IQWiG Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

Institute for Quality and Efficiency in Health Care

General Methods[a]

Version 4.2 of 22 April 2015

[a] This translation is based on the German document *Allgemeine Methoden* (Version 4.2) of 22 April 2015. Please note: The translation is provided as a service by IQWiG to English-language readers. However, solely the German original text is absolutely authoritative and legally binding.

Address of publisher:

Institute for Quality and Efficiency in Health Care
Im Mediapark 8 (KölnTurm)
50670 Cologne
Germany

Tel.: +49 (0)221 – 35685-0
Fax: +49 (0)221 – 35685-1
E-mail: methoden@iqwig.de
Internet: www.iqwig.de

Preamble

The Institute for Quality and Efficiency in Health Care (IQWiG[b]) is an establishment of the Foundation for Quality and Efficiency in Health Care. IQWiG is a professionally independent scientific institute. Information on the structure and organization of the Foundation and the Institute is available on the website www.iqwig.de.

The *General Methods* explain the legal and scientific basis of the Institute. Its tasks are described in this document, as are the scientific tools applied in the preparation of its products. Hence the Institute's methods paper provides an important contribution towards transparency in the Institute's mode of operation.

The *General Methods* are primarily directed at researchers. In order to make the information on the Institute's mode of operation accessible to as many interested persons as possible, the authors have aimed to produce a comprehensible document. However, as with any scientific text, a certain level of prior knowledge on the topic is assumed.

The *General Methods* aim to describe the Institute's procedures in a general manner. What specific individual steps the Institute undertakes in the assessment of specific medical interventions depend, among other things, on the research question posed and the available scientific evidence. The *General Methods* should therefore be regarded as a kind of framework. How the assessment process is designed in individual cases is presented in detail for each specific project.

The Institute's methods are usually reviewed annually with regard to any necessary revisions, unless errors in the document or relevant developments necessitate prior updating. Project-specific methods are defined on the basis of the methods version valid at that time. If changes are made to the general methodological procedures during the course of a project, then it will be assessed whether project-specific procedures need to be modified accordingly. In order to continuously further develop and improve its mode of operation, the Institute presents its *General Methods* for public discussion. This applies to the currently valid version, as well as to drafts of future versions.

[b] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen

What is new?

In comparison with Version 4.0 of the Institute's *General Methods* of 23 September 2011, in Version 4.1 minor errors were corrected and a few editorial changes made. The following changes to content were made:

- description of the external review for preliminary reports as an optional step in Sections 2.1.1 and 2.2.3
- division of the previous Section 3.1.4 into the new Sections 3.1.4 and 3.1.5 and specification of the requirements for the evidence base to formulate conclusions on benefit with different certainties of conclusions
- operationalization of the determination of the extent of added benefit, as well as the corresponding rationale, in Section 3.3.3 and in the new Appendix
- use of prediction intervals for meta-analyses with random effects in Section 8.3.8 (previous section 7.3.8)

In comparison with Version 4.1 of the Institute's *General Methods* of 28 November 2013, in the present draft for Version 4.2 minor errors were corrected, editorial changes made, and current literature citations added. The following changes of content were made:

- addition of a section on health economic standards in Chapter 1 (new Section 1.3)
- addition of the new product "assessment of potential" according to § 137e Social Code Book V (SGB V[c]), as well as the corresponding methods in Sections 1.1, 2.1, 2.2, 3.8, 7.2 (previously Section 6.2), 8.1.3 (previously Section 7.1.3), and 8.1.4 (previously Section 7.1.4)
- revision of the text on health information after changes of formats following the relaunch of the website informedhealthonline.org (gesundheitsinformation.de) of 13 February 2014 in Section 2.1.7 and Chapter 6 (previously Chapter 5)
- new version of Section 3.5 on diagnostic tests with integration of the old Section 3.8 on prognosis studies
- integration of methods for health economic evaluations (HEEs) as new Chapter 4 and connected revisions in Section 3.1.5
- amendment on the handling of data provided without a prior request in Chapter 7 (previously Chapter 6)

[c] Sozialgesetzbuch: regulates the statutory health care services.

- amendments on the hierarchy of evidence of non-randomized studies in Section 8.1.3 (previously Section 7.1.3)
- amendment on patient-relevant outcomes in Section 8.3.3 (previously 7.3.3)

Table of contents

List of tables

Page

List of figures

Page

List of abbreviations

Abbreviation	Definition
AGREE	Appraisal of Guidelines Research and Evaluation in Europe
AHP	analytic hierarchy process
AMNOG	Arzneimittelmarktneuordnungsgesetz (Act on the Reform of the Market for Medicinal Products)
AMSTAR	A Measurement Tool to Assess Systematic Reviews
ANV	AM-NutzenV. Arzneimittel-Nutzenbewertungsverordnung (Regulation for Early Benefit Assessment of New Pharmaceuticals)
AWMF	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V. (Association of the Scientific Medical Professional Societies)
BIA	budget impact analysis
CA	conjoint analysis
CONSORT	Consolidated Standards of Reporting Trials
DELB	Deutsches Leitlinien-Bewertungsinstrument (German Instrument for Methodological Guideline Appraisal)
DMP	disease management programme
DRG	diagnosis related groups
EBM	evidence-based medicine
EMA	European Medicines Agency
G-BA	Gemeinsamer Bundesausschuss (Federal Joint Committee)
GKV	gesetzliche Krankenversicherung (statutory health insurance, SHI)
GoR	grade of recommendation
GRADE	Grading of Recommendations Assessment, Development and Evaluation Working Group
HEE	health economic evaluation
HTA	health technology assessment
ICD	International Statistical Classification of Diseases and Related Health Problems
IPD	individual patient data
IQR	interquartile region
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)
ISPOR	International Society For Pharmacoeconomics and Outcomes Research
ITT	intention to treat

LoE	level of evidence
MTC	mixed treatment comparison
NHB	net health benefit
PRO	patient-reported outcome
PSA	probabilistic sensitivity analysis
QALY	quality-adjusted life year
RCT	randomized controlled trial
SGB V	Sozialgesetzbuch – Fünftes Buch – Gesetzliche Krankenversicherung (Social Code Book – Book V – Statutory Health Insurance)
SHI	statutory health insurance
STARD	Standards for Reporting of Diagnostic Accuracy
STE	surrogate threshold effect
WHO	World Health Organization

A chief cause of poverty in science is mostly imaginary wealth. The aim of science is not to open a door to infinite wisdom but to set a limit to infinite error.

Bertolt Brecht. Life of Galileo. Frankfurt: Suhrkamp. World premiere, first version, Zurich theatre, 1943.

1 The Institute for Quality and Efficiency in Health Care

1.1 Legal responsibilities

The Institute was founded within the framework of the German Health Care Reform of 2004 [135] as an establishment of the Foundation for Quality and Efficiency in Health Care. The legal basis and responsibilities of the Institute have been anchored in Social Code Book Fifth Book – Statutory Health Insurance (SGB V[4]) [2] and adapted and extended several times in the course of further health care reforms. More information on the Institute's structure and organization is available on the website www.iqwig.de.

The Institute addresses issues of fundamental relevance for the quality and efficiency of statutory health insurance (SHI) services. Its specific responsibilities are outlined in detail in § 139a SGB V:

- search for, assessment and presentation of current scientific evidence on diagnostic and therapeutic procedures for selected diseases
- preparation of scientific reports, expert opinions, and comments on quality and efficiency issues of SHI services, taking age, gender, and personal circumstances into account
- appraisal of evidence-based clinical practice guidelines (CPGs) on the most relevant diseases from an epidemiological point of view
- issue of recommendations on disease management programmes (DMPs)
- assessment of the benefit and cost of drugs
- provision of easily understandable information for all patients and consumers on the quality and efficiency of health care services, as well as on the diagnosis and treatment of diseases of substantial epidemiological relevance

The modalities of the commissioning and performance of tasks are specified in §139b SGB V. According to this law, only the Federal Joint Committee (G-BA[5]) or the Federal Ministry of

[4] Sozialgesetzbuch: regulates the statutory health care services.

[5] Gemeinsamer Bundesausschuss: The G-BA is the decision-making body of the self-government of the German health care system. More information on the Committee's responsibilities is provided at http://www.english.g-ba.de/.

Health[6] may commission the Institute. In the case of commissioning by the Ministry, the Institute can reject a commission as unfounded, unless the Ministry funds the project.

The Institute must ensure that external experts are involved in the work on commissions. In order to ensure the Institute's scientific independence, these experts are required to disclose all connections to associations and contract organizations, particularly in the pharmaceutical and medical devices industries, including details on the type and amount of any remuneration received (see Section 2.2.2).

The Institute submits the results of the work on commissions awarded by the G-BA to this body in the form of recommendations. According to the law, the G-BA must consider these recommendations in its decision-making processes.

The Institute is largely funded by contributions of SHI members. For this purpose, a levy is determined by the G-BA in accordance with §139c SGB V. This levy is paid by all German medical practices and hospitals treating SHI-insured patients.

Within the framework of the Act on the Reform of the Market for Medicinal Products (AMNOG[7]), at the beginning of 2011, the Institute's responsibilities were extended to the assessment of the benefit of drugs with new active ingredients shortly after market entry [136]. For this purpose manufacturers must submit dossiers summarizing the results of studies. The G-BA is responsible for this "early benefit assessment"; however, it may commission the Institute or third parties to examine and assess the dossiers.

The new regulations in §35a SGB V are the basis for these assessments. They are supplemented by a legal decree of the Federal Ministry of Health [80], which has also been effective since the beginning of 2011, and the G-BA's Code of Procedure [211].

In connection with a benefit assessment the G-BA can also commission the Institute to conduct a health economic evaluation (HEE). The framework of these HEEs is specified in §35b SGB V and §139a SGB V.

In this context, cost-effectiveness ratios of medical technologies are compared with the aim of providing information on the basis of which the appropriateness and reasonableness of cost coverage by the community of SHI insurants can be considered.

The HEE itself is based on a comparison with other drug or non-drug interventions. In particular, the following criteria to determine the benefit for patients are named in the law: increase in life expectancy, improvement in health status and quality of life (QoL), and reduction in disease duration and adverse effects. The definition of a "patient-relevant

[6] Bundesministerium für Gesundheit, BMG

[7] Arzneimittelmarktneuordnungsgesetz

benefit" valid for the Institute is inferred from the above specifications in the law (see Section 3.1).

Within the framework of the Structure of Health Care Act, in 2012 changes were made to §137c SGB V and §137e SGB V was added. This gives the G-BA the option to initiate clinical studies on new examination and treatment methods (testing), provided that the benefit of a method has not yet been sufficiently proven but its potential as a necessary treatment alternative can be recognized. External applicants (e.g. manufacturers of medical devices) can also apply for a testing procedure by submitting informative documents to the G-BA on the potential of the method. The determination of the potential of a method is the responsibility of the G-BA, which has specified criteria for this purpose [211]. The G-BA usually commissions the Institute to evaluate testing applications according to §137e (7) SGB V in view of whether a potential of the method can be inferred from the application documents.

According to §139a (4) Sentence 1 SGB V, the Institute is legally obliged to ensure the "assessment of the medical benefit [of interventions] following the internationally recognized standards of evidence-based medicine and the economic evaluation following the relevant internationally recognized standards for this purpose, in particular of health economics".

Depending on the commission, the Institute determines the methods and criteria for the preparation of assessments on the basis of the international standards of evidence-based medicine (EBM) and health economics recognized by the relevant experts. The term "evidence-based medicine", its development and the underlying concept are described in detail in Section 1.2. The term "health economics" and the underlying concept are described in detail in Section 1.3.

During the preparation of its reports, the Institute ensures the high transparency of procedures and appropriate involvement of third parties. In all important phases of report preparation the law obliges the Institute to provide the opportunity of comment to experts, manufacturers and relevant organizations representing the interests of patients and self-help groups of chronically ill and disabled persons, as well as to the Federal Government Commissioner for Patients' Affairs. The Institute goes beyond this obligation by allowing all interested persons and institutions the opportunity to submit comments on its reports, and considers these comments in its assessments.

The implementation of these regulations is described in Section 2.1.1 in connection with the production of report plans (protocols) and preliminary reports.

In addition, the Institute publishes the results of its work and supplementary information on its publicly accessible website. Those interested can also subscribe to the Institute's e-mail service (info service), where subscribers themselves can specify what type of information they would like to receive from the Institute.

1.2 Evidence-based medicine

EBM refers to patient health care that is not only based on opinions and consensus, but considers “evidence” – i.e. proof (e.g. of the benefit of a medical intervention) determined with the most objective scientific methods possible. EBM comprises tools and strategies designed to safeguard against false decisions and false expectations. In this context, a false decision can mean that beneficial interventions are not implemented in health care (or implemented with delay), or that useless or even harmful interventions are widely applied [17,178,231,236].

However, tools designed to prevent subjective (and therefore often biased) assessments (see also Chapter 8) were not first invented with the introduction of the term “EBM”, but originated decades ago. In Germany, as early as 1932 Paul Martini described the main elements of a fair assessment of drug effectiveness in his monograph *Methodology of Therapeutic Studies* [383]. In the early 1960s, the method of randomly allocating study participants to comparator groups (randomization) in order to assess the effectiveness and safety of medical interventions became the internationally accepted standard [267]. Starting in the United States, in this period this type of study became the precondition for the approval of drugs and (in some cases) medical devices regulated by authorities, legislation and other regulations [33]. About 20 years later, clinical epidemiologists attempted to establish this methodology in clinical practice [183]. Accompanied at times by serious controversy, this was not actually achieved until the 1990s, at the same time as the concept was defined as “EBM”. Since this time, clinical studies and the systematic search for and assessment of these studies (systematic reviews) have formed the basis of the international scientific standard for health technology assessments (HTAs) [31].

EBM is not a rigid concept: which standard tool is to be applied, and when, depends on the question to be answered and the decision to be made. Despite the application of standards, decisions for which no international specifications are (as yet) available have to be made repeatedly in the search for, and the processing and assessment of studies. EBM also includes the freedom to define one’s own specifications in such situations. However, this freedom is linked to the obligation to define such specifications preferably a priori, and to explain assessments in a transparent manner, so that the rationale is comprehensible. The following sections explain that in the implementation of EBM and the definition of specifications, an institution such as IQWiG is in a different situation from clinicians seeking support for a treatment decision.

1.2.1 Practical evidence-based medicine

The EBM concept is a strategy for physicians who, from a range of possible interventions, seek the most promising alternatives suited best to the needs of their patients, and who aim to offer prospects of success in an objective manner. This implementation of EBM in daily clinical practice for “individual patients” was defined by David Sackett et al. [474] as follows: “EBM is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of the individual patient. It means integrating individual clinical expertise with the best available external clinical evidence from systematic research" (1996).

However, the "best available evidence" is often incomplete or unreliable. EBM has developed instruments to assess uncertainty; evidence levels are often used for illustration. In this way, EBM helps physicians and patients to recognize the type and degree of uncertainty; they can then discuss how to deal with this. Especially in uncertain situations, personal preferences are important and determine what option patients choose. Apart from being based on evidence, decisions are also ideally based on the clinical condition and circumstances of the individual patient, as well as on his or her preferences and actions [255]. At the same time, the description of the identified gaps in knowledge creates the precondition for medical research targeted towards patients' needs.

EBM is based on a critical approach [328]. The importance of scepticism is underlined by the fact that over the past few decades, several insufficiently tested but widely applied therapies have been assessed with EBM methods; these assessments have shown that a hasty, overoptimistic approach to a new intervention can have dangerous consequences for patients [157,457]. It is the Institute's task to assess objectively with what certainty the benefit of medical interventions has been demonstrated, in order to counter inappropriate judgements.

1.2.2 The relevance of evidence-based medicine for the Institute

The Institute's main task is to provide the most reliable answer possible to the question specified by the contracting agency as to whether evidence is available of the benefits or harms from an intervention. The aim is to present sufficiently reliable proof that "Treatment A" is better for patients than "Alternative B" for a specific disease. In short: What is the benefit of A compared with B?

The Institute's remit is therefore intentionally not aimed towards treating individual patients with their potential specific characteristics, but towards determining for which patient groups proof of a benefit of an intervention is available. In its decisions, the G-BA then considers aspects of patient care that are beyond the scope of a benefit assessment [211].

1.2.3 Strategies of evidence-based medicine

A characteristic standard element of EBM is the structured and systematic approach to the search for a response to a medical question:

1) The medical question must be worded precisely. Medicine (nearly) always deals with the choice between at least 2 alternatives. This can refer to treatments, diagnostic tests or complex changes in life style. From this, the following question is always inferred: Is Option A better than Option B? In this context, the decision not to undergo treatment can also be an option that should be thoroughly reviewed. However, it should be stressed that such an option (e.g. "watchful waiting") is not the same as "doing nothing".

2) It must be defined how the benefit of treatment (or diagnosis or lifestyle change) should be measured. A standard element of EBM is the question about relevant consequences for patients: Can life expectancy be increased? Can symptoms and quality of life be improved?

3) In EBM it is explicitly noted that in medicine, only probability statements or only conclusions about groups of patients are usually possible with regard to the benefit of treatment, diagnostic procedures, or lifestyle changes. Benefit is demonstrated by showing that an intervention increases the probability of a beneficial outcome and/or reduces the risk of a non-beneficial outcome. In order to prove the benefit of an intervention, studies in sufficiently large groups of suitable patients are required. International researchers have developed a range of rules and tools for the planning, conduct, and analysis of such studies. The most important aim is to minimize (or, if this is impossible, at least document) factors that can distort the results of a comparison. The effects of such confounding factors are referred to as “bias”. The rules and tools that are internationally accepted as the prevailing standard, and are under continuous development, are the methodological basis of EBM and the Institute’s work.

4) A further key EBM strategy is to identify all “appropriate” studies (i.e. whose design and conduct are of appropriate quality) on a question and, in this way, to summarize the reliable evidence available. In this context, if large differences are shown between the results of individual studies (heterogeneity), an attempt should be made to explain them. The findings of these summaries and assessments are referred to as systematic reviews; the statistical analyses are referred to as meta-analyses.

1.2.4 The relevance of certainty of results

A specific characteristic of EBM is that it allows assessment as to what extent the available evidence is reliable. Decisions made by the G-BA must be based on highly reliable scientific evidence, as they have far-reaching consequences for all SHI members (e.g. exclusion of services from reimbursement).

The assessment of the certainty of results therefore plays a key role in the Institute’s reports. Numerous details on how studies are planned, conducted, analysed, and published have an impact on how reliable the available results are. It is an international EBM standard to test and assess these aspects critically. However, how the certainty of results needed to answer a question can be achieved also depends on the disease and on the effect size of an intervention: If 2 athletes pass the finishing line of a fair race with a great distance between them, no stopwatch is needed to identify the winner. For example, the benefit of a new therapy that results in the cure of a previously always fatal disease can be proven by a relatively small number of surviving patients. In this case, the judgement is also ultimately based on a comparison, but in interventions with such dramatic effects, the comparison between historical and current patients may already provide sufficient certainty. However, therapies that show such dramatic benefits are very rare in modern medicine.

In chronically ill patients in particular, differences between 2 therapy alternatives are mostly smaller and may be easily confounded by a fluctuant course of disease. In these cases, precise methods and appropriate study designs are required in order to be able to recognize therapy effects under such fluctuations.

It can be assumed that the Institute will be specifically commissioned to compare such interventions where it is not immediately recognizable which alternative will be more beneficial. However, the smaller the expected differences between 2 alternatives are, the more reliable the studies must be in order to be sufficiently certain that an observed effect is not caused by chance or measurement errors (a world record over 100 metres can no longer be measured with an hourglass). In the event of small differences, their clinical relevance must also be judged.

The following requirements for precision and reliability determine the Institute’s mode of operation:

1) For every question investigated, it is an international EBM standard to specify the study type (measuring tool) that minimizes the risk of unjustifiably discriminating against one of the alternatives.
2) The Institute’s assessments on the benefits and harms of interventions are therefore normally based only on studies with sufficient certainty of results. This ensures that the decisions made by the G-BA, which are based on the Institute’s recommendations, are supported by a sound scientific foundation. Moreover, an assessment that includes a literature search for studies with insufficient certainty of results would be costly and time consuming.
3) If it emerges that studies of the required quality and precision are generally lacking, it is the core task of the Institute to describe the circumstances and conclude that on the basis of the “currently best available” evidence, it is not possible to make reliable recommendations.
4) It is the G-BA’s responsibility to take this uncertainty into account in its decision-making processes. In addition to considering scientific evidence, the G-BA also considers other aspects in its decisions, such as the efficiency of interventions as well as the needs and values of people [222]. In an uncertain scientific situation, such aspects become more important. In addition, the G-BA also has the option to call for or initiate studies in order to close the evidence gaps identified.

1.2.5 The connection between certainty of results and proximity to everyday conditions

The great value placed on the assessment of the certainty of results is often criticized. One argument is that studies with a high certainty of results (especially randomized controlled trials, RCTs) may have high internal validity, but often do not represent patient care under everyday conditions, and are therefore not transferable, i.e. have only low external validity. In this context it must be examined how well the patient population investigated in the studies, the interventions applied, and the outcome criteria analysed are in accordance with everyday conditions in health care. This criticism is then often connected to the call to include other study types without randomization, in order to better consider everyday conditions.

However, this criticism conflates levels of arguments that should be clearly separated. The following aspects should be taken into account:

1) The basis of a benefit assessment is the demonstration of causality. An indispensable precondition for such a demonstration is a comparative experiment, which has to be designed in such a way that a difference between intervention groups – an effect – can be ascribed to a single determining factor – the intervention tested. This goal requires considerable efforts in clinical trials, as there are numerous confounding factors that feign or mask effects (bias). The strongest of these distorting influences are unequal baseline conditions between comparator groups. Randomization (together with careful concealment) is currently the best available tool to minimize this type of bias. Random allocation of participants to groups ensures that there are no systematic differences between groups, neither regarding known factors (e.g. age, gender, disease severity), nor unknown factors. For this reason, RCTs provide a basic precondition for the demonstration of causality. However, randomization alone does not guarantee high certainty of results. To achieve this, the unbiased assessment, summarization and publication of results, for example, are also required.
2) Study types other than RCTs are usually not suited to demonstrate causality. In non-randomized comparative studies, as a matter of principle structural equality of groups cannot be assumed. They therefore always provide a potentially biased result and mostly cannot answer with sufficient certainty the relevant question as to whether a difference observed is caused by the intervention tested. The use of non-randomized studies as proof of the causality of an intervention therefore requires particular justification or specific preconditions and special demands on quality.
3) It is correct that many randomized studies do not reflect aspects of everyday patient care, for example, by excluding patients with accompanying diseases that are common in everyday life. However, this is not a consequence of the randomization technique, but of other factors (e.g. definition of narrow inclusion and exclusion criteria for the study, choice of interventions or outcome criteria). In addition, patients in randomized studies are often cared for differently (more intensively and more closely) than in everyday practice. However, these are intentional decisions made by those persons who wish to answer a specific question in a study. Dispensing with randomization does not change these decisions. There is also a selection of participants in non-randomized studies through inclusion and exclusion criteria and other potential design characteristics, so that external validity is not given per se in this study type any more than in RCTs.

4) Even if patient groups in an RCT differ from everyday health care, this does not mean the external validity of study results must be questioned. The decisive issue is in fact whether it is to be expected that a therapy effect determined in a population varies in a different population.
5) It depends on the individual case how the intensity of care provided in a study influences outcomes. For example, it is conceivable that a benefit of an intervention actually exists only if patients are cared for by specially qualified physicians, as under everyday conditions too many complications may otherwise occur. However, it is also possible that intensified care of patients is more likely to reduce differences between groups. For example, differences in treatment adherence may be smaller in studies where, as a matter of principle, patients are cared for intensively.
6) However, the initiators of a clinical trial are responsible for the specification of study conditions. They can define research questions and outcomes rated as so relevant that they should be investigated in a study. If, for example, a drug manufacturer regards treatment adherence to be an important aspect of the benefit of a product, the obvious consequence would be to initiate studies that can measure this aspect with the greatest possible certainty of results and proximity to everyday conditions, and at the same time demonstrate its relevance for patients.

The above remarks show that certainty of results and proximity to everyday conditions (or internal and external validity) have no fixed relationship. High certainty of results and proximity to everyday conditions do not exclude one another, but only require the appropriate combination of study type, design and conduct.

Even if criticism of the lack of proximity to everyday practice may actually be justified for many studies, nothing would be gained by dispensing with high certainty of results in favour of greater proximity to everyday practice, because one would thereby be attempting to compensate one deficit by accepting another, more serious, one [253].

Studies that combine proximity to everyday conditions and high certainty of results are both desirable and feasible. RCTs are indeed feasible that neither place demands on patients beyond everyday health care nor specify fixed study visits. Such studies are being discussed at an international level as “real world trials”, “practical trials” or “pragmatic trials” [199,201,218,381,561]. However, such pragmatic trials may themselves also lead to interpretation problems. For example, if very broad inclusion criteria are chosen, the question arises as to whether the (overall) study results can be applied to the overall study population [596], which, at least to some extent, would ultimately have to be answered by means of appropriate subgroup analyses.

1.2.6 Benefit in individual cases

The aim of a benefit assessment is to make robust predictions for future patients using results of studies suited to demonstrate causal effects. The conclusions drawn always apply to groups of patients with certain characteristics. Conclusions on the benefit of an intervention in terms of predictions of success for individual cases are, as a matter of principle, not possible. Vice versa, experiences based on individual cases (except for specific situations, e.g. dramatic effects) are unsuitable for a benefit assessment, as it is not possible to ascribe the results of an individual case (i.e. without a comparison) to the effect of an intervention.

For certain research questions (therapy optimization in individual patients) so-called (randomized) single patient trials (or "n-of-1" trials) can be conducted [232,238,315,492]. However, these are usually not suited to assess the benefit of a treatment method for future patients.

1.3 Health economics

Two issues can be expressed with the term "health economics".

In the wider sense it is about "the analysis of economic aspects of the healthcare system using concepts of economic theory" [495]. For this purpose, among other things, concepts are used from the areas of microeconomic behavioural theory, competition theory, economic theory of politics, and management theory [495]. The subject of such a study could be, for example, how players in the healthcare system change their behaviour after the setting of incentives (e.g. practice charges)[8] or whether the results of price negotiations following AMNOG actually prevent excessive prices for new drugs. It can be discussed both from a methodological and from an ethical point of view to what extent such studies can and should be used to steer the healthcare system, but this is not a subject of this short presentation.

In the narrower sense health economics is understood to be health economic evaluation (HEE) in the form of comparative and also non-comparative studies, for example, cost-of-illness studies or budget impact analyses. These analyses serve to inform decision makers on the cost--effectiveness ratios of interventions and, in addition to the benefit assessment of interventions, thus represent an area of HTA.

1.3.1 Relevance of health economics for the Institute

With the establishment of the Institute in 2004, the G-BA and the Federal Ministry of Health were free to commission an HEE. Until the change in the law in 2007 an HEE of drugs was not intended. With the SHI Act to Promote Competition[9] the HEE of drugs was anchored in §35b SGB V to gain information on the recommendation for a so-called ceiling price. New drugs were to be reimbursed up to this ceiling price, as this price was to represent the appropriate costs for the added benefit of a new drug in comparison with other drugs and treatment forms in a therapeutic indication. The precondition for the commissioning of an

[8] In Germany, previously a quarterly flat-rate charge of €10 to SHI patients for outpatient treatment (abolished in 2013).

[9] Gesetzliche Krankenversicherung (GKV)-Wettbewerbsstärkungsgesetz

HEE was thus to be proof of the added benefit of a new drug, which had to have been shown in an IQWiG benefit assessment. The development of the methods resulting from this health economic question has been extensively documented [285,287,288,290-292,294,295].

With the Act on the Reform of the Market for Medicinal Products (AMNOG[10]), which became effective on 1 January 2011, the relevance of the HEE shifted within the procedure of the early benefit assessment of drugs. An HEE is primarily intended for cases where price negotiations fail between the SHI umbrella organization[11] and pharmaceutical companies and where no agreement is reached in the subsequent arbitration procedure. However, the question of the HEE remains: according to §35b (1) Sentence 4 SGB V in connection with the 5th Chapter §32 (3) of the G-BA's Code of Procedure [211], the appropriateness and reasonableness of cost coverage by the community of SHI insurants must be considered. For the G-BA to consider these factors in an appropriate manner, it must receive the corresponding information. This information is provided by the HEE (appropriateness) and the budget impact analysis (reasonableness). The assessment of the appropriateness and reasonableness of cost coverage is conducted with regard to whether, under observance of the principle of proportionality, a justifiable relation between the costs and the benefit of the drug exists. In this context, according to the 5th Chapter §32 (2, 3) of the G-BA's Code of Procedure, IQWiG is to present a recommendation on the basis of which the G-BA is to make a decision [211]. The presentation of a justifiable relation between the costs and the benefit must thus ensue from the HEE.

Even if the question as to how health economics is to be understood (see Section 1.3) is not addressed anywhere in the law or in the subordinate regulations, it results from the practical application that it is about HEE and thus about health economics in the narrower sense.

1.3.2 The international standards of health economics

As in every science, international standards also exist in health economics. These include the classification of HEE into the study types of cost-effectiveness analysis, cost-utility analysis, and cost-benefit analysis (in the narrower sense). Sometimes the cost-cost, cost-consequences, and cost-minimization analyses are also named as separate types; however, they are rarely used. With regard to the latter type it is also discussed whether it represents an independent type [159].

International standards also exist for the methods applied in HEE. On the side of the benefit assessment the Institute follows the principles of evidence-based medicine and the resulting specifications that are established as international standards. Before one speaks of international standards in the area of health economics, one must distinguish between clearly methodological questions and questions based on value judgements, opinions or surveys. This

[10] Arzneimittelmarktneuordnungsgesetz

[11] Spitzenverband Bund der Krankenkassen, GKV-Spitzenverband

can be illustrated using the example of the discounting rate. With a discounting rate, benefits and costs incurred in different periods are discounted to one period so that they are now comparable for a decision. The pure performance of discounting is clearly regulated mathematically and thus a methodological question. The choice of discounting rate and particularly the decision as to whether the costs and benefits are to be discounted with the same rate or possibly even with a non-constant rate is, among other things, subject to issues concerning the appraisal of the future economic development and intergenerational fairness [103,250,409,421,423,433,446], and is thus a value judgement.

As shown by internationally recognized instruments for the evaluation of health economic analyses [101,158,280,440], there are many steps and aspects for which methodological requirements exist and which must be processed in a transparent and comprehensible way. These include:

- Definition of the interventions under assessment and their comparators. A choice must be justified to prevent wrong decisions on the basis of an interest-driven choice of comparators.
- Perspective of the HEE.
- Time horizon of the HEE.
- Type of HEE (see above) and preferably justification of the study type.
- Costs with presentation of resource use and resource evaluation.
- Adjustment for inflation and conversion of currency (if necessary).
- Development and explanation of the model and preferably also justification of the choice of model, e.g. decision tree, Markov model.
- Discounting rate.
- Presentation of results, e.g. in an aggregated or a disaggregated form.
- Investigation of the uncertainty of results by means of deterministic and probabilistic sensitivity analyses.
- Presentation of uncertainty, e.g. with cost-effectiveness-acceptance curves or the net benefit.

For some of these topics or subtopics, requirements for good methodological practice are available in textbooks and also, for example, in the guidelines of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR).

1.3.3 Methodological standards in health economics

Furthermore, other aspects of an HEE are also understood to be the international standard. For instance, in the healthcare system a decision based on a threshold per quality-adjusted life year (QALY) gained is often presented as the international standard in health economics.

However, this should be seen critically. On the one hand, the vast majority of countries in which HEE contributes to decision-making do not have a (fixed) threshold. On the other, this would be a value judgement, and would thus not fall under the international methodological standards following §139 a (4) Sentence 1 according to which only methodological standards apply in the assessments of the Institute.

Furthermore, the question of the measure of overall benefit arises not only as a methodological question, but also always under the aspect of a value judgement. In this context, the research question and aim of a health economic analysis have an influence on which instrument should be used as a measure of overall benefit. This also means that the question as to whether the QALY should be used must be highlighted under ethical, legal, and cultural aspects. In turn, from a scientific or methodological point of view it can be discussed which assumptions are considered in the QALY concept, for example, the assumption that the assessment of a state is independent of its duration (= constant proportional (time) trade off), and whether these assumptions are sustainable. Likewise, it can be investigated methodologically whether the various measurement methods applied, for example, indirect versus direct or various indirect and direct procedures in different combinations, lead to different results and what this can mean. A question of value judgement on the basis of legal requirements (e.g. SGB V) is again the question in which persons the utility values to generate QALYs should be elicited; in those actually affected by a disease or in the general population.

Ultimately, the following question needs to be raised: How should resources in the health care system be distributed? That is, on the basis of which rights, claims or needs, with which aim, and which impact on the allocation of goods or services? This question is only understood as a value judgement, and this in turn determines which scientific standards and methods should be applied.

1.4 Evidence-based decision-making in health care

The Institute’s reports are to serve the G-BA as a basis for decisions that in principle apply to all SHI insurants. Other products of the Institute are, for example, to serve as information for the G-BA. The type of decisions made by institutions like the G-BA has an effect on the application of methods of EBM and of health economics.

2 The Institute’s products

According to its legal remit, the Institute prepares a variety of products in the form of scientific reports and easily understandable health information for consumers and patients. This chapter describes procedures and general methods applied in the preparation of the Institute’s products. At first the individual products are named and product-specific procedures presented (Section 2.1). The next section outlines further aspects independent of products (Section 2.2).

2.1 Product-specific procedures

The Institute‘s products include

- report
- rapid report
- dossier assessment
- health economic evaluation (HEE) according to §35b SGB V
- assessment of potential
- addendum
- health information
- working paper

The preparation of reports and rapid reports is conducted on the basis of the award of individual commissions through the G-BA or Federal Ministry of Health. The basis of this are the Institute’s responsibilities described in §139a SGB V (see also Section 1.1). Accordingly, reports and rapid reports can be prepared on the benefit assessment of drug and non-drug interventions, on HEEs, and on the appraisal of CPGs. The main difference between reports and rapid reports is that commenting procedures (hearings) are only conducted for reports, but not for rapid reports. Accordingly, rapid reports are particularly intended for recommendations at short notice, for which, from the point of view of the contracting agency, no hearings by the Institute are required.

Dossier assessments are commissioned by the G-BA. The foundation for this is §35a SGB V, which regulates the assessment of the benefit of new active ingredients on the basis of a dossier by the pharmaceutical company (see also Section 3.3.3). No hearing by the Institute is intended for dossier assessments according to §35a SGB V; this is conducted in the further procedure by the G-BA.

Furthermore, according to § 35b SGB V, the Institute can be commissioned by the G-BA to conduct HEEs of drugs. For such evaluations, it is intended that a commenting procedure (hearing) is conducted by IQWiG. A further commenting procedure is conducted at the G-BA.

Assessments of the potential of new examination and treatment methods are commissioned by the G-BA and refer to applications for testing according to §137e SGB V. No hearing procedure is conducted at the Institute. If testing is performed, the G-BA conducts a commenting procedure on the testing directive.

Addenda can be commissioned by the G-BA or Federal Ministry of Health in cases where, after the completion of a product, the need for additional work on the commission arises during the course of consultations.

Health information can be prepared on the basis of an individual commission; it can also be the consequence of a commission in other areas of the Institute’s work (easily understandable version of other products of the Institute, e.g. a report) or be prepared within the framework of the general legal remit to provide health information.

Working papers are prepared under the Institute’s own responsibility; specific commissioning by the G-BA or Federal Ministry of Health is not required. This takes place either on the basis of the general commission (see Section 2.1.8), with the aim of providing information on relevant developments in health care, or within the framework of the legal remit to develop the Institute’s methods. The Institute’s *General Methods* are not to be understood as a working paper in this sense, and are subjected to a separate preparation and updating procedure, which is outlined in the preamble of this document.

An overview of the Institute’s various products is shown in Table 1 below. Product-specific procedures are described in the subsequent Sections 2.1.1 to 2.1.8.

Table 1: Overview of the Institute's products

Product	Objective	Procedure	Commissioned by
Report	Recommendations on tasks described in §139a SGB V, including hearing	Described in Section 2.1.1	G-BA, Federal Ministry of Health
Rapid report	Recommendations on tasks described in §139a SGB V, insofar as no hearing on interim products is required; in particular provision of information at short notice on current topics	Described in Section 2.1.2	G-BA, Federal Ministry of Health
Dossier assessment	Assessment of the benefit of drugs with new ingredients according to §35a SGB V	Described in Section 2.1.3	G-BA
Health economic evaluation according to §35b SGB V	Assessment of the relation of the cost and benefit of drugs according to §35b SGB V	Described in Section 2.1.4	G-BA
Assessment of potential	Assessment of the potential of new examination and treatment methods according to §137e SGB V	Described in Section 2.1.5	G-BA
Addendum	Supplementary information provided at short notice by the Institute on issues that have arisen during the consultation on its completed products	Described in Section 2.1.6	G-BA, Federal Ministry of Health
Health information	Easily understandable information for consumers and patients; wide scope of topics	Described in Section 2.1.7	G-BA, Federal Ministry of Health/own initiative of the Institute
Working paper	Information on relevant developments in health care or methodological aspects	Described in Section 2.1.8	Own initiative of the Institute
G-BA: Gemeinsamer Bundesausschuss (Federal Joint Committee); SGB: Sozialgesetzbuch (Social Code Book)			

2.1.1 Report

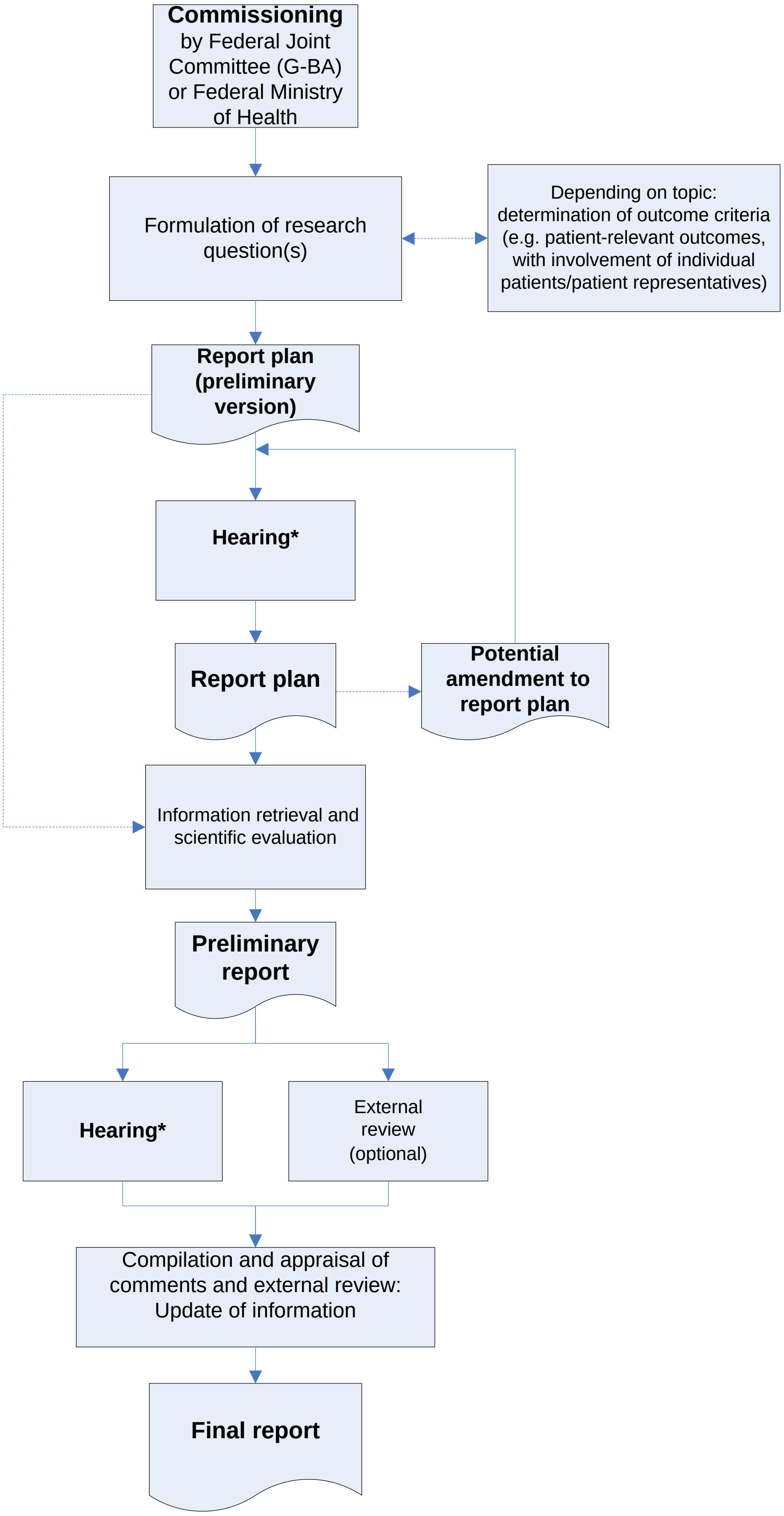
A) Procedure for report production

The procedure for report production is presented in Figure 1. All working steps are performed under the Institute’s responsibility and regularly involve external experts (see Section 2.2.1). If necessary, the Institute’s Scientific Advisory Board is also involved. The internal quality assurance process is not outlined in this flow chart.

After **commissioning** by the G-BA or Federal Ministry of Health, the research question is formulated. Depending on the topic, the determination of outcome criteria is also required (e.g. in benefit assessments). As a rule, relevant patient organizations are involved, especially in the definition of patient-relevant outcomes; the opinion of individual affected patients may also be heard. Subsequently, the report plan (protocol) is prepared.

The **report plan** contains the precise scientific research question, including the outcome criteria (e.g. patient-relevant outcomes), as well as the inclusion and exclusion criteria of the information to be used in the assessment. This plan also includes a description of the project-specific methodology applied in the retrieval and assessment of information. The preliminary version of the report plan is first forwarded to the contracting agency as well as to the Foundation’s Board of Directors, the Foundation Council and the Board of Trustees. It is normally published on the Institute’s website 5 working days later.

For a period of at least 4 weeks, the public is given the **opportunity to submit written comments (hearing)**. This opportunity particularly refers to the project-specific methodological approach applied to answer the research question. The research question itself is usually specified by the commission, and is not an object of the commenting procedure. Optionally, an oral scientific debate including persons submitting comments may be held. This debate serves the potentially necessary clarification of aspects of the written comments and aims at improving the scientific quality of the report plan.


Commissioning
by Federal Joint Committee (G-BA) or Federal Ministry of Health
Formulation of research question(s)
Depending on topic: determination of outcome criteria (e.g. patient-relevant outcomes, with involvement of individual patients/patient representatives)
Report plan (preliminary version)
Hearing*
Report plan
Potential amendment to report plan
Information retrieval and scientific evaluation
Preliminary report
Hearing*
External review (optional)
Compilation and appraisal of comments and external review: Update of information
Final report

* The hearing is conducted by inviting written comments. In addition, an optional oral scientific debate may be held to discuss any unclear aspects of the written comments.

Figure 1: Procedure for the production of a report

After the analysis of the comments and (if appropriate) the conduct of the oral debate, the revised report plan, together with the documentation of the hearing on the report plan, are first forwarded to the contracting agency, the Foundation's Board of Directors, the Foundation Council and Board of Trustees. This document is usually published on the Institute's website 5 working days later. The revised report plan is the basis for the preparation of the preliminary report. If further relevant methodological changes are required in the course of the preparation of the preliminary report, these are usually presented in one or more amendments to the report plan. An opportunity to submit comments is usually also provided after publication of an amendment, following the conditions outlined above.

The results of the information retrieval and the scientific assessment are presented in the **preliminary report**. In order to avoid undue delay in the Institute's work, the retrieval and assessment of information already start before completion of the hearing on the report plan on the basis of the criteria formulated in the preliminary report plan. However, the result of the hearing is explicitly not anticipated, as these criteria may be modified on grounds of the hearing on the preliminary version of the report plan. This may also lead to supplementation and/or modification of the retrieval and assessment of information.

The preliminary report includes the preliminary recommendation to the G-BA. After completion it is first forwarded to the contracting agency as well as to the Foundation's Board of Directors, the Foundation Council and the Board of Trustees. The preliminary report is usually published on the Institute's website 5 working days after it is sent to the contracting agency.

For a period of at least 4 weeks, the public is then given the **opportunity to submit written comments (hearing)**. The results of the retrieval and assessment of information presented in the preliminary report are in particular the subject of the commenting procedure. Optionally, an oral scientific debate with those submitting comments may be held. This debate serves the potentially necessary clarification of aspects of the written comments and aims at improving the scientific quality of the final report.

The **final report**, which is based upon the preliminary report and contains the assessment of the scientific findings (considering the results of the hearing on the preliminary report), represents the concluding product of the work on the commission. The final report and the documentation of the hearing on the preliminary report are first forwarded to the contracting agency, as well as to the Foundation's Board of Directors and Foundation Council, and subsequently (usually 4 weeks later) forwarded to the Foundation's Board of Trustees. These documents are then published on the Institute's website (usually a further 4 weeks later). If comments are received on final reports containing substantial evidence not considered, or if the Institute receives information on such evidence from other sources, the contracting agency will be sent well-founded information on whether, in the Institute's opinion, a new commission on the topic is necessary (if appropriate, a report update) or not. The contracting agency then decides on the commissioning of the Institute. Such an update is conducted according to the general methodological and procedural requirements for the Institute's products.

B) General remarks on the commenting procedure (hearing)

Organizations entitled to submit comments

In accordance with §139a (5) SGB V, the Institute must ensure that the following parties are given the opportunity to submit comments in all important phases of the assessment procedure: medical, pharmaceutical, and health economic experts (from research and practice), drug manufacturers, relevant organizations representing the interests of patients and self-help groups for the chronically ill and disabled, as well as the Federal Government Commissioner for Patients' Affairs. Their comments must be considered in the assessment. These requirements are taken into account by the fact that hearings on the report plan and preliminary report are conducted and that the circle of people entitled to submit comments is not restricted. Moreover, all the Institute's products, in accordance with §139a SGB V, are sent to the Board of Trustees before publication. The following parties are represented in the Board of Trustees: patient organizations, the Federal Government Commissioner for Patients' Affairs, organizations of service providers and social partners, as well as the self-government bodies of the supporting organizations of the G-BA.

Formal requirements

In order to avoid undue delay in the Institute's work, the comments must fulfil certain formal requirements. Further information on the commenting procedure, including the conditions for participation in a scientific debate, can be found in a guideline published on the Institute's website.

Publication of comments

Comments that fulfil the formal requirements are published in a separate document on the Institute's website (*Documentation and appraisal of the hearing*). In order to ensure transparency, documents that are submitted together with the comments and are not publicly accessible (e.g. manuscripts) are also published.

Submission of documents within the framework of the hearing

During both the hearing on the report plan and the one on the preliminary report, the opportunity is provided to submit any document of appropriate quality, which, according to the person submitting comments, is suited to answer the research question of the report. If the search strategy defined in the report plan is restricted to RCTs, for example, non-randomized studies may nevertheless be submitted within the framework of the commenting procedure. However, in such cases, appropriate justification of the validity of the causal interpretation of the effects described in these studies is also required.

2.1.2 Rapid report

The procedure for the production of a **rapid report** is presented in Figure 2. All working steps are performed under the responsibility of the Institute, involving external experts where appropriate (see Section 2.2.1). If necessary, the Institute's Scientific Advisory Board is also involved. The internal quality assurance process is not presented in this flow chart.

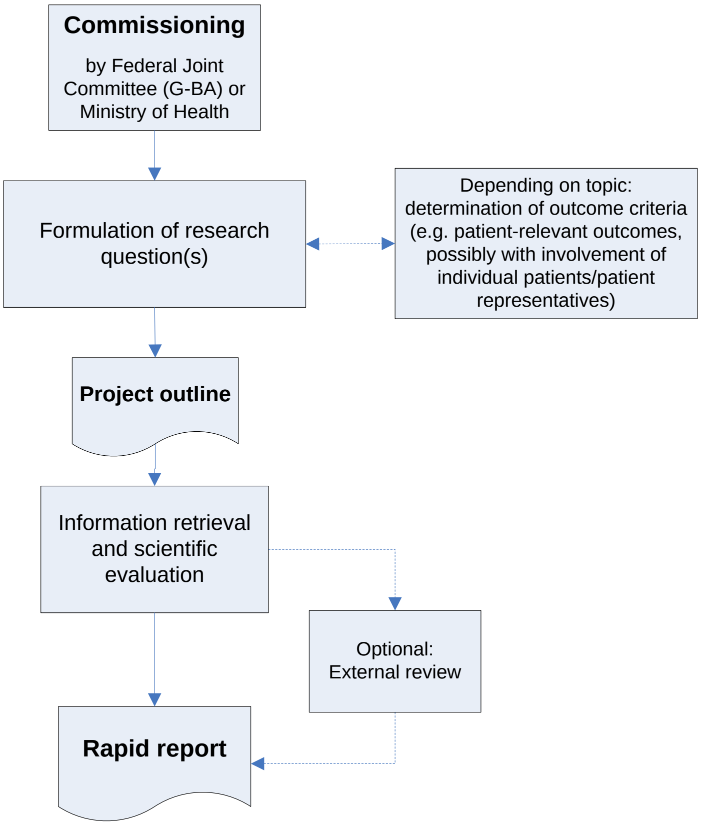

Commissioning
by Federal Joint Committee (G-BA) or Ministry of Health
Formulation of research question(s)
Depending on topic: determination of outcome criteria (e.g. patient-relevant outcomes, possibly with involvement of individual patients/patient representatives)
Project outline
Information retrieval and scientific evaluation
Optional: External review
Rapid report

Figure 2: Procedure for the production of a rapid report

Rapid reports are primarily produced with the aim of providing information at short notice on relevant developments in health care (e.g. new technologies, publication of milestone studies). A shorter production period is usually required here. Interim products are therefore not published and are not the subject of a hearing.

After **commissioning** by the G-BA or Federal Ministry of Health, the research question is formulated. Depending on the topic, the determination of outcome criteria is also required (e.g. in benefit assessments). In this context, patient organizations may be involved or the opinion of individual affected patients sought, especially for the definition of patient-relevant outcomes. Subsequently, the project outline is prepared.

The **project outline** summarizes the main steps of the information retrieval and scientific assessment. It forms the basis for the production of the rapid report. The project outline is not published.

The **rapid report** presents the results of the information retrieval and scientific assessment. Before completion, as a further quality assurance step, optionally a draft of the rapid report may be reviewed by one or more external reviewers (see Section 2.2.3) with proven methodological and/or topic-related competence. After completion the rapid report is then sent to the contracting agency, the Foundation's Board of Directors and Foundation Council, as well as (usually a week later) to the Board of Trustees. The rapid report is usually published on the Institute's website 4 weeks after it is sent to the contracting agency and Board of Directors. If comments on rapid reports are received that contain substantial evidence not considered, or if the Institute receives such evidence from other sources, the contracting agency will be provided with well-founded information on whether, in the Institute's opinion, a new commission on the topic is necessary (if appropriate, a rapid report update) or not. The contracting agency then decides on the commissioning of the Institute. Such an update is conducted according to the general methodological and procedural requirements for the Institute's products.

2.1.3 Dossier assessment

The procedure for the production of a dossier assessment is presented in Figure 3. All working steps are performed under the Institute's responsibility and regularly involve external expertise (see Section 2.2.1). If necessary, the Institute's Scientific Advisory Board is also involved. The internal quality assurance process is not outlined in this flow chart.

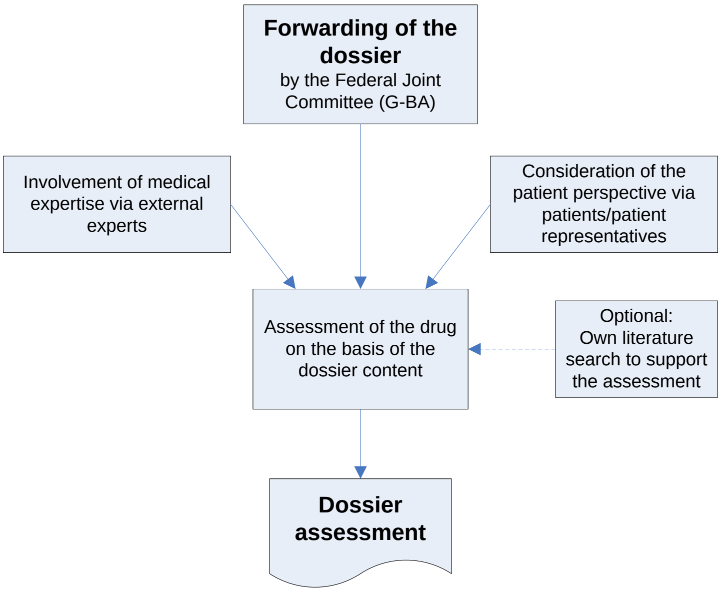

Forwarding of the dossier
by the Federal Joint Committee (G-BA)
Involvement of medical expertise via external experts
Consideration of the patient perspective via patients/patient representatives
Assessment of the drug on the basis of the dossier content
Optional: Own literature search to support the assessment
Dossier assessment

Figure 3: Procedure for the production of a dossier assessment

After the **forwarding of the dossier** by the G-BA, the assessment of the dossier content is conducted under the responsibility of the Institute. In this context, medical expertise and the patient perspective are as a rule involved via external experts and patients/patient organizations respectively.

Medical expertise is primarily involved on the basis of a questionnaire sent to external experts at the beginning of the assessment. In its assessment the Institute considers the external experts’ feedback. In addition, external experts may if necessary be drawn upon to clarify specific questions arising during the course of the assessment. External experts are identified via the Institute’s own database for external experts (see Section 2.2.1).

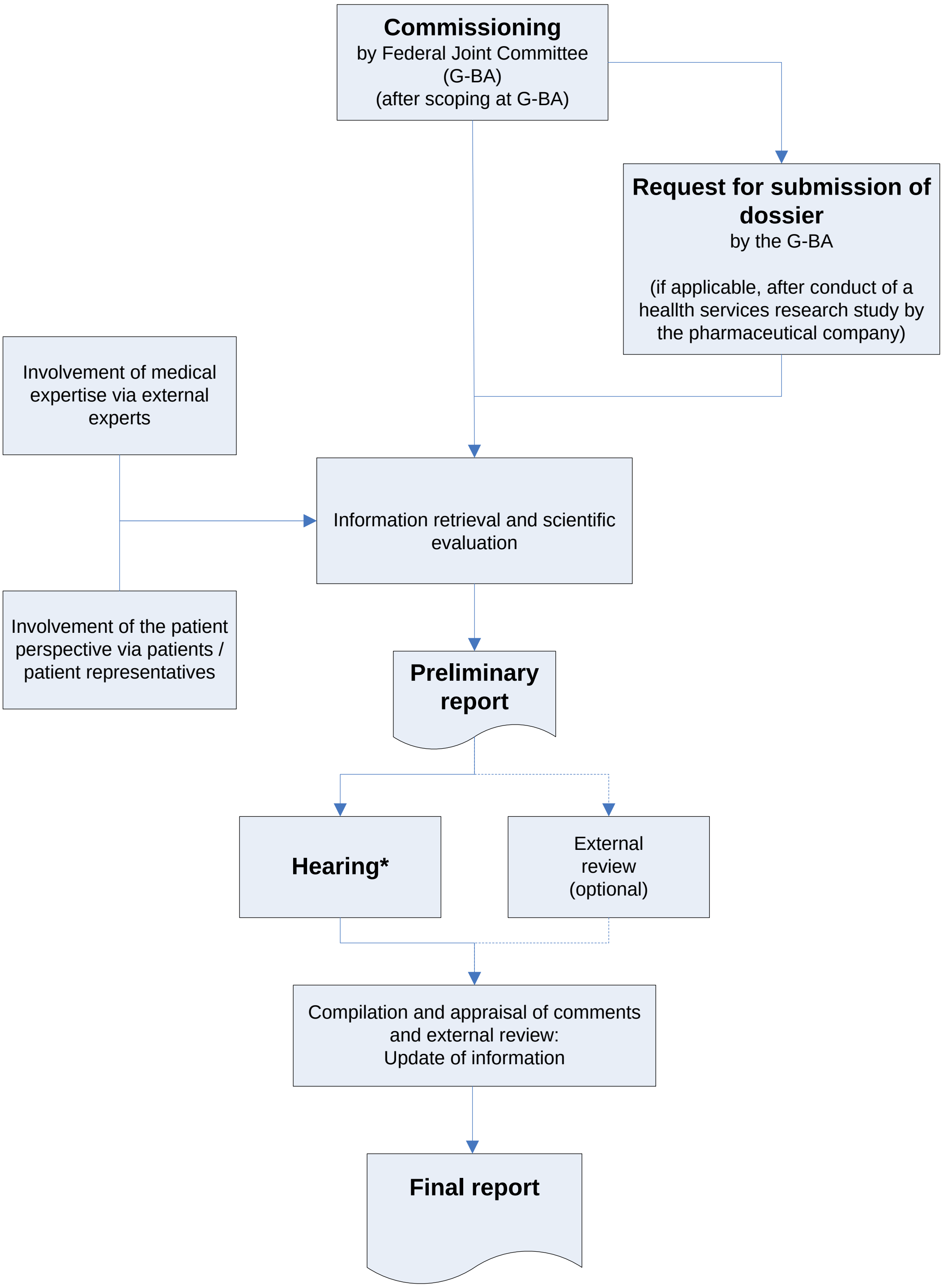
The **patient perspective is considered** on the basis of a questionnaire sent to patients/patient organizations at the beginning of the assessment. In its assessment the Institute considers the information provided in this questionnaire, e.g. on relevant outcomes and important subgroups. Patients/patient organizations are identified via the relevant organizations named in §140f SGB V.

The basis of the assessment is the dossier submitted to the G-BA by the pharmaceutical company and then forwarded to the Institute. The Institute may optionally perform its **own literature search** to support the assessment.

The preparation of the **dossier assessment** is the final step in the process. In accordance with §35a SGB V, the assessment must be completed no later than 3 months after the relevant date for the submission of the dossier. After its completion the dossier assessment is delivered to the G-BA. Shortly afterwards it is subsequently forwarded to the Foundation’s Board of Directors, the Foundation Council and the Foundation’s Board of Trustees and then published on the Institute’s website.

2.1.4 Health economic evaluation according to §35b SGB V

The procedure for an HEE according to §35b SGB V is presented in Figure 4. All working steps are performed under the Institute’s responsibility. The procedure regularly involves external experts. If necessary, the Institute’s Scientific Advisory Board may also be involved. The internal quality assurance process is not outlined in this flow chart.


Commissioning
by Federal Joint Committee
(G-BA)
(after scoping at G-BA)
Request for submission of dossier
by the G-BA
(if applicable, after conduct of a heallth services research study by the pharmaceutical company)
Involvement of medical expertise via external experts
Information retrieval and scientific evaluation
Involvement of the patient perspective via patients / patient representatives
Preliminary report
Hearing*
External review (optional)
Compilation and appraisal of comments and external review:
Update of information
Final report

* The hearing is conducted by inviting written comments. In addition, an optional oral scientific debate may be held to discuss any unclear aspects of the written comments.

