

Qualität der Evidenz und Stärke von Empfehlungen für medizinische Entscheidungen

Yngve Falck-Ytter, Gerd Antes, Andrew Oxman, Gordon Guyatt und Holger Schünemann für die GRADE Working Group.

Einleitung

Die Bewertung medizinischer Evidenz und die daraus resultierenden Empfehlungen sind komplex. Betrachtet man zum Beispiel in der Behandlung einer mittelschweren Depression die Wahl zwischen einem selektiven Serotonin-Wiederaufnahmehemmer (SSRIs) oder trizyklischen Antidepressiva (TZAs), muss der Kliniker eine Fülle von Entscheidungen treffen: Welche Outcomes (Zielgrößen) sind entscheidend? Welche Art von Evidenz ist für diesen Outcome notwendig? Wie muss die Qualität der zugrunde liegenden Studien im Einzelnen bewertet werden? Darauf aufbauend muss der Kliniker entscheiden, ob SSRIs im Vergleich zu TZAs tatsächlich in der Praxis mehr nutzen als schaden. Aufgrund des allgegenwärtigen Kostendrucks im Gesundheitswesen muss immer auch eine Kosten-Nutzen Abwägung erfolgen: d.h. ist der mögliche Vorteil von SSRIs im Vergleich zu TZA auch wirtschaftlich vertretbar?

Ohne Hilfestellung ist es für den einzelnen Arzt und Patient praktisch unmöglich, bei jeder einzelnen klinischen Entscheidung diese Bewertungen vorzunehmen. Deshalb verwenden Kliniker wie auch Patienten zu ihrer Unterstützung in zunehmendem Maße klinische Leitlinien. Diese setzen sich aus Empfehlungen zusammen, welche systematisch von Gremien entwickelt wurden, die Zugriff auf die vorhandene Evidenz haben, das klinische Problem und die Forschungsmethoden verstehen sowie genügend Zeit zur kritischen Auseinandersetzung haben.

Den Nutzern von systematisch erstellten Leitlinien sollte deutlich werden, wie sehr sie der zugrunde liegenden Evidenz und den daraus resultierenden Empfehlungen vertrauen können. Im Folgenden möchten wir Merkmale beschreiben, auf die sich unser Vertrauen gründen kann. Auch soll eine systematische Herangehensweise dargestellt werden, welche die komplexen Bewertungen beschreibt, die in klinischen Empfehlungen von Leitlinien oder anderen Gesundheitsempfehlungen implizit oder explizit einfließen. Der Einfachheit halber werden wir nicht alle Nuancen darstellen oder detaillierte Anleitungen geben, die Leitliniengremien für die Durchführung unseres Ansatzes benötigen. Wer hierzu weitere Ausführungen benötigt, kann diese unter GradeWorkingGroup.org finden.

Ein systematischer und expliziter Ansatz zur Bewertung der Qualität der Evidenz und die daraus resultierenden Empfehlungen können Fehler vermeiden helfen, verbessern die kritische Einschätzung von Empfehlungen und fördern die Kommunikation dieser Informationen. Seit den siebziger Jahren haben eine zunehmende Anzahl von Organisationen Evidenzhierarchien verwendet, um die Qualität der Evidenz und die daraus resultierenden Empfehlungen zu klassifizieren.(1-28) Leider benützen unterschiedliche Organisationen jedoch unterschiedliche Systeme zur Beurteilung

der Qualität der Evidenz und der Stärke von Empfehlungen. Die gleiche Evidenz und Empfehlung könnte als II-2/B, C+/1 oder „strong evidence/strongly recommended“ (hochgradige Evidenz/stark empfohlen) bezeichnet werden, je nachdem welches System benutzt wird. Diese verschiedenen Bezeichnungen sind verwirrend und behindern eine effektive Kommunikation.

Die GRADE Working Group begann als eine informelle Arbeitsgruppe von Wissenschaftlern, die die Nachteile der bisherigen Systeme bzw. Hierarchien angehen wollten. Tabelle 1 fasst diese Nachteile zusammen und beschreibt die Verbesserungen, die durch das GRADE System zu erwarten sind. Das GRADE System ermöglicht konsistentere Bewertungen, und die Kommunikation solcher Bewertungen hilft, besser informiert Entscheidungen im Gesundheitssystem zu fällen. In Kasten 1 werden die Schritte erläutert, die zur Entwicklung und Implementierung von Leitlinien notwendig sind: von der Priorisierung von Gesundheitsfragen bis hin zur Evaluation der Leitlinienimplementierung. Im Folgenden beschreiben wir, wie die Qualität der Evidenz bewertet, und die daraus resultierenden Empfehlungen formuliert werden können.

Definitionen

Wir werden die folgenden Definitionen verwenden: die Qualität der Evidenz ist ein Gradmesser für die Zuversicht, dass ein ermittelter Effekt korrekt ist. Die Stärke der Empfehlung ist ein Gradmesser für die Zuversicht, dass das Umsetzen der Empfehlung mehr nutzt als schadet.

Die Einschätzung der Qualität der Evidenz erfordert die Bewertung der Validität der Ergebnisse für wichtige Outcomes in einzelnen Studien. Um diese Bewertung vorzunehmen, sollten explizite Kriterien verwendet werden.(26, 29-32) Unser Ansatz setzt diese Bewertungen voraus und erfordert die folgende stufenweise Beurteilung:

- Qualität der Evidenz bezüglich jedes wichtigen Outcomes aller verfügbaren Studien
- Welche Outcomes entscheidend sind
- Abschließende Qualitätsbeurteilung der Evidenz für alle entscheidenden Outcomes
- Abwägung von Nutzen und Schaden
- Stärke der Empfehlung

Alle diese Bewertungen erfordern eine klar definierte Fragestellung und das Einbeziehen aller Outcomes, die für die Betroffenen voraussichtlich wichtig sind. Die Fragestellung sollte die zu vergleichenden Optionen (z.B. SSRIs im Vergleich zu TZA), die Zielgruppe (z.B. Erwachsene mit mittelschwerer Depression) sowie die Praxissituation (z.B. allgemeinärztlich Praktizierende in Deutschland) beinhalten.

Die Bewertung der Qualität der Evidenz für jeden wichtigen Outcome

Eine systematische Übersichtsarbeit der vorhandenen Evidenz sollte als Grundlage dienen. Zur Bewertung werden vier zentrale Kriterien herangezogen: Studiendesign, Studienqualität, Konsistenz und Direktheit der Evidenz.

Studiendesign

Das Studiendesign beschreibt die grundlegende Methode der Studie und kann zunächst in Beobachtungsstudien und randomisierte Studien unterteilt werden. Logische Argumente wie auch empirische Evidenz machen solch eine Unterteilung sinnvoll.(33-36) Obwohl Beobachtungsstudien häufig ähnliche Ergebnisse wie randomisierte Studien zeigen, trifft das nicht immer zu. Ein Beispiel für eine ausgeprägte Diskrepanz zwischen Beobachtungsstudien und randomisierten

Studien stellt die Hormonersatztherapie dar: Große Beobachtungsstudien zeigten in der Vergangenheit ein deutlich verringertes Risiko einer koronaren Herzkrankung, doch darauf folgende randomisierte Studien konnten dies nicht bestätigen und zeigten sogar eine Zunahme.(37, 38) Leider ist es unmöglich vorauszusehen, ob Beobachtungsstudien später durch randomisierte Studien bestätigt werden können. Sind Ergebnisse aus qualitativ hochwertigen randomisierten Studien vorhanden, würden wohl nur wenige dafür plädieren, Empfehlungen weiterhin aus nicht-randomisierten Studien mit widersprüchlichen Ergebnissen abzuleiten.

