1) [29].

Interessenkonflikte sind auch ein Problem bei Leitlinien und anderen Standardwerken, in denen diagnostische und therapeutische Standards definiert werden. So weisen z.B. knapp 70% der AutorInnen des *Diagnostic and Statistical Manual of Mental Disorders* (DSM 5) – einem wichtigen Standardwerk in der Psychiatrie – finanzielle Interessenkonflikte auf [30]. Die Initiative leitlinienwatch.de untersucht, wie transparent deutsche und europäische Leitlinien mit Interessenkonflikten umgehen, und bewertet Leitlinien mit diesem und weiteren Kriterien (www.leitlinienwatch.de).

Da finanzielle Interessenkonflikte Ergebnisse beeinflussen können, verlangen die meisten Fachzeitschriften mittlerweile, dass diese bei Veröffentlichungen offen gelegt werden. Ähnliche Regeln gelten für Vorträge auf vielen nationalen und internationalen Kongressen. Die Bundesvertretung der Medizinstudierenden in Deutschland e.V. (bvmd) und das Studierenden- und WissenschaftlerInnennetzwerk Universities Allied for Essential Medicines (UAEM) fordern, dass solche Regeln auch für Vorlesungen im Medizinstudium gelten sollten [31]. Eine 2019 veröffentlichte Studie konnte jedoch nur an zwei medizinischen Fakultäten in Deutschland (Aachen und Dresden) klar formulierte Regeln zum Umgang mit Interessenkonflikten identifizieren [32].

5.4. Störfaktoren (engl. *confounder*)

Viele wissenschaftliche Studien versuchen *Kausalzusammenhänge*, d.h. Zusammenhänge zwischen Ursache und Wirkung, aufzuklären. Das zuverlässigste Studiendesign für Fragen zur Wirksamkeit von Maßnahmen sind RCTs. Alternativ sind in manchen Fällen sog. quasi-experimentelle Studiendesigns möglich (siehe Abschnitt 3.1). In manchen Fällen sind jedoch sowohl RCTs als auch quasi-experimentelle Studiendesigns praktisch schwer durchführbar. Ein Beispiel sind die langfristigen gesundheitlichen Wirkungen langjähriger Ernährungsgewohnheiten: Es ist sehr schwierig und auch ethisch problematisch, sicherzustellen, dass StudienteilnehmerInnen über viele Jahre eine ihnen nach dem Zufallsprinzip zugeteilte Ernährungsweise befolgen. In diesen Fällen wird oft versucht, aus Beobachtungsstudien (Quer- und Längsschnittstudien sowie Fall-Kontroll-Studien) Schlüsse auf mögliche Kausalzusammenhänge zu ziehen.

Wenn aus Beobachtungstudien auf Kausalzusammenhänge geschlossen werden soll, können so genannte Störfaktoren (engl. *confounder*) als häufige Fehlerquelle auftreten. Dies sei anhand eines Beispiels erklärt: In einer Querschnittstudie zeigt sich, dass Menschen, die in der Nähe des Bodensees leben, statistisch signifikant häufiger an Alzheimer Demenz erkranken als die Durchschnittsbevölkerung. Bedeutet dies, dass vom Bodensee ein gefährlicher, noch unbekannter Einfluss ausgeht, der das Risiko für Alzheimer Demenz erhöht? Vermutlich eher nicht. Wahrscheinlicher ist, dass hier Störfaktoren eine Rolle spielen. Eine genauere Analyse zeigt, dass ältere Menschen häufig den Bodensee als Altersruhesitz wählen, und es am Ufer des Bodensees eine größere Zahl an Seniorenwohnheimen gibt. Ältere Menschen erkranken jedoch auch häufiger an Alzheimer Demenz. Alter ist somit ein Störfaktor, der den falschen Eindruck eines direkten Kausalzusammenhangs oder Scheinzusammenhangs zwischen Leben am Bodensee und dem Risiko für Alzheimer erzeugt (siehe Abbildung 15).

