



Schwerpunkt

GRADE: Von der Evidenz zur Empfehlung Beschreibung des Systems und Lösungsbeitrag zur Übertragbarkeit von Studienergebnissen

Holger J. Schünemann*

Department of Clinical Epidemiology & Biostatistics and Department of Medicine, McMaster University, Hamilton, Kanada

Zusammenfassung

Die Arbeitsgruppe "Grading of Recommendations Assessment, Development and Evaluation" (GRADE) ist aus einer internationalen Kooperation von Leitlinienentwicklern, Klinikern, Versorgungsforschern und Methodologen entstanden. GRADE wird von führenden internationalen Organisationen, inklusive der Weltgesundheitsorganisation (WHO), als offizielles System angewendet, da es einen wesentlichen Fortschritt in der Evidenzbeurteilung und Entwicklung von Handlungsempfehlungen im Gesundheitswesen darstellt. Das GRADE-System unterscheidet zwischen der Qualität der Evidenz und dem Grad einer Empfehlung für Handlungen im Gesundheitswesen. GRADE definiert die Qualität der Evidenz als Gradmesser, dass ein ermittelter Effekt korrekt ist, wenn es sich um eine Beurteilung der Evidenz von einzelnen Endpunkten handelt. Im Kontext der Entwicklung von Handlungsempfehlungen definiert GRADE die Qualität der Evidenz als einen Gradmesser für das Vertrauen in das Zutreffen eines ermittelten Effekts, der eine Handlungsempfehlung unterstützt. Der Grad einer Handlungsempfehlung, untergliedert in starke und abgeschwächte Empfehlungen für oder gegen eine Maßnahme, wird definiert als das Ausmaß an Sicherheit, dass die wünschenswerten

Konsequenzen einer Behandlung ihre unerwünschten Folgen überwiegen. Eine Handlungsempfehlung nach GRADE bedarf der weiteren besonderen Berücksichtigung der Größe des möglichen Nutzens und Schadens einer Intervention mit Fokussierung auf patientenrelevante Endpunkte, der damit verbundenen Wertvorstellungen und der Integration von Überlegungen zum Ressourcenverbrauch. Dabei ist eine Abwägung aller relevanten Endpunkte gefordert. GRADE liefert insbesondere einen systematischen Ansatz zur Beurteilung der Übertragbarkeit von Studienergebnissen auf die Praxis. Diese Übertragbarkeit wird bei GRADE Direktheit ("directness") genannt und unterscheidet zwischen dem Vorliegen von direkten Vergleichen von alternativen Interventionen, die begutachtet werden, und dem Ausmaß der direkten Beziehung zwischen der Population (P), der Intervention (I), der Vergleichsintervention ("comparator", C) und den Endpunkten ("outcomes", O) (PICO) der vorliegenden Evidenz und der eigentlichen Fragestellung. Zusätzlich zur übersichtlichen Darstellung des GRADE-Systems wird der Ansatz der Bewertung der Übertragbarkeit in diesem Artikel anhand von Beispielen beschrieben.

Schlüsselwörter: Leitlinien, Evidence-basierte Medizin, GRADE, Evidenzstärke

*Korrespondenzadresse: Prof. Dr. Dr. Holger J. Schünemann, M.Sc., Leiter, Department of Clinical Epidemiology & Biostatistics, Michael Gent Chair in Healthcare Research, Professor of Clinical Epidemiology, Biostatistics and Medicine, McMaster University Health Sciences Centre, Room 2C10B, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada. Tel.: +1 905 525 9140 x 24931; fax: +1 443 339 0565.
E-Mail: schuneh@mcmaster.ca

GRADE: From grading the evidence to making a recommendation

A description of the system and a proposal regarding the transferability of the results of clinical trials to clinical practice

Summary

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group represents an international collaboration of guideline developers, clinicians, health services researchers and methodologists. Many leading organizations, including the World Health Organization (WHO), use the GRADE approach because it has led to progress in the assessment of evidence and in the development of healthcare recommendations. The GRADE system distinguishes the quality of evidence from the strength of a recommendation. The quality of evidence reflects the extent of confidence that an estimate of effect is correct if it is used in the context of single endpoints. In the context of giving guidance, it reflects the extent to which confidence in an estimate of the effect is adequate to support recommendations. The strength of a recommendation, separated into strong and weak or conditional recommendations for or against an intervention, is defined as the extent to which one can be confident that

the desirable effects of an intervention outweigh the undesirable effects. A recommendation for action requires consideration for the magnitude of the expected benefit and downsides of an intervention for all patient-important endpoints, the associate values and preferences and resource use. The GRADE system includes a systematic approach to evaluate the generalizability of study results to healthcare practice. Judgments about generalizability, better termed directness, are separated into judgments about the availability of direct comparisons between two alternative management strategies and judgments about differences between the population, intervention, comparator to the intervention, and outcomes (PICO) of interest for a given question, and those included in the relevant studies. In addition to providing an overview of the GRADE system, this article focuses on the approach to assessing directness or generalizability.

Key words: guidelines, evidence-based medicine, GRADE, level of evidence

Einleitung

Die Grading of Recommendations Assessment, Development and Evaluation (GRADE) Arbeitsgruppe ist aus einer internationalen Kooperation von Leitlinienentwicklern, Klinikern, Versorgungsforschern und Methodologen entstanden [1]. GRADE (www.grade-workinggroup.org) wird von führenden internationalen Organisationen, inklusive der Weltgesundheitsorganisation (WHO), als offizielles System angewendet [2], da es einen wesentlichen Fortschritt in der Evidenzbeurteilung und Entwicklung von Handlungsempfehlungen im Gesundheitswesen darstellt (Tabelle 1). Das GRADE System unterscheidet zwischen der Qualität der Evidenz und dem Grad einer Empfehlung für Handlungen im Gesundheitswesen [3–8]. GRADE definiert die Qualität der Evidenz als ein Gradmesser, dass ein ermittelter Effekt korrekt ist, wenn es sich um eine Beurteilung der Evidenz von einzelnen Endpunkten handelt. Im Kontext der Entwicklung von Handlungsempfehlungen definiert GRADE die Qualität der Evidenz als ein Gradmesser für das Vertrauen in das Zutreffen eines ermittelten Effekts, der eine Handlungsempfehlung unterstützt. Der Grad einer Handlungsempfehlung, untergliedert in starke und abgeschwächte Empfehlungen für oder

gegen eine Maßnahme, wird definiert als das Ausmaß an Sicherheit, dass die wünschenswerten Konsequenzen einer Behandlung ihre unerwünschten Folgen überwiegen. Eine Handlungsempfehlung nach GRADE bedarf der weiteren besonderen Berücksichtigung der Größe des möglichen Nutzen und Schaden einer Intervention mit Fokussierung auf Patienten-relevante Endpunkte, der damit verbundenen Wertvorstellungen, und der Integration von Überlegungen zum Ressourcenverbrauch. Dabei ist eine Abwägung aller

relevanten Endpunkte gefordert, insbesondere derer, die für die Entscheidungsfindung von kritischer Bedeutung sind.