Figure 4: Procedure for the health economic evaluation according to §35b SGB V

Before **commissioning** by the G-BA, the G-BA prepares the main contents of the commission (during the course of “scoping”, see Section 4.9.1) and gives those entitled to comment the opportunity to do so. Simultaneously to commissioning, in its decision the G-BA discloses whether health services research studies that the G-BA agreed upon with the pharmaceutical company are to be considered.

In parallel the G-BA requests the **submission of the dossier** by the pharmaceutical company. This dossier is considered in the assessment.

The results of the information retrieval and the scientific assessment are presented in the **preliminary report**. In the assessment of content, as a rule medical expertise is involved via external experts and the patient perspective is involved via patients and/or patient organizations.

Medical expertise is primarily obtained on the basis of a questionnaire sent to external experts at the beginning of the assessment. The feedback provided by external experts is considered in the assessment. Moreover, if necessary, external experts may be involved to clarify specific questions arising during the course of the assessment. External experts are identified via the Institute’s own database for external experts (see Section 2.2.1).

The **patient perspective** is determined on the basis of a questionnaire sent to patients and/or patient organizations at the beginning of the assessment. The information provided in this questionnaire (e.g. on relevant outcomes and on important subgroups) is considered in the assessment. Patients and/or patient organizations are identified via the relevant organizations named in § 140f SGB V.

The **preliminary report** includes the preliminary recommendation to the G-BA. After completion it is first forwarded to the G-BA, the Foundation’s Board of Directors, the Foundation Council, and the Board of Trustees. The preliminary report is published on the Institute’s website soon after it is sent to the G-BA. For a period of 3 weeks, the public is then given the **opportunity to submit written comments (hearing)**. In particular the results of the retrieval and assessment of information presented in the preliminary report are the subject of the commenting procedure. Optionally, an oral scientific debate with those submitting comments may be held. This debate serves the potentially necessary clarification of aspects of the written comments and aims at improving the scientific quality of the final report.

The **final report**, which is based upon the preliminary report and contains the assessment of the scientific findings (considering the results of the hearing on the preliminary report), represents the concluding product of the work on the commission. The final report must be forwarded to the G-BA within 3 months after the initiation of the commenting procedure on the preliminary report (see the G-BA’s Code of Procedure 5th Chapter §31 [211]). The final report and the documentation of the hearing on the preliminary report are first forwarded to the G-BA, as well as to the Foundation’s Board of Directors and Foundation Council, and subsequently forwarded to the Foundation's Board of Trustees. These documents are then published on the Institute's website. If comments are received on final reports that contain substantial evidence not considered, or if the Institute receives information on such evidence from other sources, the G-BA will be sent well-founded information on whether, in the Institute's opinion, a new commission on the topic is necessary (if appropriate, a report update). The G-BA then decides on the commissioning of the Institute. Such an update is conducted according to the general methodological and procedural requirements for the Institute's products.

2.1.5 Assessment of potential

The procedure for the production of an assessment of the potential of a non-drug intervention is presented in Figure 5. All working steps are performed under the responsibility of the Institute. External experts can be involved in the procedure (see Section 2.2.1). The internal quality assurance process is not presented in this flowchart.

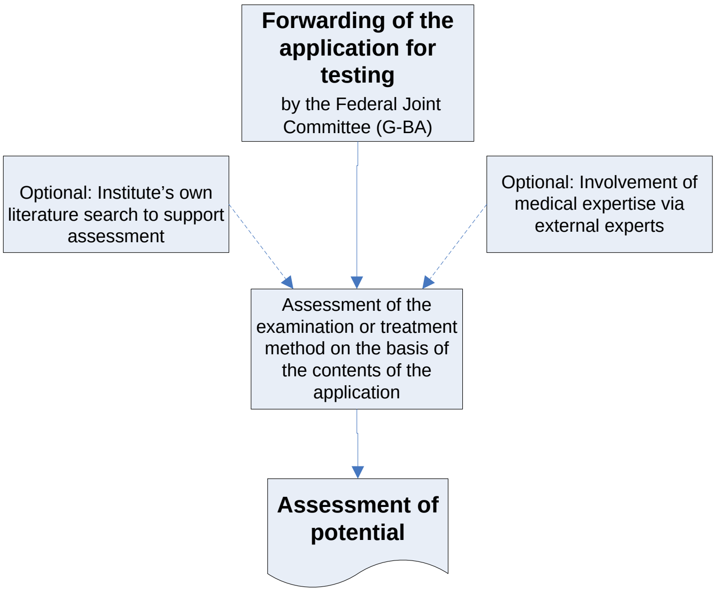

Forwarding of the application for testing
by the Federal Joint Committee (G-BA)
Optional: Institute's own literature search to support assessment
Optional: Involvement of medical expertise via external experts
Assessment of the examination or treatment method on the basis of the contents of the application
Assessment of potential

Figure 5: Procedure for the production of an assessment of potential

After the **forwarding of the application for testing** by the G-BA, the assessment of the content of the application is performed under the Institute's responsibility. External medical expertise can be involved for this purpose. This is done in the same way as in dossier assessments, but under consideration of the specific requirements for the protection of the strict confidentiality within the framework of assessments of potential.

The basis of the assessment is the application submitted by the applicant to the G-BA and then forwarded to the Institute. To support the assessment the Institute may optionally conduct its **own literature search**. As the key points of the testing study are an optional part of the application, the Institute may specify these points if the applicant provides no corresponding information.

The process is completed by the preparation of the **assessment of potential**. According to §137e SGB V, within 3 months the G-BA must make a decision on the potential of the examination or treatment method applied for. As a rule, assessments of potential are therefore completed by the Institute within 6 weeks. After completion, the assessment of potential is sent to the G-BA. The assessment is not published as, according to §137e SGB V, the assessment procedure is subject to strict confidentiality. The assessment of potential is only published if the G-BA issues a testing directive during the further course of the procedure.

2.1.6 Addendum

The procedure for the production of an addendum is presented in Figure 6. All working steps are performed under the responsibility of the Institute, involving the Institute's Scientific Advisory Board where appropriate. The internal quality assurance process is not outlined in this flow chart.

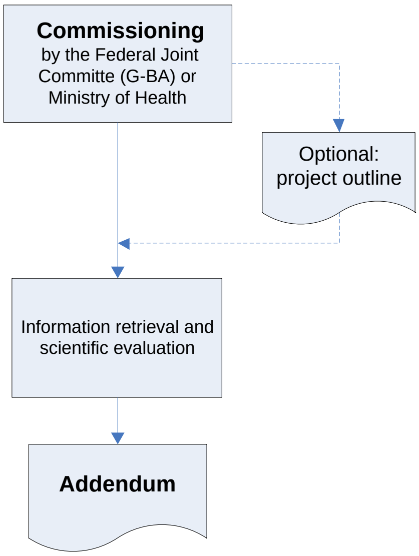

Commissioning
by the Federal Joint Committe (G-BA) or Ministry of Health
Optional: project outline
Information retrieval and scientific evaluation
Addendum

Figure 6: Procedure for the production of an addendum

An addendum can be commissioned if the need for additional work on the commission arises during the consultations on products completed by the Institute. Depending on the type and extent of the research question it may be meaningful to prepare a **project outline** in which the main steps of the information retrieval and scientific assessment are summarized. The project outline is not published.

In the work on the **addendum**, depending on the type and extent of the research question, it may be meaningful to involve those external experts who were involved in preparing the underlying product of the Institute.

The procedure for publication of an addendum follows that of the original product of the Institute. For example, an addendum on reports is first sent to the contracting agency, as well as to the Foundation Council and the Board of Directors. It is usually forwarded to the

Foundation's Board of Trustees 1 week later and published on the Institute's website a further 3 weeks later.

2.1.7 Health information

The Institute produces **health information** for the general public in various formats, which are presented in more detail in Section 6.4.

This information is provided to the public primarily via the website www.gesundheitsinformation.de (and the English-language version www.informedhealthonline.org). The website's main focus is on topics related to health and illness. Depending on breadth and depth, one topic may combine several different types of article formats.

The production process for health information is presented in Figure 7. External experts are involved in the production of the health information at various stages. Their tasks are described in more detail in Chapter 6.

The Institute's health information is produced

- in response to commissions received directly from the G-BA or Federal Ministry of Health
- as easily understandable summaries (accompanying information) of other products published by the Institute
- to fulfil its legislative responsibility to provide consumers with health information, as well as on its own initiative within the framework of the G-BA's general commission

The Institute's general commission (see Section 2.1.8) was amended in July 2006 and in March 2008 regarding the production of health information, to specifically include informing the general public. The process of selecting health information topics is described in Section 6.3.1. After deciding on the aspects the topic is to cover, the next step is the **gathering of information,** followed by **scientific review** of the identified publications. Chapter 6 describes the methodology concerning the gathering of information for the production of health information, the scientific review, and patient involvement.

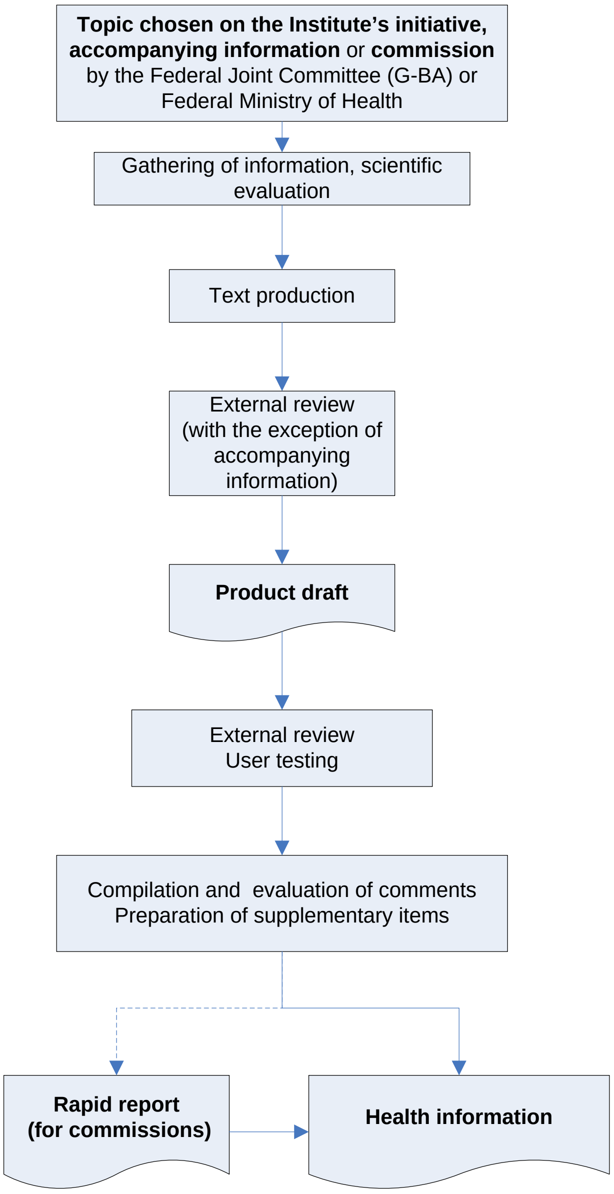

Topic chosen on the Institute's initiative, accompanying information or commission by the Federal Joint Committee (G-BA) or Federal Ministry of Health
Gathering of information, scientific evaluation
Text production
External review (with the exception of accompanying information)
Product draft
External review User testing
Compilation and evaluation of comments Preparation of supplementary items
Rapid report (for commissions)
Health information

Figure 7: Procedure for the production of health information

When it comes to the production of accompanying information, the underlying IQWiG review provides the evidence it is based on. Additional gathering of information and scientific review are optional in this case, for example as regards supplementary background information or aspects of qualitative research.

After the **text** has been produced and a departmental quality assurance process performed, the drafts are sent out for **external review**. Information accompanying IQWiG reports is reviewed internally by the respective department's project management.

Once the draft is considered by the Institute to be finished, it is sent to the contracting agency, the Board of Trustees, and the other bodies of the Institute, for limited **submission of comments** within a 1-month consultation period. The Board of Trustees includes organizations of service providers and of social partners, and the self-regulatory bodies of the supporting organizations of the G-BA; as well as representatives of organizations responsible for representing the interests of patients and self-help groups of chronically ill and disabled persons, and the Federal Government Commissioner for Patients' Affairs. Before publication, a health information article undergoes external user testing – generally at the same time as the commenting procedure. In user testing, a group of patients or potential users comment on the texts regarding their content and understandability.

The comments submitted during the consultation period and the results of the user testing are reviewed, commented on, and summarized. They may be cause for a revision of the health information submitted.

If directly commissioned by the G-BA or the Federal Ministry of Health, the health information is produced in the form of a rapid report. In this case the production and publication of the information follows the IQWiG's standard method illustrated in Section 2.1.2. The rapid report is initially sent to the contracting agency, the Board of Directors and the Foundation Board and then forwarded to the Board of Trustees, usually 4 weeks later. It is then published on the Institute's website www.iqwig.de (usually a further 4 weeks later). Usually, the corresponding health information is subsequently published on www.gesundheitsinformation.de/www.informedhealthonline.org. The readily understandable information explaining the G-BA directives is published on www.gesundheitsinformation.de/www.informedhealthonline.org only after publication of the directives themselves.

Corrections, improvements, and updates of published health information are primarily made internally. If extensive or substantive changes to content are made, external experts may be involved. A more detailed description of the updating mechanisms is provided in Chapter 6.

2.1.8 Working paper

The procedure for the production of a **working paper** is presented in Figure 8. All working steps are performed under the responsibility of the Institute, involving external experts or the Institute's Scientific Advisory Board, where appropriate. The internal quality assurance process is not presented in this flow chart.

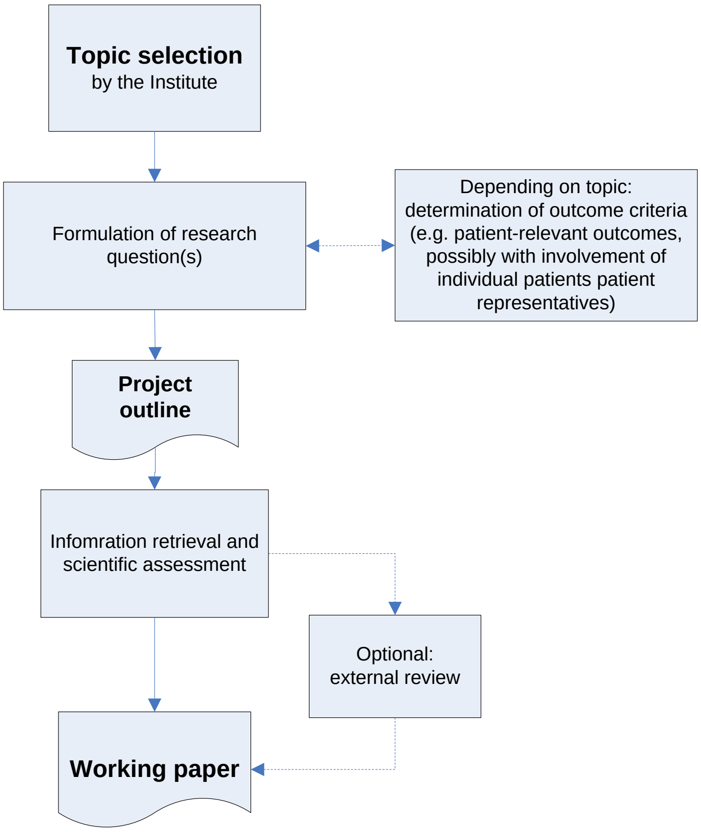

Topic selection
by the Institute
Formulation of research question(s)
Depending on topic: determination of outcome criteria (e.g. patient-relevant outcomes, possibly with involvement of individual patients patient representatives)
Project outline
Infomration retrieval and scientific assessment
Optional: external review
Working paper

Figure 8: Procedure for the production of a working paper

The production of working papers is conducted (among other things) within the framework of the **general commission** awarded by the G-BA on 21 December 2004. This commission was further specified and adapted in July 2006 and March 2008 with regard to the production of health information. According to the general commission, the Institute was commissioned "by means of documenting and analysing the relevant literature, continuously to study and assess medical developments of fundamental importance and their effects on the quality and efficiency of health care in Germany, and to relay its findings to the G-BA on a regular basis. In this context, the G-BA assumes that, within the framework of the tasks assigned in accordance with §139a (3) SGB V, the Institute will work not only on individual commissions awarded by the G-BA, but will also take on scientific projects on its own responsibility, and relay essential information on relevant health care developments to the G-BA so that it can fulfil its legal obligations. Against the background of this information, the Institute will also develop concrete proposals for individual commissions that it considers relevant."

The need to conduct independent scientific projects therefore results from the Institute's legal remit and the general commission. This also includes projects on the further development of methods, which can also be published as working papers.

The **topic selection** takes place within the Institute, particularly on the basis of the criteria defined in the general commission. The formulation of the research question may take place by involving patient organizations or seeking the opinion of individual affected patients, especially for the definition of patient-relevant outcomes. The project outline is then prepared.

The **project outline** summarizes the main steps in the information retrieval and scientific assessment. It forms the basis for the preparation of the working paper. The project outline is not published.

The working paper presents the results of the information retrieval and scientific assessments. The quality assurance process can (optionally) include an external review. After completion the working paper is first sent to the G-BA as well as to the Foundation's Board of Directors and Foundation Council. It is then forwarded to the Foundation's Board of Trustees (usually a week later) and after 3 further weeks published on the IQWiG website. If comments on working papers are received that contain substantial unconsidered evidence, or if the Institute receives such evidence from other sources, the Institute assesses whether it considers it necessary to update the document or not. The general methodological and procedural requirements for the Institute's products apply to such an update.

2.2 General aspects in the preparation of products

The following procedures and aspects that are valid for all products are presented in this chapter:

- selection of external experts for collaboration in the preparation of products
- guarantee of scientific independence in the preparation of products
- review of products
- publication of products

2.2.1 Selection of external experts

In accordance with its legal remit, the Institute involves external experts in its work. External experts are persons who are awarded research commissions within the framework of the preparation of the Institute's products or their review or who advise the Institute on medical or other topic-related research questions. The Institute awards these commissions following general procurement principles in a transparent and non-discriminating competition.

Announcements for research commissions according to §139b (3) SGB V are published on the Institute's website. Exceptions are possible, for example, in the case of a particularly urgent commission.

Commissions with a volume above the current threshold value of the procurement regulations of the European Union (EU) are advertised throughout the EU. The specific requirements regarding the suitability of applicants are published in the corresponding announcements or tendering documents.

The commissioning of external experts for benefit assessments according to §35a SGB V, HEEs according to §35b SGB V, assessments of potential according to §137e SGB V, and the production of health information is conducted on the basis of information provided by interested persons in a database for external experts. For inclusion in the database for external experts, the Institute's website offers an access point via which interested experts can enter their profile, including details of their specialty and professional expertise. For the projects to be awarded, in each case the most suitable applicant of the relevant specialty is selected from this expert database by means of a criteria list and then commissioned. Further information on the selection procedure is published on the Institute's website.

2.2.2 Guarantee of professional independence

The scientific and professional independence of the Institute and of the products it is responsible for and publishes have their legal basis in §139a SGB V, as well as in the *Charter* of the Foundation.

A) Guarantee of internal professional independence

The Institute's scientific staff are prohibited from performing paid external assignments that could in principle query their professional independence. All external assignments must be approved by the Institute's Management. External assignments in the broadest sense also include unpaid honorary positions such as positions on boards in organizations and societies.

B) Guarantee of the independence of external experts

Before a contract is signed between the Institute and an external expert or external institution with regard to the preparation of a product, in accordance with §139b SGB V, "all connections to associations and contract organizations, particularly in the pharmaceutical and medical devices industries, including details on the type and amount of any remuneration received" must be disclosed to the Institute.

Following the usual practice in research to disclose such connections as potential conflicts of interest [359,366], within the framework of the selection of external experts, the Institute interprets this regulation as a responsibility to assess these disclosures with regard to the professional independence and impartiality of applicants. In this context, the Institute assesses whether a conflict of interest for the specific topic of a commission exists due to the financial connections reported. If this is the case, in a second step it is assessed whether this conflict of interest leads to serious concerns with regard to appropriate collaboration on the commission. If this is the case, collaboration on the topic of this commission is usually not possible or only possible under specific provisions. As this assessment is performed in relation to a specific commission, collaboration on topics of other commissions is indeed possible. The further process for the selection of external experts is outlined in Section 2.2.1.

The main basis of the assessment of conflicts of interest is self-disclosure using the *Form for disclosure of potential conflicts of interest*, which is published on the Institute's website. Self-disclosure refers to the following 6 types of financial connections:

- dependent occupation/employment
- advisory activities
- payments, e.g. for presentations, comments, as well as organization and/or participation in conferences and seminars
- financial support for research activities, other scientific services or patent registrations
- other financial or other cash-value support (e.g. for equipment, staff or travel expenses, without providing scientific services in return)
- shares, equity warrants or other shares in a business

The Institute reserves the right to draw upon additional information and verify the completeness and correctness of the reported information.

The names of external experts involved in the preparation of the Institute's products are usually published in these products. As a matter of principle, these publications are freely accessible via the Institute's website. The information on conflicts of interest is only published in a summarized form. In this context, for the types of connections covered by the *Form for disclosure of potential conflicts of interest*, it is only stated whether this type of connection existed or not. Specific details, for example, concerning business partners or the amount of any remuneration received, are not published.

2.2.3 Review of the Institute's products

The review of the Institute's products aims in particular at ensuring their high scientific quality. Moreover, other aims may be relevant for individual products, such as comprehensibility for the general public.

All products (including interim ones) are subjected to a comprehensive multi-stage internal quality assurance process. In addition, during the preparation of products, an external review procedure may be performed as an optional further quality assurance step. The choice of internal and external reviewers is primarily made on the basis of their methodological and/or professional expertise.

External reviewers can be identified by a literature search, the expertise of the project group, by contacting scientific societies, or by application during the tendering procedure for work on a commission, etc. In each case, external reviewers must also disclose potential conflicts of interest.

External reviewers are selected by the Institute and their number is not limited. The external reviews are assessed with regard to their relevance for the particular product; they are not published. The names of the external reviewers of final reports and rapid reports are usually published in these documents, including a presentation of their potential conflicts of interests, in analogy to the procedure for external experts.

In addition to the external quality assurance processes described above with the involvement of reviewers selected and commissioned by the Institute, an open and independent reviewing process is guaranteed by the publication of the Institute's products and the associated opportunity to submit comments.

2.2.4 Publication of the Institute's products

One of the Institute's main tasks is to determine the available evidence on a topic by performing a careful assessment of the information available, and to publish the results of this assessment. It is legally specified that the Institute "must at regular intervals publicly report on its working processes and results, including the bases for decision-making" (§139a (4) SGB V).

To maintain the Institute's independence, it must be ruled out that the contracting agencies or any other interested third parties can exert any influence on the content of the reports. This could lead to conflation of scientific findings with political and/or economic aspects and/or interests. At the same time, it must be avoided that the Institute itself withholds certain findings. All the results obtained by the Institute within the framework of its legal responsibilities are therefore published as soon as possible (with the exception of assessments of potential, see §19 of the G-BA's Code of Procedure [211]). In the case of reports, this also includes the report plan. Product-specific features are noted in those sections in which the procedures are described. In justified exceptional cases, timelines may deviate from the stipulated norm (period between completion and publication of a document).

Unless otherwise agreed, all copyright is held by the Institute.

3 Benefit assessment of medical interventions

3.1 Patient-relevant medical benefit and harm

3.1.1 Definition of patient-relevant medical benefit and harm

The term **“benefit”** refers to positive causal effects, and the term **“harm”** refers to negative causal effects of a medical intervention on patient-relevant outcomes (see below). In this context, “causal” means that it is sufficiently certain that the observed effects can be ascribed solely to the intervention to be tested [595]. The terms “benefit” and “harm” refer to a comparison with a placebo (or another type of sham intervention) or no treatment.

In the case of a comparison between the medical intervention to be assessed and a clearly defined alternative medical intervention, the following terms are used in the comparative assessment of beneficial or harmful aspects (the terms are always described from the point of view of the intervention to be assessed):

- Beneficial aspects:
  - In the case of a greater benefit, the term “added benefit” is used.
  - In the case of lesser or comparable benefit, the terms “lesser” or “comparable” benefit are used.
- Harmful aspects:
  - The terms “greater”, “comparable” and “lesser” harm are used.

The assessment of the evidence should preferably come to a clear conclusion that either there is proof of a(n) (added) benefit or harm of an intervention, or there is proof of a lack of a(n) (added) benefit or harm, or there is no proof of a(n) (added) benefit or harm or the lack thereof, and it is therefore unclear whether the intervention results in a(n) (added) benefit or harm.

In addition, in the case of (added) benefit or harm that is not clearly proven, it may be meaningful to perform a further categorization as to whether at least “indications” or even only “hints” of an (added) benefit or harm are available (see Section 3.1.4).

As the benefit of an intervention should be related to the patient, this assessment is based on the results of studies investigating the effects of an intervention on patient-relevant outcomes. In this connection, “patient-relevant” refers to how a patient feels, functions or survives [44]. Consideration is given here to both the intentional and unintentional effects of the intervention that in particular allow an assessment of the impact on the following patient-relevant outcomes to determine the changes related to disease and treatment:

1) mortality
2) morbidity (symptoms and complications)

3) health-related quality of life

These outcomes are also named in SGB V as outcomes primarily to be considered, for example, in § 35 (1b) SGB V. As supplementary information, consideration can be given to the time and effort invested in relation to the disease and the intervention. This also applies to patient satisfaction, insofar as health-related aspects are represented here. However, a benefit or added benefit cannot be determined on the basis of these 2 outcomes alone.

For all listed outcomes it may be necessary that an assessment is made in relation to information on how other outcomes are affected by the intervention. In the event of particularly serious or even life-threatening diseases, for example, it is usually not sufficient only to demonstrate an improvement in quality of life by application of the intervention to be assessed, if at the same time it cannot be excluded with sufficient certainty that serious morbidity or even mortality are adversely affected to an extent no longer acceptable. This is in principle consistent with the ruling by the highest German judiciary that certain (beneficial) aspects must be assessed only if therapeutic effectiveness has been sufficiently proven [81]. On the other hand, in many areas (particularly in palliative care) an impact on mortality cannot be adequately assessed without knowledge of accompanying (possibly adverse) effects on quality of life.

In accordance with §35b (1) Sentence 4 SGB V, the following outcomes related to patient benefit are to be given appropriate consideration: increase in life expectancy, improvement in health status and quality of life, as well as reduction in disease duration and adverse effects. These dimensions of benefit are represented by the outcomes listed above; for example, the improvement in health status and the reduction in disease duration are aspects of direct disease-related morbidity; the reduction in adverse effects is an aspect of therapy-related morbidity.

Those outcomes reliably and directly representing specific changes in health status are primarily considered. In this context, individual affected persons as well as organizations of patient representatives and/or consumers are especially involved in the topic-related definition of patient-relevant outcomes. In the assessment of quality of life, only instruments should be used that are suited for application in clinical trials and have been evaluated accordingly [174]. In addition, valid surrogate endpoints can be considered in the benefit assessment (see Section 3.1.2).

Both beneficial and harmful aspects can have different relevance for the persons affected; these aspects may become apparent through qualitative surveys or the Institute’s consultations with affected persons and organizations of patient representatives and/or consumers in connection with the definition of patient-relevant outcomes (examples of corresponding methods are listed at the end of Section 3.1.4). In such a situation it may be meaningful to establish a hierarchy of outcomes. General conclusions on benefit and harm are then primarily based on proof regarding higher-weighted outcomes. Planned subgroup and sensitivity analyses are then primarily conducted for higher-weighted outcomes, whereas such analyses are not routinely conducted for the remaining ones.

Diagnostic tests can be of indirect benefit by being a precondition for therapeutic interventions through which it is possible to achieve an effect on the patient-relevant outcomes mentioned above. The precondition for the benefit of such tests is therefore the existence and the proven benefit of a treatment for patients, depending on the test result.

Interventions can also have consequences for those indirectly affected, for example, relatives and carers. If appropriate, these consequences can also be considered within the framework of the Institute’s reports.

The term **“benefit assessment”** refers to the whole process of the assessment of medical interventions with regard to their positive and negative causal effects compared with a clearly defined alternative treatment, a placebo (or a different type of sham intervention), or no treatment. In this context, beneficial and harmful aspects of an intervention are initially assessed on an outcome-specific basis and then presented. In addition, a combined evaluation of outcome-related beneficial and harmful aspects is possible (see Section 3.1.4) so that, for example, when the effects on all other outcomes have been analysed, the outcome-specific “lesser harm” from an intervention (in terms of a reduction in adverse effects) can lead to the balanced conclusion of an “added benefit”.

3.1.2 Surrogates of patient-relevant outcomes

Surrogate endpoints are frequently used in medical research as a substitute for patient-relevant outcomes, mostly to arrive at conclusions on patient-relevant (added) benefits earlier and more simply [15,194,444]. Most surrogate endpoints are, however, unreliable in this regard and can be misleading when used in a benefit assessment [102,219,227]. Surrogate endpoints are therefore normally considered in the Institute’s benefit assessments only if they have been validated beforehand by means of appropriate statistical methods within a sufficiently restricted patient population and within comparable interventions (e.g. drugs with a comparable mode of action). A surrogate endpoint can be regarded as valid if the effect of an intervention on the patient-relevant outcome to be substituted is explained to a sufficient degree by the effect on the surrogate endpoint [28,586]. The necessity to evaluate surrogate endpoints may have particular relevance within the framework of the early benefit assessment of drugs (see Section 3.3.3), as regulatory approval procedures primarily investigate the efficacy of a drug, but not always its patient-relevant benefit or added benefit.

There is neither a standard procedure for surrogate endpoint validation nor a general best estimation method nor a generally accepted criterion which, if fulfilled, would demonstrate validity [380]. However, the current methodological literature frequently discusses correlation-based procedures for surrogate validation, with estimation of correlation measures at a study level and individual level [286]. The Institute‘s benefit assessments therefore give preference to validations on the basis of such procedures. These procedures usually require a meta-analysis of several randomized studies, in which both the effects on the surrogate endpoint and those on the patient-relevant outcome of interest are investigated [86,400]. Alternative methods [586] are only considered in justified exceptional cases.

For correlation-based procedures the following conditions are normally required to demonstrate validity: on the one hand, a high correlation between the surrogate and the patient-relevant outcome at the individual level, and on the other hand, a high correlation between effects on the surrogate and effects on the patient-relevant outcome at a study level [86,88]. As in the Institute’s benefit assessments, conclusions related to groups of patients are drawn, the assessment of the validity of a surrogate endpoint is primarily based on the degree of correlation at the level of treatment effects, i.e. the study level. In addition to the degree of correlation, for the assessment of validity of a surrogate endpoint the reliability of results of the validation process is considered. For this purpose, various criteria are drawn upon [286]. For example, associations observed between a surrogate endpoint and the corresponding patient-relevant outcome for an intervention with a specific mode of action are not necessarily applicable to other interventions used to treat the same disease, but with a different mode of action [193,219,227,380]. The studies on which the validation was based must therefore have been conducted in patient populations and interventions that allow conclusions on the therapeutic indication investigated in the benefit assessment as well as on the test intervention and comparator intervention. In order to assess transferability, in validation studies including various disease entities or interventions, analyses on heterogeneity should at least be available.

In the event that a surrogate endpoint cannot be validated conclusively (e.g. if correlation is not high enough), it is also possible to apply the “surrogate threshold effect (STE) concept” [85,286]. For this purpose, the effect on the surrogate resulting from the studies included in the benefit assessment is related to the STE [88,400].

For the Institute’s benefit assessments, conclusions on patient-relevant outcomes can be drawn from the effects on the surrogate, depending on verification of the validity of the surrogate or the evaluation of the STE. The decisive factor for the first point is the degree of correlation of the effects on the surrogate and the patient-relevant outcome and the reliability of validation in the validation studies. In the evaluation of an STE, the decisive criterion is the size of the effect on the surrogate in the studies included in the benefit assessment compared with the STE. In the case of a statistically significant effect on the surrogate endpoints, all gradations of conclusions on the (added) benefit with regard to the corresponding patient-relevant outcome according to Section 3.1.4 are possible, depending on the constellation.

Surrogate endpoints that are not valid or for which no adequate validation procedure was conducted can nevertheless be presented in the Institute’s reports. However, independent of the observed effects, such endpoints are not suited to provide proof of verification of an (added) benefit of an intervention.

Depending on the proximity to a corresponding patient-relevant outcome, the literature uses various other terms to describe surrogate endpoints (e.g. intermediate endpoint). However, we dispense with such a distinction here, as the issue of the necessary validity remains unaffected by this. In addition it should be considered that an endpoint can at the same time represent a patient-relevant outcome and, beyond this, can also be regarded as a surrogate (i.e. a substitute) for a different patient-relevant outcome.

3.1.3 Assessment of the harm of medical interventions

The use of any type of medical intervention (drug, non-drug, surgical, diagnostic, preventive, etc.) carries per se the risk of adverse effects. In this context, the term “adverse effects” refers to all events and effects representing individually perceived or objectively detectable physical or mental harm that may to a greater or lesser extent cause a short- or long-term reduction in life expectancy, an increase in morbidity, or impairment in quality of life. It should be noted that if the term “adverse effects” is used, a causal relationship to the intervention is assumed, whereas the issue of causality still remains open with the term “adverse events” [109].

The term “harm” describes the occurrence of adverse effects when using a medical intervention. The description of harm is an essential and equal component in the benefit assessment of an intervention. It ensures the informed, population-related, but also individual weighing of benefit and harm [602]. A prerequisite for this is that the effect sizes of a medical intervention can be described by means of the data available, both for its desired as well as its adverse effects, and compared with therapy alternatives, for example.

However, in a systematic review, the analysis, assessment, and reporting of the harm of a medical intervention are often far more difficult than those of the (added) benefit. This applies in particular to unexpected adverse events [109]. Studies are typically designed to measure the effect of a medical intervention on a few predefined outcomes. In most cases, these are outcomes representing effectiveness, while adverse effects are concomitantly recorded as adverse events. The results for adverse events depend heavily on the underlying methods for data collection. For example, explicit queries on defined adverse events normally result in the determination of higher event rates than do general queries [41,304]. To detect unexpected adverse events in particular, general queries about the well-being of patients are however required. In addition, studies designed to specifically detect rare, serious adverse effects (including the description of a causal relationship to the medical intervention) are considerably underrepresented in medical research [48,164,303]. Moreover, reporting of adverse events in individual studies is of poor quality, which has also led to amendment of the CONSORT[12] statement for RCTs [302]. Finally, the systematic assessment of the adverse effects of an intervention is also made more difficult by the fact that the corresponding coding in bibliographic databases is insufficient, so that the specific search for relevant scientific literature often produces an incomplete picture [127].

[12] Consolidated Standards of Reporting Trials

The obstacles noted above often make the investigation of harm more difficult. In cases where complete clinical study reports are available for the assessment, at least sufficient data transparency is also given for adverse events. However, it is still necessary to find a meaningful balance between the completeness of the evaluation of aspects of harm and the resources invested. Consequently, it is necessary to limit the evaluation and reporting to relevant adverse effects. In particular, those adverse effects can be defined as relevant that may

- completely or almost completely offset the benefit of an intervention
- substantially vary between 2 or more otherwise equivalent treatment options
- occur predominantly with treatment options that may be particularly effective
- have a dose-effect relationship
- be regarded by patients as especially important
- be accompanied by serious morbidity or even increased mortality, or be associated with substantial impairment in quality of life

The Institute observes the following principles when evaluating and reporting adverse effects: In the benefit assessment, the initial aim is to compile a selection of potentially relevant adverse effects that are essential in deciding for or against the use of the intervention to be assessed. In this context, the selection of adverse effects and events is made in accordance with the criteria outlined above. This compilation is made within the framework of the preliminary literature search for the particular research question, especially on the basis of data from controlled intervention studies in which the benefit of the intervention was specifically investigated. In addition, and if appropriate, the compilation is made on the basis of available epidemiological data (e.g. from cohort or case-control studies), as well as pharmacovigilance and regulatory data, etc. In individual cases, data obtained from animal trials and experiments to test pathophysiological constructs may be useful. The compilation of potentially relevant adverse effects described above forms the foundation for assessment of harm on the basis of the studies included in the benefit assessment. In this context, if possible and meaningful, pooled analyses (e.g. overall rates of serious adverse events) may also be drawn upon.

3.1.4 Outcome-related assessment

The benefit assessment and the estimation of the extent of the (un)certainty of results generally follow international EBM standards as developed, for example, by the GRADE[13] group [23].

[13] Grading of Recommendations, Assessment, Development and Evaluation

Medical interventions are compared with other interventions, sham interventions (e.g. placebo), or no intervention in respect of their effects on defined patient-relevant outcomes, and their (added) benefit and harm are described in summary. For this purpose, on the basis of the analysis of the scientific data available, for each predefined patient-relevant outcome separately a conclusion on the evidence base of the (added) benefit and harm is drawn in 4 levels with regard to the respective certainty of the conclusion: The data provide either "proof" (highest certainty of conclusions), an "indication" (medium certainty of conclusions), a "hint" (weakest certainty of conclusions) in respect of the benefit or harm of an intervention, or none of these 3 situations applies. The latter is the case if no data are available or the data available do not allow any of the other 3 conclusions to be drawn.

Depending on the research question, the conclusions refer to the presence or lack of a(n) (added) benefit or harm. The prerequisite for conclusions on the lack of a(n) (added) benefit or harm are well-founded definitions of irrelevance ranges (see Section 8.3.6).

The certainty of results is an important criterion for the inference of conclusions on the evidence base. In principle, every result from an empirical study or systematic review of empirical studies is potentially uncertain and therefore the certainty of results must be examined. In this context, one distinguishes between qualitative and quantitative certainty of results. The qualitative certainty of results is impaired by systematic errors (bias; see Section 8.3.11) such as information errors, selection errors and confounding. The quantitative certainty of results is influenced by random errors caused by sampling (statistical uncertainty).

The qualitative certainty of results is thus determined by the study design, from which evidence levels can be inferred (see Section 8.1.3). It is also determined by (outcome-related) measures for further prevention or minimization of potential bias, which must be assessed depending on the study design (see Section 8.1.4). Such measures include, for example, the blinded assessment of outcomes, an analysis based on all included patients (potentially supported by the application of adequate imputation methods for missing values), and, if appropriate, the use of valid measurement instruments.

The quantitative certainty of results is directly connected to the sample size (i.e. the number of patients investigated in a study or the number of [primary] studies included in a systematic review), as well as to the variability observed within and between studies. If the underlying data allow for this, the statistical uncertainty can be quantified and assessed as the standard error or confidence interval of parameter estimates (precision of the estimate).

The Institute uses the following 3 categories to grade the degree of qualitative certainty at the individual study level and outcome level:

- **high qualitative certainty of results:** results on an outcome from a randomized study with a low risk of bias

- **moderate qualitative certainty of results**: results on an outcome from a randomized study with a high risk of bias
- **low qualitative certainty of results:** results on an outcome from a non-randomized comparative study

In the inference of the evidence base for an outcome, the number of available studies, their qualitative certainties of results, as well as the effects found in the studies are of crucial importance. If at least 2 studies are available, it is first distinguished whether, due to existing heterogeneity within a meta-analysis (see Section 8.3.8), a common effect estimate can be meaningfully formed or not. In the case of homogenous results that can be meaningfully pooled, the common effect estimate must be statistically significant to infer proof, an indication or a hint according to the existing certainty of results. If the estimated results are too heterogeneous to meaningfully form a pooled common effect estimate, one distinguishes between effects that are "not in the same direction", "moderately in the same direction" and "clearly in the same direction". These are defined as follows:

Effects in the same direction are present if the prediction interval for displaying heterogeneity in a meta-analysis with random effects (see Section 8.3.8) is presented and does not cover the zero effect. In other cases (no presentation of the prediction interval or this interval covers the zero effect) effects in the same direction are present in the following situation:

The effect estimates of 2 or more studies point in the same direction. For these "directed" studies, all of the following conditions apply:

- The overall weight of these studies is ≥ 80%.
- At least 2 of these studies show statistically significant results.
- At least 50% of the weight of these studies is based on statistically significant results.

In this context, the weights of these studies generally come from a meta-analysis with random effects (see Section 8.3.8). If no meta-analysis is meaningful, the relative sample size corresponds to the weight.

If effects in the same direction are moderately or clearly in the same direction, if possible, a decision is made on the basis of the location of the prediction interval. As the prediction interval is generally only presented if at least 4 studies are available (see Section 8.3.8), the classification into effects that are moderately or clearly in the same direction depends on the number of studies.

- <u>2 studies:</u> Effects in the same direction are always clearly in the same direction.
- <u>3 studies:</u>
  - All studies show statistically significant results. The effects in the same direction are clearly in the same direction.

  - Not all of the 3 studies show statistically significant results. The effects in the same direction are moderately in the same direction.

- <u>4 or more studies:</u>
  - All studies show statistically significant results in the same direction of effects: The effects in the same direction are clearly in the same direction.
  - The prediction interval does not cover the zero effect: The effects in the same direction are clearly in the same direction.
  - The prediction interval covers the zero effect: The effects in the same direction are moderately in the same direction.

For the case that the available studies show the same qualitative certainty of results or only one study is available, with these definitions the regular requirements for the evidence base to infer conclusions with different certainties of conclusions can be specified. As described above, the Institute distinguishes between 3 different certainties of conclusions: “proof”, “indication” and “hint”.

A conclusion on proof generally requires that a meta-analysis of studies with a high qualitative certainty of results shows a corresponding statistically significant effect. If a meta-analysis cannot be conducted, at least 2 studies conducted independently of each other and showing a high qualitative certainty of results and a statistically significant effect should be present, the results of which are not called into question by further comparable studies with a high certainty of results (consistency of results). These 2 studies do not need to have an exactly identical design. Which deviations in design between studies are still acceptable depends on the research question. Accordingly, a meta-analysis of studies with a moderate qualitative certainty of results or a single study with a high qualitative certainty of results can generally provide only an indication, despite statistically significant effects.

On the basis of only one study, in exceptional cases proof can be inferred for a specific (sub)population with regard to an outcome. This requires the availability of a clinical study report according to the International Conference on Harmonization (ICH) guidelines and the fulfilment of the other requirements stipulated for proof. In addition, the study must fulfil the following specific requirements:

- The study is a multi-centre study with at least 10 centres.
- The effect estimate observed has a very small corresponding p-value ($p < 0.001$).
- The result is consistent within the study. For the (sub)population of interest, analyses of different further subpopulations are available (particularly subpopulations of study centres), which in each case provide evaluable and sufficiently homogeneous effect estimates. This assessment of consistency is only possible for binary data if a certain minimum number of events has occurred.

- The analyses for the subpopulations addressed above are available for all relevant outcomes, i.e. these analyses are not restricted to individual selected outcomes.

It is possible that in the case of the existence of only one study, which alone provides only an indication or a hint, the evidence base may be changed by additional indirect comparisons. However, high methodological demands must be placed on indirect comparisons (see Section 8.3.9). In addition, in the case of a homogeneous data situation, it is possible that by adding indirect comparisons the precision of the effect estimate increases, which plays an important role when determining the extent of added benefit (see Section 3.3.3).

A meta-analysis of studies with a low qualitative certainty of results or an individual study with a moderate qualitative certainty of results (both with a statistically significant effect) generally only provides a hint.

An overview of the regular operationalization is shown in Table 2. In justified cases further factors influence these evaluations. The assessment of surrogate endpoints (see Section 3.1.2), the presence of serious deficiencies in study design or justified doubts about the transferability to the treatment situations in Germany may, for example, lead to a reduction in the certainty of conclusions. On the other hand, great effects or a clear direction of an existing risk of bias, for example, can justify an increase in certainty.

Table 2: Certainty of conclusions regularly inferred for different evidence situations if studies with the same qualitative certainty of results are available

		Number of studies				
		1 (with statistically significant effect)	≥ 2			
			Homogeneous	Heterogeneous		
			Meta-analysis statistically significant	Effects in the same direction[a]		
				Clear	Moderate	No
Qualitative certainty of results	High	Indication	Proof	Proof	Indication	–
	Moderate	Hint	Indication	Indication	Hint	–
	Low	–	Hint	Hint	–	–

a: See text for explanation of term.

If several studies with a different qualitative certainty of results are available, then first only the studies with the higher-quality certainty of results are examined, and conclusions on the evidence base are inferred on this basis according to Table 2. In the inference of conclusions on the evidence base for the whole study pool the following principles then apply:

- The conclusions on the evidence base, when restricted to higher-quality studies, are not weakened by the addition of the other studies, but at best upgraded.

- The confirmation (replication) of a statistically significant result of a study with a high qualitative certainty of results, which is required to infer proof, can be provided by one or more results of moderate (but not low) qualitative certainty of results. In this context the weight of the study with a high qualitative certainty of results should have an appropriate size (between 25 and 75%).
- If the meta-analytical result for the higher-quality studies is not statistically significant or if no effects in the same direction are shown in these studies, then conclusions on the evidence base are to be inferred on the basis of results of the whole study pool, whereby the certainty of conclusions is determined by the minimum qualitative certainty of results of all studies included.

According to these definitions and principles, a corresponding conclusion on benefit is inferred for each outcome separately. Considerations on the assessment across outcomes are presented in the following section (see Section 3.1.5).

3.1.5 Summarizing assessment

These conclusions, drawn separately for each patient-relevant outcome within the framework of the deduction of conclusions on the evidence base, are then summarized (as far as possible) in one evaluating conclusion in the form of a weighing of benefits and harms. If proof of a(n) (added) benefit and/or harm exists with regard to Outcomes 1 to 3 of Section 3.1.1, the Institute presents (insofar as is possible on the basis of the data available)

1) the benefit
2) the harm
3) (if appropriate) the weighing of benefit and harm

In this context, characteristics related to age, gender, and personal circumstances are considered.

One option in the conjoint evaluation of benefit and harm is to compare the outcome-related beneficial and harmful aspects of an intervention.

In this context, the effects on all outcomes (qualitative or semi-quantitative as in the early benefit assessment according to §35a SGB V) are weighed against each other, with the aim of drawing a conclusion across outcomes with regard to the benefit or added benefit of an intervention.

A further option in the conjoint evaluation is to aggregate the various patient-relevant outcomes into a single measure or to reach an overall conclusion by weighting them. The conjoint evaluation of benefit and harm is specified depending on the topic of interest (see also Section 4.3.3).

3.2 Special aspects of the benefit assessment

3.2.1 Impact of unpublished study results on conclusions

An essential prerequisite for the validity of a benefit assessment is the complete availability of the results of the studies conducted on a topic. An assessment based on incomplete data or possibly even selectively compiled data may produce biased results [179,295] (see also Section 8.3.11).

The distortion of published evidence through publication bias and outcome reporting bias has been described comprehensively in the literature [160,390,522]. In order to minimize the consequences of such bias, the Institute has extended information retrieval beyond a search in bibliographic databases, for example, by screening trial registries. In addition, at the beginning of an assessment the Institute normally contacts the manufacturers of the drugs or medical devices to be assessed, and requests the transfer of complete information on studies investigating these interventions (see also Section 7.1.5).

This transfer of information by manufacturers can only solve the problem of bias caused by unpublished evidence if the transfer is itself not selective but complete. An incomplete transfer of information carries a risk of bias for the result of the benefit assessment. This risk should be considered by the Institute in the conclusions of a benefit assessment.

Table 3 below describes what constellations carry a risk of bias for assessment results, and what consequences arise for the conclusions of a benefit assessment.

If the data transfer was complete and no evidence is available that a relevant amount of data is missing, bias seems improbable (Scenario 1). The inferences drawn from the assessment of data can therefore be adopted without limitation in the conclusions of the benefit assessment.

Table 3: Scenarios for data transfer by third parties and consequences for the conclusions of a benefit assessment

Scenario	Data transfer by third parties (e.g. manufacturer data)	Evidence that a relevant amount of data is missing	Bias	Assessment/Impact on the conclusions
1	Complete	No	Improbable	No limitation of the conclusions of the benefit assessment
2	Incomplete	No	Possible	Conclusions are made with reservations
3	Incomplete	Yes	Probable	Description of the available and missing data; no proof (or indication or hint) of benefit or harm
4	Complete	Yes (e.g. other manufacturers, investigator-initiated trials)	Possible	Conclusions are drawn with reservations

If the data transfer is incomplete, the consequences for the conclusions depend on whether additional search steps demonstrate that a relevant amount of data is missing. If this is not the case (Scenario 2), bias may still be possible, as data transfer may have been selective and further unpublished data may exist that were not identified by the search steps. In such cases the conclusions are therefore drawn with reservations. If it was demonstrated that a relevant amount of data is missing (Scenario 3), it can be assumed that the data transfer was selective. In this situation, further analysis of the available limited data and any conclusions inferred from them with regard to benefit or harm are probably seriously biased and therefore do not form a valid decision-making basis for the G-BA. Consequently, no proof (nor indication nor hint) of a benefit or harm of the intervention to be assessed can be determined in this situation, independently of whether the available data show an effect of the intervention or not.

If the manufacturer completely transfers data and additional literature searches demonstrate that a relevant amount of data from studies inaccessible to the manufacturer is missing (Scenario 4), then no selective data transfer by the manufacturer is evident. In this situation, bias caused by missing data is still possible. The conclusions are therefore drawn with reservation.

3.2.2 Dramatic effect

If the course of a disease is certainly or almost certainly predictable, and no treatment options are available to influence this course, then proof of a benefit of a medical intervention can also be provided by the observation of a reversal of the (more or less) deterministic course of the disease in well-documented case series of patients. If, for example, it is known that it is highly probable that a disease leads to death within a short time after diagnosis, and it is described in a case series that, after application of a specific intervention, most of those affected survive for a longer period of time, then this “dramatic effect” may be sufficient to provide proof of a benefit. An example of such an effect is the substitution of vital hormones in diseases with a failure of hormone production (e.g. insulin therapy in patients with diabetes mellitus type 1). An essential prerequisite for classification as a “dramatic effect” is sufficiently reliable documentation of the fateful course of the disease in the literature and of its diagnosis in the patients included in the study to be assessed. In this context, possible harms of the intervention should also be taken into account. Glasziou et al. [214] have attempted to operationalize the classification of an intervention as a “dramatic effect”. In a first approach they propose to regard an observed effect as not explicable solely by the impact of confounding factors if it was significant at a level of 1% and, expressed as the relative risk, exceeded the value of 10 [214]. This magnitude serves as orientation for the Institute and does not represent a rigid threshold. Glasziou et al. [214] made their recommendation on the basis of results of simulation studies, according to which an observed relative risk of 5 to 10 can no longer be plausibly explained only by confounding factors. This illustrates that a corresponding threshold also depends on the attendant circumstances (among other things, the quality of studies used to determine the existence of a dramatic effect). This dependence is also reflected in the recommendations of other working groups (e.g. the GRADE group) [342].

If, in the run-up to the work on a specific research question, sufficient information is available indicating that a dramatic effect caused by the intervention to be assessed can be expected (e.g. because of a preliminary literature search), then information retrieval will also include a search for studies that show a higher uncertainty of results due to their design.

3.2.3 Study duration

Study duration is an essential criterion in the selection of studies relevant to the benefit assessment. In the assessment of a therapeutic intervention for acute diseases where the primary objective is, for example, to shorten disease duration and alleviate acute symptoms, it is not usually meaningful to call for long-term studies, unless late complications are to be expected. On the other hand, in the assessment of therapeutic interventions for chronic diseases, short-term studies are not usually suitable to achieve a complete benefit assessment of the intervention. This especially applies if treatment is required for several years, or even lifelong. In such cases, studies covering a treatment period of several years are particularly meaningful and desirable. As both benefits and harms can be distributed differently over time, in long-term interventions the meaningful comparison of the benefits and harms of an intervention is only feasible with sufficient certainty if studies of sufficient duration are available. However, individual aspects of the benefits and harms may quite well be investigated in short-term studies.

With regard to the selection criterion “minimum study duration”, the Institute primarily follows standards for demonstrating the effectiveness of an intervention. In the assessment of drugs, the Institute will in particular resort to information provided in guidelines specific to therapeutic indications, which are published by regulatory authorities (e.g. [176]). As the benefit assessment of an intervention also includes aspects of harm, the generally accepted standards in this respect are also relevant when determining the minimum study duration. Moreover, for long-term interventions as described above, the Institute will resort to the relevant guidelines for the criterion “long-term treatment” [282]. In individual cases, the Institute may deviate from this approach (and will justify this deviation), for example, if a topic requires longer follow-up, or if specific (sub)questions apply to a shorter period. Such deviations may also be indicated if short-term effects are a subject of the assessment (e.g. in the assessment of newly available/approved interventions and/or technologies where no appropriate treatment alternative exists).

3.2.4 Patient-reported outcomes

The patient-relevant dimensions of benefit outlined in Section 3.1.1 can also include patient-reported outcomes (PROs). In addition to health-related quality of life, PROs can also cover other dimensions of benefit, for example, disease symptoms. As in the assessment of quality of life, instruments are required that are suitable for use in clinical trials [174]. In the selection of evidence (especially study types) to be considered for the demonstration of an effect, the same principles as with other outcomes usually apply [198]. This means that also for PROs (including health-related quality of life, symptoms, and treatment satisfaction), RCTs are best suited to demonstrate an effect.

As information on PROs is subjective due to their nature, open studies in this area are of limited validity. The size of the effect observed is an important decision criterion for the question as to whether an indication of a benefit of an intervention with regard to PROs can be inferred from open studies. Empirical evidence shows a high risk of bias for subjective outcomes in open studies [600]. This should be considered in their interpretation (see also Sections 8.1.4 and 8.3.4). However, situations are conceivable where blinding of physicians and patients is not possible. In such situations, if possible, other efforts are required to minimize and assess bias (e.g. blinded documentation and assessment of outcomes). Further aspects on the quality assessment of studies investigating PROs are outlined in [198].

3.2.5 Benefits and harms in small populations

In small populations (e.g. patients with rare diseases or special subgroups of patients with common diseases), there is no convincing argument to deviate in principle from the hierarchy of evidence levels. In this connection, it is problematical that no international standard definition exists as to what is to be understood under a “rare” disease [598]. Independent of this, patients with rare diseases also have the right to the most reliable information possible on treatment options [171]. Non-randomized studies require larger sample sizes than randomized ones because of the need of adjustment for confounding factors. However, due to the rarity of a disease it may sometimes be impossible to include enough patients to provide the study with sufficient statistical power. A meta-analytical summary of smaller studies may be particularly meaningful in such cases. Smaller samples generally result in lower precision in an effect estimate, accompanied by wider confidence intervals. Because of the relevance of the assumed effect of an intervention, its size, the availability of treatment alternatives, and the frequency and severity of potential therapy-related harms, for small sample sizes it may be meaningful to accept a higher p-value than 5% (e.g. 10%) to demonstrate statistical significance, thus increasing quantitative uncertainty. Similar recommendations have been made for other problematical constellations [173]. Such an approach must, however, be specified a priori and well justified. Likewise, for small sample sizes it may be more likely that is necessary to substitute a patient-relevant outcome that occurs too rarely with surrogate endpoints. However, these surrogates must also be valid for small sample sizes [175].

In the case of extremely rare diseases or very specific disease constellations, the demand for (parallel) comparative studies may be inappropriate [598]. Nevertheless, in such cases it is also possible at least to document and assess the course of disease in such patients appropriately, including the expected course without applying the intervention to be assessed (e.g. using historical patient data) [82]. The fact that a situation is being assessed involving an extremely rare disease or a very specific disease constellation is specified and explicitly highlighted in the report plan.

3.3 Benefit assessment of drugs

One main objective of the benefit assessment reports on drugs is to support the G-BA’s decisions on directives concerning the reimbursement of drugs by the SHI. For this purpose, it is necessary to describe whether a drug’s benefit has been demonstrated (or whether, when compared with a drug or non-drug alternative, a higher benefit [added benefit] has been demonstrated).

The G-BA’s decisions on directives do not usually consider particular cases, but the general one. Consequently, the Institute’s reports do not usually refer to decisions on particular cases.

Because of the objective of the Institute’s benefit assessments, these assessments only include studies with an evidence level principally suited to demonstrate a benefit of an intervention. Thus, studies that can only generate hypotheses are generally not relevant for the benefit assessment. The question as to whether a study can demonstrate a benefit mainly depends on the certainty of results of the data analysed.

3.3.1 Relevance of the drug approval status

The commissioning of the Institute by the G-BA to assess the benefit of drugs usually takes place within the framework of the approval status of the drug to be investigated (therapeutic indication, dosage, contra-indications, concomitant treatment, etc.). For this reason, the Institute’s recommendations to the G-BA, which are formulated in the conclusions of the benefit assessment report, usually refer to the use of the assessed drug within the framework of the current approval status.

It is clarified on a project-by-project basis how to deal with studies (and the evidence inferred from them) that were not conducted according to the use of a drug as outlined in the approval documents. In principle, it is conceivable that studies in which a drug was used outside the scope of the approval status described in the Summary of Product Characteristics ("off-label use"), over- or underestimated a drug's benefit and/or harm. This may lead to a misjudgement of the benefit and/or harm in patients treated within the framework of the drug's approval status. However, if it is sufficiently plausible or has even been demonstrated that the results obtained in these studies are applicable to patients treated according to the drug's approval status, these results can be considered in the benefit assessment.

Therefore, for studies excluded from the assessment only because they were off-label studies (or because it was unclear whether they fulfilled the requirements of the approval status), each case is assessed to establish to what extent the study results are applicable to patients treated according to the approval requirements.

Results from off-label studies are regarded as "applicable" if it is sufficiently plausible or has been demonstrated that the effect estimates for patient-relevant outcomes are not greatly affected by the relevant characteristic of the drug approval status (e.g. pretreatment required). As a rule, the equivalence of effects should be proven with appropriate scientific studies. These studies should be targeted towards the demonstration of equivalence of the effect between the group with and without the characteristic. Results applicable to patients treated according to a drug's approval status can be considered in the conclusion of the assessment.

Results from studies are regarded as "not applicable" if their applicability has not been demonstrated and if plausible reasons against the transferability of results exist. As a rule, study results are regarded to be "not applicable" if, for example, the age range or disease severity treated lay outside the approved range or severity, if off-label combinations including other active ingredients were used, or if studies were conducted in patients with contra-indications for the intervention investigated. The results of these studies are not presented in the reports, as they cannot be considered in the assessment

If results from off-label studies are regarded as applicable, this is specified in the report plan. As a rule the results of studies showing the following characteristics are discussed, independently of the applicability of study results to the use specified in the approval of the drug:

- They refer to patients with the disease specified in the commission.
- They refer to patients treated with the drug to be assessed.
- They are of particular relevance due to factors such as sample size, study duration, or outcomes investigated.

3.3.2 Studies on the benefit assessment of drugs

The results of the Institute's benefit assessment of drugs may have an impact on patient health care in Germany. For this reason, high standards are required regarding the certainty of results of studies included in the benefit assessment.

The certainty of results is defined as the certainty with which an effect (or the lack of an effect) can be inferred from a study. This refers to both "positive" aspects (benefit) as well as "negative" aspects (harm). The certainty of results of an individual study is essentially influenced by 3 components:

- the study design
- the internal validity (which is design-specific and determined by the specific way the study was conducted)
- the size of an expected or observed effect

In the benefit assessment of drugs, not only individual studies are assessed, but the results of these studies are incorporated into a systematic review. The certainty of results of a systematic review is in turn based on the certainty of results of the studies included. In addition, it is determined in particular by the following factor:

- the consistency of the results of several studies

The study design has considerable influence on the certainty of results insofar as a causal association between intervention and effect cannot usually be shown with prospective or retrospective observational studies, whereas controlled intervention studies are in principle suited for this purpose [226]. This particularly applies if other factors influencing results are completely or almost completely eliminated. For this reason, an RCT represents the gold standard in the assessment of drug and non-interventions [422].