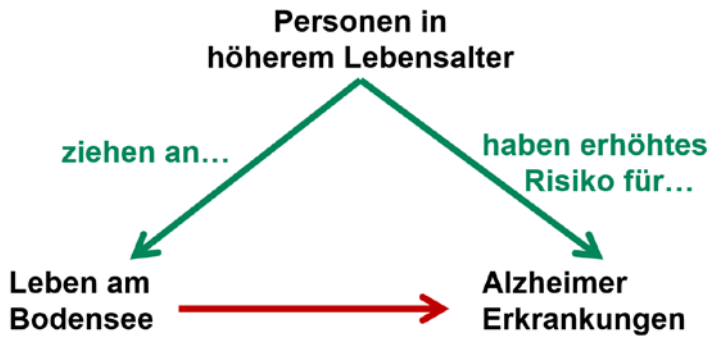


Abbildung 15: *Scheinzusammenhang zwischen Leben am Bodensee und dem Risiko für Alzheimer Demenz, erklärt durch Alter als Störfaktor.*

Ein komplexeres Beispiel wird in Abbildung 16 dargestellt: Die renommierte Fachzeitschrift *New England Journal of Medicine* publizierte einen Artikel, welcher den Zusammenhang zwischen dem Schokoladenkonsum pro Kopf und der Zahl der Nobelpreise pro 10 Millionen Einwohner untersucht [33]. Schokoladenkonsum und Nobelpreiszahl korrelieren miteinander – wobei dies wahrscheinlich nicht auf einen Kausalzusammenhang zwischen den beiden Größen hindeutet, sondern durch einen Störfaktor erklärbar ist: In Ländern mit hohen Einkommen wird mehr Schokolade verzehrt, und mehr Geld in Forschung investiert [33].

