Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany

Maren Dreier[1]
Birgit Borutta[1]
Jona Stahmeyer[1]
Christian Krauth[1]
Ulla Walter[1]

1 Institute of Epidemiology, Social Medicine and Health System Research, Hannover Medical School, Hannover, Germany

Abstract

Health care policy background

Findings from scientific studies form the basis for evidence-based health policy decisions.

Scientific background

Quality assessments to evaluate the credibility of study results are an essential part of health technology assessment reports and systematic reviews. Quality assessment tools (QAT) for assessing the study quality examine to what extent study results are systematically distorted by confounding or bias (internal validity). The tools can be divided into checklists, scales and component ratings.

Research questions

What QAT are available to assess the quality of interventional studies or studies in the field of health economics, how do they differ from each other and what conclusions can be drawn from these results for quality assessments?

Methods

A systematic search of relevant databases from 1988 onwards is done, supplemented by screening of the references, of the HTA reports of the German Agency for Health Technology Assessment (DAHTA) and an internet search. The selection of relevant literature, the data extraction and the quality assessment are carried out by two independent reviewers. The substantive elements of the QAT are extracted using a modified criteria list consisting of items and domains specific to randomized trials, observational studies, diagnostic studies, systematic reviews and health economic studies. Based on the number of covered items and domains, more and less comprehensive QAT are distinguished. In order to exchange experiences regarding problems in the practical application of tools, a workshop is hosted.

Results

A total of eight systematic methodological reviews is identified as well as 147 QAT: 15 for systematic reviews, 80 for randomized trials, 30 for observational studies, 17 for diagnostic studies and 22 for health economic studies. The tools vary considerably with regard to the content, the performance and quality of operationalisation. Some tools do not only include the items of internal validity but also the items of quality of reporting and external validity. No tool covers all elements or domains. Design-specific generic tools are presented, which cover most of the content criteria.

Discussion

The evaluation of QAT by using content criteria is difficult, because there is no scientific consensus on the necessary elements of internal validity, and not all of the generally accepted elements are based on empirical evidence. Comparing QAT with regard to contents neglects the operationalisation of the respective parameters, for which the quality and precision are important for transparency, replicability, the correct assessment and interrater reliability. QAT, which mix items on the quality of reporting and internal validity, should be avoided.

Conclusions

There are different, design-specific tools available which can be preferred for quality assessment, because of its wider coverage of substantive elements of internal validity. To minimise the subjectivity of the assessment, tools with a detailed and precise operationalisation of the individual elements should be applied. For health economic studies, tools should be developed and complemented with instructions, which define the appropriateness of the criteria. Further research is needed to identify study characteristics that influence the internal validity of studies.

Keywords: quality assessment, assessment quality, quality assessment tools, assessment tools, study quality, study assessment, clinical trials, evaluation criteria, methodologic quality, validity, quality, science, risk of bias, bias, confounding, systematic reviews, health technology assessment, HTA, health economics, health economic studies, critical appraisal, quality appraisal, checklists, scales, component ratings, components, tool, studies, interventional studies, observational studies, diagnostic studies, item, meta-analysis, QAT, EBM, evidence-based medicine, standard, epidemiology

Summary

1. Health political background

Healthcare policy decisions should be based on the best available scientific evidence. Scientific evidence is based on the synthesis of study results, which are if possible unbiased and thus have a high credibility.

2. Scientific background

Quality assessments to evaluate the credibility of studies is an inherent component of HTA reports (HTA = Health Technology Assessment) and systematic reviews. There are various quality assessment tools (QAT) that rate the extent of systematic distortion in study results by confounding or bias (internal validity).
There is no gold standard for assessing the study quality, since the true associations of exposures/interventions and outcomes are unknown. The existing tools for assessing study quality can be classified into scales, checklists and component ratings. In a scale, each item receives a numerical rating that will be added to a sum score. Scales are no longer recommended, because they do not reflect the correct extent of validity. A checklist consists of at least two items without a numerical rating system. The component rating includes components like “randomization” and “blinding”, which are also not evaluated numerically, but qualitatively.
In this report methodological quality that is used synonymously with the expression study quality and must be distinguished from the reporting quality, which is not part of this report.
The quality of health economic studies is determined by (a) the validity of study results, (b) the compliance with methodological standards of health economic evaluation and (c) the access to appropriate cost data. The methodological standards of health economic evaluations are described in health economic literature and international guidelines for providing health economic evaluations. Health economic evaluations are based on the theoretical concepts of welfare economics and decision analysis. The standards of economic evaluation have reached a broad consensus regarding the constitutive elements of health economic evaluation and approaches to cost analysis and outcome determination. Nevertheless, some guidelines recommend different approaches to be used. The elements of health economic evaluation contain (1) the justification and the choice of the evaluation type, (2) the identification and the selection of comparators, (3) the perspective, (4) the identification of resource use and costs, (5) the identification of all relevant effects and

benefits, (6) the declaration of the time horizon, (7) modelling, (8) discounting, (9) incremental analysis, (10) uncertainty analysis.