GRADE liefert insbesondere einen systematischen Ansatz zur Beurteilung der Übertragbarkeit von Studienergebnissen auf die Praxis. Diese Übertragbarkeit wird bei GRADE Direktheit ("directness") genannt und unterscheidet zwischen dem Vorliegen von direkten Vergleichen von alternativen Interventionen, die begutachtet werden, und dem Ausmaß der direkten Bezie-

Tabelle 1. Auswahl an Organisationen und Initiativen, die das GRADE System unterstützen, oder anwenden.

World Health Organization (WHO)
National Institute for Health and Clinical Excellence, UK (NICE)
Cochrane Collaboration
British Medical Journal
European Respiratory Society (ERS)
American College of Physicians (ACP)
Agency for Health Care Research and Quality (AHRQ)
Allergic Rhinitis in Asthma Guidelines (ARIA)
American College of Chest Physicians (ACCP)
American Thoracic Society (ATS)
COMPUS at The Canadian Agency for Drugs and Technologies in Health (CADTH) – Canada
EBM Guidelines-Finland/International
Infectious Disease Society of America (IDSA)
European Society of Thoracic Surgeons
BMJ Clinical Evidence
Socialstyrelsen National Board of Health and Welfare – Sweden
Spanish Society for Family and Community Medicine
UpToDate®

hung der Population (P), der Intervention (I), der Vergleichsintervention ("comparator", C) und der Endpunkte ("outcomes", O) (PICO) aus der vorliegenden Evidenz und der eigentlichen Fragestellung. Die Beschreibung des Ansatzes zur Übertragbarkeit ist das eigentliche Anliegen dieses Artikels weshalb ich zunächst die Beurteilung der Übertragbarkeit von Studienergebnissen nach GRADE beschreibe. Desweiteren stelle ich das GRADE System insgesamt in einer Übersicht dar, wobei ich die Leser auch auf andere Publikationen verweise [3–8], die der Verbreitung des Systems und der Vorbereitung für diesen Artikel dienen. Insbesondere verweise ich auf einen vorherigen Artikel in dieser Zeitschrift und eine weitere deutsche Publikation [3,9]. Eine detaillierte Beschreibung aller Aspekte des GRADE Systems mit vielfachen illustrativen Beispielen in deutscher Sprache wird als eine Artikelserie in dieser Zeitschrift in naher Zukunft erscheinen.

Fragestellungen: Evidenz existiert für (nahezu) jede relevante Frage in der Gesundheitsversorgung

Die Entstehung einer Handlungsempfehlung beginnt mit dem Stellen der Frage, die beantwortet werden soll. Nehmen wir als Beispiel die medikamentöse Behandlung der Vogelgrippe, der Infektion mit dem Influenza H5N1 Virus. Als die WHO im Jahr 2006 mit der Frage konfrontiert wurde, wie bei den sporadisch auftretenden Fällen gehandelt werden sollte, wurde eine Leitliniengruppe gebildet, um relevante Fragen zu diesem Thema zu beantworten [10,11]. Eine der Fragen, die sich in Abwesenheit von Transmission des Virus von Mensch zu Mensch stellte, war die, ob bei der Behandlung von Patienten mit H5N1 Infektion (Population) mit Oseltamivir (Intervention) im Vergleich zu keiner antiviralen Behandlung (Comparator), eine Mortalitäts-senkung eintritt, Krankenhauseinweisungen vermindert werden, Pneumonien verhindert werden und ob dieser eventuelle Nutzen den Schaden durch Nebenwirkungen und die entstanden

Kosten überwiegt (Outcomes). Eine oberflächliche und vorschnelle Betrachtung der Frage, hätte schnell zu dem Schluss geführt, dass für die Beantwortung dieser Frage „keine Evidenz“ vorliegt. Eine konträre Auffassung ist die, dass wenn sich eine Frage stellt, in der Regel Evidenz vorliegt. Die eigentliche Aufgabe, die es somit zu bewältigen gilt, ist zu beurteilen, wie gut und vollständig diese Evidenz ist. Eine differenzierte Betrachtung der Problemstellung zur Vogelgrippe führte dann zu der Erkenntnis, dass zwar nahezu keine direkte Evidenz für die Behandlung dieser Patienten vorlag (nur 36 Patienten hatten Oseltamivir zu diesem Zeitpunkt unkontrolliert erhalten), aber dass indirekte Evidenz aus Studien (nämlich 5 randomisierte klinische Studien), die Patienten mit der gewöhnlichen Influenza untersucht

haben, relevant sein kann. Die Beurteilung, die in solchen Fällen vorgenommen werden muss, ist, wie indirekt oder übertragbar die Evidenz von den jeweiligen Studien oder von unsystematischen Beobachtungen auf die Beantwortung der Frage ist. Das PICO Schema ist dafür hervorragend geeignet indem man die Übertragbarkeit sequentiell für diese Komponenten beurteilt (Tabelle 2). Man begutachtet dann z.B. wie übertragbar die Ergebnisse in den Patienten, die in den Studien zur gewöhnlichen Grippe behandelt wurden, auf die Patienten mit H5N1 Infektion sind und wie übertragbar die Intervention (Gabe von Medikamenten unter verschiedenen Bedingungen), der Vergleich und die kritischen Endpunkte von den Studien auf die zu beantwortende Frage sind.

Tabelle 2. Übertragbarkeit am Beispiel der Vogelgrippe.

Komponente	Fragestellung	Vorhandene Evidenz in Studien	Übertragbarkeit?
P(opulation)	Influenza A(H5N1) infizierte Patienten	Gewöhnliche Influenza	Sehr fraglich (indirekt), da anderes Virus mit anderem Resistenzmuster und schwerere Erkrankung
I(ntervention)	Oseltamivir	Oseltamivir	Für alleinige Gabe des Medikaments relativ sicher (Absorption bei intubierten Patienten mit H5N1 Infektion relativ sicher) (direkt)
C(omparator)	Keine medikamentöse Intervention mit Oseltamivir	Keine medikamentöse Intervention mit Oseltamivir	Gesichert (direkt)
O(outcomes) (Beispiele)	Hospitalisierung	Hospitalisierung	Fragliche Übertragbarkeit, da Schweregrad der Erkrankung Hospitalisierung von H5N1 Patienten wahrscheinlicher macht (indirekt)
	Mortalität	Mortalität	Wenn es Mortalitätsdaten gäbe wären diese wohl übertragbar (direkt), da die Bestimmung der Mortalität sich in den eingeschlossenen Studien wenig von der im klinischen Alltag unterscheiden könnte. Allerdings bleibt die sehr indirekte Evidenz bezüglich der Population.
	Pneumonien	Pneumonien	Fragliche Übertragbarkeit für Patienten mit H5N1 Infektion, da Pneumonien bei H5N1 Patienten mit grösserer Mortalität einhergehen und sich anders präsentieren (indirekt)
	Neurologische Nebenwirkungen	Neurologische Nebenwirkungen	Übertragbarkeit gegeben, da die Nebenwirkungen wohl ähnlich wären. Die Diagnose dieser Nebenwirkungen wäre bei Schwerkranken aber komplizierter.

