General Methods[a]

Version 7.0
of 19 September 2023

[a] This translation is based on the German document *Allgemeine Methoden* (Version 7.0) of 19 September 2023. 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:

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
Im Mediapark 8
50670 Köln
Germany

Tel.: +49 221 35685-0
Fax: +49 221 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 6.1 of the Institute's General Methods of 24 January 2022, in Version 7.0, minor errors were corrected, editorial changes were made, and a literature update was conducted. The following changes to content and structure were made:

- Renaming of the product "HTA report" to "ThemenCheck report".
- Fundamental revision of Section 1.3 on evidence-based medicine
- Fundamental revision of Section 1.4 on health economics
- Presentation of the extended product range with regard to assessments according to §35a in Chapter 2
- Fundamental revision and renaming of Section 3.2.1 on dealing with possible reporting bias
- Addition to Section 3.3.4 on describing the content of concepts for routine practice data collection according to §35a (3b) SGB V (RPDC concepts)
- Fundamental revision and expansion of Section 3.5 on diagnostic tests
- Fundamental revision of Chapter 4 on health economic evaluations of medical interventions
- Change of instrument for the quality assessment of systematic reviews in Chapter 7 (AMSTAR 2 instead of Oxman-Guyatt Index)
- Additions to Section 8.1.1 on the use of machine learning
- Addition of the new EMA database "Clinical Trials Information System" in Section 8.1.2
- Additions and deletions in Section 8.2 on standards for focused information retrieval
- Additions to Section 8.4 on focused information retrieval for guidelines
- Revision of Section 9.1.1 on criteria for the inclusion of studies
- Revision and renaming of Section 9.3.2 on statistical significance and confidence intervals
- Additions to Section 9.3.6 on adjustment for confounders in the context of routine practice data collection
- Additions to Section 9.3.7 B on the assessment of heterogeneity in meta-analyses
- Additions to Section 9.3.12 on the effect of systematically shortened observation periods in the recording of patient-reported outcomes (PROs)

Table of contents

List of tables

Page

List of figures

Page

List of abbreviations

Abbreviation	Definition
AdViSHE	Assessment of the Validation Status of Health-Economic Decision Models
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 (Association of the Scientific Medical Societies)
BIA	budget impact analysis
CA	conjoint analysis
CENTRAL	Cochrane Central Register of Controlled Trials
CERQual	Confidence in the Evidence from Reviews of Qualitative Research
CONSORT	Consolidated Standards of Reporting Trials
CTIS	Clinical Trials Information System
DMP	disease management programme
DMP-A-RL	DMP-Anforderungen-Richtlinie (DMP Requirements Directive)
DRG	diagnosis-related group
EBHC	evidence-based health care
EBM	evidence-based medicine
EMA	European Medicines Agency
EU-CTR	EU Clinical Trials Register
EUnetHTA	European network for Health Technology Assessment
FDA	Food and Drug Administration
G-BA	Gemeinsamer Bundesausschuss (Federal Joint Committee)
GoR	grade of recommendation
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HEE	health economic evaluation
HTA	health technology assessment
ICER	incremental cost-effectiveness ratio
ICTRP	International Clinical Trials Registry Platform Search Portal
IPD	individual patient data
IQR	interquartile range

Abbreviation	Definition
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
MedDRA	Medical Dictionary for Regulatory Activities
MID	minimally important difference
NHB	net health benefit
NICE	National Institute for Health and Care Excellence
NUB	Neue Untersuchungs- und Behandlungsmethoden (new examination and treatment methods)
NVL	Nationale VersorgungsLeitlinie (National Health Care Guideline)
OECD	Organization for Economic Co-operation and Development
PICO	population, intervention, comparison, outcome
PRESS	Peer Review of Electronic Search Strategies
PRISMA(-S)	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (-for Searching)
PRO	patient-reported outcome
PSA	probabilistic sensitivity analysis
QALY	quality-adjusted life year
RECORD	Reporting of Studies Conducted Using Observational Routinely Collected Health Data
RECORD-PE	RECORD for Pharmacoepidemiology
RCT	randomized controlled trial
ROBIS	Risk of Bias in Systematic Reviews
RPD	routine practice data
RPDC	routine practice data collection
SG	standard gamble
SGB V	Sozialgesetzbuch – Fünftes Buch – Gesetzliche Krankenversicherung (Social Code Book – Book V – Statutory Health Insurance)
SHI	statutory health insurance
SMD	standardized mean difference
STE	surrogate threshold effect
TTO	time trade-off
VAE	Visual Analogue Scale
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 [2] 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[1] V) [8] 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
- search for evidence on the current state of medical knowledge as a basis for the development or further development of clinical practice guidelines
- appraisal of evidence-based clinical practice guidelines 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
- involvement in international projects on the collaboration and further development in the field of evidence-based medicine (EBM)

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

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[2]) or the Federal Ministry of Health[3] 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. According to §139b (6) SGB V, this option does not apply to the commissioning of evidence searches for the development or updating of clinical practice guidelines.

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 2.2.3).

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 G-BA may withdraw, suspend or change commissions to the Institute. The procedure is described in Chapter 1, §16b (3) of the Rules of Procedure of the G-BA [287]. Accordingly, these decisions are taken in consultation with the Institute.

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[4]), 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 [4]. 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 [119], which has also been effective since the beginning of 2011, and the G-BA's Code of Procedure [287].

[2] 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 https://www.g-ba.de/english/.

[3] Bundesministerium für Gesundheit, BMG

[4] Arzneimittelmarktneuordnungsgesetz

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 SHI-insured community 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, and reduction in disease duration and adverse effects. The definition of a "patient-relevant 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 [287]. 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.

With the Promotion of Health Care Act, in 2015 an assessment of new examination and treatment methods (NUB[5]) with high-risk medical devices was introduced via §137h SGB V. This assessment refers to methods that a) follow a new theoretical-scientific concept [121,287], b) are particularly invasive [121,287], and c) lead to a first request according to §6 of the Hospital Reimbursement Act ("NUB request"). The G-BA receives documents on such methods from hospitals and medical device manufacturers. The G-BA generally commissions the Institute to assess documents according to §137h SGB V in respect of whether a benefit or harmfulness or ineffectiveness can be recognized in them.

Due to the Promotion of Health Care Act, in 2015 §139b SGB V was extended by Paragraph 5. §139b (5) SGB V specifies that insured persons and other interested individuals can propose assessments of medical examination and treatment methods for selected diseases as well as assessments of questions of the quality and efficiency of services provided within the framework of SHI. According to §139b (5) SGB V, this excludes topic proposals where the focus

[5] Neue Untersuchungs- und Behandlungsmethoden

is on the separate assessment of a drug. It is IQWiG's task to select topics from these proposals that are particularly important for the health care of patients and for which reports commissioned by IQWiG ("ThemenCheck reports") are then produced (see Chapter 6).

With the Digital Health Care Act (DVG[6]), §139b (3) SGB V was supplemented with the new No. 3 and §139b SGB V with Section 6. This passage stipulates that the Association of the Scientific Medical Societies (AWMF[7]) can propose clinical practice guidelines to the Federal Ministry of Health for which IQWiG is to support guideline development or updating with evidence searches.

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 1.3. The term "health economics" and the underlying concept are described in detail in Section 1.4.

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 2.2.5.

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.

[6] Digitale-Versorgung-Gesetz

[7] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften

1.2 General commission of the G-BA

In December 2004, the G-BA awarded IQWiG a so-called general commission [288]. With the general commission, the Institute was commissioned "to continuously monitor and assess medical developments of fundamental importance and their effects on the quality and efficiency of medical care in Germany by identifying and assessing the relevant literature, and to inform the G-BA regularly about these developments. With this commission, the G-BA assumes that the Institute will not only process individual commissions from the G-BA in the fields of work assigned to the Institute under §139a (3) SGB V, but will also, on the basis of its independent scientific work, provide the G-BA with information on medical developments relevant to health care that is necessary for the G-BA's statutory tasks, and will develop concrete proposals for individual commissions that the Institute considers relevant against the background of this information."

In July 2006 and March 2008, the general commission was further specified and adapted with regard to the production of health information [289]. IQWiG "produces and publishes information on its own responsibility within the framework of its scientific work, without the need for an individual commission in each case. The Institute for Quality and Efficiency in Health Care is solely responsible for the content of the information, including any consequences arising from it."

1.3 Evidence-based medicine

Evidence-based medicine (EBM) is an approach to answering health questions that does not rely solely on opinions and consensus, but is based on evidence – information that has been collected using scientific methods that are as objective as possible. EBM is based on a collection of different scientific tools and strategies that aim to protect against wrong decisions and false expectations and to provide reliable knowledge. It includes 2 complementary components [254,321,322]:

- The planning, conduct and publication of meaningful scientific studies, particularly to compare different treatment options.
- The systematic search for and assessment of studies relevant to a research question.

However, tools for the objective assessment of medical interventions were not only invented with the introduction of the term "evidence-based medicine", but their roots go back much further. In Germany, for example, Paul Martini described the main components of a fair assessment of the efficacy of drugs in his monograph "Methodology of Therapeutic Studies" in 1932 [518]. In the early 1960s, the method of randomly assigning study participants to comparator groups (randomization) in order to assess the efficacy and safety of medical interventions increasingly became the internationally accepted standard [365]. 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 [45]. About 20 years later, clinical epidemiologists attempted to establish this

methodology in clinical practice [261]. 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 then, clinical studies and their systematic identification and assessment (systematic reviews) have formed the basis of the international scientific standard for assessing the benefits and harms of medical interventions in the context of health technology assessments (HTAs) [44,563].

Although EBM requires the use of standard methods, it is not a rigid concept. Which standard to use and when to use it depends on the question to be answered and the decision to be made. In addition, decisions for which there are (as yet) no international specifications have to be made repeatedly in the search for, processing, and assessment of studies. EBM also includes the requirement to define one's own specifications in such situations. However, this is linked to the obligation to define such specifications preferably a priori, and to explain them in a transparent way.

This scientific approach has become the basis of the provision of health care in Germany [37,287,557]. The systematically processed evidence can then be used for decision-making at various levels, in particular:

- decisions about the treatment of individual patients (practical EBM)
- decisions about the design of the health care system (evidence-based health care, EBHC [527])

1.3.1 Practical evidence-based medicine

Practical EBM supports doctors who want to find the most meaningful alternatives among possible interventions for their patients and present the prospects of success in an objective way. This knowledge is then the basis for reaching a joint decision with patients.

This implementation of EBM in daily clinical practice for individual patients was defined by David Sackett et al. in 1996 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" [631].

However, the "best available evidence" is often incomplete or unreliable. EBM has developed tools to assess uncertainty. In this way, EBM helps doctors and patients to recognize the type and degree of uncertainty. They can then discuss how to make a decision under uncertainty. It is precisely in uncertain situations that personal preferences determine which option patients choose. A decision is therefore evidence-based if it combines clinical expertise with 3 components: the evidence, the clinical condition and circumstances of the individual patient, and his or her preferences and attitudes [354].

1.3.2 Evidence-based health care

In evidence-based health care (EBHC), the principles of EBM are applied to all areas of health care – including decisions about the design of health care systems [527]. Besides clinical effectiveness and safety, health economic, organizational, ethical, social and legal aspects can also be considered in the context of HTA (see Chapter 6) [563].

In Germany, overarching decisions on the design of health care are largely in the hands of the Institute's contracting agencies. The Institute's main task is therefore to provide the most reliable answers possible to the questions of its contracting agencies about the benefits and harms of a medical intervention. This Institute's remit is therefore deliberately not aimed towards treating individual patients with their possible specific characteristics, but towards determining which patient groups are likely to benefit from an intervention.

The reports of the Institute serve, for example, as a recommendation to the G-BA for decisions that apply in principle to all persons covered by statutory health insurance (SHI). In its decisions, the G-BA in turn then considers, among other things, health care aspects that are beyond the scope of a benefit assessment [287]. In this way, the results of the Institute's work feed into the design of health care.

1.3.3 Basic EBM strategies

A characteristic standard component of EBM is the structured and systematic approach to finding answers to health questions:

- The question must be precisely formulated. Medicine is often about choosing between 2 alternatives. This can apply to treatments, complex lifestyle changes as well as diagnostic procedures and screening tests. This always leads to the following question: Is Option A better than Option B?
- It must be defined how the benefit (or harm) of a medical intervention (whether it is therapeutic, diagnostic, preventive or rehabilitative) is to be measured (see also Section 3.1.1). A standard component of EBM is the question of relevant consequences for patients: Can life be extended? Can symptoms and quality of life be improved? Can disabling side effects be reduced?
- To answer these questions as reliably as possible, it is necessary to systematically search for and assess the available information on clinical studies. This means that all studies on a particular question that are of appropriate quality in terms of their design and conduct are identified and summarized in a transparent process. If large differences are found between the results of individual studies (so-called heterogeneity), an attempt must be made to explain these differences. The results of these summaries and assessments are called systematic reviews; the statistical analyses are called meta-analyses.

- It is explicitly stated in EBM that only probability statements or conclusions about groups of patients are usually possible with regard to the benefit (and harm) of medical interventions. Benefit is demonstrated by increasing the probability of a favourable outcome and/or reducing the risk of an unfavourable outcome. To show this, studies are needed in sufficiently large groups of suitable patients.

1.3.4 The relevance of certainty of results

A specific feature of EBM is that it allows an evaluation of the extent to which the available evidence is reliable. The certainty of results of clinical studies depends primarily on 3 aspects:

- Internal validity, which is determined by the chosen study design and the actual conduct of the study.
- External validity, which depends, for example, on the choice of interventions and the selection of the participating populations.
- Statistical precision, which is influenced, for example, by the size of a study and the reliability of outcome recording.

Assessing the certainty of results and describing knowledge gaps therefore play a key role in the Institute's reports. Many details on how studies are planned, conducted, analysed, and published have an impact on the reliability of the available results.

It is an international EBM standard to critically examine and assess these aspects. This includes the following steps:

1) For every question, it is an international EBM standard to specify and preferably consider the study type (measurement tool) that minimizes the risk of unjustifiably discriminating against one of the alternatives. The Institute's recommendations are therefore normally based only on studies with sufficient certainty of results. This ensures that the decisions that are based on the Institute's recommendations are supported by a sound scientific foundation.
2) If the situation arises that studies with the required validity and precision are generally lacking, it is the Institute's core task to describe and classify the situation.

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 [312]. 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.

1.4 Health economics

In a broader sense, the term health economics refers to "the analysis of economic aspects of the health care system using concepts of economic theory" [655]. This includes concepts from microeconomic behavioural theory, competition theory, economic theory of politics, and management theory [655]. The subject of such a study could be, for example, how stakeholders in the health care system change their behaviour after the introduction of incentives (e.g. co-payments). The extent to which such studies can and should be used to steer the health care system can be discussed from both a methodological and an ethical perspective, but is beyond the scope of this brief presentation.

In a narrower sense, health economics is understood here as health economic evaluations. These analyses aim, among other things, to inform decision makers about the cost-effectiveness ratios of interventions and thus represent one area of HTA, alongside the benefit assessment of interventions [309].

In health economic terminology, the term cost-benefit analysis refers to an analysis in which both the costs and the benefits are expressed in monetary units. This type of analysis is relatively rarely used. In SGB V, the term "cost-benefit assessment" (German: *Kosten-Nutzen-Bewertung*) is therefore intended more as an umbrella term for comparative health economic analyses in general. We therefore use the term "health economic evaluation" (HEE) in the following text. The relevant types of analysis are primarily cost-effectiveness analyses and cost-utility analyses. In English-speaking countries, the latter is often considered a subtype of cost-effectiveness analyses [210]. In the following, the term "cost-effectiveness ratio" will therefore be used consistently as the outcome of HEEs. Budget impact analyses (BIAs), which are often conducted in addition to one of the types of HEEs, are also discussed [147].

With the establishment of the Institute in 2004, the G-BA and the Federal Ministry of Health were free to commission an HEE. With the SHI Act to Promote Competition[8] of 2007, the HEE of drugs was anchored in §35b SGB V in order to obtain information on the recommendation of 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 compared to other drugs and forms of treatment in a therapeutic indication. A prerequisite for the commissioning of an HEE was therefore that the added benefit of a drug had been demonstrated.

The German Act on the Reform of the Market for Medicinal Products (AMNOG[9]), which became effective on 1 January 2011, shifted the relevance of the HEE within the procedure for the early benefit assessment of drugs. An HEE is primarily intended for cases in which negotiations on the reimbursement price (in short: "price negotiations") between the SHI

[8] Gesetzliche-Krankenversicherung(GKV)-Wettbewerbsstärkungsgesetz

[9] Arzneimittelmarktneuordnungsgesetz

umbrella organization[10] and the pharmaceutical companies fail and no agreement is reached in the subsequent arbitration procedure. According to §35b (1) Sentence 4 SGB V in conjunction with Chapter 5 §32 (3) of the G-BA's Code of Procedure [287], the economic evaluation must take into account the appropriateness and reasonableness of cost coverage by the SHI-insured community [7,287]. In order for the G-BA to take these factors into account appropriately, it must be provided with the relevant information. This information is provided by the HEE (appropriateness) and the budget impact analysis (BIA) (reasonableness). The assessment of the appropriateness and reasonableness of cost coverage is based on whether there is a justifiable relationship between the costs and benefits of the drug, taking into account the principle of proportionality. In this context, according to the Chapter 5, §32 (2,3) of the G-BA's Code of Procedure, IQWiG should submit a recommendation on the basis of which the G-BA should make a decision [7,287].

The aim of the HEE is to summarize economic information to complement the benefit assessment, particularly for price negotiations. The explicit recommendation of an appropriate price based on the consideration of all available comparators within an entire therapeutic indication is not the focus. The aim of an HEE is not to infer conclusions about the added benefit of an intervention. Because HEEs use decision-analytic modelling and therefore extrapolations, it is not possible to achieve the certainty of conclusions required to conclude that there is proof, an indication or a hint of added benefit (see Section 3.1.4). The benefit assessment using EBM methods provides a basis for subsequent HEEs. The HEE is concerned with the depiction of possible scenarios. This requires a moderate degree of integration of different data sources [713]. Overall, the aim of the HEE is to provide the decision-maker with additional information beyond the benefit assessment to inform price negotiations [551]. However, it should not be seen as a simple supplement in the form of costs, but as an independent analysis with its own objectives and methods. In the context of an HEE, among other factors, longer-term economic consequences and other cost components beyond drug costs can also be considered [551].

The use of a universal threshold is not intended, as this is not the subject of IQWiG's assessment of scientific information; moreover, there is no generally accepted threshold for Germany [667]. Therefore, no dichotomous conclusion on cost-effectiveness can be drawn and no single decisive price recommendation can be given. The decision-maker is responsible for weighing up the various pieces of information and assessing affordability.

[10] Spitzenverband Bund der Krankenkassen, GKV-Spitzenverband (National Association of SHI Funds)

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
- assessments according to §35a SGB V (dossier assessment, concept for routine practice data collection [RPDC concept], search for disease registries, estimation of patient numbers)
- dossier assessment
- HEE according to §35b SGB V
- assessment of potential
- assessment according to §137h SGB V
- addendum
- health information
- working paper
- ThemenCheck report
- evidence report

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 Section 1.1). The main difference between reports and rapid reports is that a public commenting procedure (hearing) is 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.

§35a SGB V regulates the assessment of the benefit of drugs with new active ingredients on the basis of a dossier submitted by the pharmaceutical company. These assessments include

dossier assessments (see Section 3.3.3), RPDC concepts (see Section 3.3.4), searches for disease registries to prepare the G-BA's decision-making process with regard to the evaluation of the need to require an RPDC, as well as estimations of patient numbers (see Section 2.1.3.4). The G-BA is the contracting agency in each case. No hearings by the Institute are planned for assessments according to §35a SGB V. For dossier assessments and RPDC concepts, a commenting procedure is conducted in the further process 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 IQWiG conducts hearings. A further commenting procedure is conducted at the G-BA.

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

Assessments according to §137h SGB V are commissioned by the G-BA and refer to new examination and treatment methods with high-risk medical devices. No hearing by the Institute is intended. If a directive is decided on, the G-BA conducts a commenting procedure in this regard.

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 1.2), 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.

ThemenCheck reports are produced on topics proposed by insured persons or other interested individuals. The Institute selects topics from the proposals that are particularly important for the health care of patients in Germany. In this context, the perspectives of both consumers and patients as well as the scientific perspective are considered. ThemenCheck

reports on the selected topics are produced. This is based on §139b (5) SGB V. A hearing by the Institute is intended for the ThemenCheck reports.

An evidence search according to §§139 (3) No. 3, 139b (6) SGB V includes the production of several evidence reports to support the development or updating of a clinical practice guideline by the AWMF and is commissioned by the Federal Ministry of Health. There is no provision for a commenting procedure.

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 2.1.11.

Table 1: Overview of the Institute's products

Product	Objective	Procedure	Commissioning
Report	Recommendations on tasks described in §139a SGB V, including hearing	Section 2.1.1	G-BA, Federal Ministry of Health
Rapid report	Recommendations on tasks described in §139a SGB V, insofar as no hearing is required; in particular provision of information at short notice on current topics	Section 2.1.2	G-BA, Federal Ministry of Health
Assessments acc. to §35a SGB V ▪ Dossier assessment	▪ Assessment of the benefit (as well as information on epidemiology and costs) of drugs with new active ingredients according to §35a SGB V	Section 2.1.3	G-BA
▪ RPDC concept	▪ Concept for routine practice data collection acc. to §35a (3b) SGB V		
▪ Search for disease registries ▪ Estimation of patient numbers	▪ Preparation of the G-BA's decision-making process regarding the evaluation of the need to require RPDC acc. to §35a (3b) SGB V		
Health economic evaluation acc. to §35b SGB V	Assessment of the relation of the cost and benefit of drugs acc. to §35b SGB V, including a hearing	Section 2.1.4	G-BA
Assessment of potential	Assessment of the potential of new examination and treatment methods acc. to §137e SGB V	Section 2.1.5	G-BA
Assessment acc. to §137h SGB V	Assessment of the benefit, harmfulness or ineffectiveness of new examination and treatment methods with high-risk medical devices acc. to §137h SGB V	Section 2.1.6	G-BA
Addendum	Supplementary information provided at short notice by the Institute on issues that have arisen during the consultation on its completed products	Section 2.1.7	G-BA, Federal Ministry of Health
Health information	Easily understandable information for consumers and patients; wide scope of topics	Section 2.1.8	General commission of the G-BA
Working paper	Information on relevant developments in health care or methodological aspects	Section 2.1.9	General commission of the G-BA
ThemenCheck report	Assessment of medical examination and treatment methods acc. to §139b (5) SGB V, including a hearing	Section 2.1.10	Institute's initiation based on proposals of interested individuals
Evidence report	Evidence presentation acc. to §§139a (3) No. 3, 139b (6) SGB V	Section 2.1.11	Federal Ministry of Health
G-BA: Gemeinsamer Bundesausschuss (Federal Joint Committee); RPDC: routine practice data collection; SGB V: Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung (Social Code Book V – Statutory Health Insurance)			

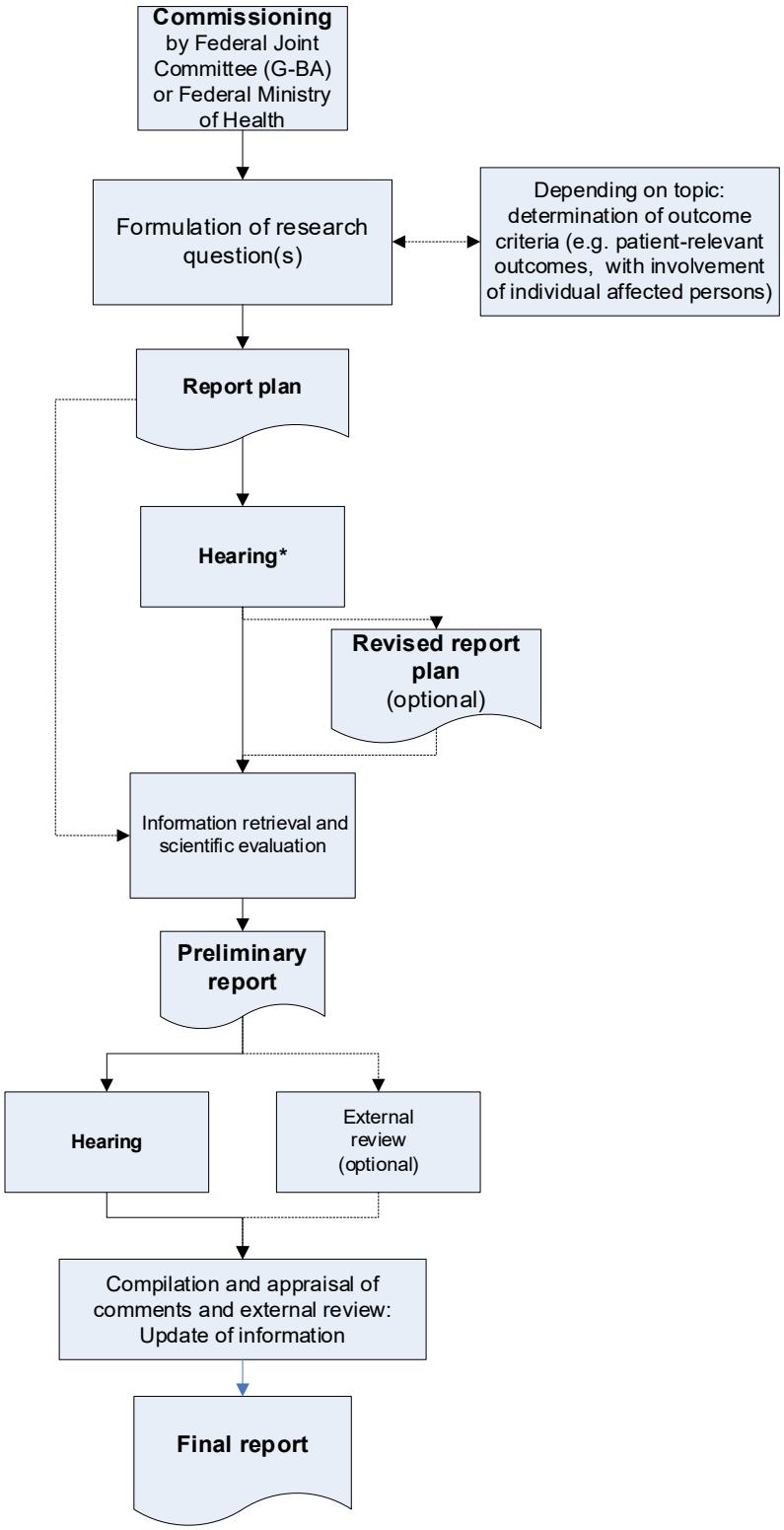
2.1.1 Report

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 2.2.2). 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, affected persons are involved, especially in the definition of patient-relevant outcomes. Affected persons may include patients (potentially represented by their parents or other relatives) as well as potential participants in prevention measures. Subsequently, the report plan (protocol) is prepared.

The **protocol (called report plan)** forms the basis for the production of the preliminary report and 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 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.

A **hearing** on the report plan is conducted according to the legal regulations in §139a (5) SGB V. The hearing is conducted by means of written comments, which can be submitted within a period of at least 4 weeks from the time of notification on the Institute's website. The hearing 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 hearing. The comments are analysed and published in order to document the hearing. If no change in the methodology of the report is required, the documentation of the hearing is published together with the preliminary report. If a change in the methodology of the report is required, a revised new version of the report plan is prepared, which is first forwarded to the contracting agency, the Foundation's Board of Directors, the Foundation Council and Board of Trustees, together with the documentation of the hearing on the report plan. This document is usually published on the Institute's website 5 working days later. The revised version of the report plan is the basis for the preparation of the preliminary report.



*A hearing procedure for the report plan is conducted according to the legal regulations in §139a (5) SGB V.

Figure 1: Procedure for the production of a report

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 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 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.

The preliminary report is subject to a **hearing**. As a matter of principle, the hearing is conducted by means of written comments, which can be submitted within a period of at least 4 weeks from the time of notification on the Institute's website. 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. The hearing in particular refers to the results of the retrieval and assessment of information presented in the preliminary 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, the Foundation's Board of Directors, the Foundation Council as well as 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.

If an examination or treatment method is assessed according to §139a (3) No. 1 SGB V, the final report is to be completed within one year after the award of the commission, according to §4 of the Regulation on Procedures for the Assessment of Methods (MBVerfV[13]) [120].

[13] Methodenbewertungsverfahrensverordnung

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 Institute's responsibility, involving external experts where appropriate (see Section 2.2.2). If necessary, the Institute's Scientific Advisory Board is also involved. The internal quality assurance process is not presented in this flow chart.

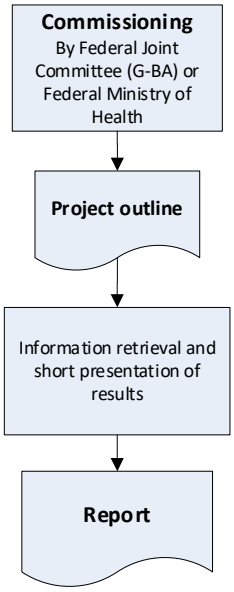


Figure 2: Procedure for the production of a rapid report

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. Therefore, no hearings are conducted during the course of the project.

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, the opinion of affected individuals can be 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 2.2.4) 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 Assessments according to §35a SGB V

2.1.3.1 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 2.2.2). If necessary, the Institute's Scientific Advisory Board is also involved. The internal quality assurance process is not outlined in this flow chart.

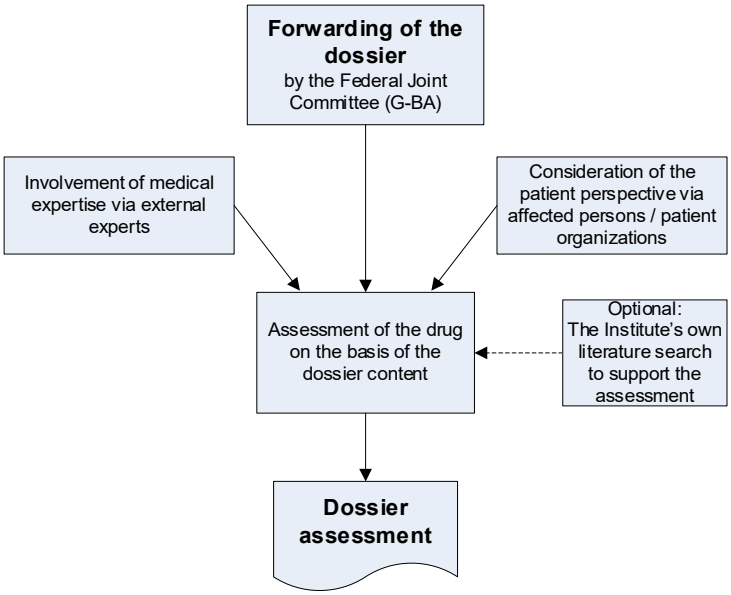


Figure 3: Procedure for the production of a dossier assessment

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 affected persons/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 2.2.2).

The **patient perspective is considered** on the basis of a questionnaire sent to affected persons/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. Affected persons/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 the Foundation's Board of Directors, the Foundation Council and the Foundation's Board of Trustees are informed about the dispatch and the dossier assessment is published on the Institute's website.

2.1.3.2 RPDC concept

The procedure for developing an RPDC concept is shown in Figure 4. All working steps are performed under the Institute's responsibility. The procedure regularly involves external experts (see Section 2.2.2). If necessary, the Institute's Scientific Advisory Board may also be involved. The internal quality assurance process is not outlined in this flow chart.

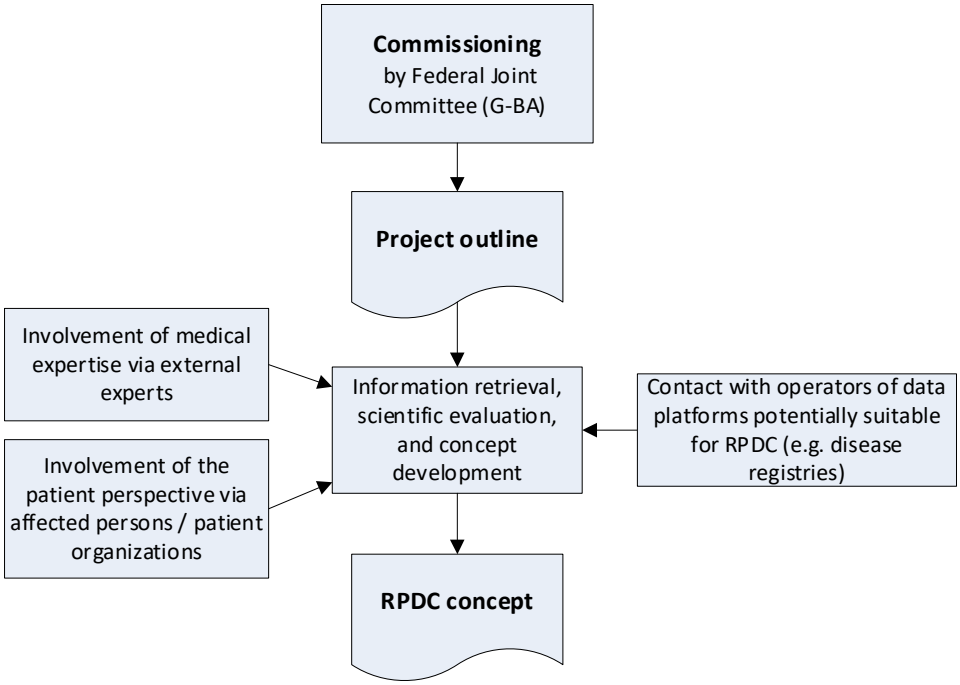


Figure 4: Procedure for development of an RPDC concept

The G-BA's **commission** describes the research question(s) for which an RPDC is to be commissioned according to the G-BA‘s decision. These questions described in the commission provide the framework for the RPDC concept.

The **project outline** summarizes the key steps for information retrieval and scientific evaluation. The project outline is the basis for the development of the RPDC concept.

During the development of the **RPDC concept**, external experts are involved for the medical expertise and affected persons or patient organizations are involved for the patient perspective. This is primarily done by means of questionnaires, similar to the involvement in dossier assessments (see Section 2.1.3.1). As part of the information retrieval process, the operators of existing data platforms that are potentially suitable for the conduct of the planned RPDC according to an initial evaluation (in particular disease registries) are also contacted. Once finalized, the RPDC concept is sent to the G-BA. The Foundation Board, the Foundation Council, and the Board of Trustees of the Foundation are informed of the dispatch shortly afterwards. The RPDC concept is published on the Institute's website following the G-BA's decision on the RPDC.

2.1.3.3 Search for disease registries

The procedure for searching for disease registries is shown in Figure 5. The internal quality assurance process is not outlined in this flow chart. All working steps are performed under the Institute’s responsibility.

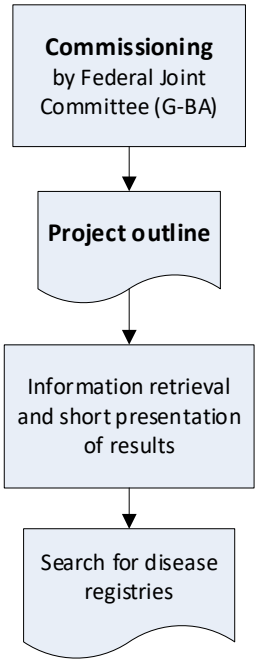


Figure 5: Procedure for the search for disease registries

The G-BA's **commission** describes the therapeutic indication(s) for which a search for disease registries is to be conducted.

The **project outline** summarizes the main steps for information retrieval.

When **searching for disease registries**, focused information retrieval is conducted in various information sources within 4 weeks. Once finalized, the document is sent to the G-BA. It is published on the Institute's website when the G-BA initiates a procedure to require an RPDC in the therapeutic indication.

2.1.3.4 Estimation of patient numbers

The procedure for estimating patient numbers within the context of an RPDC is shown in Figure 6. All working steps are performed under the Institute's responsibility.

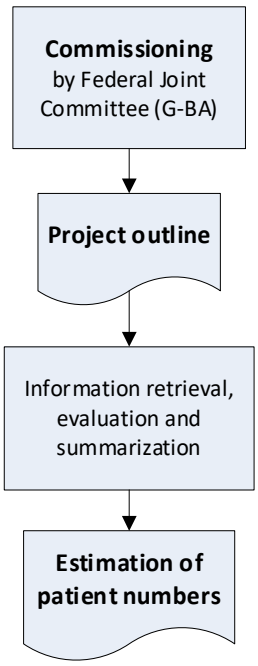


Figure 6: Procedure for the estimation of patient numbers

The G-BA‘s **commission** describes the therapeutic indication of a drug for which a search for patient numbers is required.