3. Research questions

What QAT are available to assess the quality of systematic reviews/HTA reports, intervention studies, observational studies, diagnostic studies and health economic studies, how do they differ among each other and what conclusions can be drawn from these results for quality assessments?

4. Methods

A systematic search of relevant electronic databases from 1988 onwards is done to identify QAT, supplemented by screening of the references of the HTA reports of the German Agency for Health Technology Assessment (DAHTA) and in addition an internet search. Formal characteristics and substantive elements of the tools are extracted. The substantive elements of the QAT are extracted specific to systematic reviews, intervention studies, observational studies, diagnostic studies, and health economic studies. The literature search, the data extraction and the quality assessment are carried out independently by two reviewers. Different ratings of the reviewers are solved by consensus.

The content of tools for the quality assessment of systematic reviews, intervention studies, observational studies, and diagnostic studies is extracted by using modified criteria lists. The elements of the lists are made up of study characteristics, which have either empirically demonstrated evidence of an effect on the level of the study results or its distorting effect on study results is generally accepted. The elements for study characteristics of systematic reviews, intervention studies, and observational studies are summarised in several domains. Out of all elements, those elements with empirical evidence as a potential source of bias or elements being classified on a theoretical basis as essential for internal validity are defined as relevant elements.

In order to provide a basis for the selection of a tool, only generic tools and their elements of internal validity are considered. Furthermore, the presence of sufficient operationalisation is required. The tools are distinguished by the total number of covered elements, covered relevant elements, and covered domains. Tabular summaries of the results are prepared for each study design and the results across the QAT are assessed qualitatively to identify more and less comprehensive tools.

For the data extraction of the basic elements of health economic studies, a form is developed, because there are no systematic reviews that can provide a basis for the data extraction. In the first step of the development process health economic literature and current national and international guidelines for creating health and pharmacoeconomic studies are screened. Literature and guidelines address mainly similar topics (elements of health economic evaluation). In the second step, the key elements are worked out to investigate the relation to study quality (internal validity) of health economic studies. Domains and items are developed based on the elements of health economic evaluations adapted from literature and guidelines. Domains and items are transferred into a form for analysing the quality assessment tools for health economic evaluation studies. This form helps to extract the various tools. In the development of domains and items, effort is made to ensure that items relate primarily to the internal validity.

In the health economic extraction form a gradation for rating the different items is made as such: “appropriate”, “justified”, “reported” and “missing”. If a quality assessment tool asks for a special item addressed in a study, a rating with “reported” is made (e. g. perspective of analysis, outcome parameter or discount rate). An item is rated with “justified”, the quality assessment tool asks for the rationale for choosing a special specification. The rating “appropriate” is assigned when the quality assessment tool asks for the adequacy of used methodology in an item.

In order to find out about problems in the practical application of tools, a workshop is conducted. Objectives of the workshop are to exchange and discuss user experiences with quality assessment tools for intervention studies, requirements, and content of tools on the quality of intervention studies. These discussions will examine practical issues that are rarely discussed in the literature. A consensus on individual aspects is not pursued. The target audience include authors of the German HTA reports and systematic reviews of the German Institute for Medical Documentation and Information (DIMDI) and the Institute for Quality and Efficiency in Health Care (IQWiG), experts in the field of methodology, researchers (from the disciplines of medicine conducting public health, epidemiology, prevention, health economics), involved in healthcare policy-relevant evaluations, as well as institutes/associations conducting systematic reviews. Topics are introduced by presentations of invited experts followed by moderated discussions. Presentations and discussions are documented by audio recordings and transcriptions.

5. Results

The extensive literature search yields a total of 147 tools to assess the study quality: 15 for systematic reviews/HTA reports/meta-analysis, 80 for intervention studies, 30 for observational studies, 17 for diagnostic studies and 22 for health economic studies. Among the QAT are 16 tools that can be used both for intervention and observational studies.

An initial screening of HTA reports in the DAHTA database indicates that a quality assessment was reported in 87% of the identified documents. However, in only half of these reports the chosen QAT was mentioned.

The tools show a wide variation of the formal and content characteristics. Some tools contain not only items of internal validity, but also of reporting quality and external

validity. Design-specific generic tools for the assessment of systematic reviews/HTA reports/meta-analysis, intervention studies, observational studies and diagnostic studies are identified, which cover most elements for internal validity, most of the domains with at least one, or 50% of the contained elements as well as the most relevant elements. More and less comprehensive tools can be distinguished.

The tools that examine the quality of health economic studies also reveal significant differences both in the consideration of various topics, as well as in the assessment of quality. In addition, substantial differences exist in the operationalisation of the items. Across all study designs, none of the included tools meet all elements.

A total of 27 people from HTA and EBM-associated (EBM = evidence-based medicine) institutions take part in the workshop. The following discussion points are suggested by the participants: the external validity as a part of assessment tools, the subjectivity of the assessment process, dealing with low reporting quality, endpoint versus study related quality assessment and incorporation of the results of the quality assessment. As consensus at the workshop is not intended, individual opinions are presented. External and internal validity should be assessed separately from each other. Items, which leave much room for subjective ratings, lead to a lack of inter-rater reliability and result in a high need for discussions. This can be avoided by a precise and detailed operationalisation of the items.