Als Nebenbemerkung lässt sich anfügen, dass die Varianten des PICO Schemas wie die Anfügung des Zeitrahmens der Beobachtung durch die sogenannte PICOT Fragestellung mit dem "t" für time, nur begrenzt hilfreich sind, da diese Größe in die Beurteilung des "outcomes" eingegliedert werden kann und der Einfachheit wegen auch sollte (z.B. Mortalität innerhalb welchen Zeitraums).

Parallelen zur Versorgungsforschung

Aehnliche Situationen, wie die geschilderte, muss die Versorgungsforschung bewältigen, gelegentlich auf Systemebene, in anderen Situationen bei der Behandlung einzelner Patienten. Die Unterscheidung von mechanistischen und praktischen Studien ist dabei hilfreich [12,13]. Eine Studie kann als *mechanistisch* bezeichnet werden, wenn es einen Sachverhalt zu erklären versucht und beispielsweise die biologische oder psychologische Wirksamkeit einer Intervention beschreibt. Eine Studie kann als *praktisch* bezeichnet werden, wenn eine Studie detaillierte und vollständige Informationen enthält, die direkt auf die zu beantwortende spezifische Frage in der Gesundheitsversorgung anwendbar sind. Zwischen diesen Extremen liegt ein Kontinuum an Evidenz und Studien, die entweder direkt oder weniger direkt auf die Patientenversorgung anwendbar sind. Wenn das Ziel einer Studie die größtmögliche Übertragbarkeit auf Entscheidungen auf einzelne Patienten ist, dann würde man alle Patienten ausschließen, die mit aller Wahrscheinlichkeit keinen Nutzen erfahren würden (z.B. Patienten, die die Intervention nicht genau befolgen oder befolgen können, die eine unsichere Diagnose haben, die von Klinikern oder in Systemen behandelt werden, deren unterschiedliche Erfahrung und Aufmachung die Endpunkte beeinflussen würden oder die Ko-Interventionen erhalten würden, die die Effektivität der Behandlung beeinflussen würden). Studien der Versorgungsforschung dienen als Beispiele, bei denen es zwar um

solche praktischeren Fragestellungen geht, die allerdings häufig an mangelnder Übertragbarkeit leiden, da die Patienten oder Systeme, die den Interventionen exponiert sind, häufig heterogen und z.T. indirekt sind. Die Beurteilung dieser "indirectness" wird durch das PICO Schema vereinfacht. Nehmen wir das Beispiel der chronisch obstruktiven Lungenerkrankung (COPD), der vierthäufigsten Todesursache weltweit mit einer geschätzten Anzahl von 3 bis 4 Millionen Patienten in Deutschland [14]. Die pulmonale Rehabilitation ist eine wirksame Inter-

vention für diese Patienten, die in vielen klinischen Studien beschrieben wurde. Die pulmonale Rehabilitation führt zur deutlichen Verbesserung der Lebensqualität bei diesen Patienten. Zunehmend werden Studien, die sich mit der Patientenversorgung auf Systemebene beschäftigen, durchgeführt und als Versorgungsforschung beschrieben. Eine separate Beurteilung der "directness" solcher Studien, die von Schwarz und Lellouch als *pragmatisch* bezeichnet wurden [13], ist zwar nötig, kann aber nach den gleichen PICO Prinzipien erfolgen. Zudem ist eine noch diffe-

Tabelle 3. Übertragbarkeit in mechanistischen Studien und Versorgungsforschungsstudien.

Komponente	Mechanistische Studien	Pragmatische Versorgungsforschung	Übertragbarkeit auf einzelne Patienten (praktisch?)
P (opulation)	Patienten mit schwerer COPD ohne Ko-Morbidität, die optimale medikamentöse Behandlung erhalten	Patienten mit moderater und schwerer COPD und Ko-Morbidität (z.B. Herzinsuffizienz), z.T. nicht mit optimaler medikamentöser Behandlung der COPD	Sind die Populationen in der Praxis ähnlich genug um den gleichen biologischen und psychologischen Effekt zu erhalten wenn?
I (ntervention)	Pulmonale Rehabilitation: Körperliches Trainingsprogramm, psychologische Betreuung, hohe Intensität in spezialisierten Zentren	Pulmonale Rehabilitation: Weniger intensives Trainingsprogramm, gelegentliche psychologische Betreuung, nicht alle Patienten nehmen an allen Sitzungen in weniger spezialisierten Zentren oder durch niedergelassene Kliniker teil	Kann die Intervention so in der Praxis angewandt werden?
C (omparator)	Keine pulmonale Rehabilitation, aber Nachfolgeuntersuchungen	Keine pulmonale Rehabilitation	Übertragbarkeit gegeben?
O (utcomes)	Lebensqualität gemessen mit mehreren, z.T. speziellen Instrumenten (CRQ, SF-36) Mortalität Hospitalisierungen Funktionale Kapazität, gemessen anhand der 6 Minuten Gehdistanz Studienressourcen	Lebensqualität gemessen mit generischen Instrumenten (z.B. SF-36) Mortalität Hospitalisierungen Funktionale Kapazität anhand von Patientenbefragung nach der bewältigten Gehstrecke im Alltag Systemressourcen	Sind die Instrumente tatsächlich Patientenrelevant und erfassen sie die relevanten Domänen? Selten indirekt Sind die Hospitalisierungen ähnlich (z.B. gleicher Aufwand) Spiegeln die Messungen den Patientenalltag wieder und sind sie direkt Wie wirkt sich eine Behandlung von einzelnen Patienten auf die jeweilige Jurisdiktion und den Patienten aus?

*alle Komponenten und Messungen können in der Versorgungsforschung und in einzelnen Studien identisch sein. Hier handelt es sich um ein illustratives Beispiel.