On the basis of a **project outline**, the criteria for selecting information are specified and information retrieval is conducted. This is followed by an assessment and final discussion and summarization of the estimated patient numbers.

The **estimation of patient numbers** is performed within 4 weeks. After finalization, the document is sent to the G-BA. It is published on the Institute's website when the G-BA initiates a procedure to require an RPDC in the therapeutic indication.

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 7. 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.

Before **commissioning** by the G-BA, the G-BA prepares the main contents of the commission (during the course of scoping [287]) 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 affected persons 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 2.2.2).

The **patient perspective** is determined on the basis of a questionnaire sent to affected persons 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. Affected persons 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. It is published on the Institute's website soon after it is sent to the G-BA.

The preliminary report is subject to a public hearing. As a matter of principle, the hearing is conducted by means of written comments, which can be submitted within a period of 3 weeks from the time of notification on the Institute's website. 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. The hearing refers in particular to the results of the retrieval and assessment of information presented in the preliminary 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 [287]). 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.

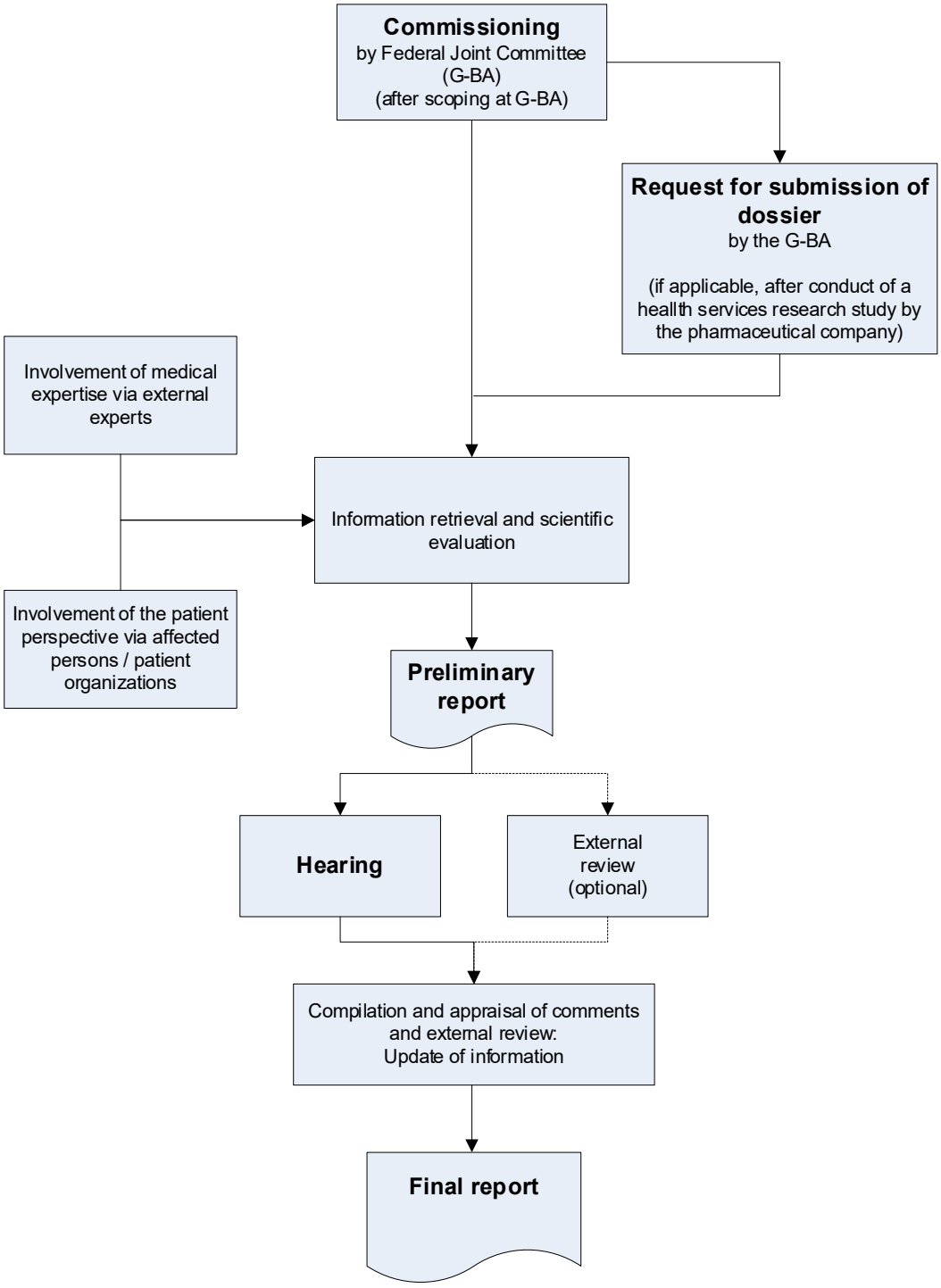


Figure 7: Procedure for the health economic evaluation according to §35b SGB V

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 8. All working steps are performed under the Institute's responsibility. External experts can be involved in the procedure (see Section 2.2.2). The internal quality assurance process is not presented in this flowchart.

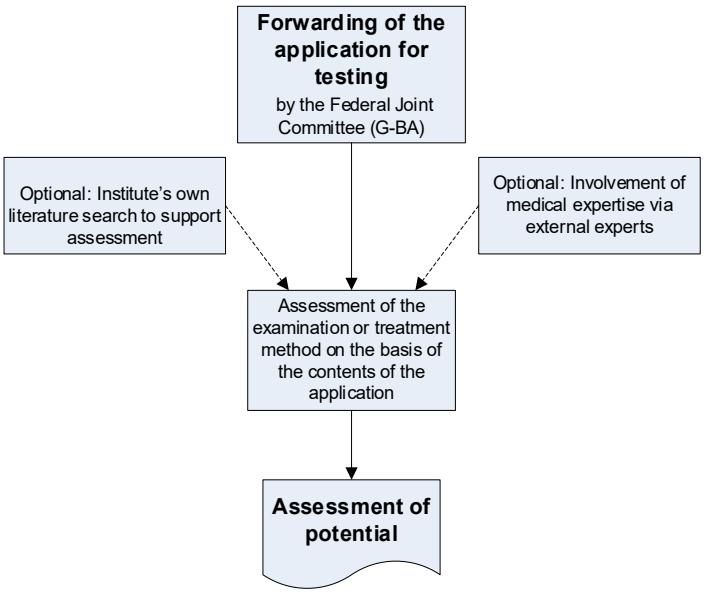


Figure 8: Procedure for the production of an assessment of potential

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 Assessment according to §137h SGB V

The procedure for the production of an assessment according to §137h SGB V largely corresponds to that of an assessment of potential (Figure 8): All working steps are performed under the Institute's responsibility; in this context, external experts may be involved (see Section 2.2.2). However, in contrast to assessments of potential, for assessments according to

§137h SGB V, neither the topic of the assessment nor the main documents are confidential. The internal quality assurance process is not presented in the flow chart.

A hospital submits documents to the G-BA on a new examination or treatment method[14] that is largely based on the use of a high-risk medical device. These documents are made public by the G-BA. After potential supplementation of the documents by further hospitals and affected medical device manufacturers, the G-BA transfers all of the documents providing the basis of the assessment to the Institute.

The Institute conducts an assessment with regard to benefit, harmfulness or ineffectiveness. The basis of the assessment if formed by the documents submitted to the G-BA by a hospital or medical device manufacturer. Optionally, the Institute can conduct its own literature search to support the assessment. External medical expertise can be involved to clarify specific questions. For this purpose, external experts are identified via the Institute's own expert database.

The Institute evaluates whether either a benefit or harmfulness or ineffectiveness can be recognized by means of the documents. As the G-BA is legally obliged to decide on the benefit or harmfulness or ineffectiveness of the method within 3 months, according to §137h SGB V the Institute prepares its assessments within 6 weeks. The respective report is transferred to the contracting agency, the G-BA. The assessment is generally published 4 to 6 weeks after the report was transferred to the G-BA.

2.1.7 Addendum

The procedure for the production of an addendum is presented in Figure 9. All working steps are performed under the Institute's responsibility, involving the Institute's Scientific Advisory Board where appropriate. The internal quality assurance process is not outlined in this flow chart.

[14] Neue Untersuchungs- oder Behandlungsmethode, NUB

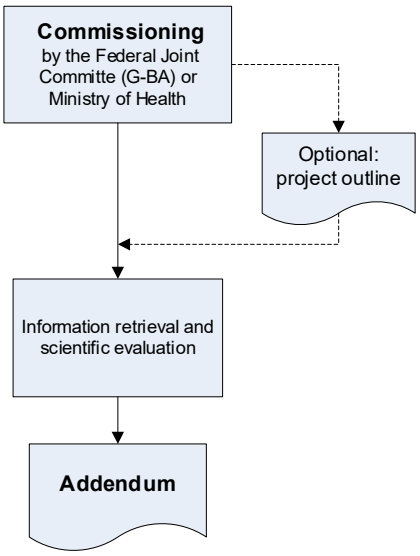


Figure 9: Procedure for the production of an addendum

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.

2.1.8 Health information

The Institute produces **health information** for the general public in various formats, which are presented in more detail in Section 7.16.

The production of health information is based on the legal mandate to provide readily understandable information for the general public in accordance with §139a SGB V (see Section 1.1) and the general commission of the G-BA (see Section 1.2).

This information is provided to the public primarily via the website http://www.gesundheitsinformation.de (and the English-language version

http://www.informedhealth.org). The website's main focus is on topics related to health and illness. Depending on the breadth and depth of a topic, it may combine several different types of article formats.

The process for the production of health information is presented in Figure 10. All working steps are performed under the Institute's responsibility. The internal quality assurance process is not presented in this flow chart.

After deciding on the aspects the topic is to cover, the next step is the **systematic information retrieval**, followed by the **scientific evaluation** of the identified publications.

After the **production of the text** and **editorial work**, as a rule, the draft health information is sent out for an **external review** by an external expert and the draft is then adjusted accordingly, if necessary.

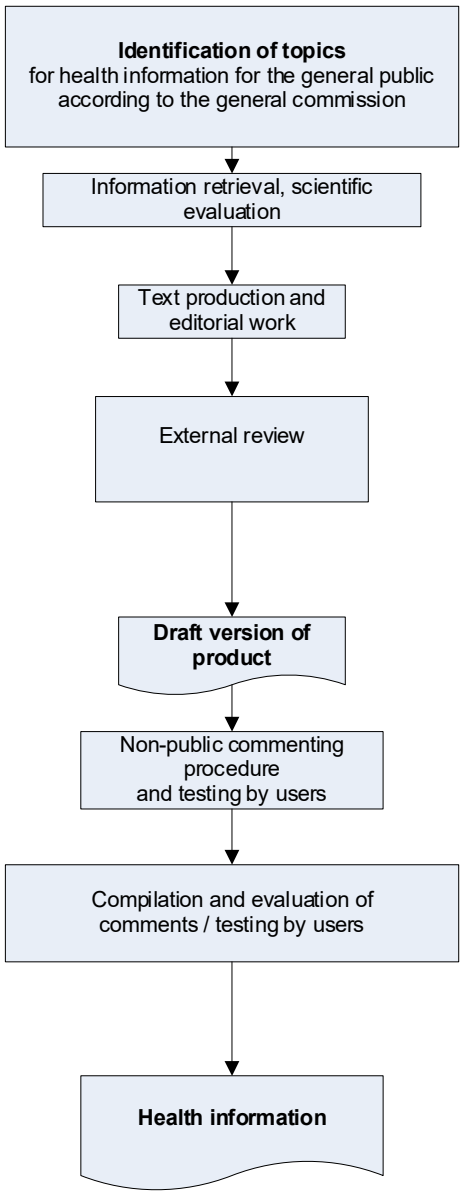


Figure 10: The process for the production of health information

The draft version of a new piece of health information is submitted to those who it was commissioned by, the Board of Trustees, the Foundation's Board of Directors and the Foundation Council as well as the Scientific Advisory Board for non-public **comments**. Before publication, a health information article undergoes standardized external testing by users – generally at the same time as the commenting procedure. The comments submitted during the 4-week commenting period and the results of the testing by users are summarized and reviewed with regard to any resulting necessary changes to the health information, either content-wise or editorial.

Chapter 7 describes details on the process of topic selection, the information retrieval for the production of health information, the scientific evaluation, and patient involvement.

As a rule, corrections, improvements, and updates of published health information are carried out internally. If extensive or substantial changes to content are made, an external review is performed again. In this case, another non-public commenting procedure and further external testing by users may also follow.

If directly commissioned by the G-BA or the Federal Ministry of Health, the health information is produced in the form of a report, rapid report or addendum. The production and publication of the information follow IQWiG's standard procedure. Usually, the corresponding health information is subsequently published on www.gesundheitsinformation.de / www.informedhealth.org. Where relevant, publication is delayed until the corresponding G-BA guidelines become effective.

A variant of health information is information that can be produced to accompany all other products of the Institute in order to present their results in a way that is readily understandable to the general public. The following adjustments in the production process apply:

- There is no external review by experts, as the relevant expertise was already included in the creation of the IQWiG product. There is a review of the health information in the Institute by the department that created the product.
- In the case of health information on "early" benefit assessments of drugs according to §35a SGB V (dossier assessments and related addenda), no commenting procedure takes place due to the tight deadlines, as stipulated in the Institute's statutes.

2.1.9 Working paper

The procedure for the production of a **working paper** is presented in Figure 11. All working steps are performed under the Institute's responsibility, involving external experts or the Institute's Scientific Advisory Board, where appropriate. The internal quality assurance process is not presented in this flow chart.

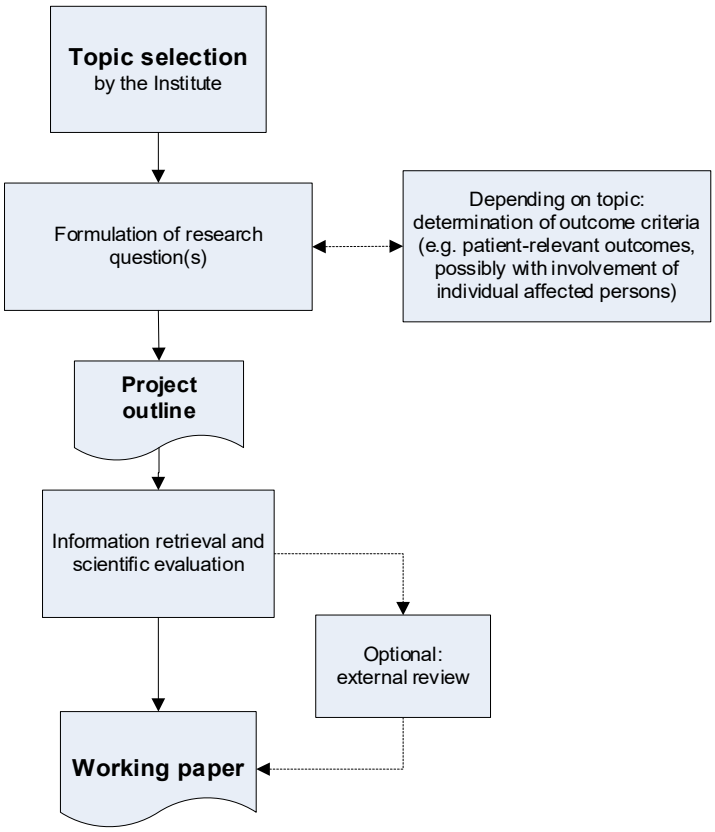


Figure 11: Procedure for the production of a working paper

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.

2.1.10 ThemenCheck report

The procedure for the production of ThemenCheck reports according to §139b (5) SGB V is shown in Figure 12. The ThemenCheck report is prepared by the external experts (see Section 2.2.2) using IQWiG's methods and supplemented by IQWiG with a publisher's comment. The internal quality assurance processes are not included in this flow chart.

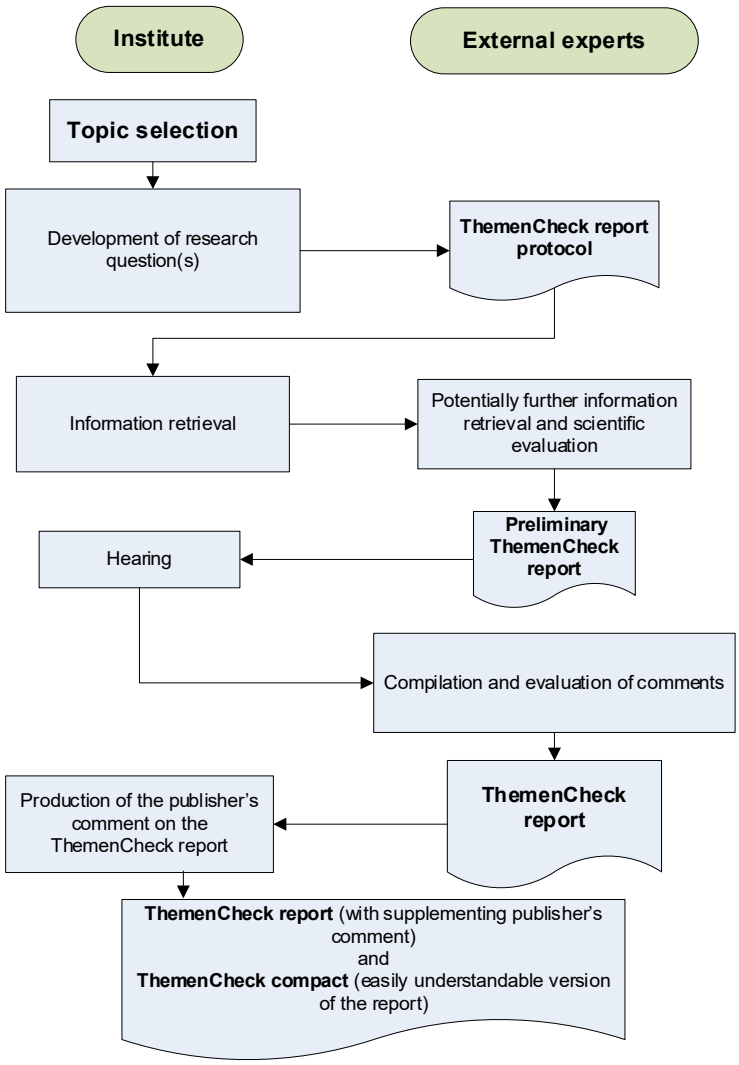


Figure 12: Procedure for the production of a ThemenCheck report

After the Institute has completed the topic selection for the ThemenCheck reports (see Section 6.3), the research question(s) is (are) formulated.

The **ThemenCheck report protocol** is prepared by external experts. It contains the precise research question including the outcome criteria (e.g. patient-relevant outcomes), the inclusion and exclusion criteria for the information to be used in the assessment, as well as the presentation of the project-specific methods for the retrieval and assessment of this information according to the Institute's methods. It forms the basis of the production of the preliminary ThemenCheck report. The ThemenCheck report protocol is published on the Institute's website and the Foundation Board, the Foundation Council, the Board of Trustees, the G-BA and the Federal Ministry of Health are informed about the publication.

In the **preliminary ThemenCheck report** the external experts present the results of the information retrieval and the scientific assessment, including their own conclusion. After

completion, it is also published on the Institute's website and the Foundation Board, the Foundation Council, the Board of Trustees, the G-BA, and the Federal Ministry of Health are informed about the publication.

The preliminary ThemenCheck report is subject to a hearing. The hearing is as a matter of principle conducted by means of written comments, which can be submitted within a period of at least 4 weeks from the time of notification on the Institute's website. Optionally, an oral scientific debate with persons who submitted comments is conducted. If necessary, this debate serves to clarify issues of the written comments. Besides the project-specific methodological approach to answer the research question, the hearing particularly addresses the results of the retrieval and assessment of information presented in the preliminary ThemenCheck report. The hearing is administered and conducted by IQWiG. The comments are evaluated and appreciated by the external experts in the ThemenCheck report.

Building on the preliminary ThemenCheck report, the **ThemenCheck report** contains the assessment of scientific findings, under consideration of the results of the hearing on the preliminary ThemenCheck report, and represents the final product of the external experts. The ThemenCheck report is preceded by a **publisher's comment** in which the Institute classifies the results. In addition, the Institute produces **ThemenCheck compact**, an easily understandable summary of the ThemenCheck report.

The ThemenCheck report, ThemenCheck compact, and the documentation of the hearing on the preliminary ThemenCheck report are first sent to the Foundation Board, the Foundation Council and the Board of Trustees as well as to the G-BA and the Federal Ministry of Health. These documents are generally published on the Institute's website 2 weeks later.

2.1.11 Evidence report

The procedure for the production of an evidence report is shown in Figure 13.

Evidence reports are produced as part of a commission by the Federal Ministry of Health to conduct an evidence search for a selected clinical practice guideline and serve guideline groups as an evidence basis for the development of recommendations for action. To this end, the guideline group, with advice from the AWMF, formulates specific PICO[15] questions for which IQWiG produces separate evidence reports. The internal quality assurance process is not presented in this flow chart.

[15] Population, intervention, comparison, outcome

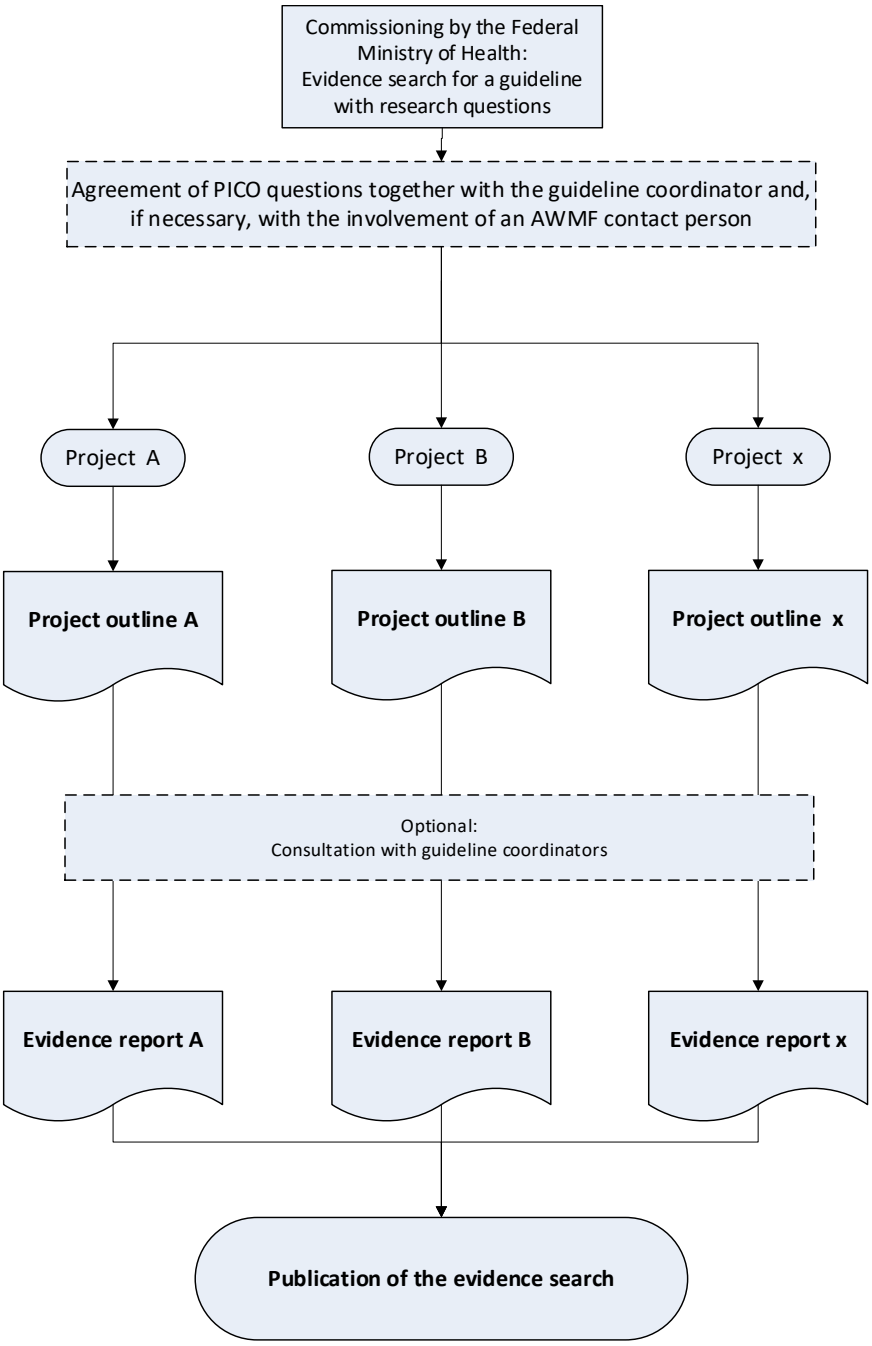


Figure 13: Procedure for the production of an evidence report within an evidence search

After **commissioning** by the Federal Ministry of Health, the guideline coordinator is provided with the **general methodology for the production of evidence reports**. Each PICO question (and the inclusion criteria for the studies relevant to answering it) is then finally agreed upon together with the guideline coordinator, if necessary, with the involvement of a contact person from the AWMF. Together with the methodology of the evidence reports, they form the basis for the production of an evidence report (**project outline**).

The information retrieval results and the evidence available are presented in the **evidence report.** After completion of an evidence report, it is sent to the guideline coordinator of the relevant guideline, the contact person at the AWMF, and the Federal Ministry of Health.

After completion of all evidence reports for a commission on an evidence search for a guideline, they are sent together to the Federal Ministry of Health, the Board of Directors of the Foundation, the Foundation Council as well as the AWMF and then (usually 1 week later) to the Board of Trustees. The reports are usually published on the Institute's website 4 weeks after it is sent to the contracting agency ("publication of the search for evidence").

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:

- the involvement of affected persons in the preparation of IQWiG products
- selection of external experts for collaboration in the preparation of products
- guarantee of scientific independence in the preparation of products
- review of products
- the commenting procedure
- publication of products

Moreover, the provision of scientific advice in relation to the preparation of products is described.

2.2.1 Involvement of affected persons in the preparation of IQWiG products

The involvement of affected persons within the preparation of systematic reviews and HTA reports is now an established international standard [155,257,462]. The involvement of affected persons at IQWiG primarily takes place in the beginning of the work on a project within the framework of the specification of patient-relevant outcomes and relevant subgroups. Moreover, within the framework of the hearing procedure, affected persons have the option of being involved in the preparation of products (see Section 2.2.5). Within the framework of the production of health information, self-help organizations can be questioned about the need for information of affected persons and about the challenges arising in coping with the disease. The drafts of the health information products regularly undergo external user testing. Moreover, topic-specific experience reports by affected persons are recorded (see Section 7.9).

Affected persons can particularly be patients (if appropriate, represented by their parents or other relatives) as well as potential participants in prevention measures.

In the selection of participants the focus is placed on persons actually affected, as patient representatives or representatives of self-help groups are sometimes not patients themselves and cannot assess in a comparable way how affected persons perceive symptoms, functions and activities, or are impaired in their quality of life.

In order to find affected persons the patient representation in the G-BA is regularly asked to name affected persons for a topic directly or via its member organizations. In addition, the Institute may contact or search for affected persons via national or local self-help organizations or groups, hospitals or medical practices, external experts or other routes.

Two different ways exists for involving affected persons: Firstly, a personal consultation meeting can be held where affected persons and IQWiG employees talk about a disease in a small circle of people. Secondly, the perspectives and experiences of affected persons can be requested in writing. The names of affected persons who participated in the consultation meeting or filled in the questionnaire on the consultation are as a matter of principle not published, unless they explicitly approve that their names are published.

In what form affected persons can be involved in the Institute's work primarily depends on how much time is available for the work on the commission and whether the topic of the commission is confidential. Furthermore, the renewed involvement of affected persons can be dispensed with if a topic of a commission has already been addressed within an earlier consultation.

For all of the Institute's products, Table 2 shows the types of potential involvement of affected persons that go beyond the product-specific commenting procedures. The respective product-specific processes are shown in Section 2.1.

Table 2: The Institute's products and potential types of involvement of affected persons

Institute's product	**Type of involvement of affected persons**
Report Rapid report ThemenCheck report	Oral consultation
Dossier assessment RPDC concept Health economic evaluation according to §35b SGB V	Written consultation
Search for disease registries Assessment of potential Assessment according to §137h SGB V Addendum Estimation of patient numbers	No involvement
Health information	Oral consultation, user testings, experience reports
Working paper	Oral or written consultation, as required
Evidence report	Involvement through the respective guideline coordinator
RPDC: routine practice data collection; SGB V: Sozialgesetzbuch (Social Code Book V – Statutory Health Insurance)	

2.2.2 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 as well as §139b (5) SGB V are published on the Institute's website. Exceptions are possible, particularly in the case of urgent commissions. 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 dossier assessments, HEEs according to §35b SGB V, assessments of potential, assessments according to §137h 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.

2.2.3 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 external assignments that could in principle query their professional independence. Details are specified in employment contracts and internal regulations.

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 [493,497], 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, or more precisely, whether there are serious doubts about appropriate collaboration due to conflicts of interest. 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 2.2.2.
- The main basis of the assessment of relationships is self-disclosure using the *Form for disclosure of relationships*, which is published on the Institute's website. Self-disclosure refers to the following types of connections:
  - employment relations/self-employed activities/voluntary activities
  - 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, patents, registered designs
- other circumstances that, from the perspective of an unbiased observer, could be assessed as a conflict of interest (e.g. activities in health-related interest groups or self-help groups)

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 relationships is only published in a summarized form. In this context, for the types of connections covered by the *Form for disclosure of relationships*, it is only stated whether this type of connection existed or not. Specific details, for example, concerning the names of business partners or the amount of any remuneration received, are not published.

2.2.4 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 all project-specific publications before conclusion of the commission) 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.

2.2.5 Commenting procedure

A) 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 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.

B) Course of the public commenting procedure (hearings)

The start date and the deadline for submitting comments are announced on the Institute's website. Interested parties can then submit written comments to the Institute. The submission is electronic; further details are available on the Institute's website. Optionally, a scientific debate with those submitting comments may additionally be held with the aim of clarifying the content of written comments. On request, employees of the G-BA office can participate in the scientific debate.

In order to avoid inappropriate delays in the Institute's work, the comments must fulfil certain formal requirements. The deadlines are outlined in the respective sections on the product-specific procedures (see Section 2.1). Further information on the commenting procedure, including the conditions for participation in a scientific debate, can be found on the Institute's website.

Comments that fulfil the formal requirements are published in a separate document on the Institute's website (*Documentation of the hearing*). In order to ensure transparency, documents that are submitted together with the comments and are not publicly accessible (e.g. manuscripts) as well as the verbatim meeting minutes of the oral scientific debate – if it took place – are also published.

Within the framework of the hearing there is the option to submit documents of adequate quality of any type that, from the perspective of the respective person submitting comments, are suitable to answer the research question. For example, if a search strategy defined in a report plan is restricted to RCTs, non-RCTs can still be submitted within the framework of the commenting procedure. But in such cases, adequate justification is additionally required of the validity of the causal interpretation of the effects described in such studies.

2.2.6 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 Second Chapter §19 of the G-BA's Code of Procedure [287]). 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.

2.2.7 Scientific advice

In special cases the Institute is involved in the provision of scientific advice to study sponsors, for example within the framework of its collaboration in the European network for Health Technology Assessment (EUnetHTA). In this context the primary goal is to support the design of studies that provide informative data for benefit assessments. In order to ensure the independence of the assessment, including those cases in which a benefit assessment contains studies on which the Institute has provided advice, an appropriate organizational separation of advice and assessment is ensured.

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 [790].

If a comparison is not explicitly named, 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 an advantage, the term "greater benefit" in comparison with the other intervention is used. Dossier assessments are an exception; in these cases the term "added benefit" is used instead of the term "greater benefit".
  - In the case of a disadvantage, the term "lesser benefit" is used.
  - In the case of comparable effects, the term "comparable benefit" is used.
- Harmful aspects:
  - The terms "greater harm", "comparable harm" 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 patients, 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 [68]. 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.

All of the outcomes mentioned need to be evaluated in context for a conclusive assessment of patient-relevant benefit. 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 [122]. 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 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 [240]. 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 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 performed as a rule (see Section 3.1.5) 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 [24,273,597]. Most surrogate endpoints are, however, unreliable in this regard and can be misleading when used in a benefit assessment [148,307,317]. 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 [42,775]. 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 [513]. However, the methodological literature frequently discusses correlation-based procedures for surrogate validation, with estimation of correlation measures at a study level

and individual level [392]. 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 [126,534]. Alternative methods [775] 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 [126,127]. 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 [392]. 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 [271,307,317,513]. The studies on which the validation was based must therefore have been conducted with 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” [125,392]. For this purpose, the effect on the surrogate resulting from the studies included in the benefit assessment is related to the STE [127,534].

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 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 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" [154].

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 [805]. 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 effects [154]. 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 [63,405]. 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 events (including the description of a causal relationship to the medical intervention) are considerably underrepresented in medical research [73,223,404]. Moreover, reporting of

adverse events in individual studies is of poor quality, which has also led to amendment of the CONSORT[16] statement for RCTs [403]. 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 [179].

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. In addition, especially for drugs, the data are recorded using MedDRA[17] in accordance with a standardized coding system. 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

Evaluating and reporting aspects of harm in benefit assessments

The Institute thus 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.

[16] Consolidated Standards of Reporting Trials

[17] Medical Dictionary for Regulatory Activities

- As a matter of principle, the overall rates of serious adverse events are used and severe adverse events, if applicable (e.g. those with CTCAE[18] Grade ≥ 3) as well as discontinuations due to adverse events.
- In addition, potentially relevant specific adverse effects are selected for the assessment. This is done in 2 different ways:
  - On the one hand, those specific adverse effects are selected that are of particular importance for the clinical picture or the interventions used in the study or studies. For benefit assessments according to §139a SGB V, this is the primary way of identifying such adverse effects. In this case, the compilation is conducted within the framework of the preliminary search on the respective research question and by involving patients and other people affected.
  - On the other hand, specific adverse effects relevant to the benefit assessment are identified based on the adverse events that occurred in the relevant study or studies. This path is particularly important for benefit assessments of drugs according to §35a SGB V (see Section 3.3.3). Firstly, due to the procedure, no preliminary searches are conducted and secondly, new drugs are as a rule assessed, so that the identification of unknown or unexpected adverse effects is of greater importance here. In this case, the selection is based on the adverse events, serious adverse events and severe adverse events (if applicable) that are submitted in the dossier by the pharmaceutical company in accordance with the requirements of the G-BA's Rules of Procedure [286]. The criteria for the selection are the relevance of the events to patients and the differences between the treatment arms. For non-serious or non-severe adverse events, additional requirements on the minimum frequency of events may also be specified, provided that this does not influence the overall weighing of benefits and harms to determine the added benefit.

For the specific adverse events selected, in each case the most appropriate operationalization with regard to the certainty of measurements and results is applied. For example, it may be the case that a particular adverse effect is better captured by a Standardized MedDRA Query (SMQ) or by an outcome operationalized outside MedDRA than by a single Preferred Term (PT). It is also checked whether the events selected are consistent with the same content construct. For example, results can be considered inconsistent if a PT (e.g. nasopharyngitis) shows an advantage of the intervention, but a disadvantage in a similar PT (e.g. rhinitis). If no suitable operationalization is available or if the results are inconsistent, they are not usually used to draw conclusions on the harm of an intervention.

[18] Common Terminology Criteria for Adverse Events

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[19] group [30].

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 9.3.5).

A) Certainty of study results

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 9.3.13) 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 9.1.3). 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 9.1.4). 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

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

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

B) Inference of the evidence base and certainty of conclusions

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 examined whether within a meta-analysis (see Section 9.3.7) a common effect estimate can be meaningfully formed. In this case, the common effect estimate must be statistically significant to infer a proof, an indication or a hint according to the existing certainty of results.

There are situations in which a common effect estimate is not meaningful (see Section 9.3.7). On the one hand, the study results can be too heterogeneous. On the other hand, homogeneous results from a few studies can also lead to common effect estimates that are not informative or very imprecise for models with random effects [56,657]. In such situations, the results are summarized qualitatively or (particularly in the case of heterogeneous results with more than 4 studies) the prediction interval is used. If the qualitative summary or the position of the prediction interval allow a conclusion in the sense of the research question, conclusive effects are present. Conclusive effects are understood to mean a data situation in which it is possible to infer an effect in the sense of the research question, even though a common estimate of the effect is not meaningfully possible. Here, a distinction is made between moderately conclusive and clearly conclusive effects (see below). If the present data situation allows no conclusion in the sense of the research question, the effects are not conclusive.