6. Discussion

The quality of studies can be defined in various ways. It is a dominating view that an assessment of study quality can either express the level of internal validity or the possibility of distortion. However, the inventory of the numerous identified tools shows that many of them include the assessment of reporting quality. Mixing the reporting quality and the internal validity can lead to a misinterpretation of the study quality, if the elements of the reporting quality are used as a surrogate for assessing the methodological quality.

Based on the tabular presentation of covered content items, the identified QAT can be compared. However, this approach has limitations, since there is no scientific consensus on the necessary elements of the internal validity, and not all of the generally accepted elements are based on empirical evidence. Therefore, the highest possible number of covered elements is not necessarily an indication of an appropriate tool.

For further differentiation of the QAT, the number of covered relevant elements is presented. While for relevant elements of intervention and diagnostic studies only evidence based elements affecting the internal validity are selected, this is true for only some of the relevant elements of observational studies and systematic reviews. Overall, the performance of relevant elements should be used cautiously to identify tools that are more or less comprehensive. Depending on the topic, it should be examined, whether all items of a chosen tool are relevant, and whether additional quality items should be assessed as a part of the assessment.

Some elements of QAT cannot be clearly assigned to the reporting quality, the internal or external validity. For example, the calculation of the required sample size is only associated with the precision of the results without affecting the size of the effect estimator. However, the precision of the effect estimates may affect the significance of the results.

Not all the tools ever used have been found. However, the possibility of having missed important and frequently applied tools is low, since different data sources including the internet were screened.

In general, the higher the scope for subjective assessments, the lower the agreement between the reviewers is. Therefore, every item of a tool should be operationalised as detailed and precisely as possible. Where necessary, the instructions can be adjusted to ensure that all reviewers are clear on how to score study quality. About 40% of the included tools provide more detailed guidance for assessment.

The quality assessment of health economic studies is an essential part of creating HTA reports. A total of 22 health economic QAT is identified. There are considerable differences regarding:

- the number of included items of the health economic extraction form (elements of health economic evaluation)
- the assessment quality: appropriate – justified – reported
- the diversity of quality sampling

None of the analysed QAT covers the whole range of relevant themes (elements of health economic evaluation). Only few consider most domains of the extraction form. Only three tools check the adequacy of the methodological procedures. Many tools ask for the methodological adequacy in few items. None of the QAT defines what is meant with adequacy. Most tools demand a justification for the methodological procedures or analyse, which items are reported.

Significant differences also exist in the sophistication of the quality assessment. The question how differentiated an assessment tool discusses the different elements of health economic evaluation can be answered by the number of items in a QAT. Because a tool is based only on few items, questions have to be more generally introduced. Reviewers will have a considerable scope for interpretation. For extensive tools with a great number of items, they can be operationalised to be more specific, so the scope for interpretation will be significantly reduced and more objective assessments are supported.

7. Conclusions

The quality assessment of studies is a mandatory part of systematic reviews, and has to be documented transparently. There are different, design-specific QAT available

that can be selected according to their substantive coverage of the elements of internal validity.
There is consensus that scales should not be used for quality assessments or should be used without quantitative assessment. To minimise the subjectivity of the evaluation, tools with a detailed and precise operationalisation of the items are preferable. If possible, the chosen tool should be tested in a few studies in advance to check if the operationalisation of the items needs to be supplemented or clarified to minimise the subjectivity of the evaluation and to ensure uniform scoring of all reviewers. Further research is needed to identify study characteristics that influence the internal validity of studies, especially for observational studies. So far, there is no evidence that qualitative overall assessment of study quality is correctly associated with the internal validity.
For assessing the quality of health economic studies, tools should be developed, which (1) cover all relevant elements of health economic evaluation, (2) assess the appropriate use of methodological procedures and (3) differentiate the various topics sufficiently. The adequacy should be based on the standards of health economic evaluation (defined by standard literature and international guidelines). Advice for filling in and operationalisations should be part of the assessment tools and, in addition, adequacy should be accurately described and defined.

Corresponding author:
Dr. med. Maren Dreier, MPH
Institute of Epidemiology, Social Medicine and Health System Research, Hannover Medical School,
Carl-Neuberg-Str. 1, 30625 Hannover, Germany, Phone: +49(0)511/5322192
Dreier.Maren@mh-hannover.de

Please cite as
Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. GMS Health Technol Assess. 2010;6:Doc07.
DOI: 10.3205/hta000085, URN: urn:nbn:de:0183-hta0000855

This article is freely available from
http://www.egms.de/en/journals/hta/2010-6/hta000085.shtml

Published: *2010-06-14*

The complete HTA Report in German language can be found online at: http://portal.dimdi.de/de/hta/hta_berichte/hta260_bericht_de.pdf

Vergleich von Bewertungsinstrumenten für die Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum

Maren Dreier[1]
Birgit Borutta[1]
Jona Stahmeyer[1]
Christian Krauth[1]
Ulla Walter[1]

1 Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, Medizinische Hochschule Hannover, Deutschland

Zusammenfassung

Gesundheitspolitischer Hintergrund

Erkenntnisse aus wissenschaftlichen Studien bilden die Grundlage für evidenzbasierte gesundheitspolitische Entscheidungen.