Andererseits sind randomisierte Studien nicht immer durchführbar und in manchen Fällen können Beobachtungsstudien besser geeignet sein, wie das z.B. bei Studien zu seltenen Nebenwirkungen grundsätzlich der Fall ist. Darüber hinaus können Ergebnisse aus randomisierten Studien nicht immer anwendbar sein – z.B. wenn im Vergleich zur Patientenpopulation die Studienteilnehmer extrem selektiert und motiviert waren. Deshalb ist es von besonderer Wichtigkeit, die Studienqualität, die Konsistenz der Ergebnisse unterschiedlicher Studien, sowie die Direktheit der Evidenz (in Bezug auf Patientenkollektive, Interventionen und Outcomes, siehe unten) zusätzlich zur Angemessenheit des Studiendesigns zu bewerten. So können z.B. gut geplante Fallserien qualitativ hochgradige Evidenz darstellen, wenn es um die Erfassung der intraoperativen Mortalität oder der Perforationsrate bei Koloskopien geht, was direktere Relevanz hat als vergleichbare Daten aus randomisierten Studien. Gleichmaßen können Kohortenstudien beim Mammographie Screening hochgradige Evidenz darstellen, wie z.B. die Rate an zusätzlicher Bildgebung (recall rate) oder die durch falsch-positiven Screeningbefunde verursachte Anzahl an weiteren Untersuchungen (Biopsie Rate).

Kasten 1: Schrittweiser Prozess der Leitlinienerstellung

Die ersten Schritte

1. *Den Arbeitsprozess einrichten* – Z.B. Priorisierung der Fragestellungen, Zusammenstellung eines Leitlinien-Gremiums, Offenlegen der Interessenskonflikte sowie Festlegen des Vorgehens

Vorbereitende Schritte

2. *Systematische Übersichtsarbeit* – Zunächst muss die beste und verfügbare Evidenz für alle wichtigen Outcomes identifiziert, kritisch beurteilt oder durch eine systematische Übersichtsarbeit zusammengefasst werden
3. *Evidenz-Tabellen mit allen wichtigen Outcomes erstellen* – Evidenz-Tabellen werden für jede Subpopulation oder Risiko-Gruppe auf Grundlage der systematischen Übersichtsarbeit aufgestellt und sollten jeweils eine Tabelle „Qualitätsbewertung“ und eine Tabelle „zusammenfassende Ergebnisse“ beinhalten

Bewertung der Qualität der Evidenz und die Stärke der Empfehlung

4. *Qualität der Evidenz für jedes Outcome* – Die Bewertung beruht auf den Daten der Evidenz-Tabelle unter Verwendung der Kriterien aus Tabelle 2
5. *Relative Wichtigkeit der Outcomes* – Nur wichtige Outcomes sollten in Evidenz-Tabellen enthalten sein. Die aufgelisteten Outcomes werden als „entscheidend“ oder „wichtig“ (aber nicht entscheidend) klassifiziert
6. *Abschließende Qualitätsbewertung der Evidenz* – Die abschließende Qualitätsbewertung der Evidenz sollte bezüglich aller entscheidenden Outcomes erfolgen und verwendet die niedrigste ermittelte Qualitätsbewertung entscheidender Outcomes als Ausgangspunkt
7. *Abwägen von Nutzen und Schaden* – Das Abwägen von Nutzen und Schaden sollte abschließend in „Netto-Nutzen“, „Abwägung notwendig“, „Abwägung unsicher“ und „kein Netto-Nutzen“ eingeteilt werden
8. *Kosten-Nutzen-Abwägung* – Ist der inkrementelle Gesundheitsnutzen auch wirtschaftlich vertretbar? Weil Ressourcen immer begrenzt sind, ist es wichtig, auch die Kosten (Nutzung der Ressourcen) bei der Formulierung von Empfehlungen einzubeziehen
9. *Stärke der Empfehlung* – Empfehlungen sollten so formuliert werden, dass ihre Stärke ersichtlich wird, d.h. ein Gradmesser für die Zuversicht sein, dass das Umsetzen der Empfehlung mehr nutzt als schadet

Anschließende Schritte

10. *Implementierung und Evaluation* – Verwendung von effektiven Implementierungsstrategien, wie z.B. die Berücksichtigung der Hürden für eine erfolgreiche Umsetzung, die Evaluation der Implementierung, sowie die regelmäßige Aktualisierung der Evidenz

Studienqualität

Mit der Studienqualität sind die im Detail verwendeten Methoden sowie die Art der Studiendurchführung gemeint. Für die Bewertung der Studienqualität sollten angemessene Kriterien bezüglich jedes wichtigen Outcomes verwendet werden.(26, 29-32) Bei randomisierten Studien können dies z.B. eine Gewährleistung der Geheimhaltung der Behandlungsfolge (allocation concealment), Verblindung und Vollständigkeit der Nachbeobachtung (follow-up) sein. Wenn die Qualität der Studie wegen Mängeln herabgestuft wird, so sollten die Gründe dafür explizit beschrieben werden. Zum Beispiel könnte man argumentieren, dass bei einer schmerzlindernden Intervention ein Unterlassen der Verblindung von Patienten und Ärzten die Qualität der Studie verringert und dies damit eine schwerwiegende Limitierung darstellt.

Konsistenz der Evidenz (Consistency)

Die Konsistenz der Evidenz bezieht sich auf die Ähnlichkeit der Ergebnisse verschiedener Studien in Bezug auf die Richtung und Größe des Effekts. Wenn eine wichtige und nicht erklärbare Inkonsistenz der Ergebnisse beobachtet wird, verringert sich unser Vertrauen in den ermittelten Effekt des entsprechenden Outcomes. Unterschiede in der Richtung des Effekts, die Größenordnung der Unterschiede der Ergebnisse und die Signifikanz der Unterschiede bedingen die (unausweichlich etwas willkürliche) Entscheidung, ob eine