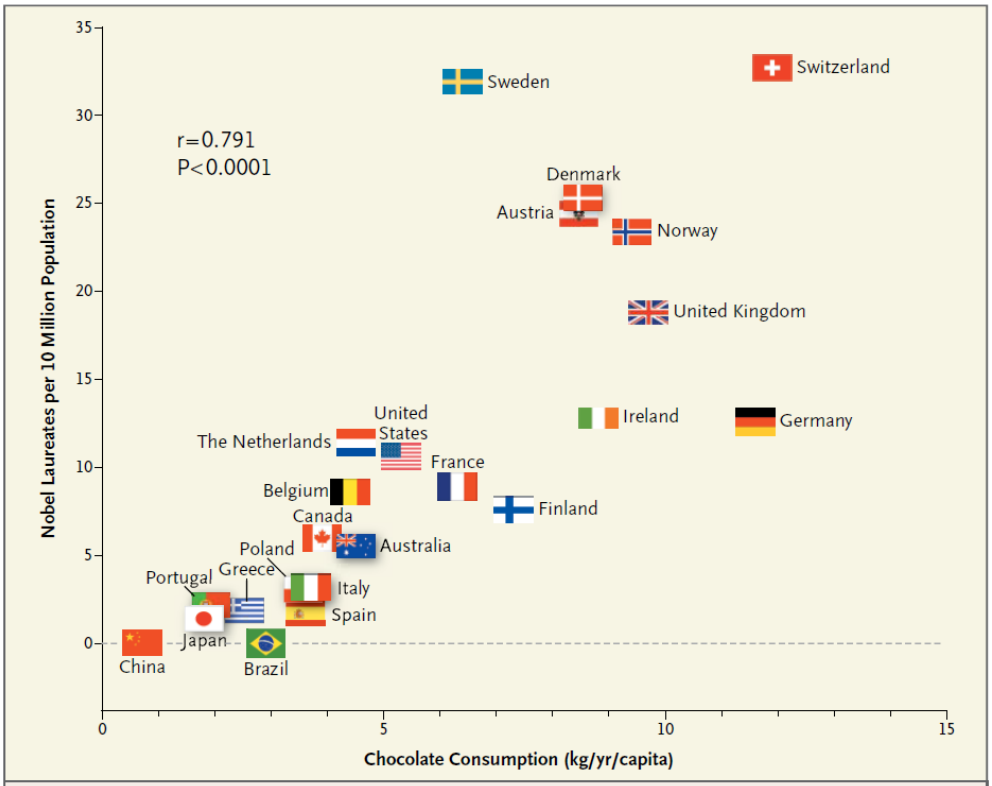


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Abbildung 16: *Statistischer Zusammenhang zwischen Nobelpreisen pro 10 Millionen Einwohnern und Schokoladenkonsum pro Kopf. Bildquelle: [33].*

5.5. Manipulation

Eine weitere mögliche Fehlerquelle in wissenschaftlichen Studien ist die bewusste Manipulation durch WissenschaftlerInnen. Das genaue Ausmaß ist unbekannt, doch weisen Befragungen darauf hin, dass bewusste Manipulation ein relevantes Problem darstellt. Eine Meta-Analyse von Befragungen von WissenschaftlerInnen zu dem Thema kam zu dem Ergebnis, das zwischen 0,9 – 4,5 % aller Befragten angaben, mindestens einmal Daten oder Ergebnisse fabriziert, gefälscht oder modifiziert zu haben [34]. In Umfragen, die nach dem Verhalten von Kollegen fragten, lagen diese Rate zwischen 9,9 und 19,7 % [34].

Immer wieder erzeugen einzelne große Wissenschaftsskandale aus verschiedenen Disziplinen mediale Aufmerksamkeit. Bei vielen Fällen sind es Whistleblower aus den Laboren und Arbeitsgruppen, welche auf Ungereimtheiten hinweisen. Unser Wissenschaftssystem basiert auf klar definierten Vorgehensweisen des wissenschaftlichen Arbeitens und auf dem Grundvertrauen, dass diese befolgt werden. So fallen geschickte Fälschungen von einzelnen Datenpunkten oder das freie Erfinden ganzer Datensätze anderen WissenschaftlerInnen beim kritischen Lesen der Publikation nicht notwendigerweise auf. Bei geschickten Fälschungen ist auch das Peer Review – die Kontrolle eines Manuskripts durch andere WissenschaftlerInnen vor der Veröffentlichung – oft kein ausreichender Schutzmechanismus [35]. Möglichst große Transparenz, ein detailliertes Berichten aller Methoden und Ergebnisse und der öffentliche Zugang zu relevanten Datensätzen (Open Science-Prinzipien) können dazu beitragen, das Grundvertrauen zu stärken und wissenschaftliches Fehlverhalten zu verhindern.

Kasten 7: Manipulation wissenschaftlicher Daten – der Fall Hermann/Brach

Fallbeispiel: Der Fall von Friedhelm Herrmann (ehemals Professur in Ulm) und Marion Brach (ehemals Professur in Lübeck) und deren Umfeld war einer der größten Skandale in der medizinischen Forschung in Deutschland. Den Krebsforschern wurde vorgeworfen Ergebnisse eigener Experimente gefälscht, sowie Ideen und Ergebnisse anderer Forscher in großem Umfang gestohlen zu haben. Eine Untersuchung der Deutschen Forschungsgemeinschaft (DFG) ergab, dass „die Professoren Herrmann und Brach über einen langen Zeitraum, mindestens von 1988 bis 1996, in ihren wissenschaftlichen Arbeiten Ergebnisse und Aussagen in erheblichem Umfang gefälscht haben“ [36] . Die Analyse von 347 Publikationen der beiden Professoren ergab, dass bei 94 Publikationen ein konkreter Verdacht auf Fälschung oder Hinweise auf Datenmanipulation gefunden wurden [36]. Die fälschungsbehafteten oder konkret fälschungsverdächtigen Publikationen entstanden an den Universitäten in Mainz, Freiburg, Berlin und Ulm [36] .