Wissenschaftlicher Hintergrund

Zur Einschätzung der Glaubwürdigkeit von Studien sind Qualitätsbewertungen von Studien immanenter Bestandteil von HTA-Berichten (HTA = Health Technology Assessment) und systematischen Übersichtsarbeiten. Diese prüfen, inwieweit die Studienergebnisse systematisch durch Confounding oder Bias verzerrt sein können (interne Validität). Es werden Checklisten, Skalen und Komponentenbewertungen unterschieden.

Forschungsfragen

Welche Instrumente zur Qualitätsbewertung von systematischen Reviews, Interventions-, Beobachtungs-, Diagnose- und gesundheitsökonomischen Studien gibt es, wie unterscheiden sich diese und welche Schlussfolgerungen lassen sich daraus für die Qualitätsbewertung ableiten?

Methodik

Es wird eine systematische Recherche in einschlägigen Datenbanken ab 1988 durchgeführt, ergänzt um eine Durchsicht der Referenzen, der HTA-Berichte der Deutschen Agentur für Health Technology Assessment (DAHTA) sowie eine Internetrecherche. Die Literaturauswahl, die Datenextraktion und die Qualitätsbewertung werden von zwei unabhängigen Reviewern vorgenommen. Die inhaltlichen Elemente der Qualitätsbewertungsinstrumente (QBI) werden mit modifizierten Kriterienlisten, bestehend aus Items und Domänen spezifisch für randomisierte, Beobachtungs-, Diagnosestudien, systematische Übersichtsarbeiten und gesundheitsökonomische Studien extrahiert. Anhand der Anzahl abgedeckter Items und Domänen werden umfassendere von weniger umfassenden Instrumenten unterschieden. Zwecks Erfahrungsaustausch zu Problemen bei der praktischen Anwendung von Instrumenten wird ein Workshop durchgeführt.

Ergebnisse

Es werden insgesamt acht systematische, methodische Reviews und HTA-Berichte sowie 147 Instrumente identifiziert: 15 für systematische Übersichtsarbeiten, 80 für randomisierte Studien, 30 für Beobachtungs-, 17 für Diagnose- und 22 für gesundheitsökonomische Studien. Die Instrumente variieren deutlich hinsichtlich der Inhalte, deren Ausprägung und der Güte der Operationalisierung. Einige Instrumente enthalten neben Items zur internen Validität auch Items zur Berichtsqualität und zur externen Validität. Kein Instrument deckt alle abgefragten Kriterien

ab. Designspezifisch werden generische Instrumente dargestellt, die die meisten inhaltlichen Kriterien erfüllen.

Diskussion

Die Bewertung von QBI anhand inhaltlicher Kriterien ist schwierig, da kein wissenschaftlicher Konsens über notwendige Elemente der internen Validität bzw. nur für einen Teil der allgemein akzeptierten Elemente ein empirischer Nachweis besteht. Der Vergleich anhand inhaltlicher Parameter vernachlässigt die Operationalisierung der einzelnen Items, deren Güte und Präzision wichtig für Transparenz, Replizierbarkeit, die korrekte Bewertung sowie die Interrater-Reliabilität ist. QBI, die Items zur Berichtsqualität und zur internen Validität vermischen, sind zu vermeiden.

Schlussfolgerungen

Es stehen unterschiedliche, designspezifische Instrumente zur Verfügung, die aufgrund ihrer umfassenderen inhaltlichen Abdeckung von Elementen der internen Validität bevorzugt zur Qualitätsbewertung eingesetzt werden können. Zur Minimierung der Subjektivität der Bewertung sind Instrumente mit einer ausführlichen und präzisen Operationalisierung der einzelnen Elemente anzuwenden. Für gesundheitsökonomische Studien sollten Instrumente mit Ausfüllhinweisen entwickelt werden, die die Angemessenheit der Kriterien definieren. Weitere Forschung ist erforderlich, um Studiencharakteristika zu identifizieren, die die interne Validität von Studien beeinflussen.

Schlüsselwörter: Validität von Ergebnissen, Bias (Epidemiologie), Verzerrung, statistische (Epidemiologie), Verzerrung, systematische (Epidemiologie), Methoden, epidemiologische, Methodik, Evaluations-, Studienqualität, Studienbewertung, klinische Studien, Bewerungskriterien, Qualitätsbewertunginstrument, Qualitätsbewertung, Verzerrungspotenzial, Verzerrung, Bias, Validität, Confounding, Health Technology Assessment, HTA, systematische Übersichtsarbeiten, Gesundheitsökonomie, gesundheitsökonomische Studien, Checklisten, Skalen, Komponentenbewertung, Komponenten, Instrument, Studien, Interventionsstudien, Beobachtungsstudien, Diagnosestudien, Item, Metaanalyse-, Metaanalyse, Meta-Analyse, QBI, EBM, evidenzbasierte Medizin, Bewertungsqualität, Bewertungsinstrumente, Standard, Qualität, Wissenschaft, methodische Qualität, Epidemiologie

Kurzfassung

1. Gesundheitspolitischer Hintergrund

Gesundheitspolitische Entscheidungen sollen evidenzbasiert auf der Grundlage von wissenschaftlichen Erkenntnissen getroffen werden. Evidenz basiert auf der Synthese von Studienergebnissen, die möglichst unverzerrt sind und damit eine hohe Glaubwürdigkeit aufweisen.