Tabelle 1: GRADE im Vergleich zu anderen Systemen

Faktor	Andere Systeme	GRADE	Vorteil des GRADE Systems*
Definitionen	Implizite Definitionen der Qualitätsbeurteilung der Evidenz und Stärke der Empfehlung	Explizite Definitionen	Verdeutlicht die Qualitätsstufen und was bei diesen Bewertungen berücksichtigt werden muss
Bewertungen	Implizite Bewertungen	Schrittweise, explizite Bewertungen bezüglich der Wichtigkeit von Outcomes, der Qualität der Evidenz für jeden wichtigen Outcome, der abschließende zusammenfassende Qualitätsbewertung der Evidenz, der Abwägung von Nutzen und Schaden und der Wertschätzung des inkrementelle Nutzens	Beschreibt die einzelnen Bewertungsschritte und verringert die Wahrscheinlichkeit von Fehlern und Verzerrungen, die bei impliziten Bewertungsschritten auftauchen können
Zentrale Komponenten der Qualität der Evidenz	Wird nicht für jeden einzelnen wichtigen Outcome erwogen. Bewertungen der Qualität der Evidenz werden meist nur auf das Studiendesign bezogen.	Systematische und explizite Einbeziehung von Studiendesign, Studienqualität, Konsistenz und Direktheit der Evidenz wenn Qualitätsbewertungen der Evidenz vorgenommen werden	Stellt sicher, dass diese Faktoren angemessen berücksichtigt werden
Andere Faktoren, die die Qualität der Evidenz beeinflussen können	Nicht explizit berücksichtigt	Explizites Einbeziehen einer unpräzisen oder spärlichen Datenlage, Publikationsbias, Stärke der Assoziation, Vorliegen einer Dosis-Wirkungs-Beziehung und Vorhandensein eines plausiblen Confoundings	Stellt sicher, dass andere Faktoren berücksichtigt werden
Abschließende, zusammenfassende Qualität der Evidenz	Basiert implizit auf der Qualität der Evidenz des Nutzens	Basiert auf der geringsten Qualitätsstufe von Outcomes, die als entscheidend eingestuft wurden	Verringert die Wahrscheinlichkeit einer Fehleinschätzung der abschließenden Qualitätsbewertung der Evidenz, wenn die für einen entscheidenden Outcome notwendige Evidenz fehlt
Relative Wichtigkeit von Outcomes	Implizit erwogen	Explizite Bewertung, welche Outcomes entscheidend, welche wichtig (aber nicht entscheidend) und welche nicht wichtig sind und somit ignoriert werden können	Stellt sicher, dass bei der Bewertung der Qualität der Evidenz und Stärke der Empfehlung jeder Outcome entsprechend angemessen berücksichtigt wird
Abwägung zwischen Schaden und Nutzen für die Gesundheit	Nicht explizit erwogen	Explizite Betrachtung einer Abwägung von wichtigem Nutzen und Schaden mit zugrunde liegender Evidenz, der Möglichkeit einer Umsetzung der Evidenz in eine spezifische Praxissituation und der Gewissheit bezüglich des Grundrisikos	Verdeutlicht und verbessert die Transparenz der Bewertung des Schadens und Nutzens
Ob der inkrementelle Nutzen für die Gesundheit auch wirtschaftlich zu vertreten ist	Nicht explizit erwogen	Explizit einbezogen, nachdem zunächst ein gesundheitlicher Netto-Nutzen ermittelt wurde	Stellt eine transparente Einschätzung des Werts vom gesundheitlichen Netto-Nutzen sicher
Zusammenfassungen der Evidenz und Ergebnisse	Inkonsistente Präsentation	Konsistente GRADE Evidenz-Tabellen, welche die Tabellen der Qualitätsbewertung und die der zusammengefassten Ergebnisse beinhalten	Stellt sicher, dass alle Mitglieder des Gremiums ihren Bewertungen die gleichen Information zugrunde legen und dass diese Information anderen zur Verfügung steht
Verbreitung	Selten von mehr als einer Organisation verwendet und nur, wenn überhaupt, geringe empirische Evaluation	Internationale Zusammenarbeit vieler verschiedener Organisationen in der Entwicklung und Evaluation	Auf die bisherige Erfahrungen gründend ein System zu ermöglichen, dass einleuchtend, zuverlässig und breit anwendbar ist

* Obwohl manche Systeme einige dieser Vorteile aufweisen, so sind bei den meisten Systeme keiner dieser Vorteile vorhanden.

ernstzunehmende Inkonsistenz besteht. Separate Abschätzungen der Größe des Effekts in verschiedenen Patienten-Untergruppen sollten erfolgen, wenn gute Gründe für die bestehende Inkonsistenz identifiziert wurden. So wäre es z.B. sinnvoll, die beobachteten Effektunterschiede bei der operativen Behandlung der Karotisstenose so aufzuarbeiten, dass separate Effektgrößen für Patientengruppen mit geringen oder hochgradigen Stenosen ermittelt werden.

Direktheit der Evidenz (Directness)

Die Direktheit der Evidenz beschreibt, inwieweit die Studienpopulation, die Interventionen und die Outcomes den Patienten in der Praxis, der Behandlung und dem zu erzielenden Outcome ähnlich sind. Zum Beispiel können Bedenken bezüglich der Direktheit entstehen, wenn die Zielgruppe älter oder kränker ist, oder eine ausgeprägtere Komorbidität aufweist als die Patienten in den Studien.(39) Um zu entscheiden, ob diesbezüglich eine bedeutende Unsicherheit besteht, müssen wir nach guten Argumenten dafür suchen, dass dieses Merkmal die Größenordnung des Effekts

beeinflussen könnte. Weil viele Interventionen mehr oder weniger den gleichen relativen Effekt auf unterschiedliche Patientengruppen aufweisen, sollten wir bei der Entscheidung, ob die Evidenz direkt ist, nicht übermäßig enge Kriterien anlegen. Für manche Interventionen (z.B. Verhaltenstherapie, bei der kulturelle Unterschiede von Wichtigkeit sind) könnten stringenter Kriterien angemessen sein.

In gleicher Weise können Bedenken über die Direktheit entstehen, wenn Medikamente bewertet werden sollen, für die nur Studien mit wirkungsähnlichen Medikamenten der gleichen Substanzklasse existieren. Ähnliche Situationen können bei anderen Interventionen entstehen: kann man eine im Vergleich zur Studienintervention weniger intensive Beratungsintervention oder eine abweichendes operatives Verfahren sinnvoll verallgemeinern? Solche Bewertungen können schwierig sein,(40) und es ist wichtig, dass die Schlussfolgerungen, die in solchen Situationen gezogen, ausreichend begründet werden.

Kasten 2: Kriterien für die Qualitätszuordnung der Evidenz

Spektrum: Hohe, mittlere, geringe, sehr geringe Qualität

Art der Evidenz

Randomisierte Studien = hohe Qualität
Beobachtungsstudien = geringe Qualität
Andere Evidenz = sehr geringe Qualität

Qualitätsbewertung herunterstufen:

- Schwerwiegende (- 1) oder sehr schwerwiegende (- 2) Limitierung der Studienqualität
- Wichtige Inkonsistenz der Ergebnisse (- 1)
- Ungewissheit (- 1) oder ausgeprägte Ungewissheit (- 2) bezüglich der Direktheit der Evidenz
- Unpräzise oder spärliche Datenlage (- 1)
- Hohes Risiko eines Publikationsbias (- 1)

Qualitätsbewertung heraufstufen:

- Vorhandensein einer starken Assoziation – ein signifikantes relatives Risiko von > 2 ($< 0,5$), wenn mehr als zwei Beobachtungsstudien ohne plausible Confounder und mit konsistenten Ergebnissen vorliegen (+1)(46)
- Vorhandensein einer sehr starken Assoziation – ein signifikantes relatives Risiko von > 5 ($< 0,2$), wenn Beobachtungsstudien mit direkter Evidenz und keine Bedenken bezüglich der Validität vorliegen (+2)(46)
- Vorhandensein einer Dosis-Wirkungs-Beziehung (+1)
- Alle verbleibenden, plausiblen Confounder haben den beobachteten Effekt bereits reduziert (+1)