In the assessment of drugs, RCTs are usually possible and practically feasible. As a rule, the Institute therefore considers RCTs in the benefit assessment of drugs and only uses non-randomized intervention studies or observational studies in justified exceptional cases. Reasons for exception are, on the one hand, the non-feasibility of an RCT (e.g. if the therapist and/or patient have a strong preference for a specific therapy alternative) or, on the other, the fact that other study types may also provide sufficient certainty of results for the research question posed. For diseases that would be fatal within a short period of time without intervention, several consistent case reports may provide sufficient certainty of results that a particular intervention prevents this otherwise inevitable course [358] (dramatic effect, see also Section 3.2.2). The special obligation to justify a non-randomized design when testing drugs can also be found within the framework of drug approval legislation in the directives on the testing of medicinal products (Directive 2001/83/EC, Section 6.2.5 [332]).

In the preparation of the report plan (see also Section 2.1.1), the Institute therefore determines beforehand which study types can be regarded as feasible on the basis of the research question posed, and provide sufficient certainty of results (with high internal validity). Studies not complying with these minimum quality standards (see also Section 8.1.4) are not given primary consideration in the assessment process.

Sections 3.1.4 and 8.1 present information on the assessment of the internal validity of studies, as well as on further factors influencing certainty of results, such as the consistency of the results of several studies and the relevance of the size of the effect to be expected.

In addition to characterizing the certainty of results of the studies considered, it is necessary to describe whether – and if yes, to what extent – the study results are transferable to local settings (e.g. population, health care sector), or what local study characteristics had (or could have had) an effect on the results or their interpretation. From this perspective, studies are especially relevant in which the actual German health care setting is represented as far as possible. However, the criteria for certainty of results outlined above must not be ignored. Finally, the transferability of study results (generalizability or external validity) must be assessed in a separate process initially independent of the study design and quality.

3.3.3 Benefit assessment of drugs according to §35a SGB V

A benefit assessment of a drug according to §35a SGB V is based on a dossier of the pharmaceutical company in which the company provides the following information:

1) approved therapeutic indications
2) medical benefit
3) added medical benefit compared with an appropriate comparator therapy
4) number of patients and patient groups for whom a therapeutically relevant added benefit exists
5) cost of treatment for the SHI
6) requirements for quality-assured usage of the drug

The requirements for form and content of the dossier are outlined in dossier templates, which are a component of the G-BA’s Code of Procedure [211]. In the dossier, specifying the validity of the evidence, the pharmaceutical company must describe the likelihood and the extent of added benefit of the drug to be assessed compared with an appropriate comparator therapy. The information provided must be related both to the number of patients and to the extent of added benefit. The costs for the drug to be assessed and the appropriate comparator therapy must be declared (based on the pharmacy sales price and taking the Summary of Product Characteristics and package information leaflet into account).

The probability of the added benefit describes the certainty of conclusions on the added benefit. In the dossier, the extent of added benefit should be described according to the categories of the Regulation for Early Benefit Assessment of New Pharmaceuticals (ANV[14]) (major, considerable, minor, non-quantifiable added benefit; no added benefit proven; benefit of the drug to be assessed smaller than benefit of the appropriate comparator therapy) [80].

In the benefit assessment the validity and completeness of the information in the dossier are examined. It is also examined whether the comparator therapy selected by the pharmaceutical company can be regarded as appropriate in terms of §35a SGB V and the ANV. In addition, the Institute assesses the effects described in the documents presented, taking the certainty of results into account. In this assessment, the qualitative and quantitative certainty of results within the evidence presented, as well as the size of observed effects and their consistency, are appraised. The benefit and cost assessments are conducted on the basis of the standards of evidence-based medicine described in this methods paper and those of health economic standards, respectively. As a result of the assessment, the Institute presents its own conclusions, which may confirm or deviate from those arrived at by the pharmaceutical company (providing a justification in the event of deviation).

The operationalization for determining the extent of added benefit comprises 3 steps:

1) In the first step the probability of the existence of an effect is examined for each outcome separately (qualitative conclusion). For this purpose, the criteria for inferring conclusions on the evidence base are applied (see Section 3.1.4). Depending on the quality of the evidence, the probability is classified as a hint, an indication or proof.
2) In the second step, for those outcomes where at least a hint of the existence of an effect was determined in the first step, the extent of the effect size is determined for each outcome separately (quantitative conclusion). The following quantitative conclusions are possible: major, considerable, minor, and non-quantifiable.
3) In the third and last step, the overall conclusion on the added benefit according to the 6 specified categories is determined on the basis of all outcomes, taking into account the probability and extent at outcome level within the overall picture. These 6 categories are as follows: major, considerable, minor, and non-quantifiable added benefit; no added benefit proven; the benefit of the drug under assessment is less than the benefit of the appropriate comparator therapy.

The quality of the outcome, as well as the effect size, are essential in determining the extent at outcome level in the second step. The rationale for this operationalization is presented in the Appendix A *Rationale of the methodological approach for determining the extent of added benefit*. The basic approach aims to derive thresholds for confidence intervals for relative

[14] Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV

effect measures depending on the effects to be achieved, which in turn depend on the quality of the outcomes and the extent categories.

It will not always be possible to quantify the extent at outcome level. For instance, if a statistically significant effect on a sufficiently valid surrogate is present, but no reliable estimate of this effect on a patient-relevant outcome is possible, then the (patient-relevant) effect cannot be quantified. In such and similar situations, an effect of a non-quantifiable extent is concluded, with a corresponding explanation.

On the basis of the case of a quantifiable effect, the further approach depends on the scale of the outcome. One distinguishes between the following scales:

- binary (analyses of 2x2 tables)
- time to event (survival time analyses)
- continuous or quasi-continuous, in each case with available responder analyses (analyses of mean values and standard deviations)
- other (e.g. analyses of nominal data)

In the following text, first the approach for binary outcomes is described. The other scales are subsequently based on this approach.

On the basis of the effect measure “relative risk”, denominator and numerator are always chosen in such a way that the effect (if present) is realized as a value < 1, i.e. the lower the value, the stronger the effect.

A) Binary outcomes

To determine the extent of the effect in the case of binary outcomes, the two-sided 95% confidence interval for the relative risk is used; if appropriate, this is calculated by the Institute itself. If several studies are pooled quantitatively, the meta-analytical result for the relative risk is used.

Depending on the quality of the outcome, the confidence interval must lie completely below a certain threshold for the extent to be regarded as minor, considerable or major. It is thus decisive that the upper limit of the confidence interval is smaller than the respective threshold.

The following 3 categories for the quality of the outcome are formed:

- all-cause mortality
- serious (or severe) symptoms (or late complications) and adverse events, as well as health-related quality of life
- non-serious (or non-severe) symptoms (or late complications) and adverse events

The thresholds are specified separately for each category. The more serious the event, the bigger the thresholds (in terms of lying closer to 1). The greater the extent, the smaller the thresholds (in terms of lying further away from 1). For the 3 extent categories (minor, considerable, major), the following Table 4 shows the thresholds to be undercut for each of the 3 categories of quality of the outcomes.

Table 4: Thresholds for determining the extent of an effect

		Outcome category		
		All-cause mortality	Serious (or severe) symptoms (or late complications) and adverse events, as well as health-related quality of life[a]	Non-serious (or non-severe) symptoms (or late complications) and adverse events
Extent category	Major	0.85	0.75 and risk $\geq 5\%$[b]	Not applicable
	Considerable	0.95	0.90	0.80
	Minor	1.00	1.00	0.90

a: Precondition (as for all patient-reported outcomes): use of a validated or established instrument, as well as a validated or established response criterion.

b: Risk must be at least 5% for at least 1 of the 2 groups compared.

The relative risk can generally be calculated in 2 ways, depending on whether the risk refers to events or counter-events (e.g. survival vs. death, response vs. non-response). This is irrelevant for the statement on significance specified in Step 1 of the approach (conventional, non-shifted hypotheses), as in such a case the p-value of a single study is invariant and plays a subordinate role in meta-analysis. However, this does not apply to the distance of the confidence interval limits to the zero effect. To determine the extent of effect for each binary outcome (by means of content criteria under consideration of the type of outcome and underlying disease), it must therefore be decided what type of risk is to be assessed, that of an event or counter-event.

B) Time to event

The two-sided 95% confidence interval for the hazard ratio is required to determine the extent of the effect in the case of outcomes representing a “time to event”. If several studies are pooled quantitatively, the meta-analytical result for the hazard ratio is used. If the confidence interval for the hazard ratio is not available, it is approximated on the basis of the available information, if possible [553]. The same limits as for the relative risk are set for determining the extent (see Table 4).

If a hazard ratio is neither available nor calculable, or if the available hazard ratio cannot be interpreted meaningfully (e.g. due to relevant violation of the proportional hazard assumption), it should be examined whether a relative risk (referring to a meaningful time point) can be calculated. It should also be examined whether this operationalization is adequate in the case of transient outcomes for which the outcome "time to event" was chosen. If appropriate, the calculation of a relative risk at a time point is also indicated here.

C) Continuous or quasi-continuous outcomes, in each case with available responder analyses

Responder analyses are used to determine the extent of added benefit in the case of continuous or quasi-continuous outcomes. For this purpose, a validated or established response criterion or cut-off value is required. On the basis of the responder analyses (2x2 tables) the relative risks are calculated directly from them. According to Table 4 the extent of the effect is then determined.

D) Other outcomes

In the case of other outcomes where no responder analyses with inferable relative risks are available either, it should be examined in the individual case whether relative risks can be approximated [116] to set the corresponding thresholds for determining the extent. Otherwise the extent is to be classified as "non-quantifiable".

For the third step of the operationalization of the overall conclusion on the extent of added benefit, when all outcomes are examined together, a strict formalization is not possible, as no sufficient abstraction is currently known for the value judgements to be made in this regard. In its benefit assessment the Institute will compare the conclusions on probability and on the extent of the effects and provide a justified proposal for an overall conclusion.

3.4 Non-drug therapeutic interventions

Even if the regulatory preconditions for the market access of drugs and non-drug therapeutic interventions differ, there is nevertheless no reason to apply a principally different standard concerning the certainty of results in the assessment of the benefits and harms of an intervention. For example, the G-BA's Code of Procedure [211] envisages, as far as possible, the preferential consideration of RCTs, independent of the type (drug/non-drug) of the medical intervention to be assessed. For medical devices, this is weakened by the conformity evaluation in the current DIN EN ISO Norm 14155 (Section A.6.1 [138]), where RCTs are not presented as the design of choice; however, the choice of design must be justified.

Compared with studies on drug interventions, studies on non-drug interventions are often associated with specific challenges and difficulties [389]. For example, the blinding of the staff performing the intervention will often be impossible, and the blinding of patients will either be difficult or also impossible. In addition, it can be assumed that therapists' and patients' preferences for certain treatment options will make the feasibility of studies in these areas particularly problematic. In addition, it may be necessary especially in the assessment of complex interventions to consider the possibility of contamination effects. It may also be necessary to consider the distinction between the effects caused by the procedure or (medical) device to be assessed on the one hand, and those caused by the expertise and skills of those applying the intervention on the other. Moreover, depending on the time of assessment, learning effects need to be taken into account.

In order to give consideration to the aspects outlined above, studies of particularly good quality are required in order to achieve sufficient certainty of results. Paradoxically, the opposite has rather been the case in the past; i.e. sound randomized studies are often lacking, particularly in the area of non-drug interventions (e.g. in surgery [389]). In order to enable any conclusions at all to be drawn on the relevance of a specific non-drug therapeutic intervention, it may therefore also be necessary to consider non-randomized studies in the assessment. Nonetheless, quality standards also apply in these studies, in particular regarding measures taken to ensure structural equality. However, such studies will usually at best be able to provide hints of a(n) (added) benefit or harm of an intervention due to their inherently lower certainty of results. The inclusion of studies with lower evidence levels is consistent with the corresponding regulation in the G-BA’s Code of Procedure [211]. However, the specific obligation to provide a justification is emphasized. In this regulation it is noted: “However, in order to protect patients, recognition of a method’s medical benefit on the basis of documents with lower evidence levels requires all the more justification the greater the deviation from evidence level 1 (in each case, the medical necessity of the method must also be considered). For this purpose, the method’s potential benefit for patients is in particular to be weighed against the risks associated with the demonstration of effectiveness based on studies of lower evidential value” [211]. This means that the non-availability of studies of the highest evidence level alone cannot generally be viewed as sufficient justification for a benefit assessment based on studies with lower evidence levels.

In the assessment of non-drug therapeutic interventions, it may also be necessary to consider the marketability or CE marking (according to the German Medical Devices Act) and the approval status of drugs (according to the German Pharmaceutical Act), insofar as the test interventions or comparator interventions comprise the use of medical devices or drugs (see Section 3.3.1). The corresponding consequences must subsequently be specified in the report plan (see Section 2.1.1).

3.5 Diagnostic tests

Diagnostic tests are characterized by the fact that their health-related benefit (or harm) is in essence only realized if the tests are followed by therapeutic or preventive procedures. The mere acquisition of diagnostic information (without medical consequences) as a rule has no benefit from the perspective of social law.

This applies in the same way both to diagnostic information referring to the current state of health and to prognostic information (or markers) referring to a future state of health. In the following text, procedures to determine diagnostic or prognostic information are therefore jointly regarded as diagnostic tests.

In general, the evaluation process for diagnostic tests can be categorized into different hierarchy phases or levels, analogously to the evaluation of drugs [204,329]. Phase 4 prospective, controlled diagnostic studies according to Köbberling et al. [329], or Level 5 studies according to Fryback and Thornbury [204] have an (ideally random) allocation of patients to a strategy with or without application of the diagnostic test to be assessed or to a group with or without disclosure of the (diagnostic) test results. These studies can be seen as corresponding to Phase 3 (drug) approval trials ("efficacy trials"). Accordingly, they are allocated to the highest evidence level (see, for example, the G-BA's Code of Procedure [211]). The US Food and Drug Administration also recommends such studies for specific indications in the approval of drugs and biological products developed in connection with diagnostic imaging techniques [197]. Examples show that they can be conducted with comparatively moderate effort [16,568].

The Institute follows this logic and primarily conducts benefit assessments of diagnostic tests on the basis of studies designed as described above that investigate patient-relevant outcomes. The main features of the assessment comply with the explanations presented in Sections 3.1 and 3.4. In this context, patient-relevant outcomes refer to the same benefit categories as in the assessment of therapeutic interventions, namely mortality, morbidity, and health-related quality of life. The impact of diagnostic tests on these outcomes can be achieved by the avoidance of high(er) risk interventions or by the (more) targeted use of interventions. If the collection of diagnostic or prognostic information itself is associated with a high(er) risk, a lower-risk diagnostic test may have patient-relevant advantages, namely, if (in the case of comparable test quality) the conduct of the test itself causes lower mortality and morbidity rates, or fewer restrictions in quality of life.

Conclusions on the benefit of diagnostic tests are ideally based on randomized studies, which can be conducted in various ways [50,51,188,360,378,484]. In a study with a strategy design including 2 (or more) patient groups, in each case different strategies are applied, which in each case consist of a diagnostic measure and a therapeutic consequence. A high informative value is also ascribed to randomized studies in which all patients initially undergo the conventional and the diagnostic test under investigation; subsequently, only those patients are randomized in whom the latter test produced a different result, and thereby a different therapeutic consequence, than the former test (discordance design). Studies in which the interaction between the diagnostic or prognostic information and the therapeutic benefit is investigated also have a high evidence level and should as a matter of priority be used for the benefit assessment of diagnostic tests (interaction design [484,541]). Many diagnostic or prognostic characteristics – especially genetic markers – can also be determined retrospectively in prospective comparative studies and examined with regard to a potential interaction (so-called "prospective-retrospective" design [516]). The validity of such "prospective-retrospective" designs depends especially on whether the analyses were planned prospectively (in particular also the specification of threshold values). Moreover, in all studies with an interaction design it is important that the treatments used correspond to the current standard, that the information (e.g. tissue samples) on the characteristic of interest is completely available for all study participants or at least for a representative sample, and that if several characteristics are analysed the problem of multiple testing for significance is adequately accounted for (see also Section 8.3.2 [485]).

Overall, it is less decisive to what extent diagnostic or prognostic information can determine a current or future state of health, but rather that this information is of predictive relevance, namely, that it can predict the greater (or lesser) benefit of the subsequent treatment [188,517]. For this – necessarily linked – assessment of the diagnostic and therapeutic intervention it is important to note that overall, a benefit can normally only arise if both interventions fulfil their goal: If either the predictive discriminative capacity of the diagnostic intervention is insufficient or the therapeutic intervention is ineffective, a study will not be able to show a benefit of the diagnostic intervention.

Besides a strategy and interaction design, a third main form of RCTs on diagnostic questions is available with the enrichment design [379,541]. In this design, solely on the basis of the diagnostic test under investigation, only part of the patient population is randomized (and thus included); for example, only test-positive patients, who then receive 1 of 2 treatment options. In comparison with an interaction design, such a design lacks the investigation of a potential treatment effect in the remaining patients (e.g. in the test-negative ones). Robust conclusions can thus only be drawn from such designs if, on the basis of other information, it can be excluded that an effect observed in the randomized patient group could also have existed in the non-randomized group.

The comments above primarily refer to diagnostic tests that direct more patients towards a certain therapeutic consequence by increasing the test quality (i.e. sensitivity, specificity or both). In these cases, as a rule it is necessary to examine the impact of the diagnostic test on patient-relevant outcomes by covering the whole diagnostic and therapeutic chain. However, it is possible that the diagnostic test under investigation is only to replace a different and already established diagnostic test, without identifying or excluding additional patients. If the new test shows direct patient-relevant advantages, for example, is less invasive or requires no radiation, it will not always be necessary to re-examine the whole diagnostic-therapeutic chain, as the therapeutic consequences arising from the new test do not differ from those of the previous test [42,51,394]. To demonstrate benefit, in these cases test quality studies could be sufficient in which it is shown that the test result of the previous test (= reference standard) and that of the test under investigation (= index test) is identical in a sufficiently high proportion of patients (one-sided question of equivalence).

On the other hand, for a comparison of 2 or more diagnostic tests with regard to certain test quality characteristics, studies with the highest certainty of results, and thus primarily considered in the Institute’s reports, are studies with a random allocation of the sequence of the conduct of the tests (which are independent of each other and preferably blinded) in the same patients or with a random allocation of the tests to different patients.

If a study is to provide informative data on the benefit, diagnostic quality or prognostic value of a diagnostic test, it is essential to compare it with the previous diagnostic approach [542]. Only in this way can the added value of the diagnostic or prognostic information be reliably determined. For studies on test accuracy this means that, besides sensitivity and specificity of the new and previous method, it is of particular interest to what extent the diagnostic measures produce different results per patient. In contrast, in studies on prognostic markers multifactorial regression models often play a key role, so that Section 8.3.7 should be taken into account. When selecting non-randomized designs for diagnostic methods, the ranking of different study designs presented in Section 8.1.3 should as a rule be used.

In the assessment of the certainty of results of studies on diagnostic accuracy, the Institute primarily follows the QUADAS-2[15] criteria [592,593], which, however, may be adapted for the specific project. The STARD[16] criteria [52,53] are applied in order to decide on the inclusion or exclusion of studies not published in full text on a case-by-case basis (see also Sections 8.1.4 and 8.3.11). Despite some individual good proposals, there are no generally accepted quality criteria for the methodological assessment of prognosis studies [11,251,252,515]. Only general publication standards exist for studies on prognostic markers [579], however, there are publication standards for prognostic markers in oncology [14,393].

Level 3 and 4 studies according to Fryback and Thornbury [204] are to investigate the effect of the (diagnostic) test to be assessed on considerations regarding (differential) diagnosis and/or subsequent therapeutic (or other management) decisions, i.e. it is investigated whether the result of a diagnostic test actually leads to any changes in decisions. However, such studies or study concepts have the major disadvantage that they are not sharply defined, and are therefore of rather theoretical nature. A principal (quality) characteristic of these studies is that it was clearly planned to question the physicians involved regarding the probability of the existence of the disease (and their further diagnostic and/or therapeutic approach) *before* the conduct of the diagnostic test to be assessed or the disclosure of results. This is done in order to determine the change in attitude caused by the test result. In contrast, retrospective appraisals and theoretical estimates are susceptible to bias [204,239]. The relevance of such ultimately uncontrolled studies within the framework of the benefit assessment of diagnostic (or prognostic) tests must be regarded as largely unclear. Information on management changes alone cannot therefore be drawn upon to provide evidence of a benefit, as long as no information on the patient-relevant consequences of such changes is available.

It is also conceivable that a new diagnostic test is incorporated in an already existing diagnostic strategy; for example, if a new test precedes (triage test) or follows (add-on test) an established test in order to reduce the frequency of application of the established test or new test, respectively [50]. However, against the background of the subsequent therapeutic (or other types of) consequences, it should be considered that through such a combination of

[15] Quality Assessment of Diagnostic Accuracy Studies
[16] Standards for Reporting of Diagnostic Accuracy

tests, the patient populations ensuing from the respective combined test results differ from those ensuing from the individual test results. This difference could in turn influence subsequent therapeutic (or other types of) consequences and their effectiveness. If such an influence cannot be excluded with sufficient certainty – as already described above – comparative studies on diagnostic strategies including and excluding the new test may be required [197,367].

Several individual diagnostic tests or pieces of information are in part summarized into an overall test via algorithms, scores, or similar approaches. In the assessment of such combined tests the same principles should be applied as those applied for individual tests. In particular, the validation and clinical evaluation of each new test must be performed independently of the test development (e.g. specification of a threshold, weighting of scores, or algorithm of the analysis) [531].

Biomarkers used within the framework of “personalized” or better “stratified” medicine should also be evaluated with the methods described here [268,541]. This applies both to biomarkers determined before the decision on the start of a treatment (or of a treatment alternative) and to those determined during treatment in order to decide on the continuation, discontinuation, switching, or adaptation of treatment [520,567]. Here too, it is essential to distinguish between the prognostic and predictive value of a characteristic.

Prognostic markers provide information on the future state of health and normally refer to the course of disease under treatment and not to the natural course of disease without treatment. The fact that a biomarker has prognostic relevance does not mean that it also has predictive relevance (and vice versa).

Finally, in the assessment of diagnostic tests, it may also be necessary to consider the result of the conformity assessment procedure for CE marking and the approval status of drugs used in diagnostics (see Section 3.3.1). The corresponding consequences must subsequently be specified in the report plan (see Section 2.1.1).

3.6 Early diagnosis and screening

Screening programmes are composed of different modules, which can be examined either in part or as a whole [120,513]. The assessment of a screening test generally follows internationally accepted standards and criteria, for example, those of the UK National Screening Committee (UK NSC [564]), the US Preventive Services Task Force (US PSTF [247,437,490]), or the New Zealand National Health Committee (NHC) [406].

According to the criteria outlined above, the Institute primarily assesses the benefit of screening tests by means of prospective comparative intervention studies on the whole screening chain, which include the (ideally random) allocation of participants to a strategy with or without application of the screening test (or to different screening strategies) and which investigate patient-relevant outcomes. In this context, the main features of the assessment comply with the explanations outlined in Sections 3.1 to 3.4.

If such studies are not available or are of insufficient quantity or quality, an assessment of the single components of the screening chain can be performed. In this context, the accuracy of the diagnostic test is assessed by means of generally applied test quality criteria, determined in studies showing sufficient certainty of results (usually Phase 3 according to Köbberling et al. [329]), (see Section 3.5), and it is reviewed to what extent it is proven that the consequences resulting from the test outcomes are associated with a benefit. In the case of therapeutic consequences (which are mostly assumed), proof can be inferred from randomized intervention studies in which an early (earlier) intervention was compared with a late(r) one. The benefit of an early (earlier) vs. a late(r) intervention may also be assessed by means of intervention studies in which the interaction between the earliness of the start of the intervention and the intervention's effect can be investigated. This can be performed either directly within a study or indirectly by comparing studies with different starting points for the intervention, but with otherwise comparable study designs. Here too, the main features of the assessment comply with the explanations outlined in Sections 3.1 to 3.4.

3.7 Prevention

Prevention is directed at avoiding, reducing the probability of, or delaying health impairment [581]. Whereas primary prevention comprises all measures employed before the occurrence of detectable biological impairment in order to avoid the triggering of contributory causes, secondary prevention comprises measures to detect clinically asymptomatic early stages of diseases, as well as their successful early therapy (see also Section 3.6). Primary and secondary prevention measures are characterized by the fact that, in contrast to curative measures, whole population groups are often the focus of the intervention. Tertiary prevention in the narrowest sense describes specific interventions to avoid permanent (especially social) functional deficits occurring after the onset of disease [254]. This is not the focus of this section, but is addressed in the sections on the benefit assessment of drug and non-drug interventions (see Sections 3.3 and 3.4).

The Institute also primarily performs benefit assessments of prevention programmes (other than screening programmes) by means of prospective, comparative intervention studies that have an (ideally random) allocation of participants to a strategy with or without application of the prevention measure, and that investigate patient-relevant outcomes. Alternatively, due to potential "contamination" between the intervention and control group, studies in which clusters were allocated to the study arms may also be eligible [554].

In individual cases, it needs to be assessed to what extent the consideration of other study designs is meaningful [308]. For example, mass-media campaigns are often evaluated within the framework of "interrupted time-series analyses" (e.g. in [572]), and the use of this study design is also advocated for community intervention research [43]. In the quality assessment of these studies, the Institute uses for orientation the criteria developed by the Cochrane Effective Practice and Organisation of Care Review Group [106].

For the benefit on the population level, not only the effectiveness of the programme is decisive, but also the participation rate. In addition, the question is relevant as to which persons are reached by prevention programmes; research indicates that population groups with an increased risk of disease participate less often in such programmes [343]. Special focus is therefore placed on both of these aspects in the Institute's assessments.

3.8 Assessment of potential

In contrast to benefit assessments, assessments of potential aim to investigate whether new examination or treatment methods potentially show a benefit. In this context, "potential" means that firstly, the evidence available so far indicates that a potential benefit may exist, and secondly, that on the basis of this evidence a study can be planned that allows an assessment of the benefit of the method on a sufficiently reliable evidence level; see §14 (3, 4) of the G-BA's Code of Procedure [211].

An assessment of potential according to §137e (7) SGB V is based on an application for which the G-BA has defined the form and required content. Those entitled to apply are manufacturers of a medical device on which the technical application of a new examination or treatment method is largely based, as well as companies that in another way as a provider of a new method have an economic interest in providing their service at the expense of the health insurance funds. The application must contain informative documents especially referring to the current evidence on and the expected benefit of the new examination and treatment method (see §20 (2) No. 5 of the G-BA's Code of Procedure [211]). Optionally a proposal can be submitted on the key points of a testing study. An application for a method can refer to one or several therapeutic indications.

Within the framework of the assessment of potential the Institute evaluates the plausibility of the information provided by the applicant. This evaluation especially refers to the meaningfulness of the medical question(s) presented in the application, the quality of the literature searches conducted by the applicant (see Section 7.2), the assessment of the certainty of results of the relevant studies, and the correctness of the results presented in the application. The assessment leads to a conclusion on the potential of the examination or treatment method applied for. If a potential is determined from the Institute's point of view, the testing study proposed by the applicant is evaluated; if the application does not contain such a proposal or an unsuitable one, the Institute specifies the key points of a possible testing study.

Due to the particular aim, considerably lower requirements for the evidence are imposed in assessments of potential compared with benefit assessments. Ultimately, the aim of testing is first to generate an adequate data basis for a future benefit assessment. Accordingly, a potential can be justified, in particular also on the basis of non-randomized studies. Moreover, further methodological principles of benefit assessments are not used or only used to a limited extent in assessments of potential, as described in the following text.

In contrast to benefit assessments, due to lower requirements for the evidence, in assessments of potential an extended assessment of the qualitative certainty of results of non-randomized studies is performed. In this context, besides the levels mentioned in Section 3.1.4 for randomized studies (high or moderate certainty of results) the following grades are used:

- **low qualitative certainty of results:** result of a higher quality non-randomized comparative study with adequate control for confounders (e.g. quasi-randomized controlled studies, non-randomized controlled studies with active allocation of the intervention following a preplanned rule, prospective comparative cohort studies with passive allocation of the intervention),
- **very low qualitative certainty of results:** result of a higher quality non-randomized comparative study (see point above), but without adequate control for confounders or result of another non-randomized comparative study (e.g. retrospective comparative cohort studies, historically controlled studies, case-control studies),
- **minimum qualitative certainty of results:** result of a non-comparative study (e.g. one-arm cohort studies, observational studies or case series, cross-sectional studies or other non-comparative studies).

An important aspect of the certainty of results is the control for confounders, which can in particular be achieved through the use of multifactorial statistical methods – as described in Section 8.3.7. Further factors are also taken into account in the assessment of the certainty of results (see Section 8.1.4).

High-quality non-randomized studies may also show a considerable risk of bias. When deriving the potential of an intervention from such studies, it must therefore be evaluated whether the available studies show differences regarding the intervention of interest in a magnitude suggesting that a benefit can be demonstrated in suitable future studies, and that these differences cannot be solely explained by the average expected influence of bias. A potential thus particularly arises if studies of a low certainty of results show at least small effects, if studies of a very low certainty of results show at least medium effects, and if studies with a minimum certainty of results show at least large effects. For the relative risk, values of 0.8 and 0.5 can serve as rough thresholds between small, medium and large effects [150,434]. Deviating from the procedure in benefit assessments (see Section 3.1.2), in assessments of potential, surrogate endpoints are also considered for which no sufficient validity has yet been shown. However, surrogate endpoints should be established and plausible so as to be able to justify a potential.

If the potential of diagnostic methods is to be evaluated, data on test accuracy are also considered. In this context, the certainty of results of the underlying studies must be examined (see Sections 3.5 and 8.3.11). In a second step, an evaluation of the plausibility of the diagnostic method is performed with regard to the effects postulated by the applicant in respect of patient-relevant outcomes, that is, possible direct effects of the method, as well as therapeutic consequences via which the diagnostic method could influence patient-relevant outcomes.

4 Health economic evaluation of medical interventions

4.1 Introduction

According to SGB V, in relation to the specific commission the Institute determines the methods and criteria for the preparation of health economic evaluations (HEEs) on the basis of the international standards of evidence-based medicine and health economics recognized by the respective experts in these fields. For each HEE, decisions must be made on the perspective, the time horizon, the choice of comparators, the underlying care pathway, the model, the data basis, and the presentation of uncertainty. These basic criteria for an HEE are briefly explained against the background of commissioning by the Federal Joint Committee (G-BA). All deviations from the methods presented here must be justified in the individual case.

4.1.1 Legal basis for a health economic evaluation according to SGB V

According to §139a (3) No. 2 SGB V, the Institute can be commissioned with regard to questions concerning the quality and efficiency of services provided within the framework of statutory health insurance (SHI). HEEs of drugs can also be commissioned by the G-BA according to §35b SGB V. Furthermore, an HEE can be commissioned by the Federal Ministry of Health according to §139b (2) SGB V [79].

In the following text, at first methodological aspects generally applying to HEEs are addressed. In Section 4.9 the deviations are then explained that arise from HEEs of drugs performed according to §35b SGB V.

4.1.2 Perspective

Depending on the commission, the following perspectives can be considered: the (pure) SHI perspective, the perspective of the community of SHI insurants (short: “SHI insurant perspective”), the social insurance perspective or the perspective of individual social insurance branches, as well as the societal perspective. In contrast to the pure SHI perspective, in the SHI insurant perspective the costs borne by the insurants (e.g. from co-payments) are also considered (see Section 4.4.1). Depending on the commission, it can be required for an HEE to consider the perspective of individual social insurance branches in addition to the SHI insurant perspective. The decision on whether further perspectives are included in an HEE depends solely on the question as to whether this is relevant for the decision maker. The results of the assessment from an extended perspective are presented separately to the decision maker.

4.1.3 Time horizon

The time horizon must at least represent the average study duration and thus consider the differences in costs and benefits between the interventions of an HEE that are relevant for the reimbursement decision. A longer time horizon should preferably be chosen for chronic diseases, particularly if survival gains are expected [68,159,377,555]. Costs and benefits should always be modelled over the same time horizon.

The appropriate time horizon is often longer than the period covered by the available primary data from prospective studies. In these cases, under consideration of the duration of the studies, a time horizon appropriate for the disease should be chosen [259,555].

4.1.4 Choice of comparators

For the derivation of an efficiency frontier, the presentation form chosen by the Institute for results of an HEE (see Section 4.6), all healthcare-relevant interventions in a therapeutic area should be considered in an HEE. Active ingredients can, for example, be pooled into drug classes, if this seems meaningful from a medical point of view and if homogeneity is sufficient (see Section 8.3.8).

4.1.5 Care pathway

For each HEE, at first one or more care pathways should be developed for the therapeutic area. A care pathway describes treatment processes for patients with one or more specific therapeutic indications in a chronological sequence and structures them according to sectors, professions involved, stages, and, if applicable, further aspects. This care pathway serves as a basis for developing the decision-analytic model (see Sections 4.1.6 and 4.2). Furthermore, the literature searches for data on costs and further data required for the model are also based on the care pathway.

At first, for each specific commission the course of disease and the provision of health care in Germany should be briefly described for the relevant therapeutic indication, together with the sources used. The relevant interventions and treatment steps in the different areas of service, including the service providers, must be rendered within the limits of the approval status and the efficiency principle. Moreover, their application must be evaluated within the specifications of the directives and treatment advice that apply in the SHI system. Furthermore, the current treatment recommendations for Germany should be presented, using valid (clinical practice) guidelines. The relevant components for the HEE should be distinguished from the healthcare context described, so that a care pathway relevant to the model can be described. If individual components are specifically not included in the care pathway, this decision should be justified.

A piggy-back study is a clinical study in which, in addition to the determination of the benefits and harms of a health technology, the costs are also simultaneously determined. Even if such a study is available, concomitantly a care pathway should also be depicted, so that the costs and further data collected in the piggy-back study can be comprehended by means of the attached care pathway.

4.1.6 Model

Piggy-back studies are very rarely available. Moreover, economic data are mostly not collected in clinical studies. The data are often insufficient to comprehensively analyse the costs of an intervention. This is because on the one hand, clinical studies rarely provide information on the long-term economic consequences accompanying the introduction of a new intervention. On the other, they do not always adequately and completely address the healthcare aspects relevant for the cost side in Germany. Moreover, a protocol-induced use of resources within the context of clinical studies can also induce misjudgements on the cost side. For these reasons, the modelling of the costs of an intervention is an important component of an HEE (see Section 4.4). Likewise, in an HEE the benefit can be modelled if a longer time horizon than the one used in the underlying studies is supposed to be used in the HEE (see Section 4.3).

4.1.7 Data basis

The basis and the assessment of data considered in an HEE on the benefit side are explained in Section 3.3. Individual studies or data from surveys are used for the measure of overall benefit (see also Section 4.3.3).

Data considered in the HEE to illustrate the provision of health care, epidemiology, and costs, can be collected in various ways and originate from various sources. These include secondary data, guidelines, expert surveys as well as price catalogues or lists (see also Sections 4.4.4 and 4.5.2).

Analyses of secondary data should follow guidelines and recommendations on the good practice of secondary data analysis [19]. In particular, the choice of data basis, the size and relevant characteristics of the sample and the study population (incl. inclusion and exclusion criteria), the statistical methods, and the control of confounding factors, should be described transparently and justified. The generalizability and representativeness of results should be explained. The individual analysis steps must be comprehensible; plausibility checks should be ensured. If guidelines are used, they should originate from the German healthcare system and preferably be evidence-based. This refers to guidelines whose recommendations are based on a systematic literature search, are provided as a matter of principle with a level of evidence (LoE) and/or grade of recommendation (GoR), and are linked to the citations of the underlying primary and/or secondary literature (modified according to AGREE[17]) [4].

Expert surveys follow the generally recognized methods and procedures of quantitative social research. This means that in expert surveys, explicit information should be provided on the recruitment, number and expertise of experts, the research question, individual answers (not only mean values), the manner of achieving a consensus, as well as the presentation and

[17] Appraisal of Guidelines Research and Evaluation in Europe

handling of results. Price catalogues or lists must be current and represent the prices relevant for the SHI.

4.1.8 Uncertainty

Following common international practice, one distinguishes between the following types of uncertainty in health economic decision analysis [63]:

Table 5: Concepts of uncertainty in health economic decision analysis[a]

Term	Concept	Other terms sometimes employed	Analogous concept in regression
Stochastic uncertainty	Random variability in outcomes between identical patients	Variability, first-order uncertainty	Error term
Parameter uncertainty	The uncertainty in estimation of the parameter of interest	Second-order uncertainty	Standard error of the estimate
Structural uncertainty	The assumptions inherent in the decision model	Model uncertainty	The form of the regression model (e.g., linear, log-linear)
a: Table content = extract from Briggs et al. [63].			

Due to its complexity, the investigation of uncertainty must be considered in all areas of an HEE. For this purpose, the Institute follows the classification of uncertainty (see Table 5).

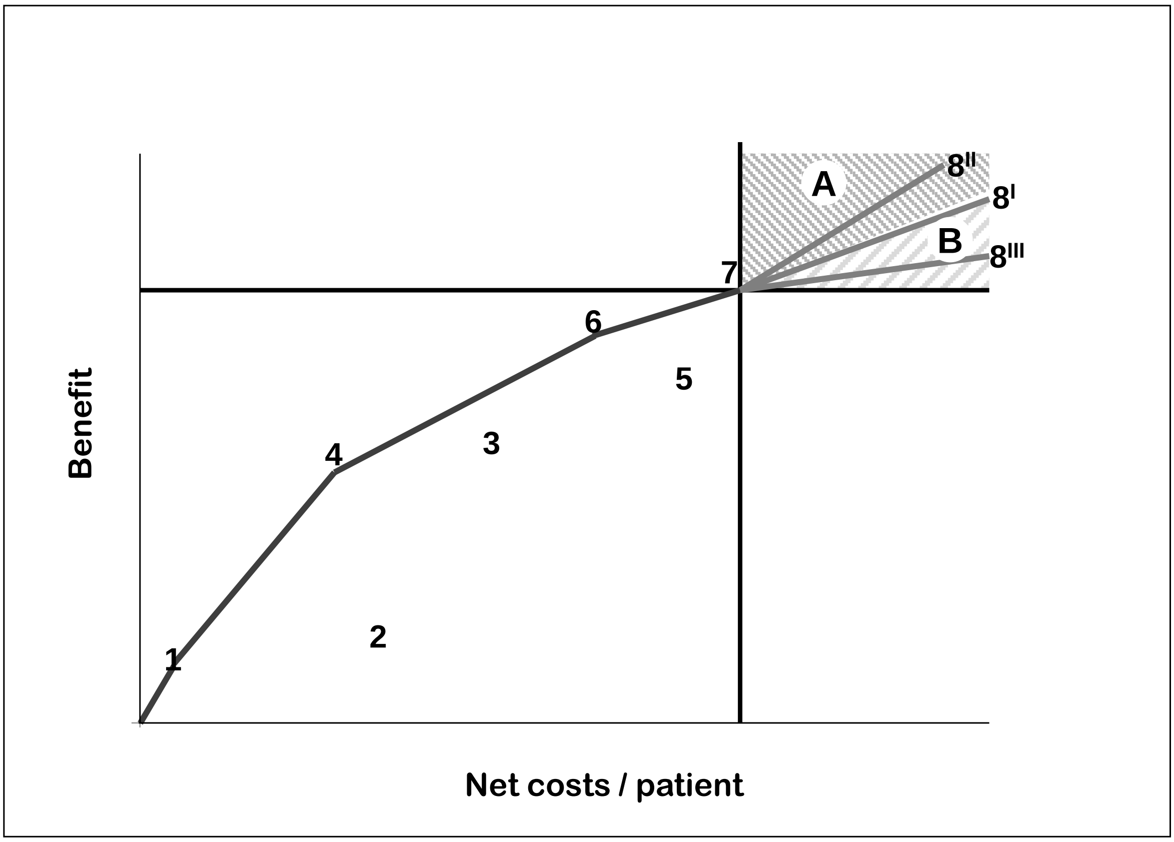
To this end, basic comments on uncertainty and the distribution assumptions are already made in Sections 4.2, 4.3, 4.4, and 4.5. The conduct and presentation of the investigation of uncertainty are then presented in Section 4.7.

4.1.9 Interpretation of results

The results are presented in tables and graphs in the form of an efficiency frontier.

The efficiency frontier is displayed in Figure 9. Interventions 1 to 7 are plotted as comparators with their cost-effectiveness ratios. Interventions 1, 4, 6, and 7 span an efficiency frontier. The last segment of the efficiency frontier can be linearly extrapolated on the assumption that it represents the reciprocal of the current willingness-to-pay. Then the following applies: Interventions that, related to an outcome, lie on Area A (see Intervention 8^{II}) have, according to their cost-effectiveness ratio, a better cost-effectiveness ratio versus the extrapolated last segment of the efficiency frontier and can thus be reimbursed at the price specified. Interventions on Area B (see Intervention 8^{III}) have, according to their cost-effectiveness ratio, a less favourable cost-effectiveness ratio versus the extrapolated last segment of the efficiency frontier, so that their price based on the efficiency frontier can be regarded as inappropriate, since the existing efficiency would deteriorate. Under consideration of the criterion of the appropriateness of the costs of interventions on Area B, the decision maker can negotiate a reimbursement price. Interventions with a constant benefit-cost relation (see Intervention 8^I) also fulfil the criterion that their price would be appropriate in comparison with the extrapolated last segment of the efficiency frontier.

The efficiency frontier illustrates which interventions show the comparatively greatest benefit in relation to the costs incurred. Inefficient interventions are, for example, both more expensive and in relation to an outcome of lesser benefit than other interventions. If both costs and the benefit generated by the new intervention are higher than those already depicted in the efficiency frontier, the costs appropriate for this intervention are not directly inferable from the efficiency frontier itself. Further criteria must thus be drawn upon to assess whether the use of a new treatment that produces an added benefit, but is also more cost-intensive, is appropriate. The Institute assumes that the deterioration of efficiency in a therapeutic indication through inclusion of new interventions is inappropriate. This efficiency is implemented through the linear extrapolation of the gradient of the last segment of the efficiency frontier. Hence, in the event of a given benefit of an intervention under assessment, those cost-effectiveness ratios are regarded to be appropriate that, as measured by the efficiency frontier, do not lead to a deterioration of efficiency in a given therapeutic area (see Figure 9).



Net costs = costs arising from health care, adjusted by cost savings, so-called cost offsets.

Figure 9: Presentation of the areas relevant to decisions

If a measure of the overall benefit is specified (see Section 4.3.3), this is to be regarded as the primary result. If the determination of several efficiency frontiers is required for the assessment of an intervention, the decision maker is entitled to conduct a weighting of results, under observation of the relevance of patient-relevant outcomes. A similar approach can also be considered for the approval of an intervention in several therapeutic areas.

The reasonableness of cost coverage by the community of SHI insurants depends on the one hand on the appropriateness of an intervention’s price, but on the other, also on the associated future overall costs depending on the financial capacity and willingness-to-pay of the community of SHI insurants. As neither the financial capacity nor the willingness-to-pay of the community of SHI insurants is assessed, no specific recommendation is issued on the reasonableness of cost coverage. To depict the future financial impact of a cost coverage, a budget impact analysis should be conducted that can serve as an information basis for the decision maker concerning the decision on reasonableness.

4.2 Modelling

4.2.1 Basic principles

In a health economic decision model ("model") as the key component of an HEE, data on benefits and costs are merged from different sources in order to calculate the cost-effectiveness ratios of interventions during the course of the disease. Merging of data from different sources by means of a model is often required for numerous reasons. In most cases not all variables relevant for the decision are recorded in a study. A health economic model is also explicitly used to extrapolate benefits and costs beyond the period covered by a study.

Health economic models, like mathematically-formalized models, are thus a simplified depiction of reality. Moreover, analytic clarity is achieved by intentionally reducing complexity to the decision factors and variables relevant for the decision problem.

Thorough documentation is of key importance for health economic models. This documentation should generally consist of 2 parts: on the one hand, a general descriptive documentation of the approach, in which the decisions made and the data (sources) chosen are presented and justified, and on the other, a technical documentation in which the functional/mathematical relations of the model components are presented, so that an expert third party can replicate the results of the model independently of a specific software.

The degree of complexity or the degree of the reduction of a model always depends on the research question posed and cannot be specified a priori. For this reason, besides the internal validity of a model, the applicability should be described and proven. The model structure (e.g. health states), which must be covered by the approval status of the intervention and the framework of the provision of services according to SGB V, is as a rule developed in agreement with external clinical experts to ensure external validity.

4.2.2 Basic aspects of model development

The results of the models must provide a detailed depiction of the benefits and costs incurred in Germany for the intervention under assessment. For this purpose, the following information must be included in the model:

- results on effects (benefits and harms) of the interventions
- complete recording of disease costs and
- all aspects of the disease and treatment that may have a relevant impact on the benefit or cost components of the model, e.g. in the areas of demographics, epidemiology, and care pathway(s)

As data on individual aspects are often lacking, it is particularly important to explore the impact of assumptions and of the model input on the results by means of sensitivity analyses.

The following conditions must be fulfilled to ensure the validity and formal/content-related comprehensibility of a model:

- complete transparency with clearly described and justified model input and assumptions
- sufficient depth to adequately depict the disease modelled as well as the associated costs and the respective care pathways
- sufficient flexibility to calculate multiple scenarios for varying assumptions and settings
- option of determining uncertainty in the predicted cost and benefit components

4.2.3 Influence diagram and model concept

On the basis of deliberations and information leading to the creation of the care pathway, the basic principles of the model are presented in an influence diagram and a model concept.

An influence diagram graphically depicts the essential relationships of the model, i.e. course of the disease, patient characteristics, pathophysiological processes, and treatment. It displays the factors that have or might have an influence on the research question(s) to be modelled. Despite its name, the influence diagram does not per se show causal associations.

The model concept is based on the influence diagram and presents the intended design in much greater depth. As even the most sophisticated models are simplifications of reality, with required assumptions and limitations referring to the content included [3,159], the model can only be properly understood if the model concept is specified and documented in a comprehensible manner.

4.2.4 Choice of modelling technique

The choice of the appropriate modelling technique depends on the research question posed, the characteristics of the intervention under assessment, the respective disease, and the general framework. When choosing the modelling technique the guiding principle for the Institute is that the economic model should be as sophisticated and complex as required to adequately answer the research question(s) posed. The evidence base itself should not determine the choice of modelling technique. If the choice of a modelling technique requires a modification of the model concept, this choice should be re-evaluated [87].

The modelling technique chosen must also be compared with those techniques already conducted/published for the same or closely related decision problems. If the model applied deviates from the models already existing, this should be discussed and justified. However, as the appropriate modelling technique always depends on the underlying research question, fixed requirements specified a priori are not meaningful, the more so as the international standard of health economics is being continuously further developed [300]. As a matter of principle, the following key problem areas should be considered:

- Temporal dimension: For which time horizon are conclusions drawn and extrapolations performed, and how is time structured within the model (e.g. continuously/discretely, length of cycle)?
- Analysis unit: Which analysis or observation unit is depicted (e.g. individuals, cohorts) and which characteristics are considered (age, sex, etc.)?
- Interactions: Which interactions are depicted between the analysis units themselves, i.e. patients or other elements of the model?

As data from different sources are often merged for modelling, it might be necessary to transform these data into the same format, e.g. relating to the same period of time.

4.2.5 Model documentation and model validation

A) Model validation and structural uncertainty

A simulation model that is valid for one research question might not be valid for another [350]. The external validation process must therefore cover each intended use of the model and, if used for other research questions, the model must be validated again. There is disagreement on the appropriate approach for model validation; however, there are some basic steps that must be followed [486].

A key element of validation addresses the question as to whether the model adequately depicts the reality of the course of the disease and of treatment. The plausibility check (face validity) refers to the influence diagram, the model concept, data acquisition, the processing of functional relationships, and the choice of modelling technique.

A further key element of validation is the correct technical implementation of the model (internal or technical validation). This aspect refers to the question as to whether the technical implementation actually implements the model concept correctly, for example, whether the results are numerically correct and robust.

A third element of validation is the predictive validity. To what extent does the model predict the future, that is, are the predicted results reflected in the "real world"? This is certainly the most desirable form of validity, but the most difficult to prove, if at all possible [571]. However, a comparison of the model's results with previous, comparable studies is meaningful and differences should be explicable. This also applies to comparisons with other health economic models (cross validity).

A specific form of uncertainty in model development or validation is the so-called structural uncertainty, with regard to which it is scrutinized to what extent the functional relationships underlying the model are actually valid and whether other functional forms would not be more appropriate. If it becomes obvious in the planning and development of a model that the structural uncertainty is relevant for the underlying research question, it may be necessary to develop several (alternative) models in order to quantify the consequences of this form of uncertainty on the result [540].

B) General documentation

The Institute prepares a detailed technical report describing all the modelling steps from the development of the influence diagram to the final validation. In addition, a fully executable version of the model must be made available, along with a user manual. In line with other suggested guidelines [107,410,584], the documentation of the model should include the following:

- The influence diagram used to guide model development.
- Details of the model concept
  - description of the target population(s) considered in the evaluation, including subgroups
  - description of the interventions evaluated
  - choice of the model settings (simulation size, time horizon, discounting rates, etc.) and justification
  - overview of HEEs in the therapeutic area investigated.
- Description of all data sources. Justification for choice of data sources must be provided.
- Details of all functional relationships used in the model. If they were custom-developed for the model, detailed information on the methods used must be provided.
- Listing of all assumptions with regard to data sources and model structure. Especially important is a detailed account of any assumption and technique used to project beyond the period to which the data apply.
- Rationale for the modelling technique adopted
  - description of how the technique conforms to the required features.
- Overview of the validation techniques used and their results.
- Detailed presentation of results, including an assessment of the impact of the
  - use of the intervention in relevant subgroups
  - uncertainty in input data (see Section 4.7 on sensitivity analyses).
- Interpretation of the results, including a description of the limitations of the approach used.

C) Technical documentation and electronic version of the model

The technical documentation is crucial for the understanding and the assessment of the underlying health economic model. All variables used should be named and defined. The functional/mathematical relationships of the model components should be presented and, if applicable, justified. The formal-mathematical relationships should connect all input variables considered in the model (e.g. health states) with the respective operators (e.g. age-specific transition probabilities). In addition, the derivation of interim and final values must still be presented.

All calculation steps within the software should be documented in a comprehensible manner. This is generally performed by documentation of the program code with which the electronic version of the model is implemented. In table calculation programs (e.g. Excel), the sequence of the calculation steps cannot be directly obtained from the electronic version. If applicable, these steps must then be documented in writing in a way that the sequence of the calculation steps is evident.

An electronic version of the model must be made available with the agreement that the model will be made publicly available and, if required, can be adapted for future evaluations. The electronic version of the model must be fully accessible and enable the reviewers, as well as the public, to view all formulae and relationships used in the model and to execute the model with different input data. To facilitate the review of the model, the electronic version should include a user manual describing which software and hardware is required, how the inputs into the model can be changed, how these inputs can be found in the model, how the model can be executed, and how results can be extracted.

4.3 Benefit

The methods used to determine the benefit of interventions within the framework of benefit assessments are described in Chapter 3. If the time horizon of the HEE is longer than the one used in the studies that are included in the benefit assessment and that form the basis of the HEE, the benefit proven by studies is to be distinguished from the modelled benefit.

4.3.1 Transfer and presentation of the benefit

For the integration of the benefit into the HEE by means of the efficiency frontier, the benefit needs to be approximately cardinally scaled. In the HEE the approximately cardinally scaled benefit (derived directly from study results when applicable) or a transformed approximately cardinally scaled benefit can be plotted on the vertical axis. Limiting the condition that a benefit “only” has to be approximately cardinally scaled is based on the following consideration: A scale used to measure benefit does not have to be cardinally scaled across its entire range. It is sufficient if it fulfils the criterion of being cardinally scaled across the range relevant for the definition of the patient-relevant added benefit. For instance, different measurement instruments often show so-called floor or ceiling effects at the margins of their value ranges, yet are cardinally scaled across the remaining range [57,182,452].

No specific approach to determine the valuation of benefit on a cardinal scale is recommended here, as each therapeutic area can offer different options that fulfil the requirement of assessing benefit on a cardinal scale.

4.3.2 Outcomes

The benefit can be presented on the vertical axis of the efficiency frontier by means of individual or aggregated patient-relevant outcomes (see Section 3.1.1 for the definition of patient-relevant medical benefit or harm). If several patient-relevant outcomes are presented next to each other, a separate efficiency frontier is created for each patient-relevant outcome. Alternatively, the benefit is aggregated into a single measure of overall benefit, which is subsequently plotted in an efficiency frontier. In a very general definition, a measure of overall benefit is an aggregation of the assessment of benefit and harm into one dimension, whereby different patient-relevant outcomes are summarized into a single measure. It can be considered both in the benefit assessment and in the HEE. The requirements presented in this chapter for the determination of a measure of overall benefit also apply if it is used within the framework of the benefit assessment.

4.3.3 Measure of overall benefit

On an international level, different measures exist to express or determine the overall benefit. These include the quality-adjusted life year (QALY) and the disability-adjusted life year (DALY). Other measures such as the saved young life equivalent [416] or healthy years equivalent (HYE) [206] were introduced with the objective of correcting weaknesses in the QALY, the most widely distributed instrument.

In this context, depending on the methodological approach or economic theory, the terms “preferences”, “utilities” or “values” are used in the scientific literature [159]. We refer to the further debate of the terms and relevance of measurement instruments in relation to the issue of a “welfarist” versus an “extra-welfarist” framework [69], but do not discuss this issue further here. Following SGB V, the following text speaks of weights by means of which individual patient-relevant outcomes can be transferred into a measure of overall benefit.

If the G-BA specifies the measure of overall benefit for an HEE according to § 35b (1) Sentence 2 SGB V (see Section 4.9), a respective instrument and, if applicable, the measurement methods specified for this purpose or an already specified weighting of outcomes are used following the requirements of the commission. The results should be made available to the decision maker together with the weighting of outcomes. The option hereby arises for the decision maker to negotiate a reimbursement price weighted by means of several added benefit-based reimbursement prices.

A) QALY as a measure of overall benefit

To calculate QALYs, weights for health states are determined. In this context respondents balance how they perceive or appraise these health states. The result is then an index score for each health state. Under integration of the duration of the corresponding health states, these weights, largely referred to as utilities (or utility values), can be transformed into QALYs. The determination and calculation of utility values is, for example, presented in Puhan et al. [445], Lipscomb et al. [363], and Tierney et al. [553].

The Institute does not rule out the possibility of using QALYs as a measure of overall benefit. QALYs should only be used if the incorporated values on the health states are first determined in affected persons who currently or in the past experienced these health states. The data should have been collected from participants of clinical studies. If generic index instruments are used, a scale validated in Germany must be used for the determination of the utility value. The use of QALYs, as well as their determination and conversion into a German scale, must in each case be presented in a comprehensible manner and justified. Apart from that, all usual standards for the respective procedures and instruments apply: i.e. evidence of objectivity, reliability, validity, and responsiveness must be available. Parallel to the use of a generic instrument, disease-specific instruments to determine quality of life in clinical studies should be applied. The mapping of disease-specific to generic instruments is therefore discouraged.

In view of the ongoing discussion on the advantages and disadvantages of different instruments, particularly the multi-attribute utility instruments (MAUI), with which quality of life, subjective well-being or utility values can be (or are supposed to be) determined or depicted, one has to say that no general recommendation can be issued. The choice of instrument depends on which of these 3 concepts is to be the most prominent one and which dimensions of quality of life are preferably to be determined [450].

There is no resumption here to the scientific debate about the ethical and methodological problems of the QALY concept itself and their solution or a linked willingness-to-pay threshold in an HEE, nor of the use of the QALY for the pure weighing of benefit and harm. In this context we refer to a number of publications [137,153,154,246,363,375,391,417,573].

B) Determination of preferences to establish a measure of overall benefit

If a measure of overall benefit for the comparison of interventions is to be determined, in addition to the disease-spanning measures named above, procedures for multi-criteria decision-making or determining preferences can be applied. For outcomes weighted by means of these procedures, all requirements according to SGB V and the Regulation for Early Benefit Assessment of New Pharmaceuticals (ANV[18]) apply. Surrogates can only be used if validity is proven. In the area of health care, the analytic hierarchy process (AHP) and the conjoint analysis (CA) have largely established themselves as methods for multi-criteria decision-making or determining preferences [62,121,277,382,466]. In relation to a specific therapeutic indication, the Institute can thus resort to these procedures to generate a measure of overall benefit. However, there are still unsolved methodological problems in the use of these procedures, so that currently it is not planned to use them routinely.

For the AHP [151,152] a problem in decision-making is broken down into so-called criteria. These are then arranged in a hierarchy. For example, a drug can be assessed by means of the criteria “mortality”, “morbidity”, and “quality of life”. The criteria can be broken down into

[18] Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV

further subcriteria that can correspond to outcomes [276]. Participants in the AHP then respond to questions about the criteria in a binary way, i.e. on a specified scale they choose how much more a certain criterion means to them than another. By means of a procedure for matrix multiplication [468,470,471] the weights for the criteria and subcriteria can be determined via a so-called “right eigenvector”; these weights must add up to 1. A further development of the method, the analytic network process (ANP), also allows to weight criteria that are dependent of each other [467,469].

The CA belongs to the group of stated-preference techniques [62]. A decision is broken down into so-called attributes that can correspond to outcomes. For each attribute levels are specified. For a discrete choice experiment (DCE = choice-based CA), the choice alternatives (stimuli) are compiled from the attributes with different levels. The respondents are then confronted with a set of (theoretical) scenarios (choice scenario = choice set) consisting of at least 2 stimuli. From the choice of scenarios, coefficients for the levels of the attributes are then determined in a regression model. The influence of the attributes on the decision can be presented by subsequently forming weights for the attributes. These weights can in turn be standardized to 1.

In its development, the AHP was targeted towards decision-making in the event of opposing aims in committees, for example, the management of a company, and the CA was targeted towards determining preferences to predict purchasing decisions and enable product adaption. Meanwhile, both procedures play a role in the identification and prioritization of patient-relevant outcomes, for example, before planning a study, and in the determination of the net benefit (measure of overall benefit) of interventions [118,402].

A clear allocation with regard to which procedure should be preferred in which situation can thus hardly be inferred. An AHP seems to be more suitable if a decision is to be made in a closed group [276,278]. In contrast, one would conduct a CA if one also wanted to consider compensation for lost benefit if an intervention is not reimbursed. Incidentally, it is also possible to calculate QALYs by means of CA [196,229]. However, when choosing either procedure the following criteria should be used: For the CA a maximum of 6 to 7 attributes can be included; no such limit applies to the AHP. Furthermore, the AHP seems to require lower cognitive effort from the respondents, which, depending on the therapeutic indication, could be considered. These evaluations can currently only partly be based on empirical data so that an evidence-driven choice of either procedure is not currently possible. In addition, there is a need for research on some issues, such as the reliability of both procedures.