2. Wissenschaftlicher Hintergrund

Zur Einschätzung der Glaubwürdigkeit von Studien sind Qualitätsbewertungen immanenter Bestandteil von HTA-Berichten (HTA = Health Technology Assessment) und systematischen Übersichtsarbeiten. Diese prüfen, inwieweit die Studienergebnisse systematisch durch Confounding oder Bias verzerrt sein können (interne Validität).
Es gibt keinen Goldstandard für die Bewertung der Studienqualität, da die wahren Zusammenhänge von Exposition/Intervention und Outcome unbekannt sind. Die eingesetzten Instrumente können als Skalen, Checklisten und Komponentenbewertungen klassifiziert werden. Bei einer Skala erhält jedes Item eine numerische Bewertung, die zu einem Summenscore addiert wird. Skalen werden nicht mehr empfohlen, da sie die Höhe der Validität nicht korrekt abbilden. Eine Checkliste besteht aus mindestens zwei Items ohne numerisches Bewertungssystem. Die Komponentenbewertung enthält als Items Komponenten wie „Randomisierung“ und „Verblindung“, die ebenfalls nicht numerisch, sondern qualitativ bewertet werden. Von der methodischen Qualität, die in diesem Bericht

synonym zum Begriff Studienqualität verwendet wird, muss die Berichtsqualität abgegrenzt werden, die nicht Bestandteil dieses Berichts ist.
Die Qualität gesundheitsökonomischer Studien wird bestimmt durch (a) die Validität der Studienergebnisse, (b) die Einhaltung methodischer Standards der gesundheitsökonomischen Evaluation und (c) den Zugang zu belastbaren Kosten- und Outcomedaten. Die methodischen Standards der gesundheitsökonomischen Evaluation sind in Standardlehrbüchern und gesundheitsökonomischen Leitlinien beschrieben. Gesundheitsökonomische Evaluation basiert auf den theoretischen Konzepten der Wohlfahrtsökonomik und Entscheidungsanalyse. Bei den Standards der gesundheitsökonomischen Evaluation hat sich ein Konsens über konstitutive Elemente der gesundheitsökonomischen Evaluation und über zulässige Ansätze der Kostenanalyse und Outcomebestimmung herausgebildet. Teilweise wird in Leitlinien explizit gefordert, alternative Ansätze zu kalkulieren. Die Elemente der gesundheitsökonomischen Evaluation umfassen (1) die begründete Auswahl der Studienform, (2) die Identifizierung und Festlegung der Vergleichsalternativen, (3) die Perspektive der Evaluation, (4) die Bestimmung von Ressourcenkonsum und Kosten, (5) die Identifizierung und Bestimmung der relevanten Effekte und Nutzen, (6) die Festlegung des Zeithorizonts, (7) die Modellierung, (8) die Diskontierung, (9) die Inkremental- und (10) die Unsicherheitsanalyse.

3. Fragestellung

Welche Instrumente zur Qualitätsbewertung von systematischen Übersichtsarbeiten, Interventions-, Beobachtungs-, Diagnose- und gesundheitsökonomischen Studien gibt es, wie unterscheiden sich diese und welche Schlussfolgerungen lassen sich daraus für die Qualitätsbewertung ableiten?