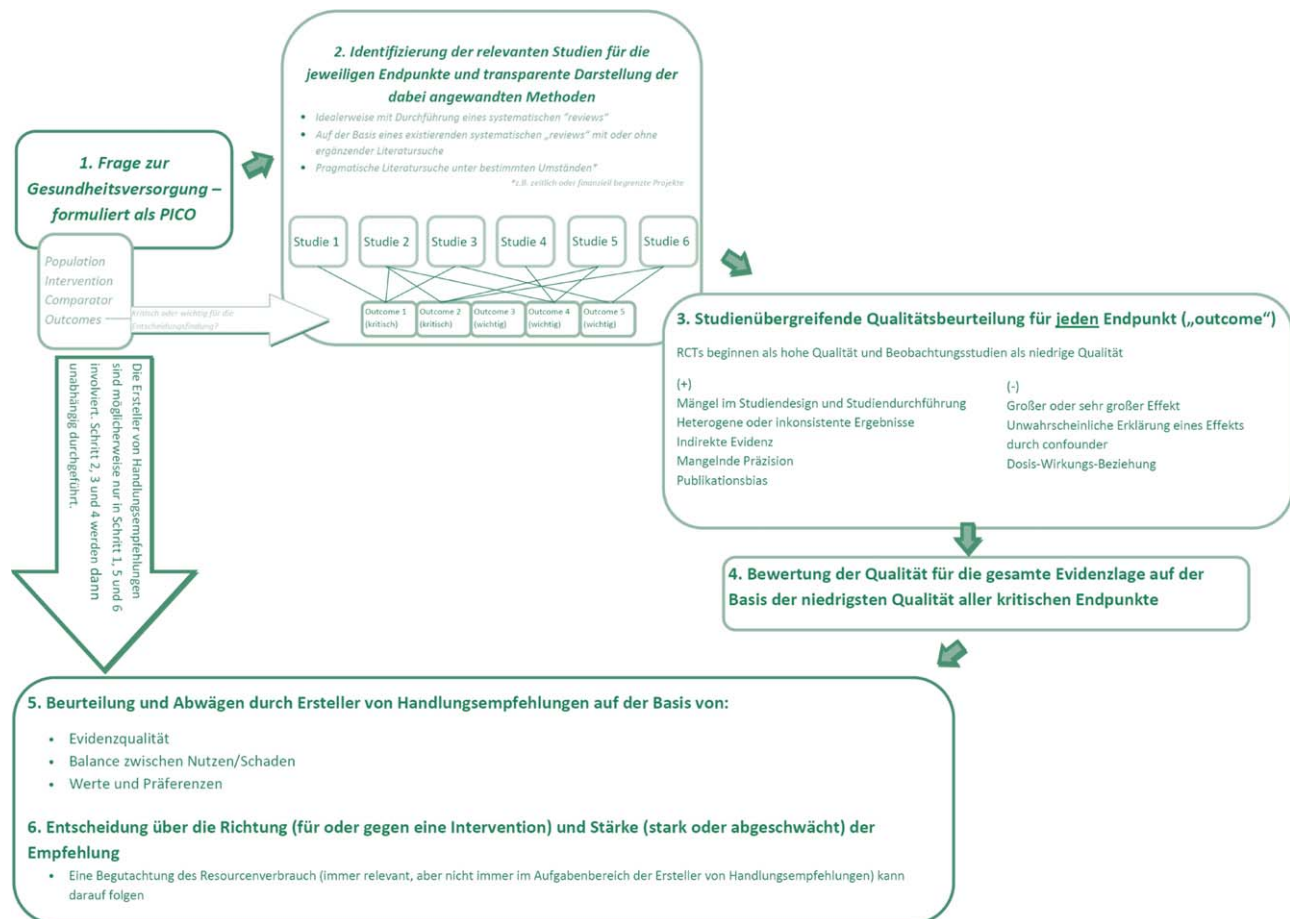


Abb. 1. Zeigt den GRADE Prozess zum Erstellen von Handlungsempfehlungen. Der Prozess wird im Text ausführlicher beschrieben - modifiziert nach Norris, S, GRADE Working Group.

renzierte Unterteilung hilfreich, nämlich, wie praktisch diese Studie nach der obigen Definition waere, d.h. wie genau sie auf die Entscheidungssituation der Patienten zutrifft. Tabelle 3 beschreibt ein Beispiel einer solchen Beurteilung für mechanistische Studien, die sich mit der Frage, ob pulmonale Rehabilitation effektiv ist, befassen, und pragmatische Studien, die sich zwar mit der Anwendbarkeit auf Systemebene befassen, aber letztlich doch nur der Information des Patienten-Klinikerin Gespanns dienen. Diese Studien sollten deshalb auch den Nettonutzen für einzelne Patienten darlegen damit ihre Praktikabilität und Übertragbarkeit bei der Entscheidungsfindung beurteilt werden kann. Diese Komponenten sollten auch beim Studiendesign berücksichtigt werden, um möglichst patientenrelevante Ergebnisse zu erhalten. Tabelle 3 beschreibt auch diese Übertragbarkeit mit dem

PICO Ansatz. Da sich Übertragbarkeit auf einem Kontinuum abspielt, wird leicht verständlich, dass nicht alle möglichen Situationen durch Forschungsergebnisse abgedeckt werden können und Beurteilungen fuer Einzelsituationen unumgänglich werden. Deshalb ist es wichtig, dass diese Beurteilungen von geschulten Wissenschaftlern und / oder Klinikern durchgeführt und transparent dargestellt werden. Es folgt auch, dass fehlende methodologische Schulung leicht ins Ungewisse bei der Beurteilung der Evidenz führen kann. Vielmehr ist die Beurteilung der Übertragbarkeit im GRADE System nur eine von mehreren Phasen bei der Evidenzbeurteilung. Die übrigen Phasen und Kriterien bei der Evidenzbeurteilung und ihre Integration beim Erstellen von Handlungsempfehlungen ist das Thema der folgenden Sektionen dieses Artikels.

Entwicklung von Handlungsempfehlungen nach GRADE

Die GRADE Arbeitsgruppe hat einen Ansatz herausgearbeitet, der angewendet werden kann, um Handlungsempfehlungen zu treffen. Abbildung 1 zeigt eine vereinfachte und schematische Darstellung dieses Prozesses. Der Leser wird auch auf andere Quellen mit detaillierter Beschreibung verwiesen [8]. Im Wesentlichen werden 4 Kriterien herangezogen. Eine Handlungsempfehlung nach GRADE bedarf der Beurteilung 1) der Evidenz, 2) der Größe des möglichen Nutzen und Schaden einer Intervention mit Fokussierung auf Patienten-relevante Endpunkte, 3) den damit verbundenen Wertvorstellungen, und 4) der Integration von Überlegungen zum Ressourcenverbrauch. Es wurde bereits erwähnt, dass eine Handlungsempfeh-

lung mit der richtigen Fragestellung beginnt (Abb. 1,1.), die alle wichtigen und kritischen Endpunkte beinhaltet. Kritische Endpunkte sind solche, für die entweder direkte oder indirekte Information unbedingt benötigt wird, um Handlungsempfehlungen zu erstellen. In der Abbildung sind 5 hypothetische Endpunkte ("outcomes") aufgeführt (Abb. 1,2.).