Andererseits liefern Studien, die Surrogat-Outcomes betrachten, im Allgemeinen weniger direkte Evidenz als Studien, die patientenrelevante Outcomes verwenden. Bei Surrogat-Outcomes ist daher Vorsicht geboten, und es müssen viel stringenter Kriterien für die Direktheit der Evidenz angewandt werden. Als Beispiele indirekter, auf Surrogat-Outcomes basierender Evidenz aus Studien, die später durch Studien mit patientenrelevanten Outcomes als irreführend erkannt wurden, sind zu nennen: Suppression von kardialen Arrhythmien bei Patienten nach Myokardinfarkt als Surrogat für Mortalität (41), Veränderungen von Lipoproteinen als Surrogat für die koronare Herzkrankheit (37) und Knochendichte von Frauen in der Menopause als Surrogat für die Verringerung von Knochenbrüchen. (42)

Die Genauigkeit eines diagnostischen Tests ist ebenfalls ein Surrogat für wichtige Outcomes, die durch die Genauigkeit einer Diagnose beeinflusst werden, wie z.B. Verbesserungen der Gesundheit durch angemessene Behandlung und geringere Risiken durch eine Reduktion von falsch-positiven Ergebnissen. Aufgrund des unterschiedlichen Studiendesigns müssen bei der Bewertung von Diagnosestudien andere Kriterien Anwendung finden. Jedoch basiert die Einschätzung der Direktheit der Evidenz auf unser Vertrauen in die richtige (richtig-positiv oder richtig-negativ) oder falsche (falsch-positiv oder falsch-negativ) Klassifikation und den daraus resultierenden entscheidenden Konsequenzen. Zum Beispiel gibt es übereinstimmende Evidenz aus gut durchgeführten Studien, dass ein Helikal-CT (ohne Kontrastgabe) bei Verdacht auf Urolithiasis weniger falsch-negative Ergebnisse aufweist als eine intravenöse Pyelographie.(43) Dennoch besteht Unklarheit darüber, inwieweit dies einen positiven Einfluss auf den Verlauf der Erkrankung hat.(44) Daher könnte

für Empfehlungen die Qualität solcher Evidenz als gering betrachtet werden.

Eine weitere Situation von indirekter Evidenz ist gegeben, wenn kein direkter Vergleich von Interventionen vorliegt und Ergebnisse aus verschiedenen Studien verglichen werden müssen. Dies wäre z.B. der Fall, wenn es nur Studien gäbe, die SSRI mit Placebo und TZA mit Placebo und nicht direkt SSRI gegen TZA verglichen hätten. Indirekte Vergleiche erhöhen immer die Unsicherheit gegenüber direkten Vergleichen aufgrund der vielen anderen Faktoren, die das Ergebnis von Studien beeinflussen können.(45)

Zusammenfassen der vier Komponenten

Die Qualität der Evidenz für jeden wichtigen Outcome kann aufgrund der oben genannten Elemente ermittelt werden: Studiendesign, Studienqualität, Konsistenz und Direktheit der Evidenz. Unser Ansatz teilt die Evidenz zuerst in die zwei Kategorien des Studiendesigns ein: randomisierte Studien und Beobachtungsstudien (z.B. Kohortenstudien, Fall-Kontrollstudien, Zeitreihenstudien, kontrollierte Vorher-Nachher-Studien). Danach beurteilen wir, ob die Studien qualitativ schwerwiegende Limitierungen aufweisen, eine wichtige Inkonsistenz der Ergebnisse besteht, oder Unsicherheit bezüglich der Direktheit der Evidenz existiert (Kasten 2). Wir empfehlen die folgende Einteilung für die abschließenden Qualitätsbeurteilung der Evidenz:

Hohe Qualität = Es ist sehr unwahrscheinlich, dass zukünftige Forschungsergebnisse unsere Einschätzung des Effekts verändern

Mittlere Qualität = Es ist wahrscheinlich, dass zukünftige Forschungsergebnisse einen wichtigen Einfluss auf unsere Einschätzung des Effekts haben werden und sich der Effekt möglicherweise verändert

Geringe Qualität = Es ist sehr wahrscheinlich, dass zukünftige Forschungsergebnisse einen entscheidenden Einfluss auf unsere Einschätzung des Effekts haben werden und sich der Effekt wahrscheinlich verändert

Sehr geringe Qualität = Jegliche Einschätzung des Effekts ist sehr unsicher

Limitierungen in der Studienqualität, wichtige Inkonsistenz der Ergebnisse oder auch Bedenken, inwieweit die Evidenz vergleichbar (direkt) ist, verringern die Qualitätsstufe. Zum Beispiel, wenn alle vorhandenen Studien schwerwiegende Limitierungen aufweisen, fällt das Niveau um eine Stufe. Beim Vorhandensein von sehr schwerwiegenden Limitierungen fällt das Niveau um zwei Stufen. Ausgeprägt fehlerhafte Studien können ausgeschlossen werden.

Zusätzliche Faktoren, die die Qualität der Evidenz verringern, sind eine unpräzise oder spärliche Datenlage (Kasten 3) und ein hohes Risiko für Publikationsbias. Andererseits gibt es auch Faktoren, die die Einstufung der Qualität der Evidenz explizit erhöhen können:

- Eine sehr starke Assoziation (z.B. eine im Vergleich zu Serotonin-Wiederaufnahmehemmer (SSRIs) fünfzigfach erhöhtes Risiko durch eine Überdosierung an trizyklischen Antidepressiva (TZAs) zu versterben (siehe Tabelle 2) oder starke Assoziation (z.B. ein dreifach erhöhtes Kopf-

verletzungsrisiko bei Fahrradfahrern ohne Helm, im Vergleich zu Helmträgern (47))

- Vorhandensein einer Dosis-Wirkungs-Beziehung
- Das Vorhandensein aller verbleibenden, plausiblen Confounder hat den beobachteten Effekt bereits reduziert. (Zum Beispiel: plausible Faktoren, die in Studien zum Vergleich von Mortalitätsraten von profit- und nicht-profitorientierten Krankenhäusern nicht zur Adjustierung verwendet wurden, hätten den beobachteten Effekt bereits reduziert.(48) Dadurch gewinnt die Evidenz, dass profitorientierte Krankenhäuser eine tatsächlich höhere Mortalität aufweisen, an Überzeugung.)

Diese Herauf- und Herabstufungen wirken kumulativ. Wenn z.B. bei der Bewertung randomisierter Studien sowohl schwerwiegende Limitierungen als auch Bedenken bezüglich der Vergleichbarkeit (Direktheit) der Evidenz auftreten, dann fällt die Einstufung von hoher auf geringe Qualität.

Für die Bewertungen der Qualität der Evidenz für Risiken sollten die gleichen Regeln wie für die Bewertung des Nutzens angewandt werden. Wichtige potentielle Risiken können und sollten in Zusammenfassungen der Evidenz berücksichtigt werden. Darin aufgelistet wird die indirekte Evidenz, welche diese Risiken plausibler macht. Wenn z.B. keine Studienergebnisse einer potentiell angstzeugenden Wirkung eines bevölkerungsweiten Melanomscreenings vorliegen, so kann es doch sinnvoll sein, Ergebnisse von anderen Screeningstudien zu berücksichtigen.