The strength and weaknesses of both methods cannot be described in detail here [413]. Comprehensibility in the planning, conduct, analysis, and evaluation of each implementation is thus crucial. For the CA there is a basic list of criteria to ensure high quality, transparency, and reliability of the results of a CA [61]; several of the requirements also apply to the conduct of an AHP.

The following requirements should be fulfilled in detail in the planning, conduct, analysis, and evaluation of the results of surveys using either procedure:

- completeness of the criteria or attributes
- comprehensive documentation of the approach of selecting the respondents and description of the extent to which they are representative (based on sociodemographic and disease-specific factors) for the collective of affected persons

It must be reported not only who participates in the survey, but also how they were recruited. Furthermore, a sample size must be planned. For the CA there are rules of thumb for a sample size estimation [312]. For the AHP there is currently no method for estimating a sample size; however, at least criteria of representativeness can be used here that are also used for other surveys (sample size, type of drawing of the sample, etc.):

- investigation of the population surveyed with regard to homogeneity
- comprehensive documentation of the analysis, together with the handover of raw data, including the verbatim questions
- language, selection and supervision of the implementation, including an assessment of bias through the type of design (a language appropriate for the respondents should be chosen)
- investigation of the consistency and uncertainty of the results by conduct of suitable analyses (e.g. sensitivity analyses).

4.3.4 Uncertainty and distribution of benefit data

For estimated effects within the framework of a benefit assessment, confidence intervals or credible intervals (if Bayesian methods are chosen, see Sections 8.3.2 and 8.3.9) can generally be calculated that indicate the precision or uncertainty of the point estimates. Appropriate assumptions should be made for the further investigation of uncertainty, as many effects are not normally distributed.

Estimates from indirect comparisons (see Section 8.3.9) are more subject to uncertainty than estimates from direct comparisons; this is pointed out in the assessment of uncertainty. For estimates from indirect comparisons that, for example, deviate from each other due to different assumptions on a-priori distributions, scenario analyses are potentially performed.

Also in particular for the overall measure of benefit, the investigations of uncertainty (sensitivity analyses) stipulated in Section 4.7 must be conducted.

4.4 Costs

4.4.1 Perspective and costs to be considered

Depending on the commission, the following perspectives can be considered: the (pure) SHI perspective, the SHI insurant perspective, the social insurance perspective or the perspective of individual social insurance branches, as well as the societal perspective. In the following text the relevant costs to be considered are distinguished according to perspectives.

From the (pure) SHI perspective all direct reimbursable costs and transfer payments (e.g. sickness allowance) are considered. Furthermore, insofar as relevant for the HEE, the proportions can be considered of contributions to pension insurance, long-term care insurance, and unemployment insurance that the SHI must bear in the case of disease after 6 weeks of incapacity for work, as well as losses in contributions (during the payment of sickness allowance).

In the SHI insurant perspective, in addition to the direct reimbursable costs, insurants' own non-reimbursable co-payments need to be considered (see Section 4.4.2). In contrast, sickness allowance is not calculated, as the payments are merely redistributed from the SHI to the insurants, so that no additional costs for the community of insurants are incurred [465]. Likewise, losses of contributions for the SHI due to sickness are not considered.

Table 6: Perspective and relevant costs to be considered[19]

Cost category / Perspective	Direct medical costs		Direct non-medical costs		Indirect costs	Transfer payments
	Reimbursable	**Non-reimbursable**	**Reimbursable**	**Non-reimbursable**	**-**	**-**
Society	Yes	Yes	Yes	Yes	Yes	No
Social insurance	Yes	No	Yes	No	No	Yes
Community of SHI insurants	Yes	Yes	Yes	Yes	No	No
SHI	Yes	No	Yes	No	No	Yes

SHI: statutory health insurance

In contrast to the SHI insurant perspective, in the social insurance perspective or that of individual social insurance branches, no co-payments of insurants are calculated. Disease-related, reimbursable expenses including transfer payments are considered.

[19] Depending on the perspective adopted, the content of the respective cost category can differ. In a narrower interpretation of the community of SHI insurants, co-payments, for example, are considered, but no further expenditure of the insurants. This is specified in the G-BA's commissions.

In the societal perspective, cost components are considered independently of who bears them and who is affected by the effects of an intervention. In general, costs should be considered that are incurred by all social insurance branches and other affected parties (see Table 6). Time expenditure of patients (and/or potentially their relatives) representing loss of working time is not considered once again as time expenditure. Together with the consideration of productivity losses this would lead to double counting. Likewise, transfer payments and SHI-funded contributions to social insurance branches are not considered, as they are merely redistributed and no additional costs are incurred from an economic point of view [465].

In general, when determining costs it should be evaluated in each perspective whether these costs and, if applicable, cost savings, are relevant for the interventions, therapeutic areas, and patient groups investigated.

4.4.2 Distinction of costs

A) Direct costs

Direct medical costs refer to the resource use in the current and future provision of healthcare services. They are further divided into direct medical and direct non-medical costs. Direct medical costs are understood to be resource use arising from the provision of health care in the healthcare sector. They include costs, for example, for hospital stays, outpatient contacts with physicians, drugs, and medical remedies and aids. Direct non-medical costs comprise resources supporting the provision of healthcare services in the healthcare sector, for example, travel costs to clinics where medical interventions are performed or the evaluated disease-related time expenditure of affected patients and their care-providing relatives.

Reimbursable costs comprise expenditure for healthcare services funded by the SHI or other social insurance branches. Non-reimbursable medical costs are services directly borne by the insurants, such as co-payments for drugs, medical remedies and aids, and outpatient contacts with physicians. Non-reimbursable non-medical costs are, for example, disease-related net losses of income[20] (e.g. financial losses of patients receiving sickness allowance below their net income) or the time expenditure of affected patients and relatives.

Most empirical studies do not consider the effects on the leisure time of affected patients and relatives. In this respect the Institute does not regularly consider the time expenditure of these persons in the societal perspective. In the event that representative and valid information sources on time expenditure are nevertheless available, this expenditure can be considered in sensitivity analyses from the societal perspective. The quality of life of relatives is generally

[20] Strictly speaking the disease-related losses of net income refer to the difference between the net income of healthy persons and that of sick persons, taking into account co-payments for healthcare services to treat the disease. However, within the framework of the SHI insurant perspective, co-payments are not considered as reimbursable costs, so that the net losses of income can be determined from the difference between the sickness allowance paid and the net income of a healthy person.

not considered on the benefit side. If their losses of leisure time are investigated, they should also be assessed on the cost side [68,311,425,585].

B) Indirect costs

Indirect costs refer to productivity losses in the event of incapacity for work, occupational invalidity (in the event of long-term disease or disability), and premature death.

The Institute primarily considers productivity losses on the cost side. This is also largely recommended by the literature [70,71,94,159,311,500,501]. To avoid double counting, productivity losses due to premature death (mortality costs) should not be recorded on the cost side if mortality is already considered on the benefit side. Mortality costs are only represented on the cost side for those cases in which the outcome investigated does not refer to mortality or survival time. Costs for society (losses of taxes and contributions to social insurance) are always represented on the cost side [311,500,501].

On an international level it is being discussed whether unpaid work (e.g. housework) should also be taken into account in an HEE. As a rule, this is currently not considered by the Institute.

C) Transfer payments

Transfer payments can be considered, insofar as relevant for the HEE. Generally, transfer payments should not be considered if payments are only redistributed and thus no additional costs are incurred for the perspective selected.

D) Intangible costs

Intangible costs are experiences not directly calculable as resource use or evaluable in monetary units, such as pain or anxiety on the part of the patients treated. Following international health economic standards they should be reported on the benefit side, insofar as data on these details are available.

E) Future costs

Furthermore, the health economic literature often proposes a distinction between intervention-associated and non-intervention-associated (future) costs. Intervention-associated costs are, for example, costs incurred for drugs or check-ups after a heart attack, whereas non-intervention-associated costs would be, for example, treatment costs for cancer occurring years later, where treatment has no connection to that for the heart attack.

The consideration of non-intervention-associated costs is the subject of controversial debate [68,159,207,371]. Intervention and non-intervention-associated costs are distinguished from each other depending on the commission. If the extension of life is relevant for the HEE, in the base case the intervention-associated future costs are considered (both for given life expectancy and for life years gained). Non-intervention-associated future costs can be considered in separate sensitivity analyses (not for given life expectancy, as this is identical for all strategies, but for life years gained).

F) Investment and implementation costs

If one-off costs to finance the provision or implementation of healthcare services arise explicitly for the SHI or the community of SHI insurants, the investment and implementation costs should be appropriately considered. This should be investigated via sensitivity analyses.

4.4.3 Steps for cost estimation

In principle, costs should be determined as precisely as possible. Methods and sources used, as well as results, should be described for the individual steps of cost estimation. The estimation of the costs considered in the model usually follows a 4-step process:

- identification of resources
- quantification of resources
- assessment of resources and
- calculation of the costs considered in the model according to health states and, if applicable, cycles

A) Identification of resources

Within the framework of the identification of resources the healthcare services used for treating the disease must be determined (see Section 4.1.5). The information should preferably be up to date and can be obtained from the sources described in Section 4.4.4.

B) Quantification of resources

The frequency of use, the proportion of the relevant patient populations using each service, and the duration of the service must be determined. Costs for services that are used very infrequently and/or have only a slight impact on the results should be described, but are not necessarily considered in the calculation [159].

Both a micro- or macro(gross)-costing approach [543,544] can be applied and combined to quantify resource use. The degree of precision of quantification is determined, among other things, by the reimbursement system and the corresponding degree of aggregation of the services.

Both approaches can be applied as a bottom-up or top-down approach [494,543,544] if either the resources used are measured on the basis of the individual patients or an (average) distribution to patients is performed on the basis of highly aggregated data (expenditure for a disease).

C) Evaluation of resources

SHI insurant perspective

In general, regulated and negotiated prices (i.e. prices that have not been exclusively developed via market mechanisms) determine expenditure and represent the opportunity costs of the community of SHI insurants. As described before, the reimbursement system determines the maximum degree of precision in the determination of expenditure of reimbursable costs. For instance, from the SHI insurant perspective, diagnosis related groups (DRGs) and the Uniform Value Scale[21] represent the best-possible evaluation for the inpatient and outpatient sector, respectively.

In the cost estimation for drugs, one distinguishes between the inpatient and outpatient sector. In the inpatient sector, drugs are normally part of the corresponding lump sum reimbursement. If additional fees are negotiated for relevant drugs or these are reimbursed via "new examination and treatment methods"[22], the corresponding costs should be determined and considered in the HEE. In the outpatient sector, at first the pharmacy retail prices are used as a basis for price calculation. If reference prices are available, these must be provided and are reduced by pharmacy and manufacturer discounts. Discounts that were negotiated by a single SHI fund or a group of funds and thus subject to confidentiality are not depicted in the HEE. As a general rule, following the principle of efficiency the most economical representative of a drug or drug class is selected. Relevant price changes over time must be considered.

Non-reimbursable costs are partly regulated, so that here one can draw upon the corresponding standardization in the evaluation of resources (e.g. co-payment regulations in the inpatient sector and for drugs). These costs are presented separately in the SHI insurant perspective.

Specific features of further perspectives

Only aggregated data may be available in the social insurance perspective, depending on the insurance branch. In this case the resources should be assessed by means of a top-down approach on the basis of the respective statistics.

When calculating costs from the societal perspective, theoretically one should consider that the societal opportunity costs normally differ from the administrative prices, as these prices only represent the perspective of the payer. For instance, case fees do not include costs for the building of hospitals; from a societal perspective, these costs would need to be allocated to each case fee. The Institute is aware of this theoretical discussion. However, it follows international standards of other HTA organizations, which also use administrative prices in the societal perspective, as a different approach (due to missing data, e.g. on the actual costs

[21] Einheitlicher Bewertungsmaßstab

[22] Diagnostic or therapeutic interventions that are not usually reimbursed by the SHI; however, they may be reimbursed in exceptional and justified cases.

that would need to be allocated to case fees for the building of hospitals) would be subject to great uncertainty. It is usually international practice in HEE only to additionally investigate indirect costs. If the time expenditure of affected persons or relatives is considered in the cost estimation, this is evaluated with the net wage.

Evaluation of indirect costs

For productivity losses, in the base case the Institute considers the friction cost approach [225,334], as the human capital approach is based on some unrealistic assumptions (particularly full employment on the labour market). This estimation can be compared with the human capital approach in sensitivity analyses.

In the HEE the evaluation of indirect costs is based on individual labour costs (i.e. gross wage rate and non-wage labour costs – in Germany, employer contributions to social insurance) or the average labour costs. The calculation of the average labour costs per working day is based on the weighted average labour costs of people employed full-time and part-time in Germany. Approximatively, the "employee remuneration in Germany per year" divided by the "number of employees x 365" can be used (whereby Sundays and public holidays must be considered in the work incapacity days). Applying this approach to self-employed people should be discussed [220]. The friction costs are assumed as being 80% of the wage costs (as in the Netherlands [334]). The friction period is, insofar as no current data are available, set at 82 days; this corresponds to the average period in Germany for the year 2012 [58] within which a position can be filled. If the human capital approach is to be investigated in a sensitivity analysis, the future productivity losses are calculated on the basis of the average age of patients up to attainment of the standard retirement age.

D) Presentation of the costs considered in the model according to states or cycles

Before the costs can be fed into the model they must be available as average costs per patient according to health states and, depending on the model, also according to cycles.

Depending on the therapeutic indication, intervention, outcomes, and model, no direct information on the costs of the respective health states in the model is possibly available. On the basis of assumptions from further sources (see Section 4.4.4), the average costs of an intervention per patient and cost category (service areas and indirect costs) for the observation period can then be distributed to the different health states and cycles of the model.

For absorbing states in a Markov model it may be necessary to calculate transition costs that are incurred only once in the transition to this health state. This is then to be recommended if it is to be assumed that the costs in this state are considerably higher in the first cycle than in the subsequent cycles.

4.4.4 Data basis

Costs to be fed into the model must, as described above, be calculated for the different health states and, if applicable, for cycles of a model. The procedure for data collection and analysis, as well as all calculations and results, should be presented in a transparent manner.

To identify and quantify resources, information can be obtained from 3 types of sources: secondary data (primarily from SHI routine data), guidelines, and expert opinions. To determine prices, the Institute uses the respective relevant regulated or negotiated prices, for example, from the "Lauer-Taxe" (Pharmacy Price Schedule), the Uniform Value Scale, the DRG Catalogue, or statistics from the pension insurance or Federal Statistical Office.

In this context, secondary data in the form of analysed SHI routine data based on a representative sample are the data source of first choice. If current analyses cannot be obtained from the literature, those applying the model should preferably perform their own analyses.

Guidelines or results from expert surveys can be used as supplementary information if routine data do not sufficiently depict the provision of health care in all states of the model. Evidence-based guidelines from the German healthcare system should preferably be used (see Section 4.1.7). If they are not available in the therapeutic area to be investigated, it should be carefully considered and presented transparently whether other German guidelines can be used or if expert surveys should be drawn upon. Expert surveys are an option only if data cannot be obtained from more representative sources or if these data do not fully cover the level of detail required in the health states (see also Section 4.1.7).

Due to system differences, the transferability of care pathways and cost data from other healthcare systems is rarely given and is only possible under very strict preconditions [354,511]. The transferability of cost data from the following countries is not excluded as a matter of principle, as their inpatient and outpatient healthcare sectors are similar to those of the German system: Austria, Switzerland, the Netherlands, Belgium, and France. However, the use of data from these countries must in each case be justified and discussed. Cost data from other countries must not be used in an HEE.

4.4.5 Uncertainty and distribution of cost data

The uncertainty in cost data should be addressed in an adequate manner. Cost data are inherently continuous, positive, without an upper limit and generally not normally distributed, but skewed to the right [159].

4.4.6 Adjustment for inflation and discounting

A) Adjustment for inflation

If cost data originate from different time periods, they must be adjusted for inflation. The Harmonised Index of Consumer Prices (HICP) of the Federal Statistical Office should be used as the source for annual inflation [525]. Further price increase rates for individual areas of health care (e.g. drugs) can be considered from other sources within the framework of a sensitivity analysis.

B) Discounting

If costs and benefits are incurred in periods lasting longer than a year, in the base case they are discounted after the first year with an identical constant rate of 3% to the current period [32,89,129,159,364]. Likewise, identical constant rates of 0 and 5% should be used in sensitivity analyses; any deviations must be justified.

4.5 Epidemiological data

4.5.1 Data

Current epidemiological data are indispensable for an HEE. Besides being used to estimate disease burden, data on the prevalence and incidence in Germany are also used to quantify changes in the SHI budget in a budget impact analysis. Statements are therefore required on whether changes in incidence, prevalence or mortality are to be expected within the next 5 years. Furthermore, data on mortality are important in order to illustrate disease-related mortality and so-called background mortality.

The basic probabilities for events play a special role in modelling. In a model, details on the outcome-related event frequencies or probabilities are required for each outcome, which are considered as baseline values in the decision-analytic model.

4.5.2 Data basis

Epidemiologic data may be obtained from secondary data such as public data collections and SHI routine data (see Section 4.4.4) as well as registry data and, if applicable, scientific publications (see Section 4.1.7). If available and obtainable in an appropriate form (e.g. suitable age groups), public data collections (e.g. from the Robert Koch Institute) should be primarily considered due to their high methodological consistency. Registry data are a special case. Independently of the assessment of the quality of a registry, these data are often related only to a specific region. Their transferability must therefore be evaluated. If scientific publications are available in which epidemiological data can be determined, these data can potentially be used directly. Usability must be clarified in the individual case, as the studies often use approaches that are methodologically different. Cohort studies or sufficiently large and representative samples are to be preferred. The methodological quality of the underlying study can, among other things, be assessed by means of the requirements of “good epidemiological practice”.

4.5.3 Uncertainty and distribution of epidemiological data

The uncertainty in epidemiological data should be addressed in an adequate manner. In particular the uncertainty of data on the baseline risk and on mortality must be adequately considered both in the sensitivity analyses and in the distributions.

4.6 Presentation of results as an efficiency frontier

An efficiency frontier is drawn on the basis of the economic evaluation of interventions within a therapeutic area. It is generated from the most efficient interventions of the available comparators and can serve to infer recommendations on decisions for the intervention(s) under assessment. It can provide information on the negotiation of reimbursement prices without recurring to a threshold for the willingness-to-pay, for which there is currently no consent in Germany.

4.6.1 Definition

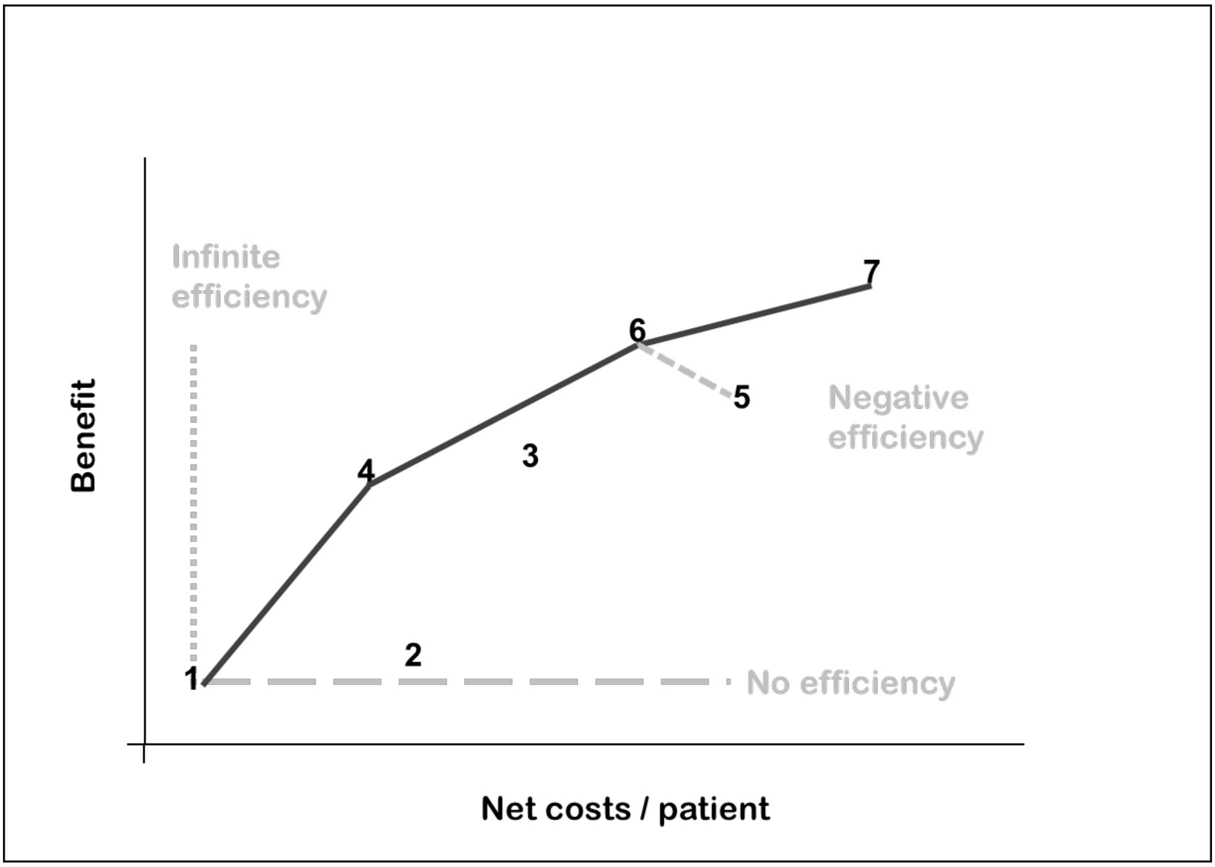
The efficiency frontier graphically compares the outcome-related benefit of available interventions within a therapeutic area with the net costs of these interventions. In this context, if required, the benefit is transferred into an approximately cardinally scaled measure.[23] Those interventions that are most efficient in respect of benefits and costs form the efficiency frontier.

4.6.2 Course of the procedure

In the procedure it must be distinguished between the new intervention(s) under assessment for price negotiations, for instance, and the interventions that form the efficiency frontier (comparators). The latter are those interventions currently used and reimbursed in Germany for the therapeutic area under assessment. Their costs and benefits are determined and depicted graphically.

In the presentation of the efficiency frontier, the interventions with greater efficiency are plotted from left to right. The gradient of the theoretical connecting line between 2 interventions (the line segment) provides the incremental benefit per incremental costs (see Figure 10).

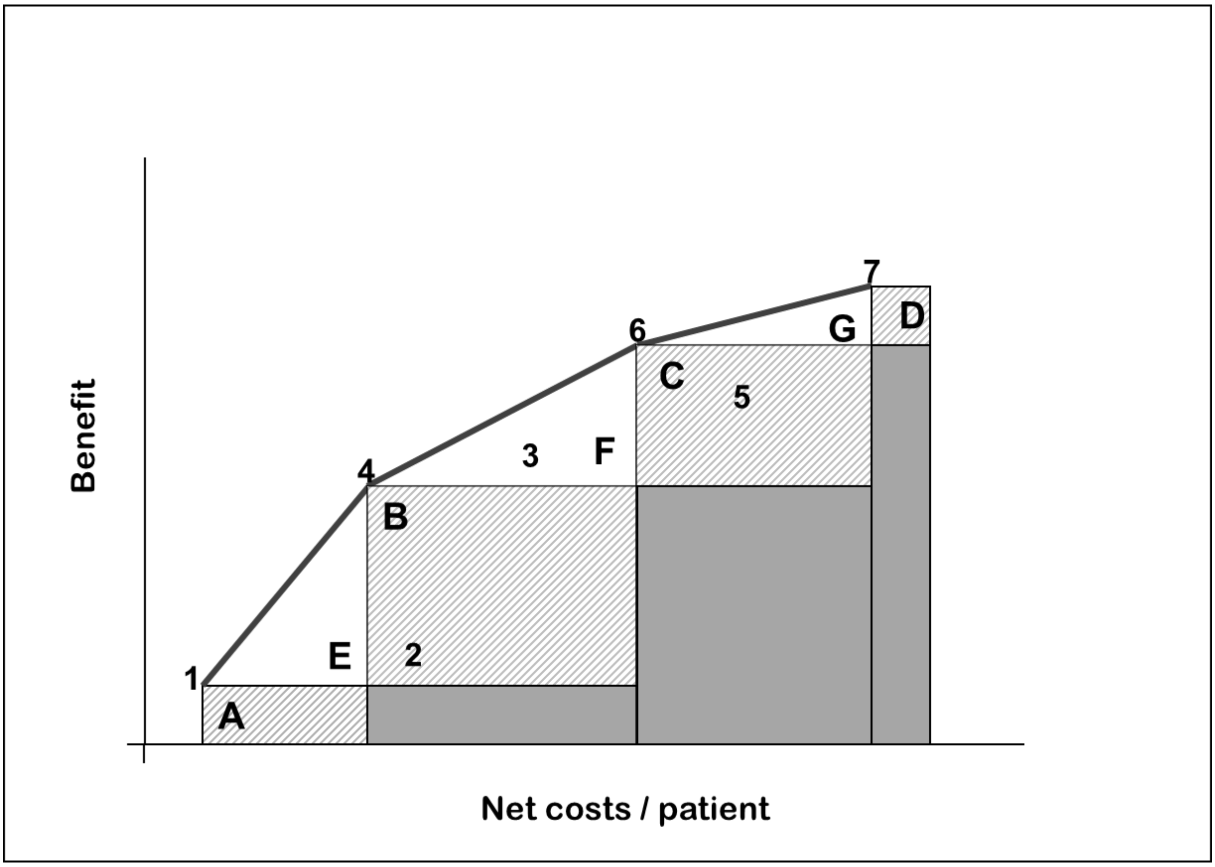
[23] If the patient-relevant added benefit determined in the prior benefit assessment already shows approximately cardinally scaled characteristics, it may be directly transferred into the HEE.



A horizontal line (gradient angle = 0°) indicates no efficiency, while a vertical line (gradient angle = 90°) indicates infinite efficiency. A positive gradient in ascending order (e.g. between Intervention 6 and Intervention 7) indicates an incremental benefit with higher costs, whereas a negative gradient (e.g. between Intervention 6 and Intervention 5) indicates lesser benefit with higher costs.

Figure 10: Interpretation of the gradient of the theoretical efficiency frontier

The positions of the interventions, such as Intervention 3 in Figure 10, require further interpretation, as they do not show negative efficiency in comparison with interventions already introduced (e.g. Intervention 4). In Figure 11, the area below the theoretical efficiency frontier is further divided by a series of rectangles (A to D). Each of these rectangles contains all interventions showing negative efficiency (higher costs with lesser benefit) on the theoretical efficiency frontier versus at least one intervention already available in the market. Interventions in these subareas (e.g. Intervention 2 or Intervention 5 in Figure 11) are clearly inefficient. This leaves 3 triangles (E, F and G) in which the interventions are not clearly inefficient. Usually, interventions plotted in these triangles are not part of the efficiency frontier because a theoretical combination of both interventions forming the hypotenuse of the triangle will provide a greater benefit with lower costs (so-called extended dominance).



The theoretical efficiency frontier (solid line) joins those interventions that are efficient relative to any other intervention or to their combinations. Interventions in Rectangles A to D (e.g. Intervention 2 or Intervention 5) are clearly inefficient. Intervention 3 is in one of the remaining triangular areas (E to G) and is not clearly inefficient. Theoretically an extended dominance would result from the combination of Intervention 4 and Intervention 6, but this may not be feasible in practice.

Figure 11: Absolute versus extended dominance

Such a combination is not always possible in practice. This would imply that if the price of Intervention 3 is fixed, then the beneficiaries would need to be redistributed to Intervention 4 and Intervention 6 to achieve greater efficiency. This may be clinically undesirable and difficult to justify, since it would lead to those receiving Intervention 4 being in a worse position. The alternative of allowing beneficiaries to switch between both therapies over time is clearly not possible with most surgical interventions, and presumably not for many drug interventions either. Thus, there may be many situations where interventions within the triangular areas constitute part of the practical efficiency frontier. If the criterion of extended dominance is not applied, then this results in a stepped absolute efficiency frontier arising from the connection of the upper segments of the shaded rectangles as opposed to the triangular areas. However, in this context it needs to be considered that the absolute efficiency frontier no longer provides a gradient in the sense of a reciprocal of the willingness-to-pay and thus no threshold values would be determined.

4.6.3 Construction of the efficiency frontier

The efficiency frontier is constructed in such a way that it represents the relevant interventions in a given therapeutic area. This involves:

- Full, detailed specification of the therapeutic area of interest. This may include the specific disease, the conditions of treatment (e.g. inpatient care), target population, sequence of therapy (first, second-line therapy, etc.), and the information on whether it is a mono-therapy or combination therapy.
- Positioning of existing therapies on the basis of their benefits and costs.
- Plotting of the interventions on a coordinate system with the benefit on the vertical (y-) axis and the costs on the horizontal (x-) axis.[24] In this context, in accordance with good scientific practice, one should ensure constant scaling (at least per outcome) of the axes.
- Drawing of the efficiency frontier.

When evaluating new interventions, their health effects and costs in the therapeutic area in question are then additionally presented.

A) Vertical axis

- Benefit and harm are plotted on a vertical axis. In this context, one should observe a positive value range, so that the efficiency frontier depicts the increased benefit or decreased harm (if applicable, e.g. multiplication with “-1” may be required or conversion to the complementary event “1-harm”).
- Benefit or harm is presented by means of patient-relevant outcomes that must be operationalized in an appropriate manner (e.g. quality-of-life scores).
- Benefit or harm is transferred onto the vertical axis. This transfer can be performed with inclusion of modelling.

B) Horizontal axis

- The total net costs per patient are plotted on the horizontal axis.
- As a rule the costs are calculated from the SHI insurant perspective. Depending on the commission, they may contain additional costs arising from extended perspectives, (e.g. social insurance perspective, societal perspective).
- The costs currently to be expected are used as costs.

In order to estimate the costs of each intervention and plot them on the coordinate system of the efficiency frontier, several conditions must be met. The costs should correspond to those

[24] This could also be presented as a table. However, the relationships would not be so graphically visible.

that would be incurred in current practice. The total net cost per patient must be plotted on the efficiency frontier.

To determine the cost-effectiveness ratio of (new) interventions with more benefit and more costs than the comparators, the last segment of the efficiency frontier is extended (see Section 4.1.9, as well as Figure 9 and Figure 12).

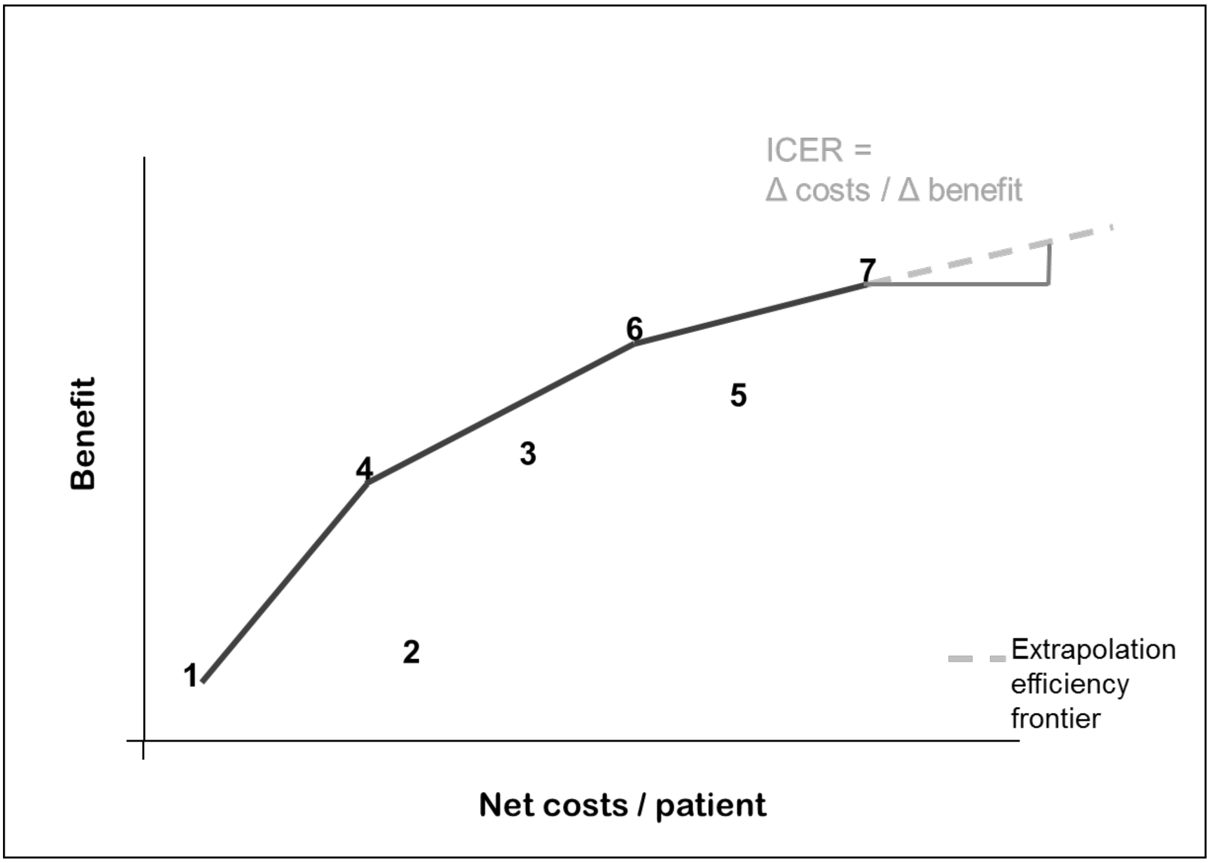
Depending on the number of outcomes taken from the benefit assessment previously conducted, several efficiency frontiers can be derived and presented.[25] If outcome weighting was performed, this is also presented. If a measure of overall benefit was specified, this is to be regarded as the primary result.

C) Definition of the origin of coordinate system

The point “no intervention” (i.e. the natural course) also requires an assessment. Although it could possibly be regarded to be the coordinate origin (zero benefit and zero costs), this is rarely appropriate, as the non-conduct of an intervention may still produce costs and health effects, for example, due to the untreated disease, monitoring, etc. Data on the natural course should therefore also be collected. In this context a common assumption is that placebo most likely corresponds to the natural course. This should be assessed in relation to each commission.

If the origin of the efficiency frontier does not correspond to the zero point, the efficiency frontiers (at least per outcome) must be plotted in equally scaled coordinate systems. The intervention that lies the furthest down and to the left will generally become the origin of the efficiency frontier (see Figure 12). For reasons of comparability of the presentation of different efficiency frontiers, a shifting of the zero point (of the coordinate system) should be rejected.

[25] This also refers to the separate presentation of divergent aspects of harm in distinction from the patient-relevant added benefit.



The efficiency frontier starts in a different origin from the zero point of the coordinate system. The extension shows the incremental cost-effectiveness ratio (ICER) at which a (new) intervention with more benefit and more costs than the comparators is measured.

Figure 12: Presentation of the efficiency frontier

4.6.4 Special constellations

There are 2 special constellations in which, despite complete information, a recommendation for a new intervention cannot be directly inferred on the basis of the efficiency frontier:

1) The last intervention on the efficiency frontier dominates all other interventions and generates the same costs as the reference scenario. The gradient would thus be infinite (see presentation in Figure 10).
2) The last intervention on the efficiency frontier before the introduction of the innovation is more cost-efficient and has more benefit than all comparators, including the origin.

Both cases would result in a new origin, on which in each case the last intervention before the introduction of the innovative intervention would lie.

The budget impact analysis might deliver further data here by depicting the impact on the budget (see Section 4.1.9 and Section 4.8).

4.7 Uncertainty (sensitivity analyses)

The types of uncertainty are presented above (see Section 4.1.8). The uncertainty of many model parameters results from the fact that their value is estimated from samples. This type of uncertainty is often captured by confidence intervals or other statistical approaches for describing variability.

4.7.1 Quantification of uncertainty

For costs, uncertainty may exist regarding assumptions on resource use, for example, dosage of a drug over time. The model can also be of a stochastic design (it uses random numbers in the Monte Carlo draws). Different techniques can be applied to restrict this type of uncertainty [349,456,508].

Uncertainty also arises from the type of potential variability in the model structure described in Section 4.2, which needs to be considered in the investigation. Finally, even input parameters specified a priori, such as the discounting rate, can be varied to depict uncertainty arising from different discounting rates (see Section 4.4.6).

4.7.2 Sensitivity analyses

Parameter uncertainty as well as types of uncertainty that cannot be reduced are quantified. The Institute considers both univariate and multivariate deterministic as well as probabilistic sensitivity analyses (PSAs), and in its work follows the recommendations of the conjoint Modeling Good Research Practices Task Force Working Group of ISPOR and the Society for Medical Decision Making (SMDM) [63].

All analyses performed for this purpose should be fully documented, with minimum and maximum values for the parameters used and underlying assumptions. The following aspects must be specified for PSAs: probability distributions used and their sources, correlations between input parameters, and any structural variants.

Structural sensitivity analyses are performed to investigate the impact of a variation of assumptions in the model structure, for example, the number or type of the model states.

Presentation of the results of the sensitivity analyses

For the deterministic sensitivity analysis, extreme levels of the input parameters should be provided for which the new intervention possibly saves costs or lies above or below the efficiency frontier. For univariate and multivariate analyses the results must be presented in a table and in a tornado diagram in which the levels of the results are displayed as an interval for the corresponding intervals of the input parameters.

For PSAs the proportion of simulations for which cost savings or a position above or below the efficiency frontier arises is provided as a percentage. In the case of PSAs the results are presented as cumulative cost distributions.

4.7.3 Presentation of uncertainty by means of the net health benefit

When presenting results of sensitivity analyses, attention should be paid to the fact that the consideration of parameter uncertainty can on the one hand change the position of several or all interventions forming the efficiency frontier. On the other, the position of the intervention under assessment, which is contrasted with this efficiency frontier, can also change.

The net health benefit (NHB) is an established procedure for presenting results from PSAs [532]. By calculating the NHB, this problem is accounted for, as the NHB is a function both of the added benefit and added costs, and also of the efficiency frontier, and depicts the position of the intervention under assessment as the distance to the shifting efficiency frontier or to the shifting last segment of the efficiency frontier. For this reason, both the base case analyses, as well as the deterministic and probabilistic sensitivity analyses, should be conducted on the basis of the concept of the NHB calculation.

4.8 Budget impact analysis

A budget impact analysis (BIA) is an assessment of the direct financial consequences related to the reimbursement of an intervention in a healthcare system [558]. In a calculation model for a BIA, the proportion of patients who will potentially receive a new intervention is considered, as well as the dissemination of the intervention in the healthcare system, including its use in previously untreated patients. In particular, a BIA predicts how a change in the mix of interventions used for a certain disease might in future influence expenditure for a therapeutic area [386].

The purpose of a BIA is not so much to produce exact estimates of the financial consequences of the use of an intervention, but rather to provide a reliable calculation framework that allows the decision maker to understand the possible expenditure effects of a new intervention (or of a change in the usage of existing interventions) [386]. Such a model is necessary, as many of the parameters vary depending on the constellation and are also subject to uncertainty. Thus, the result of the BIA is not a single value for the estimation of expenditure but rather a range resulting from the model.

4.8.1 Perspective in the budget impact analysis

The BIA should be undertaken from the perspective of the SHI or another relevant payer (see also Section 4.4.1). Any expenditure incurred or cost savings achieved outside this perspective are not included.

4.8.2 Time horizon in the budget impact analysis

The BIA should cover the time horizon most relevant to payers in view of their expenditure [386]. Since the impact on expenditure is likely to change over time after the new intervention has been introduced – both because of market adjustment and of long-term effects on the disease in question – this horizon should be estimated and presented for a period of 1 and 3 years [385]. The results must be presented as expenditure and cost savings per year instead of in the form of a single "net current value" [386]. Thus in this case no discounting of financial flows is allowed to be performed. If the result is presented as a total amount of costs for 3 years, the costs can be discounted accordingly (see Section 4.4.6).

4.8.3 Scenarios in the budget impact analysis

A BIA compares health care scenarios – each defined by a compilation of interventions – rather than specific individual interventions [386]. At least 2 scenarios must be considered: on the one hand the reference scenario, defined by the current mix of interventions, and on the other, the predicted new mix of interventions.

4.8.4 Population in the budget impact analysis

The size of the insured population likely to use the new intervention is one of the key factors determining the expected expenditure for the new intervention. The anticipated number of users results from the predicted utilization of the intervention within the target population. Any expected off-label use of the new intervention should not be considered in the primary BIA, but may be considered in sensitivity analyses [426]. When predicting the number of users, both the substitution of existing interventions and induced demand need to be taken into account.

4.8.5 Costs to be considered in the budget impact analysis

The costs (net costs, i.e. adjusted for cost savings, so-called cost-offsets) should be estimated according to the methods described in Section 4.4.

For the BIA, investment and implementation costs are – as far as possible and borne by the SHI – identified and quantified. They should be presented separately and organized according to cost categories, whereby a complete explanation of the method and the sources used for cost estimation must be included.

4.8.6 Presentation of results in the budget impact analysis

The results (in €) should be presented as a value range and not as single point estimates. Furthermore, both the total amount and the proportion related to annual expenditure should be displayed.

4.9 Specific aspects of a health economic evaluation according to §35b SGB V

4.9.1 Legal requirements and course of procedure

Some specific requirements apply for the HEE according to §35b SGB V. By default there are 2 constellations that can lead to an HEE within the framework of the benefit assessment of drugs according to § 35a SGB V:

1) If a pharmaceutical company disagrees with the decision by the G-BA that the drug under assessment has no added benefit or does not represent a therapeutic improvement,

according to §35a (5a) SGB V, the pharmaceutical company can demand that the G-BA commissions an HEE according to §35b SGB V or to §139a (3) No. 5 SGB V.

2) After a decision by the arbitration board, according to §130b (8) SGB V, both the pharmaceutical company and the SHI umbrella organization[26] can commission an HEE according to §35b SGB V.

If a pharmaceutical company and/or the SHI umbrella organization submit an application to the G-BA for an HEE according to §35b SGB V, further specific aspects arise during the course of the procedure, which are described in Section 2.1.4.

According to §130b (8) Sentence 3 SGB V, an HEE of drugs according to §35b SGB V serves the purpose of negotiating a reimbursement price that is to be negotiated in comparison with (an) appropriate comparator therapy or therapies. According to §35b SGB V, the G-BA specifies the following points in its commission of an HEE:

- appropriate comparator therapy and other drugs and treatment forms with which the drug under assessment is to be compared
- patient groups
- time period
- type of benefit and costs and
- measure of overall benefit

The basis of the HEE are 1) the results of clinical studies, 2) the results of health services research studies agreed upon with the G-BA or recognized by the G-BA after application by the pharmaceutical company, and 3) evidence provided by the pharmaceutical company (see §35b (1) Sentence 3 SGB V). Moreover, due to the legal situation in Germany (§35b (1) SGB V), as a rule the SHI insurant perspective is adopted. More details are described in the G-BA's Code of Procedure [211].

4.9.2 The net health benefit for calculation of added benefit-based reimbursement prices

As explained in Section 4.7.3, the NHB can be used to present uncertainty. On the basis of the expected value of the NHB of the intervention under assessment, an added benefit-based reimbursement price can also be derived via the further calculation of the cost-adjusted (added) benefit of the intervention under assessment [533].

The incremental NHB is calculated by means of the effect estimate for the benefits and the costs of the respective interventions as well as a threshold value. In this application the threshold value corresponds to the reciprocal of the gradient of the last (and potentially

[26] Spitzenverband Bund der Krankenkassen, GKV-Spitzenverband

extrapolated) segment of the efficiency frontier for cost-effective interventions (see Figure 13). If the NHB were about zero, then Intervention 8 would lie on the efficiency frontier determined by the gradient (1/ʎ) of the last segment of the efficiency frontier, and can also be assessed as cost-effective in comparison with the (per definition cost-effective) interventions forming the efficiency frontier. Accordingly, an added benefit-based reimbursement price is determined by means of the NHB by conversion and calculation of the maximum intervention costs that are necessary to ensure that the NHB is at least zero. The NHB can be estimated practically with the help of the model through iterative calculations.

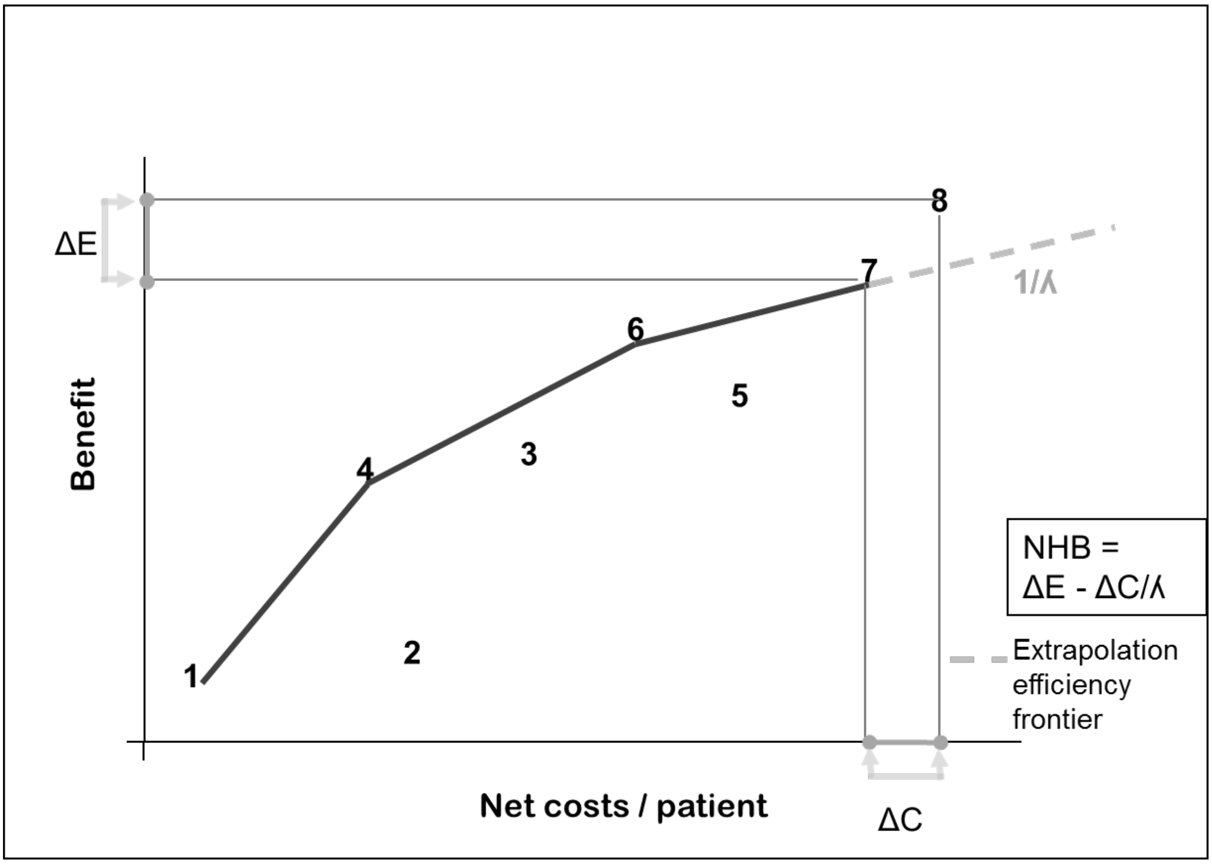


Figure 13: Presentation of an NHB > 0

4.9.3 Sensitivity analyses for calculation of added benefit-based reimbursement prices

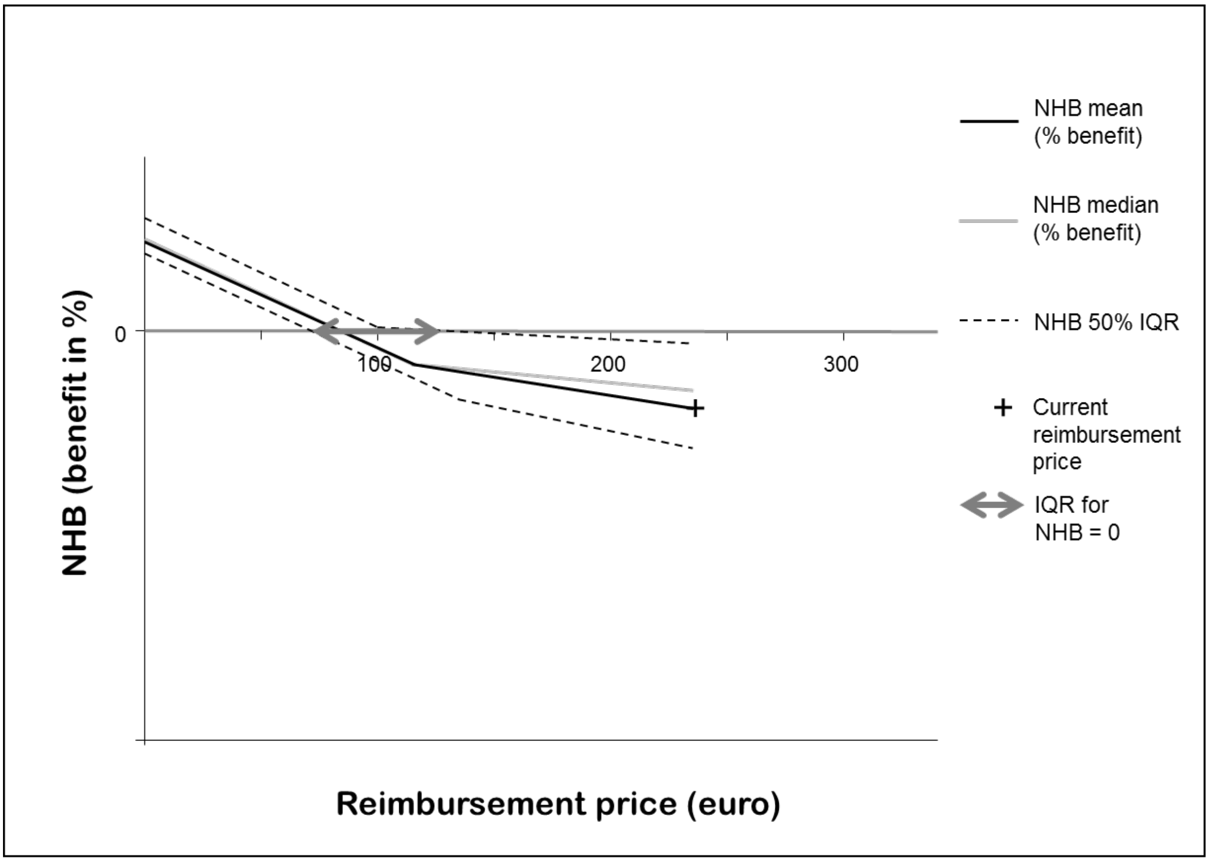
For the added benefit-based reimbursement price, price acceptance curves [187] and/or NHB values can be presented per efficiency frontier (see Section 4.9.2).

When using the NHB the results of the PSAs should be presented via the calculation and averaging of the respective expected NHB values for the intervention under assessment for a sufficiently large number of runs. In each run both the efficiency frontier and the position of the intervention under assessment relative to the efficiency frontier, and thus the respective NHB value, can change. From these values, the averaged NHB value of the intervention under assessment, as well as an interquartile region (IQR), can be calculated (see Section 4.9.4). In combination with the IQR, the expected NHB value indicates how large according to expectation the cost-adjusted (added) benefit is for the current added benefit-based reimbursement price, under consideration of the model uncertainty.

4.9.4 Interquartile region as a measure of dispersion for price negotiations

An IQR is provided to give the SHI umbrella organization and the pharmaceutical company a measure of dispersion for the negotiations on the basis of the results of the sensitivity analyses (see Section 4.7). The IQR includes all values of the NHB from the simulations margined by the lower and upper quartile (see Section 4.9.3). This means that the IQR covers those 50% of simulations in the PSAs that lie above the 25% lowest results and below the 25% highest results (see Figure 14). In principle it can also be meaningful to provide other regions with other measures.

Under consideration of the total uncertainty (implemented through PSAs), the IQR allows room to open possible reimbursement price negotiations within whose margins the uncertainty of the effect estimates and the costs are also considered.



For each possible reimbursement price the solid line indicates the average NHB to be expected (x-axis). At the position where the solid line crosses the x-axis, an added benefit-based reimbursement price can be read off in which the average NHB to be expected is zero, that is, neither positive nor negative.

Figure 14: Interquartile region of possible added benefit-based reimbursement prices (based on PSA) as a measure of dispersion for price negotiations

5 Clinical practice guidelines and health care analysis

5.1 Background

CPGs are systematically developed decision aids for service providers and patients enabling an appropriate approach to specific health problems. Their aim is to improve patient care. Their recommendations are informed by a systematic review of the evidence and an assessment of the benefits and harms of alternative treatment options [191,221]. CPGs can normatively describe standards in all areas of the health care chain, i.e. in diagnosis, treatment, rehabilitation or after-care. These health care standards contain essential information on the quality of care aimed for in a health care system. Determining a health care standard is a key precondition for drawing conclusions on the quality of care in a health care system.

The identification and description of health care standards by means of high-quality CPGs serve as the basis for different scientific analyses, for example, as a starting point for the development and update of DMPs (see Section 5.3). Likewise, by comparing these standards with specific health care structures, processes and outcomes, gaps in health care and potential for improvement can be detected (see Section 5.4). In the following text, this is described as a "health care analysis". Such an analysis enables conclusions on quality and efficiency issues of services provided within the framework of SHI (see §139a SGB V (3) No. 2).

The focus is on providing an overview of the whole picture of a disease. In addition, individual procedures or technologies may be examined, for example as a basis for further assessment in systematic reviews.

The aim is to present current (or to document lacking) health care standards for decision makers and other players in the health care system and, depending on the research question, to compare them with the specific health care situation in order to enable well-founded decisions to improve the quality of care in the health care system.

5.2 Identification of health care standards by means of clinical practice guidelines

5.2.1 Health care standards in clinical practice guidelines

Medical standard is defined by medical practice that, according to medical and scientific evidence and/or clinical experience, is accepted in the profession [248]. A CPG is a means of establishing a medical standard scientifically and institutionally.

Evidence-based CPGs are normally drawn upon in our department's reports to answer questions on health care standards. Evidence-based CPGs refer to CPGs whose recommendations are based on a systematic literature search, and are linked as a matter of principle to a level of evidence (LoE) and/or grade of recommendation (GoR), as well as to citations of the underlying primary and/or secondary literature (modified according to

AGREE[27] [4]. An evidence-based CPG does not assume that each individual recommendation included is linked to a high LoE. In general, CPGs that were prepared systematically and transparently, and are therefore evidence-based, also include recommendations founded on a weak evidence base [557].

5.2.2 Methodological appraisal of clinical practice guidelines

Information retrieval is conducted according to the procedures described in Chapter 7.

On an international level different instruments are used for the methodological appraisal of CPGs [577]. The AGREE instrument [4,374] and its revised version AGREE II [5,72-74] were developed and validated by a network of researchers and health policy makers and are the most widespread tools internationally. The German-language DELB[28] instrument of the Association of the Scientific Medical Professional Societies (AWMF[29]) and the Agency for Quality in Medicine (ÄZQ[30]) is also based on the appraisal tool of the AGREE Collaboration. To simplify any potential future comparison between the results of a CPG appraisal by the Institute and CPG appraisals published in other studies, AGREE is used as a rule in the Institute’s methodological appraisal of CPGs. The Institute is actively involved in the further development of the DELB instrument.

When preparing the report plan, the Institute specifies a priori whether, on the grounds of a research question, a methodological appraisal of CPGs should be performed with the AGREE instrument [5]. This tool consists of 23 key items assessed by means of a scale and organized in 6 domains. Each domain covers a separate dimension of CPG quality:

- Domain 1: scope and purpose
- Domain 2: stakeholder involvement
- Domain 3: rigour of development
- Domain 4: clarity and presentation
- Domain 5: applicability
- Domain 6: editorial independence

Each CPG appraisal is performed by 2 reviewers independently of each other.

A) Standardized domain scores

The domain scores are independent of each other, which is why for each CPG sum scores are calculated separately for the individual domains. As specified in the AGREE instrument, a

[27] Appraisal of Guidelines Research and Evaluation in Europe
[28] Deutsches Leitlinien-Bewertungs-(Instrument) (German Instrument for Methodological Guideline Appraisal)
[29] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften
[30] Ärztliches Zentrum für Qualität in der Medizin

standardization of the calculated domain scores is performed for better comparability between domains. These are presented in tables in the reports.

The instrument does not allow thresholds to be set for the assessment of domains. However, the individual standardized domain scores can be used for the comparison of CPGs.

B) Overall appraisal of methodological quality of clinical practice guidelines

In addition to calculation of the standardized domain scores, according to the procedures specified in the report plan, an overall assessment of CPG quality can be performed with the AGREE instrument [5].

5.2.3 Structured processing and evaluation of recommendations

A) Clinical practice guideline recommendations, levels of evidence and grades of recommendation

A recommendation is defined as a proposal for action for clinical or health system decisions. The recipients are generally professionals. In principle, in CPGs those statements are identified as recommendations that are formally indicated as such by the CPG authors. In addition, depending on the research question, recommendations not formally indicated may be identified by a linguistic label (e.g. “is recommended, must, should, can, could, be considered”, including negations or negative recommendations).

The developers of evidence-based CPGs use different systems to classify the LoE underlying recommendations and grade the strength of recommendations [24,161,235,341,499]. LoE should inform the reader or user of a CPG in brief about the strength (quality and quantity) of the evidence underlying the recommendation. “Evidence” is understood here to be the primary and secondary literature systematically searched for and evaluated by CPG developers. LoE with regard to the (benefit) assessment of medical interventions are generally based on a hierarchy of evidence and study types.

GoR provide the reader or user of a CPG with information on the strength of a CPG recommendation. They go beyond LoE, as they consider not only the strength of the underlying evidence, but in general also include the balancing of medical, social, patient-specific and/or economic benefits and corresponding risks of a recommendation [24,235,499]. They may also refer to the specific health care situation in a health care system.

B) Structured processing of recommendations

Within the framework of structured processing, the recommendations from the CPGs included are first listed in tables in their original language and separately for the health care aspects “prevention”, “diagnosis”, “therapy”, and/or “rehabilitation or after-care”. In addition, the LoE and/or GoR for a recommendation are listed in the extraction tables, insofar as they have been awarded by the CPG developers. Depending on the research question, further information may be presented.

As there is to date no internationally consented standardization of grading systems for evidence and recommendations, the LoE and GoR used by the individual CPG developers are generally noted and the corresponding grading systems documented. In order to better compare the systems of different CPG developers, if possible or planned, comparable LoE and GoR from different developers are summarized in comprehensive evidence and recommendation categories.

C) Evaluation of the recommendations extracted

Summarization of the recommendations

The evaluation of the extraction tables initially consists of a summarization of the content of the extracted recommendations on the individual aspects of care. In this context, if noted in the CPGs, both the LoE and GoR of the corresponding recommendations are presented.

Synthesis of key statements

If specified in the report plan, the information on content can be summarized from the recommendations of different CPGs on the same aspect of health care in a comprehensive "key statement". Key statements are presented in tables together with the information as to which CPG supports the particular statement with what evidence and/or recommendation category.