The following situations lead to conclusive effects:

- Conclusive effects are present if the prediction interval for displaying heterogeneity in a meta-analysis with random effects (see Section 9.3.7) is presented and does not cover the zero effect.
- If the prediction interval is not presented or it covers the zero effect, unidirectional effects are present in the following situation. The effect estimates of 2 or more studies are conclusive and for these studies, all of the following conditions apply:
  - The overall weight of these studies is 80% or greater.
  - 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 9.3.7).

If conclusive effects are moderately or clearly conclusive, 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 9.3.7), the classification into effects that are moderately or clearly conclusive depends on the number of studies.

- **2 studies:** conclusive effects are always clearly conclusive.
- **3 studies:**
  - All studies show statistically significant results. The conclusive effects are clearly conclusive.
  - Not all of the 3 studies show statistically significant results. The conclusive effects are moderately conclusive.
- **4 or more studies:**
  - All studies show statistically significant results in the same direction of effects: The conclusive effects are clearly conclusive.
  - The prediction interval does not cover the zero effect: The conclusive effects are clearly conclusive.
  - The prediction interval covers the zero effect: The conclusive effects are moderately conclusive.

In the case of heterogeneous situations, moderately and clearly conclusive effects correspond to moderately and clearly unidirectional effects, respectively. In homogeneous situations, the use of the unidirectionality concept is not meaningful, so that the term “conclusive effects” must be used here to describe that a data situation is present in which a conclusion in the sense of the research question is possible without a quantitative summary of the study results.

Impact of the certainty of results and further factors

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 ICH[20] guidelines [400] 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, at least 1000 patients were included in each study arm and there are at least 10 centres. The number of 1000 patients as well as the number of 10 centres serve as orientation for the Institute and do not mean rigid limits.
- 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

[20] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use

9.3.8). 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 3. 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 3: 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			
			Common effect estimate meaningful	Common effect estimate not meaningful		
			Meta-analysis statistically significant	Conclusive effects[a]		
				Clear	Moderate	No
Qualitative certainty of results	High	Indication	Proof	Proof	Indication	–
	Moderate	Hint	Indication	Indication	Hint	–
	Low	–	Hint	Hint	–	–
a. See text in Section 3.1.4 B 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 3. 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 within the framework of a conjoint meta-analysis. 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 conclusive effects 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 lowest 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 Consequences of incomplete data availability

An essential prerequisite for the validity of a benefit assessment is the availability of as complete information as possible on the methods and results of the studies conducted. An assessment where relevant information is missing can provide biased results [256,391] (see also Section 9.3.13). The bias in published evidence through reporting bias, including

publication bias and outcome reporting bias, has been described comprehensively in the literature [214,524,696,753].

In order to be able to assess the extent of potential bias and reduce the effects of this bias, information retrieval is therefore necessary that is also aimed at identifying potentially previously unpublished studies. Moreover, such information retrieval must aim to obtain all relevant results and information on study methods for studies that have not yet been published at all and to supplement the missing information for studies that have already been published.

The main features of such information retrieval are described in Section 8.1. For the identification of previously unpublished studies, topic-independent searches in study registries are of particular importance. However, complete information on previously unpublished studies or studies published in scientific journals can generally only be obtained from clinical study reports [452].

The request to study sponsors (in particular manufacturers of drugs or medical devices) to submit clinical study reports or comparable documents on the studies for which they are responsible therefore represents a central pillar for avoiding relevant bias due to reporting bias (see Section 8.1.3 on requests to manufacturers). If the clinical study reports are not submitted, or not submitted in full by the study sponsors despite a request, and if these are not studies that are only of minor relevance for the overall assessment, this usually means that no summarizing conclusion is drawn for the intervention to be examined on the basis of the incomplete evidence. In particular, no advantage (e.g. no added benefit) is then derived for the intervention(s) examined.

For studies where it is highly likely that no clinical study report was produced, the further procedure depends on the relevance of the studies and the extent of incompleteness (e.g. completely unpublished studies, lack of relevant results, unclear methods). In these cases too, the incomplete evidence may not allow a summarizing conclusion to be drawn for the intervention examined. In any case, the risk of bias resulting from missing information is taken into account in the conclusion of a benefit assessment.

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 the benefit of a medical intervention can also be inferred from 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 infer 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. [297] 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 [297]. This magnitude serves as orientation for the Institute and does not represent a rigid threshold. Glasziou et al. [297] 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 or consistent results on an outcome category). This dependence is also reflected in the recommendations of other working groups (e.g. the GRADE group) [325].

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 of 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. [242]). 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 of long-term treatment [383]. 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 [240,751]. In the selection of evidence (especially of study types) to be considered for the demonstration of an effect, the same principles as with other outcomes usually apply [751]. 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 (i.e. non-blinded) 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 [798]. This should be considered in their interpretation (see Section 9.1.4). 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 a US Food and Drug Administration (FDA) guideline [751].

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 [792]. Independent of this, patients with rare diseases also have the right to the most reliable information possible on treatment options [232]. 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 [239]. 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 [241].

In the case of extremely rare diseases or very specific disease constellations, the demand for (parallel) comparative studies may be inappropriate [792]. 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) [123]. 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. the 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 of the drug.

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 (see Section 3.1.4).

The study design has considerable influence on the certainty of results. This is because 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 [315]. 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-drug interventions [565].

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. A reason for exception can be 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, for example, may provide sufficient certainty of results that a particular intervention prevents this otherwise inevitable course [492] (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 5.2.5 [455]).

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 9.1.4) are not given primary consideration in the assessment process.

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 [287]. 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[21]) (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) [119].

[21] Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV

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 as well as in Skipka et al. [691]. The basic approach aims to derive thresholds for confidence intervals for different effect measures depending on the effects to be achieved, which in turn depend on the quality of the outcomes and the extent categories. Depending on the quality of the outcome, the confidence interval must be completely below (in the case of relative effect sizes) or above (in the case of the standardized mean difference) a certain threshold in order to consider the extent to be minor, considerable or major.

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

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

The thresholds are set separately for each category. The more serious the event, the lower the thresholds (in relation to the effect size). The higher the extent of added benefit, the higher the thresholds (in relation to the effect size).

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
- continuous or quasi-continuous, in each case with available standardized mean differences (SMDs)
- 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.

A) Binary outcomes

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 smaller than 1, i.e. the lower the value, the stronger the effect.

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.

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. The **upper** limit of the 95% confidence interval must be smaller than the respective threshold in order to reach the corresponding extent category.

Table 4: Thresholds for determining the extent for the relative risk

		Outcome category		
		All-cause mortality	**Serious (or severe) symptoms (or late complications) and adverse effects, as well as health-related quality of life[a]**	**Non-serious (or non-severe) symptoms (or late complications) and adverse effects**
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 an appropriate 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 the outcome 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 [742]. 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 of 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, an appropriate response criterion or cut-off value is required (see Section 9.3.3). On the basis of the responder analyses (2x2 tables) the relative risks are calculated directly from them. The extent of the effect is then determined by means of Table 4.

D) Continuous or quasi-continuous outcomes, in each case with available SMDs

In order to also assess the extent of added benefit for the effect measure SMD (Cohen's d or Hedges'g), comparable thresholds for SMDs are defined following the system of thresholds for relative risks. Orientation is provided here by the usual classification of Cohen's d into small (SMD between 0.2 and 0.5), medium (SMD between 0.5 and 0.8), and large effects (SMD 0.8) as well as the extension by Rosenthal [615], who describes the result SMD $\geq$ 1.3 as a very large effect. Furthermore, it should be kept in mind that the irrelevance threshold SMD = 0.2 has been established to ensure a relevant effect (see Section 9.3.3). Furthermore, the obvious rules of determining the extent of added benefit in the case of binary data should also apply to continuous data, namely (1) that thresholds become more stringent the greater the extent, and (2) thresholds become more stringent the less serious a symptom is (with the exception of the category "minor extent"). Based on SMD = 0.2 as the threshold for a minor added benefit for both serious and non-serious symptoms, thresholds for SMD can also be defined according to Table 5, taking into account these rules and the ranking of all thresholds according to Table 4. In contrast to the thresholds for the relative risk, it should be noted that for the SMD the 95% confidence interval must lie completely **above** the corresponding threshold and that the zero effect is not given by the number 1, but by the number 0. The **lower** limit of the 95% confidence interval must therefore be larger than the respective threshold in order to reach the corresponding extent category. It seems plausible that true small, medium and very large effects according to Cohen and Rosenthal [615] can be statistically detected by using the thresholds of Table 5 for the lower limit of the 95% confidence interval.

Table 5: Thresholds for determining the extent for the SMD

		Outcome category	
		Serious (or severe) symptoms (or late complications) and adverse effects as well as health-related quality of life[a]	**Non-serious (or non-severe) symptoms (or late complications) and adverse effects**
Extent category	major	0.5	Not applicable
	considerable	0.3	0.4
	minor	0.2	0.2
a. Precondition (as for all patient-reported outcomes): use of a validated or established instrument.			

E) 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 [167] 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.

3.3.4 Concept for routine practice data collection according to §35a (3b) SGB V (RPDC concept)

According to §35a (3b) SGB V, the G-BA can require the responsible pharmaceutical company to collect routine practice data (RPD), including the associated analysis, for drugs for rare diseases (orphan drugs), with conditional approval and with approval under exceptional circumstances [6].

A concept for RPD collection (RPDC concept) is developed in preparation for a decision by the G-BA on an RPDC to be conducted. On the one hand, the RPDC concept should describe the methodological requirements for conducting the RPDC, including the analysis for the purpose of the benefit assessment according to §35a SGB V in comparison with an appropriate comparator therapy (see Section 3.3.3). On the other hand, the RPDC concept should be used to assess whether the planned RPDC is feasible in principle. For this purpose, the RPDC concept addresses the following topics in detail:

- Description of the information requirements for the benefit assessment of the respective drug compared to the appropriate comparator therapy according to §35a SGB V: This results from the evidence gaps in the approval studies and can concern all

categories of patient-relevant outcomes (mortality, morbidity, health-related quality of life, adverse effects [397]).

- Definition of the research questions arising from the information requirements for the benefit assessment according to §35a SGB V: These are described in the PICO[22] format.
- Evaluation whether the research questions are partially or completely covered by already planned or ongoing data collections on the drug in question: Based on this, the research questions are derived for which this is not the case and for which an RPDC should therefore be conducted (RPDC questions).
- Evaluation whether data platforms already exist that can, in principle, serve as a primary data source for conducting studies to answer the RPDC questions: The analysis of concepts for the generation of RPD and their analysis for the purpose of benefit assessments of drugs according to §35a SGB V presented in the Institute's Rapid Report A19-43 [397] showed that, in addition to study-specific data collection, (disease) registries in particular can be a suitable data source for studies to answer the RPDC questions (registry studies). This requires that the respective registry can provide the necessary data in sufficient quality. In addition to data collection, this includes the planning, analysis and publication of the results of the associated registry study [369]. Potentially suitable registries are therefore identified in the RPDC concept (see Section 8.2). Potentially suitable registries are generally those that are functional and currently analysing data from patients in the patient population to be investigated. In addition, patient data from Germany must also be documented in the registries. The respective registry operators are asked for further information about the registry using a structured questionnaire. On the basis of this and other publicly available information, the registries are described and their suitability for conducting a registry study to answer the RPDC questions is assessed, also on the basis of nationally and internationally used quality criteria [245,251,301,510,707]. The need for any necessary modifications and expansions is derived from this suitability evaluation in order for the registry to serve as the primary data source for the registry study for the purpose of an RPDC.
- Description of the type of RPDC: This includes the study design and the primary data source for data collection. Due to the basic research question of the benefit assessment (comparison with an appropriate comparator therapy), this is usually a comparative study, which is conducted under the limitation of [245] §35a SGB V (3b) without randomization [6,397]. With regard to the study design, aspects of target trial emulation (including the time of the start of observation) must be taken into account [357]. The selection of the primary data source is based on the results of the previous suitability evaluation of existing data sources.

[22] Population, intervention, comparison, outcome

- Description of the duration of the RPDC: On the one hand, this includes the duration of observation of the individual patients, which should ensure that relevant outcomes can be assessed in the present therapeutic indication and situation of use. On the other hand, the duration of the RPDC is also determined by the fact that sufficient patients or events (the necessary sample size) can be enrolled and followed up to collect meaningful data for the RPDC. For this purpose, epidemiological data such as the prevalence and incidence of the disease and the degree of completeness of the data source are taken into account.
- Description of the scope of the RPDC: Because of potentially unknown confounders that may influence the effects observed in the study, a conclusion about the benefit or harm of an intervention investigated in a non-randomized study can only be drawn from a certain effect size (see Sections 3.1.4 and 3.2.2), e.g. in the range of 2 to 5 for the relative risk (or 0.2 to 0.5, e.g. for interventions that reduce mortality) in relation to the lower or upper limit of the confidence interval [397]. This is taken into account when answering the question of whether or not the planned RPDC is feasible in principle. Based on the available evidence on the drug and the appropriate comparator therapy, it is assessed whether sufficient patients can in principle be recruited and followed-up within an acceptable period of time to be able to generate meaningful results for a benefit assessment with an RPDC. This also takes into account how the patients are likely to be distributed across the treatment groups.
- Description of the basic requirements for data analysis: The concepts described in Rapid Report A19-43 for the analysis of RPD for the benefit assessment of drugs according to §35a SGB V [397] form the basis for this. Due to the necessary non-randomized study design for the RPDC, an adequate adjustment for confounders is an essential part of these requirements (see Section 9.3.6).

Finally, the RPDC concept includes a conclusion on the following aspects:

- the question(s) that should be the subject of the data collection and analyses
- the requirements for the type, duration and scope of data collection
- the data collection methods
- the analysis methods of the pharmaceutical company, and
- an evaluation of whether the proposed RPDC is feasible in principle

The conclusion represents the recommendation to the G-BA for the further procedure in which the G-BA makes the final decision on the requirement of an RPDC [287]. If the G-BA requires an RPDC, the company must prepare a study protocol and a statistical analysis plan (SAP). These documents must describe the planning, conduct and analysis of the study for the

purpose of the RPDC. IQWiG can be involved both in the consultation prior to the development of the study protocol and the SAP, as well as in the suitability evaluation of these documents on behalf of the G-BA.

3.4 Non-drug therapeutic interventions

Non-drug therapeutic interventions are assessed in detail within the framework of reports (see Section 2.1.1) or rapid reports (see Section 2.1.2) or in the form of ThemenCheck reports (see Section 2.1.10). The procedure and methods of these assessments are applied according to the general principles of the methods paper. Furthermore, new examination and treatment methods that are largely based on the use of a high-risk medical device can also be evaluated within the framework of an assessment according to §137h SGB V (see Section 2.1.6). As §137h assessments are primarily based on documents submitted (and not on information retrieval conducted by the Institute itself) and have to be produced in a markedly shorter period of time, the methodological depth of the work on the assessment and the certainty of a potential conclusion on benefit or harm is as a rule lower than is the case for reports and rapid reports. In particular, §137h assessments do not distinguish between different certainties of conclusions (proof, indication, hint). However, the requirements for a benefit are the same for reports and rapid reports in comparison with assessments according to §137h SGB V and generally require the availability of RCT results on patient-relevant outcomes (see Sections 3.1 and 3.2).

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 [287] envisages, as far as possible, the preferential consideration of RCTs, independent of the type (drug/non-drug) of the medical intervention to be assessed. The principles of evidence-based medicine and of benefit assessments are therefore also valid for specific fields of medicine, such as surgery, dentistry, psychotherapy, alternative/complementary medicine or digital health applications (e-health). For medical devices, within the framework of a conformity assessment, the RCT is not presented as the study design of choice according to the current EU Medical Device Regulation (MDR) [187], but clinical investigations must be "performed on the basis of an appropriate plan of investigation" [234], so that randomized controlled designs tend to be used more and more frequently in this area [648]. In contrast to the European conformity assessment procedure for new medical devices, the Institute's assessments always refer to examination and treatment methods, not to individual medical devices.

Compared with studies on drug interventions, studies on non-drug interventions are often associated with specific challenges and difficulties [523], but these can usually be overcome at least in part:

- Blinding of the staff performing the intervention is often impossible and blinding of patients is difficult or also impossible. However, a reduction in the informative value of a study can be mitigated by the blinded recording of outcomes or by the use of objective outcomes.
- The skills and abilities of the users of the intervention are often essential for an identifiable benefit of a non-drug intervention in a study. Both doctors and patients can be users. Ideally, adequate selection and training will help to minimize the negative impact of learning curve effects on the study results.
- Many non-drug interventions change in the course of a study. Numerous small changes (often referred to as “step-by-step innovations”) are common, especially with medical devices. Changes in the intervention during the course of a study usually do not require a new study, but can be investigated in sufficient detail by using appropriate statistical methods to assess the influence of these changes on the study results.

In addition, non-drug interventions are sometimes more or less complex interventions [160]. 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 [523]). 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 Chapter 2 §13 (2) of the G-BA’s Code of Procedure [287]. 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 informative value” [287]. 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

The assessment of diagnostic tests is subject to the same socio-legal and formal conditions as that of non-drug interventions of a therapeutic nature (see Section 3.4). For example, the result of the conformity assessment procedure for the CE marking of a relevant medical device (or an in vitro diagnostic test) and the approval status of drugs used for diagnostic purposes may also need to be taken into account. The differences in the assessment of therapeutic and diagnostic interventions are instead of a fundamental methodological nature.

In the simplest case, diagnostic tests consist of a single test that provides a dichotomous result (test positive or test negative). If the result is not dichotomous but, for example, continuous, the complete diagnostic test will include corresponding cut-off values, which are used to decide from what level the test result should be considered positive. In some cases, diagnostic tests consist of several individual diagnostic tests/pieces of information, the individual results of which are combined into an overall test using a test-specific algorithm. In principle, the methods described below are suitable for diagnostic tests of all types (dichotomous or other test result, single or repeated testing), all diagnostic areas (laboratory tests, imaging, etc.) and purpose of use (primary diagnosis, biomarkers, staging, etc.).

In the context of a benefit assessment, it should be considered when defining the research question that the diagnostic test to be evaluated ("index test") may either (at least partially) replace or complement standard diagnostics. Often, a new diagnostic test is intended to be embedded in an existing diagnostic strategy, for example by placing a new test upstream (triage test) or downstream (add-on test) of an established test in order to reduce the use of the downstream test in each case [77]. The precise definition of the question for a diagnostic test is very important, as this may result in different evidence requirements. For more complex questions, flowcharts can help to systematically record the different consequences of diagnostic interventions [642].

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 medical benefit and also no benefit relevant to social law [576,616,665].

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.

3.5.1 Study designs for the benefit assessment

In the methodology of clinical research, it has been increasingly recognized in recent decades that diagnostic interventions can and must be evaluated in a very similar way to therapeutic interventions in terms of benefit and harm [1,280,544,665]. This is because, in essence, it is always a matter of demonstrating that an intervention – including a diagnostic one – has effects on patient-relevant outcomes at the end of the treatment chain. This as a rule requires comparative intervention studies. The Institute follows this logic, which is also laid down in the G-BA's rules of procedure [287], and conducts benefit assessments on diagnostic tests primarily on the basis of comparative intervention studies with patient-relevant outcomes. The main features of the assessment correspond to the explanations given in Sections 3.1 and 3.4.

The impact of diagnostic tests on patient-relevant outcomes can be achieved by avoiding high(er) risk interventions or by the (more) targeted use of interventions. If the collection of diagnostic information itself is associated with a high risk, a lower-risk diagnostic test may have patient-relevant advantages, namely, if (in the case of comparable test accuracy) the performance of the test itself leads to lower mortality and morbidity rates, or to fewer limitations in quality of life. A distinction must therefore be made between direct effects caused by the diagnostic test itself and indirect effects caused by the subsequent treatment. In most cases, the indirect effects (e.g. avoidance of hypoglycaemia) are much more important in the weighing of benefits and harms than the direct (adverse) effects of a diagnostic test (e.g. pain during capillary blood glucose self-monitoring).

Because 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 [265,690]. Diagnostic tests should therefore be used to guide treatment. 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.

3.5.1.1 “Test-treatment studies”: studies of the diagnostic-therapeutic chain

Conclusions about the benefit of diagnostic tests are therefore ideally based on randomized trials on patient-relevant outcomes, which can be conducted in a wide variety of ways [77,78,265,494,511,644]. Depending on the diagnostic test and the medical context, different study designs can be considered, whereby a distinction should be made between strategy design, interaction design and enrichment design [722,730]:

A) Strategy design

In a study with a strategy design, 2 (or more) patient groups each receive different treatment strategies, each consisting of a diagnostic intervention and the therapeutic consequence. The strategy design makes it possible to determine the direct and indirect effects of a diagnostic test because each person as a rule receives either one or the other, but not both, diagnostic tests. One variant of the design focuses on the selection of certain subgroups based on the test. In this case, all patients in the target population first undergo the conventional diagnostic test followed by the diagnostic test to be evaluated; then only those patients for whom the test to be evaluated gives a different result (and thus a different therapeutic consequence) than the conventional test are randomized to treatment (discordance design).

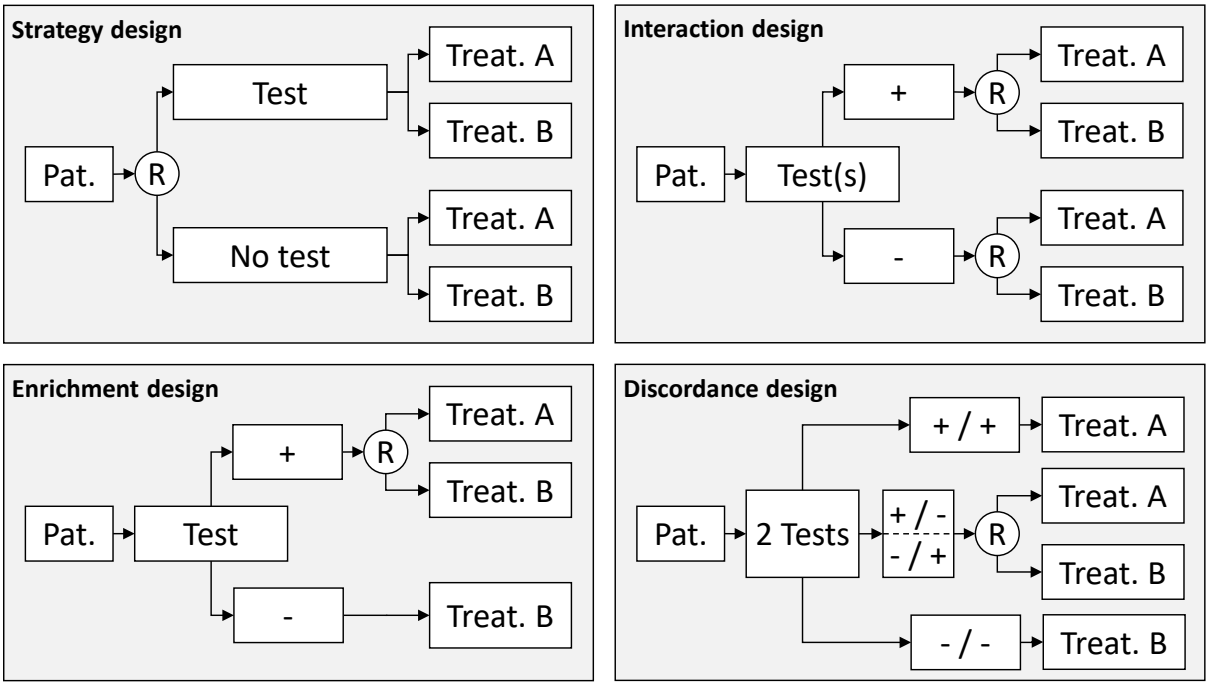
B) Interaction design

Complete information about the interaction between diagnostic information and the benefit of a subsequent therapeutic intervention, i.e. the predictive relevance of the test result for the benefit of a treatment, is provided by studies with a so-called interaction design (also known as "randomize-all design" [644,730]). For this purpose, the entire patient collective is screened using conventional and new diagnostic interventions before all patients are randomly assigned to receive 1 of 2 (or more) treatments. In this way, it is possible to analyse (in the sense of a statistical interaction) whether and to what extent the benefit of the treatment differs between the subgroups that can be formed according to the results of the diagnostic tests. For example, if a positive treatment effect can only be demonstrated in test-positive individuals, but not in test-negative individuals, then this indicates that the diagnostic test is useful in sparing some patients unnecessary treatment and using the treatment only in test-positive individuals. The predictive value of diagnostic characteristics – especially genetic markers – can also be investigated by combining information from prospective randomized treatment studies with the results of markers recorded retrospectively (e.g. from tissue samples); this is known as a prospective-retrospective design [689]. The validity of such designs depends in particular on the existence of a prospective planning of the analyses (especially also the specification of thresholds). For all studies with an interaction design, it is also important that the treatments used correspond to the current standard, that complete information (e.g. tissue samples) on the characteristic of interest is available for all study participants or at least for a clearly described sample for which the structural equality between the groups is still given, and that the problem of multiple significance testing is adequately considered in an analysis of several characteristics [645] (see also Section 9.3.2).

C) Enrichment design

Besides a strategy and interaction design, a third main form of RCTs on diagnostic questions is available with the enrichment design [512,730]. 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.



Pat.: patient collective; Treat.: treatment or further treatment; R: randomization; +: test positive; -: test negative

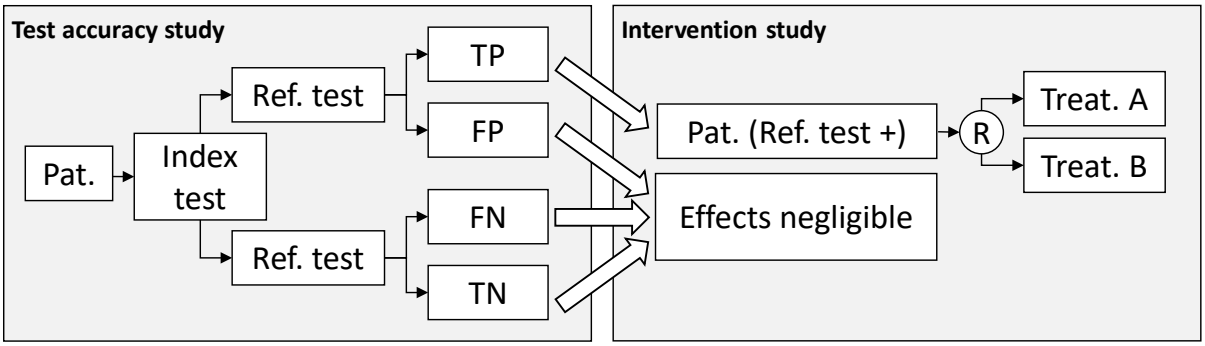
Figure 14: Simplified presentation of special study designs for the evaluation of diagnostic tests

Biomarkers used within the framework of personalized or better stratified medicine should also be evaluated with the methods described here [247,366,730]. 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 [308,694,755]. Here too, it is essential to focus on the predictive rather than the diagnostic/prognostic value of a characteristic. In oncology in particular, the term "companion diagnostic" has become established for specific, mostly genetic, markers that are required as diagnostic tests for the use of targeted anti-tumour therapy according to the approval status [308,350].

3.5.1.2 Linked evidence

Ideally, the diagnostic-therapeutic chain is investigated as a whole in a study ("test-treatment study"). In this case, the diagnostic intervention is directly linked to patient-relevant outcomes via the subsequent treatment. However, it may be appropriate and useful to investigate the

individual parts of the diagnostic-therapeutic chain in separate studies [528,664]. For example, treatment studies may demonstrate that a particular treatment is beneficial for a specific well-defined health condition. Test accuracy studies could then be used to show that a particular diagnostic test can accurately identify this specific health condition (or more accurately than with other diagnostic tests). Logically, linking the respective study results (in terms of demonstrating benefit) leads to the conclusion that the diagnostic test examined is necessary to identify the patient group who could benefit from the treatment (Figure 15). Otherwise – without the use of the test – additional patients might be treated for whom the treatment has no effect or could even be harmful.



Pat.: patient collective; Tre.: treatment or further treatment; Ref. test: reference test; TP: true positive; FP: false positive; FN: false negative; TN: true negative; R: randomization; +: test positive

Figure 15: Simplified presentation of the linking of test accuracy and intervention evidence (linked-evidence approach) for the evaluation of diagnostic tests

For this logical link ("linked evidence") to be viable, it must be ensured that the patient spectrum that can be identified with the diagnostic test is actually the same as, or sufficiently similar to, the patient spectrum that was examined in the treatment studies [528,665]. If the therapeutic knowledge relates to a patient spectrum that was selected using a different diagnostic test, then this knowledge cannot be readily transferred to a patient population that can be identified using a new, possibly more sensitive or more specific diagnostic test. Ideally, with comparable patient selection, the same reference test was used in the test accuracy study to verify the results and as an inclusion criterion in the intervention study. If it is permissible to link evidence on treatment interventions with evidence on diagnostic interventions, this can be used to demonstrate the benefit of a diagnostic test (see e.g. [394,529,628]). However, as in the enrichment design, the derivation of an overall benefit requires that there is no treatment benefit in test-negative individuals and that there are at most negligible intervention effects in the groups of individuals with false-positive, false-negative or true-negative test results. In individual cases, it may be necessary to confirm this requirement by considering further evidence.

A special case of linked evidence arises when a new diagnostic test is only intended to replace an established test without influencing further treatment through higher test accuracy. This is because the previous explanations relate primarily to diagnostic tests which, by increasing test accuracy (i.e. sensitivity, specificity or both), guide patients more specifically to a particular therapeutic consequence. In these cases, it is as a rule necessary to investigate the influence of the diagnostic test on patient-relevant outcomes by recording the entire diagnostic-therapeutic chain in intervention studies. However, it is possible that the diagnostic test to be evaluated is merely intended to replace another, already established diagnostic test without the new test identifying or excluding additional or different patients. If the new test has direct patient-relevant advantages, such as being less invasive or radiation-free, it is not necessary to re-examine the entire diagnostic-therapeutic chain, because the consequences resulting from the new test do not differ from those of the previous test [66,78,528]. In these cases, studies may be appropriate to demonstrate benefit in which it is shown that the test result of the test to be evaluated (= index test) agrees with that of the previous test (= reference test) in a sufficiently high proportion of patients (1-sided equivalence question). This question on the concordance of test results in individual patients is also referred to as a concordance question [472].

In special cases, an interaction between the diagnostic marker and the treatment effect can be inferred with sufficient certainty, even if the treatment effect is only known for the whole group (i.e. test-positive and -negative individuals together). In the (theoretical) extreme case, a test result allows the disease to be ruled out with certainty, so that the treatment of a disease would be useless and would at most have adverse effects. It can never be demonstrated with absolute certainty that a specific test result rules out a specific health condition. However, if it can be shown for a test that test-negative individuals have a sufficiently low risk (or test-positive individuals have a sufficiently high risk) of reaching key outcomes, then the test can allow a meaningful weighing of the benefits and harms of a treatment and thus provide a sufficiently reliable basis for deciding against (or in favour of) a treatment [576]. For example, the benefit of a treatment that has a positive benefit-harm ratio in the overall group of all patients may nevertheless be questionable in a subgroup of test-negative patients, namely if the absolute treatment effect is only very small in this subgroup solely because of the low probability of a clinical event. For such a linked consideration of the treatment effect in the overall group and the outcome risk in a "low-risk subgroup" to be viable, firstly the (relative) treatment effect must be supported by studies. Secondly, to define appropriate thresholds in the assessment of an acceptable benefit-harm ratio, data on patient preferences can be included. In addition, it may be useful to define a topic-specific minimum percentage size of the subgroup of test-negative or test-positive patients.

In general, the different approaches to using linked evidence are only applied when it is clear from the research question and initial literature searches that direct evidence for the

diagnostic-therapeutic chain cannot be expected. In individual cases, both methodological approaches, direct test-treatment evidence and indirect "linked evidence", may be used in parallel to allow for a mutual comparison and therefore to support the conclusions on benefit. The certainty of conclusions on benefit depends on the certainty of the results of the underlying studies and the type of evidence available, with linked evidence generally considered weaker than evidence from direct comparative intervention studies.

3.5.2 Quality assessment of studies on diagnostic and prognostic accuracy

Test accuracy studies – i.e. studies describing sensitivity and specificity – are primarily relevant to benefit assessments in the context of the linked evidence approach. When comparing 2 or more diagnostic tests for specific test accuracy characteristics, the highest certainty of results is obtained from cohort and cross-sectional studies in which the diagnostic tests (and also the reference test) are conducted independently of each other in the same patients and assessed under mutual blinding [378,495,732,783]. In the case of rapidly progressing disease, a randomized sequence of index tests may also be important. In addition to such studies, which allow an intra-individual comparison of the test results and are most commonly conducted [800], RCTs are also conceivable in which a proportion of patients are tested with only one or the other index test before preferably all results are verified with a standardized reference test. This study design also allows for the determination of test accuracy characteristics with the highest certainty of results.

In studies of prognostic markers, which (unlike diagnostic tests) can be investigated in cohort studies rather than in cross-sectional studies, multifactorial regression models often play a key role, so that Section 9.3.6 should be considered. When selecting non-randomized interventional study designs for diagnostic methods, the ranking of different study designs presented in Section 9.1.3 is as a rule applied.

When assessing the certainty of results from diagnostic accuracy studies, the Institute is primarily guided by the QUADAS-2 criteria [782,783], which can be adapted to specific projects. The QUADAS-C tool may also be helpful in evaluating comparative test accuracy studies [799]. The criteria of the STARD statement [80,81] are used in individual cases to decide on the inclusion or exclusion of studies not published in full text. The PROBAST instrument is mainly used for the methodological assessment of prognostic studies [794]. The criteria of the TRIPOD statement [152] are used to decide on the inclusion or exclusion of studies not published in full text. Explanations on meta-analyses of studies on diagnostic quality can be found in Section 9.3.7.

3.6 Early diagnosis and screening

At this point, early diagnosis and screening refers to the use of diagnostic tests in people who are asymptomatic for the target disease [316]. Screening programmes consist of different

components that are (or can be) evaluated either as a whole or in parts [171,686]. Evaluation is based on internationally accepted standards and criteria, such as those of the UK National Screening Committee (UK NSC [598]), the US Preventive Services Task Force (US PSTF [340,584,649]) or the New Zealand National Health Committee (NHC) [545].

According to the criteria outlined in Section 3.5, 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 3.1 to 3.1.5.

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. As explained in Section 3.5.1.2, the linked evidence approach [1] is used to examine the extent to which it is proven that the consequences of the test results 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 3.5.

A particular aspect of harm from screening is the fact that screening produces overdiagnoses. An overdiagnosis is defined as an actually true-positive diagnosis which, however, without screening would not have caused symptoms during a person's life [139]. For instance, overdiagnoses occur in screening for slowly progressing diseases, because in these cases, there is a high probability that a person will die of a different cause before developing symptoms. Since overdiagnoses are inevitable as a harmful effect of any screening, but can only be recorded indirectly, specific methods for recording this outcome are required [138,231].

3.7 Prevention

Prevention is directed at avoiding, reducing the probability of, or delaying health impairment [772]. 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 [353]. 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 3.3.4).

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 [743].

In individual cases, it needs to be assessed to what extent the consideration of other study designs is meaningful [411]. For example, mass-media campaigns are often evaluated within the framework of interrupted time-series analyses (e.g. Vidanapathirana et al. [764]), and the use of this study design is also advocated for community intervention research [67]. In the quality assessment of these studies, the Institute uses for orientation the criteria developed by the Cochrane Effective Practice and Organization of Care Review Group [151].

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 [473]. Special focus is therefore placed on both of these aspects in the Institute's assessments.

3.8 Assessments of potential and §137h assessments

The following text first addresses assessments of potential according to §137e SGB V, before assessments according to §137h SGB V are explained. The common feature of both assessment procedures is that certain non-drug examination or treatment methods are assessed on the basis of documents submitted to the Institute by external third parties. Another common feature of both assessment procedures is that they can lead to the conduct of a testing study. The study's key points must therefore be reviewed or developed by the Institute itself.

3.8.1 Assessments of potential

In contrast to benefit assessments, assessments of potential according to §137e SGB V 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 Chapter 2 §14 (3, 4) of the G-BA's Code of Procedure [287].

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 [287]). 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 information retrieval conducted by the applicant (see Section 8.5), the assessment of the certainty of results of the relevant studies, and the correctness of the results presented in the application. The Institute can conduct its own literature searches to support the assessment; however, it is not the Institute's responsibility or goal to complete the documents presented. 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. If the existing (or soon to be expected) evidence indicates that a benefit assessment is already meaningfully possible even without a testing study, the report refers to this and does not describe a 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 categories mentioned in Section 3.1.4 for randomized studies (high or moderate certainty of results) the following categories 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), which allows an indirect comparison with the results of other studies (literature controls).

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 9.3.6. Further factors are also taken into account in the assessment of the certainty of results (see Section 9.1.4).

Deviating from the procedure aimed at inferring conclusions on benefit in terms of proof, indications or hints (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 9.3.13). 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.