Literatursuche

Der Fragestellung sollte idealerweise eine systematische Übersichtsarbeit (SR) folgen, die der Identifizierung der Studien dient, die für die jeweiligen Endpunkte relevante Information liefern. Wenn die verfügbare Zeit oder die Ressourcen es nicht zulassen, kann ein pragmatischer Ansatz des Verwendens von existierenden SRs oder einer pragmatischen Literatursuche gewählt werden. Wert wird hierbei auf die transparente Darstellung der Literaturerfassung gelegt. Eine genaue Beschreibung der Methoden der Literaturrecherche und Zusammenfassung ist unverzichtbar. Abbildung 1 zeigt, dass in dem hypothetischen Beispiel 6 Studien die Einschlusskriterien erfüllen und dass für einen Endpunkt (outcome 3) keine Studie direkte Information liefert, was ebenfalls vermerkt werden sollte. Es folgt eine endpunktspezifische und studienübergreifende Beurteilung der Qualität der Evidenz für die einzelnen Endpunkte.

Qualitätsbeurteilung

Wie bereits erwähnt, unterscheidet GRADE zwischen vier Abstufungen der Qualität der Evidenz, obwohl die Qualität der Evidenz ein Kontinuum darstellt (Tabelle 4). Die Kategorisierung ist nötig, da eine Darstellung eines Kontinuums weniger verständlich ist. Expertenmeinung ist keine Evidenzstufe oder Evidenzqualität sondern eine Interpretation von vorliegender Evidenz, die von hoher oder sehr niedriger Qualität sein können (z.B. einzelne, nicht systematisch beschriebene Fälle). Die Beurteilung konzentriert sich auf die Beurteilung der Wahrscheinlichkeit, dass systematische Fehler vorliegen oder die Übertragbarkeit der Ergebnisse ungewiss ist. Tabelle 5 beschreibt die Faktoren anhand von Beispielen, die zur Beurteilung der Qualität der Evidenz herangezogen werden und in der Literatur weitgehend validiert oder beschrieben sind. Dies ist in Abbildung 1 als 3. Schritt dargestellt. Eine gesonderte Anmerkung zu randomisierten Studien in der Versorgungsforschung folgt hier.

Randomisierte Studien und die Versorgungsforschung

Da Randomisierung die sicherste Methode zur Vermeidung von systematischen Fehlern und "Confounding" darstellt, ist sie als Idealfall anzusehen, ob bei der Erforschung der Wirksamkeit einzelner Komponenten oder

in der Versorgungsforschung. Selbst wenn Randomisierung aus ethischen oder anderen Gründen nicht durchgeführt wird, sollte dieses Qualitätsmerkmal als Referenzstandard für die Beurteilung der Qualität herangezogen werden. Dieser Ansatz beruht auf der Erkenntnis, dass die vorliegende Evidenz oder Situation nicht die Erstellung des Referenzstandards rechtfertigt. Die bestmögliche Evidenz ("lowest risk of bias for a given situation") sollte hingegen als Standard gelten, wobei es sich bei der bestmöglichen Evidenz um randomisierte Studien handelt, die systematische Fehler unwahrscheinlicher machen. Das GRADE System stellt diesen Grundsatz explizit dar; es stellt aber auch klar, dass Evidenz aus nicht-randomisierten Studien unter besonderen Umständen als hohe Qualität eingestuft werden kann und dass diese Studien zu starken Empfehlungen führen können.

Bewertung der gesamten Evidenzlage für eine Fragestellung

Hat die Beurteilung der Qualität der Evidenz für alle kritischen Endpunkte mit den Kriterien der Tabelle 5 stattgefunden, bedarf es einer vorsichtigen Betrachtung der Gesamtqualität der Evidenz (Abb. 1,4.). Das GRADE System fordert, dass für eine integrative Beurteilung der zugrundeliegenden Qualität der Evidenz alle kritischen Endpunkte beachtet werden sollten. Dabei wird die Gesamtqualität auf der Basis der niedrigsten Qualität der einzelnen kritischen Endpunkte bewertet, um eine falsche Sicherheit bei der integrativen Beurteilung der Evidenz zu vermeiden.

Die Empfehlung

Die Schritte 5 und 6 in Abbildung 1 vervollständigen den Prozess des Erstellens von Handlungsempfehlungen. Nehmen wir erneut das schon in der vorherigen Arbeit beschriebenen Beispiels der Human Papilloma Virus (HPV) Impfung [9] und die Frage, ob sich 12 jährige Mädchen einer Impfung unterziehen sollten, um das Risiko der Gebärmutterhalskrebsmortalität und Warzenbildung in Anbetracht der mög-

Tabelle 4. Interpretation der verschiedenen Qualitätsstufen der Evidenz.

Evidenzstufe*	Definition
Hohe Qualität	Es ist sehr unwahrscheinlich, dass weitere Forschung das Vertrauen in den beobachteten Behandlungseffekt verändert.
Moderate Qualität	Weitere Forschung wird sich vermutlich erheblich auf unser Vertrauen in den beobachteten Behandlungseffekt auswirken. Möglicherweise ändert sich der Behandlungseffekt.
Niedrige Qualität	Weitere Forschung wird sich sehr wahrscheinlich auf unser Vertrauen in den beobachteten Behandlungseffekt auswirken. Wahrscheinlich ändert sich der Behandlungseffekt.
Sehr niedrige Qualität	Der beobachtete Behandlungseffekt ist mit sehr großer Unsicherheit behaftet.

*Expertenmeinung stellt kein Evidenzstufe dar, es handelt sich vielmehr um eine Interpretation von vorliegender Evidenz, die sowohl randomisierte Studien als auch beobachtende Studien (manchmal nur in Form von wenigen beobachteten Fällen) einschließt.

Tabelle 5. Kriterien für das Herabstufen der methodischen Qualität von RCTs und Beobachtungsstudien nach GRADE (adaptiert von [3,5,6]).