D) Identification of gaps or discrepancies in the presentation of health care standards

The structured processing and evaluation of CPG recommendations enable the identification of gaps, deviations or consistencies in the presentation of existing health care standards.

Depending on their objective, CPGs address certain health care areas up to the complete care chain of a disease. If recommendations are lacking in individual CPGs on the addressed areas of the care chain (e.g. on rehabilitation or treatment), a gap exists in the presentation of the health care standard. This may have various causes. The specifics of the evidence base are a crucial factor (e.g. missing, deficient, insufficiently transferable evidence). Aspects of health care are also of importance, such as the approval and reimbursement status in a health care system or opportunities available in the corresponding context.

Differences in CPG recommendations or in the allocation to LoE/GoR constitute deviations in the presentation of health care standards. Such deviations may, for example, occur if the content of recommendations differs or if nearly identical recommendations or key statements on the same health care aspect are supported by very different LoE/GoR. The detection of deviations may, among other things, indicate an uncertain evidence base or state of consensus for a distinct aspect or the influence of context-specific factors.

5.3 Recommendations on disease management programmes

The health care standards identified by means of the procedure described in Section 5.2 can serve as a basis for the preparation of DMPs. In addition, comparison of health care standards with existing DMP recommendations can determine a potential need for revision of the DMP. Recommendations that are consistent in content and were allocated a high GoR in the underlying CPGs are particularly suitable as a basis for the preparation or revision of DMPs. If GoR are lacking, a high LoE is taken.

5.4 Health care analysis

5.4.1 Background

A) Health care

Health care is defined as the medical and psychosocial care of sick people, as well as measures for prevention and health promotion offered by medical and non-medical providers of health care services. Medical care comprises diagnosis, treatment, nursing care, rehabilitation and after-care. The care provided offers all measures within the health care system that are directly or indirectly targeted towards improving or sustaining the health status (mortality, morbidity, and quality of life) of certain individuals or populations [20].

B) Health care standard

Medical standard is defined by medical practice that, according to medical and scientific evidence and/or clinical experience, is accepted in the profession [248]. This standard is referred to as the health care standard, which may be specified by laws, regulations and directives, or identified in CPGs (see also Section 5.2). The reference values of quality indicators can also be interpreted as health care standards [190].

C) Quality of health care

For the assessment of quality of care the actual health care situation referring to structures, processes and outcomes is compared with the particular health care standard specified through norms, directives and CPGs [249,297]. By comparing the target status with the actual status, conclusions on the current quality of health care become possible; in this connection the current health care situation represents the “actual status” and the current health care standard represents the “target status”, whereby the latter describes the goals to be achieved in health care, i.e. “ideal” health care. This analysis/evaluation is conducted for an area of health care defined by the research question of the commission. The precondition for determining the quality of health care is the availability of health care data that were systematically collected and analysed and that a corresponding health care standard can be determined. In this context, “systematic” is understood to be a planned data collection with uniform documentation instructions (coding instructions, e.g. International Classification of Diseases [ICD] version), standardized data collection forms, as well as a complete, and, if possible, comprehensive collection of data (depending on the research question).

5.4.2 Content aspects of a health care analysis

The health care analysis comprises the current and systematic description, analysis and assessment of health care aspects of a defined population group with regard to a specific medical or system-related research question (see §139a SGB V (3) Nos. 1 and 2). How detailed the analysis is depends on the type of commission.

The analysis usually examines the German health care situation, potentially supplemented by international comparison. The health care analysis allows the examination of complex interventions referring to both patient-relevant outcomes and outcomes related to the health care system. For the health care analysis, different individual medical as well as population and health system-related data and studies can be compiled in a modular system. In health sciences the term “individual medicine” is used for “classical” medicine involving the patient, in order to make a distinction from the term “population medicine”; the latter term is a component of public health.

The health care analysis can describe and assess different levels and/or several health care aspects. Basically one distinguishes between 2 areas: an epidemiological area and one comprising the social organization of health care. The first area describes the distribution and frequency of diseases in the population. If one examines a health care problem, this area is important for estimation of how many and what type of patients are affected and whether, in the attempt to solve the health care problem, certain subgroups need to be focused on, e.g. elderly people or socially disadvantaged persons. The second area addresses, for example, issues of health care-related structures and processes.

The health care analysis can examine different resources of the health care system (input), structures and processes (throughput), health care services (output), and/or results (outcomes) [439]. In order to assess the quality of health care, the health care situation is compared with a normative standard, the health care standard, insofar as such a standard exists.

5.4.3 Aims of a health care analysis

The superordinate aim of a health care analysis is to assess the quality of care.

The following points can be subgoals of the health care analysis:

- examination of the implementation of standards within health care and identification of possible potential for improvement
- investigation of the effects of health care models or measures of quality assurance on the population or on patient groups/population groups
- provision of (background) information for the development of quality indicators or for the prioritization of research questions
- presentation of references to a potential over-, under- or inappropriate provision of health care [472] and, if applicable, formulation of suggestions for improvement in terms of the optimized use of available resources
- identification of a potential need for research (e.g. clinical research, HTA, health care system research)

For feasibility reasons, the focus within the framework of a project is usually on one or a small number of the aims described above with regard to a certain disease.

5.4.4 Research question of a health care analysis

The precondition for the systematic description, investigation, and assessment of health care areas is the formulation of a specific research question. The definition of the research question comprises the specification of the following points:

- population (age; gender; disease; if relevant, subgroup or severity of disease)
- the interventions to be investigated (e.g. care of diabetic patients in general practice)
- outcome measures/patient-relevant outcomes (e.g. structural characteristics or health-related quality of life)
- health care setting (e.g. outpatient care, acute inpatient care or cross-sector care)

When formulating the research question it needs to be specified from which perspective (e.g. patients, society, cost carriers, etc.) health care is to be described and assessed, as the focus of the investigation and the selection of outcomes may change depending on the perspective. In this context, specific attention may be paid to the interests of vulnerable groups.

Regional variations (disparities), international comparisons, as well as temporal developments (trends) may also be addressed according to the research question.

5.4.5 Potential health care parameters

Different parameters can be used within the framework of a health care analysis. Health care parameters are, for example, epidemiological indices or indicators that help describe various areas of the health care system (see Table 7).

Table 7: Examples of potential health care parameters

Examples of potential health care parameters	
Indicators	**Health care parameters**
Incidence, prevalence, morbidity	Disease burden
Case fatality rate	Disease severity
Impairments and disabilities according to International Classification of Functioning (ICF), early retirements, mortality	Consequences of disease
Number of doctors per 1000 inhabitants, number of service providers per spatial unit, number of hospital beds per 1000 inhabitants etc.	Structure of the health care system (e.g. in Germany)
Utilization of services or service provision	Volume of services
Quality indicators for inpatient/outpatient sector, e.g. for patient safety, guideline-compliant care of patients	Quality of health care
E.g. neonatal and/or maternal mortality, vaccination rates, length of hospital stays	Structures, processes, and outcomes of health care in an international comparison

Epidemiological indices, for example, prevalence of a disease, can be drawn upon to obtain an overview of the extent of a health care problem. They provide information on the frequency of disease [346]. Disease severity can be estimated by means of the case fatality rate [256]. The consequences of a disease can be assessed by means of data according to the International Classification of Functioning (ICF) and pension fund data (e.g. on invalidity pensions) [133,526]. Health care studies, as well as data from cost carriers or service providers (health insurance funds, associations of SHI physicians, etc.), can identify patients' utilization of health care services. They thus provide information on how often such services are requested, provided or made use of. Quality indicators for the structural, process and outcome quality of inpatient and/or outpatient care may supplement the data pool. They serve quality assurance purposes and may indicate specific health care problems related to structural characteristics, process steps or individual outcomes. In addition, patient safety data from hospital quality reports and registries, as well as clinical and qualitative studies (as far as available), may be incorporated into a health care analysis. For example, they may disclose avoidable adverse events. Evaluation reports on model projects according to §63 SGB V may indicate potential new health care paths. At a system level, further parameters can be used to describe the health care situation and compared with international data. Examples are vaccination rates, disease-specific life expectancy, the number of hospital beds per 1000 inhabitants, and the proportion of expenditure on health care services in relation to the gross domestic product [320,325,587].

Depending on the research question, the above-mentioned parameters (and possibly others) can be combined and thus enable a comprehensive overview of individual health care areas. The health care standards allocated to these areas are identified as described in Section 5.4.8.

5.4.6 Procedure for a health care analysis

An example of a procedure for a health care analysis is presented in Figure 15.

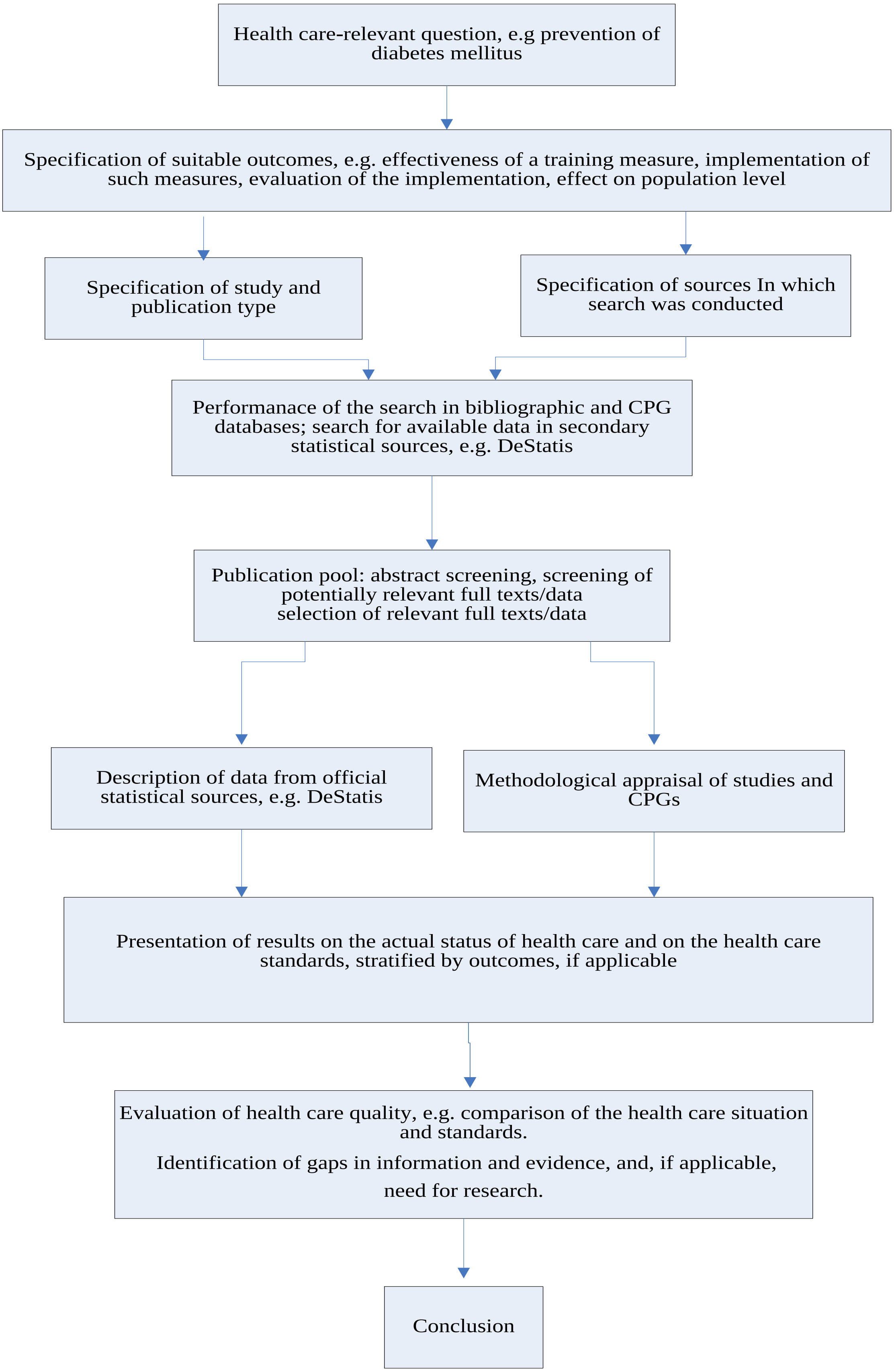

Health care-relevant question, e.g prevention of diabetes mellitus
Specification of suitable outcomes, e.g. effectiveness of a training measure, implementation of such measures, evaluation of the implementation, effect on population level
Specification of study and publication type
Specification of sources In which search was conducted
Performanace of the search in bibliographic and CPG databases; search for available data in secondary statistical sources, e.g. DeStatis
Publication pool: abstract screening, screening of potentially relevant full texts/data selection of relevant full texts/data
Description of data from official statistical sources, e.g. DeStatis
Methodological appraisal of studies and CPGs
Presentation of results on the actual status of health care and on the health care standards, stratified by outcomes, if applicable
Evaluation of health care quality, e.g. comparison of the health care situation and standards.
Identification of gaps in information and evidence, and, if applicable, need for research.
Conclusion

Figure 15: Example of a procedure for a health care analysis

5.4.7 Levels of a health care analysis

Health care can be described by means of the above-mentioned parameters relating to 3 different levels: that of individual medicine, of population medicine and of the health care system.

The first level refers to individual patients or patient groups in a clinical setting. Typical outcome measures on this level are patient-relevant outcomes such as mortality, morbidity and health-related quality of life.

The second level refers to population-based studies in the sense of evidence-based health care (population medicine) [260,346]. Outcome measures on this level are also patient-relevant outcomes such as mortality and morbidity, referring, however, to the general population [346]. Beyond this further outcome measures can be investigated, for example, rates of or reasons for participation in screening or vaccination programmes or in health care models such as DMPs.

The third level is the health care system [21,125]. Outcome measures on this level can be the utilization of health care services or the provision of services in different settings (inpatient/outpatient care) or by different professions/providers.

After a description of areas in German health care an international comparison may be meaningful. Depending on the research question, the description of health care in a modular system can refer to information from all 3 levels (individual medicine, population medicine, health care system). In addition, at all levels, temporal developments and regional variations (disparities) can be investigated [131]; for this purpose, geographic information systems can be used, amongst other things.

5.4.8 Methodological features of a health care analysis

With regard to the complexity of the health care system and the above-mentioned levels (see Sections 5.4.5 and 5.4.7), different study and publication types may be considered within the framework of a health care analysis.

In addition, it may be necessary to examine different research questions on health care with different quantitative and qualitative methods (pluralism of methods). Moreover, data from several sources are drawn upon (see Section 5.4.9) and processed with different methods. As far as possible, the methodological assessment is performed with suitable instruments (see Section 5.4.9).

In addition, the consideration of sociocultural and ethical aspects may be necessary in the assessment of quality of health care in certain groups of patients, for example, access to health care.

5.4.9 Information retrieval

Depending on the research question, different sources may be searched. The search is developed according to the requirements of the source. Both the literature search and the search for CPGs are conducted according to the Institute's *General Methods* (see Section 7.1).

A) Determination of the health care standard

The type of health care standard is inferred from the research question for the health care analysis. The first preference is to identify health care standards via evidence-based CPGs. The systematic approach to identify health care standards via CPGs is described in Section 5.2. Laws, regulations and directives define the legally binding framework of health care/medical care.

Structures and processes are mostly assessed by means of quality indicators. High-quality CPGs designate quality indicators, among other things. These refer to measures that indirectly represent the quality of health care. They can be applied to the quality of structures, processes and outcomes. The reference range of the quality indicator specifies the health care goal, i.e. the standard. An indicator always only refers to one health care area, therefore it is meaningful to combine several indicators in order to assess quality [10]. Table 8 provides an overview of potential sources for identifying health care standards.

Table 8: Information sources for identifying German health care standards

Information on	**Examples of data providers**
Health care/medical standards (clinical practice guidelines)	Association of the Scientific Medical Professional Societies (AWMF) Guidelines International Network (G-I-N) National Guideline Clearinghouse (NGC)
Laws (Social Code Book, SGB) and regulations	Federal Ministry of Justice and Consumer Protection (BMJV) Federal Ministry of Health (BMG)
Directives	Federal Joint Committee (G-BA German Medical Association (BÄK)
Indicators for the quality of structures, processes and outcomes	National Association of Statutory Health Insurance Physicians (KBV), e.g. Ambulatory Quality Indicators and Key Measures (AQUIK) Federal Office for Quality Assurance (BQS) Institute for Applied Quality Promotion and Research in Health Care (AQUA)

AQUA: Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen; AQUIK: Ambulante Qualitätsindikatoren und Kennzahlen; AWMF: Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften; BÄK: Bundesärztekammer; BMJ: Bundesministerium der Justiz und für Verbraucherschutz; BMG: Bundesministerium für Gesundheit; BQS: Institut für Qualität und Patientensicherheit (previously Bundesgeschäftsstelle Qualitätssicherung); G-BA: Gemeinsamer Bundesausschuss; KBV: Kassenärztliche Bundesvereinigung; SGB: Sozialgesetzbuch

B) Identification of data sources for health care data

All relevant data sources for the particular research question should be identified and, as far as possible, used to describe the provision of health care. Following the general principles of topic-related information retrieval (see Section 7.1), selection of data sources is specified in the report plan and is binding (sources: e.g. bibliographic databases, databases of organizations providing official statistics, morbidity registries, handsearch in selected professional journals, contacts with experts, patient organizations, and, if applicable, industry). Potential data sources for identifying health care data are named below (see Table 9).

Table 9: Data sources for identifying health care data

Information on	**Examples of sources**
Morbidity and mortality, e.g. incidence and prevalence rates (population level)	▪ Health report of federal and state organizations (e.g. child and youth health survey of the Robert Koch Institute) ▪ Report of the Federal Statistical Office (e.g. hospital discharge diagnoses, statistics on causes of death). ▪ Morbidity registries (e.g. epidemiological cancer registries) ▪ Routine data, e.g. of health care funds and Associations of Statutory Health Insurance Physicians
Health care needs (e.g. regional needs analyses)	▪ Health care studies
Utilization and prescription behaviour	▪ Drug prescription report (Research Institute of the Local Health Care Fund, WidO) ▪ Hospital report (WidO) ▪ Remedy report (WidO) ▪ ICD-10 key codes according to specialty groups (Central Institute for Health Care provided by Statutory Health Insurance Physicians) ▪ Routine data, e.g. of health care funds or Associations of Statutory Health Insurance Physicians
Patient safety	▪ Arbitration boards of the Regional Medical Associations ▪ Quality indicators of the OECD ▪ Further publications of the statutory health insurance funds
Measurement of health care quality with indicators	
▪ Quality of health care at a system level	▪ OECD (e.g. access to health care)
▪ Quality of outpatient medical care	▪ Quality reports of the Associations of Statutory Health Insurance Physicians
▪ Quality of inpatient care	▪ Hospital quality reports according to §137 ▪ Publications of the Federal Office for Quality Assurance (BQS)/Institute for Applied Quality Promotion and Research in Health Care (AQUA)
▪ Quality of nursing care	▪ Nursing care reports of the Medical Review Board of the Statutory Health Insurance Funds (MDK)
▪ DMP	▪ Evaluation reports of DMPs
Health care system/Comparison of systems	▪ e.g. WHO publications (e.g. World Health Report)

AQUA: Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen; BQS: Institut für Qualität und Patientensicherheit (previously Bundesgeschäftsstelle Qualitätssicherung); DMP: disease management programme; ICD: International Classification of Diseases; MDK: Medizinischer Dienst der Krankenversicherung; OECD: Organization for Economic Co-operation and Development; WidO: Wissenschaftliches Institut der Allgemeinen Ortskrankenkasse; WHO: World Health Organization

C) Search procedure for health care data

The search procedure follows the usual approach in the Institute. Regarding the search for CPGs and the bibliographic search, this is described in a different section (see Section 7.1).

Health care data, e.g. from official statistical sources or morbidity registries, are specifically searched for. The search and search result are documented. For relevant data published exclusively on the Internet, the search strategy to be used is determined by the structure and options of the relevant websites.

Depending upon the specific research question, different data sources/study types are drawn upon to describe health care.

5.4.10 Assessment of the data identified

The following aspects need to be considered in the assessment of the data identified:

- Assessment of the study and publication quality of the studies included: Studies are assessed following the Institute's *General Methods*. Supplementations, e.g. regarding evaluation studies or qualitative studies [213], need to be justified.
- Assessment of studies with constructs as endpoints: For patient-relevant outcomes that are constructs, e.g. health-related quality of life, the validity of the survey instrument is assessed. Non-validated instruments are not suitable for comparison.
- Assessment of official statistics: Such data, e.g. from DeStatis, are not assessed methodologically, as it is often impossible to assess such statistics (e.g. on mortality) in this way. In addition, they are already subjected to strict quality criteria by the issuing organizations [77,431]. Publication of these data through third parties, e.g. in journal articles, are assessed according to the Institute's *General Methods*.
- Assessment of registry data: If endpoints are presented by means of registry data, the validity of the registry should be addressed (data quality [i.e. completeness and plausibility of data sets], completeness, currentness) [431,563].
- Conclusive assessment of study and publication quality: The evaluation of the potential risk of bias in the studies/publications to be assessed is conducted following the Institute's *General Methods*.
- Assessment of the methodological quality of the CPGs: see Section 5.2.3.

5.4.11 Information synthesis and analysis

The synthesis and analysis of information is conducted as follows: At first the available literature is checked for relevant information on the outcomes specified in the report plan, assessed according to the Institute's *General Methods*, and then described. The results are subsequently summarized. On the basis of the results of the health care analysis an assessment of health care quality is conducted.

5.4.12 Description and assessment of health care quality

The assessment of health care quality comprises 3 steps:

Step 1: description of the actual status

The actual status of health care is described as specified in the report plan. In this context, the following questions need to be considered:

- Are data, studies or publications available on the goals or health care aspects of the investigation?
- How reliable are the results found?

Step 2: description of the target status

In a further step, health care standards are identified and described. Here too, the availability and the methodological quality of standards are checked.

- Does a health care standard exist for the goals/health care aspects stated in the report plan?
- How reliable are the results found?

Step 3: comparison between actual and target status

Then the actual health care status is compared to the health care standards. Taking the following questions into account, this leads to the assessment of health care quality:

- Is the health care standard implemented in everyday health care?
- How great are the deviations between the actual and the target status? In which direction does the actual status deviate from the target status (over- or underprovision of health care)?
- What conclusions can be drawn from the above comparison?

A final evaluation is made in the conclusions of the report. The evaluation enables us to judge whether gaps in information and/or evidence exist, whether there is a need for research, and/or whether potential for improvement exists.

5.5 Validity of clinical practice guideline recommendations

5.5.1 Background

Even though the methodology for CPG development is increasingly being further developed [18,233]. CPGs still differ considerably in their development process, presentation, methodological quality, and not least in their content [83,84,92,269,281,384,388,392]. In addition, high methodological quality of CPGs does not necessarily correlate with the content quality of the recommendations included in them [582]. For many research questions the examination of the methodological quality of a CPG is therefore insufficient to assess the value of individual recommendations [233]. This results in the necessity to analyse and review the contents of CPGs, particularly with regard to the validity of individual recommendations.

5.5.2 Definitions of internal and external validity

One distinguishes between the “internal” and “external” validity of CPG recommendations. These are defined as follows:

Internal validity: ensures the minimization of potential bias in the development of CPG recommendations.

External validity: describes the applicability of a recommendation under the conditions of the health care setting described in the CPG. This can comprise both the clinical setting as well as the use of a CPG at a system level.

The appraisal of the internal validity of CPG recommendations is understood to be the appraisal of the handling of the literature underlying the recommendation, as well as the appraisal of the consensus process. The appraisal of external validity involves consideration of context aspects (e.g. availability, patient preferences, and ethical aspects) or costs in the generation and formulation of the recommendation.

External validity is distinguished from the term “transferability”, which describes to what extent a recommendation is transferable to a different context. This can refer to transferability between different health care systems, as well as within a system (e.g. different setting, different target groups of patients).

5.5.3 Aim of the analysis and appraisal of clinical practice guideline recommendations

The aim of the methodological approach is to appraise the internal validity of CPG recommendations. Aspects of external validity are only considered if they are helpful in the appraisal of internal validity. For example, the designated context for the CPGs or the user target group may need to be taken into account when assessing the underlying evidence. Closer examination of external validity going beyond this is not conducted.

5.5.4 Approach to the analysis and appraisal of internal validity

The appraisal of the internal validity of individual CPG recommendations is conducted by means of

1) Identification and documentation of potentially biasing factors that might impair the internal validity of a CPG recommendation. Such factors are identified and documented at the level of the characteristics and structure of the CPG or its recommendation, the evidence base of the recommendation, and the consensus process.
2) Identification of the need for adjustment of the CPG recommendation: This results from the potential risk of bias identified under Point 1. In this context, depending on the severity of the determined deficiencies a distinction must be made between a potential and mandatory need for adjustment. A recommendation is classified as “not valid” if a mandatory need for adjustment is identified for this recommendation.

5.5.5 Potential research questions

The methods for analysis and appraisal of the internal validity of CPG recommendations are applicable to various research questions. Potential research questions are

- appraisal of individual recommendations from several CPGs on a disease or on a health care aspect comprising several interventions: e.g. “Appraisal of the internal validity of CPG recommendations from evidence-based CPGs on preoperative diagnostics”
- appraisal of recommendations from several CPGs on an intervention: e.g. “Appraisal of the internal validity of CPG recommendations from evidence-based CPGs on the treatment of diabetes mellitus type 2 with long-acting insulin analogues”
- appraisal of recommendations from a specific CPG

Moreover, the analysis and appraisal of the internal validity of CPG recommendations can also contribute to the appraisal of evidence-based CPGs for diseases with the greatest epidemiological relevance (see §139a SGB V (3) No. 3).

6 Evidence-based health information for consumers

6.1 Background and goals

The Institute has a legislative responsibility to provide health information to consumers, but not direct advice to individuals. Its goal is to improve health and patient autonomy by providing health information that aims to advance health and scientific literacy [35,104,143,322,336]. The health information should therefore

- facilitate active and informed decision-making about health issues
- promote the critical use of health care services
- improve understanding of physical, mental and emotional health
- improve understanding of medical and scientific information, including the concept of evidence-based medicine, and
- enable support of patients by family and friends

To achieve these goals, the Institute needs to be a reliable, trusted and patient-centred information provider. The integration of patient values in decision-making is core to the concept of evidence-based medicine [475], and thus to evidence-based health information. In addition to descriptions of benefits and harms, evidence-based health information must also include facts on the uncertainties and unknowns associated with health interventions [113,139,169,482,527]. Further requirements are that:

- the content is based on clear scientific evidence, particularly systematic reviews
- the information is developed according to systematic methods which aim to minimise bias and maintain neutrality
- evidence-based communication techniques are used to meet the goals of informing, supporting and empowering users
- uncertainties as well as the potential for benefit and harm of medical interventions are described in a manner that is readily understandable
- language and framing are neutral and non-directive, so that people can make their decisions in accordance with their own values
- the information is kept up-to-date so that it remains evidence-based

The Institute’s primary medium for communication is the Internet.

6.2 Patient-centred communication

6.2.1 Communication standards

A key challenge for evidence-based health information is to communicate in a way that is widely understandable while remaining scientifically accurate and objective. Additionally, the

Institute's health information is aimed at a heterogeneous target group in terms of health literacy, individual health, education, personal background, age, and sex, and the information should be readily understandable for all of them.

The primary means of quality assessment for understandability are reviews of drafts by groups of test readers that reflect the differences mentioned above, as well as reader ratings (see below).

Explaining evidence and remaining objective in communicating health-related information raise additional challenges [170,322,498,556]. To be objective and non-directive, the health information should provide a balanced description of what is considered to be scientifically known. No recommendations are made. This is achieved by preparing the information in a non-directive way, using neutral language.

Drawing on the evidence that has accumulated on communicating research findings, the Institute aims to

- present information in consistent formats to aid understanding, supplemented with additional formats to enhance the understandability of medical words and numerical information
- communicate the degree of uncertainty associated with evidence
- indicate to which groups of people the evidence is applicable, in order to enhance transferability
- distinguish very clearly and carefully between "absence of evidence" and "evidence of no effect"
- describe risks as absolute risks, supplementing them with other information such as relative risk where suitable, if there are data that are reliable and relevant for decision-making, and
- avoid biasing information in favour of or to the detriment of the products of any particular company by using generic names of products whenever possible and only using brand names when it is essential for understanding

There is evidence from screenings that the presentation of absolute risk estimates can help provide orientation for patients in their individual decision-making process [163]. If there are reliable data available that can be helpful for individual risk assessment, this information is presented. One method to help people individually weigh benefit and harm is offering decision aids. Although uncertainties remain about formal decision aids [420], especially for individual use on the Internet [98,167], the Institute may develop decision aids for some topics, particularly by drawing on the experience of specific decision aids which have been shown to be effective in randomized trials. Decision aids are developed using the International Patient Decision Aid Standards (IPDAS) for orientation [167,270].

Giving people information is not the only purpose of health communication. An important role of health information is to provide users with emotional support [186], and it can also play a role in enhancing patient autonomy. Health communication needs to be patient-centred if it is to be empowering and emotionally supportive. According to the definition of the World Health Organization (WHO), empowerment in health includes the ability to make choices and take actions in accordance with your own goals [419]. These abilities enable consumers to think and act autonomously. Empowering health communication addresses what consumers want to know, shows interest and respect for what patients think and respects their competence [145,321,575]. As well as seeking to be understandable, objective and accurate in its information, the Institute aims to

- demonstrate sensitivity and respect for user knowledge, values and concerns, autonomy, cultural differences as well as gender, age and disability-related interests
- maintain a patient-centred, non-judgmental, non-directive and neutral style of language
- respect readers' time

A neutral style of language should ensure that information reaches both women and men and that both genders feel addressed in the same way. Continuously referring to people in the masculine form (generic masculine) leads to a mental underrepresentation of women, which is to be seen as linguistic discrimination of women [305]. The information published at Gesundheitsinformation.de/Informed Health Online uses a gender-neutral style of language, which does not use generic masculine forms whenever possible. Both genders are explicitly named if both are meant, or gender-neutral expressions are used.

On the basis of available evidence and the experience of other groups, a health information style guide was developed for editorial staff to aid in text production, to increase awareness and for other purposes [408]. This guide undergoes continuous development based on the evaluation of our information and emerging evidence in the field of evidence-based communication.

6.2.2 Method of multidimensional patient pathways

Patient-centred health information is oriented towards users' questions, takes patients' experience and views into account and acknowledges their competency. Patient-centred health information not only aims at answering medical-scientific questions and making an informed decision possible, but also at offering emotional support. To do this, it is necessary on the one hand to know the questions users might be interested in. On the other hand, it is important for the authors to approach the patients' or their family members' perspectives and to develop an understanding of what it means to be living with a certain illness. To do this, the Institute uses a method that traces the possible paths patients with a certain illness can follow. This method is called "method of multidimensional pathways" (in German: "Methode der Patientenwege") in the following.

Multidimensional patient pathways summarize and illustrate the different social, emotional, cognitive and clinical dimensions that can be associated with an illness. The method follows the medical-sociological “illness trajectory” model [110] and “patient career” model [212,340] as well as different models of “patient’s journey” [345].

Medical sociology started early to look into the effects of illnesses on the patients’ lives. In this context the term “patient career” (in German: “Patientenkarriere”) was coined in Germany. Some of the contributions to be mentioned here are the ones made by Goffman, Gerhardt and Dörner [212,340]. Another approach is the “illness trajectory” as described by Corbin and Strauss [110].

The method of multidimensional patient pathways is derived from these different approaches. Even though patients can become experts at living with a certain illness, and therefore do, in a certain sense, pursue a kind of “career”, the term “multidimensional patient pathways” (in German: “Patientenwege”) is preferred. This term comprises as many potential patient pathways as possible. Multidimensional patient pathways are developed with the aim of determining the different possible paths for a particular medical condition and the different challenges and decisions the patient will face.

One of the aims of developing multidimensional patient pathways is to set the framework for the contents of the Institute’s health information. The following questions can be used to help determine this framework:

- Who might read the information?
- What content-related questions might the readers have?
- What might be the emotional state of the reader?
- Which information might be used at what point during the course of the disease?
- What decisions are patients faced with, and when will they have to make these decisions?
- What effects might health information on this topic have?

This method mainly aims to help the authors of the Institute’s health information systematically develop a good understanding of patients and their relatives as well as of their interaction with information. The orientation towards the dimensions given in Table 10 supports this aim.

Table 10: Different dimensions of a patient pathway

(Everyday) life	Effects of the disease on social relations and roles: family and relationship, job, quality of life, "performance", etc.
Doing/coping	Anything that is done with regards to the illness, such as visiting a doctor, taking medications, looking for information, self-help
Feeling	Feelings that come up during the course of the disease and the treatment, such as grief, fears, worries, etc.
Knowledge	What do consumers know already? What information might they need?
Decisions	What decisions must the person affected make in each phase?
Clinic	Description of the medical phases, such as risk factors, symptoms, diagnosis, treatment, rehabilitation
Contact point in the health care system	Who in the health care or social welfare system can be contacted in each phase, for example, doctors, nurses, physiotherapists, psychotherapists, social workers, counselling centres, insurance funds?

Multidimensional patient pathways can be mapped for the more comprehensive products. This can help to systematically identify the effects an illness may have on the lives of patients. The method should be comprehensible and reproducible, and reflect everyday health care. The sources used include rapid reviews of qualitative studies, personal accounts from the healthtalk.org database [144], literature on factors influencing adherence, literature on patients' informational needs systematic reviews on communication and information on health care issues.

(Clinical) care pathways can help identify important diagnostic and therapeutic steps and other "milestones" on a patient's pathway. Care pathways are multidisciplinary care and treatment plans. They describe how basic diagnostic and therapeutic steps in typical patients with a certain diagnosis or illness are optimally coordinated and organized. They aim to connect evidence and practice, and to detect patients' expectations and preferences in order to eventually facilitate optimal care [403,459].

6.2.3 Consumer involvement

There is some evidence that involving the people affected by a particular health problem in the development of health information can increase its relevance [415]. One of the requirements of evidence-based health information is its orientation towards the consumers' perspective and their informational needs [139]. This is a key element when producing health information [603]. The different measures implemented here include the following: When prioritising and narrowing down a selected topic, ideas proposed by website users, experiences from consultations with self-help groups and results of online polls on www.gesundheitsinformation.de/www.informedhealthonline.org are taken into account (see

Section 6.3.1). Additionally, the individual stories of patients, as well as those close to them, are presented in order to enable patients and other interested people to find out about the different aspects of living with a condition and nursing care. This is intended to complement the other health information (see Section 6.4.3). As part of the quality assurance, patients or patient representatives are asked to review certain text drafts. Another procedure to include the public is the limited commenting procedure. The Board of Trustees and others are given the opportunity to comment on drafts. The Board of Trustees also includes representatives of patients' interests and representatives of self-help organizations of chronically ill and disabled people. All our health information usually also undergoes external user testing when it is submitted for comments. In user testing, a group of people affected by the given condition or disease or potential users comment on the texts regarding their content and understandability. In addition, users of the website www.gesundheitsinformation.de/ www.informedhealthonline.org can contact the publisher with their feedback. They can comment online on the individual pieces of health information. There is also an ongoing user survey on the website (see Section 6.2.4).

6.2.4 Visual communication and multi-media

Text alone may not be as understandable and memorable as information where explanations are supported by images [170,274,333,362,556]. Spoken text may also enhance understanding [274,483]. Explanations where text, images and sound are combined may be the most understandable form of communication, especially for people of lower literacy [274]. Where appropriate, texts are supported with visuals and sound to enhance the effectiveness of its information and reach a wider audience. These include anatomical diagrams and short animated films on key topics that combine visuals, text and sound (videos). Graphics and pictograms also help many people understand numerical data and other aspects of scientific evidence [162,362,498]. Visual and multi-media elements should not replace text, but enhance the material covered in the texts. This ensures that the information is also accessible to people who are visually or hearing impaired.

The Internet enables health information to be presented in multi-media formats. As the technology of the Internet constantly improves and access to the Internet is no longer limited only to computers, communicating effectively with vision and sound on websites is becoming increasingly feasible for more users. The Internet also enables interactivity with users, so that communication need not flow only towards them. Showing an interest in what is important to patients is a critical element in patient-centred and empowering communication [145,321,575]. While the Institute cannot provide individual health advice, there are nevertheless multiple ways in which the Institute offers its users the opportunity to share their views and concerns, including

- the opportunity to leave comments on individual articles
- topic suggestions and a general online contact form
- an ongoing survey of the website's usability, and

- occasional online polls on specific health topics [326]

6.2.5 Accessibility

Various factors can restrict accessibility to the Internet and its use, including the following:

- disabilities, particularly (but not only) visual and hearing impairment
- poor reading skills
- insufficient computer skills
- technological capacity (affecting speed and access to multi-media)
- language (user's native language)

It is ensured that the website meets internationally accepted disability accessibility criteria [583], and the German Barrier-Free Information Technology Regulation (BITV[31]) [78]. The usability of the website is also continuously evaluated and optimized.

Publishing press releases helps health information content reach people who do not use the Internet to look for information about health topics through other media.

The health information is published in both German and English. The best possible quality assurance requires broad international involvement. Publication of health information in English makes feedback from international researchers, and in particular systematic review authors, possible. The availability of an English version also broadens possibilities for translation into further languages.

6.3 Topic selection, gathering of information and evaluation of evidence

6.3.1 Topic selection

The Institute's health information is produced

- in response to direct commissions received from the G-BA or Federal Ministry of Health
- to summarize other products published by the Institute and as accompanying information for these products
- to fulfil its legislative responsibility to provide consumers with health information, as well as on its own initiative within the framework of the G-BA's general commission

Health information is potentially limitless in scope, and informing everyone about everything is not possible. As with other health care priority-setting decisions, deciding on priorities for health information involves simultaneous analysis of multiple sources of information [29,30].

[31] Barrierefreie Informationstechnik-Verordnung

Section 139a, Paragraph 3, Sentence 6 of the German Social Code Book V (SGB V) sets the following task for the Institute: “provision of easily-understandable general information to citizens on the quality and efficiency of health care services, as well as on the diagnosis and therapy of diseases of substantial epidemiological relevance”. The Institute’s general commission was amended in July 2006. According to this, it must “continuously monitor and evaluate fundamentally important developments in medicine” and report on these. This general commission was adapted to the Institute’s health information in 2008 [210].

It is not possible to arrive at a broadly acceptable definition or a clear list of “diseases of substantial epidemiological relevance”. The epidemiological relevance of a disease with practical impact can only be determined using factors for which burden of disease data can be identified. Epidemiologically relevant factors could include the following:

- mortality
- prevalence/incidence
- frequency of utilization of health care services
- treatment costs
- absence from work due to illness
- impairment of quality of life and other consequences that have a significant impact on the lives of those affected by the condition

When deciding on which diagnoses or group of diseases to cover, those that affect at least one percent of the public at any given time (prevalence) or within a given year (incidence) are prioritised. Where justified, this catalogue can be expanded, for example in response to commissions received by IQWiG. When compiling the catalogue of topics, the primary source is the health care supply report (*Versorgungsreport*) produced by the AOK[32] Research Institute (WIdO), which is published periodically at www.wido.de/vsreport.html. The report is updated annually and contains information on prevalence and hospitalization rates for the 1500 most common illnesses (grouped according to ICD-10), based on about 24 million members of the statutory health insurance fund “AOK”. The topic catalogue for health information is reviewed regularly and subsequently amended as needed. The current status of implementation can be viewed on the website.

When prioritizing topics, a number of additional sources can be used to find out what consumers, healthy or ill, would like to know:

- surveys, primary qualitative research and reviews of qualitative research on people’s information needs

[32] Allgemeine Ortskrankenkasse (Local Healthcare Fund)

- enquiries made to call centres of the SHI funds
- experiences of other information providers, patient advice services and self-help groups
- enquiries made to the Federal Government Commissioner for Patients' Affairs [493]
- topics that are entered into the search engine of the IQWiG website www.gesundheitsinformation.de/www.informedhealthonline.org as well as other data concerning Internet use
- topics suggested by the website users
- results of the Institute's own online polls about information needs and interests

Further scientific, editorial, and user-oriented aspects are also considered. An overview is provided in Table 11.

Table 11: Possible aspects to consider when prioritising topics

Evidence	Editorial considerations	Patient/user interest
Systematic reviews of the benefit of health-related interventions	Balanced range of topics	Patient/user interest
Evidence on the effect of information on the topic	Current relevance of topic	Information searched for by users
	Possible adverse effects of the health information	Topic arousing interest in the reader/user
	Priorities of contracting agencies	Unfulfilled information needs
	Workload and resources	Burden of disease
		Information needs from an expert's point of view

6.3.2 Gathering of information for health information

The health information is predominantly based on systematic reviews and qualitative research. When researching a topic in depth, the following information is generally gathered to help identify issues of interest and concern for patients and consumers:

- rapid appraisals of primary qualitative studies as well as reviews of qualitative studies (see Section 8.4)
- reviews of the effects of communication
- reviews of adherence studies
- freely accessible patient information on the Internet as well as self-help group websites

The Internet and other sources are also searched to identify the interventions being used by, or offered to, consumers.

The results of this primary assessment of patients' and information needs help form a picture of the different stages a person affected with a certain health problem has to go through, the associated psychological and emotional problems that can occur, and the points at which decisions need to be made. For specific topics, patient representatives can also be interviewed to identify further issues and discuss the relevance of research findings in Germany.

The scientific basis for health information is usually a topic-specific systematic search for systematic reviews on the benefits and harms of medical interventions such as diagnostic measures, including screening or treatment options with or without medication. Research includes but is not limited to the Database of Abstracts of Reviews of Effects (other reviews) (DARE), the Cochrane Database of Systematic Reviews (Cochrane reviews), the Health Technology Assessment Database (technology assessments), as well as MEDLINE. Systematic reviews with searches performed during the last 3 years are the preferred source of information [509,510]. Depending on the topic, the search period can be extended.

If necessary, a search is also done for information on the causes, course, prognosis, and epidemiology of the medical condition. This usually covers the entire disease, with an internal scoping exercise conducted later to focus on the areas that the health information will cover.

In exceptional cases and primarily as part of the updating process, a search for primary studies is considered. If there is no recent review regarding an important topic, an update search for primary studies may be done. Update searches for studies to examine how current a review is are usually done in the Cochrane Controlled Trials Register, MEDLINE, and EMBASE [254]. Other databases can also be consulted.

6.3.3 Evaluation of evidence

The health information is based mainly on systematic reviews (see Section 8.2). Systematic reviews on the effects of an intervention need to fulfil certain minimum requirements for them to be used for health information: they are only allowed to have few methodological flaws according to the Oxman and Guyatt Index [309,428,430]. To provide the basis for a statement on the benefits and harms of a treatment, the review should include at least two studies judged to be of adequate quality by the review's authors, and include data on at least one patient-relevant outcome. The relevance and applicability of the evidence are also taken into consideration, particularly in terms of gender and age (see Section 8.4).

When more than one systematic review of adequate methodological quality addresses a particular subject or outcome, a further quality assessment is carried out. The aim is to determine whether there are qualitative differences between the different reviews or whether individual reviews are less suitable. The aspects compared include the following:

- main content of the review, especially as regards its relevance for patient information
- quality, depth and date of search
- sensitivity analyses and handling of heterogeneity
- management of any bias potential

The results of the highest-quality review for a particular topic are taken as the source of the numerical data used in the health information. Where reviews come to contradictory conclusions, the possible reasons are explored [310].

For issues concerning aspects like the aetiology or prognosis of a condition, or qualitative descriptions of patients' experiences, other types of primary studies are suitable for inclusion in systematic reviews [216]. When assessing such systematic reviews, criteria of the Oxford Centre for Evidence-Based Medicine and the McMaster University evidence-rating system are used [96,254]. The methods for assessing qualitative research are described in Section 8.4.

6.3.4 Updating

A critical part of evidence-based health information is making sure that its conclusions are not out-of-date. Regular updating is one of the quality criteria determined by the EU for health-related websites [108] and which the German position paper *Gute Praxis Gesundheitsinformation* (Good Practice Health Information) [139] describes. Evidence is growing exponentially. This is the case for both trials [34,559] and systematic reviews [34,399]. New evidence can render existing reviews obsolete or out-of-date [203,480,510,578], although new evidence often leads to no change or a strengthening of the original conclusions [301,443,538].

A study of guideline recommendations concluded that after 3 years, over 90% of recommendations may still be current, while after 6 years, about 50% of the recommendations in guidelines may be obsolete [509]. For some topics, for example where the evidence is very strong, the half-life of evidence can be much longer, and in other areas it can be less than 3 years [510]. However, as evidence continues its exponential growth, the half-life of information is likely to shorten: that is, information will become out-of-date more quickly. The Institute sees 3 years as the usual time after which its information requires review. On the basis of this period, a deadline is specified for the update of the health information. In addition, the following sources are checked regularly during evidence scanning: Cochrane Database of Systematic Reviews (Cochrane Reviews), McMaster Online Rating of Evidence (MORE) and PubMed. German, European, and U.S. regulatory agencies are monitored for health warnings as well. The content of regularly updated and evidence-based information for medical professionals is also considered, including Clinical Evidence and EBM Guidelines. If a systematic review, study, or announcement found through evidence scanning is identified as being relevant, its effect on the need for an update to health information on related topics is ascertained and evaluated. This evaluation may confirm the initial schedule or trigger a revision to the scheduled update.

6.4 Information products

6.4.1 Main types of articles

The website gesundheitsinformation.de/informedhealthonline.org focuses primarily on the presentation of topics relating to health and illness. One topic may be comprised of different types of articles. These different formats are intended to cover the most important aspects of a topic and answer central questions users may have. They are also intended to meet the different information needs of different audiences.

The main types of articles include the following:

- **Overview**: The overview introduces the topic and provides fundamental facts and links to the types of articles (described below) that further explore the topic. The overview has a fixed structure.
- **Learn More**: This provides further information on more specific aspects of the topic, such as treatment options with or without medication or certain diagnostic tests. If possible, a “Learn More” will also describe the advantages and disadvantages of individual treatment options or, if there is not enough evidence, the resulting uncertainties. A “Learn More” may also describe life with an illness, and consider both the perspective of those directly affected by the illness and their families and people close to them.
- **Research Summaries**: These articles are objective summaries of the current state of knowledge about the question posed in the title. As a rule, they are based on the results of good-quality, systematic evidence syntheses. They provide in-depth descriptions of the studies and explain how the answer to the research question was found.

These various articles, together with the supplementary items presented below in 6.4.2, constitute an evidence-based health encyclopaedia.

Section 2.1.7 describes how health information is produced. Information on health research should have a similar level of quality assurance as the research report itself [460]. Quality assurance for content is provided by external medical experts, and in some cases by patients as well, depending on the topic. In a limited commenting procedure, drafts are given to the Institute’s Board of Trustees, among others. This ensures that the patient representatives in this body will also have the opportunity to comment on these drafts. In addition, external user testing is done simultaneously. The patients whose interviews have served as the basis of the real-life stories are also invited to comment on the patient information drafts that correspond to their stories (see Section 6.4.3).

6.4.2 Supplementary items

In addition to the main types of articles, various supplementary items can be produced. These aim to make the key messages more understandable and interesting. For example, the inclusion of images, sound and animated films may increase the understandability of the website, especially for people with lower literacy levels (see Section 6.2.4). The supplementary items include the following:

- real-life stories by persons affected, see Section 6.4.3 for more information
- illustrations, photos and other images
- animated films with text and sound
- quizzes
- glossary of medical and scientific terms
- “in brief” – general articles explaining anatomy, bodily functions, treatment and diagnostic measures, as well as the principles and methods of evidence-based medicine
- calculators

The goals of these supplementary items are the following:

- promote general understanding of health and medical issues
- help users to understand and weigh up the potential benefits and harms of medical interventions
- facilitate self-management strategies

As a rule, interactive items are also tested for usability by external users. Accessibility is a particular focus.

6.4.3 Real-life stories

Patients may trust health websites more if they include the experiences of people affected by the respective condition [512].

Many patients would like to hear or read about the experiences of people affected by the same health condition as them [257,539]. Real-life stories are commonly used to impart information in both the fields of journalism and patient information. They represent one means of conveying scientific evidence and making it accessible to the general public [217]. The importance of real-life stories in medical practice and in health care is increasingly recognized [223,528,601].

Real-life stories provide the following functions [539]:

- They offer the opportunity to compare people’s own experiences with those of others.
- Reading about the feelings of others might “allow” acceptance of similar emotions.
- They can show people who are affected that they are not alone with their experiences.

By presenting the individual stories of patients as well as those close to them, the Institute would like to enable patients and other interested people to find out about the different aspects of living with a condition and nursing care. This is intended as a complementary source of health information, in addition to the other products. The content of the real-life stories should not contradict the evidence-based health information.

One example of real-life stories associated with evidence-based health information is the multimedia website healthtalk.org [144], which is available free of charge on the Internet [257,258,601]. The page's content originates from the Database of Personal Experience of Health and Illness (DIPEx) and the Health Experiences Research Group (HERG) at Oxford University.

The methods for gathering, editing and publishing real-life stories are based on DIPEx's established approach.

Real-life stories are prepared using the following process:

1) Interview partners are found, most often via self-help organisations.
2) Informed consent is sought regarding the interview procedure and how the story will be used.
3) The interviews are carried out.
4) The interviews are documented and edited, and the interview partners give their informed consent regarding the publication of the final version.
5) The real-life story is published on the website with the permission of the interview partner.

Particular importance is placed on extensively briefing the interview partners before the interview, on the fact that they can withdraw their informed consent to publish the story at any time, on preparing the interviews well, on carrying out the interviews based on predefined criteria, as well as on the anonymity of the interviews. If possible, every feature article should be accompanied by at least 2 real-life stories.

6.4.4 Website

The primary dissemination vehicle for the health information is the bilingual website, www.gesundheitsinformation.de/www.informedhealthonline.org. High website standards are to be maintained in the following areas:

- usability and accessibility [283,339,414] (see Section 6.2.5)
- privacy and data protection [289]
- transparency
- search engine visibility [550]

- attractiveness to users
- user interactivity

The website also includes a free electronic newsletter, with the choice of bi-weekly or monthly subscription. The newsletter contains information on what is new on the website, including when information is updated. RSS feeds provide core information to enable individuals to subscribe by RSS. The contents of the website can also be automatically integrated into other websites.

The Institute’s website is certified by the “Health On the Net” (HON) Foundation, fulfilling the 8 requirements of the HON Code of Conduct (HONcode) for medical and health-related websites, and it complies with the principles of good practice in health information development set by the German *Gute Praxis Gesundheitsinformation*.

7 Information retrieval

Various types of information form the basis of the Institute's reports (e.g. results from scientific studies, CPGs, registry data and other data collections, documents from regulatory authorities, and dossiers from pharmaceutical companies). This chapter describes the process of a topic-related search for scientific literature.

In the following Section 7.1 the Institute's approach to conducting its own information retrieval is described. The approach to examining information retrieval conducted by others is presented in Section 7.2.

If data are submitted to the Institute that are not allowed to be published, then these data cannot be considered in the Institute's assessments, as this would contradict the principle of transparency.

Likewise, the unrequested submission of data, that is, outside commenting procedures or outside other existing regulations (e.g. queries to manufacturers), are not considered. The unrequested submission of study data bears the risk of selective submission and subsequent bias in the result of the benefit assessment.

7.1 Information retrieval conducted by the Institute itself

A systematic literature search aims to identify all publications **relevant** to the particular research question (i.e. publications that contribute to a gain in knowledge on the topic). The search for primary literature is normally orientated towards the aim of achieving high sensitivity.

If a benefit assessment is based on systematic reviews, completeness in terms of complete consideration of all available primary studies is not aimed for. If the completeness of the pool of primary studies used in a systematic review is in doubt and the robustness of results is no longer ensured, a benefit assessment is conducted on the basis of primary studies. In this context, robustness is understood to be sufficient certainty that the result would not be considerably changed by the inclusion of additional information or studies.

The following aspects have to be defined a priori in the systematic literature search:

- The inclusion criteria of the report plan or project outline with regard to
  - medical criteria (e.g. target population, intervention)
  - the study design or type of guideline
  - formal characteristics of the publication (e.g. abstract publication, language)
- The data sources to be included (e.g. bibliographic databases, guideline databases, handsearching in selected scientific journals, contacts with experts/industry/patient organizations)

Studies and examples on this topic are provided by numerous publications [271,272,398,491,522]. The relevance of the above criteria varies, depending on the different research questions. The type of product to be prepared (e.g. report, rapid report, working paper) and the resulting timeframe also have an impact on the approach to information retrieval.

7.1.1 Search procedure

The search in bibliographic databases, trial registries as well as guideline databases and websites of guideline providers consists of the following steps:

1) if necessary, specification of the research question posed
2) modification of the research question to a searchable research question
3) formulation of a search concept
4) selection of databases
5) identification of search terms
6) formulation of the search strategies
7) quality assurance (in the case of a bibliographic search)
8) performance of the search
9) storage of the search results in text files and import into a reference management software programme (if a standardized export is possible)
10) documentation of the search

Relevant publications identified in the preliminary search are usually drawn upon to identify search terms and formulate the search strategy for bibliographic databases. As a quality assurance step, it is tested whether the search strategy developed in this way identifies known relevant primary publications (test set) with sufficient certainty. The test set is generated by using previous publications by other working groups (systematic reviews on the topic of interest). In addition, a formal internal quality assurance is performed taking the review by Sampson into account [478,479].

7.1.2 Bibliographic databases

A) Search for primary literature

The selection of databases for each product is generally based on the focus (i.e. regarding content, methods, and region) of the bibliographic databases. At least 2 large biomedical databases (e.g. MEDLINE and EMBASE) are always selected. For the preparation of health information a search for qualitative studies is additionally conducted in CINAHL and PsycInfo.

B) Search for systematic reviews

In the search for systematic reviews, some different sources from those used in the search for primary literature need consideration. As a rule, databases are searched that exclusively or largely contain systematic reviews. In addition, a selection of biomedical databases also (but not primarily) containing systematic reviews is searched (e.g. MEDLINE and EMBASE).

Depending on the topic investigated, it is decided what databases or other sources (e.g. websites of individual HTA agencies) are also relevant and should be searched. HTAs that are not free of charge are considered in exceptional cases, if it is assumed, for example, that additional relevant information can be retrieved from them, or if no information is otherwise available.

7.1.3 Search in trial registries

The systematic search should identify not only published but also unpublished studies. In this context, “unpublished” means that the studies (or individual data sets) have not been published (or only partly) in scientific journals. Study publications are generally identified by means of a search in bibliographic databases such as MEDLINE or EMBASE. Trial registries can be drawn upon in the search for unpublished studies or data [352].

As a rule, the Institute’s benefit assessments involve a search in large general trial registries, as well as meta-registries thereof. In particular these include the trial registry ClinicalTrials.gov of the US National Institute of Health (NIH), the WHO’s meta-registry “International Clinical Trials Registry Platform Search Portal” (ICTRP), as well as the EU Clinical Trials Register of the European Medicines Agency (EMA). In the benefit assessment of drugs, two German meta-registries (“Pharmnet.BUND Klinische Prüfung” and “Arzneimittel-Informationssystem”, AMIS) as well as trial registries of the pharmaceutical industry (individual companies and meta-registries) are additionally screened. Searches in disease-specific trial registries are only performed in exceptional cases.

In addition to information on the existence of a study, some registries are also increasingly including study results. This applies, for example, to ClinicalTrials.gov and trial registries of the pharmaceutical industry. Providing the study in question is in principle relevant to the assessment, results from trial registries can be considered in the Institute’s reports.

7.1.4 Clinical practice guideline databases and providers

If the aim of the search is to identify CPGs, it is primarily conducted in guideline databases (e.g. Guidelines International Network [G-I-N], the AWMF, or the National Guideline Clearinghouse [NGC]), and may be followed by a search on the websites of providers of specialist and multi-disciplinary guidelines. Whether a supplementary search for guidelines is performed in bibliographic databases depends on the type of report to be prepared.

For the search in guideline databases or websites of guideline providers, the search strategy to be applied is targeted towards the structure and options of the particular websites. Only a few websites allow a search with key words, so that generally the complete list of a website's published guidelines is screened. In addition, for the search in guideline databases or websites of guideline providers, a standardized export is often not possible. For this reason, the search and number of hits are documented in a standardized search protocol. The potentially relevant hits are documented in a literature management programme. Otherwise, the procedure is followed as described in Section 7.1.1.

Within a benefit assessment, guidelines are not categorically excluded as a source of information. However, a systematic search for guidelines is not usually conducted.

7.1.5 Requests to manufacturers

Within the framework of the Institute's benefit assessments, the manufacturers of the technologies to be assessed are usually asked to provide previously unpublished information. The aim of this request is to identify all studies and other information relevant to the benefit assessment, independent of their publication status. For drug assessments this request is usually made in 2 steps. In the first, the Institute asks the manufacturer to supply a complete overview of all studies conducted by the manufacturer on the drug to be assessed. If appropriate, the Institute defines the project-specific inclusion criteria for this overview. In the second, the Institute identifies studies relevant to the benefit assessment from the overview, and requests detailed information on these studies. This may refer to a request for unpublished studies, or for supplementary, previously unpublished information on published studies. Previously unpublished information considered in the benefit assessment will also be published in the Institute's reports in order to ensure transparency. The basis for the incorporation of previously unpublished information into the benefit assessment is the conclusion of an agreement on the transfer and publication of study information. This agreement is made between the Institute and the manufacturer involved before the submission of data (see sample contract [293]). It specifies the procedure, the requirements for the documents to be submitted, as well as their confidential and non-confidential components. If the manufacturer concerned does not agree to this contract and therefore does not agree in particular to the complete transfer of all information requested by the Institute, or does not completely transfer the information requested despite conclusion of the agreement, no further requests to the manufacturer will be made. This is to prevent biased results due to selective provision of information.

7.1.6 Other data sources for the search

A) Proceedings of abstracts and selected scientific journals

Besides bibliographical database searches, it can be useful (depending on the research question) to conduct a handsearch in selected scientific journals and proceedings of abstracts from scientific meetings. This is decided on a case-by-case basis.

B) Publicly accessible documents from regulatory authorities

In the case of drug assessments, but also of assessments of specific (non-drug) medicinal products, publicly accessible drug approval databases or correspondence with regulatory authorities are further potential sources of information.

C) Information from authors of individual publications

Within the framework of guideline appraisals or benefit assessments it may be meaningful in individual cases to contact authors of publications or guidelines. For example, the requests may refer to specific details on individual guidelines or to unpublished information on journal publications.

D) Documents transferred by the G-BA or Federal Ministry of Health

If documents are provided by the contracting agency (G-BA or Federal Ministry of Health), they are regarded as a component of information retrieval. In the subsequent procedure, these documents are handled following the other principles of the search for and assessment of information.

7.1.7 Selection of relevant publications

Due to the primarily sensitive approach, the literature search in bibliographic databases results in a large number of citations that are not relevant to the assessment. The selection of relevant publications is made in several steps:

- Exclusion of definitely irrelevant publications (i.e. publications not fulfilling the inclusion or exclusion criteria of the report plan or project outline) through perusal of the titles, and, if available, the abstracts. This step can be divided into 2 in order to distinguish completely irrelevant publications from topic-related ones which, however, do not fulfil the inclusion or exclusion criteria. “Topic-related” refers, for example, to studies investigating the topic of interest but with a different study design or duration from that specified in the report plan or project outline.
- The full texts of the remaining potentially relevant publications are obtained. The decision on the inclusion of the study in the assessment concerned is then made on the basis of these documents.
- Depending on the research question, a supplementary third step is performed for the search in (clinical practice) guideline databases and on websites of guideline providers, where it is examined whether a methodological approach was adopted in the development and formulation of the guideline. This usually refers to the evidence base of the guideline (see Section 5.2). When preparing the report plan the Institute specifies a priori whether on the basis of the research question only evidence-based guidelines are to be considered in the particular report.