Bewertungen der Qualität der Evidenz für wichtige Outcomes verschiedener Studien können und sollten im Kontext systematischer Übersichtsarbeiten, wie z.B. der Cochrane Collaboration, erfolgen. Einschätzungen

Kasten 3: Unpräzise oder spärliche Datenlage

Es gibt keine empirische Grundlage zur Definition unpräziser oder spärlicher Daten (sparse data). Zwei mögliche Definitionen sind:

- Daten sind spärlich, wenn die Ereignisrate oder die Zahl der Beobachtungen so niedrig ist, dass sie nicht informativ sind
- Es handelt sich um unpräzise Daten, wenn das Konfidenzintervall ausreichend weit ist, so dass ein Effektschätzer sowohl mit einem klinisch relevanten Risiko als auch Nutzen vereinbar ist

Diese unterschiedlichen Definitionen können zu unterschiedlichen Bewertungen führen. Auch wenn es vielleicht nicht möglich ist, diese Unterschiede in Übereinstimmung zu bringen, so empfehlen wir für die Frage einer möglichen Herabstufung der Qualität der Evidenz aufgrund einer unpräzisen oder spärlichen Datenlage folgende Vorgehensweise:

- Der Schwellenwert, Daten als unpräzise oder spärlich zu bezeichnen, sollte bei Einzelstudien niedriger ausfallen. Ergebnisse von Einzelstudien mit kleinen Fallzahlen (oder niedrigen Ereignisraten) und damit großen Konfidenzintervallen – die daher einen potentiellen klinischen Nutzen wie auch Schaden mit einschließen – sollten als unpräzise oder spärliche Daten beschrieben werden
- Effektschätzer mit ausreichend großen Konfidenzintervallen, welche, ungeachtet anderer Outcomes, zu widersprüchlichen Empfehlungen führen, sollten mit dem Attribut unpräzise oder spärliche Daten versehen werden.

bezüglich der zusammenfassenden Qualität der Evidenz, der Nutzen-Schaden Abwägung und der daraus resultierenden Empfehlungen, erfordern typischerweise weitere Informationen, die über die Ergebnisse von Übersichtsarbeiten hinausgehen.

Die abschließende Qualitätsbeurteilung der Evidenz

Bei der abschließenden Qualitätsbeurteilung der Evidenz legten bisherige Systeme üblicherweise nur den Nutzen der bewerteten Intervention zugrunde. Wenn das Risiko einer unerwünschten Wirkung entscheidend für die Bewertung ist, jedoch die Qualität dieser Evidenz schwächer ausfällt als für den Nutzen, dann ist es problematisch, diese Unsicherheit der Evidenzlage einfach zu ignorieren. Wir empfehlen daher, die für alle entscheidenden Outcomes niedrigste Qualitätsstufe als Ausgangspunkt für die abschließende Qualitätsbeurteilung zu verwenden.

Outcomes, die wichtig, aber nicht entscheidend sind, sollten zwar in Evidenz-Tabellen enthalten sein und auch bei der Abwägung zwischen Nutzen und Schaden einer Maßnahme einfließen können, jedoch nicht in der abschließenden Qualitätsbeurteilung der Evidenz Verwendung finden. Die Bewertung, ob ein Outcome „entscheidend“, „wichtig“ (aber nicht entscheidend) oder auch „nicht wichtig“ ist, stellt ein Werturteil dar. So weit es geht, sollten dabei die Wertvorstellungen derjenigen einfließen, die durch die Empfehlung der Intervention direkt betroffen sein werden.

Welche Outcomes entscheidend sind, kann schwierig zu entscheiden sein. Die Plausibilität möglicher unerwünschter Wirkungen kann ausschlaggebend bei der Bewertung sein, ob es sich um einen entscheidenden Outcome handelt. Schwache Evidenz eines unplausiblen, aber vermeintlichen Risikos sollte jedoch die abschließende Qualitätsbeurteilung nicht beeinflussen. Einschätzungen, ob ein vermeintliches Risiko plausibel ist, können aus indirekter Evidenz gewonnen werden. Sollten z.B. bei einem Medikament schwerwiegende Bedenken aus Tierversuchen bezüglich einer ernsthaften unerwünschten Wirkung bestehen, dann müsste die abschließende Qualitätsbeurteilung der Evidenz von Studien am Menschen in Bezug auf diese unerwünschte Wirkung niedriger eingestuft werden. Aufgrund des Mangels an Evidenz bei putativen, plausiblen Risiken ist es manchmal unmöglich, den Netto-Nutzen einer Intervention einzuschätzen. In solchen Situationen kann z.B. ein Leitliniengremium empfehlen, dass weitere Forschung notwendig ist.

Wenn die Evidenz für alle entscheidenden Outcomes übereinstimmend die Durchführung einer Intervention unterstützt, jedoch nicht für alle entscheidenden Outcomes eine hohe Qualität der Evidenz vorliegt, so sollte dennoch die Qualität der Evidenz als hoch eingestuft werden. Zum Beispiel gibt es Evidenz hoher Qualität, dass bei Postinfarktpatienten Thrombozytenaggregationshemmer das Risiko nicht-tödlicher Schlaganfälle und nicht-tödlicher Myokardinfarkte verringern. Obwohl es nur eine mittlere Qualität der Evidenz für eine Reduktion der Gesamtmortalität gibt, wäre es angebracht, die abschließende Qualitätsbeurteilung der Evidenz weiterhin als hoch einzustufen, auch wenn man den Outcome Gesamtmortalität als entscheidend für eine Empfehlung einstuft.

Empfehlungen

Verursacht die Intervention mehr Nutzen als Schaden?

Empfehlungen beinhalten ein Abwägen zwischen Nutzen und Schaden. Um diese Abwägung durchzuführen, muss, ob implizit oder explizit, jedem Outcome eine relative Wertschätzung beigemessen werden. Es ist häufig schwer zu entscheiden, welche Gewichtung einzelnen Outcomes zuzuschreiben ist – unterschiedliche Menschen haben häufig auch unterschiedliche Wertvorstellungen. Werden Bewertungen im Namen anderer (z.B. Patienten) durchgeführt, so werden diese auf eine solideren Grundlage gestellt, wenn man die Erkenntnisse über die Wertvorstellungen dieser Patienten mit einbezieht. Empfiehlt man z.B. eine Chemotherapie für Frauen mit Brustkrebs im Früh-

Kasten 4: Werte sind weder richtig noch falsch

Das folgende Beispiel zeigt, wie Menschen aufgrund unterschiedlicher Wertvorstellungen unterschiedliche Empfehlungen formulieren, obwohl sie die zugrunde liegende Evidenz gleich einschätzen.

Fragestellung: Sollte ein bevölkerungsweites Melanomscreening durchgeführt werden?

Praxissituation: Allgemeinärztlich Praktizierende in den USA

Grundrisiko: Allgemeine Population (Inzidenz des Melanoms 1995: 13,3 in 100 000)

Quelle: Helfand et al. Screening for skin cancer. Systematic evidence review No 2. Rockville, MD: Agency for Healthcare Research and Quality. April 2001. (AHRQ Publication No 01-S002.)

Für die Genauigkeit des Screenings und den Outcome tödlich verlaufender Melanome liegt nur eine Evidenz sehr geringer Qualität vor. Falsch-positive Screeningtests sind eine mögliche negative Konsequenz des Screenings, aber es gibt keine Evidenz dazu. Daher liegt der Schluss nahe, die abschließende Qualität der Evidenz als gering und den Nutzen eines Melanomscreenings als ungewiss einzustufen. Basierend auf einer einzigen Fall-Kontroll-Studie ergab sich für tödlich verlaufende Melanome ein odds ratio von 0,37 beim Vergleich von Screening- zur Nicht-Screening-Population. Das Lebenszeitrisiko für weiße Männer an einem Melanom zu versterben, wurde auf 0,36% geschätzt.