Kriterien	Erklärung	Beispiele
Herabstufen der Qualität		
Mängel im Studiendesign und Studiendurchführung	Fehlende Maskierung bei der Randomisierung; fehlende Verblindung; hohe Patientenverluste; unvollständige Nachbeobachtung, keine Intention-to-Treat (ITT)-Analyse; keine verblindete Endpunkterhebung, u.a.m.	Die Evidenz für einen Effekt von sublingualer Immunotherapie bei Kindern mit allergischer Rhinitis basiert auf einer randomisierten Studie mit fehlender Beschreibung der Randomisierung und Verdeckung der Randomisierungsfolge, keiner Verblindung und 21% der Kinder ohne Erfassung des Endpunktes ("lost to follow up"). Dieser schweren Mängel würden zu einer Herabstufung der Qualität führen.
Heterogene oder inkonsistente Ergebnisse	Große, nicht erklärte Unterschiede in den Behandlungseffekten der aufgefundenen Studien.	Beobachtende Studien, die den Effekt von nicht-steroidalen Antiphlogistika auf das Pankreaskarzinomrisiko untersuchen zeigen sehr unterschiedliche Ergebnisse, die nicht erklärlich sind [19].
Indirekte Evidenz	Indirekter Vergleich: Nur Placebo-kontrollierte (oder anders kontrollierte) Vergleiche zwischen 2 Interventionen, aber kein Kopf-an-Kopf-Vergleich; <i>aber auch</i> : Studienpopulation, Interventionen und Studienendpunkte der gefundenen Studien entsprechen nicht genau der PICO-Frage (siehe Tabelle 2 und 3).	<ol style="list-style-type: none"> 1) Für die Entscheidung zwischen oraler und intravenöser Gabe von Kortikosteroiden zur Behandlung des akuten Asthmas gibt es nur Studien, die die jeweilige Gabe mit Placebo oder keiner Gabe vergleichen. Randomisierte Studien für den direkten Vergleich gibt es nicht und deshalb ist eine sichere Aussage, welche Intervention effektiver, ist unsicher. 2) Thromboseprophylaxe im Krankenhaus: 2-mal versus 3-mal tägliche Gabe von Heparin. Obwohl viele Studien existieren, die 2 x oder 3 x tägliche Gabe mit Placebo verglichen haben, gibt es keinen direkten Vergleich. 3) Für die meisten kompetitiven Pharmaka der gleichen Klasse gibt es keine direkt vergleichenden Studien.
	Population	Siehe auch Tabelle 2 und 3 Oseltamivir zur Behandlung der Vogelgrippe durch Influenza A(H5N1) Virus: Randomisierte Studien mit Oseltamivir sind vorhanden, aber für die gewöhnliche Grippe, nicht für die Vogelgrippe.
	Intervention/Comparator	Sigmoidoskopie "Screening" zur Prävention von Maglinomen des Kolons: Randomisierte Studien, die das Screening nach okkultem Blut ("fecal occult blood screening") anwenden sind indirekt für die direkter Methode der Sigmoidoskopie
	"Outcome"	Thromboseerkennung: Erkennung von oberflächlichen Venenthrombosen mit weniger direkten Methoden als Surrogat für die Erkennung von symptomatischen tiefen Venenthrombose (mit direkteren Methoden oder anhand unzweideutiger klinischen Kriterien).
Fehlende Präzision	(Zu) breite Vertrauensintervalle verursacht durch kleine Patientenzahlen und/oder wenige Ereignisse in den Studien.	Beobachtende Studien, die den Effekt von exklusiver Fütterung von Säuglingen mit Muttermilch auf die Entwicklung von allergischer Rhinitis untersuchten, fanden ein relatives Risiko von 0,87 (95% Konfidenzintervall: 0,48–1,58). Die Resultate (auf der Basis des Konfidenzintervalls) schließen weder einen wichtigen Nutzen noch einen wichtigen Schaden aus.
Publikationsbias	Große Wahrscheinlichkeit für fehlende Publikationen mit negativen Studienergebnissen. Das Risiko ist erhöht bei Meta-Analysen aus kleinen und/oder Studien, die ausschließlich von Profit-orientierten Unternehmen finanziert werden.	RCTs über die Wirksamkeit von chinesischen Kräutern ("herbal medicine") zur Behandlung von Dysmenorrhoe im Vergleich zu konventioneller Therapie zeigen ungewöhnlich wenige Studien mit negativem oder kleinem Effekt der Schmerzreduzierung [20].

Tabelle 5. (Fortsetzung)

Kriterien	Erklärung	Beispiele
Hochstufen der Qualität		
Großer oder sehr großer Effekt	Hochwertige Beobachtungsstudien mit direkter Evidenz. Ein relatives Risiko/Odds Ratio von >2 oder <0.5 entspricht einem großen Effekt; ein relatives Risiko/Odds Ratio von >5 oder <0.2 entspricht einem sehr großen Effekt.	In Fall-Kontrollstudien zeigte sich durch das Tragen eines Fahrradhelms eine erhebliche Senkung der schweren Kopfverletzungen [21].
Unwahrscheinliche Erklärung eines Effekts durch confounder	Alle verbleibenden, plausiblen "Confounder" haben den beobachteten Effekt bereits reduziert oder einen abwesenden Effekt möglicherweise verstärkt.	1) Plausible Faktoren, die in Studien zum Vergleich von Mortalitätsraten von Profit- und nicht-profitorientierten Krankenhäusern nicht zur Adjustierung verwendet werden konnten, hätten den beobachteten Effekt bereits reduziert [22]. 2) Das Diabetes Medikament "Fenformin" verursacht "lactic acidosis". Ein verwandtes Medikament, Metformin, wurde verdächtigt die gleiche Nebenwirkung zu haben. Große Beobachtungsstudien haben das nicht gezeigt trotzdem Kliniker gegenüber dieser Nebenwirkung aber nach Einführung des Metformin beobachtet haben [23] alarmiert waren.
Dosis-Wirkungsbeziehung	Nachweis einer Dosis-Wirkungsbeziehung	1) Anstieg des Blutungsrisiko bei zunehmenden INR-Werten. 2) Anstieg des Malignomrisikos mit ansteigender Strahlendosis bei der prophylaktischen Schädelbestrahlung bei Leukämiepatienten .

lichen Nebenwirkungen (Impfreaktionen) zu senken. Abbildung 2 zeigt ein GRADE Evidenzprofil mit den Ergebnissen einer SR von Rambout et al. [15]. Die ersten 4 Endpunkte in diesem Beispiel ("cancer mortality", "cancer incidence", "dysplasia" und "Grade 2 cervical intraepithelial neoplasia") sind als sehr indirekt eingestuft und die Qualität für diese Endpunkte ist deshalb von hoch auf niedrig heruntergestuft worden. Der Effektschätzer suggeriert allerdings einen deutlichen Nutzen für diese Endpunkte. "Persistent HPV Infection at 12 months" ist als moderate Qualität und genitale Warzen als niedrige Qualität eingestuft, aber beide zeigen einen deutlich positiven Effekt. Schwere Nebenwirkungen treten auf, aber das Risiko ist in der tatsächlich behandelten Gruppe nicht signifikant höher als in der Placebo-gruppe (RR 1.07, 95% Konfidenzintervall 0.94 bis 1.22). Wenn wir nun die 4 Kriterien, die die Empfehlungstärke beeinflussen, heranziehen, muss man feststellen, dass zwar die Gesamtevidenz nur von niedriger Qualität ist,

aber der potenzielle, wenngleich nicht ganz sichere Nutzen deutlich grösser als der Schaden erscheint und die Wertvorstellungen der Patienten ähnlich sein würden. Die meisten Patientinnen würden einen sehr großen Wert auf die Verhinderung von Gebärmutterhalskrebs und genitalem Warzen legen und den unsicheren Nebenwirkungen einen relativ geringeren Wert beimessen. Damit könnte trotz der nur niedrigen Evidenzqualität eine starke Empfehlung für diese Patientengruppe getroffen werden. Wenn der Ressourcenverbrauch (Abb. 1,6.) zusätzlich herangezogen wird, kann die Empfehlungstärke möglicherweise vom Basisrisiko der Infektion abhängen. Eine solche Betrachtung bedarf der Integration von direkten Kosten (des Impfstoffes) also auch der Kosten, die durch weniger Therapien des Gebärmutterhalskrebses vermieden werden, und anderer Kosten, die entstehen können. Wenn die Nettokosten bei geringem Basisrisiko hoch wären, würde eine Empfehlung aber eher schwach ausfallen. Wenn die Nettokosten gerechtfertigt

scheinen, würde die Empfehlung dennoch stark ausfallen.