3.8.2 §137h assessments

In §137h assessments, the basis of the assessment is not an application by a medical device manufacturer or another company, which is the case for the procedure following §137e SGB V; assessments according to §137h SGB V are based on documents compiled by a hospital. The assessment refers to a new examination or treatment method whose technical application is largely based on the use of a high-risk medical device. The hospital must reach an agreement with the medical device manufacturer before submitting the information on the benefit of the method to the G-BA. This information can then be supplemented at the G-

BA by further hospitals and medical device manufacturers before being forwarded to IQWiG for assessment.

A §137h assessment can lead to 3 possible outcomes according to §137h (1) Sentence 4 SGB V:

- The method to be assessed shows a benefit in terms of §137h (1) Sentence 4 SGB V. In this case, there is no need to deal with a possible testing study.
- The harmfulness or ineffectiveness of the method to be assessed can be identified in terms of §137h (1) Sentence 4 SGB V. In this case, too, there is no need to deal with a possible testing study.
- Neither the benefit nor the harmfulness nor the ineffectiveness of the method can be identified. In this case, the key points of a possible testing study must be examined or newly developed.

The assessment of the first 2 items (identifiable benefit, harmfulness) follows the principles described in Section 3.1. Harmfulness corresponds to greater harm from the new method compared with no treatment. Ineffectiveness consists of the new method recognizably showing only effects comparable to those of no treatment. For this purpose, in addition to results on patient-relevant outcomes, results on surrogate outcomes may be sufficient.

4 Health economic evaluation of medical interventions

4.1 Introduction

According to SGB V, a health economic evaluation (HEE) can take place in 3 situations:

1) In the context of the AMNOG procedure, the HEE is a tool that can be used to inform the pricing of a new drug after an early benefit assessment of drugs has been completed. Following an arbitration decision, the SHI umbrella organization[23] or the pharmaceutical company (in short: the company) can apply for an HEE (§130b (8) SGB V) [7]. According to §35b (3) SGB V, the reimbursement amount then has to be agreed upon again on the basis of the HEE [7].
2) If no added benefit was determined by the G-BA, the company can apply for an HEE – provided that the company bears the costs for the HEE [7].
3) In general, the Institute can be commissioned to assess questions of the quality and efficiency of services provided in the SHI system, which can also take the form of an HEE. According to §139b (2) SGB V, such a commission can also be awarded by the Federal Ministry of Health [7].

In addition, HEEs can be carried out in the context of the ThemenCheck reports (acc. to §139b (5) SGB V); see Chapter 6 and Section 6.5.2.

The aim of the HEE is to summarize economic information as a supplement to the benefit assessment, in particular for negotiations on the reimbursement price (in short: price negotiations) in the sense of an information synthesis. In addition to the calculation of the incremental cost-effectiveness ratio (ICER) and its contextualization, the HEE according to §35b SGB V also includes a budget impact analysis (BIA). The explicit recommendation of an appropriate price based on the consideration of all available comparators within an entire therapeutic indication is not the focus. The HEE is intended to create transparency about the cost-effectiveness ratios and is to be understood as an additional technical basis to support price negotiations. The HEE can complement the benefit assessment with an additional health economic perspective with regard to the main results of the benefit assessment.

For an HEE of drugs according to §35b SGB V, the G-BA on the one hand commissions IQWiG to conduct an HEE (Chapter 5, Section 2, §26 of the G-BA's Code of Procedure) [287]. At the same time, the G-BA requires the company to submit a complete dossier for the HEE (Chapter 5, Section 2, §27 of the G-BA's Code of Procedure) [287].

[23] Spitzenverband Bund der Krankenkassen, GKV-Spitzenverband (National Association of SHI Funds)

In other countries, HEEs of drugs in the context of reimbursement and pricing decisions are also mainly based on submitted dossiers [11,131,177,293,351,401,548,556,587,588]. In the context of an HEE according to §35b SGB V, modelling is usually carried out. Model development often requires information that is not covered by the content of the preceding benefit assessment (e.g. determination of utilities, additional data to determine transition probabilities, possibly available clinical study data relevant to resource use). In order for an HEE to be conducted in a meaningful and timely manner, the dossiers prepared by the company for the HEE must contain the necessary data.

4.2 Health economic evaluation of drugs according to §35b SGB V

According to SGB V, depending on the commission, IQWiG determines the methods and criteria for conducting an HEE on the basis of the international standards of evidence-based medicine and health economics recognized in the relevant expert groups [7].

The following text first describes the main components that an HEE in particular should contain in the context of the AMNOG procedure (acc. to §35b SGB V) in the form of a so-called reference case. Special features of HEEs outside the context of §35b SGB V are discussed in Section 4.15.

A reference case is a set of methodological standards for conducting HEEs, the use of which serves to ensure the comparability and quality of such analyses [302,643]. For example, the reference case is intended to provide guidance through standardized components, regardless of who conducts the HEE.

In contrast to this, the term “base case” in the context of HEEs describes the analysis with fixed, non-varying characteristics of the model parameters [681]. The base case forms the basis for the sensitivity analyses required to account for uncertainty [681].

Table 6 below provides an overview of the main components of the reference case. These are described in more detail in the following sections.

Table 6: Overview over the components of a reference case (multipage table)

HEE component	Explanation
Question / decision problem	▪ This is defined in the G-BA's commission[a], which specifies in particular the following components:
Population	▪ E.g., the population (or individual subpopulations or subgroups) eligible for treatment with the drug to be assessed
Intervention and comparator(s)	▪ Comparison between the drug to be assessed and the treatment option of the appropriate comparator therapy used in the preceding benefit assessment procedure for the demonstration of added benefit ▪ Separate analyses are conducted taking into account the treatment costs of other treatment options covered by the appropriate comparator therapy (sensitivity analyses in the sense of a cost variation). ▪ Depending on the commission, other comparators may be considered in the HEE. A new benefit assessment including the other comparators may be necessary.
Health economic outcome	▪ ICER: difference in costs divided by difference in the outcome assessed in the benefit assessment (weighted if applicable) or difference in QALYs
Perspective	▪ SHI-insured community
Time horizon	▪ Depending on the research question or commission, the available evidence and the perspective of the decision-maker, an appropriate time horizon is chosen that is sufficient to depict all relevant medical and economic consequences. ▪ In addition, a 5-year time horizon is considered in a sensitivity analysis.
Study type HEE	▪ Cost-utility analysis or cost-effectiveness analysis, each with an additional BIA.
Effectiveness and other clinical input parameters	▪ A benefit assessment according to §35a SGB V should already be available. Its results are included in the decision-analytic model, particularly for determining transition probabilities. ▪ The result of the HEE does not call into question the result of the benefit assessment. ▪ Development of the decision-analytic model usually requires additional clinical input parameters to depict the health care situation.
Utilities	▪ For the calculation of QALYs, the utilities used in the decision-analytic model should be based on valuations by patients. ▪ Utilities based on valuations by the general population are particularly useful if the valuations do not differ from those of patients.
Costs	▪ The costs should be as up-to-date as possible: The resources used should be relevant and justified. ▪ Direct (medical and non-medical) reimbursable and non-reimbursable costs should be recorded. ▪ The quantification and valuation of resource use may require a combination of different approaches (bottom-up, top-down, micro- and macro[gross]-costing approaches).
Inflation adjustment and discounting	▪ If cost data are from different time periods, they need to be adjusted for inflation. ▪ After the first year in the base case, costs and benefits that last longer than 1 year should be discounted at the identical constant rate of 3% per year.

Table 6: Overview over the components of a reference case (multipage table)

HEE component	Explanation
Study design	▪ A decision-analytic model is usually developed. ▪ The model structure is developed on the basis of, among other things, the decision on the benefit assessment and the outcomes that inform the decision, as well as on the basis of a health care pathway and previously published models. ▪ The choice of model type depends in particular on the research question, the characteristics of the treatments to be assessed, the disease considered, and the health care context. ▪ For extrapolation, e.g., of results from clinical studies, at least one conservative and one justified optimistic scenario should be considered. ▪ Full documentation of the model, consisting of a detailed description of the process and technical documentation, should enable the model to be fully reproduced.
Presentation of results	▪ The calculation of the ICER between the two treatment alternatives considered forms the main result, which is presented in tabular and graphical form. The tabular presentation also includes the absolute results per treatment and the corresponding incremental differences between the treatments. ▪ In addition, the results of separate analyses are presented, taking into account the treatment costs of other treatment options covered by the appropriate comparator therapy (sensitivity analyses in terms of cost variation).
Handling of uncertainty / sensitivity analyses	▪ Univariate and multivariate deterministic sensitivity analyses (DSAs) and, if necessary, scenario analyses or structural sensitivity analyses are performed and presented in tabular form and as a tornado plot. ▪ A probabilistic sensitivity analysis (PSA) is also performed. The result is presented as a scatter plot of the incremental cost-effectiveness. The scatter plot can also be presented as a cost-effectiveness acceptance curve (CEAC) or extended to include a confidence ellipse. ▪ In addition, sensitivity analyses are performed on methodological aspects, such as the definition of a fixed time horizon of 5 years, the assumption of constant identical discount rates between 0 and 5% and, if necessary, the investigation of differential discounting (costs 3%, benefit 1.5%).
Classification of results	▪ The classification includes the presentation of uncertainty in the form of the results of the sensitivity analyses. The aim is also to present the results of other HEEs with comparable questions.
Budget impact analysis	▪ The BIA is mandatory from the SHI perspective as a supplement to the HEE. ▪ The effects on the annual costs of the treatments considered for the SHI system are estimated over a period of at least 3 years, taking into account 2 scenarios (health care situation without the drug to be assessed; health care situation with the drug to be assessed).
Information retrieval	▪ The basic information retrieval procedure follows the procedure described in Chapter 8. ▪ Focused information retrieval is conducted for HEEs and, if necessary, for utilities. ▪ Exploratory searches can be conducted to determine costs in order to derive further input parameters relevant to the model or BIA. ▪ Other optional sources can be consulted (e.g. analyses of secondary data).

a. According to §35b SGB V, the G-BA's commission for the HEE must specify in particular the appropriate comparator therapy and patient groups for which the assessment is to be conducted, as well as the time period, the type of benefits and costs and the measure of overall benefit to be considered in the assessment [7].

BIA: budget impact analysis; G-BA: Gemeinsamer Bundesausschuss (Federal Joint Committee); HEE: health economic evaluation; SGB: Sozialgesetzbuch (Social Code Book); SHI: statutory health insurance; ICER: incremental cost-effectiveness ratio; QALY: quality-adjusted life year

4.3 Decision problem

The decision problem or the research question of the HEE is stipulated by the commission. In particular, the following content is specified: population, comparators, outcome, perspective, and time horizon [287].

4.3.1 Population

The population includes patients who are eligible for treatment with the drug to be assessed according to the approval status (target population). The extent to which the whole target population or, where appropriate, individual subpopulations or subgroups should be considered (e.g. the proportion of the target population of the drug to be assessed for which an added benefit was determined) is part of the specification of the commission.

If an HEE is commissioned for several subpopulations (questions) or subgroups, a separate analysis is performed in the HEE. An ICER is therefore calculated for each of these groups. When categorizing the results, additional information such as the size or proportion of these subpopulations or subgroups can be provided. IQWiG does not aggregate the results of these different groups into one cost-effectiveness ratio.

4.3.2 Intervention and comparator(s)

The AMNOG procedure focuses in particular on the assessment of the added benefit of drugs with new active ingredients. The added benefit of the drug to be assessed is demonstrated in comparison to the appropriate comparator therapy specified by the G-BA. If several alternatives are equally suitable as comparators, it is sufficient to demonstrate the added benefit over one comparator. In AMNOG, this can be the basis for a decision on the added benefit over the entirety of the (equally suitable) comparators specified by the G-BA. In contrast, the cost of all treatment options specified by the G-BA as the appropriate comparator therapy must be included in the dossier for the benefit assessment according to §35a SGB V (Chapter 5, Section 1, §9 of the G-BA's Code of Procedure) [287].

In the base case of an HEE, the comparison is made between the drug to be assessed and the appropriate comparator therapy used to demonstrate the added benefit. In the case of equally appropriate treatment options within the appropriate comparator therapy, separate analyses (sensitivity analyses in the sense of a cost variation) are conducted that take into account the range of costs of the additional treatment options covered by the appropriate comparator therapy (in particular the cheapest and the most expensive treatment; see Section 4.10).

While in practice HEEs often aim to compare the new drug with the most cost-effective therapeutic alternative, a comparison with the standard of care is a pragmatic, albeit simplistic, approach [248,803]. The appropriate comparator therapy is a treatment that is appropriate for the therapeutic indication according to the general state of knowledge.

Among other things, it should preferably be a treatment that is established in practice and is not excluded by the efficiency principle (Chapter 5, Section 1, §6 of the G-BA's Code of Procedure) [287].

Depending on the commission, other comparators that are not part of the appropriate comparator therapy may be considered in the HEE. If the HEE is to include a comparison with multiple comparators and not just with the appropriate comparator therapy, this valuation will not necessarily be made on the basis of the existing benefit assessment; a new benefit assessment including the other comparators may be required. In particular, this may also affect an HEE of orphan drugs for which a comparator may first have to be specified by the G-BA in the context of the commission for an HEE.

With regard to the interventions to be compared in the HEE, orphan drugs are a special case. According to §35a (1), Sentence 11, SGB V, the added medical benefit of an orphan drug is deemed to be proven by its approval. At the time of market access, therefore only a limited benefit assessment is conducted, without specification of an appropriate comparator therapy; the G-BA must always infer an added benefit, irrespective of the available evidence. According to §35a (1) Sentence 12, a regular benefit assessment and the associated designation of an appropriate comparator therapy will only subsequently take place if the turnover in the last 12 calendar months exceeds an amount of €30 million. According to the current legal situation, an HEE for orphan drugs is at least not excluded (see §35b SGB V, §130b (8) SGB V, and §35a (5a) SGB V). As with other drugs, an HEE can theoretically be conducted in 2 situations:

- following an arbitration decision as a result of a failure to agree on a reimbursement amount (at the request of the SHI umbrella organization or the company), or
- in the case of a commission initiated by the company following the conclusion that no added benefit has been proven.

While the former is possible both with a regular or a limited benefit assessment without specification of an appropriate comparator therapy, the latter is only possible if a regular benefit assessment has been conducted beforehand and on condition that the company bears the costs of the procedure itself. An HEE – in the sense of a comparative HEE – for orphan drugs also requires the designation of at least one comparator intervention by the G-BA in the associated commission. However, the challenge with a limited benefit assessment is that relevant comparative data are often not available. The frequent lack of data on the natural course of the disease and epidemiological data for rare diseases is equally problematic [579].

4.3.3 Outcome of the health economic evaluation

According to the legal requirements, among other things, it must be specified in the commission which type of benefit and which measure of overall benefit should be considered in the HEE [7]. For patient benefit, appropriated consideration should be given to improvement in health status, reduction in disease duration, increase in life expectancy, reduction in adverse effects, and improvement in quality of life [7].

Given these requirements, it makes sense to conduct the HEE in the form of a cost-effectiveness analysis, where costs can be determined in relation to individual outcomes of the benefit assessment. The focus here should be on calculating the ICER in relation to selected outcomes that are relevant for informing the decision. Particularly in cases where the decision on added benefit according to §35a SGB V is largely based on one dominant outcome from the benefit assessment, the costs in relation to this outcome could form the outcome of the HEE. However, it should be considered that not every outcome of the benefit assessment can be meaningfully represented in a decision-analytic model.

If several outcomes are commissioned as an outcome of the HEE, the decision-maker itself can weight the results for the different outcomes as part of the decision-making process. Another option is to weight the results taking into account methods to elicit patient preferences. The applicability of these preference elicitation methods, as well as the challenges and aspects to be considered are currently the subject of scientific debate [417,514,515,538,539]. An HEE according to §35b SGB V does not include preference elicitation surveys conducted by the Institute. Preference elicitation surveys submitted as part of an HEE dossier can be assessed depending on the commission, provided they contain the outcomes to be considered in the HEE.

Cost-utility analyses are a specific form of cost-effectiveness analysis, in which the costs are primarily set in relation to quality-adjusted life years (QALYs). The QALY concept combines the health state valuations in the form of utilities with quantitative aspects (remaining life expectancy or length of stay per health state) [64,246,248,586,779]. A multiplicative link is used to aggregate health state valuations with the length of stay per state into a one-dimensional outcome measure [586,779]. Various limitations of the QALY concept are discussed in the literature, in particular ethical, methodological and conceptual limitations [64,163,186,269,475,550,586,601,779]. This also includes a discussion about the question of whose preferences should be used as the basis for determining QALYs (see Section 4.6). Nevertheless, the QALY concept is currently the most widely used method that, explicitly and with a theoretically justified construct, enables the assessment of the effectiveness of treatments in HEEs, taking preferences into account [246,248,355,399,453].

4.3.4 Perspective

In the reference case, the HEE is conducted from the perspective of the SHI-insured community (see §35b SGB V) [7].

Depending on the commission, the (pure) SHI perspective, the social insurance perspective or the perspective of individual social insurance providers, as well as the societal perspective, may also be taken into account. The results of the valuation from an extended perspective are provided separately to the decision-maker.

The cost components to be included depending on the perspective are specified in Section 4.7. Care must be taken to ensure that the perspective is applied consistently in the analysis and that double counting is avoided.

4.3.5 Time horizon

The time horizon is specified in the commission for the HEE. It should be chosen so that it captures decision-relevant cost and benefit differences between the treatments of an HEE [560,609]. The appropriate time horizon is often longer than the period covered by the available primary data from prospective studies [445]. For chronic diseases in particular, even a lifetime horizon (or remaining life expectancy) is often considered appropriate [248,445]. When considering longer time horizons, assumptions must be made about the extent to which the treatment effect observed in the clinical studies will persist beyond the end of the study (see Section 4.9) [46,248]. Therefore, conclusions based on models with very long time horizons and extrapolations are associated with greater uncertainty [252]. The choice of time horizon therefore represents a trade-off between the desire to capture all expected or observed long-term effects of treatments and the uncertainty associated with extrapolation [222,352].

Therefore, an appropriate time horizon should be chosen, taking into account the therapeutic indication, the duration of the study, the long-term effects of the treatments considered, the period over which costs are incurred, the dynamics of treatment, and the perspective of the decision-maker [222,248,352,609]. In addition, a 5-year time horizon is considered in a sensitivity analysis.

4.4 Study type

Depending on the chosen outcome for reflecting the health change, different types of comparative HEEs are distinguished [210]. Because of the consideration of the difference in costs in relation to the differences in selected clinical outcomes or the difference in QALYs as the outcome for the reference case, the HEE is therefore conducted as a cost-effectiveness analysis or in the form of a cost-utility analysis, in each case with an additional BIA.

4.5 Effectiveness and further clinical input parameters

For the reference case, it is assumed that a decision on a benefit assessment according to §35a SGB V already exists and forms the basis for the HEE. The results of this assessment, in particular of the comparison between the drug to be assessed and the appropriate comparator therapy used by the company to demonstrate added benefit, therefore form the basis for the decision-analytic model, e.g. to derive transition probabilities.

The methods of the benefit assessment are described Chapter 3.

The result of the HEE does not call into question the result of the preceding benefit assessment according to §35a SGB V. This means, for example, that in situations in which no added benefit was demonstrated in the benefit assessment procedure, no added benefit is suggested by the HEE.

The development of a decision-analytic model usually requires additional clinical input parameters. These include, for example, data to determine the so-called transition probabilities for experiencing certain events or data on disease-related deaths or deaths due to another cause (so-called background mortality), see Section 4.14.

4.6 Utilities (in the case of a cost-utility analysis)

Utilities are needed to determine QALYs. These reflect health state valuations or transient events such as adverse effects [502]. Utilities can be calculated using direct or indirect methods [87]. With direct methods (e.g. time trade-off [TTO], standard gamble [SG]), respondents (e.g. patients, general population, health professionals, proxies) value their own (experienced) or a hypothetical health state [246,703]. Indirect methods consist of a classification system (e.g. EQ-5D, Short Form 12 [SF-12], SF-36, Health Utility Index [HUI]-2 or -3 questionnaire) to describe health states and an algorithm to calculate utilities for all states that can be described by the classification system, which is usually determined on the basis of a sample of the general population [87].

The systematic review by Helgesson et al [356], compares various arguments in favour of an assessment by the general population or by patients [356]. The respective justifications are partly based on similar, overarching arguments with different interpretations. There are structural similarities between the two positions, particularly with regard to difficulties in the assessment due to adaptation, focusing effects and distortions due to perspective [356]. The authors emphasize that no clear superiority of one of the two perspectives can be derived [356]. Regarding the question of which collective is the most accurate source of information for health state valuations, they ultimately argue in favour of those respondents with experience, i.e. who are either currently affected by the state to be assessed or who have been affected in the past [356].

Various studies show differences in health state valuations by patients compared with the general population [580]. Other confounding factors for differences in valuations appear to be the method used (TTO, SG), the time of the survey (early stage, ongoing disease process), and disease severity [87,91].

However, following the argumentation of Helgesson et al. [356], patients seem to be the most accurate source of information, as they can be expected to be better informed about their own health state. Furthermore, the inclusion of the patient perspective is of great importance to the Institute. The legal framework stipulates that certain aspects of "patient benefit" must be adequately considered in HEEs [7]. From this it can be deduced that for HEEs in the context of the AMNOG procedure, the utilities considered in the analysis should be based primarily on valuations by patients in order to be relevant to the decision-making context. If it is not possible to obtain patient utilities, for example due to the nature of their disease (e.g. dementia), these can be obtained by involving the patient's immediate environment (e.g. family members or carers). However, it should be noted that the evidence on valuations by patients may be limited due to the common international practice of preferring to use valuations by the general population. Nevertheless, studies to determine utilities in patient populations are possible.

One possible alternative is the visual analogue scale (VAS). The VAS is an instrument that is being discussed in the literature [90,574,721,744]. It is used, for example, as part of the EQ-5D instrument and allows for an assessment of one's own health state. Among other things, it has been pointed out that the VAS does not allow for a choice-based assessment of health states [89,90,204,574,744] and that it has limited usability if the scale is not anchored to the state of death [487]. Assessments of one's own health state using the VAS can be summarized into an aggregated measure by multiplication with the duration of the assessed state. However, because of the conceptual differences described above, this aggregated measure should not be readily labelled as a QALY [487]. Despite the discussions about the VAS described above, it is a commonly used method in clinical studies as part of the EQ-5D instrument for patients to assess their own health state and is therefore potentially useful for the HEE [133].

Another alternative is the use of utilities based on health state valuations by the general population. This is particularly useful if no differences in the valuations between the general population and patients have been demonstrated. If differences can be assumed to exist, there is an uncertainty in the results of analyses based on valuations by the general population that the results based on valuations by patients may be different. This must be taken into account when classifying the results of the HEE.

Valuations based on indirect methods should only be used if a validated tariff is available for Germany.

Furthermore, wherever possible, the same method should be used to determine utilities for all health states considered in a decision-analytic model.

Mapping is usually applied to approximate utilities from non-preference-based surveys using regression analyses. Ara et al. [25] point out that mapping can introduce additional uncertainty into an analysis and that this method should therefore be considered secondary to obtaining utilities. Therefore, mapping disease-specific instruments to generic instruments is generally not recommended for the HEE.

4.7 Costs

In principle, the costs should be as up-to-date as possible and the resources used should be relevant and justified in terms of their selection and quantity. In addition, the calculations and results must be presented in a transparent manner.

4.7.1 Cost categories

A) Direct costs

Direct costs refer to resource use in the current and future provision of health care services. They are further subdivided into direct medical costs and direct non-medical costs. Direct medical costs are understood to be resource use resulting from the provision of health care in the health care sector. They include, for example, costs of hospital stays, outpatient visits, drugs, and medical remedies and aids. Direct non-medical costs include resources that support the provision of health care services in the health care sector, such as travel costs for medical interventions or the estimated disease-related time invested by affected patients and their relatives [302,459,460].

Reimbursable costs include expenditure on health care services financed by the SHI or other social insurance providers. Non-reimbursable medical costs are services directly borne by the insured persons, such as co-payments for drugs, medical remedies and aids, and outpatient visits. Non-reimbursable non-medical costs are, for example, disease-related net income losses[24] (e.g. financial losses of patients receiving sickness allowance below their net income) or the time invested by affected patients and their relatives.

As time costs – depending on the perspective considered – cannot always be clearly allocated to direct non-medical costs or indirect costs, they should be reported separately [459].

[24] Strictly speaking the disease-related net income losses refer to the difference between the net income of healthy persons and that of sick persons, taking into account co-payments for health care services for the treatment of the disease. However, from the perspective of the SHI-insured community, co-payments are not considered as reimbursable costs, so that the net income losses can be calculated from the difference between the sickness allowance paid and the net income of a healthy person.

B) Indirect costs

Indirect costs refer to productivity losses due to incapacity to work, occupational invalidity (in the case of long-term disease or disability), and premature death or the reduced productivity in the case of paid and unpaid labour [465,466].

The Institute considers productivity losses primarily on the cost side. This is also widely recommended by the literature [104,210,422,466]. To avoid double counting, productivity losses due to premature death (mortality costs) should not be included on the cost side if mortality is already included on the benefit side [302,459].

Effects on leisure time in terms of reduced quality of leisure time due to the disease should only be presented in a sensitivity analysis if the available evidence allows and it can be assumed that such differences are not already depicted in the benefit parameter [103].

C) Transfer payments

Transfer payments are not costs in the economic sense, as they are not offset by any direct use of resources [210,655].

If the transfer payments overall do not represent additional expenditure within the selected perspective (both payers and recipients of transfer payments are included in the perspective), they should generally be omitted. Otherwise, transfer payments should be taken into account and shown separately in the presentation of results.

D) Intangible costs

Intangible costs are effects that cannot be directly calculated as resource use or valued in monetary units, such as pain or anxiety experienced by patients. These should be reported on the benefit side, if the data are available [713].

E) Future costs

Future costs are defined as resource use in terms of additional life years gained. In addition, the health economic literature often proposes a distinction between related and unrelated (future) costs. Related costs are, for example, the costs of drugs or check-ups after a heart attack, whereas unrelated costs are, for example, the costs of treating cancer that develops later and whose treatment is unrelated to that of the heart attack [173,713].

Depending on the time horizon considered, and if relevant differences are expected as a result of the intervention, the related future costs are included in the reference case. Unrelated future costs can be considered in separate sensitivity analyses [173].

F) Investment and implementation costs

If one-off costs are incurred explicitly for the SHI or the SHI-insured community to finance the provision or implementation of health care services, the investment and implementation costs should be taken into account in an appropriate way. This should be done through sensitivity analyses.

4.7.2 Steps for cost estimation

The estimation of the costs considered in the model usually follows a 4-step process:

- identification of resources
- quantification of resources
- valuation of resources and
- calculation of the costs considered in the model by health state and cycle, if applicable

Identification of resources

Identifying resources involves determining the health services used to treat the disease. The information should be as up-to-date as possible.

From the perspective of the SHI-insured community, direct (medical and non-medical) reimbursable costs as well as non-reimbursable expenses of the insured persons or their relatives (time costs) must be taken into account. Transfer payments that are merely redistributed from the SHI system to the insured persons (e.g. sickness allowance) are not taken into account [302].

As described in Section 4.3.4, further perspectives may lead to the consideration of deviating or additional cost components, depending on the commission.

In the (pure) SHI perspective, all direct reimbursable costs, including payments such as sickness allowance, are considered. In addition, and as far as relevant to the HEE, the shares of contributions to pension insurance, long-term care insurance, and unemployment insurance that the SHI has to bear for a sick insured person after 6 weeks of incapacity to work, as well as losses in contributions (during the payment of sickness allowance), can be taken into account.

In the social insurance perspective or in the perspective of individual social insurance providers, no co-payments by insured persons are calculated. Disease-related reimbursable expenses and transfer payments that do not cancel each other out within the selected perspective are taken into account.

In the societal perspective, cost components are considered regardless of who bears them and who is affected by the intervention. In general, costs should be considered that are incurred for all social insurance providers and other affected parties. Time invested by patients (and/or possibly their relatives) that represents a loss of working time is not counted again in terms of direct non-medical costs. This would lead to double counting if productivity losses were considered as indirect costs [713]. Likewise, transfer payments, such as SHI-funded contributions to social insurances, are not considered because they are merely redistributed and do not represent a resource use from an economic perspective [302].

Quantification of resources

The quantification of resources involves determining the frequency of use, the proportion of the relevant patient population that has used the respective service, and the duration of use. The costs of services that are used very infrequently or have only little impact on the results should be described, but do not necessarily need to be included in the calculation [210].

Both a micro- or macro(gross)-costing approach [697,698,733,734] can be used and combined to quantify resource use. The degree of precision of the quantification is determined, among other things, by the reimbursement system and the corresponding degree of aggregation of the services.

Both bottom-up or top-down approaches can be applied [654,697,698,733,734], i.e. 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).

Valuation of resources

Wherever possible, standardized methods should be used to valuate cost components, adapted as necessary to the perspective [70,86,461,656,666]. Depending on the cost area, resources are valued using regulated prices (e.g. per-case flat rates for the inpatient sector) or, for example, negotiated prices (drug prices).

Non-reimbursable costs are partly regulated, so that the valuation of resources can be based on the corresponding standardization (e.g. co-payment regulations in the inpatient sector and for drugs). These costs are shown separately in the perspective of the SHI-insured community. If the time invested by patients or their relatives is included in the cost estimation, it is generally valued at the average net wage.

If other perspectives are taken into account, in the sense of a pragmatic approach (e.g. in the societal perspective [656]), the reimbursement amounts of the SHI can be used as an approximation of the actual direct costs and supplemented by indirect costs (e.g. productivity losses).

Only aggregated data may be available in the social insurance perspective, depending on the insurance branch. In this case, the resources should be valued using a top-down approach based on the relevant statistics.

If the societal perspective is taken, the friction cost approach should as a rule be used to value productivity losses from paid labour [103,456,591]. Depending on the research question or in the context of sensitivity analyses, the alternative valuation of productivity losses using the human capital approach can be examined.

The valuation is generally based on average labour costs. The calculation of the average labour costs per working day is based on the weighted average labour costs of full-time and part-time employees in Germany. As an approximation, the "annual compensation of employees in Germany per year" divided by the "number of employees x 365" can be used (with Sundays and public holidays included in the days of incapacity to work). Applying this approach to the self-employed should be discussed [310]. Friction costs are assumed to be 80% of wage costs [456,465]. The duration of the friction period is determined on the basis of the average actual period within which a position can be filled in Germany [387]. If the human capital approach is to be examined in a sensitivity analysis, future productivity losses are calculated on the basis of the average age of patients until the normal retirement age is reached.

Productivity losses due to unpaid labour are generally valued at the average net wage (opportunity cost approach). In the context of sensitivity analyses, productivity losses can be valued using the gross wage rate of an adequate professional replacement (replacement cost approach) [302,459,465].

Calculation of the costs included in the model by health state and cycle, if applicable

The costs for the model are presented as average costs per patient by health state and, depending on the model, also by cycle and are calculated on the basis of several sources (see Section 4.14). Depending on the therapeutic indication, intervention and model, there may be no direct information available on the costs of the respective health states. In this case, the average cost of an intervention per patient and cost component (health services area and indirect costs, if applicable) for the observation period can be distributed across the different health states and cycles of the model using assumptions from other sources (see Section 4.14).

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

4.8 Inflation adjustment and discounting

If cost data are from different time periods, they must be adjusted for inflation. The consumer price index (*Verbraucherpreisindex*, VPI) of the German Federal Statistical Office should be used as a source for annual inflation [33,210,706]. Further price increase rates for individual health care areas may be considered from other sources within the context of sensitivity analyses.

If costs and benefits are incurred in periods that last longer than 1 year, after the first year in the base case they are discounted to the current period at an identical constant rate per year. [33,651]. This rate is set at 3%.

Likewise, identical constant rates of 0 and 5% should be used in sensitivity analyses. The literature also discusses the use of differential discounting in certain cases – especially when the consumption value for health is assumed to increase over time [150,423]. The effects of differential discount rates (3% for costs and 1.5% for benefits) can therefore also be examined in sensitivity analyses [423].

4.9 Health economic study design

HEEs can be conducted either alongside a clinical study (piggy-back study) [592] or by pooling data within a decision-analytic model [49]. However, individual clinical studies rarely contain all the information needed to conduct an HEE [380,668]. Often, they do not include any economic data or the available data are not sufficient to make a comprehensive estimate of the costs of an intervention. On the one hand, clinical studies rarely provide information on the long-term economic consequences of introducing a new intervention. On the other hand, they do not always adequately and completely address the health care aspects relevant to the cost side of the HEE. In addition, protocol-driven resource use in the context of the accompanying collection of economic data in a clinical study can lead to misinterpretations on the cost side. Because of the option to combine data from different data sources within a decision-analytic model (including the consideration of a piggy-back study) and thus to depict long-term consequences, HEEs based on a decision-analytic model are the standard for the HEE.

A decision-analytic model combines data on effectiveness, costs and other (e.g. clinical or epidemiological) parameters from different sources to calculate cost-effectiveness ratios of the treatments considered. In addition to combining data from different sources, decision-analytic models also explicitly aim to extrapolate results on effectiveness and costs beyond the study period [681] (see below).

Like mathematically formalized models, decision-analytic models are therefore also a simplified representation of reality. This means that the complexity of the factors and

variables relevant to the decision problem is deliberately reduced. In this way, a clear structure of the decision problem is achieved. The degree of complexity or the extent to which complexity is reduced depends on the respective question and cannot be determined a priori. The decision-analytic model should be as differentiated and complex as necessary to adequately answer the question(s) posed [134,683].

It is important to note that decision-analytic methods, including health economic modelling, are tools for decision-making under uncertainty [683]. This means that these tools are based on the explicit acceptance that decisions must be made despite the presence of uncertainty [210]. Therefore, decision-analytic modelling is more important when a decision needs to be made, but empirical studies are not available. Decision-analytic models and clinical studies are not competing alternatives; rather, decision-analytic models build on, among other things, what is measured in clinical studies, and also provide a framework for assumptions that are unavoidable in the context of decision-making [210]. Decision-analytic models do not aim to accurately predict reality, but allow interventions to be compared in well-defined artificial scenarios to obtain an estimate of their potential cost-effectiveness [294]. Nevertheless, it should be noted that there are credible and less credible models [112], so their critical evaluation plays an important role [210]. It should also be noted that, according to §35b SGB V, the decision on an HEE determines the added benefit and the treatment costs when using the respective drug. However, as described above, the objective of decision-analytic models differs from that of clinical studies. While clinical studies are based on empirical data, decision-analytic modelling uses a combination of different data sources and assumptions. Empirical data are used to support the validity of decision-analytic models. However, for the reasons described above, decision-analytic modelling is not suitable as a basis for demonstrating added benefit in the sense of §35a SGB V.

Basis for the concept and structure of the model

As a result, models must comprehensively depict the effectiveness and costs that arise in Germany for the interventions to be assessed. The following information is included in the model for this purpose:

- results on the effects (benefits and harms) of the interventions
- the disease- and intervention-related costs, and
- all aspects of the disease and treatment that may have a relevant influence on the effectiveness or cost components of the decision-analytic model (e.g. in the areas of demographics and epidemiology or aspects related to the health care pathway)

An important basis for any HEE is the development of a health care pathway (or several health care pathways) for the therapeutic indication in question. A health care pathway describes treatment processes for patients with specific therapeutic indications and structures them

according to sectors, professional groups involved, stages and other aspects, where appropriate. A health care pathway contains a brief description of the course of the disease and the health care provided in Germany for the therapeutic indication considered, including sources. For the relevant interventions and treatment steps in the various areas of health care services, the efficiency principle and – if applicable – the limits of the drug approval status must be taken into account. In addition, the assessment of the situation of use must be carried out within the requirements applicable in the SHI system. Furthermore, current treatment recommendations for Germany based on valid clinical practice guidelines must be included, if available. The HEE-relevant components in terms of a model-relevant health care pathway should emerge from the described health care context. If individual components are decidedly not taken into account, this should be justified.

An important basis for the development of the model structure is also the decision of the G-BA about the benefit assessment and the outcomes described in it.

In addition, published decision-analytic models dealing with similar questions are examined as a possible basis for model development [608].

Since even the most sophisticated decision-analytic models represent simplifications of reality with necessary assumptions and limitations regarding the content included [210], a model concept that represents the intended design in sufficient detail is necessary to ensure understanding and reproducibility.

Choice of modelling technique

The design of the decision-analytic model usually determines which modelling technique is used (decision tree, Markov model, discrete event simulation, etc.) [434,593,608,682]. The choice depends on the research question, the characteristics of the interventions to be assessed, the disease considered, and the general setting. The decision-analytic model should be as sophisticated and complex as necessary to answer the question(s) posed. The available evidence alone should not determine the choice of modelling technique [210]. The chosen modelling technique should also be compared with previously conducted/published decision-analytic models or closely related decision problems and, if deviations from existing models are identified, discussed. As the appropriate modelling technique always depends on the underlying question, fixed a-priori specifications are not useful.