Aufgrund dieser Evidenz kämen viele zu einer Einschätzung, ein Melanomscreening eher „nicht zu tun“, weil sie die Vermeidung von unbekanntem, aber potentiellen Risiken im Vergleich zum ungewissen Nutzen als wichtig empfinden. Dennoch würden manche ein bevölkerungsweites Melanomscreening eher „wahrscheinlich tun“, weil sie einen geringen, aber potentiell klinisch relevanten Nutzen von Screening im Vergleich zu den unbekanntem Risiken als wichtig empfinden. Unter diesen Umständen und nach Abschätzen der Kosten, könnte ein Leitliniengremium entscheiden, keine Empfehlung für die Praxis auszusprechen, sondern spezifische Empfehlungen bezüglich der klinischen Forschung zu geben, die notwendig wäre, um diese Unsicherheit zu verringern und eine Abwägung zwischen Nutzen und Schaden zu ermöglichen.

Dieses Beispiel ist typisch für die Wertvorstellungen, die den Empfehlungen für Screeningprogrammen zugrunde liegen. Aber auch bei Empfehlungen zu Behandlungen akuter und chronischer Erkrankungen muss immer das Abwägen des erwarteten Nutzens gegen den erwarteten Schaden in Anbetracht der generellen Ungewissheit und der relativen Wertschätzung wichtiger Outcomes vorgenommen werden.

stadium, dann kann solch eine Empfehlung wesentlich fundierter erfolgen, wenn man Erkenntnisse darüber besitzt, welche Wichtigkeit Frauen dem Rezidivrisiko im relativen Vergleich zu den unerwünschten Wirkungen einer Chemotherapie zusprechen.

Wir empfehlen explizite Abwägungen zwischen dem vornehmlichen Gesundheitsnutzen und Schaden vorzunehmen, bevor Kosten erwogen werden. Verursacht die Intervention mehr Nutzen als Schaden? Variiert der Nutzen im Vergleich zum möglichen Schaden bei bestimmten Patienten oder in bestimmten Situationen, so müssen Empfehlungen auf definierte Praxis-situationen und Patientengruppen zugeschnitten werden. Empfiehlt man z.B. den Einsatz von Marcumar bei Patienten mit Vorhofflimmern zur Verringerung des Schlaganfallsrisikos, so steigt damit auch gleichzeitig das Blutungsrisiko. Empfehlungen oder die Stärke der Empfehlung werden natürlicherweise unterschiedlich ausfallen in Situationen, in denen eine regelmäßige Überwachung der Gerinnung möglich ist, im Vergleich zu Situationen, in denen dies nicht der Fall ist. Darüber hinaus werden aufgrund der unterschiedlichen absoluten Risikoreduktion Empfehlungen (oder die Stärke der Empfehlung) bei Patienten mit einem sehr niedrigem Schlaganfallrisiko (Patienten unter 65 Jahre ohne Komorbidität) wahrscheinlich anders ausfallen als bei Patienten mit einem weitaus höheren Risiko (wie z.B. ältere Patienten mit Herzinsuffizienz). Empfehlungen müssen daher eindeutig in Bezug auf Patientengruppen und Praxissituation sein. Bei der Formulierung von Empfehlungen ist es besonders wichtig, Umstände von benachteiligten Gruppen zu berücksichtigen und, wo geboten, die Empfehlungen dementsprechend zu modifizieren.

Beim Abwägen zwischen Nutzen und Schaden einer Intervention empfehlen wir die folgenden Definitionen zur Kategorisierung:

Netto-Nutzen = Eindeutig größerer Nutzen als Schaden

Abwägung notwendig = Eine Abwägung ergibt sowohl wesentlichen Nutzen als auch wesentlichen Schaden

Abwägung unsicher = Es ist nicht klar, inwieweit der Nutzen den Schaden überwiegt

Kein Netto-Nutzen = Eindeutig größerer Schaden als Nutzen

Verfasst man Empfehlungen, so sollte man vier Hauptfaktoren berücksichtigen:

- Beim Abwägen von Nutzen und Schaden ist die Größe des Effekts der wesentlichen Outcomes zu berücksichtigen, wie auch die Konfidenzintervalle um den Effektschätzer sowie die relative Wertschätzung, die diesen Outcomes zugeteilt wird
- Die abschließende Qualität der Evidenz
- Wichtige Faktoren die bei der Umsetzung der Evidenz in die spezifische Praxissituation die Größenordnung des Effekts verändern könnten, wie z.B. die Nähe zu einem Krankenhaus oder das Vorhandensein von notwendiger Expertise
- Unsicherheit über die Größenordnung des Grundrisikos in der Zielpopulation

Bei Ungewissheit, inwieweit die Evidenz in die Praxis umsetzbar ist, oder wenn Unklarheit über das Grund-

Tabelle 2: Qualitätsbewertung der Studien bezüglich der Therapie einer mittelschweren Depression mit Serotonin-Wiederaufnahmehemmer (SSRIs) oder trizyklischen Antidepressiva (TZAs). Praxissituation: Allgemeinärztlich Tätige²

Bewertung der Qualität der Evidenz						Zusammenfassung der Ergebnisse					
Anzahl der Studien	Design	Qualität	Konsistenz	Direktheit	Andere Faktoren*	Anzahl der Patienten		Effektschätzer			
						SSRIs	TZAs	Relativ (95% KI)	Absolut	Qualität	Wichtigkeit
Ausprägung der Depression (auf der Hamilton Depressionsskala nach 4 - 12 Wochen Therapie)											
Citalopram (8) Fluoxetine (38) Fluvoxamine (25) Nefazodone (2) Paroxetine (18) Sertraline (4) Venlafaxine (4)	Randomisierte, kontrollierte Studien	Keine schwerwiegenden Limitierungen	Keine wichtige Inkonsistenz	Ungewissheit bezüglich der Direktheit der Evidenz (Outcome)†	Keine	5044	4510	WMD 0,0034 (-0,007; 0,075)	Kein Unterschied	Mittlere Qualität	Entscheidend
Vorübergehende unerwünschte Wirkungen die zur Beendigung der Therapie führten											
Citalopram (8) Fluoxetine (50) Fluvoxamine (27) Nefazodone (4) Paroxetine (23) Sertraline (6) Venlafaxine (5)	Randomisierte, kontrollierte Studien	Keine schwerwiegenden Limitierungen	Keine wichtige Inkonsistenz	Direkt	Keine	1948/7032 (28%)	2072/6334 (33%)	RRR 13% (5%; 20%)	5/500	Hohe Qualität	Entscheidend
Tödlich verlaufende Überdosierungen[§]											
UK Office for National Statistics (1)	Beobachtungsstudien	Schwerwiegende Limitierung‡	Nur eine Studie	Direkt	Sehr starke Assoziation	1/100.000 Behandlungsjahre	58/100.000 Behandlungsjahre	RRR 98% (97%; 99%)§	6/10.000	Mittlere Qualität	Entscheidend

WMD = gewichtete Mittelwertdifferenz (weighted mean difference), RRR = Relative Risiko Reduktion, KI = Konfidenzintervall

* Unpräzise oder spärliche Datenlage, eine starke oder sehr starke Assoziation, hohes Risiko eines Publikationsbias, Vorliegen einer Dosis-Wirkungs-Beziehung oder eines plausiblen, verbleibenden Confoundings

† Bei der Bewertung ergab sich aufgrund der kurzen Dauer der Studien eine Ungewissheit bezüglich der Direktheit der Evidenz

‡ Es ist möglich, dass Patienten mit weniger ausgeprägter Depression eher ein SSRI gegeben wurde und es ist ungewiss, ob das Wechseln von Antidepressiva Suizidversuche verhindert hätte

§ Es herrscht Ungewissheit bezüglich des Grundrisikos für tödlich verlaufende Überdosierungen

risiko besteht, kann dies unser Vertrauen in die Empfehlung verringern. Wenn zum Beispiel eine Intervention schwerwiegende unerwünschte Wirkungen, aber auch entscheidende Vorteile aufweist, dann ist eine Empfehlung wesentlich unsicherer, wenn die Größe des Grundrisikos unklar ist, als wenn sie bekannt wäre.