Anmerkung

Die Zusammenhänge für das Erstellen von Handlungsempfehlungen sind komplex und Modellierungen der vorliegenden Daten sind nötig, um Nutzen und Schaden abzuwiegen. Die Integration aller relevanten Komponenten in "decision analysis" bietet bislang leider keine vollständig zufriedenstellende Lösung, insbesondere wegen der fehlenden Daten und inkompletten Methodik für eine adäquate Beurteilung der Patientenwerte und Patientenpräferenzen. Dennoch müssen Entscheidungen getroffen werden und die explizite Beachtung und Beschreibung der angenommenen Größen, wie der Werte und Präferenzen, in Handlungsempfehlungen stellt einen Fortschritt dar [3,10,11,16,17]. Durch eine transparentere Darstellung der Annahmen kann aber die Implementierung von Handlungsempfehlungen durch die Patienten-Klinikerin Partnerschaft vereinfacht

		Quality assessment						Summary of findings			Quality	Importance
No of studies	Design	Limitations	Inconsistency	Indirectness	Imprecision	Other considerations	No of patients	Relative (95% CI)	Absolute	Quality	Importance	
5	Cancer Mortality (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	very serious ¹	no serious imprecision	none ²	142/18096 (0.8%)	OR 0.52 (0.43 to 0.63)	7259 fewer per 1,000,000 (from 5586 fewer to 8632 fewer)	6900 LOW	CRITICAL	
5	Cancer incidence (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	very serious ¹	no serious imprecision	none	142/18096 (0.8%)	OR 0.52 (0.43 to 0.63)	7259 fewer per 1,000,000 (from 5586 fewer to 8632 fewer)	6900 LOW	CRITICAL	
5	Dysplasia (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	very serious ¹	no serious imprecision	none	142/18096 (0.8%)	OR 0.52 (0.43 to 0.63)	7259 fewer per 1,000,000 (from 5586 fewer to 8632 fewer)	6900 LOW	CRITICAL	
5	GRADE 2 Cervical Intraepithelial Neoplasia (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	very serious ³	no serious imprecision	none	142/18096 (0.8%)	OR 0.52 (0.43 to 0.63)	7259 fewer per 1,000,000 (from 5586 fewer to 8632 fewer)	6900 LOW	CRITICAL	
2	Persistent HPV infection at 12 month (follow-up mean 12 months) randomised trials	no serious limitations	no serious inconsistency ⁴	no serious indirectness ⁵	no serious imprecision ⁶	reporting bias ⁷	12/3967 (0.3%)	OR 0.26 (0.16 to 0.41)	2 fewer per 1000 (from 1 fewer to 2 fewer)	6900 MODERATE	CRITICAL	
2	External genital lesions (warts) (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	no serious indirectness ⁵	serious ⁸	reporting bias ⁹	63/2512 (2.5%)	OR 0.13 (0.08 to 0.22)	22 fewer per 1000 (from 19 fewer to 23 fewer)	6900 LOW	CRITICAL	
6	Serious adverse effects (follow-up mean 5 years) randomised trials	no serious limitations	no serious inconsistency	serious ⁵	no serious imprecision	reporting bias ⁷	446/19823 (2.2%)	RR 1.07 (1.04 to 1.12)	2 more per 1000 (from 5 more to 224 more)	6900 LOW	CRITICAL	
							415/19786 (2.1%)		1 more per 1000 (from 4 more to 199 more)	6900 LOW	CRITICAL	
							2.1%		1 more per 1000 (from 5 more to 209 more)	6900 LOW	CRITICAL	

¹ The outcome was considered indirect given the duration of follow-up and the long term effect on mortality that was not evaluated.
² Publication bias was considered likely, but the panel did not downgrade for this factor because the large studies were consistent and the availability of a few studies with smaller effects would have not significantly altered the results.
³ There were important concerns about indirectness given the different population and the indirect outcome (CI12 or worse related to mortality).
⁴ There was inconsistency (I² of 68%), but it was not considered to be relevant as the larger studies were consistent in their results and contributed many more patients.
⁵ The panel considered the evidence indirect for the outcome, but decided to not downgrade for indirectness.
⁶ The OR was 0.26 (0.16 to 0.41).
⁷ Publication bias is likely.
⁸ There was important imprecision. The Confidence interval were wide and in the context of decision making of other outcomes considered too imprecise to make sufficiently certain estimates of the effect measure.
⁹ No explanation was provided.

Abb. 2. Zeigt ein GRADE Evidenzprofil zur HPV Impfung, das sowohl erwünschte als auch unerwünschte Konsequenzen einer Entscheidung oder Handlungsempfehlung darstellt. Die linke Seite der Abbildung zeigt sowohl die Studienlage als auch die Evidenzbeurteilung während die rechte Seite die numerischen Daten dieser Evidenz und eine integrative Evidenzbewertung beschreibt (Summary of Findings). Die Fußnoten ("footnotes") sind ein kritischer Bestandteil von Evidenzprofilen, da sie die Beurteilung transparenter machen. Das Beispiel ist illustrativ und sollte nicht zur tatsächlichen Entscheidungsfindung herangezogen werden. GRADE Evidenzprofile und Summary of Findings Tabellen können mit der GRADEpro Software hergestellt werden (<http://www.cc-ims.net/revman/gradepro>).

werden. Erfahrungen mit detaillierten Beurteilungen des Netto-Nutzen und der Anwendung von Werte-sensitiven Handlungsempfehlungen und eine partnerschaftliche Forschungsagenda könnte dieses Forschungsgebiet in den nächsten Jahren entscheidend voranbringen.