Validation

In the field of decision-analytic models, the term “validation” refers to the process of assessing the model in question in the context of a specific situation of use to determine whether it adequately and sufficiently represents the system it is intended to depict [758-760]. In contrast to the previous sections on benefit assessment, which are concerned with demonstrating validity (see in particular Section 3.1.2), the validation status of decision-

analytic models is more a relative rather than an absolute judgement [260]. The judgement depends on what validation efforts have been made and how these affect confidence in the model [260]. Model validation helps to distinguish credible from less credible models based on the presence of errors and biases or a limited depiction of the decision problem [635].

A decision-analytic model that is valid for one question may not be valid for another [480]. Therefore, the validation process must cover each intended use of the model, and validation must be repeated if the model is used for other questions.

A key component of validation is whether the content of the decision-analytic model adequately depicts the reality of disease progression and treatment. The plausibility check (face validity) relates to the model concept, data collection, the development of functional relationships, and the choice of the modelling technique. To ensure face validity, it is particularly useful to involve clinical experts in the development of the model.

Another important aspect of validation is the correct technical implementation (internal or technical validity). This aspect refers to the question of whether the technical implementation correctly realizes the model concept.

Predictive validity should also be mentioned: To what extent does the decision-analytic model predict the future, i.e., do the predicted results correspond to reality? This is certainly the most desirable form of validity, but also the most difficult to realize, if it is possible at all [219]. However, a comparison with previous, comparable observations is useful, and differences should be explainable. This also applies in comparison with other decision-analytic models (so-called cross-validation).

In HEEs, the responsible analysts sometimes already carry out extensive model validation measures themselves for procedural reasons [758]. However, as validation plays an important role in building confidence in the decision-analytic model, it is also necessary for the users of the models to carry out validation measures – particularly in view of the fact that modelling may be influenced by financial interests [260]. The so-called AdViSHE tool (Assessment of the Validation Status of Health-Economic decision models) [260] provides a structured overview of specific validation measures. On the one hand, it can provide users with an overview of potential measures that they could take to improve their confidence in the model results. On the other hand, AdViSHE provides the analysts who develop the decision-analytic model with a way to transparently document the measures they have taken, thus also contributing to the credibility of the model.

Extrapolation

As the time horizon considered in the decision-analytic model is usually longer than the duration of the clinical studies used for the benefit assessment, it may be necessary to

extrapolate the results of these clinical studies. For this extrapolation, at least the following two scenarios should be considered:

- a conservative scenario in which it is assumed that the effect/difference that was shown in the clinical studies between the treatments considered will not persist beyond the end of the clinical study, and
- a more optimistic scenario in which it is assumed that the effect/difference that was shown in the clinical studies between the treatments considered will persist for a certain period of time after the end of the clinical study

The underlying assumptions must be justified taking into account the available evidence.

Depending on the time horizon considered and the evidence available, one of these scenarios forms the base case of the analysis. The other scenario or scenarios are then presented as part of the sensitivity analysis.

Documentation

Thorough documentation is of major importance for decision-analytic models. A clear and transparent presentation allows interested parties to understand the model and thus contributes decisively to the credibility and usability of the model [220].

On the one hand, the documentation includes a general, clear presentation of the approach, including the justification of the decisions made and the choice of data (sources). On the other hand, technical documentation is required that describes the functional / mathematical relationships of the decision-analytic model. This must allow an expert to fully reproduce the model results [220].

If an HEE is to be conducted on the basis of a submission of a decision-analytic model, an electronic version of the model must be available. This must be fully accessible and allow all formulas and relationships of the analysis to be viewed as part of the evaluation. In addition, the electronic version must allow for adjustments (e.g. to perform additional sensitivity analyses or to include other data sources). To facilitate the exploration of the decision-analytic model, the electronic version should be accompanied by a user guide that describes, for example, how to change input data, where to find the respective input data, and how to run the model and extract the results.

4.10 Presentation of results

The ICER is the result of the HEE. In the base case, the ICER is first determined for the comparison between the drug to be assessed and the appropriate comparator therapy used by the company to demonstrate added benefit. It can be presented in both tabular and graphical form. The tabular presentation includes both the absolute results – i.e. costs and

outcome separately for the treatments considered – and the corresponding incremental differences between the treatments. Due to the particular relevance to the question of pricing in the context of the AMNOG procedure, the effects of a variation in the price of the drug to be assessed on the result are to be examined first.

The appropriate comparator therapy may include treatment options with markedly different treatment costs. However, it is possible that the appropriate comparator therapy used by the company is not the most efficient of the treatment options included. According to §130b (3) SGB V, if several alternatives are specified as the appropriate comparator therapy, the one that represents the most efficient alternative in terms of annual treatment costs is to be used when agreeing on a reimbursement amount. Against this background, the range of costs of different options covered by the appropriate comparator therapy should also be considered in an HEE. For this purpose, separate analyses are performed when several equally appropriate treatment options have been specified as the appropriate comparator therapy. These are technically performed as sensitivity analyses in which, in contrast to the base case described above, the costs of the comparator are varied, taking into account the treatment costs of the treatments that are also covered by the appropriate comparator therapy (in particular, taking into account the cheapest and most expensive of the treatments viewed to be equally appropriate). However, the analyses described are not primarily intended to examine the robustness of the results, but are relevant analyses in the context of the AMNOG procedure.

4.11 Handling of uncertainty / sensitivity analyses

The literature distinguishes between different forms of uncertainty that need to be addressed in decision analyses [95]. In the context of an HEE, this is dealt with in particular in the form of sensitivity analyses (see below). Table 7 presents different concepts of uncertainty in the context of HEEs.

Table 7: Concepts of uncertainty in health economic decision analysis (adapted and translated from Briggs et al. [95])

Designation	Concept	Other designations in the literature	Comparable concept for regression models
Stochastic uncertainty	Random variability in the result for the same patients	Variability, first-order uncertainty	Error term
Parameter uncertainty	Uncertainty in the estimation of the respective parameter	Second-order uncertainty	Standard error of the estimator
Structural uncertainty	Underlying assumptions in the model	Model uncertainty	Type of regression model, e.g. linear, log-linear

The uncertainty of many model parameters arises from the fact that their value is itself an estimate based on a sample. This type of uncertainty is often expressed by means of confidence intervals or other statistical approaches to describing variability.

- For estimated effects in the context of a benefit assessment, confidence intervals or credible intervals (if Bayesian methods are used; see Sections 9.3.2 and 9.3.8) can usually be calculated to indicate the precision or uncertainty of the point estimates. Appropriate assumptions need to be made for further exploration of uncertainty, as many effects are not normally distributed. Estimates from indirect comparisons (see Section 9.3.8) are subject to greater uncertainty than estimates from direct comparisons; this needs to be noted in the assessment of uncertainty. Scenario analyses may need to be performed for estimates from indirect comparisons that differ, e.g. due to different assumptions about a-priori distributions.
- Uncertainty in cost data needs to be addressed appropriately. Costs are inherently continuous, positive, without an upper limit, and generally not normally distributed, but usually right-skewed [41]. Costs may also be subject to uncertainty regarding assumptions about resource use, such as the dosage of a drug over time.
- Utilities are important parameters in HEEs that consider QALYs. The uncertainty associated with utilities also needs to be adequately taken into account. It should be noted that utilities are special parameters in terms of their range of values, as they can theoretically take negative values (conditions that are perceived as worse than death) and have a maximum value of 1 (perfect health) [94].
- The uncertainty of epidemiological data should also be taken into account. This includes, for example, data on baseline risk and mortality.

In addition, there may be variability between different patients due to their characteristics [95,414]. As already discussed in Section 4.3.1, differences between patient groups in HEEs may need to be considered separately.

In cases where a model is also stochastic (i.e. based on random numbers and Monte Carlo draws), various techniques can be used to limit this type of uncertainty [481,678].

Uncertainty can also arise from the variability in the model structure (structural uncertainty). The extent to which the functional relationships of the model are actually valid and whether other functional forms may be more appropriate should be scrutinized here [96,329].

Similarly, predefined input parameters, such as the discount rate, can be varied in order to reflect the uncertainty arising from different assumptions.

Sensitivity analyses

In HEEs, both univariate and multivariate deterministic sensitivity analyses (DSAs) and probabilistic sensitivity analyses (PSAs) should be performed to account for parameter uncertainty. The recommendations of the joint Modeling Good Research Practices Task Force

Working Group of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the Society for Medical Decision Making (SMDM) can be used as a guide [95].

In particular, the aim of DSAs is to identify the parameters that have a particular influence on the result of the analysis, e.g. transition probabilities, utilities, costs, etc. As mentioned above, sensitivity analyses can also be used to examine the influence of certain methodological aspects on the model result.

All DSAs are fully documented with the minimum and maximum values for the parameters used and the underlying assumptions. For PSAs, the following aspects in particular are specified: probability distributions and their sources and correlations between input parameters.

The results of the DSA should be presented in tabular form. In addition, the influence of the varied parameters can be presented in a tornado plot [96].

The results of the PSA are presented in the form of a scatter plot of the (incremental) cost effectiveness (see, e.g., [94]). In addition, the result of the PSA can be presented in the form of a cost-effectiveness-acceptability curve (CEAC) or the scatter plot can be extended to include a confidence ellipse [94].

Structural sensitivity analyses aim to determine the impact of varying assumptions in the model structure, for example on the number or type of health states.

4.12 Classification of the ICER

For the HEE, the appropriateness and reasonableness of cost coverage by the SHI-insured community must be taken into account in accordance with the legal requirements [7,287].

To assess appropriateness, the HEE provides the following information to the decision-maker: The main result is the ICER, which is also categorized. This includes the presentation of uncertainty in the form of the results of the sensitivity analyses (see Section 4.11). The aim is also to present the results of other HEEs with comparable questions. These should then preferably include modelling for the German context. The presentation of HEEs from other countries should be examined on a commission-specific basis.

The reasonableness of cost coverage by the SHI-insured community depends on the one hand on the appropriateness of the price of a treatment, but on the other also on the associated future total expenditure, depending on the financial ability and willingness to pay of the SHI-insured community. In order to illustrate the future financial effects of cost coverage, a BIA should be carried out, which can serve the decision-maker as an information basis for the decision on reasonableness.

4.13 Budget impact analysis

In addition to the HEE, the BIA provides information that supports the decision-maker in assessing the financial reasonableness of an intervention. A BIA is an assessment of the direct financial consequences associated with the reimbursement of an intervention in a health care system [747]. In particular, a BIA predicts how a change in the mix of treatments for a particular disease may affect future spending in the health care system [717].

The BIA estimates the expected impact of the reimbursement of the new intervention on annual expenditure from the perspective of the SHI or another relevant payer over a period of at least 3 years. In principle, the costs incurred over this period are compared for at least 2 health care scenarios:

- a scenario in which the target population defined in the commission is treated only with the existing treatment options and not with the drug to be assessed (often referred to as the reference scenario), and
- a further scenario in which the drug to be assessed is also available and is part of the health care service provided

Both the total expenditure per scenario and the incremental expenditure for each year of the analysis period are determined.

As the input parameters of a BIA often cannot be determined without various assumptions and / or estimates, the benefit of such an analysis lies less in the calculation of a single figure, but rather in the outline of a robust calculation framework that can be used to analyse the influences of the various parameters [717]. In particular, the determination of the number and characteristics of patients in the target population, their change over the observation period, the current proportions of the different treatment options in the health care services provided, the expected market penetration of the new treatments, and the associated change in the current proportions require assumptions and/or estimates that introduce uncertainties in the informative value of the results. The impact of these uncertainties on the BIA can be illustrated using additional scenarios. In this context, the existing uncertainties regarding the input parameters or the calculation framework are taken into account through alternative assumptions and / or estimates.

As reliable information on the full disease costs per intervention is rarely available, and a separate estimate can be very burdensome (and may not be proportionate to the additional knowledge gained), a pragmatic approach is to focus on the costs directly related to the use of the treatment. Costs should be estimated using the methods described in Section 4.7.

If there is a justified expectation that a new intervention will lead to significant differences in disease costs (excluding the direct costs associated with its use), this aspect may be considered

qualitatively or as part of further scenario analyses, depending on the availability of reliable data.

4.14 Information retrieval and data sources in the HEE

The data included in the HEE can be collected in different ways and originate from different sources.

For the reference case, it is assumed that a benefit assessment according to §35a SGB V is already available. The procedure for information retrieval is described accordingly. If a benefit assessment is required, the information retrieval procedure corresponds to the procedure described in IQWiG's General Methods (in Sections 8.1 and 8.2).

Comprehensive information retrieval for every parameter of an HEE is not efficient and generally not necessary [88,222]. Therefore, so-called focused or exploratory searches are usually conducted [88,222,432] or an iterative process is used to identify the required information.

Focused information retrieval

As a basis for selecting models and categorizing results, for the HEE mandatory, focused information retrieval must be conducted in bibliographic databases (see Section 8.2.2) and, if necessary, other sources such as searching for HEEs on the websites of other HTA agencies. Utilities may be determined within the framework of clinical studies. Alternatively, or in addition, focused information retrieval for utilities for the health states considered in the model may be necessary.

Where available, validated study filters are used. This applies, for example, to the search for HEEs [295] or for utilities [27].

Exploratory searches

Within the HEE, exploratory searches may be needed to determine costs (e.g. resource use, prices), to derive other input parameters relevant to the model (e.g. transition probabilities for experiencing certain events, background mortality, extrapolation) or for the BIA. Sources of information may include bibliographic databases (e.g. to search for cost analyses or epidemiological information), guideline databases or other sources (e.g. websites, clinical information systems and other databases).

If clinical practice guidelines are used (e.g. for the development of the health care pathway), they should be evidence-based and relate to the German health care system or originate from a country with health care structures comparable to those in Germany. If no evidence-based guidelines are available for the therapeutic indication to be investigated, it must be considered

and transparently explained whether other German guidelines can be used or whether expert surveys should be used.

Resource use needs to be determined and evaluated as part of the cost estimation process. Different sources can be used to calculate resource use [656,666]. Price catalogues or directories should be used to assess this resource use (e.g. database of the Information Service Provider for the Pharmaceutical Market (IFA[25]), Uniform Value Scale[26], Catalogue of Diagnosis-Related Groups) [70,86,461,656,666]. These must be up-to-date and reflect the prices relevant to the SHI system.

Additional sources

Analyses of secondary data in the form of routine SHI data are a potentially relevant source for the HEE, particularly in the context of cost estimation (e.g. to identify the cost components to be considered or to calculate resource use) or, if necessary, to determine additional input parameters (see Section 4.5). These analyses may, for example, be conducted by IQWiG itself within the context of the HEE or provided by the company as part of the dossier. The analyses should be based on as representative a sample as possible, follow the guidelines and recommendations on good practice for secondary data analyses [26] and be reported according to the reporting standard for secondary data analyses (STROSA[27]) [724].

The calculation of costs may be supplemented by, among other things, expert surveys, especially if secondary data do not adequately depict health care in all states of the decision-analytic model.

In addition to cost estimation, supplementary expert surveys may be used to identify additional potentially relevant factors. Expert surveys may, for example, be conducted by IQWiG itself as part of its own HEE or may be submitted by the company as part of the dossier. The conduct of expert surveys follows the generally accepted methods and procedures of quantitative social research and requirements for the reporting of these surveys [385]. In particular, details must be provided on the recruitment, number and expertise of the experts, the conduct of the survey, the aggregation of individual responses, the type of consensus finding, and the reporting and handling of the results [385,718].

The following table provides an overview of the key components of information retrieval for the reference case of the HEE.

[25] Informationsstelle für Arzneispezialitäten

[26] Einheitlicher Bewertungsmaßstab

[27] Standardisierte Berichtsroutine für Sekundärdatenanalysen

Table 8: Overview of key searches in the context of the HEE (excluding searches in the context of the benefit assessment[a])

Objective of information retrieval	Type of information retrieval			
	Focused information retrieval	Exploratory information retrieval	Further sources, e.g.	
			Secondary data analyses	Expert survey[b]
Health economic evaluation	X			
Utilities	(X)			
Cost estimation / BIA		X	(X)	(X)
Additional input parameters		X	(X)	(X)
Additional factors				(X)

a: For the reference case, it is assumed that a benefit assessment according to §35a SGB V already exists. The information retrieval process is described accordingly.
b: Potentially in addition to the other searches

BIA: budget impact analysis; X: obligatory; (X): optional

4.15 Health economic evaluations outside §35b SGB V

As explained in Section 4.1, HEEs can also be conducted outside the framework defined by §35b SGB V. Depending on the research question, the methodological requirements defined for the reference case (see Section 4.2) essentially apply to these HEEs.

Again, the health economic question must be specified at the outset, particularly with regard to the population, the treatments and comparators to be assessed, the outcome, the perspective and time horizon. The treatment to be assessed may also be compared with several comparators. It is not necessarily assumed that a completed benefit assessment already exists. This benefit assessment can also be carried out as part of the HEE. This may then need to be taken into account when determining the timeframe for the HEE. The requirements in Chapter 3 for conducting the benefit assessment apply. The use of a universal threshold value is also not intended for HEEs outside the framework defined by §35b SGB V. The results can be classified taking into account uncertainty in the form of the results of the sensitivity analyses (see Section 4.11) or the presentation of the results of other HEEs with a comparable research question. The performance of a BIA is optional.

When more than 2 treatments are considered, the calculated cost-effectiveness ratios can be plotted as one or more efficiency frontiers. The efficiency frontier graphically compares the benefits of the treatments considered with their costs [395]. The treatments that are the most efficient in terms of benefits and costs (taking into account the criterion of absolute and extended dominance) form the efficiency frontier [395]. It is not necessary to extend the last segment, which was previously intended for the derivation of a price recommendation.

5 Analyses of health care

5.1 Evidence searches for guidelines

5.1.1 Background

Clinical practice guidelines are systematically developed normative action and decision aids for service providers as well as patients enabling an appropriate approach to specific health problems. Their aim is to improve patient care [787]. Ideally, their recommendations are informed by a systematic review of the evidence and an assessment of the benefits and harms of alternative treatment options [267,311]. In addition to all relevant professional groups involved in the health care provided, guideline groups should also integrate affected persons as well as persons with methodological experience in guideline development.

With the Digital Health Care Act (DVG[28]), it was stipulated that the AWMF can propose clinical practice guidelines to the Federal Ministry of Health for which IQWiG is to support guideline development or updating with evidence searches [5]. An evidence search usually includes the production of several evidence reports.

5.1.2 Evidence reports

The evidence report answers the question as to what evidence exists regarding the effects of a specific intervention compared with a control intervention in a clearly defined population. This research question is called the PICO[29] question. It is addressed based on the methodological requirements of GRADE [660]. As a result, (quantitative) evidence profiles are generated. In contrast to the benefit assessment described in Chapter 3, in which the effects of an intervention are evaluated on an outcome-related basis and a conclusion is inferred regarding benefit or harm, such a conclusion is not inferred in the evidence reports. This final step is the responsibility of the respective guideline groups within the development of recommendations for action.

Qualitative evidence reports can also be produced with the aim of answering questions about the patient or family perspective, feasibility, or acceptance of this intervention [533]. For this purpose, on the basis of qualitative studies qualitative evidence profiles are generated roughly following the GRADE CERQual[30] instrument [488]. Instead of the PICO scheme, the PICo[31] scheme is used [498].

[28] Digitale-Versorgung-Gesetz

[29] Population, intervention, comparison, outcome

[30] Confidence in the Evidence from Reviews of Qualitative Research

[31] Population, phenomena of interest, context, others / outcome

5.2 Guideline synopses

5.2.1 Background

Guidelines can include recommendations for action in one or more areas of the health care chain (prevention, diagnosis, therapy, rehabilitation and aftercare) for one or more specific health problems. To provide an overview of the most important evidence-based recommendations for the health care of patients with a selected clinical picture across the entire health care chain, guideline synopses summarize the recommendations of national and international guidelines of high methodological quality that were identified in a systematic search. In this way, they help to bundle information on the quality of health care to be aimed for in a health care system [342].

Guideline synopses are particularly suitable to provide an overview of normative health care requirements for selected diseases. Among other things, they serve the G-BA as a basis for developing and updating the requirements for structured treatment programmes for chronically ill people (disease management programmes [DMPs]). These are described under various items in the DMP Requirements Directive (DMP-A-RL[32]) [287,290].

5.2.2 Evidence-based guidelines

Evidence-based guidelines are normally used to answer questions on health care standards within the framework of guideline synopses. Guidelines are referred to as evidence-based if their recommendations are based on a systematic literature search and selection, if their recommendations are generally linked to a grade of recommendation (GoR) and/or level of evidence (LoE), and are linked to citations of the underlying primary and/or secondary literature (modified according to AGREE[33] [13,14]). An evaluation of these formal criteria, but no evaluation of content, is performed for the inclusion of guidelines in the guideline synopses.

5.2.3 Applicability to the German health care system

If national as well as international guidelines are searched for in the production of guideline synopses, it is in principle assumed that guidelines from member states of the OECD are particularly relevant. The OECD was originally established in 1961 as the follow-up organization of the Organization for European Economic Co-operation [566]. Besides a high per-capita income by global standards, these member states have high-performing health care systems. Since 2003 the OECD has been aiming to compare the quality of health care in the

[32] DMP-Anforderungen-Richtlinie

[33] Appraisal of Guidelines Research and Evaluation in Europe

different member states via the common reporting of selected quality indicators (Health Care Quality Indicators [HCQI] project [521]).

In principle all guidelines from the OECD member states can be considered. However, the applicability of recommendations from international guidelines can be problematical if they represent specific recommendations that, for instance, are not compatible with the funding principles of the German health care system or with the special features of the SHI system in Germany with regard to the legal regulations for professional conduct or for the provision of health care services. This is considered in the synopses of the guidelines, where necessary.

5.2.4 Appraisal of methodological guideline quality

An important aspect in the interpretation and appraisal of guideline recommendations within the framework of the guideline synopsis is the appraisal of the methodological quality of the underlying guidelines. No examination of content quality is carried out.

On an international level different instruments are used for the appraisal of methodological guideline quality [684,768]. Special attention is paid to the AGREE instrument [13,508] and its further development (AGREE II instrument) [14,105-107]. It was developed by an international group of researchers and is the most widely used instrument internationally. It is therefore also used in guideline synopses to appraise the methodological quality of the guidelines included.

The AGREE II instrument [14] consists of 23 appraisal criteria (items) assessed by means of a multi-step scale. These items are organized in 6 domains, which are independent of each other; each domain covers a separate dimension of methodological guideline 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

While the AGREE instrument [14] specifies the calculation of standardized domain scores for each of the 6 domains, its use within the framework of guideline synopses is limited to Domains 2,3 and 6. This is done with a view to the objective of the guideline synopses according to SGB V and the Rules of Procedure of the G-BA [287] in order to provide an evidence-based basis for updating existing DMPs and developing new ones. The limitation to Domains 2, 3 and 6 is also undertaken by other authors [62].

Each guideline appraisal is performed by 2 reviewers independently of each other. Each standardized domain score is presented in the guideline synopsis. AGREE does not provide any thresholds to distinguish between methodologically strong and methodologically weak guidelines [14]. However, some users of the AGREE II instrument make recommendations for the use of guidelines based on the standardized domain scores, using 2- and 3-level systems. In the 3-level system, guidelines with domain scores below a specific value – but varying according to the user group – are considered weak or not recommended [368]. Following this procedure, guidelines that achieve standardized domain scores < 30% in 1 or more of the 3 domains considered are marked in the guideline synopses. This also applies to results based exclusively on such guidelines.

The results of the AGREE II appraisal are thus not a criterion for the inclusion of guidelines in a guideline synopsis, but marking is used to transparently show whether the evidence-based guidelines included in a guideline synopsis show particular methodological strengths or weaknesses.

5.2.5 Structured processing of recommendations: levels of evidence and grades of recommendation

A guideline recommendation is defined as a proposal for action concerning the clinical decision in a specific situation or for system decisions. The recipients are generally professionals. In guidelines those statements are generally identified as recommendations that are clearly formally indicated as such by the guideline authors.

The authors of evidence-based guidelines use different systems to classify the LoEs or GoRs of their recommendations [31,217,327,469,663].

GoRs provide information on the strength of a recommendation. They are generally based on a weighing of the benefit and harms of a (medical) intervention and on the specific health care context based on an evaluation of the respective evidence. LoEs focus on the internal validity of the underlying studies; in this context, systematic reviews of RCTs usually receive the highest LoE. The systems of evidence classification may attribute a differing relevance within the LoE classification to clinical and epidemiological studies, the characteristics of study conduct, and the respective risk of bias [31,109-111,327,663].

As there is to date no internationally consented standardization of grading systems for evidence and recommendations, the LoEs and GoRs used by the individual guideline developers are generally noted in the guideline synopsis and the corresponding grading systems documented. In order to better compare the different systems of different author groups of guidelines, the classification of GoRs and LoEs for the guideline synopses is simplified by transferring them into a standardized category system. In this context, a distinction is made between “high”, “not high” and “unclear” categories.

In the Institute's work, the grading system from the procedure of the National Health Care Guideline (NVL[34]) is relevant for the assessment of GoRs [116]; the evidence classification applied by the G-BA is used for the classification of LoEs [287]. A high GoR is assigned if the measure recommended can be assigned to the strength of recommendation "A" (strong recommendation). All other recommendations are assigned to the category "non-high or unclear GoR". An LoE assigned by the guideline authors is classified as high if the LoE is based on at least one RCT. This corresponds to evidence levels Ia and Ib of the evidence classification used by the G-BA.

If the guideline authors use a classification system according to the GRADE Working Group [323,324,327], the highest evidence level according to GRADE is generally assigned to the category "high LoE". All other LoEs provided by the guideline authors that cannot be assigned to the category "high" are assigned to the category "non-high or unclear LoE".

The category "unclear" is assigned if

- a GoR specified by the guideline group cannot be categorized according to the recommendation grading system of the NVL programme
- an LoE specified by the guideline group cannot be categorized according to the evidence classification of the G-BA or GRADE
- the GoR / LoE specified cannot be clearly assigned to a recommendation or
- no GoR / LoE is specified.

5.2.6 Structured synthesis of information: extraction and analysis of recommendations

The basis for the synthesis of recommendations within the framework of the guideline synopsis are current evidence-based guidelines. In the first step, the recommendations of the guidelines are compared with the content of the currently valid DMP-A-RL. In this context, the recommendations are assigned to the respective statements of the DMP-A-RL (so-called health care aspects).

In the case of existing DMPs, it is checked when assigning the recommendations whether their content is discrepant to the DMP-A-RL. For the further analysis, only those health care aspects are considered that contain recommendations discrepant to the DMP-A-RL.

The recommendations of these health care aspects are assigned to both the evidence and recommendation levels according to Section 5.2.5. The methodological assessment is based on this: Only health care aspects that contain recommendations with at least one high GoR or alternatively only recommendations with an unclear GoR (whereby at least one of these

[34] Nationale VersorgungsLeitlinie

recommendations must have a high LoE) are considered for the development of key statements and are presented in the guideline synopsis. For this purpose, the main content of the recommendations is summarized per health care aspect (key content).

The discrepancy between the health care aspect and the DMP-A-RL is explained briefly and concisely in the results tables. If necessary, these explanations are supplemented by methodological and other scientific information, for example, on relevant details that deviate from health care in Germany or scientific discourses.

In the case of guideline synopses on new DMPs, the procedure is the same, with the exception of the comparison with the DMP-A-RL.

5.3 Minimum volume regulations

5.3.1 Background

The potential correlation between defined volumes of services in (mainly inpatient) facilities and quality has been investigated for more than 40 years [520]. Since 2003, the G-BA has set so-called binding minimum volume standards for hospitals in accordance with §137 (3) Sentence 1 No. 2 SGB V – old version, now §136b (3) Sentence 1 No. 2 SGB V – for certain inpatient elective services: Hospitals may only provide and bill the corresponding services if the competent hospital operators convincingly demonstrate to the regional associations of the health insurance funds that they will at least meet the minimum volume standards in the next year on the basis of the volume of services achieved in the previous year.

These minimum volume regulations are binding for hospitals licensed according to §108 SGB V and specify in which case a hospital site may provide the services for which minimum volumes have been specified [285]. However, some exceptional rules apply. For example, emergencies, as a matter of principle, remain unaffected by the minimum volume regulation. Or the authorities of the federal states (“Länder”) responsible for hospital planning can determine exceptional rules for those services where the application of the minimum volume regulation could endanger guaranteeing the provision of area-wide health care for the population.

In order to specify a new minimum volume or to review an existing one, the G-BA can commission IQWiG with the scientific assessment of the correlation between volume of services and quality for a specific medical procedure (§136b (1) Sentence 1 No. 2 SGB V and Chapter 8 §16 (5) Sentence 1 No. 1 of the Rules of Procedure of the G-BA [287]).

5.3.2 Information basis and assessment

The investigation of the correlation between the volume of services and the quality of the treatment outcome is based on appropriate observational studies or controlled intervention studies. Results for different target variables are considered. These may relate to mortality,

morbidity, and health-related quality of life, including activities of daily living and dependence on the help of others.

Since the quality of the treatment outcome of a medical intervention can be influenced by numerous individual risk factors of the patients, as well as by the quality of the intervention itself, only those observational studies are included in which relevant confounding factors were controlled for (risk adjustment).

The informative value of the results of observational studies is assessed on the basis of quality criteria developed specifically for studies evaluating volume-outcome relationships [59,389,390,778]. Various issues are examined in this context (data sources used, adequacy of the statistical models, appropriate risk adjustment and completeness of reporting). On the basis of this qualitative assessment, a quality classification of the observational studies into studies of high and low informative value is performed.

In addition, it is explained how a binding volume of services specified for health care (per hospital, per doctor or per combination of hospital and doctor) has affected the quality of the treatment outcome.

If controlled intervention studies are included, the specification of a minimum case volume is considered to be the intervention to be tested. The risk of bias of the controlled intervention studies included is assessed as described in Chapter 9 (see Section 9.1.4).

The results on the target variables reported in the studies are compared and described in the report. Where possible, suitable meta-analytical procedures are used in addition to the comparison of the results of the individual studies (see Section 9.3.7).

5.4 Analysis of data on health care provision

5.4.1 Background

Data on health care provision cover secondary data that describe health care. Secondary data are characterized by the fact that the data are analysed beyond their original, primary purpose. The subsequent use beyond the primary reason for data collection is decisive here [726]. This includes, for example, routine data from SHI funds, data from registries on epidemiological, clinical or intervention-related research questions, or data from population-related health surveys conducted by organizations such as the Robert Koch Institute. In varying detail and depending on the commission, the analysis of data on health care provision comprise the current and systematic description and analysis of health care aspects (often from the perspective of the appropriateness [629], quality and efficiency of services provided within the SHI) of a defined population group regarding a specific medical or system-related research question (see §139a SGB V). In addition to the description of individual aspects of the health care process, individual or complex interventions can be examined. For a planned

analysis of data on health care provision, various individual medical, as well as population- and health system-related data, can be used in a modular system, if necessary from different sources.

5.4.2 Aims of an analysis of data on health care provision

The superordinate aim of an analysis of data on health care provision is to describe the provision of health care.

The following aspects can be sub-goals of an analysis of data on health care provision:

- determination of epidemiological indicators
- determination of cost of illness
- determination of health economic indicators (e.g. budget impact analysis, cost-benefit ratios)
- examination of whether the health care provided meets the demand and is appropriate, as well as examination of indications of potential over-, under- or inappropriate provision of health care [629]
- 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 or a certain health care aspect.

5.4.3 Aspects of content of an analysis of data on health care provision

An absolute prerequisite for an analysis of data on health care provision is that the research question to be investigated must be answered with the available data. The formulation of a specific question is therefore necessary for the systematic description and investigation of health care areas. The definition of research questions comprises the specification of the following points:

- population (age; gender; disease; if relevant, subgroup or severity of disease)
- the interventions to be investigated, if applicable (e.g. care of diabetic patients by a general practitioner or a specialist)
- health care setting (e.g. outpatient care, acute inpatient care, long-term care facility, or cross-sector care)
- outcome measures (e.g. number of patients, costs of SHI services)

The analysis of data on health care provision can refer to different levels and/or several health care aspects. Basically, 3 main areas are distinguished: an epidemiological, a health economical, and an area of social organization of health care. The first area covers the distribution and frequency of diseases in the population, on the basis of which the need for medical services can be derived. In this context, specific attention may be paid to certain subgroups of the population (e.g. 'What is the age structure of a disease population with a certain co-morbidity?'). The second area involves the determination of costs or the relationship between costs and benefit of health care interventions. Finally, on a third level, for example, questions concerning the design of health care-related structures and processes of service provision are addressed.

In addition, it may be necessary to take socio-cultural and ethical aspects into account when describing the care of certain patient groups, e.g. consideration of difficult access to health care.

5.4.4 Data sources

Depending on the research question, different data sources can be used to answer it.

Data sources potentially relevant to the respective research question should be defined before the analysis is conducted. Potential data sources are named below (see Table 9).

Table 9: Potential data sources for the analysis of health care

Examples of data	**Publicly available**	**Agreement with data owner required**
Health surveys, e.g.: ▪ Health report of federal and state organizations (e.g. German Health Survey for Children and Adolescents by the RKI) ▪ Report of the Federal Statistical Office (e.g. hospital discharge diagnoses, statistics on causes of death).	X	
Registry data, e.g.: ▪ RKI data ▪ Epidemiological and clinical cancer registries	X	X
Routine data, e.g. from: ▪ Statutory health insurance funds ▪ Associations of Statutory Health Insurance Physicians		X
RKI: Robert Koch Institute		

If potentially suitable data have been identified for a research question, appropriate contractual agreements may be made between the owner of the data and IQWiG, so that the data can be made available for analysis.

As a matter of principle, an assessment of the data structure should precede the analysis, for example with regard to data quality, i.e. completeness and plausibility of the data sets, data collection methods, completeness, and timeliness.

IQWiG addressed the use of data from different sources outside clinical trials for the purpose of benefit assessments of drugs according to §35a SGB V in Rapid Report A19-43 [397].

5.4.5 Methodological features of the analysis of data on health care provision

The complexity of analyses of data on health care provision fundamentally requires an interdisciplinary approach. In addition to clinical expertise, sufficient knowledge of the handling and interpretation of the corresponding data is important.

It is not possible to define a uniform approach for the analysis of data on health care provision. This is strongly dependent on the given framework, such as the main research questions or the data sources used.

Research questions are processed using quantitative methods. When conducting and planning an analysis of data on health care provision, the recommendations of good practice for secondary data analysis should be observed [26].

5.4.6 Presentation of results

A transparent presentation of the methodological approach and results is therefore essential. The method applied must be described in a comprehensible way to ensure replicability. Reporting should be in accordance with the reporting standard for secondary data analysis (STROSA) [725].

6 ThemenCheck reports

6.1 Background and aim

According to §139b (5) SGB V, persons insured in the statutory health insurance (SHI) and other interested individuals can propose assessments of medical examination and treatment methods for selected diseases as well as of questions of quality and efficiency. According to §139b (5) SGB V, this excludes topic proposals where the separate assessment of a drug is the primary focus.

The Institute's task is to select topics from these proposals that are of particular importance to the health care of patients and to process these topics in the form of ThemenCheck reports.

The following text describes the process from topic submission to report production.

6.2 Topic collection

Topic proposals for reports can be submitted by persons insured in the SHI and other interested individuals via the website https://www.iqwig.de/en/participation/themencheck-medizin-suggest-a-topic/.

6.3 Selection of topics for the ThemenCheck reports

The topics that are proposed up to a yearly deadline run through a multi-step selection procedure (Figure 16). In this context, the perspectives of both the general public and patients as well as the scientific perspective are considered.