Wir schlagen die folgenden Kategorien für Empfehlungen vor:

„Tun“ („Do it“) oder „Nicht tun“ („don't do it“) – Beschreibt eine Entscheidung, die die meisten, gut informierten Personen treffen würden.

„Wahrscheinlich tun“ („Probably do it“) oder „Wahrscheinlich nicht tun“ („probably don't do it“) – Beschreibt eine Entscheidung, die eine gut informierte Mehrheit treffen würde, jedoch eine ebenso gut informierte, deutliche Minderheit nicht.

Eine Empfehlung für oder gegen eine Intervention bedeutet nicht automatisch, dass alle Patienten identisch behandelt werden sollen. Auch bedeutet dies nicht, dass Kliniker ihre Patienten nicht in die Entscheidung mit einbeziehen oder nicht die Vor- und Nachteile von Alternativen beschreiben sollen. Da jedoch die meisten gut informierten Menschen eine gleiche Entscheidung treffen würden, bedeutet dies, dass man das Beschreiben der relativen Vor- und Nachteile von Alternativen eher kurz fassen kann. Eine Empfehlung ist dazu gedacht, zu einer angemessenen Entscheidung für individuelle Patienten, wie auch für Populationen, zu gelangen. Sie sollte daher widerspie-

geln, was Patienten aufgrund der Evidenz und ihrer eigenen Wertvorstellungen oder Präferenzen in Bezug auf die zu erwartenden Outcomes wählen würden. Eine Empfehlung eine Intervention „wahrscheinlich zu tun“ ist daher ein Anlass für Kliniker, wesentlich genauer und umfassender die Wertvorstellungen und Präferenzen ihrer Patienten zu erkunden, sollten sie die Intervention empfehlen.

In manchen Fällen, wenn eine Abwägung unsicher ist oder keine Übereinstimmung erzielt werden kann, ist es nicht angemessen, Empfehlungen auszusprechen (siehe Kasten 4). Falls dies durch einen Mangel an qualitativ ausreichender Evidenz hervorgerufen wurde, so sollte das Durchführen weiterer Studien empfohlen werden, damit die für eine informierte Empfehlung benötigte Evidenz erzeugt werden kann.

Die Kosten-Nutzen Abwägung

Da Ausgaben für eine Intervention bedeuten, dass für andere Interventionen weniger finanzielle Ressourcen zur Verfügung stehen, wird bei Empfehlungen, ob implizit oder explizit, die Bewertung bezüglich des inkrementellen Nutzens auch in Hinblick auf die zusätzlich entstehenden Kosten beeinflusst. Kosten – der monetäre Wert der verwendeten Ressourcen – sind ein wichtiger Faktor bei der Formulierung von Empfehlungen, verhalten sich aber kontextspezifisch, verändern sich mit der Zeit und ihre Größenordnung kann schwer zu ermitteln sein. Obwohl die Schwierigkeit bekannt ist, korrekte Kostenschätzungen vorzunehmen, empfehlen wir dennoch die inkrementellen Kosten der zu bewertenden Alternativen im Zusammenhang

mit dem zu erwarteten Nutzen und Schaden explizit aufzuführen. Wo relevant und vorhanden, sollten disaggregierte Kosten (Unterschiede in der Verwendung von Ressourcen) in den Evidenz-Tabellen zusammen mit den wichtigen Outcomes aufgelistet werden. Die Beurteilung der Qualität der Evidenz für Unterschiede in der Verwendung von Ressourcen sollte mit dem gleichen (oben genannten) Vorgehen wie für andere wichtige Outcomes durchgeführt werden.

Wie GRADE in der Praxis durchgeführt wird

Tabelle 2 zeigt an einem Beispiel wie das GRADE System verwendet wurde, um die Evidenz aus einer systematischen Übersichtsarbeit aus dem Jahre 1997 über den Vergleich von SSRIs zu TZAs aufzuarbeiten (49). Nach Diskussion wurde festgelegt, dass für die Wirkung von SSRIs und TZAs bezüglich der Outcomes „Schwere der Depression“ und „Tödliche Überdosierungen“ eine Evidenz mittlerer Qualität vorliegt und Evidenz hoher Qualität für „vorübergehende unerwünschte Wirkungen“. Es bestand dann Übereinkunft, dass die abschließende Qualitätsbewertung (alle entscheidenden Outcomes einschließend) „Evidenz mittlerer Qualität“ sei. Desweiteren wurde ein „Netto-Nutzen“ von SSRIs im Vergleich zu TZAs aus den Daten ermittelt (kein Unterschied in der Schwere der Depression, geringere vorübergehende unerwünschte Wirkungen und weniger tödlich verlaufende Überdosierungen). Obwohl Konsensus bezüglich des Netto-Nutzens von SSRIs bestand, wurde aufgrund der Unsicherheit in der Evidenzlage (Qualität: mittel) nur die Empfehlung SSRIs „wahrscheinlich“ eher als TZAs anzuwenden. Uns standen für diese GRADE Bewertung keine Daten über Kosten von SSRIs im Vergleich zu TZAs zur Verfügung. Wären Kosten mit eingeflossen, so hätte sich die Empfehlung vielleicht verändert.

Schlussfolgerung

Jedes System zur Bewertung der Evidenz und zur Formulierung der Stärke einer Empfehlung muss Einfachheit und Eindeutigkeit miteinander abwägen. Verringert man die Komplexität eines Systems, so reduziert sich meist auch seine Eindeutigkeit, da bei weniger komplexen Systemen Bewertungen häufiger implizit als explizit getroffen werden müssen. Auf der anderen Seite wird das System durch Bemühungen zu mehr Eindeutigkeit und Transparenz wahrscheinlich komplexer. In dem hier beschriebenen System haben wir versucht, ein Gleichgewicht zwischen Einfachheit und Eindeutigkeit zu finden. Unabhängig davon, wie einfach oder komplex ein System ist – Bewertungen müssen immer vorgenommen werden. Das GRADE Prinzip versucht ein Bezugssystem für eine strukturierte Vorgehensweise aufzustellen und hilft sicherzustellen, dass angemessene Bewertungen erfolgen. Es kann jedoch nicht die Aufgabe abnehmen, aktive Bewertungen vornehmen zu müssen.

Mitglieder der “Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group” und vertiefende Ausführungen sind auf der Website www.GradeWorkingGroup.org aufgelistet.

Interessenskonflikte: Die meisten Mitglieder der GRADE Working Group sind Vertreter bzw. Entwickler bisher verwendeter Systeme zur Bewertung der Qualität der Evidenz und Stärke der Empfehlung.