4. Methodik

Zur Identifikation von Instrumenten wird eine systematische Recherche in einschlägigen Datenbanken ab 1988 durchgeführt, ergänzt um eine Durchsicht der Referenzen, der HTA-Berichte der Deutschen Agentur für Health Technology Assessment (DAHTA) sowie eine Internetrecherche. Es werden formale Charakteristika und inhaltliche Elemente der Instrumente extrahiert. Die inhaltliche Datenextraktion wird spezifisch für Interventions-, Beobachtungs-, Diagnosestudien, systematische Übersichtsarbeiten und gesundheitsökonomische Studien durchgeführt. Die Literaturauswahl, die Datenextraktion und die Qualitätsbewertung werden jeweils von zwei unabhängigen Reviewern vorgenommen, bei Diskrepanzen erfolgt eine Konsensentscheidung.
Die Inhalte von Instrumenten zur Bewertung von randomisierten Interventions-, Beobachtungs-, Diagnosestudien und systematischen Übersichtsarbeiten werden anhand von modifizierten Kriterienlisten extrahiert. Die Elemente der Listen setzen sich aus Studiencharakteristika zusammen, für die entweder empirisch ein Einfluss auf die Höhe der Studienergebnisse nachgewiesen oder deren Einfluss allgemein akzeptiert bzw. theoretisch fundiert ist. Die Elemente für Studiencharakteristika von Interventions-, Beobachtungsstudien und systematischen Übersichtsarbeiten werden in mehrere Domänen zusammengefasst. Von den Elementen werden diejenigen als relevant definiert, für die empirische Evidenz als potenzielle Biasquelle besteht bzw. die von anderen Autoren als essenziell eingestuft werden.
Als Basis für die Auswahl eines Instruments zur Qualitätsbewertung werden designspezifisch nur generische Instrumente und ihre Elemente der internen Validität betrachtet. Außerdem wird das Vorhandensein von Ausfüllhinweisen berücksichtigt. Anhand der Anzahl abgedeckter Elemente insgesamt, abgedeckter relevanter Elemente sowie abgedeckter Domänen werden umfassendere von weniger umfassenden Instrumenten unterschieden.
Für die Datenextraktion der inhaltlichen Elemente für gesundheitsökonomische Studien wird ein Formular entwickelt, da keine Übersichtsarbeiten vorliegen, die als Referenz dienen können. Im ersten Schritt des Entwicklungsprozesses werden Standardlehrbücher sowie aktuelle nationale und internationale Leitlinien zur Erstellung gesundheits- und pharmakoökonomischer Studien gesichtet. Inhaltlich sprechen die Lehrbücher und Leitlinien weitgehend identische Themenschwerpunkte an (Elemente der gesundheitsökonomischen Evaluation). In einem zweiten Schritt werden die herausgearbeiteten Themenschwerpunkte auf den Bezug zur Studienqualität (interne Validität) gesundheitsökonomischer Studien untersucht. Es werden Domänen und Items entwickelt, die auf den Themenschwerpunkten der Lehrbücher und Leitlinien basieren. Sie werden in ein Formular zur Extraktion von gesundheitsökonomischen Qualitätsbewertungsinstrumenten (QBI) überführt, mit dessen Hilfe die verschiedenen Bewertungsinstrumente extrahiert werden. Bei der Entwicklung der Domänen und Items wird darauf geachtet, dass sich diese primär auf die interne Validität beziehen.
Im gesundheitsökonomischen Extraktionsformular wird für die Bewertung der Items der berücksichtigten QBI eine Abstufung vorgenommen: „angemessen“, „begründet“, „berichtet“ und „fehlend“. Eine Bewertung „berichtet“ wird vergeben, wenn ein QBI lediglich abfragt, ob ein Item in einer gesundheitsökonomischen Studie berichtet wird (z. B. Perspektive der Analyse, einbezogene Outcomeparameter oder Diskontierungsrate). Die Beurteilung „begründet“ bedeutet, dass das QBI explizit nach Begründungen für die Ausprägung des Items fragt. Die Bewertung „angemessen“ heißt, dass ein Instrument eine Überprüfung der Angemessenheit des Items fordert. Die Überprüfung der Angemessenheit sollte an den Standards der gesundheitsökonomischen Evaluation orientiert sein.
Zwecks Erfahrungsaustausch zu Problemen bei der praktischen Anwendung von Instrumenten wird ein Workshop durchgeführt. Ziele des Workshops sind der Austausch und die Diskussion der Erfahrungen sowie des

Umgangs mit Bewertungsinstrumenten zur Qualität von randomisierten und nicht-randomisierten Interventionsstudien, Anforderungen sowie Inhalte an/von Bewertungsinstrumente/n zur Qualität von Interventionsstudien. Der Austausch dient zur Ergänzung von wissenschaftlichen Untersuchungen um praktische Aspekte, deren Stellenwert in Publikationen oft nicht thematisiert wird. Eine Konsensbildung zu einzelnen Aspekten wird nicht angestrebt. Zielgruppe des Workshops sind Autoren von deutschsprachigen HTA-Berichten oder systematischen Reviews des Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI) und des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Experten auf dem Gebiet der Methodik, Wissenschaftler (aus den Disziplinen Medizin, Public Health, Epidemiologie, Prävention, Gesundheitsökonomie), die mit gesundheitspolitisch relevanten Evaluationen befasst sind, sowie Institute/Verbände, die systematische Reviews mit Qualitätsbewertung durchführen. Referenten werden mit ihren Vorträgen die entsprechenden Themen einleiten. Im Anschluss an die Vorträge sind jeweils 20 bis 30 Minuten für eine moderierte Diskussion vorgesehen. Zur Dokumentation wird u. a. eine Audio-Aufzeichnung mit anschließender Transkription durchgeführt.

5. Ergebnisse

Die umfassende Recherche ergibt insgesamt 147 Instrumente zur Bewertung der Studienqualität: 15 für systematische Reviews/HTA-Berichte/Metaanalysen, 80 für Interventions-, 30 für Beobachtungs-, 17 für Diagnose- und 22 für gesundheitsökonomische Studien. Unter den QBI sind 16 Instrumente, die sowohl für Interventions- als auch für Beobachtungsstudien eingesetzt werden können.

Ein initiales Screening von HTA-Berichten in der DAHTA-Datenbank zeigt, dass in 87% der Berichte die Durchführung einer Qualitätsbewertung angegeben wird. Von diesen wird ist jedoch nur bei der Hälfte das verwendete QBI dokumentiert.