All selection steps are performed by 2 persons independently of each other. Discrepancies are resolved by discussion. In the first selection step, if doubts exist as to the relevance of a study, the corresponding full text is obtained and assessed. In this step, completely irrelevant publications may also be distinguished from topic-related ones.

The languages of publication are usually restricted to those of Western Europe. However, other foreign-language publications may also be included if the available information on these publications indicates that additional and relevant information for answering the research question is to be expected.

In the search for guidelines, the steps for the full-text screening (from the second screening onwards) are performed by 2 persons independently of each other. The quality assurance of the first screening step is conducted with the help of a standardized search protocol.

7.1.8 Documentation of the search

All steps in the search in bibliographic databases are documented. This especially includes

- the search strategy for the databases selected
- the search date
- the user interface
- the number of hits
- after perusal of all hits: documentation of the publications judged relevant to the research question posed (citations)
- after perusal of the full texts: documentation of the citations not judged relevant; alternatively, documentation of the topic-related publications that were, however, irrelevant for the report (in each case providing a reason for exclusion)

All other steps in the information retrieval procedure are also documented (e.g. correspondence with authors, queries to manufacturers).

7.1.9 Benefit assessments based on systematic reviews – supplementary search

In most cases a supplementary search for current primary studies is required; this search covers the period between the last date of the search conducted in the systematic review and the date of the search conducted by IQWiG in the preparation of its report. In benefit assessments based on systematic reviews, a supplementary update search for primary literature can only be dispensed with in justified exceptional cases. This applies, for example, if it is sufficiently certain that the result of the assessment would not be considerably changed by the inclusion of additional information or studies (robustness).

In addition, it may be necessary to conduct supplementary searches for primary literature for specific research questions not addressed in the systematic review.

7.2 Evaluation of the information retrieval

In its dossier assessments and assessments of potential, the Institute does not primarily conduct its own information retrieval; instead, the information retrieval presented in the dossiers or in the applications for testing is evaluated.

In the preparation of a dossier or application, a search in bibliographic databases and a search in publicly accessible trial registries must be conducted as a matter of principle by the applicant; the precise requirements are provided in the G-BA’s Code of Procedure [209,211].

The Institute conducts an evaluation of form and content of information retrieval for dossier assessments and assessments of potential. This refers to the search in bibliographic databases and trial registries and is based on the quality assurance procedures described in Section 7.1.1, as well as on the document templates for the preparation of dossiers and applications for testing included in the requirements of the G-BA’s Code of Procedure [209,211].

For dossier assessments, depending on the results arising from the evaluation of form and content of the dossiers, the Institute subsequently conducts its own search and/or study selection in order to evaluate the completeness of information retrieval. For this purpose, various strategies are available, for example, random checks of the literature citations included in the dossier, the application of specific functions of literature databases (e.g. “related articles” feature in PubMed [481,580], as well as the potential conduct of a complete literature search). The result of the evaluation of information retrieval in the dossier and the description of the approach in this regard are part of the dossier assessment.

8 Assessment of information

In research the term “bias” means a systematic deviation between research results and the “truth” [473]. For example, this may refer to an erroneously too high (or too low) estimation of a treatment effect.

A main objective in the benefit assessment of medical services is to estimate the actual effect of therapies and interventions as reliably and unbiasedly as possible. In order to minimize bias in the benefit assessment of medical services, different approaches are adopted internationally; these include using scientifically robust methods, ensuring wide participation in the relevant studies, as well as avoiding conflicts of interest [105]. All these methods also form the legal basis of the Institute’s work.

8.1 Quality assessment of individual studies

8.1.1 Criteria for study inclusion

The problem often arises that studies relevant to a benefit assessment do not completely fulfil the inclusion criteria for the patient population and/or the test and comparator intervention defined in the systematic review. In this case the Institute usually proceeds according to the following criteria:

For the inclusion criterion with regard to the study population, it suffices if at least 80% of the patients included in the study fulfil this criterion. Corresponding subgroup analyses are drawn upon if they are available in such studies. Studies in which the inclusion criterion for the study population is fulfilled in less than 80% of the patients included in the study are only included in the analysis if corresponding subgroup analyses are available, or if it has been demonstrated with sufficient plausibility or has been proven that the findings obtained from this study are applicable to the target population of the systematic review (see Section 3.3.1 for applicability).

Studies are also included in which at least 80% of patients fulfil the inclusion criterion regarding the test intervention (intervention group of the study) and at least 80% fulfil the inclusion criterion regarding the comparator intervention (comparator group of the study). If 1 of the 2 criteria is violated in a study, it is excluded from the benefit assessment.

8.1.2 Relationship between study type and research question

Only the most relevant study designs that play a role in benefit assessments in medical research (depending on the research question posed) are summarized here.

It is primarily the inclusion of a control group that is called for in the benefit assessment of interventions. In a design with dependent samples without a control group, proof of the effect of an intervention cannot usually be inferred from a pure “before-after” comparison. Exceptions include diseases with a deterministic (or practically deterministic) course (e.g. ketoacidotic diabetic coma; see Section 3.2.2). Randomization and blinding are quality criteria that increase the evidential value of controlled studies. Parallel group studies [442], cross-over studies [314], and cluster randomized studies [155] are common designs for clinical trials. If interim analyses are planned, the use of appropriate sequential designs must be considered [590].

Case reports or case series often provide initial information on a topic. These are susceptible to all kinds of bias, so that, depending on the research question, only limited reliable evidence can be inferred from this type of study. The prevalence of diseases can be estimated from population-based cross-sectional studies. Other fundamental and classical study types in epidemiology are case-control studies [59] to investigate the association between exposures and the occurrence of rare diseases, as well as cohort studies [60] to investigate the effect of an exposure over time. Cohort studies designed for this purpose are prospective, although retrospective cohort studies are also conducted in which past exposure is recorded (this type of study is frequently found in occupational or pharmacological epidemiology). In principle, prospective designs are preferable to retrospective designs. However, case-control studies, for example, are frequently the only feasible way of obtaining information on associations between exposures and rare diseases. Newer study designs in modern epidemiology contain elements of both case-control and cohort studies and can no longer be clearly classified as retrospective or prospective [317].

Diagnostic and screening studies may have very different aims, so that the assessment depends on the choice of an appropriate design (see Sections 3.5 and 3.6).

8.1.3 Ranking of different study types/evidence levels

Different approaches exist within the framework of systematic reviews or guideline development for allocating specific evidence levels to particular study types [237,242]. These levels can be used to create a ranking with regard to the validity of evidence from different study types. However, no system of evidence assessment currently exists that is generally accepted and universally applicable to all systematic reviews [318,588]. Due to the complexity of the appraisal of studies, no conclusive judgement on quality can be inferred from the hierarchy of evidence [24,599]. In general, the Institute follows the rough hierarchy of study types, which is widely accepted and is also largely consistent with the evidence classification of the G-BA [211], and has been incorporated in the regulation on the benefit assessment of drugs according to §35a SGB V [80]. At least for the evaluation of intervention effects, the highest evidence level is allocated to RCTs and systematic reviews of RCTs. In some classifications, individual RCTs are further graded into those of higher or lower quality (see Section 3.1.4).

However, at the latest in the classification of non-randomized studies with regard to their risk of bias, the study design alone can no longer provide sufficient orientation [234,261,576], even if the basis distinction between comparative and non-comparative studies seems meaningful. As described in Section 3.8, in the classification of non-randomized studies, besides other design aspects the Institute will primarily evaluate the control for potential confounders. However, this grading refers to the risk of bias (see Section 8.1.4) and not to the evidence level of the study.

8.1.4 Aspects of the assessment of the risk of bias

One main aspect of the interpretation of study results is the assessment of the risk of bias (see qualitative uncertainty of results, Section 3.1.4). In this context, the research question, the study type and design, and the conduct of the study play a role, as well as the availability of information. The risk of bias is substantially affected by the study quality; however, its assessment is not equivalent to the quality assessment of a study. For example, individual outcomes may also be considerably biased in a high-quality study. Other studies, however, may provide high certainty of results for specific outcomes in individual cases, despite being of low quality. As a rule, the Institute will therefore estimate the extent of the risk of bias in a problem-orientated manner for all relevant results (both for the study and the specific outcomes).

In principle, a recognized standardized concept should be followed in a study; from planning to conduct, data analysis, and reporting. This includes a study protocol describing all the important methods and procedures. For (randomized) clinical trials, the usual standards are defined by the basic principles of good clinical practice (GCP) [299,331]; for epidemiological studies, they are defined by guidelines and recommendations to ensure good epidemiological practice (GEP) [132]. In this context, a key criterion to avoid bias is whether the study was actually analysed in the way planned. This cannot usually be reliably concluded from the relevant publications. However, a section on sample size planning may at least provide indications in this regard. In addition, a comparison with the study protocol (possibly previously published) or with the corresponding publication on the study design is useful.

The following important documents were developed to improve the quality of publications:

- the CONSORT statement on RCTs [496] and the corresponding explanatory document [396]
- a proposal for an extension of the CONSORT statement for randomized studies on non-drug interventions [55] and the corresponding explanatory document [54]
- the CONSORT statement on cluster-randomized trials [93]
- the CONSORT statement on the documentation of adverse events [302]
- the CONSORT statement on non-inferiority and equivalence studies [441]
- the CONSORT statement on pragmatic studies [604]
- the CONSORT PRO extension for patient-reported outcomes [91]

- the PRISMA[33] statement on meta-analyses of randomized trials [397] and the corresponding explanatory document [357]
- the TREND[34] statement on non-randomized intervention trials [128]
- the STROBE[35] statement for observational studies in epidemiology [579] and the corresponding explanatory document [570]
- the MOOSE[36] checklist for meta-analysis of observational studies in epidemiology [534]
- the STARD statement on diagnostic studies [52] and the corresponding explanatory document [53]
- die ISOQOL[37] Reporting Standards for patient-reported outcomes [75]

If a publication fails to conform to these standards, this may be an indicator of an increased risk of bias of the results of the relevant study. Additional key publications on this issue describe fundamental aspects concerning the risk-of-bias assessment [165,236,264].

Key aspects of the Institute's risk-of-bias assessment of the results of RCTs comprise

- adequate concealment, i.e. the unforeseeability and concealment of allocation to groups (e.g. by external randomization in trials that cannot be blinded)
- blinded outcome assessment in trials where blinding of physicians and patients is not possible
- appropriate application of the "intention-to-treat" (ITT) principle

There must be a more cautious interpretation of the results of unblinded trials, or of trials where unblinding (possibly) occurred, compared with the interpretation of blinded studies. Randomization and the choice of appropriate outcome variables are important instruments to prevent bias in studies where a blinding of the intervention was not possible. In studies that cannot be blinded, it is crucial to ensure adequate concealment of the allocation of patients to the groups to be compared. It is also necessary that the outcome variable is independent of the (non-blinded) treating staff or assessed in a blinded manner independent of the treating staff (blinded assessment of outcomes). If a blinded assessment of outcome measures is not possible, a preferably objective outcome should be chosen which can be influenced as little as possible (with regard to its dimension and the stringency of its recording) by the (non-blinded) person assessing it.

[33] Preferred Reporting Items for Systematic Reviews and Meta-Analyses
[34] Transparent Reporting of Evaluations with Non-randomized Designs
[35] Strengthening the Reporting of Observational Studies in Epidemiology
[36] Meta-analysis of Observational Studies in Epidemiology
[37] International Society of Quality of Life Research

In the production of reports standardized assessment forms are generally used to assess the risk of bias of study results. As a rule, for controlled studies on the benefit assessment of interventions the following items across and specific to outcomes are considered in particular:

Items across outcomes:

- appropriate generation of a randomization sequence (in randomized studies)
- allocation concealment (in randomized studies)
- temporal parallelism of the intervention groups (in non-randomized studies)
- comparability of intervention groups and appropriate consideration of prognostically relevant factors (in non-randomized studies)
- blinding of patients and treating staff/staff responsible for follow-up treatment
- reporting of all relevant outcomes independent of results

Outcome-specific items:

- blinding of outcome assessors
- appropriate implementation of the ITT principle
- reporting of individual outcomes independent of results

On the basis of these aspects, in randomized studies the risk of bias is summarized and classified as “high” or “low”. A low risk of bias is present if it can be excluded with great probability that the results are relevantly biased. Relevant bias is understood to be a change in the basic message of the results if the bias were to be corrected.

In the assessment of an outcome, the risk of bias across outcomes is initially classified as “high” or “low”. If classified as “high”, the risk of bias for the outcome is also usually classified as “high”. Apart from that, the outcome-specific items are taken into account.

The classification as “high” of the risk of bias of the result for an outcome does not lead to exclusion from the benefit assessment. This classification rather serves the discussion of heterogeneous study results and affects the certainty of the conclusion.

No summarizing risk-of-bias assessment is usually performed for non-randomized comparative studies, as their results generally carry a high risk of bias due to the lack of randomization. The Institute specifically deviates from this procedure in assessments of the potential of new examination and treatment methods (see Section 3.8).

If a project of the Institute involves the assessment of older studies that do not satisfy current quality standards because they were planned and conducted at a time when these standards did not exist, then the Institute will present the disadvantages and deficiencies of these studies and discuss possible consequences. A different handling of these older studies compared with the handling of newer studies that have similar quality deficits is however only necessary if this is clearly justifiable from the research question posed or other circumstances of the assessment.

The assessment of formal criteria provides essential information on the risk of bias of the results of studies. However, the Institute always conducts a risk-of-bias assessment that goes beyond purely formal aspects in order, for example, to present errors and inconsistencies in publications, and to assess their relevance in the interpretation of results.

8.1.5 Interpretation of composite outcomes

A "composite outcome" comprises a group of events defined by the investigators (e.g. myocardial infarctions, strokes, cardiovascular deaths). In this context the individual events in this group often differ in their severity and relevance for patients and physicians (e.g. hospital admissions and cardiovascular deaths). Therefore, when interpreting composite outcomes one needs to be aware of the consequences thereby involved [111,189,202]. The following explanations describe the aspects to be considered in the interpretation of results. However, they specifically do not refer to a (possibly conclusive) assessment of benefit and harm by means of composite outcomes, if, for example, the potential harm from an intervention (e.g. increase in severe bleeding events) is included in an outcome together with the benefit (e.g. decrease in the rate of myocardial infarctions).

A precondition for consideration of a composite outcome is that the individual components of the composite outcome all represent patient-relevant outcomes defined in the report plan. In this context surrogate endpoints can be only included if they are specifically accepted by the Institute as valid (see Section 3.1.2). The results for every individual event included in a composite outcome should also be reported separately. The components should be of similar severity; this does not mean that they must be of identical relevance. For example, the outcome "mortality" can be combined with "myocardial infarction" or "stroke", but not with "silent myocardial infarction" or "hospital admission".

If a composite outcome fulfils the preconditions stated above, then the following aspects need to be considered in the interpretation of conclusions on benefit and harm:

- Does the effect of the intervention on the individual components of the composite outcome usually take the same direction?
- Was a relevant outcome suited to be included in the composite outcome not included, or excluded, without a comprehensible and acceptable justification?
- Was the composite outcome defined a priori or introduced post hoc?

Insofar as the available data and data structures allow, sensitivity analyses may be performed by comparing the exclusion versus the inclusion of individual components.

If the relevant preconditions are fulfilled, individual outcomes may be determined and calculated from a composite outcome within the framework of a benefit assessment.

8.1.6 Interpretation of subgroup analyses

In the methodological literature, subgroup analyses are a matter of controversy [22,429]. The interpretation of results of subgroup analyses at a study level is complicated mainly by 3 factors:

- No characteristic of proof: Subgroup analyses are rarely planned a priori and are rarely a component of the study protocol (or its amendments). If subgroup analyses with regard to more or less arbitrary subgroup-forming characteristics are conducted post hoc, the results cannot be regarded as a methodologically correct testing of a hypothesis.
- Multiple testing: If several subgroups are analysed, results in a subgroup may well reach statistical significance, despite actually being random.
- Lack of power: The sample size of a subgroup is often too small to enable the detection of moderate differences (by means of inferential statistics), so that even if effects actually exist, significant results cannot be expected. The situation is different if an adequate power for the subgroup analysis was already considered in the sample size calculation and a correspondingly larger sample size was planned [67].

The results of subgroup analyses should be considered in the assessment, taking the above 3 issues into account and not dominating the result of the primary analysis, even more so if the primary study objective was not achieved. An exception from this rule may apply if social law implications (see below) necessitate such analyses. Moreover, subgroup analyses are not interpretable if the subgroup-forming characteristic was defined after initiation of treatment (after randomization), e.g. in responder analyses. These aspects also play a role in the conduct and interpretation of subgroup analyses within the framework of meta-analyses (see Section 8.3.8).

The statistical demonstration of different effects between various subgroups should be conducted by means of an appropriate homogeneity or interaction test. The finding that a statistically significant effect was observed in one subgroup, but not in another, cannot be interpreted (by means of inferential statistics) as the existence of a subgroup effect.

Analyses of subgroups defined a priori represent the gold standard for subgroup analyses, where stratified randomization by means of subgroups and appropriate statistical methods for data analysis (homogeneity test, interaction test) are applied [114].

Despite the limitations specified above, for some research questions subgroup analyses may represent the best scientific evidence available in the foreseeable future in order to assess effects in subgroups [200], since factors such as ethical considerations may argue against the reproduction of findings of subgroup analyses in a validation study. Rothwell [458] presents an overview of reasons for conducting subgroup analyses. Sun et al. [536] identified criteria to assess the credibility of subgroup analyses.

Possible heterogeneity of an effect in different, clearly distinguishable patient populations is an important reason for conducting subgroup analyses [335,458]. If a-priori information is available on a possible effect modifier (e.g. age, pathology), it is in fact essential to investigate possible heterogeneity in advance with regard to the effect in the various patient groups. If such heterogeneity exists, then the estimated total effect across all patients cannot be interpreted meaningfully [335]. It is therefore important that information on a possible heterogeneity of patient groups is considered appropriately in the study design. It may even be necessary to conduct several studies [228]. Within the framework of systematic reviews, the analysis of heterogeneity between individual studies (and therefore, if applicable, subgroup analyses) is a scientific necessity (see Section 8.3.8), but also a necessity from the perspective of social law, as according to §139a (2) SGB V, the Institute is obliged to consider characteristics specific to age, gender, and life circumstances. In addition, according to the official rationale for the SHI Modernization Act[38], the Institute is to elaborate in which patient groups a new drug is expected to lead to a relevant improvement in treatment success, with the aim of providing these patients with access to this new drug [134]. A corresponding objective can also be found in §35a SGB V regarding the assessment of the benefit of drugs with new active ingredients [136]. In this assessment, patient groups should be identified in whom these drugs show a therapeutically relevant added benefit. According to social law, a further necessity for subgroup analyses may arise due to the approval status of drugs. On the one hand, this may be the consequence of the decision by regulatory authorities that, after balancing the efficacy and risks of a drug, may determine that it will only be approved for part of the patient population investigated in the approval studies. These considerations may also be based on subgroup analyses conducted post hoc. On the other hand, studies conducted after approval may include patient groups for whom the drug is not approved in Germany; the stronger approvals differ on an international level, the more this applies. In such cases, subgroup analyses reflecting the approval status of a drug may need to be used, independently of whether these analyses were planned a priori or conducted post hoc.

8.1.7 Assessment of data consistency

To assess the evidential value of study results, the Institute will review the consistency of data with regard to their plausibility and completeness. Implausible data are not only produced by incorrect reporting of results (typing, formatting, or calculation errors), but also by the insufficient or incorrect description of the methodology, or even by forged or invented data [9]. Inconsistencies may exist within a publication, and also between publications on the same study.

[38] GKV-Modernisierungsgesetz, GMG

One problem with many publications is the reporting of incomplete information in the methods and results sections. In particular, the reporting of lost-to-follow-up patients, withdrawals, etc., as well as the way these patients were considered in the analyses, are often not transparent.

It is therefore necessary to expose potential inconsistencies in the data. For this purpose, the Institute reviews, for example, calculation steps taken, and compares data presented in text, tables, and graphs. In practice, a common problem in survival-time analyses arises from inconsistencies between the data on lost-to-follow-up patients and those on patients at risk in the survival curve graphs. For certain outcomes (e.g. total mortality), the number of lost-to-follow-up patients can be calculated if the Kaplan-Meier estimates are compared with the patients at risk at a point in time before the minimum follow-up time. Statistical techniques may be useful in exposing forged and invented data [9].

If relevant inconsistencies are found in the reporting of results, the Institute's aim is to clarify these inconsistencies and/or obtain any missing information by contacting authors, for example, or requesting the complete clinical study report and further study documentation. However, it should be considered that firstly, enquiries to authors often remain unanswered, especially concerning older publications, and that secondly, authors' responses may produce further inconsistencies. In the individual case, a weighing-up of the effort involved and the benefit of such enquiries is therefore meaningful and necessary. If inconsistencies cannot be resolved, the potential impact of these inconsistencies on effect sizes (magnitude of bias), uncertainty of results (increase in error probability), and precision (width of the confidence intervals) will be assessed by the Institute. For this purpose, sensitivity analyses may be conducted. If it is possible that inconsistencies may have a relevant impact on the results, this will be stated and the results will be interpreted very cautiously.

8.2 Consideration of systematic reviews

Systematic reviews are publications that summarize and assess the results of primary studies in a systematic, reproducible, and transparent way. This also applies to HTA reports, which normally aim to answer a clinical and/or patient-relevant question. HTA reports also often seek to answer additional questions of interest to contracting agencies and health policy decision makers [156,353,435]. There is no need to differentiate between systematic reviews and HTA reports for the purposes of this section. Therefore, the term "systematic review" also includes HTA reports.

8.2.1 Classification of systematic reviews

Relying on individual scientific studies can be misleading. Looking at one or only a few studies in isolation from other similar studies on the same question can make treatments appear more or less useful than they actually are [1]. High quality systematic reviews aim to overcome this form of bias by identifying, assessing and summarizing the evidence systematically rather than selectively [156,165,216,435].

Systematic reviews identify, assess and summarize the evidence from one or several study types that can provide the best answer to a specific and clearly formulated question. Systematic and explicit methods are used to identify, select and critically assess the relevant studies for the question of interest. If studies are identified, these data are systematically extracted and analysed. Systematic reviews are non-experimental studies whose methodology must aim to minimize systematic errors (bias) on every level of the review process [1,165,264].

For systematic reviews of the effects of medical interventions, RCTs provide the most reliable answers. However, for other questions such as aetiology, prognosis or the qualitative description of patients' experiences, the appropriate evidence base for a systematic review will consist of other primary study types [216]. Systematic reviews of diagnostic and screening tests also show some methodological differences compared with reviews of treatment interventions [122].

In the production of the Institute's reports, systematic reviews are primarily used to identify potentially relevant (primary) studies. However, an IQWiG report can be based partially or even solely on systematic reviews (see Section 8.2.2). Health information produced by the Institute for patients and consumers is to a large part based on systematic reviews. This includes systematic reviews of treatments, and reviews addressing other questions such as aetiology, adverse effects and syntheses of qualitative research (see Section 6.3.3).

The minimal prerequisite for a systematic review on the effects of treatments to be used by the Institute is that it has only minimal methodological flaws according to the Oxman and Guyatt index [309,428,430] or the AMSTAR[39] instrument [505-507]. In addition to considering the strength of evidence investigated in systematic reviews, the Institute will also consider the relevance and applicability of the evidence. This includes investigating the question as to whether the results have been consistent among different populations and subgroups as well as in different healthcare contexts. The following factors are usually considered: the population of the participants in the included studies (including gender and baseline disease risk); the healthcare context (including the healthcare settings and the medical service providers); and the applicability and likely acceptance of the intervention in the form in which it was assessed [47,119].

8.2.2 Benefit assessment on the basis of systematic reviews

A benefit assessment on the basis of systematic reviews can provide a resource-saving and reliable evidence base for recommendations to the G-BA or the Federal Ministry of Health, provided that specific preconditions have been fulfilled [112,348]. In order to use systematic reviews in a benefit assessment these reviews must be of sufficiently high quality, that is, they must

[39] Assessment of Multiple Systematic Reviews

- show only a minimum risk of bias
- present the evidence base in a complete, transparent, and reproducible manner

and thus allow clear conclusions to be drawn [23,428,594]. In addition, it is an essential prerequisite that the searches conducted in the systematic reviews do not contradict the Institute's methodology and that it is possible to transfer the results to the research question of the Institute's report, taking the defined inclusion and exclusion criteria into account.

The methodology applied must provide sufficient certainty that a new benefit assessment based on primary literature would not reach different conclusions from one based on systematic reviews. For example, this is usually not the case if a relevant amount of previously unpublished data is to be expected.

A) Research questions

In principle, this method is suited for all research questions insofar as the criteria named above have been fulfilled. The following points should be given particular consideration in the development of the research question:

- definition of the population of interest
- definition of the test intervention and comparator intervention of interest
- definition of all relevant outcomes
- if appropriate, specification of the health care setting or region affected (e.g. Germany, Europe)

The research question defined in this way also forms the basis for the specification of the inclusion and exclusion criteria to be applied in the benefit assessment, and subsequently for the specification of the relevance of the content and methods of the publications identified. On the basis of the research question, it is also decided which type of primary study the systematic reviews must be based on. Depending on the research question, it is possible that questions concerning certain parts of a commission are answered by means of systematic reviews, whereas primary studies are considered for other parts.

B) Minimum number of relevant systematic reviews

All systematic reviews that are of sufficient quality and relevant to the topic are considered. In order to be able to assess the consistency of results, at least 2 high-quality publications (produced independently of each other) should as a rule be available as the foundation of a report based on systematic reviews. If only one high-quality publication is available and can be considered, then it is necessary to justify the conduct of an assessment based only on this one systematic review.

C) Quality assessment of publications, including minimum requirements

The assessment of the general quality of systematic reviews is performed with Oxman and Guyatt's validated quality index for systematic reviews [427,428,430] or with the AMSTAR Instrument [505-507]. According to Oxman and Guyatt's index, systematic reviews are regarded to be of sufficient quality if they have been awarded at least 5 of 7 possible points in the overall assessment, which is performed by 2 reviewers independently of one another. No such threshold is defined for the AMSTAR Instrument and therefore should, if appropriate, be defined beforehand. In addition, as a rule, the sponsors of systematic reviews, as well as authors' conflicts of interests, are documented and discussed. Depending on the requirements of the project, the particular index criteria can be supplemented by additional items (e.g. completeness of the search, search for unpublished studies, for example in registries, or additional aspects regarding systematic reviews of diagnostic studies).

D) Results

For each research question, the results of a benefit assessment based on systematic reviews are summarized in tables, where possible. If inconsistent results on the same outcome are evident in several publications, possible explanations for this heterogeneity are described [310].

If the compilation of systematic reviews on a topic indicates that a new benefit assessment on the basis of primary studies could produce different results, then such an assessment will be performed.

E) Conclusion/recommendations

Reports based on systematic reviews summarize the results of the underlying systematic reviews and, if necessary, they are supplemented by a summary of up-to-date primary studies (or primary studies on questions not covered by the systematic reviews). Independent conclusions are then drawn from these materials.

The recommendations made on the basis of systematic reviews are not founded on a summary of the recommendations or conclusions of the underlying systematic reviews. In HTA reports, they are often formulated against the background of the specific socio-political and economic setting of a particular health care system, and are therefore rarely transferable to the health care setting in Germany.

8.2.3 Consideration of published meta-analyses

Following international EBM standards, the Institute's assessments are normally based on a systematic search for relevant primary studies, which is specific to the research question posed. If it is indicated and possible, results from individual studies identified are summarized and evaluated by means of meta-analyses. However, the Institute usually has access only to aggregated data from primary studies, which are extracted from the corresponding publication or the clinical study report provided. Situations exist where meta-analyses conducted on the basis of individual patient data (IPD) from relevant studies have a higher value (see Section 8.3.8). This is especially the case if, in addition to the effect caused solely by the intervention, the evaluation of other factors possibly influencing the intervention effect is also of interest (interaction between intervention effect and covariables). In this context, meta-analyses including IPD generally provide greater certainty of results, i.e. more precise results not affected by ecological bias, when compared with meta-regressions based on aggregated data [514]. In individual cases, these analyses may lead to more precise conclusions, particularly if heterogeneous results exist that can possibly be ascribed to different patient characteristics. However, one can only assume a higher validity of meta-analyses based on IPD if such analyses are actually targeted towards the research question of the Institute's assessment and also show a high certainty of results. The prerequisite for the assessment of the certainty of results of such analyses is maximum transparency; this refers both to the planning and to the conduct of analyses. Generally valid aspects that are relevant for the conduct of meta-analyses are outlined, for example, in a document published by EMA [172]. In its benefit assessments, the Institute considers published meta-analyses based on IPD if they address (sub)questions in the Institute's reports that cannot be answered with sufficient certainty by meta-analyses based on aggregated data. In addition, high certainty of results for the particular analysis is required.

8.3 Specific statistical aspects

8.3.1 Description of effects and risks

The description of intervention or exposure effects needs to be clearly linked to an explicit outcome variable. Consideration of an alternative outcome variable also alters the description and size of a possible effect. The choice of an appropriate effect measure depends in principle on the measurement scale of the outcome variable in question. For continuous variables, effects can usually be described using mean values and differences in mean values (if appropriate, after appropriate weighting). For categorical outcome variables, the usual effect and risk measures of 2x2 tables apply [36]. Chapter 9 of the *Cochrane Handbook for Systematic Reviews of Interventions* [124] provides a well-structured summary of the advantages and disadvantages of typical effect measures. Agresti [6,7] describes the specific aspects to be considered for ordinal data.

It is essential to describe the degree of statistical uncertainty for every effect estimate. For this purpose, the calculation of the standard error and the presentation of a confidence interval are methods frequently applied. Whenever possible, the Institute will state appropriate confidence intervals for effect estimates, including information on whether one- or two-sided confidence limits apply, and on the confidence level chosen. In medical research, the two-sided 95% confidence level is typically applied; in some situations, 90% or 99% levels are used. Altman et al. [13] give an overview of the most common calculation methods for confidence intervals.

In order to comply with the confidence level, the application of exact methods for the interval estimation of effects and risks should be considered, depending on the particular data situation (e.g. very small samples) and the research question posed. Agresti [8] provides an up-to-date discussion on exact methods.

8.3.2 Evaluation of statistical significance

With the help of statistical significance tests it is possible to test hypotheses formulated a priori with control for type 1 error probability. The convention of speaking of a “statistically significant result” when the p-value is lower than the significance level of 0.05 ($p<0.05$) may often be meaningful. Depending on the research question posed and hypothesis formulated, a lower significance level may be required. Conversely, there are situations where a higher significance level is acceptable. The Institute will always explicitly justify such exceptions.

A range of aspects should be considered when interpreting p-values. It must be absolutely clear which research question and data situation the significance level refers to, and how the statistical hypothesis is formulated. In particular, it should be evident whether a one- or two-sided hypothesis applies [45] and whether the hypothesis tested is to be regarded as part of a multiple hypothesis testing problem [560]. Both aspects, whether a one- or two-sided hypothesis is to be formulated, and whether adjustments for multiple testing need to be made, are a matter of repeated controversy in scientific literature [185,327].

Regarding the hypothesis formulation, a two-sided test problem is traditionally assumed. Exceptions include non-inferiority studies. The formulation of a one-sided hypothesis problem is in principle always possible, but requires precise justification. In the case of a one-sided hypothesis formulation, the application of one-sided significance tests and the calculation of one-sided confidence limits are appropriate. For better comparability with two-sided statistical methods, some guidelines for clinical trials require that the typical significance level should be halved from 5% to 2.5% [298]. The Institute generally follows this approach. The Institute furthermore follows the central principle that the hypothesis formulation (one- or two-sided) and the significance level must be specified clearly a priori. In addition, the Institute will justify deviations from the usual specifications (one-sided instead of two-sided hypothesis formulation; significance level unequal to 5%, etc.) or consider the relevant explanations in the primary literature.

If the hypothesis investigated clearly forms part of a multiple hypothesis problem, appropriate adjustment for multiple testing is required if the type I error is to be controlled for the whole multiple hypothesis problem [40]. The problem of multiplicity cannot be solved completely in systematic reviews, but should at least be considered in the interpretation of results [37]. If meaningful and possible, the Institute will apply methods to adjust for multiple testing. In its benefit assessments (see Section 3.1). The Institute attempts to control type I errors separately for the conclusions on every single benefit outcome. A summarizing evaluation is not usually conducted in a quantitative manner, so that formal methods for adjustment for multiple testing cannot be applied here either.

The Institute does not evaluate a statistically non-significant finding as evidence of the absence of an effect (absence or equivalence) [12]. For the demonstration of equivalence, the Institute will apply appropriate methods for equivalence hypotheses.

In principle, Bayesian methods may be regarded as an alternative to statistical significance tests [523,524]. Depending on the research question posed, the Institute will, where necessary, also apply Bayesian methods (e.g. for indirect comparisons, see Section 8.3.9).

8.3.3 Evaluation of clinical relevance

The term "clinical relevance" refers to different concepts in the literature. On the one hand, at a group level, it may address the question as to whether a difference between 2 treatment alternatives for a patient-relevant outcome (e.g. serious adverse events) is large enough to recommend the general use of the better alternative. On the other hand, clinical relevance is understood to be the question as to whether a change (e.g. the observed difference of 1 point on a symptom scale) is relevant for individual patients. Insofar as the second concept leads to the inspection of group differences in the sense of a responder definition and corresponding responder analyses, both concepts are relevant for the Institute's assessments.

In general, the evaluation of the clinical relevance of group differences plays a particular role within the framework of systematic reviews and meta-analyses, as they often achieve the power to "statistically detect" the most minor effects [569]. In this context, in principle, the clinical relevance of an effect or risk cannot be derived from a p-value. Statistical significance is a statement of probability, which is not only influenced by the size of a possible effect but also by data variability and sample size. When interpreting the relevance of p-values, particularly the sample size of the underlying study needs to be taken into account [461]. In a small study, a very small p-value can only be expected if the effect is marked, whereas in a large study, highly significant results are not uncommon, even if the effect is extremely small [184,279]. Consequently, the clinical relevance of a study result can by no means be derived from a p-value.

Widely accepted methodological procedures for evaluating the clinical relevance of study results do not yet exist, regardless of which of the above-mentioned concepts are being addressed. For example, only a few guidelines contain information on the definition of relevant or irrelevant differences between groups [344,546]. Methodological manuals on the preparation of systematic reviews also generally provide no guidance or no clear guidance on the evaluation of clinical relevance at a system or individual level (e.g. the *Cochrane Handbook* [264]). However, various approaches exist for evaluating the clinical relevance of study results. For example, the observed difference (effect estimate and the corresponding confidence interval) can be assessed solely on the basis of medical expertise without using predefined thresholds. Alternatively, it can be required as a formal relevance criterion that the confidence interval must lie above a certain "irrelevance threshold" to exclude a clearly irrelevant effect with sufficient certainty. This then corresponds to the application of a statistical test with a shifting of the null hypothesis in order to statistically demonstrate clinically relevant effects [597]. A further proposal plans to evaluate relevance solely on the basis of the effect estimate (compared to a “relevance threshold”), provided that there is a statistically significant difference between the intervention groups [323]. In contrast to the use of a statistical test with a shifting of the null hypothesis, the probability of a type 1 error cannot be controlled thorough the evaluation of relevance by means of the effect estimate. Moreover, this approach may be less efficient. Finally, a further option in the evaluation of relevance is to formulate a relevance criterion individually, e.g. in terms of a responder definition [324]. In this context there are also approaches in which the response criterion within a study differs between the investigated participants by defining individual therapy goals a priori [453].

Patient-relevant outcomes can also be recorded by means of (complex) scales. A prerequisite for the consideration of such outcomes is the use of validated or established instruments. In the assessment of patient-relevant outcomes that have been operationalized by using (complex) scales, in addition to evaluating the statistical significance of effects, it is particularly important to evaluate the relevance of the observed effects of the interventions under investigation. This is required because the complexity of the scales often makes a meaningful interpretation of minor differences difficult. It therefore concerns the issue as to whether the observed difference between 2 groups is at all tangible to patients. This evaluation of relevance can be made on the basis of differences in mean values as well as responder analyses [497]. A main problem in the evaluation of relevance is the fact that scale-specific relevance criteria are not defined or that appropriate analyses on the basis of such relevance criteria (e.g. responder analyses) are lacking [401]. Which approach can be chosen in the Institute’s assessments depends on the availability of data from the primary studies.

In order to do justice to characteristics specific to scales and therapeutic indications, the Institute as a rule uses the following hierarchy for the evaluation of relevance, the corresponding steps being determined by the presence of different relevance criteria.

1) If a justified irrelevance threshold for the group difference (mean difference) is available or deducible for the corresponding scale, this threshold is used for the evaluation of relevance. If the corresponding confidence interval for the observed effect lies completely above this irrelevance threshold, it is statistically ensured that the effect size does not lie within a range that is certainly irrelevant. The Institute judges this to be sufficient for demonstration of a relevant effect, as in this case the effects observed are normally realized clearly above the irrelevance threshold (and at least close to the relevance threshold). On the one hand, a validated or established irrelevance threshold is suitable for this criterion. On the other hand, an irrelevance threshold can be deduced from a validated, established or otherwise well-justified relevance threshold (e.g. from sample size estimations). One option is to determine the lower limit of the confidence interval as the irrelevance threshold; this threshold arises from a study sufficiently powered for the classical null hypothesis if the estimated effect corresponds exactly to the relevance threshold.

2) If scale-specific justified irrelevance criteria are not available or deducible, responder analyses may be considered. It is required here that a validated or established response criterion was used in these analyses (e.g. in terms of an individual minimally important difference [MID]) [449]. If a statistically significant difference is shown in such an analysis in the proportions of responders between groups, this is seen as demonstrating a relevant effect (unless specific reasons contradict this), as the responder definition already includes a threshold of relevance.
3) If neither scale-specific irrelevance thresholds nor responder analyses are available, a general statistical measure for evaluating relevance is drawn upon in the form of standardized mean differences (SMD expressed as Hedges' g). An irrelevance threshold of 0.2 is then used: If the confidence interval corresponding to the effect estimate lies completely above this irrelevance threshold, it is assumed that the effect size does not lie within a range that is certainly irrelevant. This is to ensure that the effect can be regarded at least as "small" with sufficient certainty [181].

8.3.4 Evaluation of subjective outcomes in open-label study designs

Various empirical studies have shown that in non-blinded RCTs investigating subjective outcomes, effects are biased on average in favour of the test intervention. These subjective outcomes include, for example, PROs, as well as outcomes for which the documentation and assessment strongly depend on the treating staff or outcome assessors. Wood et al. provide a summary of these studies [600]. According to this such results show a potential high risk of bias. A generally accepted approach to this problem within the framework of systematic reviews does not exist. In this situation the Institute will normally infer neither proof of benefit nor harm from statistically significant results.

One possibility to take the high risk of bias for subjective outcomes in open-label studies into account is the definition of an adjusted decision threshold. Only if the confidence interval of the group difference of interest shows a certain distance to the zero effect is the intervention effect regarded as so large that it cannot only be explained by bias. The usual procedure for applying an adjusted decision threshold is to test a shifted null hypothesis. This procedure has been applied for decades; among other things, it is required in the testing of equivalence and non-inferiority hypotheses [173]. The prospective determination of a specific threshold value is required in the application of adjusted decision thresholds. If applied, the Institute will justify the selection of a threshold value on a project-specific basis by means of empirical data from meta-epidemiological research [489,600].

8.3.5 Demonstration of a difference

Various aspects need to be considered in the empirical demonstration that certain groups differ with regard to a certain characteristic. It should first be noted that the "demonstration" (of a difference) should not be understood as "proof" in a mathematical sense. With the help of empirical study data, statements can only be made by allowing for certain probabilities of error. By applying statistical methods, these probabilities of error can, however, be specifically controlled and minimized in order to “statistically demonstrate” a hypothesis. A typical method for such a statistical demonstration in medical research is the application of significance tests. This level of argumentation should be distinguished from the evaluation of the clinical relevance of a difference. In practice, the combination of both arguments provides an adequate description of a difference based on empirical data.

When applying a significance test to demonstrate a difference, the research question should be specified a priori, and the outcome variable, the effect measure, and the statistical hypothesis formulation should also be specified on the basis of this question. It is necessary to calculate the sample size required before the start of the study, so that the study is large enough for a difference to be detected. In simple situations, in addition to the above information, a statement on the clinically relevant difference should be provided, as well as an estimate of the variability of the outcome measure. For more complex designs or research questions, further details are required (e.g. correlation structure, recruitment scheme, estimate of drop-out numbers, etc.) [46,130].

Finally, the reporting of results should include the following details: the significance level for a statement; a confidence interval for the effect measure chosen (calculated with appropriate methods); descriptive information on further effect measures to explain different aspects of the results; as well as a discussion on the clinical relevance of the results, which should be based on the evaluation of patient-relevant outcomes.

8.3.6 Demonstration of equivalence

One of the most common serious errors in the interpretation of medical data is to rate the non-significant result of a traditional significance test as evidence that the null hypothesis is true [12]. To demonstrate “equivalence”, methods to test equivalence hypotheses need to be applied [313]. In this context, it is important to understand that demonstrating exact “equivalence” (e.g. that the difference in mean values between 2 groups is exactly zero) is not possible by means of statistical methods. In practice, it is not demonstration of exact equivalence that is required, but rather demonstration of a difference between 2 groups that is “at most irrelevant”. To achieve this objective, it must, of course, first be defined what an irrelevant difference is, i.e. an equivalence range must be specified.

To draw meaningful conclusions on equivalence, the research question and the resulting outcome variable, effect measure, and statistical hypothesis formulation need to be specified a priori (similar to the demonstration of a difference). In addition, in equivalence studies the equivalence range must be clearly defined. This range can be two-sided, resulting in an equivalence interval, or one-sided in terms of an “at most irrelevant difference” or “at most irrelevant inferiority”. The latter is referred to as a “non-inferiority hypothesis” [115,298,455].

As in superiority studies, it is also necessary to calculate the required sample size in equivalence studies before the start of the study. The appropriate method depends on the precise hypothesis, as well as on the analytical method chosen [454].

Specifically developed methods should be applied to analyse data from equivalence studies. The confidence interval approach is a frequently used technique. If the confidence interval calculated lies completely within the equivalence range defined a priori, then this will be classified as the demonstration of equivalence. To maintain the level of $\alpha = 0.05$, it is sufficient to calculate a 90% confidence interval [313]. However, following the international approach, the Institute generally uses 95% confidence intervals.

Compared with superiority studies, equivalence studies show specific methodological problems. On the one hand, it is often difficult to provide meaningful definitions of equivalence ranges [344]; on the other hand, the usual study design criteria, such as randomization and blinding, no longer sufficiently protect from bias [502]. Even without knowledge of the treatment group, it is possible, for example, to shift the treatment differences to zero and hence in the direction of the desired alternative hypothesis. Moreover, the ITT principle should be applied carefully, as its inappropriate use may falsely indicate equivalence [313]. For this reason, particular caution is necessary in the evaluation of equivalence studies.

8.3.7 Adjustment principles and multi-factorial methods

Primarily in non-randomized studies, multi-factorial methods that enable confounder effects to be compensated play a key role [319]. Studies investigating several interventions are a further important field of application for these methods [387]. In the medical literature, the reporting of results obtained with multi-factorial methods is unfortunately often insufficient [38,404]. To be able to assess the quality of such an analysis, the description of essential aspects of the statistical model formation is necessary [245,462], as well as information on the quality of the model chosen (goodness of fit) [273]. The most relevant information for this purpose is usually

- a clear description and a-priori specification of the outcome variables and all potential explanatory variables
- information on the measurement scale and on the coding of all variables
- information on the selection of variables and on any interactions
- information on how the assumptions of the model were verified
- information on the goodness of fit of the model
- inclusion of a table with the most relevant results (parameter estimate, standard error, confidence interval) for all explanatory variables

Depending on the research question posed, this information is of varying relevance. If it concerns a good prediction of the outcome variable within the framework of a prognosis model, a high-quality model is more important than in a comparison of groups, where an adjustment for important confounders must be made.

Inadequate reporting of the results obtained with multi-factorial methods is especially critical if the (inadequately described) statistical modelling leads to a shift of effects to the "desired" range, which is not recognizable with mono-factorial methods. Detailed comments on the requirements for the use of multi-factorial methods can be found in various reviews and guidelines [27,39,319].

The Institute uses modern methods in its own regression analysis calculations [244]. In this context, results of multi-factorial models that were obtained from a selection process of variables should be interpreted with great caution. When choosing a model, if such selection processes cannot be avoided, a type of backward elimination will be used, as this procedure is preferable to the procedure of forward selection [244,535]. A well-informed and careful preselection of the candidate predictor variable is essential in this regard [126]. If required, modern methods such as the lasso technique will also be applied [552]. For the modelling of continuous covariates, the Institute will, if necessary, draw upon flexible modelling approaches (e.g. regression using fractional polynomials [463,488]) to enable the appropriate description of non-monotonous associations.

8.3.8 Meta-analyses

A) General comments

Terms used in the literature, such as "literature review", "systematic review", "meta-analysis", "pooled analysis", or "research synthesis", are often defined differently and not clearly distinguished [165]. The Institute uses the following terms and definitions:

- A "non-systematic review" is the assessment and reporting of study results on a defined topic, without a sufficiently systematic and reproducible method for identifying relevant research results on this topic. A quantitative summary of data from several studies is referred to as a "pooled analysis". Due to the lack of a systematic approach and the inherent subjective component, reviews and analyses not based on a systematic literature search are extremely prone to bias.
- A "systematic review" is based on a comprehensive, systematic approach and assessment of studies, which is applied to minimize potential sources of bias. A systematic review may, but does not necessarily have to, contain a quantitative summary of study results.
- A "meta-analysis" is a statistical summary of the results of several studies within the framework of a systematic review. In most cases this analysis is based on aggregated study data from publications. An overall effect is calculated from the effect sizes measured in individual studies, taking sample sizes and variances into account.
- More efficient analysis procedures are possible if IPD are available from the studies considered. An "IPD meta-analysis" is the analysis of data on the patient level within the framework of a general statistical model of fixed or random effects, in which the study is considered as an effect and not as an observation unit.

- The Institute sees a "prospective meta-analysis" as a statistical summary (planned a priori) of the results of several prospective studies that were jointly planned. However, if other studies are available on the particular research question, these must also be considered in the analysis in order to preserve the character of a systematic review.

The usual presentation of the results of a meta-analysis is made by means of forest plots in which the effect estimates of individual studies and the overall effect (including confidence intervals) are presented graphically [355]. On the one hand, models with a fixed effect are applied, which provide weighted mean values of the effect sizes (e.g. weighting by inversing the variance). On the other hand, random-effects models are frequently chosen in which an estimate of the variance between individual studies (heterogeneity) is considered. The question as to which model should be applied in which situation has long been a matter of controversy [168,503,574]. If information is available that the effects of the individual studies are homogeneous, a meta-analysis assuming a fixed effect is sufficient. However, such information will often not be available, so that in order to evaluate studies in their totality, an assumption of random effects is useful [504]. Moreover, it should be noted that the confidence intervals calculated from a fixed-effect model may show a substantially lower coverage probability with regard to the expected overall effect, even if minor heterogeneity exists when compared with confidence intervals from a random-effects model [64]. The Institute therefore primarily uses random-effects models and only switches to models with a fixed effect in well-founded exceptional cases. In this context, if the data situation is homogeneous, it should be noted that meta-analytical results from models with random and fixed effects at best show marginal differences. As described in the following text, the Institute will only perform a meta-analytical summary of strongly heterogeneous study results if the reasons for this heterogeneity are plausible and still justify such a summary.

B) Heterogeneity

Before a meta-analysis is conducted, it must first be considered whether the pooling of the studies investigated is in fact meaningful, as the studies must be comparable with regard to the research question posed. In addition, even in the case of comparability, the studies to be summarized will often show heterogeneous effects [266]. In this situation it is necessary to assess the heterogeneity of study results [215]. The existence of heterogeneity can be statistically tested; however, these tests usually show very low power. Consequently, it is recommended that a significance level between 0.1 and 0.2 is chosen for these tests [307,330]. However, it is also important to quantify the extent of heterogeneity. For this purpose, specific new statistical methods are available, such as the I^2 measure [265]. Studies exist for this measure that allow a rough classification of heterogeneity, for example, into the categories "might not be important" (0 to 40%), "moderate" (30 to 60%), "substantial" (50 to 90%) and "considerable" (75 to 100%) heterogeneity [124]. If the heterogeneity of the studies is too large, the statistical pooling of the study results may not be meaningful [124]. The specification as to when heterogeneity is "too large" depends on the context. A pooling of data is usually dispensed with if the heterogeneity test yields a p-value of less than 0.2. In this context, the location of the effects also plays a role. If the individual studies show a clear effect in the same direction, then pooling heterogeneous results by means of a random effects model can also lead to a conclusion on the benefit of an intervention. However, in this situation a positive conclusion on the benefit of an intervention may possibly be drawn without the quantitative pooling of data (see Section 3.1.4). In the other situations the Institute will not conduct a meta-analysis. However, not only statistical measures, but also reasons of content should be considered when making such a decision, which must be presented in a comprehensible way. In this context, the choice of the effect measure also plays a role. The choice of a certain measure may lead to great study heterogeneity, yet another measure may not. For binary data, relative effect measures are frequently more stable than absolute ones, as they do not depend so heavily on the baseline risk [205]. In such cases, the data analysis should be conducted with a relative effect measure, but for the descriptive presentation of data, absolute measures for the specific baseline risks may possibly be inferred from relative ones.

In the case of great heterogeneity of the studies, it is necessary to investigate potential causes. Factors that could explain the heterogeneity of effect sizes may possibly be detected by means of meta-regression [547,566]. In a meta-regression, the statistical association between the effect sizes of individual studies and the study characteristics is investigated, so that study characteristics can possibly be identified that explain the different effect sizes, i.e. the heterogeneity. However, when interpreting results, it is important that the limitations of such analyses are taken into account. Even if a meta-regression is based on randomized studies, only evidence of an observed association can be inferred from this analysis, not a causal relationship [547]. Meta-regressions that attempt to show an association between the different effect sizes and the average patient characteristics in individual studies are especially difficult to interpret. These analyses are subject to the same limitations as the results of ecological studies in epidemiology [224]. Due to the high risk of bias, which in analyses based on aggregate data cannot be balanced by adjustment, definite conclusions are only possible on the basis of IPD [438,514,547] (see also Section 8.2.3).

The Institute uses prediction intervals to display heterogeneity within the framework of a meta-analysis with random effects [230,262,451]. In contrast to the confidence interval, which quantifies the precision of an estimated effect, the 95% prediction interval covers the true effect of a single (new) study with a probability of 95%. In this context it is important to note that a prediction interval cannot be used to assess the statistical significance of an effect. The Institute follows the proposal by Guddat et al. [230] to insert the prediction interval – clearly distinguishable from the confidence interval – in the form of a rectangle in a forest plot. The use of meta-analyses with random effects and related prediction intervals in the event of very few studies (e.g. less than 5) is critically discussed in the literature, as potential heterogeneity can only be estimated very imprecisely [262]. The Institute generally presents prediction intervals in forest plots of meta-analyses with random effects if at least 4 studies are available and if the graphic display of heterogeneity is important. This is particularly the case if, due to great heterogeneity, no pooled effect is presented.

Prediction intervals are therefore particularly used in forest plots if no overall effect can be estimated and displayed due to great heterogeneity. In these heterogeneous situations, the prediction interval is a valuable aid in evaluating whether the study effects are in the same direction or not or whether in the former case these effects are moderately or clearly in the same direction (see Section 3.1.4).

C) Subgroup analyses within the framework of meta-analyses

In addition to the general aspects requiring consideration in the interpretation of subgroup analyses (see Section 8.1.6), there are specific aspects that play a role in subgroup analyses within the framework of meta-analyses. Whereas in general subgroup analyses conducted post hoc on a study level should be viewed critically, in a systematic review one still depends on the use of the results of such analyses on a study level if the review is supposed to investigate precisely these subgroups. In analogy to the approach of not pooling studies with too great heterogeneity by means of meta-analyses, results of subgroups should not be summarized to a common effect estimate if the subgroups differ too strongly from each other. Within the framework of meta-analyses, the Institute usually interprets the results of a heterogeneity or interaction test regarding important subgroups as follows: A significant result at the level of $\alpha = 0.05$ is classified as proof of different effects in the subgroups; a significant result at the level of $\alpha = 0.20$ is classified as an indication of different effects. If the data provide at least an indication of different effects in the subgroups, then the individual subgroup results are reported in addition to the overall effect. If the data provide proof of different effects in the subgroups, then the results for all subgroups are not pooled to a common effect estimate. In the case of more than 2 subgroups, pairwise statistical tests are conducted, if possible, to detect whether subgroup effects exist. Pairs that are not statistically significant at the level of $\alpha = 0.20$ are then summarized in a group. The results of the remaining groups are reported separately and separate conclusions on the benefit of the intervention for these groups are inferred [518].

D) Small number of events

A common problem of meta-analyses using binary data is the existence of so-called “zero cells”, i.e. cases where not a single event was observed in an intervention group of a study. the Institute follows the usual approach here; i.e. in the event of zero cells, the correction value of 0.5 is added to each cell frequency of the corresponding fourfold table [124]. This approach is appropriate as long as not too many zero cells occur. In the case of a low overall number of events, it may be necessary to use other methods. In the case of very rare events the Peto odds-ratio method can be applied; this does not require a correction term in the case of zero cells [56,124].

If studies do exist in which no event is observed in either study arm (so-called “double-zero studies”) then in practice these studies are often excluded from the meta-analytic calculation. This procedure should be avoided if too many double-zero studies exist. Several methods are available to avoid the exclusion of double-zero studies. The absolute risk difference may possibly be used as an effect measure which, especially in the case of very rare events, often does not lead to the heterogeneities that otherwise usually occur. A logistic regression with random effects represents an approach so far rarely applied in practice [562]. Newer methods such as exact methods [551] or the application of the arcsine difference [464] represent interesting alternatives, but have not yet been investigated sufficiently. Depending on the particular data situation, the Institute will select an appropriate method and, if applicable, examine the robustness of results by means of sensitivity analyses.

E) Meta-analyses of diagnostic studies

The results of studies on diagnostic accuracy can also be statistically pooled by means of meta-analytic techniques [140,306]. However, as explained in Section 3.5, studies investigating only diagnostic accuracy are mostly of subordinate relevance in the evaluation of diagnostic tests, so that meta-analyses of studies on diagnostic accuracy are likewise of limited relevance.

The same basic principles apply to a meta-analysis of studies on diagnostic accuracy as to meta-analyses of therapy studies [140,447]. Here too, it is necessary to conduct a systematic review of the literature, assess the methodological quality of the primary studies, conduct sensitivity analyses, and examine the potential influence of publication bias.

In practice, in most cases heterogeneity can be expected in meta-analyses of diagnostic studies; therefore it is usually advisable here to apply random-effects models [140]. Such a meta-analytical pooling of studies on diagnostic accuracy can be performed by means of separate models for sensitivity and specificity. However, if a summarizing receiver operating characteristic (ROC) curve and/or a two-dimensional estimate for sensitivity and specificity are of interest, newer bivariate meta-analyses with random effects show advantages [241,448]. These methods also enable consideration of explanatory variables [240]. Results are presented graphically either via the separate display of sensitivities and specificities in the form of modified forest plots or via a two-dimensional illustration of estimates for sensitivity and specificity. In analogy to the confidence and prediction intervals in meta-analyses of therapy studies, confidence and prediction regions can be presented in the ROC area in bivariate meta-analyses of diagnostic studies.

F) Cumulative meta-analyses

For some time it has been increasingly discussed whether, in the case of repeated updates of systematic reviews, one should calculate and present meta-analyses included in these reviews as cumulative meta-analyses with correction for multiple testing [49,65,66,418,548,589]. As a standard the Institute applies the usual type of meta-analyses and normally does not draw upon methods for cumulative meta-analyses.

However, if the conceivable case arises that the Institute is commissioned with the regular update of a systematic review to be updated until a decision can be made on the basis of a statistically significant result, the Institute will consider applying methods for cumulative meta-analyses with correction for multiple testing.

8.3.9 Indirect comparisons

“Methods for indirect comparisons” are understood to be both techniques for a simple indirect comparison of 2 interventions as well as techniques in which direct and indirect evidence are combined. The latter are called “mixed treatment comparison (MTC) meta-analysis” [368-370], “multiple treatment meta-analysis” (MTM) [90], or “network meta-analysis” [372,476].

These methods represent an important further development of the usual meta-analytic techniques. However, there are still several unsolved methodological problems, so that currently the routine application of these methods within the framework of benefit assessments is not advisable [26,208,477,521,537]. For this reason, in its benefit assessments of interventions, the Institute primarily uses direct comparative studies (placebo-controlled studies as well as head-to-head comparisons); this means that conclusions for benefit assessments are usually inferred only from the results of direct comparative studies.

In certain situations, as, for example, in assessments of the benefit of drugs with new active ingredients [136], as well as in health economic evaluations (HEEs, see below), it can however be necessary to consider indirect comparisons and infer conclusions from them for the benefit assessment, taking a lower certainty of results into account.

For the HEE of interventions, conjoint quantitative comparisons of multiple (i.e. more than 2) interventions are usually required. Limiting the study pool to direct head-to-head comparisons would mean limiting the HEE to a single pairwise comparison or even making it totally impossible. In order to enable an HEE of multiple interventions, the Institute can also consider indirect comparisons to assess cost-effectiveness ratios [284], taking into account the lower certainty of results (compared with the approach of a pure benefit assessment).

However, appropriate methods for indirect comparisons need to be applied. The Institute disapproves the use of non-adjusted indirect comparisons (i.e. the naive use of single study arms); it accepts solely adjusted indirect comparisons. These particularly include the approach by Bucher et al. [76], as well as the MTC meta-analysis methods mentioned above. Besides the assumptions of pairwise meta-analyses, which must also be fulfilled here, in MTC meta-analyses sufficient consistency is also required in the effects estimated from the direct and indirect evidence. The latter is a critical point, as MTC meta-analyses provide valid results only if the consistency assumption is fulfilled. Even though techniques to examine inconsistencies are being developed [142,369], many open methodological questions in this area still exist. It is therefore necessary to describe completely the model applied, together with any remaining unclear issues [537]. In addition, an essential condition for consideration of an indirect comparison is that it is targeted towards the overall research question of interest and not only towards selective components such as individual outcomes.

8.3.10 Handling of unpublished or partially published data

In the quality assessment of publications, the problem frequently arises in practice that essential data or information is partially or entirely missing. This mainly concerns “grey literature” and abstracts, but also full-text publications. Moreover, it is possible that studies have not (yet) been published at the time of the Institute’s technology assessment.

It is the Institute’s aim to conduct an assessment on the basis of a data set that is as complete as possible. If relevant information is missing, the Institute therefore tries to complete the missing data, among other things by contacting the authors of publications or the study sponsors (see Sections 3.2.1 and 7.1.5). However, depending on the type of product prepared, requests for unpublished information may be restricted due to time limits.