Anmerkung: Es war das Ziel, möglichst nahe an der Originalfassung des englischen Manuskripts zu bleiben. Eine Anzahl von Begriffen mussten sowohl im Original als auch in der deutschen Übersetzung neu definiert werden (z.B. consistency and directness of evidence – Konsistenz und Direktheit der Evidenz).

Literatur

1. Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979;121: 1193-254.
2. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1986;89(suppl 2): 2-3S.
3. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Arch Intern Med* 1986;146: 464-5.
4. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. *Chest* 1989;95: 2-4S.
5. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Antithrombotic therapy consensus conference. *Chest* 1992;102(suppl 4): 305-11S.
6. US Department of Health and Human Services, Public Health Service, Agency Health Care Policy and Research. Acute pain management: operative or medical procedures and trauma. Rockville, MD: Agency for Health Care Policy and Research Publications, 1992. (AHCPR Pub 92-0038.)
7. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, et al. An approach to the development of practice guidelines for community health interventions. *Can J Public Health* 1994; 85 (suppl 1): S8-13.
8. Hadorn DC, Baker D. Development of the AHCPR-sponsored heart failure guideline: methodologic and procedural issues. *Jt Comm J Qual Improv* 1994;20: 539-54.
9. Cook DJ, Guyatt GH, Laupacis A, Sackett DL, Goldberg RJ. Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* 1995;108(suppl 4): 227-30S.
10. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, et al. Users' guide to the medical literature IX: a method for grading health care recommendations. *JAMA* 1995;274: 1800-4.
11. Scottish Intercollegiate Guidelines Network (SIGN). Forming guideline recommendations. In: A guideline developers' handbook. Edinburgh: SIGN, 2001. (Publication No 50.) www.sign.ac.uk/guidelines/fulltext/50/section6.html (accessed 16 Nov 2004).
12. US Preventive Services Task Force. Guide to clinical preventive services. 2nd ed. Baltimore: Williams & Wilkins, 1996: xxxix-lv.
13. Eccles M, Clapp Z, Grimshaw J, Adams PC, Higgins B, Purves I, et al. North of England evidence based guidelines development project: methods of guideline development. *BMJ* 1996;312: 760-2.
14. Centro per la Valutazione della Efficacia della Assistenza Sanitaria (CeVEAS). Schema di grading CeVEAS. <http://web1.satcom.it/interage/ceveas/html/doc/45/GLICO.pdf> (accessed 16 Nov 2004).
15. Guyatt G, Schünemann H, Cook D, Jaeschke R, Pauker S, Bucher H. Grades of recommendation for antithrombotic agents. *Chest* 2001;119:3S-7S. www.chestjournal.org/content/vol119/1_suppl/index.shtml (accessed 16 Nov 2004).
16. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M. Levels of evidence and grades of recommendations. Oxford: Oxford Centre for Evidence-Based Medicine. www.cebm.net/levels_of_evidence.asp (accessed 16 Nov 2004).

17. National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. Canberra: AusInfo, 2000. www.health.gov.au/nhmrc/publications/pdf/cp69.pdf (accessed 16 Nov 2004).
18. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. *BMJ* 2001;323: 334-6.
19. Roman SH, Silberzweig SB, Siu AL. Grading the evidence for diabetes performance measures. *Eff Clin Pract* 2000;3: 85-91.
20. Woloshin S. Arguing about grades. *Eff Clin Pract* 2000;3: 94-5.
21. Guyatt GH, Schünemann H, Cook D, Pauker S, Sinclair J, Bucher H, et al. Grades of recommendation for antithrombotic agents. *Chest* 2001;119: 3-7S.
22. Atkins D, Best D, Shapiro EN, eds. Third US Preventive Services Task Force: background, methods and first recommendations. *Am J Prev Med* 2001;20: 3(suppl):1-108.
23. Woolf SH, Atkins D. The evolving role of prevention in health care: contributions of the US Preventive Services Task Force. *Am J Prev Med* 2001;20: 3(suppl):13-20.
24. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. *Am J Prev Med* 2001;20:3 (suppl): 21-35.
25. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Agüero L, et al. Developing an evidence-based guide to community preventive services—methods. *Am J Prev Med* 2000;18 (suppl 1): 35-43.
26. Zaza S, Wright-De A, Briss PA, Truman BI, Hopkins DP, Hennessy MH, et al. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. *Am J Prev Med* 2000;18(suppl 1): 44-74.
27. Greer N, Mosser G, Logan G, Halaas GW. A practical approach to evidence grading. *Jt Comm J Qual Improv* 2000;26: 700-12.
28. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Rockville, MD: Agency for Healthcare Research and Quality, 2002: 64-88. (AHRQ publication No 02-E016.)
29. Guyatt G, Drummond R, eds. *Users' guide to the medical literature*. Chicago, IL: AMA Press, 2002: 55-154.
30. Clarke M, Oxman AD, eds. Assessment of study quality. *Cochrane reviewers' handbook* 4.1.5 section 6. In: *Cochrane Library*. Issue 4. Oxford: Update Software, 2002.
31. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Davey Smith G, Altman DG, eds. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Books, 2001: 87-121.
32. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Rockville, MD: Agency for Healthcare Research and Quality, 2002: 51-63. (AHRQ publication No 02-E016.)
33. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in health-care trials (Cochrane methodology review). In: *Cochrane Library Issue 4*. Oxford: Update Software, 2002.
34. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;286: 821-30.
35. Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? In: Chalmers I, Maynard A, eds. *Non-random reflections on health care research: on the 25th anniversary of Archie Cochrane's effectiveness and efficiency*. London: BMJ, 1997: 93-106.
36. Lacchetti C, Guyatt G. Surprising results of randomized controlled trials. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature*. Chicago, IL: AMA Press, 2002: 247-65.
37. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280: 605-13.
38. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the women's health initiative randomized controlled trial. *JAMA* 2002;288: 321-33.
39. Dans A, McAlister F, Dans L, Richardson WS, Straus S, Guyatt G. Applying results in individual patients. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature*. Chicago, IL: AMA Press, 2002: 369-84.
40. McAlister F, Laupacis A, Wells G. Drug class effects. In: Guyatt G, Drummond R, eds. *Users' guide to the medical literature*. Chicago, IL: AMA Press, 2002: 415-31.
41. Echt DS, Liebson PR, Mitchell LB, Peters RW, Obias-Manno D, Barker AH, et al. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The cardiac arrhythmia suppression trial. *N Engl J Med* 1991;324: 781-8.
42. Riggs BL, Hodgson SF, O'Fallon WM, Chao EY, Wahner HW, Muhs JM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med* 1990;322: 802-9.
43. Worster A, Preyra I, Weaver B, Haines T. The accuracy of noncontrast helical computed tomography versus intravenous pyelography in the diagnosis of suspected acute urolithiasis: a meta-analysis. *Ann Emerg Med* 2002;40: 280-6.
44. Worster A, Haines T. Does replacing intravenous pyelography with noncontrast helical computed tomography benefit patients with suspected acute urolithiasis? *Can Assoc Radiol J* 2002;53: 241.
45. Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. *BMJ* 2003;326: 472.
46. Bross IDJ. Pertinency of an extraneous variable. *J Chron Dis* 1967;20: 487-95.
47. Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000;(2): CD001855.
48. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. *CMAJ* 2002;166: 1399-406.
49. North of England Evidence Based Guideline Development Project. Evidence based clinical practice guideline: the choice of antidepressants for depression in primary care. Newcastle upon Tyne: Centre for Health Services Research, 1997