6. Fazit

Evidenzbasierte Medizin (EbM) bezeichnet das Fällen von medizinischen Entscheidungen unter Nutzung der jeweils besten verfügbaren wissenschaftlichen Erkenntnisse, klinisch-praktischer Erfahrung, und der Werte und Präferenzen der betroffenen PatientInnen. Systematik, Partizipation, Integration, Transparenz und ein reflektierter Umgang mit Interessenkonflikten sind zentrale Prinzipien der EbM. Zu den Instrumenten der EbM zählen Primärstudien, systematische Übersichtsarbeiten, Leitlinien, evidenzbasierte klinische Referenzwerke und Evidenz-Visualisierungen wie z.B. Faktenboxen, sowie Health Technology Assessments (HTAs). Die fünf Schritte der Umsetzung von EbM in der Praxis umfassen die Formulierung einer klaren Fragestellung (z.B. anhand des PICO-Schemas – Population, Intervention, Comparison und Outcome), die systematische Evidenzsuche, die

kritische Prüfung der Evidenz anhand der Kriterien interne Validität, externe Validität und klinische Relevanz, und die Anwendung der Evidenz und die Bewertung und Evaluation der Umsetzung. Faktoren, welche die interne Validität einer Studie beeinträchtigen können, sind Zufallsfehler, systematische Fehler (bias), Störfaktoren (confounder) sowie Manipulation. Eine Kenntnis der wichtigsten Fehlerquellen von Studien kann helfen, die Zuverlässigkeit von Evidenz zu beurteilen. Richtig verstanden sind die Prinzipien, Instrumente und Methoden der EbM eine wichtige Hilfestellung in der klinischen Praxis, und ein wesentlicher Beitrag zu einer effektiven Gesundheitsversorgung.

Weiterführende Ressourcen

Zur Wiederholung und Prüfungsvorbereitung:

- Inhalte von Amboss zur medizinischen Statistik: www.amboss.com/de/wissen/Übersicht_der_Inhalte_zur_medizinischen_Statistik
- Lehrbücher:
    - Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N: Lehrbuch Evidenzbasierte Medizin. Köln: Deutscher Ärzte-Verlag; 2007.
    - Razum O, Breckenkamp J, Brzoska P: Epidemiologie für Dummies. Wiley-VCH; 2017
- Artikel: Christopher J. Cates, Elizabeth Stovold, Emma J. Welsh: How to make sense of a Cochrane systematic review. Breathe 2014 10: 134-144; Online: https://breathe.ersjournals.com/content/10/2/134

Für Klinik und Forschung:

- Klinische Referenzwerke:
    - UpToDate: www.uptodate.com
    - Amboss: www.amboss.com/de/aerztinnen-aerzte
    - Deximed: https://deximed.de/
- Leitlinien:
    - Deutschland: www.awmf.org/leitlinien/leitlinien-suche.html
    - International: www.who.int/publications/guidelines/en
- Systematische Übersichtsarbeiten:
    - Cochrane Library: www.cochranelibrary.com
    - Campbell Collaboration: www.campbellcollaboration.org
    - Epistemonikos: www.epistemonikos.org/de
    - Evidence Aid: www.evidenceaid.org
- Weltweit größte gesundheitsrelevante Datenbank:
    - PubMed https://pubmed.ncbi.nlm.nih.gov/
    - Embase (insbes. für weitergehende Suchen relevant, z.B. für systematische Übersichtsarbeiten): Zugriff an der LMU über den E-Medien Katalog der Universitätsbibliothek mittels der Datenbanksuchplattform OVID: https://login.emedien.ub.uni-muenchen.de/login (Unter dem Link einloggen und nach dem Schlagwort „EMBASE“ suchen)
- Faktenboxen:
    - Harding Center: www.hardingcenter.de/de/projekte-und-kooperationen/faktenboxen
    - Arriba: https://arriba-hausarzt.de
- Verfahren zur Bewertung der Studienqualität und des Verzerrungsrisikos:

    - Critical-Appraisal-Checklisten: https://casp-uk.net/casp-tools-checklists/
    - Cochrane Risk of Bias Tool: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials
    - ROBINS-I Tool: https://methods.cochrane.org/bias/risk-bias-non-randomized-studies-interventions
- Reporting-Guidelines: www.equator-network.org

Literaturangaben

1. Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N: **Lehrbuch Evidenzbasierte Medizin**. Köln: Deutscher Ärzte-Verlag; 2007.
2. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS: **Evidence based medicine: what it is and what it isn't**. *BMJ* 1996, **312**(7023):71-72.
3. Djulbegovic B, Guyatt GH: **Progress in evidence-based medicine: a quarter century on**. *Lancet* 2017, **390**(10092):415-423.
4. **Bundesrahmenempfehlungen nach § 20d Abs. 3 SGB V. Erste weiterentwickelte Fassung vom 29. August 2018.** [www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/P/Praeventionsgesetz/BRE_Fassung_vom_29.08.2018.pdf]
5. Yamey G, Volmink J: **An Argument for Evidence-Based Policy-Making in Global Health**. In: *The Handbook of Global Health Policy.* edn. Edited by Brown WG, Yamey G, Wamala S: John Wiley & Sons, Ltd; 2014: 133-156.
6. Gerhardus A, Breckenkamp J, Razum O, Schmacke N, Wenzel H: **Evidence-based Public Health**. Bern: Huber; 2010.
7. Brownson RC, Fielding JE, Green LW: **Building Capacity for Evidence-Based Public Health: Reconciling the Pulls of Practice and the Push of Research**. *Annual Review of Public Health* 2018, **39**(1):27-53.
8. Greenstone G: **The History of Bloodletting**. *BCMJ* 2010, **51**(1):12-14.
9. Rehfuess E, Stratil J, Scheel I, Portel A, Norris S, Baltussen R: **The WHO-INTEGRATE evidence to decision framework version 1.0: Integrating WHO norms and values and a complexity perspective**. *BMJ global health* 2018.
10. Vineis P, Saracci R: **Conflicts of interest matter and awareness is needed**. *J Epidemiol Community Health* 2015, **69**(10):1018-1020.
11. The WOMAN Collaborators: **Effect of early tranexamic acid administration on mortality, hysterectomy, and other morbidities in women with post-partum haemorrhage (WOMAN): an international, randomised, double-blind, placebo-controlled trial**. *Lancet* 2017, **389**(10084):2105-2116.
12. Craig P, Katikireddi SV, Leyland A, Popham F: **Natural Experiments: An Overview of Methods, Approaches, and Contributions to Public Health Intervention Research**. *Annual Review of Public Health* 2017, **38**(1):39-56.
13. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, Ogilvie D, Petticrew M, Reeves B, Sutton M *et al*: **Using natural experiments to evaluate population health interventions: new Medical Research Council guidance**. *Journal of Epidemiology and Community Health* 2012, **66**(12):1182-1186.
14. South EC, Hohl BC, Kondo MC, MacDonald JM, Branas CC: **Effect of Greening Vacant Land on Mental Health of Community-Dwelling Adults: A Cluster Randomized Trial**. *JAMA Network Open* 2018, **1**(3):e180298-e180298.
15. Grimes DA, Schulz KF: **An overview of clinical research: the lay of the land**. *The Lancet* 2002, **359**(9300):57-61.
16. Nabarro DN, Tayler EM: **The "Roll Back Malaria" Campaign**. *Science* 1998, **280**(5372):2067.
17. **Handbook for Guideline Development** [https://apps.who.int/iris/bitstream/handle/10665/75146/9789241548441_eng.pdf]
18. **Trends in maternal mortality 2000 to 2017** [https://apps.who.int/iris/bitstream/handle/10665/327596/WHO-RHR-19.23-eng.pdf?ua=1]
19. **WHO Recommendations on Prevention and Treatment of Postpartum Haemorrhage and the WOMAN Trial** [www.who.int/reproductivehealth/topics/maternal_perinatal/pph-woman-trial/en/]
20. **Das AWMF-Regelwerk Leitlinien** [www.awmf.org/leitlinien/awmf-regelwerk.html]
21. **CASP Checklists** [https://casp-uk.net/casp-tools-checklists/]

22. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions**. *BMJ* 2016, **355**.
23. Carey JL, Nader N, Chai PR, Carreiro S, Griswold MK, Boyle KL: **Drugs and Medical Devices: Adverse Events and the Impact on Women’s Health**. *Clinical Therapeutics* 2017, **39**(1):10-22.
24. Bots SH, Groepenhoff F, Eikendal ALM, Tannenbaum C, Rochon PA, Regitz-Zagrosek V, Miller VM, Day D, Asselbergs FW, den Ruijter HM: **Adverse Drug Reactions to Guideline-Recommended Heart Failure Drugs in Women: A Systematic Review of the Literature**. *JACC Heart Fail* 2019, **7**(3):258-266.
25. Melloni C, Berger JS, Wang TY, Gunes F, Stebbins A, Pieper KS, Rowena JD, Douglas PS, Mark DB, Newby KL: **Representation of Women in Randomized Clinical Trials of Cardiovascular Disease Prevention**. *Circulation: Cardiovascular Quality and Outcomes* 2010, **3**(2):135-142.
26. Gemmati D, Varani K, Bramanti B, Piva R, Bonaccorsi G, Trentini A, Manfrinato MC, Carè A, Bellini T: **"Bridging the Gap" Everything that Could Have Been Avoided If We Had Applied Gender Medicine, Pharmacogenetics and Personalized Medicine in the Gender-Omics and Sex-Omics Era**. *Int J Mol Sci* 2020, **21**(1).
27. Chen A, Wright H, Itana H, Elahi M, Igun A, Soon G, Pariser AR, Fadiran EO: **Representation of Women and Minorities in Clinical Trials for New Molecular Entities and Original Therapeutic Biologics Approved by FDA CDER from 2013 to 2015**. *J Womens Health (Larchmt)* 2018, **27**(4):418-429.
28. Rao SV, Kaul P, Newby LK, Lincoff AM, Hochman J, Harrington RA, Mark DB, Peterson ED: **Poverty, process of care, and outcome in acute coronary syndromes**. *J Am Coll Cardiol* 2003, **41**(11):1948-1954.
29. Ahn R, Woodbridge A, Abraham A, Saba S, Korenstein D, Madden E, Boscardin WJ, Keyhani S: **Financial ties of principal investigators and randomized controlled trial outcomes: cross sectional study**. *BMJ* 2017, **356**:i6770.
30. Cosgrove L, Krimsky S: **A Comparison of DSM-IV and DSM-5 Panel Members' Financial Associations with Industry: A Pernicious Problem Persists**. *PLoS medicine* 2012, **9**(3):e1001190.
31. **Medizinstudierende fordern Regeln zum Umgang mit der Pharmaindustrie im Studium** [www.bvmd.de/fileadmin/redaktion/Pressemitteilungen/2019-23-10_Interessenkonflikte.pdf]
32. **Conflict of Interest Policies at German medical schools - A long way to go** [www.biorxiv.org/content/10.1101/809723v1]
33. Messerli FH: **Chocolate Consumption, Cognitive Function, and Nobel Laureates**. *New England Journal of Medicine* 2012, **367**(16):1562-1564.
34. Fanelli D: **How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data**. *PLoS One* 2009, **4**(5):e5738.
35. Gross C: **Scientific Misconduct**. *Annu Rev Psychol* 2016, **67**:693-711.
36. **Task Force legt Abschlußbericht vor: Unstimmigkeiten auch im Umfeld von Friedhelm Herrmann** [www.dfg.de/service/presse/pressemitteilungen/2000/pressemitteilung_nr_26/index.html]