Die identifizierten Instrumente weisen eine große Variation hinsichtlich der formalen und inhaltlichen Charakteristika auf. Einige Instrumente enthalten neben Items zur internen Validität auch welche zur Berichtsqualität und zur externen Validität. Designspezifisch werden generische Instrumente für die Bewertung von systematischen Reviews/HTA-Berichten/Metaanalysen, Interventions-, Beobachtungs- und Diagnosestudien ermittelt, die die meisten Elemente zur internen Validität, die meisten Domänen mit mindestens einem bzw. 50% der enthaltenen Elemente sowie die meisten als relevant definierten Elemente abdecken. Es können umfassendere von weniger umfassenden Instrumenten unterschieden werden.

Die Instrumente, die die Qualität gesundheitsökonomischer Studien untersuchen, weisen ebenfalls erhebliche Unterschiede auf sowohl in der Betrachtung der verschiedenen Themenbereiche, als auch in der Bewertung der Qualität. Zudem bestehen beträchtliche Differenzen in den Operationalisierungen. Über alle Studiendesigns hinweg erfüllt keines der eingeschlossenen Instrumente alle Bereiche.

Am Workshop nehmen insgesamt 27 Personen aus HTA- und EbM-assoziierten (EbM = Evidenzbasierte Medizin) Institutionen teil. Folgende Diskussionspunkte werden von den Teilnehmern vorgeschlagen: externe Validität als Bestandteil von Bewertungsinstrumenten, Subjektivität der Bewertung, Umgang mit geringer Berichtsqualität, endpunkt- statt studienbezogene Qualitätsbewertung und Integration der Ergebnisse der Bewertung. Eine Konsensbildung ist im Rahmen des Workshops nicht vorgesehen, es werden daher Einzelmeinungen wiedergegeben. Externe und interne Validität sollten getrennt voneinander bewertet werden. Items, die einen großen Spielraum für subjektive Bewertungen lassen, führen zu mangelnder Übereinstimmung der Bewertung und hohem Diskussionsbedarf. Dies kann durch eine präzise Operationalisierung der Items vermieden werden.

6. Diskussion

Studienqualität kann unterschiedlich operationalisiert werden. Es überwiegt die Auffassung, dass eine Bewertung der Studienqualität die Höhe der internen Validität bzw. das Verzerrungspotenzial abbilden sollte. Die Bestandsaufnahme der zahlreichen identifizierten Instrumente zeigt jedoch, dass viele Instrumente auch Items der Berichtsqualität enthalten. Diese Vermischung von Berichtsqualität und interner Validität kann zu einer Fehleinschätzung der Studienqualität führen, wenn Elemente der Berichtsqualität als Surrogatparameter für die Einschätzung der methodischen Qualität herangezogen werden.

Anhand der tabellarischen Darstellung abgedeckter inhaltlicher Items können die identifizierten QBI verglichen werden. Dieses Vorgehen ist jedoch mit Einschränkungen verbunden, da kein Konsens über geeignete Kriterien existiert und nicht für alle Elemente Evidenz vorliegt, dass sie die Höhe der internen Validität einer Studie beeinflussen. Daher ist eine hohe Zahl an abgedeckten Elementen nicht notwendigerweise ein Hinweis auf ein gutes Instrument.

Zur weiteren Differenzierung der QBI wird die Anzahl der als relevant definierten Elemente dargestellt. Während für die relevanten Elemente in Interventions- und Diagnosestudien nur evidenzbasierte Biasquellen ausgewählt werden, trifft dies nur für einige der relevanten Elemente in Beobachtungsstudien und systematischen Übersichtsarbeiten zu. Insgesamt kann die Erfüllung von relevanten Elementen nur als erste Einschätzung dienen, um Instrumente zu identifizieren, die mehr oder weniger umfassend sind. Je nach Themenbereich sollte jeweils geprüft werden, ob alle Items des Instruments relevant sind bzw. ob für das jeweilige Thema zusätzliche Items einbezogen werden sollten.

Einige inhaltliche Elemente von QBI waren nicht eindeutig der Berichtsqualität, der internen oder externen Validität zuzuordnen. Beispielsweise ist die Berechnung der erforderlichen Stichprobengröße zunächst nur mit der Präzisi-

on der Ergebnisse assoziiert ohne dass die Höhe des Effektschätzers beeinflusst wird. Die Präzision der Effektschätzer kann jedoch Einfluss auf die Signifikanz der Ergebnisse haben.
Sicher werden nicht alle jemals eingesetzten Instrumente gefunden. Gleichwohl wird die Möglichkeit, bedeutsame und häufig eingesetzte Instrumente übersehen zu haben, als gering eingeschätzt, u. a. auch durch die Nutzung mehrerer Datenquellen einschließlich Internet.
Generell gilt, je höher der Spielraum für subjektive Bewertungen ist, desto geringer ist die Übereinstimmung der Reviewer. Die einzelnen Items der Instrumente sollten daher möglichst präzise und ausführlich operationalisiert sein. Ggf. sind die Ausfüllhinweise anzupassen, um eine eindeutige Bewertungsgrundlage für alle Reviewer sicherzustellen. Etwa 40% der eingeschlossenen Instrumente geben eine ausführlichere Anleitung zur Durchführung der Qualitätsbewertung.
Die Bewertung der Qualität gesundheitsökonomischer Studien ist ein zwingend erforderlicher Bestandteil bei der Erstellung von HTA-Berichten. Insgesamt werden 22 gesundheitsökonomische QBI identifiziert. Zwischen den untersuchten Instrumenten gibt es deutliche Unterschiede bezüglich:

- Anzahl der untersuchten Items aus dem Extraktionsformular (Themenschwerpunkte)
- Bewertungsqualität: angemessen – begründet – berichtet
- Differenziertheit der Qualitätsabfragen.

Keines der untersuchten Bewertungsinstrumente deckt die gesamte Bandbreite der Themenschwerpunkte (Elemente der gesundheitsökonomischen Evaluation) ab. Nur wenige Instrumente berücksichtigen fast alle Bereiche des Extraktionsbogens. Nur drei Instrumente überprüfen überwiegend die Angemessenheit der methodischen Verfahren. In vielen Instrumenten wird zumindest bei einigen Items nach der Angemessenheit der Verfahren gefragt. Für keines der Instrumente wird jedoch erläutert, was unter „angemessen“ zu verstehen ist. Die Mehrzahl der Instrumente fordert Begründungen für konkrete Ausprägungen der Items ein oder untersucht lediglich, ob und welche Items berichtet werden.
Deutliche Unterschiede bestehen auch in der Differenziertheit der Qualitätsabfragen. Wie differenziert ein Bewertungsinstrument die Themenschwerpunkte erfragt, wird über die Anzahl der Items abgebildet. Wenn sich die Qualitätsbewertung auf wenige Items stützt, müssen die Fragen global gestellt werden. Reviewern bleiben dann größere Spielräume bei der Interpretation von Items. Bei umfangreicheren Instrumenten mit großer Itemanzahl lassen sich Items stärker operationalisieren, sodass die Interpretationsspielräume deutlich eingeschränkt werden und objektivere Bewertungen unterstützt werden.

7. Schlussfolgerungen

Die Qualitätsbewertung von Studien ist ein obligatorischer Arbeitsschritt bei der Erstellung von systematischen Übersichtsarbeiten, der transparent darzustellen ist. Es stehen unterschiedliche designspezifische Instrumente zur Verfügung, die entsprechend ihrer inhaltlichen Abdeckung von Elementen der internen Validität für die Qualitätsbewertung ausgewählt werden können.
Für die Auswahl eines QBI gilt, dass Skalen nicht bzw. ohne quantitative Gesamtbewertung eingesetzt werden sollten. Zur Minimierung der Subjektivität der Bewertung sind Instrumente mit einer ausführlichen und präzisen Operationalisierung der einzelnen Elemente vorteilhaft. Wenn möglich, sollten die ausgewählten Instrumente zuvor an ausgewählten Studien getestet und bei Bedarf die Operationalisierung der Items ergänzt bzw. präzisiert werden, um die Subjektivität der Bewertung zu minimieren und eine hohe Übereinstimmung der Bewertungen sicherzustellen.
Weitere Forschung ist erforderlich, um Studiencharakteristika zu identifizieren, die die interne Validität von Studien beeinflussen. Dies gilt insbesondere für Beobachtungsstudien. Offen ist auch, inwieweit die Validität von Studien durch eine qualitative Gesamtbewertung korrekt gemessen wird.
Für die gesundheitsökonomische Qualitätsbewertung sollten Instrumente entwickelt werden, die (1) die gesamten Themenschwerpunkte abbilden, (2) die angemessene Umsetzung von Items in gesundheitsökonomischen Studien überprüfen und (3) die Themenschwerpunkte hinreichend differenziert abfragen. Die Angemessenheit sollte sich an den Standards der gesundheitsökonomischen Evaluation orientieren (definiert durch Standardlehrbücher und internationale Guidelines). Es sollten Erläuterungen und Ausfüllhinweise zu den Bewertungsinstrumenten entwickelt werden, in denen beschrieben wird, wie Angemessenheit definiert ist.

Korrespondenzadresse:
Dr. med. Maren Dreier, MPH
Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, Medizinische Hochschule Hannover, Carl-Neuberg-Str. 1, 30625 Hannover, Deutschland, Tel.: +49(0)511/5322192
Dreier.Maren@mh-hannover.de

Bitte zitieren als
Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. GMS Health Technol Assess. 2010;6:Doc07.
DOI: 10.3205/hta000085, URN: urn:nbn:de:0183-hta0000855

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/hta/2010-6/hta000085.shtml

***Veröffentlicht:** 14.06.2010*

Der vollständige HTA-Bericht in deutscher Sprache steht zum kostenlosen Download zur Verfügung unter:
http://portal.dimdi.de/de/hta/hta_berichte/hta260_bericht_de.pdf