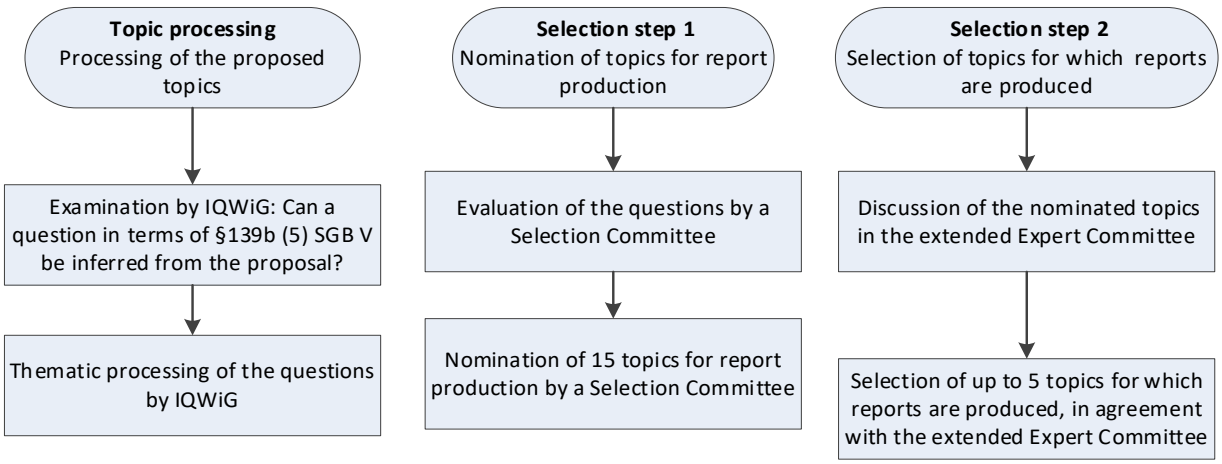


Figure 16: Schematic illustration of the multi-step selection procedure

6.3.1 Selection criteria

For the examination and weighting of topic proposals, as well as after each of the 2 selection steps, the topics proposed are evaluated by means of predefined criteria.

The aim is to select topics that are particularly relevant for the health care of patients. Likewise, as a matter of principle it is specified that medical examination and treatment methods are evaluated. Further criteria and questions that are considered in the selection are, among other things:

- How large is the number of affected persons?
- How great is the disease burden or the severity of disease?
- How comprehensive is the available evidence on the research question?
- Has the research question already been investigated in current HTA reports?
- What costs are related to an intervention?

6.3.2 Evaluation of the research question and processing of topics

The proposals submitted are assessed by the Institute in a timely manner and, if necessary, editorial and qualitative changes are made.

In accordance with §139b (5) SGB V, topic proposals where the focus is on the separate assessment of a drug are excluded from further processing. Therefore only those topic proposals from which a suitable question can be inferred in terms of §139b (5) SGB V are processed further.

For the selection criteria mentioned above (see Section 6.3.1), information on all topics with a question in terms of §139b (5) SGB V is collected and a preliminary search (see Section 8.1.1) is conducted. The results are summarized in the processing of topics.

6.3.3 First step of the selection procedure: nomination of topics

In the first step of the selection procedure, a Selection Committee nominates topics for the production of reports. For this purpose, the processed topic proposals are made available to the Committee. On this basis the Committee selects 15 topics that appear suitable for processing. In this context, the perspectives of both the general public and patients as well as the scientific perspective are considered.

The Selection Committee members comprise firstly, the representatives of organizations recognized on the federal level as relevant for the representation of the interests of patients and self-help groups of chronically ill and disabled persons, secondly, a representative of the Federal Government Commissioner for Patients' Affairs, and thirdly, members of the general public.

6.3.4 Second step of the selection procedure: selection of topics for which ThemenCheck reports are produced

The topics nominated by the Selection Committee in the first step of the selection procedure are discussed with the extended Expert Committee in the second step. The latter Committee consists of representatives of organizations forming the Foundation Council of the Foundation for Quality and Efficiency in Health Care as well as a representative of the Federal Ministry of Health. Subsequently, the Institute selects up to 5 topics from the topics nominated by the Selection Committee in consultation with the extended Expert Committee. ThemenCheck reports are produced for these topics.

6.3.5 Handling of nominated topics that are ultimately not selected

The topics nominated by the Selection Committee but not selected by the Institute are (once only) again included in the next selection procedure in the following year, where they again pass through the selection steps described in Sections 6.3.3 and 6.3.4.

6.4 Ensuring the quality of the ThemenCheck reports

The following measures, among others, ensure a high quality of the ThemenCheck reports:

- structuring of the content of the reports by the Institute as well as the provision of templates for the ThemenCheck report protocol and the preliminary ThemenCheck report
- conformity evaluation of the ThemenCheck report protocol, the preliminary ThemenCheck report, and the final ThemenCheck report by the Institute
- production of the report following the Institute's methods
- conduct of systematic searches in the Institute for information in relation to the ThemenCheck reports
- conduct of a commenting procedure
- classification of the results of the report in a publisher's comment produced by the Institute

ThemenCheck compact is produced by the Institute (see Section 2.1.10). It thus also undergoes the multistage internal quality assurance procedure provided for all product-specific publications of the Institute (see Section 2.2.4).

6.5 Processing of topics of ThemenCheck reports

As a rule, the reports address all relevant aspects. Besides the obligatory assessments of the benefit and harm of interventions, following international definitions, the economical, ethical, social, legal and organizational aspects of the intervention are also presented in the ThemenCheck reports [249,463,505,581].

The level of detail of the investigation of the health economical relevance of the respective examination or treatment method, as well as its ethical, social, legal and organizational aspects, depends on the research question and is specified in the ThemenCheck report protocol. The benefit and harm of examination and treatment methods, as well as their economic, ethical, social, legal and organizational aspects, are interlinked. In the processing of these aspects and the summarized presentation of results, the individual aspects therefore cannot be examined separately from each other.

6.5.1 Benefit assessment

The production of the sections of the reports referring to the benefit and harm of examination and treatment methods is conducted using the procedure described in Chapter 3, Section 7.13, and Chapter 9.

6.5.2 Health economics

If a separate health economic analysis is to be prepared in the ThemenCheck reports, the requirements of Section 4.15 must be considered.

A further option is to generate a systematic review of the available evidence on health economic analyses. In this context, the analyses are in particular assessed with regard to the applicability of the results to Germany, the classification of results under consideration of the methodological requirements described in Chapter 4, as well as the completeness of the data basis.

6.5.3 Ethics

In medicine, but also in public health and health services research, the 4 principles of beneficence, nonmaleficence, respect for patient autonomy, and justice following Beauchamp und Childress [48] are widely disseminated [216,330]. This approach has also been commonly used for the analysis of ethical aspects of medical interventions [47,517]. In recent years further methodological approaches have also been applied. These include, for instance, the Socratic approach, the social shaping of technology, virtue ethics or the triangular approach [28,373]. Depending on the research questions to be processed, the methodological concepts available are suitable in different ways.

The Socratic approach delivers detailed results and has been used for many different medical interventions [209,372,506].The questionnaire by Hofmann 2005 [370,371] is based on the Socratic approach and is as a rule to be used for the ThemenCheck reports. If other methodological approaches are better suited, they can also be applied, with corresponding justification in the ThemenCheck report protocol.

6.5.4 Social aspects

The social and socio-cultural aspects discussed in the report address the mutual interactions between the examination or treatment method and the social environment (e.g. distribution of resources in a society, access to technologies, patient preferences, societal norms and values). Four approaches for identifying social aspects are being discussed: checklists, literature reviews, participatory approaches, and empirical research [292,537]. The framework of Mozygemba et al. [537] is recommended for the examination or exploratory assessment of sociocultural aspects. A generic questionnaire (e.g. Gerhardus and Stich [292]) or the checklist from the HTA Core Model of EUnetHTA [249] may also possibly be helpful.

6.5.5 Legal aspects

The discussion of legal aspects refers on the one hand to the legal framework in which the examination or treatment method and its assessment are embedded, and on the other hand to the legal aspects related to the implementation and use of the health technology. One distinguishes between technology- and patient-related legal aspects [228,341,785].

6.5.6 Organizational aspects

A ThemenCheck report can also examine which interactions arise through an examination or treatment method for the organization of health care and which determinants can influence the implementation of an examination or treatment method. In principle, one can distinguish between the interaction of organizational conditions, procedures and processes and the requirements for health care professionals as well as the decrease or increase in their workload [582]. However, no methodological standard so far exists whereby the organizational interactions of examination or treatment methods in the health care system can be examined [582].

The grid template proposed by Perleth et al. [582] for the assessment of organizational consequences of examination and treatment methods can provide support for the processing of organizational questions.

7 Evidence-based health information for consumers

The Institute has a statutory responsibility to provide readily understandable health information to consumers, but not direct advice to individuals.

In the production of its health information, the Institute adheres to the principles of evidence-based medicine as presented in Section 1.3. This includes applying systematic methods for the scientific assessment of medical interventions as well as integrating the perspective of people affected.

The methods used in the production of evidence-based health information are based on a systematic approach that aims to present the current state of scientific knowledge in an easy-to-understand way, minimize systematic errors (bias) and remain neutral.

To achieve this, evidence-based information is based on the following principles:

- systematic search in the form of targeted information retrieval based on the research questions relevant to the target audience,
- justified selection of evidence appropriate to the research question,
- reasonable and – as far as possible – objective reporting of results that are relevant to patients, such as mortality, symptoms and complications (morbidity) and health-related quality of life,
- appropriate factual and linguistic communication of uncertainties and unclear issues,
- avoidance of any direct recommendations,
- consideration of current evidence concerning risk communication.

These principles apply regardless of the type of commissioning of health information. There are 2 ways of commissioning:

- direct commissions received from the G-BA or the Federal Ministry of Health to produce patient information; these commissions are processed as a report, rapid report or addendum (see Sections 2.1.1, 2.1.2 and 2.1.7),
- the mandate derived from the Institute's statutory responsibility to provide consumers with health information, specified by a general commission from the G-BA (see Sections 1.2 and 7.2)

7.1 Goals of health information

The primary goal of the Institute’s health information is to provide easily understandable medical knowledge that is relevant to decision-making. This knowledge is meant to support

the users of information in their autonomy and competency to make an informed decision between different options. This way, users are free to make their own decisions. The information is written based on the concept of "shared decision-making", according to which users aren't told how to make decisions [226]. Whether to delegate the decision to a doctor or to decide together with others or alone is left up to the individual.

Moreover, the information is to be provided in a way that promotes general health and scientific literacy.

In summary, these are the Institute's goals:

- to provide information that is relevant to medical decisions and easy to understand, even if it concerns complex scientific issues,
- to facilitate active and informed decision-making,
- to promote the critical use of health care services,
- to improve understanding of physical and mental health,
- to improve understanding of medical and scientific information, including the concept of evidence-based medicine,
- to promote the support of patients by family and friends,
- to provide information on handling practical and emotional issues in everyday life,
- to emotionally support people who are affected and
- to help navigate the health care system.

These goals can be summarized as "empowerment". According to the definition of the World Health Organization (WHO), empowerment within a health care system includes the competency to make choices and take actions in accordance with your own goals [559]. Empowering health communication addresses what consumers want to know, shows interest in and respect for patients' opinions, and acknowledges their competence [197,439,766].

The particular challenge of evidence-based health information is in meeting these requirements and goals while being attractive and understandable to users [189]. It is important not to allow the various requirements to result in an overload of information. So it is sometimes not possible to reach all the goals and fulfil all the quality requirements in a single piece of information.

One solution is to combine individual texts and other formats that have different priorities, in order to create information packages of appropriate breadth and depth. This is realized on

the websites www.gesundheitsinformation.de (in German) and www.informedhealth.org (in English).

7.2 Selection of topics and identification of information needs

§139a (3) Sentence 6 of 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 of and therapy for diseases of substantial epidemiological relevance".

7.2.1 Topic catalogue in accordance with the statutory responsibility

It is not possible to arrive at a broadly acceptable definition or a clear list of "diseases of substantial epidemiological relevance" based on the literature. One fundamental aspect of epidemiological relevance is a measure of how common a particular illness is. 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 prioritized. This threshold is applied to different subpopulations according to sex and age (0 to 17 years old, 18 to 59 years old, 60 years old and above) in order to reflect the sex- and age-specific particularities of these groups as accurately as possible [451].

The topic catalogue is primarily based on the biannually updated health care report produced by the AOK[35] Research Institute (WIdO[36]) [319]. The report contains information on prevalence and hospitalization rates for the 1500 most common illnesses (grouped according to 3-digit ICD[37]-10 codes), based on about 24 million members of the AOK. The Institute's topic catalogue is reviewed regularly and subsequently amended as needed. This topic catalogue can be expanded, e.g. depending on the topics commissioned to IQWiG or to include medical conditions whose significance and burden of disease cannot be satisfactorily reflected by prevalence or incidence data.

7.2.2 Identification of information needs / Production of information about personal experiences with medical conditions

Ideally, evidence-based health information should be guided by the information needs of the target group. In addition to the general aspects that are relevant to all target groups, such as prevalence and the course of the disease, the information should address topic-specific questions and issues, common misunderstandings and gaps in knowledge.

[35] Allgemeine Ortskrankenkasse (Local Healthcare Fund, an SHI fund)

[36] Wissenschaftliches Institut der Allgemeinen Ortskrankenkasse

[37] International Statistical Classification of Diseases and Related Health Problems

Patient-centred health information shouldn't just answer medical questions and enable an informed decision. It should also answer questions about health care services and dealing with the disease, as well as offer emotional support [263]. To do this, it is necessary to know the questions users might be interested in. It is also important that the authors at the Institute have an understanding of the situation and the burdens that an illness may cause for patients and their family members. They should develop an understanding of what it means to live with a certain illness.

Therefore, to prepare a topic, qualitative literature is typically searched for and analysed as part of targeted information retrieval (see Section 8.2) in order to identify personal experiences with the condition as well as questions and knowledge gaps that are relevant to the users. The relevant results from Germany are analysed first. If none are available [590], then information needs are derived from studies conducted in similar countries.

Furthermore, freely available health information from other sources on the Internet (e.g. statutory health insurance funds, health authorities, research institutes and commercial businesses) are also reviewed. This provides an impression of the current availability of health information and its main areas of focus. In addition, for different topics, different self-help organizations are contacted and asked about the information needs of patients and the challenges of coping with the illness.

This approach provides a picture of the different stages that a person who is affected by a certain health problem goes through, the associated psychological and emotional problems that can occur, which information is needed, and the points at which decisions need to be made.

7.2.3 Multidimensional patient pathways

Especially for chronic diseases, a sort of map is created to navigate the issues and decisions associated with a given illness. This map shows the possible "pathways" that patients with the disease might take. It should attempt to show as comprehensively as possible how the patient's life may be affected by the illness, and at what points critical decisions are expected.

This approach will be referred to as "multidimensional patient pathways" below.

The decision regarding whether a multidimensional patient pathway will be created for a specific topic will depend on the following criteria, among others:

- Is the illness a long-term or chronic disease that is associated with different stages for the patient (e.g. the processing of a serious diagnosis, unpleasant and difficult treatments, the need for aftercare)?

- Do some of these stages involve complex decisions, such as choosing between various treatment options with specific risk/benefit profiles?
- Are several physicians or other health care professionals involved in the process? Are there different health care options (outpatient or inpatient)?

Multidimensional patient pathways summarize the different social, emotional, cognitive and clinical dimensions that can be associated with an illness. The information is presented in a table. It should illustrate what challenges and decisions patients may face over the course of the illness. The multidimensional patient pathway approach shares similarities with the medical-sociological "illness trajectory" model [157] and the "patient career" model [291,468], as well as various "patient journey" models [478].

One of the aims of multidimensional patient pathways is to determine the scope of a topic and the main areas of focus when producing the Institute's health information. The following questions can be helpful here:

- Who might read the information?
- What questions might the readers have?
- What might be the emotional state of the reader?
- What information might be needed at different points during the course of the disease?
- What decisions are patients faced with, and when and where will they have to make these decisions?
- What effects might health information on this topic have?

This approach mainly aims to help the authors of the Institute's health information systematically develop a good understanding of patients and those close to them, as well as of their interaction with information. Consideration of the dimensions shown in Table 10 supports this aim.

Table 10: Different dimensions of a patient pathway

(Everyday) life	Effects of the disease on social relationships and roles: family and partners, friends, job, quality of life, "performance", etc.
Doing / coping	Any activities related to the illness, such as visiting a doctor, taking medication, 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 already know? What information might they need?
Decisions	What decisions must the person affected make in each phase?
Medical / clinical aspects	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 offer help and guidance in each phase, for example, doctors, nurses, physiotherapists, psychotherapists, social workers, information centres, insurance funds?

Various sources are used when producing draught patient pathways. These may include other qualitative literature (qualitative studies and systematic reviews), evidence-based guidelines of national and international scientific associations, evidence syntheses, literature on the information needs of patients and the health care situation, as well as reports of personal experiences with a medical condition [458].

7.3 Information retrieval for the production of health information

When producing evidence-based health information, systematic searches are a key part of the process. They involve scanning current scientific literature for information that is relevant to the research question. As a rule, 2 searches are performed for every topic:

- A search for qualitative literature (see Section 8.2.2): The results are used to identify any needs for additional relevant information, such as personal experiences with the condition. This search is meant to enable the authors to empathize as much as possible with the people affected.
- A search for systematic reviews (see Section 8.2.1): The results provide the basis for drawing conclusions about the benefits and harms of medical interventions.

In addition to these searches, simplified searches (see Section 8.3) are performed to answer specific questions, for instance about the prevalence of the condition.

Searches for systematic reviews aim to identify such overviews of research on all aspects covered in a piece of health information, such as the causes, the course of a medical condition, the prognosis, treatments and epidemiological data. In general, only systematic reviews that are based on searches that were performed in the last 3 years are accepted as a source of information [679,680]. If the searches were performed more than 3 years ago, they may be

out-of-date, considering the short half-life of medical knowledge. This time period can be adjusted for specific topics, for example if there is a lack of more recent research.

The identified reviews are then matched with the research questions. In an internal editorial and scientific scoping process, the results of the literature search are used to cover the identified information needs.

If an important information need, such as a need for long-term data, cannot be met with recent systematic overviews, a targeted search for primary studies may be considered.

7.4 Selecting evidence

When producing evidence-based health information, the available relevant scientific knowledge from current and sufficiently reliable studies is to be considered. The type of studies that are appropriate depends on the question being asked. Conclusions about the benefits and harmful effects of interventions are generally based on systematic reviews of randomized controlled trials (RCTs – see Section 8.2). Quality is assessed using the AMSTAR 2 tool ("A Measurement Tool to Assess Systematic Reviews") [677]. For a systematic review on the effect of an intervention to be used for health information, the certainty of the results ("overall confidence rating") must be at least moderate [677]. Sixteen items are assessed, including the quality of information retrieval, the selection of studies, and the synthesis of the evidence. The studies must be described in sufficient detail to assess whether they meet the defined inclusion criteria and whether the description is sufficiently informative, for example, to examine heterogeneity. The assessment of their risk of bias must be adequate. For a systematic review to be considered an appropriate basis for drawing conclusions about the benefits and harms of a medical intervention, it has to meet certain minimum requirements, i.e. it can only have minor methodological flaws. It is also assessed how the review authors handled possible publication bias. For example, did they specifically search for unpublished data? Failing to address publication bias or handling it inadequately may lead to a description of the subsequent uncertainty of the results in the health information, or even to the exclusion of the systematic review.

The selection will also be influenced by the question of whether the results are applicable in the context of the German health care system. Aspects taken into consideration include the study population, approval status and how common the intervention is. If conclusions only apply to certain groups, this will be explained in the text.

If a health information article contains conclusions that are based on evidence syntheses of varying qualitative certainty, then this will be addressed in the article.

When more than one systematic review of adequate methodological quality addresses a particular subject or outcome, the qualitatively best reviews are considered. In addition to the previously mentioned aspects, the following are also considered:

- The main areas of focus of the review, especially with regard to its relevance for patient information
- Sensitivity analyses and handling of heterogeneity

The results of the highest-quality review for a particular research question are taken as the source of the numerical data described in the health information. If reviews come to different conclusions, the possible reasons are explored [412]. These may include differences in the study pool, the statistical analysis or the interpretation of the results.

For research questions on aetiology or prognosis, for instance, systematic reviews on the basis of types of studies other than RCTs can also be used [298]. When assessing such systematic reviews, the criteria of the Oxford Centre for Evidence-Based Medicine are used [353,564]. The methods for assessing qualitative research are described in Section 8.4.

7.5 Choosing the results (endpoints) to be presented

Information on treatment outcomes is based on endpoints that are relevant to patients – in particular mortality, symptoms and complications (morbidity), as well as health-related quality of life. The principles described in Sections 3.1 and 3.2 of these Methods generally apply here.

Further information that is often important to patients includes circumstances surrounding the treatment (e.g. the time involved, physical, mental, social, as well as financial burdens).

7.6 Choosing and presenting comparisons

In order to enable users of gesundheitsinformation.de / informedhealth.org to consider the benefits and harms of an intervention concerning the patient-relevant outcomes described in Section 7.5, a comparison is drawn between having the intervention and forgoing it or using a different standard treatment.

7.7 Handling numerical data and information about risks

Producing a balanced description of the possible benefits and harms of an intervention generally requires a quantification of the effects and risks. This involves taking care to limit the use of numerical data and information about risks so as not to overload the health information, or affect the flow of reading and comprehension.

Because describing probabilities in words often does not lead to a realistic assessment, numbers are preferred [115].

The following principles are followed when providing numbers and probabilities:

- The effect of a medical intervention is described by showing the absolute event frequencies in the comparison groups. Any uncertainties are mentioned. If possible, the reference values will be selected so that the magnitude of the effect is readily apparent.
- The same reference value is used for the descriptions of the benefits and harms, if possible. The starting point for the comparison is the baseline risk, e.g. the natural course of an illness. Here we mean the likelihood that, in the absence of treatment, symptoms will improve, worsen or stay the same. The users will be informed if the symptoms may improve without treatment as well.
- If it is helpful, relative changes will be described in addition to the changes in the absolute risk.

Whether the effects are presented as a "gain" or a "loss" will depend on the particular intervention and the perspective of the person affected by the condition. Here, a uniform frame of reference concerning the benefits and harms is selected. Studies on the benefits of also using diagrams to present numerical information do not clearly show how this affects the endpoints knowledge, risk perception or understandability. The German-language evidence-based guideline *Evidence-based Health Information*[38] arrives at a recommendation that this is optional, based on the low quality of the evidence [503].

In the production of health information, illustrations are mainly used for the supplementary presentation of complex subject matter, e.g. generally in cancer screening decision aids.

For the purpose of presenting them in health information, the relative measures of effect that are used in meta-analyses are converted to absolute measures. The absolute effect measure is calculated using a similar approach to that used to produce the results tables within a Cochrane review [364]. The uncertainty of the effect measure is thereby taken into account, but not the uncertainty of the baseline risk [175,699].

The basis of the calculation is a pooled effect estimate from a sufficiently homogeneous meta-analysis. If the effect measure is already a risk difference, it will be used in the further consideration. If a relative effect measure is given, then a plausible baseline risk will first be selected to derive the absolute risk difference. This is typically based on the median of the control group risk in the included individual studies.

In justified cases, the baseline risk can also be derived from a suitable individual study (e.g. the study with the largest by far population or the highest external validity) from the included study pool or from a valid external source (e.g. a German prevalence study, or from registry

[38] Evidenzbasierte Gesundheitsinformation

data). In the event that the baseline risk is taken from an external source and shows a high degree of uncertainty, this uncertainty will be taken into account. The method described by Newcombe and Bender is applied in that case [552].

Based on the assumed baseline risk, the absolute risk in the intervention group and the absolute risk difference are calculated using the relative pooled effect size estimator (odds ratio [OR], relative risk [RR], hazard ratio [HR]) from the meta-analysis.

If the included individual studies have heterogeneous baseline risks, then each absolute effect for the various assumed baseline risks is presented (e.g. for a low and high baseline risk). If it does not make sense to include an absolute effect measure, then it can be left out.

7.8 Taking into account differences related to age and gender

The natural course of a disease, risk factors, symptoms, morbidity, mortality, effects and side effects of interventions, health-related quality of life and accompanying circumstances of an intervention can vary depending on age or gender. If the literature that has been identified as relevant to a condition describes significant differences related to age and gender, these are taken into account while producing health information.

For many topics, the epidemiology of the condition already results in aspects specific to gender and age. If a piece of information is aimed at the group that is mainly affected by a condition, this target audience is not explicitly mentioned in the information.

If no relevant differences between the different groups can be identified, this lack of difference is not specifically pointed out.

If, however, any results apply to subgroups, this is made explicit.

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 leads to a mental underrepresentation and linguistic discrimination of women [406]. The information published at gesundheitsinformation.de / informedhealth.org uses a gender-neutral style of language, which avoids the use of generic masculine forms whenever possible. Both genders are explicitly named if both are meant, or gender-neutral expressions are used.

7.9 Adaptation to the target group

A key challenge when producing evidence-based health information is communicating in a way that is understandable while remaining scientifically accurate and objective. Additionally, the Institute's health information is aimed at a heterogeneous target group in terms of literacy, abilities, current health problems, education, personal background, age, and sex.

Aspects such as native language and cultural background are also important for some target groups.

When preparing a given topic, a check is done to see whether the epidemiology of the condition results in special requirements for particular target groups.

To meet the needs of specific target groups, the information can be adapted by combining the following options:

- surveys, primary qualitative research and reviews of qualitative research on people's information needs
- experiences of other information providers, patient advice services and self-help groups
- interviews with those affected by the health problem (see Section 9.4.2)
- collection of real-life stories as a separate format (see Section 7.16.2)

7.9.1 Involvement of those affected

There is some evidence that involving the people affected by a particular health problem in the development of health information can increase its relevance [554]. One of the requirements of evidence-based health information [188] is that it takes the consumers' perspective and information needs into consideration. This is a key element when producing health information [807]. This can be achieved in various ways, including the following: When prioritising and narrowing down the aspects to be covered by a topic, ideas proposed by website users and experiences from consultations with self-help groups are taken into account.

Those affected by the medical condition are contacted through the patient representative organization in the G-BA and through contact with self-help groups.

For suitable topics, 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 7.16.2).

7.9.2 Non-public commenting procedure

After the internal quality assurance processes and the external expert peer review have been concluded, the Board of Trustees is given the opportunity to comment on the draft version in a non-public commenting procedure. The Board of Trustees also includes representatives of patients' interests and representatives of self-help organizations of chronically ill and disabled people. After the deadline for the submission of comments has passed, the members of the team involved in producing the information and the Head or Deputy Head of the department

view the comments submitted and discuss their relevance for the health information. At an editorial conference, the remarks and arguments made in the comments are discussed and – if available – the evidence is reviewed. Any need for revision is agreed upon and documented. The comments are acknowledged close to the time of publication of the final version of the information. If received prior to the deadline, suggestions regarding changes to the content are acknowledged. Comments containing only a few editorial remarks and comments that are received after the deadline are generally not acknowledged. Comments and acknowledgments are not published.

7.9.3 Testing by users

The primary means of testing the understandability of the health information are reviews of drafts by groups of test readers. The standardized external testing by users generally occurs during the same time period as the commenting procedure. It is conducted by an external service provider in the form of focus groups or individual interviews. Depending on the specific topic, patient organizations or other associations or institutions may also participate.

A mix of methods is applied in user testing, including documented individual evaluation and group discussion or individual interviews. Whether a group discussion or individual interviews are conducted is decided depending on the topic of interest. For example, individual interviews may be more suitable for taboo subjects or specific target groups. On the basis of a semi-structured guideline, usually 5 people affected or other potential users test the texts with regard to their content and comprehensibility. The testers are recruited by the external contractor on the basis of criteria specified by IQWiG. The selection criteria relate to the priorities set in the information materials to be tested and usually include information on diagnosis, gender, age, experience with certain treatment options, and, if applicable, social status.

The results of the user testing are considered in the revision of the draft health information.

7.9.4 Users' feedback comments

In addition, users of the website www.gesundheitsinformation.de / www.informedhealth.org can contact the publisher with their feedback. Various channels are offered on the website for this purpose:

- the opportunity to leave comments on individual articles
- a general online contact form
- a randomly generated survey asking individual users to rate the website.

7.9.5 Accessibility

Because the information is primarily offered on the Internet, the website gesundheitsinformation.de / informedhealth.org fulfils the German Barrier-Free Information Technology Regulation (BITV[39]) [118].

The health information is published in both German and English. The availability of an English version improves access for non-German speakers and broadens possibilities for translation into further languages.

7.10 Neutral presentation

Information related to decisions about diagnostic and therapeutic procedures should convey a realistic overview of current knowledge while using non-directive, non-judgmental language and providing an appropriate contextual frame. Any biased and in particular any inappropriately alarming phrasing is to be avoided, as is trivialization. Both content and language should appropriately reflect significant uncertainties.

In order to convey this requirement to the authors in their daily work, a style guide on text production is used [546]. This guide undergoes continuous development based on the evaluation of our information and emerging evidence in the field of evidence-based communication.

To achieve the goal of neutral presentation, our health information undergoes a multi-stage editorial process including internal quality assurance, an external review, user testing and non-public commenting (see Section 7.9).

7.11 Inferring assessments and recommendations

Individual benefit-risk assessments are based on accurate information concerning patient-relevant outcomes. This aims to enable patients to make decisions in line with their own values and preferences.

Explaining evidence and remaining neutral in communicating health-related information pose a special challenge [230,440,662,745]. As a rule, recommendations are not made in the health information. This is achieved by using non-directive and non-judgmental language. Exceptions are possible, for instance when discussing how to act in the event of an emergency.

7.12 The development of decision aids

One tool to help people individually weigh benefits and harms are decision aids. The general requirements when producing health information also apply to the content of decision aids.

[39] Barrierefreie-Informationstechnik-Verordnung

The decision aids are developed according to the International Patient Decision Aid Standards (IPDAS) [227,374].

7.13 Transparency regarding author and publisher

On the websites gesundheitsinformation.de / informedhealth.org and iqwig.de, the Institute outlines its principles and funding as a non-profit and scientifically independent publisher of health information. The information provided exceeds legal requirements and fulfils further transparency criteria.

7.14 Disclosure of relationships

The Institute requires that its staff fully discloses relationships that could influence their work.

Insofar as external reviewers are involved in the production of health information, the same general standards apply as for any other of the Institute's products (see Section 2.2).

7.15 Mentioning of medications, medical procedures and devices

When mentioning medications and medical procedures and devices, generic names are to be preferred. However, as the general public often may not know the names of active ingredients, and instead use colloquial trade names, the Institute's health information may also include trade names.

7.16 Description of typical formats and contents

The website gesundheitsinformation.de / informedhealth.org focuses primarily on the presentation of topics relating to health and illness. One topic may comprise different types of articles and information formats. These different formats are intended to cover the main 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 background information and links to the types of articles (described below) that further explore the topic. The overview has a fixed structure.
- **Learn more:** This format 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 good evidence, the resulting uncertainties. A "Learn More" may also offer insights into aspects of living or coping with a particular health problem. The article will then attempt to take into account both the perspective of the person directly affected by the illness and that of family members or

other people who are close to them. “Learn more” articles may be supplemented by illustrations or multimedia content.

- **Research summaries:** These articles are objective summaries of the current 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.

7.16.1 Supplementary formats

The main formats can be supplemented by additional formats, e.g. to expand on individual aspects of a topic or to try to present certain information in a different way. For example, using images, sound and animated films may increase the attractiveness and understandability of the website, especially for people with lower literacy levels.

The supplementary formats include the following:

- real-life stories of people affected by the medical condition (see Section 7.16.2 for more information),
- illustrations, photos and other images,
- animated films with text and sound,
- quizzes,
- glossary,
- “In brief“: general articles explaining anatomy, bodily functions, treatment approaches and diagnostic measures, as well as the principles and methods of evidence-based medicine,
- calculators.

The goals of these supplementary items include 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 subjected to external testing by users.

7.16.2 Real-life stories

The real-life stories represent one means of conveying scientific evidence and making it accessible to the general public [300]. The importance of real-life stories in medical practice

and in health care is increasingly recognized [313,708,804]. Many patients would like to hear or read about the experiences of people affected by the same health condition as them [358,727].

The essential part of a real-life story is to show how an individual experiences and deals with their situation. Real-life stories provide the following [727]:

- They offer people the opportunity to compare their 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 by the medical condition that they are not alone.

By presenting the individual stories of patients as well as those close to them, the Institute makes it possible for 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.

Some people may see the real-life stories as a recommendation to make similar decisions. This effect may be in conflict with the aim of creating neutral information. For the real-life stories published on gesundheitsinformation.de / informedhealth.org to first and foremost offer insights into how people experience a medical condition and cope with its effects, the articles are edited to ensure that they

- do not contain any passages that contradict the statements on the scientific evidence contained in the other articles,
- do not make any explicit recommendations, and
- only mention options that are commonly used.

Real-life stories are put together as follows:

- Interview partners are found, most often via self-help organizations, patient universities and doctor's offices.
- Written informed consent is sought regarding the interview procedure and how the interview will be used.
- The interview is conducted (usually by telephone).
- The interviews are documented and edited, and the interview partners give their informed consent for the publication of the final version.
- Publication on the website.

The editing of the interviews includes the transcription of the audio material into German and the reduction of the content to an amount that is suitable for the Internet. The contents are shortened and summarized on the basis of predefined areas of focus, especially in the areas of living with the condition, experiencing the symptoms, and coping with the diagnosis, course and effects of the disease. This process is carried out in close cooperation with the interview partners.

The methods used to record, edit and publish the real-life stories are guided by the established methods followed by the creators of the Database of Individual Patient Experience (DIPEx) [196].

The decision on whether to include real-life stories in the information on a particular topic is based on various criteria, including the following:

- the possible effects of the illness on patients' lives in terms of physical, emotional and social aspects,
- the possible duration of the medical condition and the likelihood of it becoming chronic,
- the extent to which aspects of the illness are taboo in society, which may make talking about them in a social environment difficult (such as mental illness or conditions affecting the genitals).

7.16.3 Website

The main tool for the dissemination of the health information is the bilingual website www.gesundheitsinformation.de / www.informedhealth.org. High website standards are to be maintained in the following areas:

- usability and accessibility [386,467,553]
- privacy and data protection [398]
- transparency
- search engine visibility [740]
- attractiveness to users

The website includes a free electronic newsletter, with the choice of a fortnightly or monthly subscription. The newsletter contains information on what is new on the website, including when information is updated. The content of the website is also available in Atom and RSS formats, and can be read using the customary readers. In addition, website content can be automatically integrated into other suitable websites through an interface.

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. It also complies with the principles of good practice in the development of health information defined by the German *Good Practice Health Information*[40] [189].

7.17 Updating content

A critical part of providing evidence-based health information is making sure that its conclusions are not out-of-date. Regular updates are one of the quality criteria determined by the EU for health-related websites [153] and described by the German position paper *Good Practice Health Information* [188].

A study of guideline recommendations concluded that after 3 years, over 90% of recommendations may still be up-to-date, compared to only about 50% after 6 years [679]. For some topics, for example where the evidence is very strong, the half-life of evidence may be much longer, and in other areas it may be less than 3 years [680]. The Institute generally considers it necessary to review its information every 3 years. On the basis of this time period, as early as the initial publication of a topic, a deadline is specified for when an update search is to be performed.

Notwithstanding the above, the Cochrane Database of Systematic Reviews (Cochrane Reviews) and the McMaster Online Rating of Evidence (MORE) are checked regularly as part of an orientating evidence scan. German, European, and U.S. regulatory agencies are monitored for health warnings as well.

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 evaluated. The consequences depend on how much the statements in a health information would have to change. This may result in an update ahead of time or go as far as a withdrawal of the health information affected.

7.18 Updating the methods of gesundheitsinformation.de / informedhealth.org

The methods used in the production of health information are reviewed as part of the general update process of the Institute's methods, and revised if there is a need for an update.

[40] Gute Praxis Gesundheitsinformation

8 Information retrieval

Information retrieval for the generation of the Institute's products is conducted in a systematic manner. It can follow quite different goals. The type of goal determines whether comprehensive (see Section 8.1) or focused (see Section 8.2) information retrieval is conducted. In addition, exploratory searches are conducted (see Section 8.3). Starting from a detailed description of the quality standards for comprehensive information retrieval in Section 8.1, only the changes are presented in Sections 8.2 and 8.3. Section 8.4 explains the search for guidelines.

In Section 8.5 the approach for assessing information retrieval is described, as in particular conducted within the framework of assessments according to §35a SGB V, assessments of potential, and assessments according to §137h SGB V.

8.1 Comprehensive information retrieval

Comprehensive information retrieval aims to identify all relevant studies and related documents for the respective research question. For this purpose, a systematic search is conducted in several databases as well as in other information sources. The search is usually not limited to specific years. However, a restriction to English and German publications can be made, since evaluations have shown that the exclusion of non-English publications does not change the conclusion of most systematic reviews [203]. It is checked on a project-specific basis whether a language restriction can be expected to affect the result of the report (e.g. complementary medicine) [531]. Information retrieval is presented in detail in the methods and results section of the report.

Preliminary note: handling reporting bias (including publication bias)

In information retrieval, superordinate aspects such as reporting bias (including publication bias) should be taken into account as they have a fundamental effect on the selection of the different information sources. In this context it should be considered that many research results are never or only partly published [132,524,696] and published studies tend to overestimate the positive effects of interventions and underestimate the negative effects [524,696] (see Sections 3.2.1 and 9.3.13).

In benefit assessments a search for unpublished data is therefore also conducted as a standard component (see Sections 8.1.2 and 8.1.3). Besides unpublished studies, this search should also identify unpublished data from published studies.