Zusammenfassung

Keine Forschungsstudie kann jede erdenkliche Situation beim Fällen von Entscheidungen im Gesundheitswesen direkt abdecken. Auf dieser Tatsache beruht die Erkenntnis, dass die Anwendung von pragmatischen und erklärenden Studien immer mit einer Restunsicherheit bezüglich der Übertragbarkeit der Ergebnisse verbunden ist. GRADE definiert die Qualität der Evidenz im Kontext der Entwicklung von Handlungsempfehlungen als ein Gradmesser für das Vertrauen in das Zutreffen eines ermittelten Effekts, der eine Handlungsempfehlung für bestimmte Populationen, Interventionen und Endpunkte unterstützt. Diese Definition beinhaltet die Beurteilung der Übertragbarkeit von Studienergebnissen explizit und stellt somit einen Lösungsansatz für die Versorgungsforschung dar. Die Anwendung von GRADE macht gesonderte Ansätze für die Beurteilung der Übertragbarkeit von Studienergebnissen für erklärende Studien und pragmatische Versorgungsstudien unnötig. Das Einbeziehen des Ansatzes der mechanistischen und praktischen Entwicklung und Interpretation von Studien leistet weitere Hilfestellung [18]. Der Beurteilung der Übertragbarkeit nach GRADE liegt zugrunde, dass es sich bei der Übertragbarkeit um ein Kontinuum handelt, das beinhaltet inwieweit die Populationen, Interventionen, Comparator und "outcomes" in einer Studie auf eine reale Fragestellung im Gesundheitswesen zutrifft. Das Schaffen der Transparenz dieser Beurteilungen, die als geringe Restunsicherheit eingestuft wird, wenn Studien dem Versorgungsalltag sehr nahe kommen und als pragmatische Studien angelegt waren, und als wesentliche Unsicherheit eingestuft werden kann, wenn die Studien erklärend oder mechanistisch angelegt

waren und dem Versorgungsalltag wenig entsprechen, ist eine entscheidender Beitrag des GRADE Systems. Ein neuer Ansatz und andere Qualitäts- und Designkriterien sind für die Versorgungsforschung aus der Sicht des Autors und der GRADE Arbeitsgruppe nicht vonnöten. Per Definitionem steht Wissenschaft aber nicht still. Durch die Kooperation und Annahme des GRADE Systems durch viele internationale Organisationen wird die explizite Weiterentwicklung des Systems gefördert. Das Konzept der offenen Zusammenarbeit und Mitarbeit in der internationalen GRADE Arbeitsgruppe und die enge Zusammenarbeit mit führenden Organisationen bietet hervorragende Voraussetzungen, um mit der konstanten Entwicklung im Bereich der Forschungsmethoden nicht nur Schritt zu halten, sondern wegweisend und integrierend tätig zu sein.

Finanzierung dieser Arbeit und Anmerkung zur Erstellung dieses Texts:

Die Finanzierung dieser Arbeit fand durch die McMaster University, Hamilton, Kanada durch reguläre Gehaltszahlungen an den Autor während der Erstellung des Textes statt. Dieses Manuskript beruht auf einem am 21.10.2008 in Berlin anlässlich des „Diskussionsforums zur Nutzenbewertung des GFR und IQWiG“ vom Autor gehaltenen Vortrages mit dem Thema „Übertragbarkeit von Studienergebnissen“.

GRADE wird in zunehmendem Masse in Deutschland populär und in seiner leitenden Funktion in der Arbeitsgruppe möchte der Autor zur Zusammenarbeit herzlich einladen, obwohl es keiner formalen Einladung bedarf.

Literatur

- [1] Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence

- and recommendations. *CMAJ* 2003; 169(7):677–80.
- [2] Global Programme on Evidence for Health Policy. Guidelines for WHO Guidelines. EIP/GPE/EQC/2003.1. World Health Organization, Geneva, 2003.
- [3] Kunz R, Burnand B, Schunemann HJ. The GRADE System. An international approach to standardize the graduation of evidence and recommendations in guidelines. *Internist (Berl)* 2008;49(6):673–80.
- [4] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coeillo P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336(7650):924–6.
- [5] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ. What is “quality of evidence” and why is it important to clinicians?. *BMJ* 2008;336(7651):995–8.
- [6] Schunemann H, Cook D, Jaeschke R, Vist G, Kunz R, Guyatt G. Grading Recommendations. In: Guyatt G MM, Cook D, Drummond R, editors. *Users’ Guide to the Medical Literature*. McGraw Hill, Chicago IL, 2008.
- [7] Schunemann H, Oxman AD, Higgins JPT, Vist GE, Glasziou P, Guyatt GH. Chapter 11: Presenting results and ‘Summary of findings’ tables. In: Higgins JPT GS, editor. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.1*. The Cochrane Collaboration, 2008, Available from www.cochrane-handbook.org, 2008.
- [8] Schunemann H, Brozek J, Oxman A. GRADE handbook for grading quality of evidence and strength of recommendation. Version 3.2 [updated March 2009]. The GRADE Working Group, 2009, Available from <http://www.cc-ims.net/gradepr>, 2009.
- [9] Schunemann H. Integrative Beurteilung der Evidenz im Gesundheitswesen: das GRADE System. *Z Evid Fortbild Qual Gesundheitswesen* 2009;103(5):261–268 (in press).
- [10] Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. *PLoS Med* 2007;4(5):e119.
- [11] Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. *Lancet Infect Dis* 2007;7(1):21–31.
- [12] Karanickolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH. A new “Mechanistic-Practical” Framework for designing and interpreting randomized trials. *J Clin Epidemiol* 2009;62:479–84.
- [13] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chron Dis* 1967;20:637–48.
- [14] Organization IfCSI-PN. Chronic obstructive pulmonary disease. 2001 Dec (revised 2005 Dec). 66.
- [15] Rambout L, Hopkins L, Hutton B, Ferguson D. Prophylactic vaccination against human papillomavirus infection and disease in women: a systematic review of randomized controlled trials. *CMAJ* 2007;177(5):469–79.
- [16] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. *BMJ* 2008;336(7652):1049–51.
- [17] Schunemann HJ, Jaeschke R, Cook DJ, Bria WF, El-Solh AA, Ernst A, et al. An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006; 174(5):605–14.
- [18] Karanickolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH. A new “mechanistic-practical” framework for designing and interpreting randomized trials. *J Clin Epidemiol* 2009;62(5):479–84.
- [19] Capurso G, Schunemann HJ, Terrenato I, Moretti A, Koch M, Muti P, et al. Meta-analysis: the use of non-steroidal anti-inflammatory drugs and pancreatic cancer risk for different exposure categories. *Aliment Pharmacol Ther* 2007;26(8): 1089–99.
- [20] Zhu X, Proctor M, Bensoussan A, Wu E, Smith CA. Chinese herbal medicine for primary dysmenorrhoea. *Cochrane Database of Systematic Reviews* 2008(2): Art. No.: CD005288. DOI:005210.001002/14651858.CD14005288.pub14651853.
- [21] Thompson D, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. *Cochrane Database Syst Rev* 2000; (CD001855).
- [22] Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schunemann 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(11):1399–406.
- [23] Salpeter S, Greyber E, Pasternak G, Salpeter E. Risk of fatal and nonfatal lactic acidosis with metformin use in type 2 diabetes mellitus. *Cochrane Database Syst Rev* 2002; CD002967(2).