A common problem is that important data required for the conduct of a meta-analysis (e.g. variances of effect estimates) are lacking. However, in many cases, missing data can be calculated or at least estimated from the data available [141,275,432]. If possible, the Institute will apply such procedures.

If data are only partly available or if estimated values are used, the robustness of results will be analysed and discussed, if appropriate with the support of sensitivity analyses (e.g. by presenting best-case and worst-case scenarios). However, a worst-case scenario can only be used here as proof of the robustness of a detected effect. From a worst-case scenario not confirming a previously found effect it cannot be concluded that this effect is not demonstrated. In cases where relevant information is largely or completely lacking, it may occur that a publication cannot be assessed. In such cases, it will merely be noted that further data exist on a particular topic, but are not available for assessment.

8.3.11 Description of types of bias

Bias is the systematic deviation of the effect estimate (inferred from study data) from the true effect. Bias may be produced by a wide range of possible causes [99]. The following text describes only the most important types; a detailed overview of various types of bias in different situations is presented by Feinstein [183].

“Selection bias” is caused by a violation of the random principles for sampling procedures, i.e. in the allocation of patients to intervention groups. Particularly in the comparison of 2 groups, selection bias can lead to systematic differences between groups. If this leads to an unequal distribution of important confounders between groups, the results of a comparison are usually no longer interpretable. When comparing groups, randomization is the best method to avoid selection bias [263], as the groups formed do not differ systematically with regard to known as well as unknown confounders. However, structural equality can only be ensured if the sample sizes are sufficiently large. In small studies, despite randomization, relevant differences between groups can occur at random. When comparing groups with structural inequality, the effect of known confounders can be taken into account by applying multifactorial methods. However, the problem remains of a systematic difference between the groups due to unknown or insufficiently investigated confounders.

Besides the comparability of groups with regard to potential prognostic factors, equality of treatment and equality of observation for all participants play a decisive role. “Performance bias” is bias caused by different types of care provided (apart from the intervention to be investigated). A violation of the equality of observation can lead to detection bias. Blinding is an effective protection against both performance and detection bias [316], which are summarized as “information bias” in epidemiology.

If not taken into account, protocol violations and study withdrawals can cause a systematic bias of study results, called “attrition bias”. To reduce the risk of attrition bias, in studies that aim to show superiority, the ITT principle can be applied, where all randomized study participants are analysed within the group to which they were randomly assigned, independently of protocol violations [316,338].

Missing values due to other causes present a similar problem. Missing values not due to a random mechanism can also cause bias in a result [365]. The possible causes and effects of missing values should therefore be discussed on a case-by-case basis and, if necessary, statistical methods should be applied to account or compensate for bias. In this context, replacement methods (imputation methods) for missing values are only one class of various methods available, of which none are regarded to be generally accepted. For example, EMA recommends comparison of various methods for handling missing values in sensitivity analyses [177].

When assessing screening programmes, it needs to be considered that earlier diagnosis of a disease often results only in an apparent increase in survival times, due to non-comparable starting points (“lead time bias”). Increased survival times may also appear to be indicated if a screening test preferably detects mild or slowly progressing early stages of a disease (“length bias”). The conduct of a randomized trial to assess the effectiveness of a screening test can protect against these bias mechanisms [195].

“Reporting bias” is caused by the selective reporting of only part of all relevant data and may lead to an overestimation of the benefit of an intervention in systematic reviews. If, depending on the study results, some analyses or outcomes are not reported or reported in less detail within a publication, or reported in a way deviating from the way originally planned, then “selective” or “outcome reporting bias” is present [97,160,263]. In contrast, “publication bias” describes the fact that studies finding a statistically significant negative difference or no statistically significant difference between the test intervention and control group are not published at all or published later than studies with positive and statistically significant results [530]. The pooling of published results can therefore result in a systematic bias of the common effect estimate. Graphic methods such as the funnel plot [166] and statistical methods such as meta-regression can be used to identify and consider publication bias. These methods can neither certainly confirm nor exclude the existence of publication bias, which underlines the importance of also searching for unpublished data. For example, unpublished information can be identified and obtained by means of trial registries or requests to manufacturers [347,373,436,529,530].

In studies conducted to determine the accuracy of a diagnostic strategy (index test), results may be biased if the reference test does not correctly distinguish between healthy and sick participants (“misclassification bias”). If the reference test is only conducted in a non-random sample of participants receiving the index test (“partial verification bias”) or if the reference test applied depends on the result of the index test (“differential verification bias”), this may lead to biased estimates of diagnostic accuracy. Cases in which the index test itself is a component of the reference test may lead to overestimates of diagnostic accuracy (“incorporation bias”) [351].

“Spectrum bias” is a further type of bias mentioned in the international literature. This plays a role in studies where the sample for validation of a diagnostic test consists of persons who are already known to be sick and healthy volunteers as a control group [361]. The validation of a test in such studies often leads to estimates for sensitivity and specificity that are higher than they would be in a clinical situation where patients with a suspected disease are investigated [591]. However, the use of the term “bias” (in the sense of a systematic impairment of internal validity) in this connection is unfortunate, as the results of such studies may well be internally valid if the study is conducted appropriately [591]. Nonetheless, studies of the design described above may have features (particularly regarding the composition of samples) due to which they are not informative for clinical questions in terms of external validity.

As in intervention studies, in diagnostic studies it is necessary to completely consider all study participants (including those with unclear test results) in order to avoid systematic bias of results [351]. While numerous investigations are available on the relevance and handling of publication bias in connection with intervention studies, this problem has been far less researched for diagnostic accuracy studies [351].

A general problem in the estimation of effects is bias caused by measurement errors in the study data collected [95,100]. In practice, measurement errors can hardly be avoided and it is known that non-differential measurement errors can also lead to a biased effect estimate. In the case of a simple linear regression model with a classical measurement error in the explanatory variable, “dilution bias” occurs, i.e. a biased estimate in the direction of the zero effect. However, in other models and more complex situations, bias in all directions is possible. Depending on the research question, the strength of potential measurement errors should be discussed, and, if required, methods applied to adjust for bias caused by measurement errors.

8.4 Qualitative methods

8.4.1 Qualitative studies

Qualitative research methods are applied to explore and understand subjective experiences, individual actions, and the social world [146,243,376,405]. They can enable access to opinions and experiences of patients, relatives, and medical staff with respect to a certain disease or intervention.

The instruments of qualitative research include focus groups conducted with participants of a randomized controlled trial, for example. Qualitative data can also be collected by means of interviews, observations, and written documents, such as diaries.

An analysis follows collection of data, which mainly aims to identify and analyse overlapping topics and concepts in the data collected. Among other things, qualitative methods can be used as an independent research method, in the preparation of or as a supplement to quantitative studies, within the framework of the triangulation or mixed-method approach, or after the conduct of quantitative studies, in order to explain processes or results. Qualitative research is seen as a method to promote the connection between evidence and practice [148].

Systematic synthesis of various qualitative studies investigating a common research question is also possible [25,337,395,549]. However, no generally accepted approach exists for the synthesis of qualitative studies and the combination of qualitative and quantitative data [148,149].

A) Qualitative studies in the production of health information

In the development of health information the Institute uses available qualitative research findings to identify (potential) information needs, as well as to investigate experiences with a certain disease or an intervention.

Relevant publications are then selected by means of prespecified inclusion and exclusion criteria, and the study quality is assessed by means of criteria defined beforehand. The results of the studies considered are extracted, organized by topic, and summarized in a descriptive manner for use in the development of health information. The Institute may also take this approach in the production of reports.

In recent years various instruments for evaluating the quality of qualitative studies have been developed [117]. The main task of the Institute in the assessment of qualitative studies is to determine whether the study design, study quality, and reliability are appropriate for the research question investigated. There is a weaker general consensus with regard to the validity of criteria for the conduct, assessment, and synthesis of qualitative studies when compared with other research areas [146,149,243,405].

B) Qualitative studies in the production of reports

Different sources of information can support the integration of systematic reviews [147,356,545]. One possible source are research results from qualitative studies [243,356,406,545]. Qualitative studies seem to be establishing themselves in systematic reviews on the benefit assessment of medical services [146,147,406].

Qualitative research can provide information on the acceptability and suitability of interventions in clinical practice [25,146]. The results of qualitative research can be helpful in the interpretation of a systematic review [545] and may be used in the context of primary studies or systematic reviews on determining patient-relevant outcomes [146,148,337,405,406].

The Institute can use qualitative research findings to identify patient-relevant outcomes, and to present background information on patients' experiences and on the patient relevance of the intervention to be assessed. The Institute can also use these findings in the discussion and interpretation of results of a systematic review.

8.4.2 Consultation techniques

The processing of research questions and tasks commissioned to the Institute often requires the consultation of patients, patient representatives, and national and international experts. To do this the Institute uses various consultation techniques.

In the production of reports, the Institute uses these techniques to identify patient-relevant outcomes and to involve national and international experts, and also uses them in the Institute's formal consultation procedure. In the development of health information, consultation techniques serve to involve patients and patient representatives in the identification of information needs, the evaluation of health information, and during consultation.

The Institute uses the following consultation techniques:

- key informant interviews [565], e.g. interviews with patient representatives to identify patient-relevant outcomes
- group meetings and consultations [407,411,412], e.g. within the framework of scientific debates on the Institute's products
- group interviews and focus groups [146,565], e.g. with patients with respect to the evaluation of health information
- surveys and polling (including online polling and feedback mechanisms), e.g. to identify information needs of readers of www.gesundheitsinformation.de/www.informedhealthonline.org

If a deeper understanding of experiences and opinions is necessary, then the Institute should use the scientific findings obtained from qualitative research. The use of consultation techniques and the involvement of experts are associated with an additional use of resources. However, the involvement of patients in research processes enables the consideration of patient issues and needs as well as the orientation of research towards these issues and needs [424].

Appendix A – Rationale of the methodological approach for determining the extent of added benefit

This appendix describes the rationale of the methodological approach for determining the extent of added benefit according to the Regulation for Early Benefit Assessment of New Pharmaceuticals (ANV[40]).

According to §5 (4) Sentence 1 of ANV, the dossier must present and consequently also assess "the extent to which there is added benefit". For this purpose, §5 (7) ANV contains a classification into 6 categories: (1) major added benefit, (2) considerable added benefit, (3) minor added benefit, (4) non-quantifiable added benefit, (5) no added benefit proven, (6) less benefit. For the Categories 1 to 3, §5 (7) ANV also provides a definition, as well as examples of criteria for particular consideration, as orientation for the presentation and assessment. These criteria describe qualitative characteristics (type of outcome) and also explicitly quantitative characteristics (e.g. "major" vs. "moderate" increase in survival time). In addition, a hierarchical ranking of outcomes is obviously intended, as sometimes the same modifier (e.g. "relevant") results in a different extent of added benefit for different outcomes. The corresponding details of the primarily relevant extent categories of added benefit (minor, considerable, major) are shown in Table 12. On the basis of these requirements, it was IQWiG's responsibility to operationalize the extent of added benefit for the benefit assessment.

The criteria provided in §5 (7) ANV for the extent of added benefit designate (legal) terms. Some of these terms are clearly defined (e.g. "survival time", "serious adverse events") and some are not (e.g. "alleviation of serious symptoms"). In addition, the criteria listed are not allocated to all categories. For instance, examples of "survival time" are given only for the categories "considerable" and "major" added benefit.

By using the wording "in particular" in §5 (7) with regard to the Categories 1 to 3, the legislator makes it clear that the criteria allocated to the categories are not to be regarded as conclusive. For instance, even if an increase in survival time is classified as less than "moderate", it cannot be assumed that the legislator would not at least acknowledge a "minor" added benefit. Furthermore, the outcome "(health-related) quality of life", which is explicitly defined as a criterion of benefit in §2 (3) ANV, is not mentioned at all in the list of criteria for the extent of added benefit.

[40] Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV

Table 12: Determination of extent of added benefit – Criteria according to the ANV

Extent category	**Major** **sustained and great improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Cure	Major increase in survival time	Long-term freedom from serious symptoms	Extensive avoidance of serious adverse events
	Considerable **marked improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Perceptible alleviation of the disease	Moderate increase in survival time	Alleviation of serious symptoms	Relevant avoidance of serious adverse events Important avoidance of other adverse events
	Minor **moderate and not only marginal improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy			Reduction in non-serious symptoms	Relevant avoidance of adverse events

ANV: Arzneimittel-Nutzenbewertungsverordnung (Regulation for Early Benefit Assessment of New Pharmaceuticals)

In a first step it is thus reasonable to extend the list of criteria by means of criteria that are qualitatively and quantitatively comparable. These amendments to the ANV requirements are shown in Table 13. In this context, the criteria “cure” and “perceptible alleviation of the disease” were not explicitly considered. The former generally requires operationalization. This should in principle be based on criteria referring to the outcomes “mortality” and “morbidity” (e.g. survival over a defined minimum period in patients with oncological diseases). As the ANV links “cure” solely to a major added benefit, the respective specific operationalization, on the basis of the outcomes used, must be examined with regard to whether this equals a relevant improvement in mortality or serious events. In this sense, a reduction in the duration of symptoms, for instance, in patients with simple infections, is not regarded as a “cure”.

On the basis of the above amendments the outcome categories are restructured to illustrate the ranking of outcomes intended in the ANV and to consider disease severity according to §5 (7) ANV. For this purpose, the outcomes are grouped as follows, according to their relevance (see Table 14):

1) all-cause mortality
2) serious (or severe) symptoms (or late complications); serious (or severe) adverse events; health-related quality of life
3) non-serious (or non-severe) symptoms (or late complications); non-serious (or non-severe) adverse events

Health-related quality of life is regarded to be of equal importance as serious (or severe symptoms), late complications and adverse events. The potential categories of extent of added benefit for non-serious outcomes are restricted to “minor” and “considerable”.

The requirements of the ANV make it clear that to determine the extent of added benefit, first the effect sizes must be described at outcome level. For each outcome separately the effect size – independent of its direction – is classified into 1 of the 3 extent categories (minor, considerable, major). Within the overall weighing of benefits and harms, these individual outcomes are then summarized into a global conclusion on the extent of added benefit. This step-by-step approach is described in Section 3.3.3.

Table 13: Determination of extent of added benefit – Criteria according to the ANV plus amendments[a]

		Outcome category			
		All-cause mortality	***Symptoms (morbidity)***	***Health-related quality of life***	***Adverse events***
Extent category	**Major** **sustained and great improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Major increase in survival time	Long-term freedom from serious (*or severe*) symptoms (*or late complications*)	*Major improvement in quality of life*	Extensive avoidance of serious (*or severe*) adverse events
	Considerable **marked improvem**ent in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Moderate increase in survival time	Alleviation of serious (*or severe*) symptoms (*or late complications*) *Important reduction in non-serious (or non-severe) symptoms (or late complications)*	*Important improvement in quality of life*	Relevant avoidance of serious *(or severe)* adverse events Important avoidance of other (*non-serious or non-severe*) adverse events
	Minor **moderate and not only marginal improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	*Any increase in survival time*	*Any reduction in serious (or severe) symptoms (or late complications)* Reduction in non-serious *(or non-severe)* symptoms *(or late complications)*	*Relevant improvement in quality of life*	*Any statistically significant reduction in serious (or severe) adverse events* Relevant avoidance of *(other, non-serious or non-severe)* adverse events

a: Amendments to the ANV in *italics*.

ANV: Arzneimittel-Nutzenbewertungsverordnung (Regulation for Early Benefit Assessment of New Pharmaceuticals)

Table 14: Determination of extent of added benefit – Ranked criteria according to the ANV plus amendments[a]

		Outcome category			
		All-cause mortality	**Serious** ***(or severe)*** **symptoms** ***(or late complications)*** **and adverse events**	***Health-related quality of life***	**Non-serious** ***(or non-severe)*** **symptoms** ***(or late complications)*** **and adverse events**
Extent category	**Major** **sustained and great improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Major increase in survival time	Long-term freedom or extensive avoidance	*Major improvement*	*Not applicable*
	Considerable **marked improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	Moderate increase in survival time	Alleviation or relevant avoidance	*Important improvement*	Important avoidance
	Minor **moderate and not only marginal improvement** in the therapy-relevant benefit, which has not previously been achieved versus the appropriate comparator therapy	*Any increase in survival time*	*Any reduction*	*Relevant improvement*	Relevant avoidance

a: Amendments to the ANV in *italics*.

ANV: Arzneimittel-Nutzenbewertungsverordnung (Regulation for Early Benefit Assessment of New Pharmaceuticals)

In accordance with §2 (3) ANV, the term “benefit” is defined as an “effect” and in §2 (4) ANV the term “added benefit” is defined as such an effect compared with the appropriate comparator therapy. It can be inferred from these definitions that the extent of added benefit must be determined by taking into account both the hierarchy of outcomes and effect sizes.

The ANV provides no details on the questions as to which effect sizes for the individual outcomes result in which extent category, or which effect measures should be chosen for the assessment. In principle, these questions can only be partly answered from a methodological point of view. Nevertheless, IQWiG is required to assess the extent of added benefit presented in the dossiers (§7 (2) ANV) and to draw its own conclusions on the extent. To restrict to a minimum at this stage the value judgements that will necessarily be made in the further deliberation process and to reveal them, the following measures are required:

- explicit operationalization to ensure a transparent approach
- abstract operationalization to achieve the best possible consistency between benefit assessments

Against this background a suitable effect measure must first be chosen. The initial focus is on the situation with binary data (analysis of 2x2 tables). In this context, relative effect measures – these mainly comprise the relative risk (RR) and the odds ratio (OR) – show the following advantages over absolute measures such as the risk difference (RD):

- The risk difference does not describe the effectiveness of therapy as such, as this difference strongly depends on the baseline risk in the control group. However, the baseline risk varies between regions, populations and over the course of time, as well as particularly between control groups receiving different comparator therapies. A risk difference should thus be interpreted as a descriptive measure of a specific study, not as a fixed measure of a specific treatment procedure; this is also and primarily a problem in meta-analyses [519]. This great susceptibility to external conditions calls into question the transferability of absolute effect measures from clinical studies to the daily healthcare setting. It is therefore common practice preferably to express effects shown in clinical studies as relative risks, odds ratios or hazard (or incidence) ratios [123].
- The degree of the risk difference is limited by the degree of the baseline risk (absolute risk in the control group). If this baseline risk is 1%, then the risk difference can never exceed 0.01 (or if it is 10%, the risk difference can never exceed 0.1 etc.). The risk difference could only reach the optimum value of 1 if the baseline risk was 100%. For instance, if an absolute risk reduction of at least 20% was defined as a substantial therapeutic improvement, then, for this example of a requirement, in diseases with (long-term) survival rates of > 80%, generally a major added benefit (for the corresponding outcome) would no longer be presentable.

- A further disadvantage of the use of the absolute risk reduction as an effect measure to operationalize the determination of the extent of added benefit is that an exact time point must be defined at which this absolute risk reduction is determined (e.g. after 1, 2, 5 or 10 years), if no generally accepted definitions are available (e.g. 30-day mortality for myocardial infarction).

In summary, absolute risk reductions may have more of an impact in a situation of individual decision making, but relative effect measures are more suitable for general conclusions in terms of an assessment of the added benefit of a drug.

Relative measures have in common that the zero effect (no group difference) is 1. In the following text we address effects below 1, from which effects above 1 can be calculated by using the reciprocal. For the result to be classified as a minor, considerable or major added benefit, the approach stipulates that the (two-sided) 95% confidence interval of the effect undercuts the respective threshold in terms of a shift in the hypothesis boundary. In comparison with the examination of point estimates, such an inferential statistical approach has 2 main advantages: (i) The precision of the estimate is considered in the assessment; and accordingly, (ii) the probability of statistical errors can be limited to the usual small values (e.g. 5%).

The thresholds vary with regard to the 2 dimensions “outcome category” and “extent category (of the effect)” displayed in Table 14. The greater the relevance ascribed to the outcome, the closer the thresholds should lie to 1 (below 1). This takes into account the ANV’s requirement to consider disease severity. In contrast, the greater the determined extent of the effect, the further the thresholds should lie from 1 (below 1).

Following the explicit and abstract operationalization above, a division of the thresholds in step sizes of 0.05 is planned [296]. The further development of the methodological approach leading to these thresholds is briefly explained in the following text. The further deliberations will show that the choice of 0.05 is applicable in practice and leads to reasonable conclusions.

The starting point was formed by the question as to how large the actual effects have to be in order to be classified, for instance, as effects of a “major” extent. For this purpose, a relative risk of 0.50 – proposed by Djulbegovic et al. [150] as a requirement for a “breakthrough” – was defined as an effect of a major extent for the outcome “all-cause mortality” [296].

For this actual effect (0.5) the question arises as to how the threshold should be chosen to really achieve the extent “major” with adequate power. Details of the corresponding considerations can be found in the first dossier assessment conducted by the Institute [296], but are also addressed again at the end of this appendix. Following these considerations, the simultaneous requirement for feasibility and stringency can be regarded as fulfilled for a threshold of 0.85.

In a next step, for the matrix of the extent, the other actual effects are specified and the corresponding thresholds determined. In this context it should be considered that, on the basis of the outcome category “mortality”, the requirements should increase for less serious outcomes, and on the basis of the extent category “major”, should decrease for lower extent categories. In this context, a division into sixths for the actual effects was shown to be a pragmatical solution. The thresholds for the respective extent categories are described in the following text.

1. All-cause mortality

With the usual significance level of 5%, any statistically significant increase in survival time is at least classified as “minor added benefit”, since for all-cause mortality the requirement that an effect should be “more than marginal” is regarded to be fulfilled by the outcome itself. The threshold referring to the 95% confidence interval is thus 1 here. An increase in survival time is classified as a “considerable” effect if a threshold of 0.95 is undercut. An increase in survival time is classified as being “major” if the threshold of 0.85 is undercut by the upper limit of the 95% confidence interval.

2. Serious (or severe) symptoms (or late complications), serious or (severe) adverse events, health-related quality of life

For serious (or severe) symptoms (or late complications) and serious (or severe) adverse events, any statistically significant reduction also represents at least a “minor” effect, as the requirement of “more than marginal” is already fulfilled by the quality of the outcome itself. In contrast to the desired effects on all-cause mortality, a “considerable” effect requires that a threshold of 0.90 must be undercut and a “major” effect requires that a threshold of 0.75 is undercut. To derive a major effect from these outcomes also requires that the risk of the examined event should be at least 5% in at least one of the groups compared. This additional criterion supports the relevance of the event at population level and allows for the special requirements for this category of added benefit.

The precondition for determining the extent of added benefit for outcomes on health-related quality of life (as for all PROs) is that both the instruments applied and the response criteria must be validated or at least generally established. If these results are dichotomous in terms of responders and non-responders, the above criteria for serious symptoms apply (risk for the category “major” should be at least 5%).

3. Non-serious (or non-severe) symptoms (or late complications), non-serious (or non-severe) adverse events

The specification of thresholds for the non-serious (or non-severe) symptoms (or late complications) and the non-serious (or non-severe) adverse events takes into account the lower severity compared with Categories 1 and 2. As a matter of principle, the effect for non-serious outcomes should not be classified as “major”. To classify an effect as “considerable” or “minor” the thresholds of 0.80 or 0.90 respectively must be undercut. In the latter case, this is based on the requirement for minor added benefit specified in §5 (7) ANV that there must be a moderate, and not only marginal, improvement. The procedure thus implies that effects (also statistically significant ones) only assessed as "marginal" lead to classification into the category "no added benefit".

The corresponding thresholds for all extent categories and outcome categories are presented in the following Table 15.

Table 15: Inferential statistical thresholds (hypotheses boundaries) for relative effect measures

		Outcome category		
		All-cause mortality	Serious (or severe) symptoms (or late complications) and adverse events, as well as quality of life[a]	Non-serious (or non-severe) symptoms (or late complications) and adverse events
Extent category	Major	0.85	0.75 and risk ≥ 5%[b]	Not applicable
	Considerable	0.95	0.90	0.80
	Minor	1.00	1.00	0.90

a: Precondition (as for all patient-reported outcomes): use of a validated or established instrument, as well as a validated or established response criterion.

b: Risk must be at least 5% for at least 1 of the 2 groups compared.

Detailed methodological rationale for determination of thresholds

The starting point is the planning of a (fictional) study to test the conventional hypotheses

$$H_0: RR \geq RR_0 \quad vs. \quad H_1: RR < RR_0$$

on the basis of the relative risk $RR_0 = 1$. The required sample size is calculated by specifying the significance level, the power, the risk in the control group, and the actual effect (RR_1).

For all hypothesis boundaries shifted from 1 ($RR_0 < 1$) a study of this sort has reduced power. In order to maintain the same power for the shifted hypothesis boundary of interest (the thresholds named above) as specified for the testing of the conventional (non-shifted) hypotheses, the sample size must be increased – either within the study or through a combination of several studies. Assuming the normal case of 2 (e.g. pivotal) studies, it can be assumed that the sample size is twice as large. The hypothesis boundary for the shifted hypotheses is then precisely selected so that the power for the conventional hypotheses of the 2 individual studies corresponds to the power for the shifted hypotheses of the combined (pooled) analysis. This hypothesis boundary serves as the threshold for the upper limit of the two-sided 95% confidence interval for the relative risk. For instance, the specification of a significance level of 5% (two-sided) and a power of 90% (both for the conventional and for the shifted hypothesis boundary), as well as a doubling of the sample size for the shifted hypothesis boundary resulted in a threshold of (rounded) 0.85 for the actual effect of 0.5 postulated for the outcome “mortality” and the extent category “major”.

The formula included in Appendix A of the benefit assessment on ticagrelor [296] for the relationship between the actual effect and the threshold is independent of the other requirements and is based on the algorithm used in the “power” procedure of the software SAS. The corresponding documentation for this algorithm [487] refers to the work by Fleiss et al. [192], A query to Mr Röhmel (former Speaker of the Working Group “Pharmaceutical Research” of the German Region of the International Biometric Society), as well as directly to the Technical Support Section of SAS, showed that documentation of the validity of this algorithm has evidently not been published. The question arises as to which actual effects are required in more precise calculations to reach the respective extent category with high probability.

The actual effects were thus determined by means of Monte Carlo simulations as follows:

1. The significance level for the above hypothesis is 2.5% and the power is 90%. The parameter RR_1 runs through all values between 0.2 and 0.95 at step sizes of 0.01. The risk in the control group p_C runs through all values between 0.05 and 0.95 at step sizes of 0.05. For each of these tuples (RR_1, p_C) the required sample size n is calculated using $RR_0 = 1$ according to the formula by Farrington and Manning [180] and then doubled $(m \coloneqq 2n)$.
2. For each triple (RR_1, p_C, m) a threshold T runs through all values between 1 and 0 in a descending order with a step size of -0.005. For each T the power for the above hypothesis is approximated with $RR_0 = T$. The significance level is 2.5%. For this purpose 50 000 2x2 tables are simulated with a random generator, the upper confidence interval limit for the relative risk is calculated by means of the normal distribution approximation and the delta method for estimation of variance. Subsequently, the proportion of simulation cycles is determined for which the upper confidence interval limit is smaller than T. The T cycle is stopped as soon as an approximated power is smaller than 90%. The corresponding triple (RR_1, p_C, T) is documented in a list.
3. After the cycle of all parameters in Steps 1 and 2, all triples are chosen from the list for which the threshold T deviates less than 0.01 from one of the values 0.75, 0.80, 0.85, 0.90 or 0.95.

Figure 16 shows the resulting (more precise) actual effects, depending on the risk in the control group for all thresholds specified above (points approximated by smoothed curves).

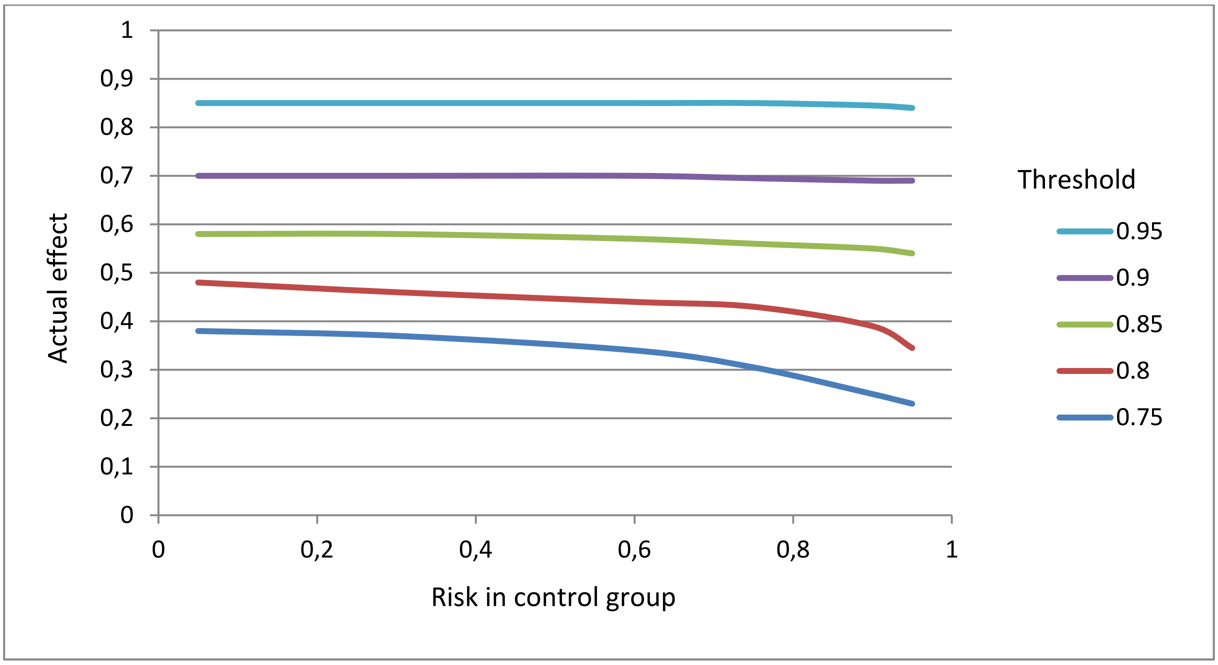


Figure 16: Actual effects depending on the baseline risk

Table 16 again contains the ranges (depending on the risk of the control group) in which the actual effects are realized, per outcome category and extent category.

Table 16: Actual effects for the relative risk

		Outcome category		
		All-cause mortality	Serious (or severe) symptoms (or late complications) and adverse events as well as quality of life	Non-serious (or non-severe) symptoms (or late complications) and adverse events
Extent category	Major	0.53–0.58	0.24–0.38	Not applicable
	Considerable	0.84–0.85	0.69–0.71	0.34–0.48
	Minor	Not applicable	Not applicable	0.69–0.71

In relation to all-cause mortality, actual relative risks of about 0.55 – i.e. still corresponding to about a halving of the risk – are to be specified for the extent "major". For the extent "considerable" the actual effect must lie at about 0.85. For serious symptoms and comparable outcomes, to be classified as a "major" extent, an actual reduction in risk to about a quarter to a third of the risk is required. Compared with the originally specified actual effects [296] good consistency is provided for thresholds lying close to 1. For the thresholds lying further away from 1, the simulation results show slightly more moderate requirements for the strength of the actual effects. The division of the thresholds as defined in Table 15 seems reasonable and practicable.

References

1. Editorial commentary: avoiding biased comparisons [online]. In: James Lind Library. 2007 [accessed: 19 April 2013]. URL: http://www.jameslindlibrary.org/essays/bias/avoiding-biased-comparisons.html.

2. SGB V Handbuch Sozialgesetzbuch V: Krankenversicherung. Altötting: KKF-Verlag; 2011.

3. Ades AE, Claxton K, Sculpher MJ. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. Health Econ 2006; 15(4): 373-381.

4. AGREE Collaboration. Appraisal of guidelines for research & evaluation: AGREE instrument. London: St. George's Hospital Medical School; 2001. URL: http://apps.who.int/rhl/agreeinstrumentfinal.pdf.

5. AGREE Next Steps Consortium. Appraisal of guidelines for research & evaluation II: AGREE II instrument [online]. May 2009 [accessed: 23 April 2013]. URL: http://www.agreetrust.org/index.aspx?o=1397.

6. Agresti A. Modelling ordered categorical data: recent advances and future challenges. Stat Med 1999; 18(18): 2191-2207.

7. Agresti A (Ed). Categorical data analysis. Hoboken: Wiley; 2002.

8. Agresti A. Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Stat Methods Med Res 2003; 12(1): 3-21.

9. Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005; 331(7511): 267-270.

10. Altenhofen L, Blumenstock G, Diel F, Döbler K, Geraedts M, Jäckel WH et al. Qualitätsindikatoren: Manual für Autoren. Neukirchen: Make a Book; 2009. (ÄZQ-Schriftenreihe; Volume 36). URL: http://www.aezq.de/mdb/edocs/pdf/schriftenreihe/schriftenreihe36.pdf.

11. Altman DG. Systematic reviews of evaluations of prognostic variables. In: Egger M, Davey Smith G, Altman DG (Ed). Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001. p. 228-247.

12. Altman DG, Bland JM. Statistic notes: absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485.

13. Altman DG, Machin D, Bryant TM, Gardner MJ. Statistics with confidence: confidence intervals and statistical guidelines. London: BMJ Publishing Group; 2000.

14. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS Med 2012; 9(5): e1001216.

15. American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol 1996; 14(2): 671-679.

16. Amir E, Seruga B, Martinez-Lopez J, Kwong R, Pandiella A, Tannock IF et al. Oncogenic targets, magnitude of benefit, and market pricing of antineoplastic drugs. J Clin Oncol 2011; 29(18): 2543-2549.

17. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. JAMA 1992; 268(2): 240-248.

18. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, Ärztliche Zentralstelle Qualitätssicherung. Das Leitlinien-Manual von AWMF und ÄZQ. Z Arztl Fortbild Qualitatssich 2001; 95(Suppl 1): 5-84.

19. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten der Deutschen Gesellschaft für Sozialmedizin und Prävention und der Deutschen Gesellschaft für Epidemiologie. Gute Praxis Sekundärdatenanalyse (GPS): Leitlinien und Empfehlungen; 3. Fassung 2012 [online]. 2012 [accessed: 18 March 2015]. URL: http://dgepi.de/fileadmin/pdf/leitlinien/GPS_fassung3.pdf.

20. Arbeitskreis Versorgungsforschung beim Wissenschaftlichen Beirat. Definition und Abgrenzung der Versorgungsforschung [online]. 8 September 2004 [accessed: 18 March 2015]. URL: http://www.bundesaerztekammer.de/downloads/Definition.pdf.

21. Arnold M. Gesundheitssystemforschung. In: Hurrelmann K, Laaser U (Ed). Gesundheitswissenschaften: Handbuch für Lehre, Forschung und Praxis. Weinheim: Beltz; 1993. p. 423-437.

22. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355(9209): 1064-1069.

23. Atkins D, Best D, Briss PA, Eccles MP, Falck-Ytter Y, Flottorp S et al. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454): 1490.

24. Atkins D, Eccles MP, Flottorp S, Guyatt GH, Henry D, Hill S et al. Systems for grading the quality of evidence and the strength of recommendations; I: critical appraisal of existing approaches. BMC Health Serv Res 2004; 4: 38.

25. Atkins S, Lewin S, Smith H, Engel M, Fretheim A, Volmink J. Conducting a meta-ethnography of qualitative literature: lessons learnt. BMC Med Res Methodol 2008; 8: 21.

26. Bafeta A, Trinquart L, Seror R, Ravaud P. Reporting of results from network meta-analyses: methodological systematic review. BMJ 2014; 348: g1741.

27. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001; 54(10): 979-985.

28. Baker SG. Surrogate endpoints: wishful thinking or reality? J Natl Cancer Inst 2006; 98(8): 502-503.

29. Baltussen R, Niessen L. Priority setting of health interventions: the need for multi-criteria decision analysis. Cost Eff Resour Alloc 2006; 4: 14.

30. Baltussen R, Stolk E, Chisholm D, Aikins M. Towards a multi-criteria approach for priority setting: an application to Ghana. Health Econ 2006; 15(7): 689-696.

31. Banta D. The development of health technology assessment. Health Policy 2003; 63(2): 121-132.

32. Barro RJ, Sala-i-Martin X. World real interest rates. In: Blanchard OJ, Fischer S (Ed). NBER Macroeconomics Annual 1990. Cambridge: MIT Press; 1990. p. 15-61.

33. Barron BA, Bukantz SC. The evaluation of new drugs: current Food and Drug Administration regulations and statistical aspects of clinical trials. Arch Intern Med 1967; 119(6): 547-556.

34. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010; 7(9): e1000326.

35. Bastian H, Kaiser T, Matschewsky S. Förderung allgemeiner Gesundheits- und Wissenschaftskenntnisse mittels Bürger- und Patienteninformationen: die Rolle des IQWiG. Z Arztl Fortbild Qualitatssich 2005; 99(6): 379-385.

36. Bender R. Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung. Med Klin 2001; 96(2): 116-121.

37. Bender R, Bunce C, Clarke M, Gates S, Lange S, Pace NL et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008; 61(9): 857-865.

38. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. BMJ 1996; 313(7057): 628.

39. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond 1997; 31(5): 546-551.

40. Bender R, Lange S. Adjusting for multiple testing: when and how? J Clin Epidemiol 2001; 54(4): 343-349.

41. Bent S, Padula A, Avins AL. Brief communication: better ways to question patients about adverse medical events; a randomized, controlled trial. Ann Intern Med 2006; 144(4): 257-261.

42. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. Ann Epidemiol 2006; 16(7): 540-544.

43. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. Prev Sci 2000; 1(1): 31-49.

44. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69(3): 89-95.

45. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309(6949): 248.

46. Bock J, Toutenburg H. Sample size determination in clinical research. In: Rao CR, Chakraborty R (Ed). Statistical methods in biological and medical sciences. Amsterdam: Elsevier; 1991. p. 515-538. (Handbook of Statistics; Volume 8).

47. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ 2006; 333(7563): 346-349.

48. Bonhoeffer J, Zumbrunn B, Heininger U. Reporting of vaccine safety data in publications: systematic review. Pharmacoepidemiol Drug Saf 2005; 14(2): 101-106.

49. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009; 62(8). 825-830, 830.e1-830.e10.

50. Bossuyt PM, Irwig LM, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332(7549): 1089-1092.

51. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000; 356(9244): 1844-1847.

52. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Ann Intern Med 2003; 138(1): 40-44.

53. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138(1): W1-W12.

54. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008; 148(4): 295-309.

55. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P. Methods and processes of the CONSORT group: example of an extension for trials assessing nonpharmacologic treatments. Ann Intern Med 2008; 148(4): W60-W66.

56. Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007; 26(1): 53-77.

57. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004; 13(9): 873-884.

58. Brenzel H, Kettner A, Kubis A, Moczall A, Müller A, Rebien M et al. Neueinstellungen im Jahr 2012: Strukturwandel und Demografie prägten die Personalsuche [online]. August 2013 [accessed: 18 March 2015]. (IAB Kurzbericht; Volume 17/2013). URL: http://doku.iab.de/kurzber/2013/kb1713.pdf.

59. Breslow NE, Day NE. Statistical methods in cancer research; volume I: the analysis of case-control studies. Lyon: International Agency for Research on Cancer; 1980. (IARC Scientific Publications; Volume 32). URL: http://www.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32.pdf.

60. Breslow NE, Day NE. Statistical methods in cancer research; volume II: the design and analysis of cohort studies. Lyon: International Agency for Research on Cancer; 1987. (IARC Scientific Publications; Volume 82). URL: http://www.iarc.fr/en/publications/pdfs-online/stat/sp82/SP82.pdf.

61. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser LA, Regier DA et al. Conjoint analysis applications in health: a checklist; a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value Health 2011; 14(4): 403-413.

62. Bridges JF, Kinter ET, Kidane L, Heinzen RR, McCormick C. Things are looking up since we started listening to patients: trends in the application of conjoint analysis in health 1982-2007. Patient 2008; 1(4): 273-282.

63. Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making 2012; 32(5): 722-732.

64. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001; 20(6): 825-840.

65. Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol 2008; 61(8): 763-769.

66. Brok J, Thorlund K, Wetterslev J, Gluud C. Apparently conclusive meta-analyses may be inconclusive: trial sequential analysis adjustment of random error risk due to repetitive testing of accumulating data in apparently conclusive neonatal meta-analyses. Int J Epidemiol 2009; 38(1): 287-298.

67. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57(3): 229-236.

68. Brouwer W, Rutten F, Koopmanschap M. Costing in economic evaluations. In: Drummond M, McGuire A (Ed). Economic evaluation in health care: merging theory with practice. Oxford: Oxford University Press; 2001. p. 68-93.

69. Brouwer WB, Culyer AJ, Van Exel NJ, Rutten FF. Welfarism vs. extra-welfarism. J Health Econ 2008; 27(2): 325-338.

70. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs in cost-effectiveness analysis: numerator or denominator; a further discussion. Health Econ 1997; 6(5): 511-514.

71. Brouwer WBF, Koopmanschap MA, Rutten FFH. Productivity costs measurement through quality of life: a response to the recommendation of the Washington Panel. Health Econ 1997; 6(3): 253-259.

72. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010; 182(18): E839-E842.

73. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. Development of the AGREE II; part 1: performance, usefulness and areas for improvement. CMAJ 2010; 182(10): 1045-1052.

74. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G et al. Development of the AGREE II; part 2: assessment of validity of items and tools to support application. CMAJ 2010; 182(10): E472-E478.

75. Brundage M, Blazeby J, Revicki D, Bass B, De Vet H, Duffy H et al. Patient-reported outcomes in randomized clinical trials: development of ISOQOL reporting standards. Qual Life Res 2013; 22(6): 1161-1175.

76. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997; 50(6): 683-691.

77. Bundesministerium der Justiz. Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz - BStatG) [online]. 25 July 2013 [accessed: 18 March 2015]. URL: http://www.gesetze-im-internet.de/bundesrecht/bstatg_1987/gesamt.pdf.

78. Bundesministerium der Justiz. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0) [online]. 12 September 2011 [accessed: 18 March 2015]. URL: http://www.gesetze-im-internet.de/bitv_2_0/BJNR184300011.html.

79. Bundesministerium für Gesundheit. Stellungnahme zur Methodik der Kosten-Nutzen-Bewertung von Arzneimitteln [online]. 6 August 2008 [accessed: 9 October 2009]. URL: http://www.bmg.bund.de/cln_117/nn_1168258/SharedDocs/Standardartikel/DE/AZ/K/Glossar-Kosten-Nutzen-Bewertung/Stellungnahme.html.

80. Bundesministerium für Gesundheit. Verordnung über die Nutzenbewertung von Arzneimitteln nach § 35a Absatz 1 SGB V für Erstattungsvereinbarungen nach § 130b SGB V (Arzneimittel-Nutzenbewertungsverordnung – AM-NutzenV) [online]. 27 March 2014 [accessed: 18 March 2015]. URL: http://www.gesetze-im-internet.de/bundesrecht/am-nutzenv/gesamt.pdf.

81. Bundessozialgericht. Urteil: Aktenzeichen B 6 A 1/08 R [online]. 6 May 2009 [accessed: 19 April 2013]. URL: http://juris.bundessozialgericht.de/cgi-bin/rechtsprechung/document.py?Gericht=bsg&Art=en&sid=965bc60820d25990f7f287c0fa2b4c2c&nr=11110&pos=0&anz=1.

82. Bundesverfassungsgericht. Leitsatz zum Beschluss des Ersten Senats: Aktenzeichen 1 BvR 347/98 [online]. 6 December 2005 [accessed: 18 March 2015]. URL: http://www.bverfg.de/entscheidungen/rs20051206_1bvr034798.html.

83. Burgers JS. Guideline quality and guideline content: are they related? Clin Chem 2006; 52(1): 3-4.

84. Burgers JS, Bailey JV, Klazinga NS, Van der Bij AK, Grol R, Feder G. Inside guidelines: comparative analysis of recommendations and evidence in diabetes guidelines from 13 countries. Diabetes Care 2002; 25(11): 1933-1939.

85. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat 2006; 5(3): 173-186.

86. Burzykowski T, Molenberghs G, Buyse M (Ed). The evaluation of surrogate endpoints. New York: Springer; 2005.

87. Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ 1997; 6(3): 217-227.

88. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000; 1(1): 49-67.

89. Cairns J. Discounting in economic evaluation. In: Drummond MF, McGuire A (Ed). Economic evaluation in health care: merging theory with practice. Oxford: Oxford University Press; 2001. p. 236-255.

90. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005; 331(7521): 897-900.

91. Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013; 309(8): 814-822.

92. Campbell F, Dickinson HO, Cook JV, Beyer FR, Eccles M, Mason JM. Methods underpinning national clinical guidelines for hypertension: describing the evidence shortfall. BMC Health Serv Res 2006; 6: 47.

93. Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012; 345: e5661.

94. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. Ottawa: CADTH; 2006. URL: http://www.cadth.ca/media/pdf/186_EconomicGuidelines_e.pdf.

95. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective. London: Chapman & Hall; 2006.

96. Centre for Evidence-based Medicine. Levels of evidence (March 2009) [online]. March 2009 [accessed: 18 March 2015]. URL: http://www.cebm.net/index.aspx?o=1025.

97. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291(20): 2457-2465.

98. Charles C, Gafni A, Whelan T, O'Brien MA. Treatment decision aids: conceptual issues and future directions. Health Expect 2005; 8(2): 114-125.

99. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol 2010; 63(11): 1205-1215.

100. Cheng CL, Van Ness JW. Statistical regression with measurement error. London: Arnold; 1999.

101. Chiou CF, Hay JW, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD et al. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care 2003; 41(1): 32-44.

102. Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ 2013; 346: f457.

103. Claxton K, Paulden M, Gravelle H, Brouwer W, Culyer AJ. Discounting and decision making in the economic evaluation of health-care technologies. Health Econ 2011; 20(1): 2-15.

104. Clement S, Ibrahim S, Crichton N, Wolf M, Rowlands G. Complex interventions to improve the health of people with limited literacy: a systematic review. Patient Educ Couns 2009; 75(3): 340-351.

105. Cochrane Collaboration. Our principles [online]. 16 January 2014 [accessed: 18 March 2015]. URL: http://www.cochrane.org./about-us/our-principles.

106. Cochrane Effective Practice and Organisation of Care Review Group. The data collection checklist [online]. June 2002 [accessed: 18 March 2015]. URL: http://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/datacollectionchecklist.pdf.

107. Collège des Économistes de la Santé. French guidelines for the economic evaluation of health care technologies [online]. September 2004 [accessed: 18 March 2015]. URL: http://www.ces-asso.org/docs/France_Guidelines_HE_Evaluation.PDF.

108. Commission of the European Communities. eEurope 2002: quality criteria for health related websites [online]. 29 November 2002 [accessed: 18 March 2015]. URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2002:0667:FIN:EN:PDF.

109. Committee of Experts on Management of Safety and Quality in Health Care, Expert Group on Safe Medication Practices. Glossary of terms related to patient and medication safety [online]. 20 October 2005 [accessed: 18 March 2015]. URL: http://www.who.int/patientsafety/highlights/COE_patient_and_medication_safety_gl.pdf.

110. Corbin JM, Strauss AL. Weiterleben lernen: Verlauf und Bewältigung chronischer Krankheit. Bern: Huber; 2003.

111. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ 2010; 341: c3920.

112. Cornell JE, Laine C. The science and art of deduction: complex systematic overviews. Ann Intern Med 2008; 148(10): 786-788.

113. Coulter A. Evidence based patient information is important, so there needs to be a national strategy to ensure it. BMJ 1998; 317(7153): 225-226.

114. Cui L, Hung HMJ, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. J Biopharm Stat 2002; 12(3): 347-358.

115. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues; the encounters of academic consultants in statistics. Stat Med 2003; 22(2): 169-186.

116. Da Costa BR, Rutjes AWS, Johnston BC, Reichenbach S, Nüesch E, Tonia T et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol 2012; 41(5): 1445-1459.

117. Daly J, Willis K, Small R, Green J, Welch N, Kealy M et al. A hierarchy of evidence for assessing qualitative health research. J Clin Epidemiol 2007; 60(1): 43-49.

118. Danner M, Hummel JM, Volz F, Van Manen JG, Wiegard B, Dintsios CM et al. Integrating patients' views into health technology assessment: Analytic Hierarchy Process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care 2011; 27(4): 369-375.

119. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature; XIV: how to decide on the applicability of clinical trial results to your patient. JAMA 1998; 279(7): 545-549.

120. Dans LF, Silvestre MA, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations; part I: general principles. J Clin Epidemiol 2011; 64(3): 231-239.

121. De Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012; 21(2): 145-172.

122. Deeks JJ. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323(7305): 157-162.

123. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002; 21(11): 1575-1600.

124. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. p. 243-296.

125. Delnoij DM, Groenewegen PP. Health services and systems research in Europe: overview of the literature 1995-2005. Eur J Public Health 2007; 17(Suppl 1): 10-13.

126. Derksen S, Keselman HJ. Backward, forward, and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 1992; 45(2): 265-282.

127. Derry S, Loke YK, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Med Res Methodol 2001; 1: 7.

128. Des Jarlais DC, Lyles C, Crepaz N. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health 2004; 94(3): 361-366.

129. Desroches B, Francis M. World real interest rates: a global savings and investment perspective [online]. March 2007 [accessed: 18 March 2015]. (Bank of Canada Working Papers; Volume 2007-16). URL: http://www.bankofcanada.ca/wp-content/uploads/2010/03/wp07-16.pdf.

130. Desu MM, Raghavarao D. Sample size methodology. Boston: Academic Press; 1990.

131. Detmer DE. Building the national health information infrastructure for personal health, health care services, public health, and research. BMC Med Inform Decis Mak 2003; 3: 1.

132. Deutsche Gesellschaft für Epidemiologie. Leitlinien und Empfehlungen zur Sicherung von guter epidemiologischer Praxis (GEP): Langversion [online]. March 2008 [accessed: 18 March 2015]. URL: http://www.gmds.de/pdf/publikationen/stellungnahmen/stell_gep_ergaenzung.pdf.

133. Deutsche Rentenversicherung Bund (Ed). Rentenversicherung in Zeitreihen: Ausgabe 2012. Berlin: DRV; 2008. (DRV-Schriften; Volume 22). URL: http://www.deutsche-rentenversicherung.de/cae/servlet/contentblob/238700/publicationFile/50912/rv_in_zeitreihen.pdf.

134. Deutscher Bundestag. Gesetzentwurf der Fraktionen SPD, CDU/CSU und BÜNDNIS 90/DIE GRÜNEN: Entwurf eines Gesetzes zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG); BT Drucksache 15/1525 [online]. 8 September 2003 [accessed: 18 March 2015]. URL: http://dipbt.bundestag.de/doc/btd/15/015/1501525.pdf.

135. Deutscher Bundestag. Gesetz zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG). Bundesgesetzblatt Teil 1 2003; (55): 2190-2258.

136. Deutscher Bundestag. Gesetz zur Neuordnung des Arzneimittelmarktes in der gesetzlichen Krankenversicherung (Arzneimittelmarktneuordnungsgesetz – AMNOG) vom 22. Dezember 2010. Bundesgesetzblatt Teil 1 2010; (67): 2262-2277.

137. Deutscher Ethikrat (Ed). Nutzen und Kosten im Gesundheitswesen: zur normativen Funktion ihrer Bewertung; Stellungnahme. Berlin: Deutscher Ethikrat; 2011. URL: http://www.ethikrat.org/dateien/pdf/stellungnahme-nutzen-und-kosten-im-gesundheitswesen.pdf.

138. Deutsches Institut für Normung. Klinische Prüfung von Medizinprodukten an Menschen: gute klinische Praxis (ISO 14155:2011 + Cor. 1:2011); deutsche Fassung EN ISO 14155:2011 + AC:2011. Berlin: Beuth; 2012.

139. Deutsches Netzwerk Evidenzbasierte Medizin. Die "Gute Praxis Gesundheitsinformation". Z Evid Fortbild Qual Gesundhwes 2010; 104(1): 66-68.

140. Devillé WL, Buntinx F, Bouter LM, Montori VM, De Vet HCW, Van der Windt DAWM et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2: 9.

141. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. Stat Med 2006; 25(13): 2299-2322.

142. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010; 29(7-8): 932-944.

143. Dierks ML, Seidel G, Horch K, Schwartz FW. Bürger- und Patientenorientierung im Gesundheitswesen. Berlin: Robert Koch-Institut; 2006. (Gesundheitsberichterstattung des Bundes; Volume 32). URL: http://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsT/buergerorientierung.pdf?__blob=publicationFile.

144. DIPEx. Healthtalk.org [online]. [Accessed: 18 March 2015]. URL: http://www.healthtalk.org/.

145. Dixon-Woods M. Writing wrongs? An analysis of published discourses about the use of patient information leaflets. Soc Sci Med 2001; 52(9): 1417-1432.

146. Dixon-Woods M, Agarwal S, Young B, Jones D, Sutton A. Integrative approaches to qualitative and quantitative evidence. London: Health Development Agency; 2004. URL: http://www.nice.org.uk/niceMedia/pdf/Integrative_approaches_evidence.pdf.

147. Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews: has established a place for itself. BMJ 2001; 323(7316): 765-766.

148. Dixon-Woods M, Fitzpatrick R, Roberts K. Including qualitative research in systematic reviews: opportunities and problems. J Eval Clin Pract 2001; 7(2): 125-133.

149. Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B et al. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. J Health Serv Res Policy 2007; 12(1): 42-47.

150. Djulbegovic B, Kumar A, Soares HP, Hozo I, Bepler G, Clarke M et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. Arch Intern Med 2008; 168(6): 632-642.

151. Dolan JG. Shared decision-making: transferring research into practice; the Analytic Hierarchy Process (AHP). Patient Educ Couns 2008; 73(3): 418-425.

152. Dolan JG, Isselhardt BJ Jr, Cappuccio JD. The Analytic Hierarchy Process in medical decision making: a tutorial. Med Decis Making 1989; 9(1): 40-50.

153. Dolan P, Edlin R, Tsuchiya A. The relative societal value of health gains to different beneficiaries: final report [online]. 31 January 2008 [accessed: 11 July 2011]. URL: http://www.haps.bham.ac.uk/publichealth/methodology/docs/publications/JH11_Social_Value_QALY_Final_Report_Paul_Dolan_et_al_2008.pdf.

154. Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: a methodological review of the literature. Health Econ 2005; 14(2): 197-208.

155. Donner A, Klar J. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

156. Draborg E, Gyrd-Hansen D, Poulsen PB, Horder M. International comparison of the definition and the practical application of health technology assessment. Int J Technol Assess Health Care 2005; 21(1): 89-95.

157. Drazen JM. COX-2 inhibitors: a lesson in unexpected problems. N Engl J Med 2005; 352(11): 1131-1132.

158. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. BMJ 1996; 313(7052): 275-283.

159. Drummond MF, Sculpher MJ, Torrance GW, O'Brian BJ, Stoddart GL. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2005.

160. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008; 3(8): e3081.

161. Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B et al. Strength of Recommendation Taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician 2004; 69(3): 548-556.

162. Edwards AGK, Elwyn GJ, Mulley A. Explaining risks: turning numerical data into meaningful pictures. BMJ 2002; 324(7341): 827-830.

163. Edwards AGK, Evans R, Dundon J, Haigh S, Hood K, Elwyn GJ. Personalised risk communication for informed decision making about taking screening tests. Cochrane Database Syst Rev 2006; (4): CD001865.

164. Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. J Pain Symptom Manage 1999; 18(6): 427-437.

165. Egger M, Davey Smith G, Altman DG (Ed). Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001.

166. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315(7109): 629-634.

167. Elwyn GJ, O'Connor A, Stacey D, Volk R, Edwards AGK, Coulter A et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ 2006; 333(7565): 417-424.

168. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000; 19(13-35): 1707-1728.

169. Entwistle VA, Sheldon TA, Sowden A, Watt IS. Evidence-informed patient choice: practical issues of involving patients in decisions about health care technologies. Int J Technol Assess Health Care 1998; 14(2): 212-225.

170. Epstein RM, Alper BS, Quill TE. Communicating evidence for participatory decision making. JAMA 2004; 291(19): 2359-2366.

171. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EG) Nr. 141/2000 des Europäischen Parlaments und des Rates vom 16. Dezember 1999 über Arzneimittel für seltene Leiden. Amtsblatt der Europäischen Gemeinschaften 2000; 43(L18): 1-5.

172. European Medicines Agency. Points to consider on application with: 1. meta-analyses; 2. one pivotal study [online]. 31 May 2001 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

173. European Medicines Agency. Guideline on the choice of the non-inferiority margin [online]. 27 July 2005 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf.

174. European Medicines Agency. Reflection paper on the regulatory guidance for the use of Health Related Quality of Life (HRQL) measures in the evaluation of medicinal products [online]. 27 July 2005 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf.

175. European Medicines Agency. Guideline on clinical trials in small populations [online]. 27 July 2006 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf.

176. European Medicines Agency. Guideline on clinical investigation of medicinal products in the treatment of diabetes mellitus: draft [online]. 20 January 2010 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/02/WC500073570.pdf.

177. European Medicines Agency. Guideline on missing data in confirmatory clinical trials [online]. 2 July 2010 [accessed: 18 March 2015]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.

178. Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. JAMA 1992; 268(17): 2420-2425.

179. Eyding D, Lelgemann M, Grouven U, Harter M, Kromp M, Kaiser T et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. BMJ 2010; 341: c4737.

180. Farrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat Med 1990; 9(12): 1447-1454.

181. Fayers P, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley; 2007.

182. Feeny D. As good as it gets but good enough for which applications? Med Decis Making 2006; 26(4): 307-309.

183. Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: Saunders; 1985.

184. Feinstein AR. Invidious comparisons and unmet clinical challenges. Am J Med 1992; 92(2): 117-120.

185. Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002; 2: 8.

186. Feldman-Stewart D, Brennenstuhl S, Brundage MD. A purpose-based evaluation of information for patients: an approach to measuring effectiveness. Patient Educ Couns 2007; 65(3): 311-319.

187. Fenwick E, O'Brien BJ, Briggs A. Cost-effectiveness acceptability curves-facts, fallacies and frequently asked questions. Health Econ 2004; 13(5): 405-415.

188. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012; 344: e686.

189. Ferreira-Gonzáles I, Busse JW, Heels-Ansdell D, Montori VM, Alk EA, Byrant DM et al. Problems with use of composite end points in cardiocascular trials: systematic review of randomized controlled trials. BMJ 2007; 334(7597): 786-792.

190. Fessler J, Fischer J, Franzen D, Geraedts M, Graf HJ, Kroegel C et al. Leitlinien-Clearingbericht "COPD": Leitlinien-Clearingverfahren von Bundesärztekammer und Kassenärztlicher Bundesvereinigung in Kooperation mit Deutscher Krankenhausgesellschaft, Spitzenverbänden der Krankenkassen und Gesetzlicher Rentenversicherung. Niebüll: Videel; 2003. (ÄZQ-Schriftenreihe; Volume 14). URL: http://www.leitlinien.de/mdb/edocs/pdf/schriftenreihe/schriftenreihe14.pdf.

191. Field MJ, Lohr KN (Ed). Clinical practice guidelines: directions for a new program. Washington: National Academy Press; 1990.

192. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980; 36(2): 343-346.

193. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. Health Aff (Millwood) 2005; 24(1): 67-78.

194. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996; 125(7): 605-613.

195. Fletcher RH, Fletcher SW. Klinische Epidemiologie: Grundlagen und Anwendung. Bern: Huber; 2007.

196. Flynn TN. Using conjoint analysis and choice experiments to estimate QALY values: issues to consider. Pharmacoeconomics 2010; 28(9): 711-722.

197. Food and Drug Administration. Guidance for industry: developing medical imaging drug and biological products; part 2: clinical indications [online]. June 2004 [accessed: 18 March 2015]. URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071603.pdf.

198. Food and Drug Administration. Guidance for industry: patient-reported outcome measures; use in medical product development to support labeling claims [online]. December 2009 [accessed: 18 March 2015]. URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.

199. Fransen GA, Van Marrewijk CJ, Mujakovic S, Muris JW, Laheij RJ, Numans ME et al. Pragmatic trials in primary care: methodological challenges and solutions demonstrated by the DIAMOND-study. BMC Med Res Methodol 2007; 7: 16.

200. Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ 2001; 322(7292): 989-991.

201. Freemantle N, Blonde L, Bolinder B, Gerber RA, Hobbs FD, Martinez L et al. Real-world trials to answer real-world questions. Pharmacoeconomics 2005; 23(8): 747-754.

202. Freemantle N, Calvert M. Weighing the pros and cons for composite outcomes in clinical trials. J Clin Epidemiol 2007; 60(7): 658-659.

203. French SD, McDonald S, McKenzie JE, Green SE. Investing in updating: how do conclusions change when Cochrane systematic reviews are updated? BMC Med Res Methodol 2005; 5: 33.

204. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11(2): 88-94.

205. Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002; 31(1): 72-76.

206. Gafni A, Birch S, Mehrez A. Economics, health and health economics: HYEs versus QALYs. J Health Econ 1993; 12(3): 325-339.

207. Garber AM, Weinstein MC, Torrance GW, Kamlet MS. Theoretical foundations of cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 25-53.

208. Gartlehner G, Moore CG. Direct versus indirect comparisons: a summary of the evidence. Int J Technol Assess Health Care 2008; 24(2): 170-177.

209. Gemeinsamer Bundesausschuss. Anlage I zum 2. Kapitel der Verfahrensordnung: Antrag zur Erprobung von Untersuchungs- und Behandlungsmethoden nach § 137e des Fünften Buches Sozialgesetzbuch (SGB V) [online]. [Accessed: 18 March 2015]. URL: http://www.g-ba.de/downloads/17-98-3627/Anlage%20I_2-Kapitel-VerfO_Erprobungsantrag_Formular.pdf.

210. Gemeinsamer Bundesausschuss. Beschluss des Gemeinsamen Bundesausschusses über die Anpassung der Beauftragung des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen zur Erstellung von Patienteninformationen [online]. 13 March 2008 [accessed: 18 March 2015]. URL: http://www.g-ba.de/downloads/39-261-650/2008-03-13-IQWiG-Anpassung-Generalauftrag.pdf.

211. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. 19 November 2014 [accessed: 19 March 2015]. URL: https://www.g-ba.de/downloads/62-492-938/VerfO_2014-06-19_iK-2014-11-19.pdf.

212. Gerhardt U. Patientenkarrieren. Frankfurt am Main: Suhrkamp; 1986.

213. Gesellschaft für Evaluation. Standards für Evaluation. Mainz: DeGEval; 2008. URL: http://www.degeval.de/fileadmin/user_upload/Sonstiges/STANDARDS_2008-12.pdf.

214. Glasziou PP, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351.

215. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Stat Med 2002; 21(11): 1503-1511.

216. Glasziou PP, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328(7430): 39-41.

217. Glenton C, Nilsen ES, Carlsen B. Lay perceptions of evidence-based information: a qualitative evaluation of a website for back pain sufferers. BMC Health Serv Res 2006; 6: 34.

218. Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R et al. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. BMC Med Res Methodol 2003; 3: 28.

219. Gøtzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. Int J Technol Assess Health Care 1996; 12(2): 238-246.

220. Graf von der Schulenburg JM, Greiner W, Jost F, Klusen N, Kubin M, Leidl R et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation: dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundheitsökonomie & Qualitätsmanagement 2007; 12(5): 285-290.

221. Graham RM, Mancher M, Miller-Wolman D, Greenfield S, Steinberg E. Clinical practice guidelines we can trust. Washington: National Academies Press; 2011. URL: http://www.awmf.org/fileadmin/user_upload/Leitlinien/International/IOM_CPG_lang_2011.pdf.

222. Gray JAM. How to get better value healthcare. Oxford: Offox Press; 2007.

223. Greenhalgh T, Hurwitz B. Narrative based medicine: why study narrative? BMJ 1999; 318(7175): 48-50.

224. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989; 18(1): 269-274.

225. Greiner W, Damm O. Die Berechnung von Kosten und Nutzen. In: Schöffski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. p. 23-42.

226. Grimes DA, Schulz K. An overview of clinical research: the lay of the land. Lancet 2002; 359(9300): 57-61.

227. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. Obstet Gynecol 2005; 105(5 Pt 1): 1114-1118.

228. Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. J Biopharm Stat 2005; 15(5): 869-882.

229. Gu Y, Norman R, Viney R. Estimating health state utility values from discrete choice experiments: a QALY space model approach [online]. 2013 [accessed: 18 March 2015]. URL: http://www.icmconference.org.uk/index.php/icmc/ICMC2013/paper/viewFile/537/210.

230. Guddat C, Grouven U, Bender R, Skipka G. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev 2012; 1: 34.

231. Guyatt GH. Evidence-based medicine. ACP J Club 1991; 14(Suppl 2): A16.

232. Guyatt GH, Jaeschke R, Roberts R. N-of-1 randomized clinical trials in pharmacoepidemiology. In: Strom BL (Ed). Pharmacoepidemiology. Chichester: Wiley; 2005. p. 665-680.

233. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A et al. Going from evidence to recommendations. BMJ 2008; 336(7652): 1049-1051.

234. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P et al. GRADE guidelines; 4: rating the quality of evidence; study limitations (risk of bias). J Clin Epidemiol 2011; 64(4): 407-415.

235. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650): 924-926.

236. Guyatt GH, Rennie D. Users' guides to the medical literature: a manual for evidence-based clinical practice. Chicago: American Medical Association; 2002.

237. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature; IX: a method for grading health care recommendations. JAMA 1995; 274(22): 1800-1804.

238. Guyatt GH, Sackett DL, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy: randomized trials in individual patients. N Engl J Med 1986; 314(14): 889-892.

239. Guyatt GH, Tugwell P, Feeny DH, Drummond MF, Haynes RB. The role of before-after studies of therapeutic impact in the evaluation of diagnostic technologies. J Chronic Dis 1986; 39(4): 295-304.

240. Hamza TH, Van Houwelingen HC, Heijenbrok-Kal MH, Stijnen T. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. J Clin Epidemiol 2009; 62(12): 1284-1291.

241. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008; 61(11): 1095-1103.

242. Harbour RT, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323(7308): 334-336.

243. Harden A, Garcia J, Oliver S, Rees R, Shepherd J, Brunton G et al. Applying systematic review methods to studies of people's views: an example from public health research. J Epidemiol Community Health 2004; 58(9): 794-800.

244. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

245. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15(4): 361-387.

246. Harris J. QALYfying the value of life. J Med Ethics 1987; 13(3): 117-123.

247. 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.

248. Hart D (Ed). Ärztliche Leitlinien im Medizin- und Gesundheitsrecht: Recht und Empirie professioneller Normbildung. Baden-Baden: Nomos; 2005. (Gesundheitsrecht und Gesundheitswissenschaften; Volume 9).

249. Harteloh P. The meaning of quality in health care: a conceptual analysis. Health Care Anal 2003; 11(3): 259-267.

250. Haute Autorité de Santé. Choices in methods for economic evaluation [online]. October 2012 [accessed: 18 March 2015]. URL: http://www.has-sante.fr/portail/upload/docs/application/pdf/2012-10/choices_in_methods_for_economic_evaluation.pdf.

251. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006; 144(6): 427-437.

252. Hayden JA, Van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med 2013; 158(4): 280-286.

253. Haynes RB. Forming research questions. J Clin Epidemiol 2006; 59(9): 881-886.

254. Haynes RB, Cotoi C, Holland J, Walters L, Wilczynski N, Jedraszewski D et al. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006; 295(15): 1801-1808.

255. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. Evid Based Med 2002; 7(2): 36-38.

256. Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company; 1987.

257. Herxheimer A, McPherson A, Miller R, Shepperd S, Yaphe J, Ziebland S. Database of Patients' Experiences (DIPEx): a multi-media approach to sharing experiences and information. Lancet 2000; 355(9214): 1540-1543.

258. Herxheimer A, Ziebland S. DIPEx: fresh insights for medical practice. J R Soc Med 2003; 96(5): 209-210.

259. Hessel F, Kohlmann T, Krauth C, Nowy R, Seitz R, Siebert U et al. Gesundheitsökonomische Evaluation in der Rehabilitation; Teil 1: Prinzipien und Empfehlungen für die Leistungserfassung. In: Verband Deutscher Rentenversicherungsträger (Ed). Förderschwerpunkt "Rehabilitationswissenschaften": Empfehlungen der Arbeitsgruppen "Generische Methoden", "Routinedaten" und "Reha-Ökonomie". Frankfurt: VDR; 1999. p. 103-193. (DRV-Schriften; Volume 16).

260. Hicks NJ. Evidence-based health care. Bandolier 1997; 4(5): 8.

261. Higgins JP, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. Res Syn Meth 2013; 4(1): 12-25.

262. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc A 2009; 172(1): 137-159.

263. Higgins JPT, Altman DG. Assessing risk of bias in included studies. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. p. 187-242.

264. Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008.

265. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11): 1539-1558.

266. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414): 557-560.

267. Hill AB (Ed). Controlled clinical trials. Oxford: Blackwell; 1960.

268. Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013; 346: e5793.

269. Hirsh J, Guyatt G. Clinical experts or methodologists to write clinical guidelines? Lancet 2009; 374(9686): 273-275.

270. Holmes-Rovner M. International Patient Decision Aid Standards (IPDAS): beyond decision aids to usual design of patient education materials. Health Expect 2007; 10(2): 103-107.

271. Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. Cochrane Database Syst Rev 2007; (2): MR000001.

272. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009; (1): MR000006.

273. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. Am J Public Health 1991; 81(12): 1630-1635.

274. Houts PS, Doak CC, Doak LG, Loscalzo MJ. The role of pictures in improving health communication: a review of research on attention, comprehension, recall, and adherence. Patient Educ Couns 2006; 61(2): 173-190.

275. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005; 5: 13.

276. Hummel JM, IJzerman MJ. The use of the Analytic Hierarchy Process in health care decision making. Enschede: University of Twente; 2009.

277. Hummel M, IJzerman M. The past and future of the AHP in health care decision making [online]. In: Proceedings of the XI International Symposium on the Analytic Hierarchy Process (ISAHP); 15.-18.06.2011; Sorrent, Italien. [Accessed: 18 March 2015]. URL: http://doc.utwente.nl/79775/1/past_and_future.pdf.

278. Hummel MJM, Steuten LMG, Groothuis-Oudshoorn KGM, IJzerman MJ. How the Analytic Hierarchy Process may fill missing gaps in early decision modeling. ISPOR Connections 2011; 17(3): 10-11.

279. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics 1997; 53(1): 11-22.

280. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. Pharmacoeconomics 2013; 31(5): 361-367.

281. Hussain T, Michel G, Shiffman RN. The Yale Guideline Recommendation Corpus: a representative sample of the knowledge content of guidelines. Int J Med Inf 2009; 78(5): 354-363.

282. ICH Expert Working Group. ICH harmonised tripartite guideline: the extent of population exposure to assess clinical safety for drugs intended for long-term treatment of non-life-threatening conditions; E1; current step 4 version [online]. 27 October 1994 [accessed: 18 March 2015]. URL: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E1/Step4/E1_Guideline.pdf.

283. Inan H. Measuring the success of your website: a customer-centric approach to website management. Frenchs Forest: Pearson Education Australia; 2002.

284. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden zur Bewertung von Verhältnissen zwischen Nutzen und Kosten [online]. 12 October 2009 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/Methodik_fuer_die_Bewertung_von_Verhaeltnissen_zwischen_Kosten_und_Nutzen.pdf.

285. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Anhang: Dokumention der Stellungnahmen zur „Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung Version 1.0“ [online]. 30 September 2008 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/08-10-14_Dokumentation_der_Stellungnahmen_KNB_Version_1_0.pdf.

286. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Aussagekraft von Surrogatendpunkten in der Onkologie: Rapid Report; Auftrag A10-05; Version 1.1 [online]. 21 November 2011 [accessed: 18 March 2015]. (IQWiG-Berichte; Volume 80). URL: https://www.iqwig.de/download/A10-05_Rapid_Report_Version_1-1_Surrogatendpunkte_in_der_Onkologie.pdf.

287. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Stellungnahmen zur „Aktualisierung einiger Abschnitte der Allgemeinen Methoden Version 4.0 sowie neue Abschnitte zur Erstellung der Allgemeinen Methoden Version 4.1“ [online]. 28 November 2013 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/Dokumentation_und_Wuerdigung_der_Stellungnahmen_IQWiG_Methoden_4-1.pdf.

288. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Entwurf einer Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung: Version 2.0 [online]. 16 March 2009 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/09-03-18_Entwurf_Methoden_Kosten-Nutzen-Bewertung_Version_2_0.pdf.

289. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Gesundheitsinformationen.de: Hinweise zur Nutzung [online]. 23 January 2014 [accessed: 19 March 2015]. URL: http://www.gesundheitsinformation.de/hinweise-zur-nutzung.2010.de.html.

290. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung: Version 1.1 [online]. 9 October 2008 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/08-10-14_Entwurf_Methoden_Kosten-Nutzen-Bewertung_Version_1_1.pdf.

291. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Technischer Anhang: Modellierung [online]. 9 October 2008 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/TA_KNB_Modellierung_v_1_0.pdf.

292. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Technischer Anhang: Unsicherheit [online]. 9 October 2008 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/TA_KNB_Unsicherheit_v_1_0.pdf.

293. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Vereinbarung über die vertrauliche Behandlung von Unterlagen [online]. 19 August 2005 [accessed: 18 March 2015]. URL: http://www.iqwig.de/download/IQWiG-VFA-Mustervertrag.pdf.

294. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Würdigung der Stellungnahmen zur „Methodik für die Bewertung von Verhältnissen zwischen Nutzen und Kosten im System der deutschen gesetzlichen Krankenversicherung Version 1.0“ [online]. 9 October 2008 [accessed: 18 March 2015]. URL: https://www.iqwig.de/download/08-10-14_Wuerdigung_der_Stellungnahmen_KNB_Version_1_0.pdf.

295. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Bupropion, Mirtazapin und Reboxetin bei der Behandlung von Depression: Abschlussbericht; Auftrag A05-20C [online]. 9 November 2009 [accessed: 18 March 2015]. (IQWiG-Berichte; Volume 68). URL: https://www.iqwig.de/download/A05-20C_Abschlussbericht_Bupropion_Mirtazapin_und_Reboxetin_bei_Depressionen.pdf.

296. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ticagrelor: Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung; Auftrag A11-02 [online]. 29 September 2011 [accessed: 18 March 2015]. (IQWiG-Berichte; Volume 96). URL: https://www.iqwig.de/download/A11-02_Ticagrelor_Nutzenbewertung_35a_SGB_V_.pdf.

297. Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington: National Academy Press; 2001. URL: http://books.nap.edu/openbook.php?record_id=10027.

298. International Conference on Harmonisation Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials; an introductory note on an international guideline. Stat Med 1999; 18(15): 1905-1942.

299. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Webauftritt [online]. [Accessed: 18 March 2015]. URL: http://www.ich.org.

300. International Society for Pharmacoeconomics and Outcomes Research. ISPOR good practices for outcomes research index [online]. [Accessed: 18 March 2015]. URL: http://www.ispor.org/workpaper/practices_index.asp.

301. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005; 294(2): 218-228.

302. Ioannidis JPA, Evans S, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004; 141(10): 781-788.

303. Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001; 285(4): 437-443.

304. Ioannidis JPA, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. Ann Intern Med 2006; 144(4): 298-300.

305. Irmen L, Linner U. Die Repräsentation generisch maskuliner Personenbezeichnungen: eine theoretische Integration bisheriger Befunde. Z Psychol 2005; 213(3): 167-175.

306. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120(8): 667-676.

307. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. Stat Med 2006; 25(15): 2688-2699.

308. Jackson N, Waters E. Criteria for the systematic review of health promotion and public health interventions. Health Promot Int 2005; 20(4): 367-374.

309. Jadad AR. Randomised controlled trials: a user's guide. London: BMJ Books; 1998.

310. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. CMAJ 1997; 156(10): 1411-1416.

311. Johannesson M. Avoiding double-counting in pharmacoeconomic studies. Pharmacoeconomics 1997; 11(5): 385-388.

312. Johnson RF. Sample size issues for conjoint analysis. In: Orme BK (Ed). Getting started with conjoint analysis: strategies for product design and pricing research. Madison: Research Publishers LLC; 2010. p. 57-66.

313. Jones B, Jarvis P, Lewis J, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996; 313(7048): 36-39.

314. Jones B, Kenward MG. Design and analysis of cross-over trials. London: Chapman and Hall; 1989. (Monographs on Statistics and Applied Probability; Volume 34).

315. Jull A, Bennett D. Do n-of-1 trials really tailor treatment? Lancet 2005; 365(9476): 1992-1994.

316. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001; 323(7303): 42-46.

317. Kass PH, Gold EB. Modern epidemiologic study designs. In: Ahrens W, Pigeot I (Ed). Handbook of epidemiology. Berlin: Springer; 2005. p. 321-344.

318. Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004; 4: 22.

319. Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med 2003; 138(8): 644-650.

320. Kelley E, Hurst J. Health Care Quality Indicators Project: conceptual framework paper [online]. 9 March 2006 [accessed: 18 March 2015]. (OECD Health Working Papers; Volume 23). URL: http://www.oecd.org/dataoecd/1/36/36262363.pdf.

321. Kettunen T, Liimatainen L, Villberg J, Perko U. Developing empowering health counseling measurement: preliminary results. Patient Educ Couns 2006; 64(1-3): 159-166.

322. Kickbusch IS. Health literacy: addressing the health and education divide. Health Promot Int 2001; 16(3): 289-297.

323. Kieser M. Assessment of clinical relevance by considering point estimates and associated confidence intervals. Pharm Stat 2005; 4(2): 101-107.

324. Kieser M, Röhmel J, Friede T. Power and sample size determination when assessing the clinical relevance of trial results by 'responder analyses'. Stat Med 2004; 23(21): 3287-3305.

325. Klusen N, Meusch M (Ed). Wettbewerb und Solidarität im europäischen Gesundheitsmarkt. Baden-Baden: Nomos Verlagsgesellschaft; 2006. (Beiträge zum Gesundheitsmanagement; Volume 16).

326. Knelangen M, Zschorlich B, Büchter R, Fechtelpeter D, Rhodes T, Bastian H. Online-Umfragen auf Gesundheitsinformation.de: Ermittlung potenzieller Informationsbedürfnisse für evidenzbasierte Gesundheitsinformationen. Z Evid Fortbild Qual Gesundhwes 2010; 104(8-9): 667-673.

327. Knottnerus JA, Bouter LM. The ethics of sample size: two-sided testing and one-sided thinking. J Clin Epidemiol 2001; 54(2): 109-110.

328. Köbberling J. Der Zweifel als Triebkraft des Erkenntnisgewinns in der Medizin. In: Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N (Ed). Lehrbuch evidenzbasierte Medizin in Klinik und Praxis. Köln: Deutscher Ärzteverlag; 2007. p. 3-14.

329. Köbberling J, Trampisch HJ, Windeler J. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990; 28(12): 873-879.

330. Koch A, Ziegler S. Metaanalyse als Werkzeug zum Erkenntnisgewinn. Med Klin 2000; 95(2): 109-116.

331. Kolman J, Meng P, Scott G. Good clinical practice: standard operating procedures for clinical researchers. Chichester: Wiley; 1998.

332. Kommission der Europäischen Gemeinschaften. Richtlinie 2003/63/EG der Kommission vom 25. Juni 2003 zur Änderung der Richtlinie 2001/83/EG des Europäischen Parlaments und des Rates zur Schaffung eines Gemeinschaftskodexes für Humanarzneimittel. Amtsblatt der Europäischen Gemeinschaften 2003; 46(L159): 46-94.

333. Kools M, Van de Wiel MW, Ruiter RA, Kok G. Pictures and text in instructions for medical devices: effects on recall and actual performance. Patient Educ Couns 2006; 64(1-3): 104-111.

334. Koopmanschap MA, Rutten FFH, Van Ineveld BM, Van Roijen L. The friction cost method for measuring indirect costs of disease. J Health Econ 1995; 14(2): 171-189.

335. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. JAMA 2006; 296(10): 1286-1289.

336. Kranich C. Patientenkompetenz: was müssen Patienten wissen und können? Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2004; 47(10): 950-956.

337. Kristensen FB, Sigmund H. Health technology assessment handbook. Kopenhagen: Danish Centre for Heath Technology Assessment; 2007. URL: http://sundhedsstyrelsen.dk/publ/Publ2008/MTV/Metode/HTA_Handbook_net_final.pdf.

338. Kristman V, Manno M, Cote P. Loss to follow-up in cohort studies: how much is too much? Eur J Epidemiol 2004; 19(8): 751-760.

339. Krug S. Don't make me think! Web Usability; das intuitive Web. Heidelberg: mitp; 2006.

340. Kulbe A. Grundwissen Psychologie, Soziologie und Pädagogik: Lehrbuch für Pflegeberufe. Stuttgart: Kohlhammer; 2009.

341. Kunz R, Djulbegovic B, Schünemann HJ, Stanulla M, Muti P, Guyatt G. Misconceptions, challenges, uncertainty, and progress in guideline recommendations. Semin Hematol 2008; 45(3): 167-175.

342. Kunz R, Lelgemann M, Guyatt GH, Antes G, Falck-Ytter Y, Schünemann H. Von der Evidenz zur Empfehlung. In: Kunz R, Ollenschläger G, Raspe H, Jonitz G, Donner-Banzhoff N (Ed). Lehrbuch evidenzbasierte Medizin in Klinik und Praxis. Köln: Deutscher-Ärzte-Verlag; 2007. p. 231-247.

343. Laaser U, Hurrelmann K. Gesundheitsförderung und Krankheitsprävention. In: Hurrelmann K, Laaser U (Ed). Handbuch Gesundheitswissenschaften. Weinheim: Juventa Verlag; 1998. p. 395-424.

344. Lange S, Freitag G. Choice of delta: requirements and reality; results of a systematic review. Biom J 2005; 47(1): 12-27.

345. Lapsley P. The patient's journey: travelling through life with a chronic illness. BMJ 2004; 329(7466): 582-583.

346. Last JM, Spasoff RA, Harris SS, Thuriaux MC (Ed). A dictionary of epidemiology. Oxford: Oxford University Press; 2001.

347. Lau J, Ioannidis JPA, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006; 333(7568): 597-600.

348. Lavis JN. How can we support the use of systematic reviews in policymaking? PLoS Med 2009; 6(11): e1000141.

349. Law AM, Kelton WD. Simulation modelling and analysis. Boston: McGraw Hill; 2000.

350. Law AM, McComas MG. How to build valid and credible simulation models. In: Peters BA, Smith JS, Medeiros DJ, Rohrer MW (Ed). Proceedings of the 2001 Winter Simulation Conference; 09.-12.09.2001; Arlington, USA. 2001. p. 22-29. URL: http://www.informs-sim.org/wsc01papers/004.PDF.

351. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149(12): 889-897.

352. Lefebvre C, Manheimer E, Glanville J. Searching for studies. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. New York: Wiley; 2008. p. 95-150.

353. Lehoux P, Tailliez S, Denis JL, Hivon M. Redefining health technology assessment in Canada: diversification of products and contextualization of findings. Int J Technol Assess Health Care 2004; 20(3): 325-336.

354. Leidl R, Graf von der Schulenburg JM, Wasem J (Ed). Ansätze und Methoden der ökonomischen Evaluation: eine internationale Perspektive. Baden-Baden: Nomos Verlagsgesellschaft; 1999. (Health Technology Assessments; Volume 9).

355. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322(7300): 1479-1480.

356. Leys M. Health care policy: qualitative evidence and health technology assessment. Health Policy 2003; 65(3): 217-226.

357. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009; 339: b2700.

358. Liberati A, Sheldon TA, Banta HD. EUR-ASSESS project subgroup report on methodology: methodological guidance for the conduct of health technology assessment. Int J Technol Assess Health Care 1997; 13(2): 186-219.

359. Lieb K, Klemperer D, Koch K, Baethge C, Ollenschläger G, Ludwig WD. Interessenskonflikt in der Medizin: mit Transparenz Vertrauen stärken. Dtsch Arztebl 2011; 108(6): A256-A260.

360. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009; 62(4): 364-373.

361. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, Van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11): 1061-1066.

362. Lipkus IM, Hollands JG. The visual communication of risk. J Natl Cancer Inst Monogr 1999; (25): 149-163.

363. Lipscomb J, Drummond M, Fryback D, Gold M, Revicki D. Retaining, and enhancing, the QALY. Value Health 2009; 12(Suppl 1): S18-S26.

364. Lipscomb J, Weinstein MC, Torrance GW. Time preference. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 214-246.

365. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley; 2002.

366. Lo B, Field MJ (Ed). Conflict of interest in medical research, education, and practice. Washington: National Academies Press; 2009.

367. Lord SJ, Irwig LM, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006; 144(11): 850-855.

368. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004; 23(20): 3105-3124.

369. Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006; 101(474): 447-459.

370. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. Stat Med 2007; 26(20): 3681-3699.

371. Luce BR, Manning WG, Siegel JE, Lipscomb J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Russell LB, Siegel JE, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 176-213.

372. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002; 21(16): 2313-2324.

373. Macaskill P, Walter SD, Irwig LM. A comparison of methods to detect publication bias in meta-analysis. Stat Med 2001; 20(4): 641-654.

374. MacDermid JC, Brooks D, Solway S, Switzer-McIntyre S, Brosseau L, Graham ID. Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. BMC Health Serv Res 2005; 5: 18.

375. Maetzel A. Der Gebrauch von Nutzwerten im gesundheitsökonomischen Vergleich von Interventionen bei verschiedenen Krankheitsbildern: eine Einführung. Z Rheumatol 2004; 63(5): 380-384.

376. Malterud K. The art and science of clinical knowledge: evidence beyond measures and numbers. Lancet 2001; 358(9279): 397-400.

377. Mandelblatt JS, Fryback DG, Weinstein MC, Russell LB, Gold MR, Hadorn DC. Assessing the effectiveness of health interventions. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 135-175.

378. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol 2009; 27(24): 4027-4034.

379. Mandrekar SJ, Sargent DJ. All-comers versus enrichment design strategy in phase II trials. J Thorac Oncol 2011; 6(4): 658-660.

380. Mangiapane S, Velasco Garrido M. Surrogatendpunkte als Parameter der Nutzenbewertung [online]. 2009 [accessed: 18 March 2015]. (Schriftenreihe Health Technology Assessment; Volume 91). URL: http://portal.dimdi.de/de/hta/hta_berichte/hta250_bericht_de.pdf.

381. March JS, Silva SG, Compton S, Shapiro M, Califf R, Krishnan R. The case for practical clinical trials in psychiatry. Am J Psychiatry 2005; 162(5): 836-846.

382. Marsh K, Lanitis T, Neasham D, Orfanos P, Caro J. Assessing the value of healthcare interventions using multi-criteria decision analysis: a review of the literature. Pharmacoeconomics 2014; 32(4): 345-365.

383. Martini P. Methodenlehre der therapeutischen Untersuchung. Berlin: Springer; 1932.

384. Matthys J, De Meyere M, Van Driel ML, De Sutter A. Differences among international pharyngitis guidelines: not just academic. Ann Fam Med 2007; 5(5): 436-443.

385. Mauskopf JA, Earnshaw S, Mullins CD. Budget impact analysis: review of the state of the art. Expert Rev Pharmacoecon Outcomes Res 2005; 5(1): 65-79.

386. Mauskopf JA, Sullivan SD, Annemans L, Caro J, Mullins CD, Nuijten M et al. Principles of good practice for budget impact analysis: report of the ISPOR Task Force on Good Research Practices; budget impact analysis. Value Health 2007; 10(5): 336-347.

387. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. JAMA 2003; 289(19): 2545-2553.

388. McAlister FA, Van Diepen S, Padwal RS, Johnson JA, Majumdar SR. How evidence-based are the recommendations in evidence-based guidelines? PLoS Med 2007; 4(8): e250.

389. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ 2002; 324(7351): 1448-1451.

390. McGauran N, Wieseler B, Kreis J, Schüler YB, Kölsch H, Kaiser T. Reporting bias in medical research: a narrative review. Trials 2010; 11(1): 37.

391. McGregor M, Caro JJ. QALYs: are they helpful to decision makers? Pharmacoeconomics 2006; 24(10): 947-952.

392. McMurray J, Swedberg K. Treatment of chronic heart failure: a comparison between the major guidelines. Eur Heart J 2006; 27(15): 1773-1777.

393. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005; 97(16): 1180-1184.

394. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care 2013; 29(3): 343-350.

395. Mills E, Jadad AR, Ross C, Wilson K. Systematic review of qualitative studies exploring parental beliefs and attitudes toward childhood vaccination identifies common barriers to vaccination. J Clin Epidemiol 2005; 58(11): 1081-1088.

396. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869.

397. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009; 339: b2535.

398. Moher D, Pham B, Lawson ML, Klassen TP. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess 2003; 7(41): 1-90.

399. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007; 4(3): e78.

400. Molenberghs G, Burzykowski T, Alonso A, Assam P, Tilahun A, Buyse M. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Stat Methods Med Res 2010; 19(3): 205-236.

401. Molnar FJ, Man-Son-Hing M, Fergusson D. Systematic review of measures of clinical significance employed in randomized controlled trials of drugs for dementia. J Am Geriatr Soc 2009; 57(3): 536-546.

402. Mühlbacher AC, Bethge S, Tockhorn A. Präferenzmessung im Gesundheitswesen: Grundlage von Discrete-Choice-Experimenten. Gesundheitsökonomie & Qualitätsmanagement 2013; 18(4): 159-172.

403. Müller HP, Schmidt K, Conen D. Qualitätsmanagement: interne Leitlinien und Patientenpfade. Med Klin 2001; 96(11): 692-697.

404. Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002; 136(2): 122-126.

405. Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998; 2(16): 1-274.

406. National Advisory Committee on Health and Disability. Screening to improve health in New Zealand: criteria to assess screening. Wellington: National Health Committee; 2003. URL: https://www.nsu.govt.nz/system/files/resources/screening-to-improve-health.pdf.

407. National Health and Medical Research Council. Statement on consumer and community participation in health and medical research. Canberra: Commonwealth of Australia; 2002. URL: http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/r22.pdf.

408. National Health and Medical Research Council. Cultural competency in health: a guide for policy, partnerships and participation. Canberra: Commonwealth of Australia; 2006. URL: http://www.nhmrc.gov.au/_files_nhmrc/file/publications/synopses/hp19.pdf.

409. National Institute for Health and Care Excellence. Guide to the processes of technology appraisal. London: NICE; 2014. URL: http://www.nice.org.uk/article/pmg19/resources/non-guidance-guide-to-the-processes-of-technology-appraisal-pdf.

410. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. London: NICE; 2013. URL: http://www.nice.org.uk/article/pmg9/resources/non-guidance-guide-to-the-methods-of-technology-appraisal-2013-pdf.

411. National Resource Centre for Consumer Participation in Health. Methods and models of consumer participation [online]. 1 September 2008 [accessed: 6 May 2013]. (Information Series; Volume 2). URL: http://www.healthissuescentre.org.au/documents/items/2008/09/231154-upload-00001.pdf.

412. National Resource Centre for Consumer Participation in Health. Feedback, participation and consumer diversity: a literature review. Canberra: Commonwealth of Australia; 2000. URL: http://www.healthissuescentre.org.au/documents/items/2008/08/226293-upload-00001.pdf.

413. Neidhardt K, Wasmuth T, Schmid A. Die Gewichtung multipler patientenrelevanter Endpunkte: ein methodischer Vergleich von Conjoint Analyse und Analytic Hierarchy Process unter Berücksichtigung des Effizienzgrenzenkonzepts des IQWiG; Diskussionspapier [online]. February 2012 [accessed: 18 March 2015]. (Wirtschaftswissenschaftliche Diskussionspapiere; Volume 02-12). URL: http://www.fiwi.uni-bayreuth.de/de/download/WP_02-12.pdf.

414. Nielsen J, Loranger H. Web Usability. München: Addison-Wesley; 2008.

415. Nilsen ES, Myrhaug HT, Johansen M, Oliver S, Oxman AD. Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. Cochrane Database Syst Rev 2006; (3): CD004563.

416. Nord E. An alternative to QALYs: the saved young life equivalent (SAVE). BMJ 1992; 305(6858): 875-877.

417. Nord E. Cost-value analysis in health care: making sense out of QALYs. Cambridge: Cambridge University Press; 1999.

418. Nüesch E, Jüni P. Commentary: which meta-analyses are conclusive? Int J Epidemiol 2009; 38(1): 298-303.

419. Nutbeam D. Health promotion glossary. Health Promot Int 1998; 13(4): 349-364.

420. O'Connor AM, Bennett CL, Stacey D, Barry M, Col NF, Eden KB et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2009; (3): CD001431.

421. O'Mahony JF, Paulden M. NICE's selective application of differential discounting: ambiguous, inconsistent, and unjustified. Value Health 2014; 17(5): 493-496.

422. Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schünemann H et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev 2011; (4): MR000012.

423. Oliver A. A normative perspective on discounting health outcomes. J Health Serv Res Policy 2013; 18(3): 186-189.

424. Oliver S, Clarke-Jones L, Rees R, Milne R, Buchanan P, Gabbay J et al. Involving consumers in research and development agenda setting for the NHS: developing an evidence-based approach. Health Technol Assess 2004; 8(15): 1-148.

425. Oostenbrink JB, Koopmanschap MA, Rutten FF. Standardisation of costs: the Dutch Manual for Costing in economic evaluations. Pharmacoeconomics 2002; 20(7): 443-454.

426. Orlewska E, Mierzejewski P. Proposal of Polish guidelines for conducting financial analysis and their comparison to existing guidance on budget impact in other countries. Value Health 2004; 7(1): 1-10.

427. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Can Med Assoc J 1988; 138(8): 697-703.

428. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 44(11): 1271-1278.

429. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116(1): 78-84.

430. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA et al. Agreement among reviewers of review articles. J Clin Epidemiol 1991; 44(1): 91-98.

431. Parkin DM, Chen VW, Ferlay J, Galceran J, Storm HH (Ed). Comparability and quality control in cancer registration. Lyon: International Agency for Research on Cancer; 1994. (IARC Technical Reports; Volume 19).

432. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998; 17(24): 2815-2834.

433. Paulden M, Claxton K. Budget allocation and the revealed social rate of time preference for health. Health Econ 2012; 21(5): 612-618.

434. Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA 2012; 308(16): 1676-1684.

435. Perleth M, Jakubowski E, Busse R. What is 'best practice' in health care? State of the art and perspectives in improving the effectiveness and efficiency of the European health care systems. Health Policy 2001; 56(3): 235-250.

436. Peters JL, Sutton A, Jones D, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006; 295(6): 676-680.

437. Petitti DB, Teutsch SM, Barton MB, Sawaya GF, Ockene JK, DeWitt T. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. Ann Intern Med 2009; 150(3): 199-205.

438. Petkova E, Tarpey T, Huang L, Deng L. Interpreting meta-regression: application to recent controversies in antidepressants' efficacy. Stat Med 2013; 32(17): 2875-2892.

439. Pfaff H, Glaeske G, Neugebauer EA, Schrappe M. Memorandum III: Methoden für die Versorgungsforschung (Teil 1). Gesundheitswesen 2009; 71(8-9): 505-510.

440. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004; 8(36): iii-iv, ix-xi, 1-158.

441. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012; 308(24): 2594-2604.

442. Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.

443. Poynard T, Munteanu M, Ratziu V, Benhamou Y, Di Martino V, Taieb J et al. Truth survival in clinical research: an evidence-based requiem? Ann Intern Med 2002; 136(12): 888-895.

444. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989; 8(4): 431-440.

445. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. BMC Med Res Methodol 2012; 12: 173.

446. Raftery JP. How should we value future health? Was NICE right to change? Value Health 2013; 16(5): 699-700.

447. Raum E, Perleth M. Methoden der Metaanalyse von diagnostischen Genauigkeitsstudien. Köln: Deutsches Institut für Medizinische Dokumentation und Information; 2003. (Schriftenreihe Health Technology Assessment; Volume 2). URL: http://portal.dimdi.de/de/hta/hta_berichte/hta025_bericht_de.pdf.

448. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005; 58(10): 982-990.

449. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008; 61(2): 102-109.

450. Richardson J, lezzi A, M.A. K, Maxwell A. Cross-national comparison of twelve quality of life instruments: MIC paper 2. Melbourne: Centre for Health Economics; 2012. (Research Papers; Volume 78). URL:
http://www.buseco.monash.edu.au/centres/che/pubs/researchpaper78.pdf.

451. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011; 342: 964-967.

452. Ringbaek T, Brøndum E, Martinez G, Lange P. EuroQoL in assessment of the effect of pulmonary rehabilitation COPD patients. Respir Med 2008; 102(11): 1563-1567.

453. Rockwood K, Fay S, Song X, MacKnight C, Gorman M. Attainment of treatment goals by people with Alzheimer's disease receiving galantamine: a randomized controlled trial. Can Med Assoc J 2006; 174(8): 1099-1105.

454. Roebruck P, Elze M, Hauschke D, Leverkus F, Kieser M. Literaturübersicht zur Fallzahlplanung für Äquivalenzprobleme. Inform Biom Epidemiol Med Biol 1997; 28(2): 51-63.

455. Röhmel J, Hauschke D, Koch A, Pigeot I. Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren: Nicht-Unterlegenheit in klinischen Studien. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2005; 48(5): 562-571.

456. Ross SM. Simulation. Amsterdam: Elsevier Academic Press; 2006.

457. Rossouw JE. Estrogens for prevention of coronary heart disease: putting the brakes on the bandwagon. Circulation 1996; 94(11): 2982-2985.

458. Rothwell PM. Treating individuals 2: subgroup analysis in randomised controlled trials; importance, indications, and interpretation. Lancet 2005; 365(9454): 176-186.

459. Rotter T, Kinsman L, James E, Machotta A, Gothe H, Willis J et al. Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. Cochrane Database Syst Rev 2010; (3): CD006632.

460. Royal Society. Science and the public interest: communicating the results of new scientific research to the public [online]. April 2006 [accessed: 18 March 2015]. URL: http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/publications/2006/8315.pdf.

461. Royall RM. The effect of sample size on the meaning of significance tests. Am Stat 1986; 40(4): 313-315.

462. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med 2000; 19(14): 1831-1847.

463. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl Stat 1994; 43(3): 429-467.

464. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Stat Med 2009; 28(5): 721-738.

465. Russell LB, Siegen JE, Daniels N, Gold MR, Luce BR, Mandelblatt JS. Cost-effectiveness analysis as a guide to resource allocation in health: roles and limitations. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 3-24.

466. Ryan M, Gerard K, Amaya-Amaya M (Ed). Using discrete choice experiments to value health and health care. Dordrecht: Springer; 2008. (The Economics of Non-Market Goods and Resources; Volume 11).

467. Saaty T, Vargas LG. Decision making with the analytic network process: economic, political, social and technological applications with benefits, opportunities, costs and risks. New York: Springer Science and Business Media; 2006. (International Series in Operations Research Management Science; Volume 95).

468. Saaty TL. A scaling method for priorities in hierarchical structures. J Math Psychol 1977; 15(3): 234-281.

469. Saaty TL. Theory and applications of the analytic network process: decision making with benefits, opportunities, costs and risks. Pittsburg: RWS Publications; 2005.

470. Saaty TL. Decision making with the Analytic Hierarchy Process. International Journal of Services Sciences 2008; 1(1): 83-98.

471. Saaty TL, Vargas LG. The Analytic Hierarchy Process: wash criteria should not be ignored. International Journal of Management and Decision Making 2006; 7(2/3): 180-188.

472. Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen. Bedarfsgerechtigkeit und Wirtschaftlichkeit; Band III: Über- Unter- und Fehlversorgung; Gutachten 2000/2001; ausführliche Zusammenfassung [online]. August 2001 [accessed: 18 March 2015]. URL: http://www.svr-gesundheit.de/fileadmin/user_upload/Gutachten/2000-2001/Kurzf-de-01.pdf.

473. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32(1-2): 51-63.

474. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023): 71-72.

475. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. Edinburgh: Churchill Livingstone; 2000.

476. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res 2008; 17(3): 279-301.

477. Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol 2009; 62(8): 857-864.

478. Sampson M, McGowan J, Cogo E, Grimshaw J, Moher D, Lefebvre C. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009; 62(9): 944-952.

479. Sampson M, McGowan J, Lefebvre C, Moher D, Grimshaw J. PRESS: Peer Review of Electronic Search Strategies. Ottawa: Canadian Agency for Drugs and Technologies in Health; 2008. URL: http://www.cadth.ca/media/pdf/477_PRESS-Peer-Review-Electronic-Search-Strategies_tr_e.pdf.

480. Sampson M, Shojania KG, McGowan J, Daniel R, Rader T, Iansavichene AE et al. Surveillance search techniques identified the need to update systematic reviews. J Clin Epidemiol 2008; 61(8): 755-762.

481. Sampson MJ. Updating searches for systematic reviews [Dissertation]. Aberystwyth: Universität; 2009.

482. Sänger S, Lang B, Klemperer D, Thomeczek C, Dierks ML. Manual Patienteninformation: Empfehlungen zur Erstellung evidenzbasierter Patienteninformationen. Berlin: Ärztliches Zentrum für Qualität in der Medizin; 2006. (ÄZQ-Schriftenreihe; Volume 25). URL: http://www.aezq.de/mdb/edocs/pdf/schriftenreihe/schriftenreihe25.pdf.

483. Santo A, Laizner AM, Shohet L. Exploring the value of audiotapes for health literacy: a systematic review. Patient Educ Couns 2005; 58(3): 235-243.

484. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005; 23(9): 2020-2027.

485. Sargent DJ, Mandrekar SJ. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. Clin Trials 2013; 10(5): 647-652.

486. Sargent RG. Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (Ed). Proceedings of the 2004 Winter Simulation Conference; 05.-08.12.2004; Washington, USA. 2004. p. 17-28. URL: http://www.informs-sim.org/wsc04papers/004.pdf.

487. SAS Institute. SAS/STAT 9.2 user’s guide: second edition [online]. 2009 [accessed: 18 March 2015]. URL: http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf.

488. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc Ser A 1999; 162(1): 71-94.

489. Savović J, Jones HE, Altman DG, Harris RJ, Jűni P, Pildal J et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess 2012; 16(35): 1-82.

490. Sawaya GF, Guirguis-Blake J, LeFevre M, Harris R, Petitti D. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. Ann Intern Med 2007; 147(12): 871-875.

491. Scherer RW, Langenberg P, Von Elm E. Full publication of results initially presented in abstracts. Cochrane Database Syst Rev 2007; (2): MR000005.

492. Schluter PJ, Ware RS. Single patient (n-of-1) trials with binary treatment preference. Stat Med 2005; 24(17): 2625-2636.

493. Schneider N, Dierks ML, Seidel G, Schwartz FW. The federal government commissioner for patient issues in Germany: initial analysis of the user inquiries. BMC Health Serv Res 2007; 7: 24.

494. Schöffski O. Grundformen gesundheitsökonomischer Evaluationen. In: Schöfski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. p. 43-70.

495. Schöffski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012.

496. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332.

497. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. Health Qual Life Outcomes 2006; 4: 62.

498. Schünemann HJ, Best D, Vist GE, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. Can Med Assoc J 2003; 169(7): 677-680.

499. Schünemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development; 9: grading evidence and recommendations. Health Res Policy Syst 2006; 4: 21.

500. Sculpher M. The role and estimation of productivity costs in economic evaluation. In: Drummond MF, McGuire A (Ed). Economic evaluation in health care: merging theorey with practice. Oxford: Oxford University Press; 2001. p. 94-112.

501. Sculpher MJ, O'Brien BJ. Income effects of reduced health and health effects of reduced income: implications for health-state valuation. Med Decis Making 2000; 20(2): 207-215.

502. Senn SJ. Inherent difficulties with active control equivalence studies. Stat Med 1993; 12(24): 2367-2375.

503. Senn SJ. The many modes of meta. Drug Inf J 2000; 34(2): 535-549.

504. Senn SJ. Trying to be precise about vagueness. Stat Med 2007; 26(7): 1417-1430.

505. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS One 2007; 2(12): e1350.

506. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 7: 10.

507. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009; 62(10): 1013-1020.

508. Shechter SM, Schaefer AJ, Braithwaite RS, Roberts MS. Increasing the efficiency of Monte Carlo cohort simulations with variance reduction techniques. Med Decis Making 2006; 26(5): 550-553.

509. Shekelle PG, Ortiz E, Rhodes S, Morton SC, Eccles MP, Grimshaw JM et al. Validity of the agency for healthcare research and quality clinical practice guidelines: how quickly do guidelines become outdated? JAMA 2001; 286(12): 1461-1467.

510. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007; 147(4): 224-233.

511. Siebert U. Entscheidungsanalytische Modelle zur Sicherung der Übertragbarkeit internationaler Evidenz von HTA auf den Kontext des deutschen Gesundheitssystems: ein Methodenbeitrag zu HTA [online]. 2005 [accessed: 18 March 2015]. (Schriftenreihe Health Technology Assessment; Volume 16). URL: http://portal.dimdi.de/de/hta/hta_berichte/hta099_bericht_de.pdf.

512. Sillence E, Briggs P, Harris PR, Fishwick L. How do patients evaluate and make use of online health information? Soc Sci Med 2007; 64(9): 1853-1862.

513. Silvestre MAA, Dans LF, Dans AL. Trade-off between benefit and harm is crucial in health screening recommendations; part II: evidence summaries. J Clin Epidemiol 2011; 64(3): 240-249.

514. Simmonds MC, Higgins JPT. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Stat Med 2007; 26(15): 2982-2999.

515. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994; 69(6): 979-985.

516. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Cancer Inst 2009; 101(21): 1446-1452.

517. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol 2014; 67(6): 612-621.

518. Skipka G, Bender R. Intervention effects in the case of heterogeneity between three subgroups: assessment within the framework of systematic reviews. Methods Inf Med 2010; 49(6): 613-617.

519. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses: sometimes informative, usually misleading. BMJ 1999; 318(7197): 1548-1551.

520. Sölétormos G, Duffy MJ, Hayes DF, Sturgeon CM, Barak V, Bossuyt PM et al. Design of tumor biomarker-monitoring trials: a proposal by the European Group on Tumor Markers. Clin Chem 2013; 59(1): 52-59.

521. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009; 338: b1147.

522. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010; 14(8): 1-193.

523. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. J R Stat Soc Ser A 1994; 157(3): 357-416.

524. Spiegelhalter DJ, Myles JP, Jones D, Abrams KR. Methods in health service research: an introduction to Bayesian methods in health technology assessment. BMJ 1999; 319(7208): 508-512.

525. Statistisches Bundesamt. Preise: Harmonisierter Verbraucherpreisindex; September 2013 [online]. 11 October 2013 [accessed: 17 October 2013]. URL: https://www.destatis.de/DE/Publikationen/Thematisch/Preise/Verbraucherpreise/HarmonisierteVerbraucherpreisindizesPDF_5611201.pdf?__blob=publicationFile.

526. Statistisches Bundesamt. Statistik der schwerbehinderten Menschen 2007: Kurzbericht [online]. January 2009 [accessed: 18 March 2015]. URL: https://www.destatis.de/DE/Publikationen/Thematisch/Gesundheit/BehinderteMenschen/SozialSchwerbehinderteKB5227101079004.pdf?__blob=publicationFile.

527. Steckelberg A, Berger B, Köpke S, Heesen C, Mühlhauser I. Kriterien für evidenzbasierte Patienteninformationen. Z Arztl Fortbild Qualitatssich 2005; 99(6): 343-351.

528. Steiner JF. The use of stories in clinical research and health policy. JAMA 2005; 294(22): 2901-2904.

529. Sterne J, Egger M, Smith GD. Systematic reviews in health care: investigating and dealing with publication and other biases in meta-analysis. BMJ 2001; 323(7304): 101-105.

530. Sterne JAC, Egger M, Moher D. Addressing reporting biases. In: Higgins JPT, Green S (Ed). Cochrane handbook for systematic reviews of interventions. Chichester: Wiley; 2008. p. 297-333.

531. Steyerberg EW, Moons KG, Van der Windt DA, Hayden JA, Perel P, Schroter S et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013; 10(2): e1001381.

532. Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making 1998; 18(2 Suppl): S68-S80.

533. Stollenwerk B, Lhachimi SK, Briggs A, Fenwick E, Caro JJ, Siebert U. Communicating the parameter uncertainty in the IQWiG efficiency frontier to decision-makers. Health Econ 04.03.2014 [Epub ahead of print].

534. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA 2000; 283(15): 2008-2012.

535. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996; 49(8): 907-916.

536. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010; 340: c117.

537. Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008; 26(9): 753-767.

538. Sutton AJ, Donegan S, Takwoingi Y, Garner P, Gamble C, Donald A. An encouraging assessment of methods to inform priorities for updating systematic reviews. J Clin Epidemiol 2009; 62(3): 241-251.

539. Swift TL, Dieppe PA. Using expert patients' narratives as an educational resource. Patient Educ Couns 2005; 57(1): 115-121.

540. Tainio M, Tuomisto JT, Hänninen O, Ruuskanen J, Jantunen MJ, Pekkanen J. Parameter and model uncertainty in a life-table model for fine particles (PM2.5): a statistical modeling study. Environ Health 2007; 6: 24.

541. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: a systematic review and classification. Clin Cancer Res 2013; 19(17): 4578-4588.

542. Takwoingi Y, Leeflang MMG, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013; 158(7): 544-554.

543. Tan SS. Microcosting in economic evaluations: issues of accuracy, feasibility,consistency and generalisability [Dissertation]. Rotterdam: Erasmus Universität; 2009. URL: http://repub.eur.nl/res/pub/17354/091127_Tan,%20Siok%20Swan.pdf.

544. Tan SS, Bouwmans CAM, Rutten FFH, Hakkaart-van Roijen L. Update of the Dutch manual for costing in economic evaluations. Int J Technol Assess Health Care 2012; 28(2): 152–158.

545. Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R et al. Integrating qualitative research with trials in systematic reviews. BMJ 2004; 328(7446): 1010-1012.

546. Thomas S. Klinische Relevanz von Therapieeffekten: systematische Sichtung, Klassifizierung und Bewertung methodischer Konzepte [Dissertation]. Duisburg/Essen: Universität; 2009.

547. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21(11): 1559-1573.

548. Thorlund K, Devereaux PJ, Wetterslev J, Guyatt G, Ioannidis JPA, Thabane L et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009; 38(1): 276-286.

549. Thorne S. The role of qualitative research within an evidence-based context: can metasynthesis be the answer? Int J Nurs Stud 2009; 46(4): 569-575.

550. Thurow S. Search engine visibility. Indianapolis: New Riders; 2003.

551. Tian L, Cai T, Pfeffer MA, Piankov N, Cremieux PY, Wei LJ. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 × 2 tables with all available data but without artificial continuity correction. Biostatistics 2009; 10(2): 275-281.

552. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996; 58(1): 267-288.

553. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 2007; 8: 16.

554. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ 2001; 322(7282): 355-357.

555. Torrance GW, Siegel JE, Luce BR, Gold MR, Russell LB, Weinstein MC. Framing and designing the cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC (Ed). Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 54-81.

556. Trevena LJ, Davey HM, Barratt A, Butow P, Caldwell P. A systematic review on communicating with patients about evidence. J Eval Clin Pract 2006; 12(1): 13-23.

557. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA 2009; 301(8): 831-841.

558. Trueman P, Drummond M, Hutton J. Developing guidance for budget impact analysis. Pharmacoeconomics 2001; 19(6): 609-621.

559. Tsay MY, Yang YH. Bibliometric analysis of the literature of randomized controlled trials. J Med Libr Assoc 2005; 93(4): 450-458.

560. Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. Science 1977; 198(4318): 679-684.

561. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003; 290(12): 1624-1632.

562. Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000; 19(24): 3417-3432.

563. Tyczynski JE, Demaret E, Parkin DM (Ed). Standards and guidelines for cancer registration in Europe: the ENCR recommendations vol.1. Lyon: IARC Press; 2003. (IARC Technical Publications; Volume 40).

564. UK National Screening Committee. Programme appraisal criteria: criteria for appraising the viability, effectiveness and appropriateness of a screening programme [online]. [Accessed: 18 March 2015]. URL: http://www.screening.nhs.uk/criteria.

565. USAID Center for Development Information and Evaluation. Conducting key informant interviews [online]. 1996 [accessed: 18 March 2015]. (Performance Monitoring and Evaluation TIPS; Volume 2). URL: http://pdf.usaid.gov/pdf_docs/PNABS541.pdf.

566. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21(4): 589-624.

567. Van Tinteren H, Hoekstra OS, Boers M. Do we need randomised trials to evaluate diagnostic procedures? Eur J Nucl Med Mol Imaging 2004; 31(1): 129-131.

568. Van Tinteren H, Hoekstra OS, Smit EF, Van den Bergh JH, Schreurs AJ, Stallaert RA et al. Effectiveness of positron emission tomography in the preoperative assessment of patients with suspected non-small-cell lung cancer: the PLUS multicentre randomised trial. Lancet 2002; 359(9315): 1388-1393.

569. Van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976) 2003; 28(12): 1290-1299.

570. Vandenbroucke JP, Von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007; 147(8): W163-W194.

571. Veerman JL, Mackenbach JP, Barendregt JJ. Validity of predictions in health impact assessment. J Epidemiol Community Health 2007; 61(4): 362-366.

572. Vidanapathirana J, Abramson MJ, Forbes A, Fairley C. Mass media interventions for promoting HIV testing. Cochrane Database Syst Rev 2005; (3): CD004775.

573. Vijan S. Should we abandon QALYs as a resource allocation tool? Pharmacoeconomics 2006; 24(10): 953-954.

574. Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. Stat Med 2001; 20(23): 3635-3647.

575. Virtanen H, Leino-Kilpi H, Salantera S. Empowering discourse in patient education. Patient Educ Couns 2007; 66(2): 140-146.

576. Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters M et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions: Agency for Healthcare Research and quality methods guide for comparative effectiveness reviews [online]. March 2012 [accessed: 18 March 2015]. URL: http://effectivehealthcare.ahrq.gov/ehc/products/322/998/MethodsGuideforCERs_Viswanathan_IndividualStudies.pdf.

577. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care 2005; 17(3): 235-242.

578. Voisin CE, De la Varre C, Whitener L, Gartlehner G. Strategies in assessing the need for updating evidence-based guidelines for six clinical topics: an exploration of two search methodologies. Health Info Libr J 2008; 25(3): 198-207.

579. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007; 335(7624): 806-808.

580. Waffenschmidt S, Janzen T, Hausner E, Kaiser T. Simple search techniques in PubMed are potentially suitable for evaluating the completeness of systematic reviews. J Clin Epidemiol 2013; 66(6): 660-665.

581. Walter U, Schwartz FW. Prävention. In: Schwartz FW, Badura B, Busse R, Leidl R, Raspe H, Siegrist J et al (Ed). Das Public Health Buch: Gesundheit und Gesundheitswesen. München: Urban und Fischer; 2003. p. 189-214.

582. Watine J, Friedberg B, Nagy E, Onody R, Oosterhuis W, Bunting PS et al. Conflict between guideline methodologic quality and recommendation validity: a potential problem for practitioners. Clin Chem 2006; 52(1): 65-72.

583. Web Accessibility Initiative. Webauftritt [online]. [Accessed: 18 March 2015]. URL: http://www.w3.org/WAI.

584. Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C et al. Principles of good practice for decision analytic modeling in health care evaluation: report of the ISPOR Task Force on Good Research Practices; modeling studies. Value Health 2003; 6(1): 9-17.

585. Weinstein MC, Siegel JE, Garber AM, Lipscomb J, Luce BR, Manning WG et al. Productivity costs, time costs and health-related quality of life: a response to the Erasmus Group. Health Econ 1997; 6(5): 505-510.

586. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med 2006; 25(2): 183-203.

587. Wendt C. Gesundheitssysteme im internationalen Vergleich. Gesundheitswesen 2006; 68(10): 593-599.

588. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF et al. Systems to rate the strength of scientific evidence: AHRQ publication no. 02-E016 [online]. March 2002 [accessed: 16 April 2014]. (Evidence Report/Technology Assessment (Summaries); Volume 47). URL: http://archive.ahrq.gov/clinic/epcsums/strengthsum.pdf.

589. Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008; 61(1): 64-75.

590. Whitehead J. The design and analysis of sequential clinical trials. Chichester: Horwood; 1983.

591. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140(3): 189-202.

592. Whiting PF, Rutjes AW, Westwood ME, Mallett S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013; 66(10): 1093-1104.

593. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8): 529-536.

594. Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008; 148(10): 776-782.

595. Windeler J. Bedeutung randomisierter klinischer Studien mit relevanten Endpunkten für die Nutzenbewertung [online]. In: Gesundheitsforschungsrat des Bundesministeriums für Bildung und Forschung (Ed). Diskussionsforum zur Nutzenbewertung im Gesundheitswesen: Begriffsdefinitionen und Einführung; Dokumentation des ersten gemeinsamen Workshops von GFR und IQWiG am 4. September 2007 in Berlin. December 2007. S. 26-31 [accessed: 18 March 2015]. URL: http://www.gesundheitsforschung-bmbf.de/_media/DLR_Nutzenbewert_07-11-22_Druckversion.pdf.

596. Windeler J. Externe Validität. Z Evid Fortbild Qual Gesundhwes 2008; 102(4): 253-259.

597. Windeler J, Conradt C. Wie können "Signifikanz" und "Relevanz" verbunden werden? Med Klin 1999; 94(11): 648-651.

598. Windeler J, Lange S. Nutzenbewertung in besonderen Situationen: seltene Erkrankungen. Z Evid Fortbild Qual Gesundhwes 2008; 102(1): 25-30.

599. Windeler J, Ziegler S. Evidenzklassifizierungen. Z Arztl Fortbild Qualitatssich 2003; 97(6): 513-514.

600. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008; 336(7644): 601-605.

601. Ziebland S, McPherson A. Making sense of qualitative data analysis: an introduction with illustrations from DIPEx (personal experiences of health and illness). Med Educ 2006; 40(5): 405-414.

602. Ziegler DK, Mosier MC, Buenaver M, Okuyemi K. How much information about adverse effects of medication do patients want from physicians? Arch Intern Med 2001; 161(5): 706-713.

603. Zschorlich B, Knelangen M, Bastian H. Die Entwicklung von Gesundheitsinformationen unter Beteiligung von Bürgerinnen und Bürgern am Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG). Gesundheitswesen 2011; 73(7): 423-429.

604. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008; 337: a2390.