Published studies usually appear in scientific journals and thus can largely be searched for via bibliographical databases such as MEDLINE or Embase.

Indications of the existence of unpublished studies and data can, for instance, arise from trial registry entries as well as from clinical study reports prepared by the manufacturers of medical interventions.

If clinical study reports are available, they represent the primary source for the benefit assessment, as only clinical study reports contain nearly all information on a study [452]. In contrast, the information from other sources is often insufficient for a targeted evaluation according to the underlying research question or exhibits discrepancies [427,596,786]. However, data from registry entries and publications can supplement each other [786] or unpublished data can be used to check the correctness of published data [43].

If data are submitted that, according to the will of the persons submitting, are not allowed to be published, these data cannot be considered in assessments, as this would contradict the principle of transparency.

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

8.1.1 Searches in bibliographic databases

Bibliographical databases (besides trial registries) represent a key source for identifying study results, especially if no clinical study reports are available. The detailed procedure is described below.

A) Preparatory searches

At the start of a project, before the development of the actual search strategy, preparatory searches are usually conducted by means of an exploratory search (preliminary search) or focused information retrieval for systematic reviews. The searches serve to prepare the project. For this purpose, a search is conducted in particular for existing systematic reviews [142,275,705], but also for potentially relevant primary studies on the topic.

A search for systematic reviews can be conducted, for example, in the Cochrane Library, the International HTA Database, and on the websites of HTA agencies such as NICE or AHRQ [250,705,784]. In addition, if appropriate, earlier IQWiG reports, published manufacturer dossiers as well as resolutions by the G-BA can be screened. To identify ongoing HTA reports and systematic reviews, databases such as the Prospective Register of Systematic Reviews (PROSPERO) [141] can be considered. Furthermore, guidelines or clinical information systems such as Dynamed or UpToDate can be used as an information source.

If no relevant systematic reviews are found, an iterative process is conducted using different search techniques (see Section 8.1.4) such as citation-based techniques [156,348] or the use of the "similar articles" function in PubMed (see Section 8.1.4) [348]. The starting point is formed by several relevant articles that are already known or were found by a very precise search. In several runs further articles are then identified and tested for relevance [604,650].

B) Structure of a search strategy

Before the development of a search strategy the structure of the search must be specified. For this purpose, a clearly formulated research question is required. Formulation based on PICOS[41] is usually a useful approach [486] to structure the search [485]. In this context only the most important search components are used to develop the search strategy [652]. The search strategy mostly contains search terms on the therapeutic indication, intervention and study type [485,486].

C) Selection of databases

A systematic search in several bibliographic databases is required for the production of systematic reviews [621,622,636,716]. Recent analyses show that most published studies can be found in a limited number of databases [9,255,331,343]. The Cochrane Handbook lists MEDLINE, Embase and CENTRAL[42] as the 3 most important bibliographic databases (for primary studies [485,486]). These are also the most frequently used databases for systematic reviews [74]. Depending on the research question, regional or topic-specific databases can also be considered.

D) Development of search strategies

A combination of subject headings (including publication type) and free-text terms is necessary for the development of search strategies [283,421,436]. An objective approach for developing search strategies is characterized by the fact that text analysis procedures are used for identifying free-text terms and subject headings [577,704]. For instance, IQWiG's objective approach is based on the analysis of relevant articles already known [346,349]. At the outset, these are divided into 2 sets (set for development and set for quality assurance of the search strategy). In this context, different text-analysis tools (e.g. searchbuildR [433]) are used. In a next step the search terms selected are allocated to the single search components of the search strategy [640,652].

If available, validated study filters (e.g. for RCTs [215,485,486,796,797] and systematic reviews [796] or validated classifiers from machine learning (e.g. RCT classifiers [516,771]) are used. For other study types or questions, it should be checked in the individual case whether

[41] Population, intervention, comparison, outcome, and study design

[42] Cochrane Central Register of Controlled Trials

validated study filters are available that can be applied reliably [420,484]. For example, there is a controversial discussion about the use of search filters in the search for diagnostic accuracy studies [65]. Study filters cannot always be used when searching for non-randomized studies [347], even though validated filters for specific study types are now available [770].

Furthermore, an additional search for non-indexed data sets in PubMed/MEDLINE can be conducted. This especially aims to identify very recent citations [211,737]. This search is based on free-text terms with an adaption of study filters [169,418], as these filters are usually optimized for a combination of subject-heading and free-text searches.

E) Quality assurance of search strategies

A high-quality search strategy is a prerequisite for ensuring the completeness of the evidence base of a benefit assessment [639,640]. Due to their complexity, search strategies for bibliographic databases are error prone [637]. The PRESS[43] checklist is therefore used to support the process of quality assurance [525].

Quality assurance with the PRESS checklist is initially a formal evaluation and is always performed before the conduct of searches. In addition to the PRESS checklist, the search strategy is tested against an independent set of relevant citations [638]. It is thus evaluated in advance whether the set of relevant citations can be found by the search strategy.

F) Study selection

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 [516].
- 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.

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.

Machine learning approaches (e.g. prioritization, application of classifiers) can be tested and used to support study selection.

[43] Peer Review of Electronic Search Strategies

The documentation of study selection is as transparent as possible and includes the decisions to include and exclude each reference (only at full-text level) [142,221]. Study selection is performed in a web-based application (EVIScreener). The documentation of the evaluated references is further carried out in the course of report production using the reference management software Endnote [149].

G) Documentation in the report

A clear and transparent presentation of all aspects of the search enables the assessment of the quality and completeness of the systematic search [573,605,641] as well as the conduct of later search updates.

As a standard, for the search in bibliographic databases, the following items adapted from the PRISMA-S[44] guideline [605] are documented:

- databases used, as well as database providers and search interfaces
- the search period and the last search date
- the search strategies with all search limitations

In addition, the selection process is presented in the results section of the report by means of a flowchart [572,573] and the citations of the studies or documents included or excluded are presented in separate reference lists.

8.1.2 Searches in trial registries

The importance of trial registries has strongly increased in the past years. For instance, since 2005, the International Committee of Medical Journal Editors has required the prospective registration of clinical studies as a precondition for publication in scientific journals [172].

In addition, since 2007 there has been a legal obligation in the United States to register nearly all clinical studies regulated by the FDA as well as their results [3]. If available, ClinicalTrials.gov now provides links to the study protocol, the statistical analysis plan (SAP) and the data sharing statement of a study [549,729,802].

Since 2022, the European Medicines Agency (EMA) has been publishing data on the registration and study reports of clinical trials in the European Union and the European Economic Area via the Clinical Trials Information System (CTIS) [236] (with a transitional period until 2025). The basis for this is the Clinical Trials Regulation (Regulation (EU) No 536/2014 [233]). From 2025, the EU Clinical Trials Register (EU-CTR) [244] will only serve as an archive.

[44] Preferred Reporting Items for Systematic Reviews and Meta-Analyses (-for Searching)

A) Structure of a search strategy

Searches in trial registries should show high sensitivity, be simple, and, if possible, only consider one component (usually the therapeutic indication or intervention) [296]. Firstly, terms for the search component that can best be depicted (generating the least number of hits, despite high sensitivity) are included in the search. The search is only further restricted if too many hits are retrieved. Due to the differing quality of the individual entries, further restrictions (e.g. according to study status) are only to be undertaken in exceptional cases.

B) Selection of trial registries

A systematic search always considers several trial registries, as no trial registry includes all studies [146,296,728]. The search is at least conducted in ClinicalTrials.gov and the International Clinical Trials Registry Platform (ICTRP) Search Portal of the WHO [43,296,485,486]. The ICTRP is a meta-registry that includes a large part of clinical studies [146,337]. However, the search functions are very limited [296,449] and the trial registry regularly produces error messages [345]. Important trial registries such as ClinicalTrials.gov are therefore searched directly, even though they are also covered via the ICTRP [296,449].

In addition, EMA databases (EU-CTR and CTIS) should be considered for benefit assessments of drugs. Furthermore, trial registries of the pharmaceutical industry (registries of individual companies) and the Drug Information System (AMIce[45]) of the Federal Institute for Drugs and Medical Devices (BfArM[46]) [117] can also be searched. This database includes results reports of studies conducted outside Europe and the United States.

C) Further procedure

The procedure for trial registries regarding quality assurance, conduct of the search, study selection, documentation, as well as search updates largely follows the search in bibliographic databases. There is one exception: the trial registry entries are selected in a one-step procedure, as the complete information is directly available via the website of the trial registry.

8.1.3 Requests to manufacturers

For the benefit assessment of drugs or procedures largely based on a medical device, 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 relevant information on these studies, independent of their publication status.

[45] Arzneimittel-Informationssystem

[46] Bundesinstitut für Arzneimittel und Medizinprodukte

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 for drugs [legally-binding German version and English translation] [388]). It specifies the procedure, the requirements for the documents to be submitted, and their confidential and non-confidential components.

This request is usually made in 2 steps. In the first step, the Institute asks the manufacturer to supply a complete overview of all studies conducted by the manufacturer. In this context, the Institute defines the project-specific inclusion criteria for this overview. In the second step, 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 is also published in the Institute's reports in order to ensure transparency.

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 the selective provision of information (see Section 3.2.1).

8.1.4 Further information sources and search techniques

To identify further relevant studies or documents, depending on the research question, further information sources are used and further search techniques applied.

With regard to study selection and documentation in the report, differences arise that result partly from the limited search options and partly from the type of data searched. For instance, certain information sources are only screened by one person with regard to studies, who then assesses these studies in respect of their relevance; a second person checks the whole process including the assessments (e.g. publicly accessible documents of regulatory authorities, screening of reference lists).

The following text presents further information sources and search techniques that are considered either as a standard or optionally in the benefit assessment.

A) Regulatory authorities

If drug assessments or treatment methods largely based on a medical device are assessed, publicly accessible documents from regulatory authorities are a potential source for information retrieval.

Documents of regulatory authorities

Information on drugs approved centrally in Europe (e.g. European Public Assessment Reports, EPARs) is searched for via the EMA website [237]. In the United States, access to the FDA's Medical Reviews and Statistical Reviews is provided via Drugs@FDA [750].

In contrast to the United States, there is no central approval procedure for medical devices in Europe. Publicly available information on medical devices is currently only occasionally available in individual countries, for example at NICE via the "List of interventional procedures" [547]. In future, the results of studies on new implantable and class III devices are to be made (partly) publicly accessible in the European Database on Medical Devices (EUDAMED) [235]. In the United States, information on medical devices evaluated by the FDA, including the information on the underlying data basis, can be searched for via Devices@FDA [749].

B) Selected scientific journals and conference proceedings

Depending on the research question, it can be useful to conduct a handsearch in selected scientific journals. This is decided on a case-by-case basis. A search of conference proceedings is usually dispensed with, as these documents usually contain little information on study methods and results [212].

C) 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 evaluated with regard to whether they refer to studies that fulfil the inclusion criteria of the assessment.

D) G-BA website and IQWiG website

Depending on the research question, it can be useful to screen the G-BA and IQWiG websites for previous IQWiG products, publicly accessible manufacturer documents (e.g. Modules 1 to 4 of dossiers on early benefit assessments pursuant to §35a SGB V), as well as G-BA resolutions with regard to whether studies are mentioned that fulfil the inclusion criteria of the assessment.

E) Application of further search techniques

Different search techniques have been established for preparatory searches at the start of projects, for research questions that are difficult to search for (e.g. complex interventions), and for the evaluation of search strategies in bibliographic databases [348]. These include citation-based search techniques (such as the screening of reference lists and citation tracking) as well as the use of the "similar articles" function. For all benefit assessments, the screening of reference lists of systematic reviews is usually conducted as an additional search technique.

F) Hearing

Studies or study information transferred within the framework of the hearing on the preliminary report plan or on the preliminary report are considered.

G) Requests to authors

If the available information on a study is incomplete, unclear, or contradictory, it may be useful to contact the authors of study publications. These requests can be undertaken to better evaluate the suitability, the methodological quality or results of a study [221,540,604]. Such requests are usually undertaken only if a relevant impact on the report can be expected.

8.2 Focused information retrieval

It is not necessary or possible to conduct information retrieval targeted towards completeness for all research questions. In such cases, so-called focused information retrieval is conducted, especially if the requirement of a systematic and transparent approach still exists.

Focused information retrieval is, for example, conducted 1) for projects with a short processing time (e.g. dossier assessments, evidence reports), 2) if a research question is not targeted towards completeness (e.g. in qualitative research), or 3) if the assessment is to be based on systematic reviews. Focused information retrieval aims to achieve a balanced relation between sensitivity (i.e. completeness) and precision (i.e. accuracy).

The approach regarding the development of the search strategy, quality assurance, conduct of the search, study selection and documentation is based on comprehensive information retrieval (see Section 8.1). However, restrictions or adaptions can be undertaken in the following areas, for example, as frequently applied in the area of rapid reviews [447,746]:

- in the selection of information sources (e.g. databases): depending on the topic, different sources of information (at least 2) that are very likely to contain relevant documents can be used [447]
- in the selection of study filters
- in the restriction of search periods (publication years) and/or languages
- in the selection of studies (performed by one person; quality assurance of the results by a second person)
- in the presentation of methods and results

8.2.1 Search for systematic reviews

In the search for systematic reviews it is sufficient if a large proportion of the high-quality and current systematic reviews on a research question is identified. In this context the search is conducted in at least the following databases:

- MEDLINE
- the HTA Database, and
- the Cochrane Database of Systematic Reviews

Precise study filters are applied in the development of search strategies (e.g. the "High specificity strategy" [795]). If necessary, the search period of the search is restricted. For example, as a rule searches for systematic reviews used as a basis for health information are restricted to the last 3 years.

In addition, systematic reviews can be used as a source for primary studies in order to conduct a benefit assessment based on these studies [611]. For this purpose, an assessment of the inclusion and exclusion criteria, as well as the information retrieval in the systematic review(s) identified, is conducted beforehand [675]. One (or potentially several) high-quality and current systematic review(s) is/are then chosen (so-called basic systematic reviews [SRs]), and the primary studies considered are extracted and then selected. In this approach only the search results of the basic SRs used are taken over, but not the assessment of the primary studies included or the data extraction. In addition, an update of the information retrieval according to Section 8.1 is mostly conducted for the period not covered by the systematic review(s) (see Section 8.2.1).

If information sources listed in the report plan or project outline were not considered in the systematic review or not searched comprehensively (e.g. trial registry), these sources can be included within the framework of information retrieval for the benefit assessment.

8.2.2 Further research questions with focused information retrieval

Focused information retrieval is usually conducted for the following research questions:

- Evidence reports
- Searches for disease registries
- Information requirements within the framework of HEEs
- Addenda to §137e or §137h assessments
- Searches for qualitative research
- Searches for confounders (see Section 9.3.6)
- Evaluation of the completeness of the study pool

8.3 Exploratory searches

A search is called an exploratory search if information on certain aspects is required. The search ends as soon as the information required is available.

The information sources are very topic dependent and often comprise clinical information systems such as Dynamed and UpToDate or guideline databases. In addition, depending on the research question, specific data collections are used, such as those of the Robert Koch Institute, the Federal Statistical Office or the AOK[47] Research Institute (WIdo[48]), as well as data from regional registries, laws, regulations or directives.

In contrast to comprehensive information retrieval, the search for and selection of data is performed by one person. Quality assurance of the result is performed by a second person. Documentation in the report is restricted to the presentation of specific results.

Examples of exploratory searches are preliminary searches at the start of projects (see Section 8.1.1) as well as searches for cost data (see Section 4.14) and for epidemiological data (see Section 4.14).

8.4 Search for guidelines

If the aim of the search is to comprehensively identify clinical practice guidelines, it is conducted in guideline databases (e.g. the AWMF, the Canadian Medical Association [CMA] Infobase, and the TRIP Medical Database) and on the websites of providers of specialist and multi-disciplinary guidelines.

If focused information retrieval for key guidelines is sufficient for certain projects (e.g. due to specific research questions), a search in MEDLINE is conducted. The following databases and websites are also taken into account, in particular the TRIP Medical Database, AWMF online, the CMA Infobase, ECRI Guidelines Trust, the Guidelines International Network (GIN) Library, and NHS Evidence Search.

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 reference management programme. The procedure of the selection of guidelines is conducted as presented in Section 5.2.2. However, within the framework of the search in guideline databases and websites of guideline providers, depending on the research question, in a supplementary step it is evaluated whether a methodological system was used in the development of the guideline. This usually means

[47] Allgemeine Ortskrankenkasse (Local Healthcare Fund, an SHI fund)

[48] Wissenschaftliches Institut der Allgemeinen Ortskrankenkasse

whether a guideline is evidence based or not (see Chapter 5). Within the framework of the production of the report plan, the inclusion and exclusion criteria are specified a priori.

Title and abstract screening is performed by one person; the quality assurance of this step is performed by a second person. The following steps (from the full-text screening) are performed by 2 persons independently of each other.

Within the framework of guideline appraisal it can be useful in the individual case to contact authors of publications on guidelines or guideline developers. The requests can refer, for example, to specific details of individual guidelines or to non-published partial aspects of publications.

8.5 Assessment of information retrieval

In the preparation of a dossier or application for testing, a search in bibliographic databases and a search in publicly accessible trial registries must as a matter of principle be conducted by the applicant; the precise requirements are provided in the G-BA's Code of Procedure [284,287]. In a similar way, hospitals preparing information on an assessment in accordance with §137h SGB V must systematically search for information on the method requested.

The Institute conducts an evaluation of the information retrieval documented in the presented documents for assessments according to §35a SGB V, assessments of potential, and assessments in accordance with §137h SGB V. For all assessment procedures, the searches in bibliographic databases and trial registries, as well as the study selection, are evaluated. This evaluation is based on the procedure described in Section 8.1.1 regarding the quality assurance of search strategies, as well as on the document templates included in the requirements of the G-BA's Code of Procedure [284,287].

Evaluation of the completeness of study pools

Depending on the results arising from the assessment of the dossiers, different strategies are available to check the completeness of information retrieval. For example, random checks of the literature citations excluded in the dossier can be performed or the Institute can conduct its own search and/or study selection by means of a check of completeness (see Section 8.2.2). If a high number of hits is retrieved, the comparison can also be performed on the basis of systematic reviews (see Section 8.3 or 8.2.1). The result of the check of the completeness of information retrieval and the description of the approach in this regard form part of the dossier assessment.

9 Assessment of information

As a matter of principle, the step of information retrieval (see Chapter 8) is followed by an assessment step in which the information retrieved is systematically assessed with regard to its informative value. This chapter primarily explains the aspects that apply to the assessment of information within the framework of benefit assessments.

In research the term "bias" means a systematic deviation between research results and the "truth" [630]. 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 disclosure of relationships. All these methods also form the legal basis of the Institute's work.

9.1 Quality assessment of individual studies

It is essential that at least the most relevant information on a study is available to allow use of the study results. Submissions to scientific meetings, published abstracts or press releases on studies usually contain far too little and unreliable information on study methods and results to allow use of these sources for benefit assessments [375].

In order to assess the usability of a source, the Institute follows the various statements that have been prepared to improve the quality of publications. These include, among others:

- the CONSORT[49] statement on RCTs [658] and the corresponding explanatory document [530]
- a proposal for an extension of the CONSORT statement for randomized studies on non-drug interventions [83] and the corresponding explanatory document [82]
- the CONSORT statement on cluster-randomized trials [130]
- the CONSORT statement on the documentation of adverse events [403]
- the CONSORT statement on non-inferiority and equivalence studies [589]
- the CONSORT statement on pragmatic studies [808]
- the CONSORT PRO extension for PROs [129]
- the CONSORT extension for multi-arm randomized trials [430]

[49] Consolidated Standards of Reporting Trials

- the TREND[50] statement on non-randomized intervention trials [181]
- the STROBE[51] statement for observational studies in epidemiology [769] and the corresponding explanatory document [757]
- the RECORD[52] statement for observational studies with routine data [51]
- das RECORD statement for pharmaco-epidemiological studies (RECORD-PE) [476]
- the TRIPOD[53] statement for prognostic studies [152] and the corresponding explanatory document [536]
- the STARD[54] statement on diagnostic studies [79,80] and the corresponding explanatory document [81]
- die ISOQOL[55] reporting standards for PROs [113]

9.1.1 Criteria for study inclusion

The problem often arises that studies relevant to a benefit assessment do not fully meet the inclusion criteria for the patient population defined in the systematic review or the defined test and comparator interventions. In general, the Institute applies the following criteria.

If analyses of the relevant subpopulation or the relevant test and comparator interventions are available for a study, these analyses are used. If such analyses are not available, the study is generally included if the criterion (population) applies to at least 80% of the patients included and the inclusion criterion for the test intervention (intervention group of the study) and the comparator intervention (comparator group of the study) is in each case met to at least 80%. Regardless of the degree of fulfilment (at least 80%, less than 80%), there may be situations in which there is suitable information on an effect modification by the inclusion criterion in question (population or interventions). In these situations, the inclusion of the study must be decided on the basis of the strength of the effect modification and the proportion of patients who do not meet the inclusion criterion or the degree of deviation of the interventions.

[50] Transparent Reporting of Evaluations with Non-randomized Designs

[51] Strengthening the Reporting of Observational Studies in Epidemiology

[52] Reporting of Studies Conducted Using Observational Routinely Collected Health Data

[53] Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

[54] Standards for Reporting of Diagnostic Accuracy

[55] International Society of Quality of Life Research

9.1.2 Relation 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 or ventricular fibrillation; see Section 3.2.2). Randomization and blinding are quality criteria that increase the informative value of controlled studies. Parallel group studies [595], cross-over studies [426], and cluster randomized studies [207] are common designs for clinical trials. If interim analyses are planned, the use of appropriate sequential designs must be considered [426].

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 [92] to investigate the association between exposures and the occurrence of rare diseases, as well as cohort studies [93] 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 [435].

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).

9.1.3 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 [328,335]. 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 [437,776]. Due to the complexity of the appraisal of studies, no conclusive judgement on quality can be inferred from the hierarchy of evidence [31,793]. According to EBM standards, the Institute follows the rough

hierarchy of study types, which is widely accepted and is also largely consistent with the evidence classification of the Fourth Chapter §7 (3) of the G-BA's Code of Procedure [287] and has been incorporated in the regulation on the benefit assessment of drugs according to §35a SGB V [119]. 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 [326,359,767], 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 9.1.4) and not to the evidence level of the study.

9.1.4 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) [400,454]; for epidemiological studies, they are defined by guidelines and recommendations to ensure good epidemiological practice (GEP) [184]. 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.

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 [711].

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.

Standardized assessment forms are used to assess the risk of bias. 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.

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 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. When assessing individual aspects of the risk of bias, the Institute follows the criteria of the ROBINS-I[56] instrument [710]. The Institute specifically deviates from this procedure in assessments of the potential of new examination and treatment methods (see Section 3.8).

The QUADAS-2[57] criteria [782,783] are considered for the assessment of diagnostic accuracy studies; these criteria are adapted to the specific project, if necessary. The PROBAST[58] instrument [794] is primarily used for the methodological assessment of prognostic studies (see Section 3.5).

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.

[56] Risk of Bias in Non-randomized Studies of Interventions

[57] Quality Assessment of Diagnostic Accuracy Studies

[58] Prediction Model Risk Of Bias Assessment Tool

9.1.5 Interpretation of composite outcomes

A so-called 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 [158,266,277]. 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.

9.1.6 Assessment of data consistency

To assess the informative value of study results, the Institute will review the consistency of data with regard to their plausibility and completeness. On the one hand, implausible data are produced by the incorrect reporting of results (typing, formatting, or calculation errors), and

on the other hand, by the insufficient or incorrect description of the methodology, or even by forged or invented data [18]. Inconsistencies may exist within a publication, and also between publications on the same study.

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 [18].

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.

9.2 Consideration of systematic reviews

9.2.1 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. High-quality systematic reviews aim to overcome this form of bias by identifying, assessing and summarizing the evidence systematically, reproducibly and transparently, rather than selectively [208,224,298,583].

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 [224,364].

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 [298]. Systematic reviews of diagnostic and screening tests also show some methodological differences compared with reviews of treatment interventions [174].

For the work of the Institute, systematic reviews are mostly used to identify potentially relevant (primary) studies. However, an assessment can be based partially or even solely on systematic reviews (see Section 9.2.2). Health information produced by the Institute for patients and consumers is to a large part based on systematic reviews. These include, in particular, systematic reviews of treatment effects, causes of disease or prognostic factors. Syntheses of qualitative research are also used (see Section 9.4).

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 [413,568,570], the AMSTAR [674-676], AMSTAR 2 [677] or the ROBIS[59] instrument [781]. 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 health care contexts. The following factors are usually considered: the population of the participants in the included studies (including gender and baseline disease risk); the health care context (including the health care settings and the medical service providers); and the applicability and likely acceptance of the intervention in the form in which it was assessed [72,170].

9.2.2 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 [159,479]. In order to use systematic

[59] Risk of Bias in Systematic Reviews

reviews in a benefit assessment these reviews must be of sufficiently high quality, that is, they must

- 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 [30,568,784]. In addition, it is an essential prerequisite that the information retrieval conducted in the systematic reviews does not contradict the Institute's methodology (see Section 8.2.1) 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 general quality of systematic reviews is assessed using the AMSTAR [674-676], the AMSTAR 2 [677] or the ROBIS instrument [781]. For these instruments, no clear thresholds are defined for when a systematic review is considered to be of sufficient quality; such thresholds may therefore need to be defined a priori. In addition, as a rule, the sponsors of systematic reviews, as well as relationships of authors, 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 accuracy 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 [412].

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

Benefit assessments 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.

9.2.3 Consideration of published meta-analyses

Following international EBM standards, the Institute's assessments are normally based on comprehensive information retrieval 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 9.3.7). 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 [687]. 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 the PRISMA statement on meta-analyses of randomized trials [572] and in the corresponding explanatory document [573], in the PRISMA-IPD statement for meta-analyses with IPD [712], in the PRISMA-P statement for protocols of systematic reviews [532] and in the corresponding explanatory document [673], in the PRISMA harms checklist, [806], the PRISMA-DTA statement for meta-analyses of diagnostic test accuracy studies [526], as well as in in a document published by EMA [238]. 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.

9.3 Specific statistical aspects

9.3.1 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 [52]. Chapter 10 of the *Cochrane Handbook for Systematic Reviews of Interventions* [176] provides a well-structured summary of the advantages and disadvantages of typical effect measures in systematic reviews. Agresti [15,16] 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. [23] 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 [17] provides an up-to-date discussion on exact methods.

9.3.2 Statistical significance tests and confidence intervals

The Institute uses statistical significance tests and confidence intervals to support decision-making. However, in the context of assessments based on existing data, the results of statistical significance tests cannot represent strictly confirmatory hypothesis tests. The Institute follows the convention of speaking of a statistically significant result when the p-value is less than 5% ($p < 0.05$). Depending on the research question and conclusion, it may be necessary to require a different significance level. The Institute always explicitly justifies such exceptions.

Furthermore, a 2-sided question is as a rule assumed. Exceptions to this are, for example, non-inferiority studies. For other 1-sided questions, the Institute follows the usual approach of lowering the threshold for a statistically conspicuous result from 5% to 2.5%.

In systematic reviews and the Institute's projects, in particular the benefit assessments, the results of numerous outcomes are described using statistical significance tests and confidence intervals. The resulting problem of multiple testing cannot be formally solved in systematic reviews, but should nevertheless at least be taken into account when interpreting the results [55]. Where appropriate and possible, the Institute applies methods to adjust for multiple testing [60]. In its benefit assessments (see Section 3.1), the Institute attempts to control for type I errors separately for the conclusions on each single aspect of benefit, for example, by primarily using the cross-time analysis for survival data. A summarizing evaluation across all outcomes is not usually conducted in a quantitative manner, so that formal methods to adjust for multiple testing are not applied here (see Section 3.1.5).

Conversely, the interpretation of non-statistically significant results also requires attention. In particular, such a result is not interpreted as evidence of the non-existence of an effect

(absence or equivalence) [21]. To demonstrate equivalence or non-inferiority, the Institute uses appropriate methods for equivalence or non-inferiority hypotheses.

In principle, Bayesian methods may be regarded as an alternative to statistical significance tests [700,701]. Depending on the research question posed, the Institute will, where necessary, also apply Bayesian methods (e.g. in meta-analyses or for indirect comparisons, see Sections 9.3.7 and 9.3.8).

9.3.3 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 effects) 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 [756]. 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 [620]. 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 [262,381]. 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 [477,736]. 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 [364]). 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 [791]. 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 [443]. 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 [444]. 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 [612].

Evaluation of relevance for scales

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 [661]. 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 [535]. Which approach can be chosen in the Institute’s assessments depends on the availability of data from the primary studies.

In recent years, responder analyses based on a response criterion in terms of an individual minimally important difference (MID) have been increasingly conducted. However, the methodological problems of this approach are becoming increasingly apparent.

In principle, empirical methods that determine an MID using patient-reported anchors are preferred for determining MIDs; in addition, distribution-based methods are also used [567,606]. Systematic compilations of empirically determined MIDs show that a large number of MIDs are often published for individual measurement instruments, which can range widely within an instrument [136,140,333,555]. This variability may be due, among other things, to the fact that studies on determining MIDs use different anchors, observation periods or analytical methods [193,555,567]. There is currently no established standard that can be used to assess the quality of these studies and to estimate the informative value of the MIDs

determined [136,193,424]. A first proposal for a quality assessment instrument of studies on determining MIDs and on the informative value of MIDs was published in 2020 [192]. Comprehensive results on applicability are still pending. However, a study on the reliability of this new instrument shows that in publications of studies on determining MIDs, essential parts of the methodology or the criteria for an assessment are mostly not reported at all [192]. Thus, an empirically determined MID for benefit assessments based on methodological quality criteria cannot be selected at present [191,193,424].

In addition to the methodological factors influencing the magnitude of the MIDs determined, another part of the variability of MIDs is caused by the fact that they depend on characteristics of the patient population in which the instrument is used and other contextual factors. For example, the severity of the disease, the type of intervention used, or the question as to whether patients experience an improvement or deterioration of their condition, may all influence the MID [19]. How to deal with this part of the variability of MIDs remains unresolved in the international discussion.

Against this background, the procedure described below is applied to ensure that suitable response thresholds are used in responder analyses within the framework of benefit assessments. On the one hand, the aim is to ensure that a response threshold reflects changes that are perceptible to patients with sufficient certainty. This response threshold is intended to represent a relatively small change, but the value may well be above a minimum threshold. This property also takes into account the empirical variability of MIDs and helps to ensure that the response criterion does not fall below the MIDs too frequently in different constellations. At the same time, selective outcome reporting is to be minimized, which could arise, for example, by selecting any one of many possible MIDs.

To develop an approach for the use of MIDs in benefit assessments, systematic reviews on MIDs were systematically identified in a bibliographic search [20,205,218,333,419,555,567,702]. These systematically compiled MIDs were related to the range of the respective scales. This approach identified a value of 15% of the range of the respective scales as a plausible threshold for a relatively small, but sufficiently certain noticeable change [396]. This as a rule results in the following approach:

1) If responder analyses using an MID are pre-specified in a study and the response criterion corresponds to **at least** 15% of the scale range of the measurement instrument used, these responder analyses are used for the assessment without further examination of the response criterion.

2) If pre-specified response criteria in terms of an MID are below 15% of the scale range, these are as a rule not used. In these cases and in those cases where no response criteria have been pre-specified at all, but analyses of continuous data are available instead, there are various possibilities. Either the analyses of continuous data can be used, in which case a general statistical measure is drawn upon in the form of standardized mean differences (SMD expressed as Hedges' g). In this context, an irrelevance threshold of 0.2 is 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 [259]. Alternatively, analyses specified post hoc with a response criterion of **exactly** 15% of the scale range can be considered.
3) If both suitable responder analyses (response criterion pre-specified at least 15% of the scale range or exactly 15% of the scale range post hoc) and analyses of continuous data are available, the responder analyses are used.

9.3.4 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 a 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) [69,183].

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.

9.3.5 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 [21]. To demonstrate "equivalence", methods to test equivalence hypotheses need to be applied [425]. 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" [166,384,614].

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 [613].

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 [425]. 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 [477]; on the other hand, the usual study design criteria, such as randomization and blinding, no longer sufficiently protect from bias [669]. 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 [425]. For this reason, particular caution is necessary in the evaluation of equivalence studies.

9.3.6 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 [438]. Studies investigating several interventions are a further important field of application for these methods [522]. A prerequisite for the application of such methods is the availability of complete information on IPD. In the medical literature, the reporting of results obtained with multi-factorial methods is unfortunately often insufficient [57,541]. To be able to assess the quality of such an analysis, the description of essential aspects of the statistical model formation is necessary [339,623], as well as information on the quality of the model chosen (goodness of fit) [376]. 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 the handling of missing data
- information on the handling of implausible data and outliers
- 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 [39,58,438].

The Institute uses common methods in its own regression analysis calculations [338]. 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 [338,719]. A well-informed and careful

preselection of the candidate predictor variable is essential in this regard [178]. For the modelling of continuous covariates, the Institute will, if necessary, draw upon flexible modelling approaches (e.g. regression using fractional polynomials [624,647]) to enable the appropriate description of non-monotonous associations.

Since a routine practice data collection (RPDC) is usually a comparative study without randomization, adequate adjustment for confounders also plays a key role for this type of study (see Section 3.3.4). Here it is necessary to identify all relevant confounders (including important interactions) a priori using a systematic approach [599], to record them completely, and to consider them in the model in an appropriate form when analysing the data [397]. It should also be noted whether biases in the effect estimate are to be expected due to different data collection times for the drug and the appropriate comparator therapy.

For data analysis with adequate adjustment for confounders, there are various causal methods (e.g. multifactorial regression models, instrument variables, propensity scores) that should be described in a detailed SAP [12,165,714,773]. In the presence of time-dependent confounding, the use of specific causal inference methods such as marginal structural models may be necessary [578,610]. The choice of the specific analysis method should be based on the objectives of the respective analysis and the available evidence. When reporting the methods used in the SAP, it is particularly necessary,

- to describe the causal model accurately, e.g. using causal graphs,
- to present the assumptions of the causal model,
- to give well-founded reasons why these assumptions can be justified in the specific case of use,
- to explain which procedures and criteria are used to adjust the model, and
- to explain which sensitivity analyses are used to check the robustness of the results.

In the context of an RPDC, the propensity score method plays a particular role in adjusting for confounders [253,397], among other things, because there are approaches here to evaluate and present the main assumptions of causal inference, namely sufficient positivity, overlap and balance [34]. However, these assumptions are also required for other causal methods. If one of these assumptions cannot be achieved with a propensity score approach, it is not a reasonable option to use a multifactorial regression model instead, for example [34,789].

Positivity means that the forms of treatment to be compared in the RPDC represent a treatment option for all patients included in the analysis. This must be evaluated and described as part of the inclusion criteria for the RPDC. In the available data basis, there must be sufficient overlap between the two groups to be compared with regard to the distribution

of the propensity score. Furthermore, it is necessary that the two groups to be compared are sufficiently balanced with regard to all relevant confounders.

Numerous methods exist to achieve sufficient overlap and balance in propensity score approaches [34,164,182,471,774]. Excluding patients in non-overlapping areas of the propensity score can improve overlap and balance, but may also change the target population to which the effect estimate ultimately relates. Therefore, this population must be described precisely and it must be investigated whether it corresponds to the original research question. If this is not the case, it may not be possible to adequately answer the original research question with the available data [182,397].

9.3.7 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 [224]. 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 – if meaningful – 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 experimental 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 [490]. 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 [229,670,765]. 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 [672]. 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 [99,305]. If the existence of heterogeneity cannot be excluded with sufficient certainty, a random-effects model should be chosen. Several methods exist for the conduct of meta-analyses with random effects [762].

According to recommendations from the literature, if a sufficient number of studies are available, as a rule the Knapp-Hartung method [344,448] should be used to conduct meta-analyses with random effects; in this context, the heterogeneity parameter is estimated using the Paule-Mandel method [402,761,763].

When using the Knapp-Hartung method for meta-analyses with random effects, it should be noted that in homogeneous data situations, misleadingly narrow confidence intervals can occur [619]. To avoid this, Knapp and Hartung [448] proposed an ad-hoc variance correction. According to recommendations in the literature, when the Knapp-Hartung method is applied, sensitivity analyses should always be performed using the fixed-effect model or the DerSimonian-Laird method [180] to investigate whether the use of the ad-hoc variance correction is appropriate [409,788]. If the confidence interval using the Knapp-Hartung method is narrower than that of the DerSimonian-Laird method, the Knapp-Hartung method with ad-hoc variance correction should be applied.

However, the use of meta-analyses with random effects reaches its limits in the case of very few studies (fewer than 5). As heterogeneity then cannot be reliably estimated [362], the use of meta-analyses with random effects can lead to very broad confidence intervals that potentially no longer allow conclusions on the evidence base. Especially in the case of very few studies, a fixed-effect model or a qualitative summary (see Section 3.1.4) should be considered [56]. If there are no arguments against the use of a fixed-effect model, this model should be chosen, especially in the case of only 2 studies. If the use of the fixed effect model

is not justifiable, it should be assessed whether a common effect estimate is meaningful or, otherwise, a qualitative summary can be made. To evaluate when a common effect estimate is meaningful, the Institute proceeds as follows. First, pooled effects are calculated according to the Knapp-Hartung method – with and without variance correction – and according to the DerSimonian-Laird method. It is checked whether the confidence interval according to Knapp-Hartung (without variance correction) is narrower than that according to DerSimonian-Laird. If this is the case, the Knapp-Hartung effect estimate with variance correction is used; if not, the one without variance correction is used. Subsequently, it is checked whether this effect estimate is informative. The estimate is considered informative if the confidence interval (of the common effect) is included in the union of the confidence intervals of the individual studies. If the effect estimate is informative, it is checked whether the result calculated by means of Knapp-Hartung is congruent with the result according to DerSimonian-Laird with regard to the conclusion on statistical significance. In the case of congruence, a common effect estimate is considered meaningful and this effect estimate (according to Knapp-Hartung) is used for the final evaluation. Otherwise, a common effect estimate is not considered meaningful. In summary, an effect estimate is not meaningful if the effect estimate is not informative or if the result calculated by means of Knapp-Hartung is not congruent with the result according to DerSimonian-Laird with respect to the conclusion on statistical significance.

Depending on the context, alternative procedures for meta-analytical summarization could also be an option, such as Bayesian approaches [40,248,278,279,618,693] or methods from the area of generalized linear models [71,470,519,594,688]. For the selection of a suitable approach and for sensitivity analyses, several alternative methods have to be applied, especially in the case of meta-analyses with very few studies [320]. 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 [361]. In this situation it is necessary to assess the heterogeneity of study results [299]. The existence of heterogeneity can be statistically tested; however, these tests usually show very low power [408,450]. In addition, it is also important to quantify the extent of heterogeneity. For this purpose, specific statistical methods are available, such as the I^2 measure [360]. 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% [176]. However, the I^2 measure is a relative measure for describing heterogeneity and not an absolute measure, which must be taken into account when interpreting it [75]. The I^2 measure is unsuitable as the sole measure for describing heterogeneity [75,627]. An absolute measure is provided by the heterogeneity parameter τ, which is also an important component of the prediction interval (see below). There are also proposals for a rough categorization of heterogeneity for the heterogeneity parameter τ (e.g. using the categories of "acceptable": 0.1 to 0.5, "fairly high": 0.5 to 1.0 and "extreme": > 1.0 [618].If the heterogeneity of the studies is too large, the statistical pooling of the study results may not be meaningful [176].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 0.05. In this context, the location of the effects also plays a role. If the individual studies show a clear conclusive effect, 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 [281]. 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 (see Section 7.7).

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 [738,754]. 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 [738]. 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 [314]. 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 [585,687,738] (see Section 9.2.3).

The Institute uses prediction intervals to display heterogeneity within the framework of a meta-analysis with random effects [318,362,607,761]. 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. [318] 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 (less than 5) is critically discussed in the literature, as potential heterogeneity can only be estimated very imprecisely [362,410]. The Institute generally presents prediction intervals in forest plots of meta-analyses with random effects if at least 5 studies are available or if the graphic display of heterogeneity is important, for example, to assess whether and to what extent the effects are conclusive (see Section 3.1.4).

If no pooled effect is displayed because heterogeneity is too large, prediction intervals are also used in order to evaluate whether the effects observed in the available studies are moderately or clearly conclusive (see Section 3.1.4). In this case, as a rule prediction intervals are used if at least 4 studies are available.

C) 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 [176]. 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 [85,176]. However, the use of this method is only adequate if the effects to be estimated are not too large and the design is not unbalanced [97,98].

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. Further potential methods comprise logistic regression models with random effects [688,748], beta-binomial models [470,519] exact methods [741] or the application of the arcsine difference [626]. 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.

D) Meta-analyses of diagnostic accuracy studies

The results of diagnostic accuracy studies can also be statistically pooled by means of meta-analytic techniques [190,407,507]. 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 diagnostic accuracy studies are likewise of limited relevance.

The same basic principles apply to a meta-analysis of diagnostic accuracy studies as to meta-analyses of therapy studies [190,602]. 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 accuracy studies; therefore it is usually advisable here to apply random-effects models [190]. Such a meta-analytical pooling of diagnostic accuracy studies can be performed by means of separate models for sensitivity and specificity, if the evidence on the use of bivariate models is insufficient [731]. 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 [334,603]. These methods also enable consideration of explanatory variables [332]. Meta-analytical methods have also been developed for complex data situations with several diagnostic thresholds at the study level [377,428,709]. 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 test accuracy studies.

E) 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 [76,100,101,558,739,777]. 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.

9.3.8 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 [499-501], multiple treatments meta-analysis (MTM) [128], or network meta-analysis [504,625,633]. These methods represent an important further development of the usual meta-analytic techniques [632]. 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 [38,282,634,695,723]. For this reason, direct comparative studies are primarily used in benefit assessments of interventions (placebo-controlled studies as well as head-to-head comparisons); this means that conclusions for benefit assessments are preferably inferred from the results of direct comparative studies. Adequate justification is required if methods for indirect comparisons are to be used. In addition, an essential precondition 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.

In certain situations, as, for example, in assessments of the benefit of drugs with new active ingredients [4], 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 (of 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, it can be necessary to regularly consider indirect comparisons to assess cost-effectiveness ratios (see Chapter 4), 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 use of non-adjusted indirect comparisons (i.e. the use of single arms from different studies) is disapproved [61]. This also applies to methods for indirect comparisons in which, via modelling with strong assumptions about unknown effects [135] or by means of methods from the area of causal models for observational studies with non-testable assumptions [685], it is attempted to enable estimations of effects despite missing common comparators. Only adjusted indirect comparisons via adequate common comparators are accepted. These particularly include the approach by Bucher et al. [114], as well as the network meta-analysis methods mentioned above.

Interpretable comparisons in unconnected networks without adequate common comparators are only possible if access to the complete IPD information is available. In this case, the methods for comparisons in non-randomized studies can be used (see Section 9.3.6), for which further requirements must be met, such as the systematic identification and recording of all relevant confounders.

Besides the assumptions of sufficient similarity and homogeneity of the pairwise meta-analyses, which must also be fulfilled here, in network meta-analyses sufficient consistency of the effects estimated from the direct and indirect evidence is additionally required. The latter is a critical point, as network meta-analyses provide valid results only if the consistency assumption is fulfilled. Several methods are available for examining the consistency assumption [195,206,501]. However, they have not yet been insufficiently investigated and no methodological standard has so far been established here [715]. In addition, consistency cannot always be examined, as a comparison of direct and indirect evidence is not possible (e.g. in the method following Bucher et al. [114]). In these cases in particular, a very careful evaluation of similarity and homogeneity is therefore required [442]. If serious doubts exist whether one or several of the basic assumptions are fulfilled to a sufficient extent, then indirect comparisons should not be used [441]. In practice it is necessary to describe completely the model applied, together with any remaining unclear issues [723]. These issues should be carefully examined in sensitivity analyses. The guidelines available in the literature on the conduct and assessment of indirect comparisons should be observed [10,367,382,415,416,441].

9.3.9 Subgroup analyses

With subgroup analyses it is examined whether the results of one or several studies differ between the different subgroups included in these studies (e.g. patient with versus patients without renal dysfunction). This difference can be qualitative (reversal of the effect in one subgroup, but not in another) or quantitative (different effect sizes).

Such subgroup analyses are useful for the targeted use of medical interventions, as with these analyses patient groups can potentially be defined for whom an intervention has a benefit or for whom the same intervention is more likely to be harmful than beneficial. This information can also lead to a restriction of the therapeutic indication of an intervention, for example in the approval of drugs. Even though subgroup analyses are useful for treatment optimization, they are in part a matter of controversy in the methodological literature [29,569].

- 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, the result within a subgroup does not necessarily have to be statistically significant [306]. 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 [102].

- Multiple testing: If several subgroups are analysed, results in a subgroup may well reach statistical significance, despite actually being random.
- Comparability between treatment groups within the subgroups: If randomization was not stratified according to the subgroup characteristic, in the event of small sample sizes within the subgroups the treatment groups could differ with regard to prognostic factors [162,720]. In this case, the comparability between treatment groups within the subgroups is jeopardized, so that (non-) existing differences between subgroups can be caused by this imbalance alone.
- Effect modification through more than one subgroup characteristic (interaction of higher order): If for one outcome there is a difference, for example, between 2 age groups, as well as between men and women, to interpret the results separate analyses are required for each age group as well as for men and women (i.e. analyses of 4 subgroups). However, such analyses are rarely available.

Moreover, it is being discussed that subgroup analyses generally have no characteristic of proof, particularly if they were not planned a priori. If subgroup analyses with regard to more or less arbitrary subgroup-forming characteristics are conducted post hoc, their results cannot be regarded as a methodologically correct testing of a hypothesis. Whereas in general subgroup analyses conducted post hoc on a study level should be viewed critically (also in view of the methodological limitations named above), 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. Such subgroup analyses are not to be designated as “post hoc” in terms of the systematic review, but correspond to a hypothesis to be tested in this review. In this respect the analysis of heterogeneity between the individual studies (and thus potentially the analyses of subgroups) are a scientific necessity. Subgroup analyses of characteristics not recorded before randomization but during the course of the study (e.g. patients with versus patients without myocardial infarction under the treatment investigated) are as a matter of principle unreliable.

On the one hand the aspects mentioned above require the assessment of the credibility of subgroup analyses; Sun et al. [720] identified criteria for this purpose. On the other hand, despite these limitations, for some research questions subgroup analyses may represent the best scientific evidence available in the foreseeable future in order to assess effects in subgroups [276], since, for example, factors such as ethical considerations may argue against the verification of the observed findings in further studies.

Expected differences in effects between different, clearly distinguishable patient populations are an important reason for conducting subgroup analyses [457,617]. 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.

Subgroup analyses can also be a necessity from the perspective of social law: According to §139a (2) SGB V, the Institute is obliged to consider characteristics specific to age, gender, and life circumstances. In addition, it should also be elaborated 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 [185]. A corresponding objective can also be found in §35a SGB V regarding the assessment of the benefit of drugs with new active ingredients [4]. In this assessment, patient groups should be identified in whom these drugs show a therapeutically relevant added benefit.

When interpreting subgroup analyses it should be considered that a statistically significant effect in one subgroup, but no effect or a reversed effect in another subgroup, cannot on its own (by means of inferential statistics) be interpreted as the existence of different effects between subgroups. Instead, first of all the statistical demonstration of different effects between different subgroups should be conducted by means of an appropriate homogeneity or interaction test. If a certain probability for such a demonstration is to exist at all, as a rule subgroup analyses are only conducted if each subgroup comprises at least 10 people and, in the event of binary data and survival times, at least 10 events occurred in one of the subgroups. An “event” means an event that occurred during the course of the observation period and was not detectable at baseline (e.g. achievement of viral clearance in infected persons).

If the result of a heterogeneity or interaction test between important subgroups is significant at the level of $\alpha = 0.05$, an effect modification (i.e. different effects between subgroups) is present. In this case the results of subgroups are not pooled to a common effect estimate. In the case of more than 2 subgroups, pairwise statistical tests to detect interactions are conducted, if meaningful. Pairs that are not statistically significant at the level of $\alpha = 0.05$ (with simultaneous statistical significance of the remaining pairs) are summarized into one group. The results of the remaining groups are reported separately and separate conclusions on the benefit of the intervention for these groups are inferred. If no pairs can be found that are not statistically significant at the level of $\alpha = 0.05$, no pairs are formed; instead a separate conclusion is drawn for each subgroup. In situations that are inconclusive on the basis of the pairwise interaction tests (e.g. when only a single pairwise interaction test is statistically significant), a decision is made on a case-by-case basis as to whether subgroup pairs should be combined into one group, and if so, which ones.

An exception to the requirement for an adequate homogeneity or interaction test exists if a necessity according to social law arises for subgroup analyses through 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 populations 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 greater the differences between approvals 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 within the study or not.

9.3.10 Handling of unpublished or partially published data

In the quality assessment of publications, in practice the problem frequently arises that essential data or information is partially or entirely missing (see Section 8.1). 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 8.1.3). 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 [194,379,575]. 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 or a study cannot be used in the analysis. In such cases, it will be noted that further data exist on a particular topic, but are not available for a quality assessment or for the analysis; this is accordingly taken into account in the assessment across studies (e.g. by downgrading the certainty of conclusions, see Section 3.2.1).

9.3.11 Handling of incomplete data

Whenever a study is followed up, it is inevitable that study participants will no longer be available for data collection on outcomes. These losses to follow-up can be due to a wide variety of reasons, e.g. death, study discontinuation or switch of intervention (by the doctor or patient), revocation of the informed consent due to adverse effects or lack of efficacy, newly occurring comorbidity or the stressful nature of the follow-up itself. The standards for reporting clinical trials stipulate that the reasons for losses to follow-up – if known – must be reported separately for each intervention group (see Section 9.1.1). Losses to follow-up, i.e. study participants who are included in the analyses with incomplete data, should be distinguished from study participants who were included in the study but are completely disregarded in the analyses (regardless of whether follow-up occurred, is complete, or is incomplete). For an intention-to-treat analysis, if possible, all randomized study participants should also be considered in the analysis [264].

Both losses to follow-up and study participants who remain completely unaccounted for in the analyses lead to missing data (partial or complete). Missing data not only increase the statistical uncertainty of the effect estimate, but also lead to a risk of bias unless it is clearly evident from the reasons for losses to follow-up or exclusions from the analyses that drop-outs were missing completely at random [496]. The risk of bias increases with an increasing proportion of missing values.

In addition to losses to follow-up, missing values can also occur sporadically or temporarily – even at baseline – and even if a study participant attends the study visit (e.g. loss of samples, technical failure or errors in the documentation of measurements). However, missing values due to such reasons are of minor importance in practice.

In a 2-step process, the missing values are assessed. In the first step, study participants who were completely excluded from the analyses are evaluated. Results are generally not included in the benefit assessment if they are based on less than 70% of the study participants to be included in the analysis, i.e., if the proportion of study participants not included in the analysis at all is greater than 30%. In some of the literature, analyses in which 20% of the study participants are not included are no longer considered meaningful [659]. Furthermore, the results are also not included in the benefit assessment if the difference in the proportions of study participants not included is greater than 15 percentage points between the groups, as this indicates non-random exclusion. Depending on the context, these rules are not rigid but should be viewed as guidance. Exceptions to these rules are made when exclusions are shown to be completely at random.

In the second step, the losses to follow-up are evaluated, i.e. the study participants included in the analyses but incompletely observed. No fixed limits are applied for the assessment of

the risk of bias due to these losses. The extent of the risk of bias is determined, depending on the context, by the number, the time points as well as the reasons for the losses to follow-up. In particular, differences between groups in these factors increase the risk of bias. If the proportion of missing values is too high or inappropriate replacement strategies were used, the results may not be taken into account.

The results are considered robust if the potential problems due to losses to follow-up are compensated by adequate replacement strategies or statistical analysis methods [50,272]. For this, it is necessary to know in detail and be able to describe the mechanism that led to the missing data. This will be impossible in practice if the proportion of missing data is too high.

9.3.12 Handling of variable observation periods

If the entry time of patients differs in clinical studies (recruitment period), but the follow-up time point is identical, the result is that the observation periods for the individual patients are different. In order to take this adequately into account, methods of survival time analysis are available [446,482]. Commonly used standard methods are the Kaplan-Meier curve for estimating the survival function, the log-rank test for comparing multiple survival functions, and the Cox model for estimating (adjusted) hazard ratios. When examining non-fatal outcomes (such as non-fatal myocardial infarction), it is important to note that death is always a competing event for non-fatal events. In this case, special methods of survival analysis for competing risks must be applied [35,431]. In this context, the application of the usual Kaplan-Meier curve leads to an overestimation of the absolute risk [474] and it should therefore not be used. Instead, the Aalen-Johansen estimator for the cumulative incidence function should be applied [653]. In contrast, the application of the Cox model for the cause-specific hazard function represents a correct analysis. However, for a complete analysis of all data, one also needs the Cox model for the competing event [653]. Another option is the application of the Fine-Gray model [268]; the corresponding explanations can be found in the literature [36,600].

Particularly when investigating adverse events and PROs in oncological studies, the additional problem often arises that they are only collected during treatment and that the corresponding observations are censored if treatment is discontinued or switched. This often results in different mean observation periods in the 2 groups to be compared. Although for the analysis of adverse events, the consideration of observation periods by using adequate survival time methods has long been demanded in drug approval [561,562], simple methods based on relative frequencies or incidence densities still dominate in this area [54,752].

In benefit assessments, however, analyses are required for all patient-relevant outcomes that allow conclusions to be drawn on the basis of adequate statistical inference. For group comparisons with different mean observation periods, this means in particular that adequate survival time methods are needed [53,54].

If the reason for the different mean observation periods is incomplete data collection (censoring when treatment is discontinued or switched), the problem of informative censoring also arises when using survival time methods, which may result in a high risk of bias. To avoid such problems, data on all outcomes relevant to the assessment, including PROs and adverse events, should be collected completely, even after discontinuation or switching of treatment [54].

The frequently applied practice of recording adverse events only up to a maximum of 30 days after discontinuation or switching of treatment is insufficient for benefit assessments, since analyses based on the treatment policy estimand (i.e. an estimate of the effect for the entire treatment strategy independent of discontinuation or switching of treatment) and the intention-to-treat principle are required [54,801].

Survival time analyses (e.g. responder analyses for the time to deterioration) are also sometimes performed for PROs with different observation periods because of incomplete data collection (e.g. due to discontinuation of observation when the disease progresses). However, these analyses lack the necessary information on PROs after treatment discontinuation or switching. In such a situation, the interpretability of PROs is severely limited due to the systematically shortened data collection period. So-called permanent changes (improvements or deteriorations) cannot be meaningfully interpreted if the observation periods differ considerably between the treatment arms. Generally, "permanent" changes arise from analysis strategies in which, once the threshold for a response has been exceeded, the threshold must be exceeded in all subsequent recordings until observation is prematurely discontinued (censoring in the case of treatment discontinuation or switching). The "permanent" change is then potentially more difficult to achieve in the longer observed treatment arm due to the need to continuously exceed or fall below the threshold (response criterion). Such analyses are therefore generally not used for the benefit assessment if the observation period differs considerably between the treatment arms. If there are no relevant differences in observation periods between the study arms for an outcome, the analysis of the time until a confirmed or permanent change is generally meaningful. However, it is inappropriate to speak of a "permanent" change if the observation period covers only a shortened (and possibly very small) part of the total study duration; rather, in this case it is appropriate to describe the change as a "multiple confirmed" change.

Overall, if the observation period varies between the groups to be compared, data analyses that do not adequately take this into account cannot be used for the benefit assessment. Furthermore, in the case of systematically shortened observation periods, the results of affected outcomes (e.g. adverse events or PROs) generally cannot dominate the overall conclusion on the benefit or added benefit of a medical intervention.

9.3.13 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 [144]. The following text describes only the most important types; a detailed overview of various types of bias in different situations is presented by Feinstein [261].

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 [363], 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 multi-factorial 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 [429], 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 [429,464].

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 [496]. 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 [243].

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 [274].

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 reporting bias or outcome reporting bias is present [143,213,303,304,571]. 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 [84]. The pooling of published results can therefore result in a systematic bias of the common effect estimate. Graphic methods such as the funnel plot [225] 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.

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) [483].

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 [495]. 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 [780]. 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 [780]. 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 [483]. 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 [483].

A general problem in the estimation of effects is bias caused by measurement errors in the study data collected [137,145]. 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.

9.3.14 Analysis of dependent data

Simple standard procedures of medical statistics assume independent experimental units within a treatment arm. In certain situations (e.g. in cluster-randomized studies or studies of eyes or teeth) in which several experimental units of a cluster or a patient within a treatment arm are included in the analysis, this independency does not apply. The application of simple standard procedures for independent experimental units within a treatment arm then leads to an underestimation of variance and possibly to incorrect conclusions on significance, as the p-values calculated are too small [22]. Evidence from studies that used simple standard procedures despite correlated data can potentially still be considered, if the related error can be properly classified. If this is not the case, then results from studies with correlated data cannot be interpreted adequately. In the event of dependent data within a treatment arm, then the related correlation structure must be considered by applying suitable methods for dependent data. In practice, the Generalized Estimating Equations [124] as well as methods from the class of mixed models [108] are commonly used methods.

Dependencies can also arise between the treatment arms to be compared (e.g. due to matching or in cross-over studies), which must be accounted for in the data analysis by applying appropriate statistical methods for paired samples or cross-over studies [426,671].

9.4 Qualitative methods

Qualitative research methods are applied to explore and understand subjective experiences, individual actions, and the social world [198,336,509,543]. They can, among other things, enable access to opinions and experiences of patients and their relatives 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 [200].

9.4.1 Qualitative research in the production of health information

In the development of health information, research findings from qualitative primary studies and reviews of qualitative studies are used to identify (potential) information needs, as well as to investigate patients' experiences with a certain disease or intervention and to investigate how they cope with a disease. In particular, the following questions should be answered:

- What questions do people have with regard to this disease/this topic?
- How do people perceive this disease/this intervention?
- When do people visit a doctor and how do they perceive their symptoms?
- What experiences, problems, challenges and questions do people have with regard to diagnostic procedures and the diagnosis?
- What experiences have people made with the treatment or the treatment decision?
- What experiences, problems, challenges and questions do people have with regard to coping with daily life?
- What information do people need?
- How can information support people (e.g. by the processing and format)?

For this purpose, focused information retrieval in bibliographic databases is conducted (see Section 8.2.2).

The study quality is assessed by means of criteria defined beforehand. In recent years various instruments for evaluating the quality of qualitative studies have been developed [168]. The assessment of qualitative studies aims to determine whether the study design, study quality, and reliability are appropriate for the research question investigated. No general consensus exists yet with regard to the criteria for the conduct, assessment, and synthesis of qualitative

studies when compared with other research areas [198,201,336,543]. The Institute monitors the methodological developments and currently uses the CERQual instrument.

The quality of the qualitative studies identified is assessed by means of the following aspects, which are based on the Critical Appraisal Skills Programme (CASP) checklist [161]:

- Are the research question and/or the aims of the study described?
- Is the sampling strategy described?
- Is the sample described and suitable for the research question?
- Are the methods for data collection described and are they suitable for the topic?
- Are the methods for data analysis described?
- Were at least 2 researchers involved in the data analysis?
- Is the process of data analysis described transparently and comprehensibly?
- Are the study results presented clearly?

After extraction of the studies included, an overarching analysis and a descriptive summary of the results are performed. Furthermore, potential information needs are derived from the results.

9.4.2 Qualitative studies in the production of other IQWiG products

Different sources of information can support the production of systematic reviews [199,491,735]. One possible source are research results from qualitative studies [336,491,545,735]. Qualitative research can, among other things, provide information on the acceptance, suitability and implementation of interventions in clinical practice [32,198,489,542]. The results of qualitative research can be helpful in the interpretation of a systematic review [735] and may be used in the context of primary studies or systematic reviews in order to determine patient-relevant outcomes [198,200,463,543,545].

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. In addition, in the ThemenCheck reports (see Sections 6.5.3 to 6.5.6), results from qualitative surveys and analyses (individual or focus group interviews) performed by external experts commissioned by IQWiG, as well as results from qualitative studies and reviews, form the basis for processing the domains of ethics and social and organizational issues.

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[60]).

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 11. 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 effects) 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.

[60] Arzneimittel-Nutzenbewertungsverordnung, AM-NutzenV

Table 11: 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 effects
	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 effects Important avoidance of other adverse effects
	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 effects
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 12. 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 13):

1) all-cause mortality
2) • serious (or severe) symptoms (or late complications)
   • serious or (severe) adverse effects
   • health-related quality of life
3) • non-serious (or non-severe) symptoms (or late complications)
   • non-serious (or non-severe) adverse effects

Health-related quality of life is regarded to be of equal importance as serious (or severe symptoms), late complications and adverse effects. 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 12: 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 effects***
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 effects
	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 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 effects Important avoidance of other *(non-serious or non-severe)* adverse effects
	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 effects* Relevant avoidance of *(other, non-serious or non-severe)* adverse effects

a. Amendments to the ANV in *italics*.

ANV: Arzneimittel-Nutzenbewertungsverordnung (Regulation for Early Benefit Assessment of New Pharmaceuticals)

Table 13: 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 effects**	***Health-related quality of life***	**Non-serious *(or non-severe)* symptoms *(or late complications)* and adverse effects**
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 [692]. This great susceptibility to external conditions calls into question the transferability of absolute effect measures from clinical studies to the daily health care setting. It is therefore common practice preferably to express effects shown in clinical studies as relative risks, odds ratios or hazard (or incidence) ratios [175].
- 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 greater than 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 13. The greater the relevance ascribed to the outcome, the closer the thresholds should lie to 1 (below 1). This takes into account the requirement from the ANV 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 [393]. 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. [202] as a requirement for a “breakthrough” – was defined as an effect of a major extent for the outcome “all-cause mortality” [393].

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 [393], but are also addressed again at the end of this appendix. Following these considerations, the

simultaneous requirements 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 effects**
- **health-related quality of life**

For serious (or severe) symptoms (or late complications) and serious (or severe) adverse effects, 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 (the 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 effects

The specification of thresholds for the non-serious (or non-severe) symptoms (or late complications) and the non-serious (or non-severe) adverse effects 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 of no added benefit.

The corresponding thresholds for all extent categories and outcome categories are presented in the following Table 14.

Table 14: Inferential statistical thresholds (hypotheses boundaries) for relative effect measures

		Outcome category		
		All-cause mortality	**Serious (or severe) symptoms (or late complications) and adverse effects, as well as quality of life[a]**	**Non-serious (or non-severe) symptoms (or late complications) and adverse effects**
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 the

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 [393] 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 [646] refers to the work by Fleiss et al. [270]. A query to the former Speaker of the Working Group "Pharmaceutical Research" of the German Region of the International Biometric Society as well as 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 a step size of 0.01. The risk in the control group p_C runs through all values between 0.05 and 0.95 at a step size 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 [258] and then doubled ($m := 2n$).
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 and 0.95.

Figure 17 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).

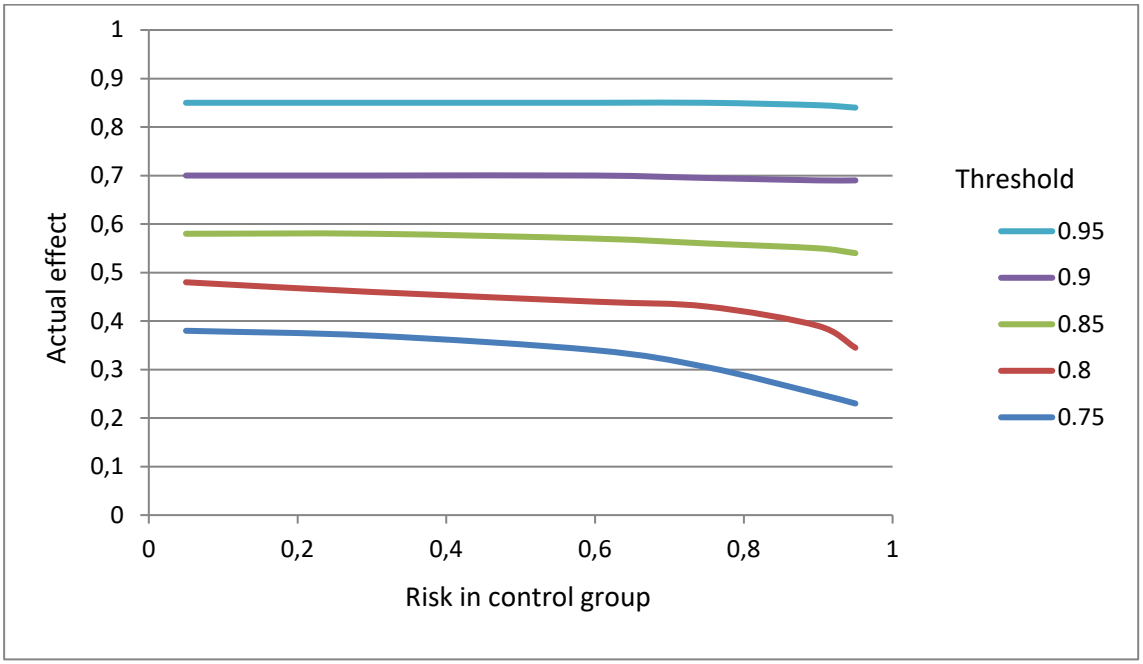


Figure 17: Actual effects depending on the baseline risk

Table 15 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 15: Actual effects for the relative risk

		Outcome category		
		All-cause mortality	**Serious (or severe) symptoms (or late complications) and adverse effects as well as quality of life**	**Non-serious (or non-severe) symptoms (or late complications) and adverse effects**
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 [393] 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 14 seems reasonable and practicable.

References

1. Memorandum for the evaluation of diagnostic measures. J Clin Chem Clin Biochem 1990; 28(12): 873-879.

2. Gesetz zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG). Bundesgesetzblatt Teil 1 2003; (55): 2190-2258.

3. Food and Drug Administration Amendments Act of 2007: Public Law 110–85 [online]. 2007 [Accessed: 01.02.2021]. URL: http://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf.

4. Gesetz zur Neuordnung des Arzneimittelmarktes in der gesetzlichen Krankenversicherung (Arzneimittelmarktneuordnungsgesetz – AMNOG) vom 22. Dezember 2010. Bundesgesetzblatt Teil 1 2010; (67): 2262-2277.

5. Gesetz für eine bessere Versorgung durch Digitalisierung und Innovation (Digitale-Versorgung-Gesetz – DVG). Bundesgesetzblatt Teil 1 2019; (49): 2562-2584.

6. Gesetz für mehr Sicherheit in der Arzneimittelversorgung (GSAV). Bundesgesetzblatt Teil 1 2019; (30): 1202-1220.

7. SGB V Handbuch; Sozialgesetzbuch V; Krankenversicherung. Altötting: KKF-Verlag; 2020.

8. Sozialgesetzbuch (SGB) Fünftes Buch (V) - Gesetzliche Krankenversicherung - (Artikel 1 des Gesetzes v. 20. Dezember 1988, BGBl. I S. 2477) [online]. 2021 [Accessed: 05.08.2021]. URL: https://www.gesetze-im-internet.de/sgb_5/SGB_5.pdf.

9. Aagaard T, Lund H, Juhl C. Optimizing literature search in systematic reviews - are MEDLINE, EMBASE and CENTRAL enough for identifying effect studies within the area of musculoskeletal disorders? BMC Med Res Methodol 2016; 16(1): 161. https://dx.doi.org/10.1186/s12874-016-0264-6.

10. Ades AE, Caldwell DM, Reken S et al. Evidence synthesis for decision making 7: a reviewer's checklist. Med Decis Making 2013; 33(5): 679-691. https://dx.doi.org/10.1177/0272989X13485156.

11. Agency for Health Technology Assessment and Tariff System. Assessment of reimbursement applications [online]. [Accessed: 28.07.2022]. URL: https://www.aotm.gov.pl/en/medicines/assessment-of-reimbursement-applications/.

12. Agoritsas T, Merglen A, Shah ND et al. Adjusted Analyses in Studies Addressing Therapy and Harm: Users' Guides to the Medical Literature. JAMA 2017; 317(7): 748-759. https://dx.doi.org/10.1001/jama.2016.20029.

13. AGREE Collaboration. Appraisal of guidelines for research & evaluation: AGREE instrument. London: St. George's Hospital Medical School; 2001.

14. AGREE Next Steps Consortium. Appraisal of Guidelines for REsearch & Evaluation II, AGREE II instrument [online]. 2017 [Accessed: 04.02.2021]. URL: https://www.agreetrust.org/wp-content/uploads/2017/12/AGREE-II-Users-Manual-and-23-item-Instrument-2009-Update-2017.pdf.

15. Agresti A. Modelling ordered categorical data: recent advances and future challenges. Stat Med 1999; 18(17-18): 2191-2207. https://dx.doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18<2191::aid-sim249>3.0.co;2-m.

16. Agresti A. Categorical Data Analysis. Hoboken: Wiley; 2002.

17. 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. https://dx.doi.org/10.1191/0962280203sm311ra.

18. Al-Marzouki S, Evans S, Marshall T et al. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005; 331(7511): 267-270. https://dx.doi.org/10.1136/bmj.331.7511.267.

19. Alma H, de Jong C, Jelusic D et al. Baseline health status and setting impacted minimal clinically important differences in COPD: an exploratory study. J Clin Epidemiol 2019; 116: 49-61. https://dx.doi.org/10.1016/j.jclinepi.2019.07.015.

20. Alma H, de Jong C, Tsiligianni I et al. Clinically relevant differences in COPD health status: systematic review and triangulation. Eur Respir J 2018; 52(3): 1800412. https://dx.doi.org/10.1183/13993003.00412-2018.

21. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311(7003): 485. https://dx.doi.org/10.1136/bmj.311.7003.485

22. Altman DG, Bland JM. Statistics notes. Units of analysis. BMJ 1997; 314(7098): 1874. https://dx.doi.org/10.1136/bmj.314.7098.1874.

23. Altman DG, Machin D, Bryant TN et al. Statistics with confidence; confidence intervals and statistical guidelines. London: BMJ Books; 2005.

24. American Society of Clinical Oncology. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol 1996; 14(2): 671-679. https://dx.doi.org/10.1200/JCO.1996.14.2.671.

25. Ara R, Rowen D, Mukuria C. The Use of Mapping to Estimate Health State Utility Values. Pharmacoeconomics 2017; 35(Suppl 1): 57-66. https://dx.doi.org/10.1007/s40273-017-0548-7.

26. Arbeitsgruppe Erhebung und Nutzung von Sekundärdaten, Deutsche Gesellschaft für Sozialmedizin und Prävention, Deutsche Gesellschaft für Epidemiologie. Gute Praxis Sekundärdatenanalyse (GPS); Leitlinien und Empfehlungen [online]. 2014 [Accessed: 04.02.2021]. URL: https://www.dgepi.de/assets/Leitlinien-und-Empfehlungen/GPS_revision2-final_august2014.pdf.

27. Arber M, Garcia S, Veale T et al. Performance of Ovid Medline Search Filters to Identify Health State Utility Studies. Int J Technol Assess Health Care 2017; 33(4): 472-480. https://dx.doi.org/10.1017/S0266462317000897.

28. Assasi N, Schwartz L, Tarride JE et al. Methodological guidance documents for evaluation of ethical considerations in health technology assessment: a systematic review. Expert Rev Pharmacoecon Outcomes Res 2014; 14(2): 203-220. https://dx.doi.org/10.1586/14737167.2014.894464.

29. Assmann SF, Pocock SJ, Enos LE et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355(9209): 1064-1069. https://dx.doi.org/10.1016/S0140-6736(00)02039-0.

30. Atkins D, Best D, Briss PA et al. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454): 1490. https://dx.doi.org/10.1136/bmj.328.7454.1490.

31. Atkins D, Eccles M, Flottorp S et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004; 4(1): 38. https://dx.doi.org/10.1186/1472-6963-4-38.

32. Atkins S, Lewin S, Smith H et al. Conducting a meta-ethnography of qualitative literature: lessons learnt. BMC Med Res Methodol 2008; 8: 21. https://dx.doi.org/10.1186/1471-2288-8-21.

33. Attema AE, Brouwer WBF, Claxton K. Discounting in Economic Evaluations. Pharmacoeconomics 2018; 36(7): 745-758. https://dx.doi.org/10.1007/s40273-018-0672-z.

34. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011; 46(3): 399-424. https://dx.doi.org/10.1080/00273171.2011.568786.

35. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. Stat Med 2017; 36(8): 1203-1209. https://dx.doi.org/10.1002/sim.7215.

36. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Stat Med 2017; 36(27): 4391-4400. https://dx.doi.org/10.1002/sim.7501.

37. Baethge C. In der Versorgung angekommen, aber noch nicht heimisch. Z Evid Fortbild Qual Gesundhwes 2014; 108(10): 594-596. https://dx.doi.org/10.1016/j.zefq.2014.10.016.

38. Bafeta A, Trinquart L, Seror R et al. Reporting of results from network meta-analyses: methodological systematic review. BMJ 2014; 348: g1741. https://dx.doi.org/10.1136/bmj.g1741.

39. 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. https://dx.doi.org/10.1016/S0895-4356(01)00372-9.

40. Bai O, Chen M, Wang X. Bayesian Estimation and Testing in Random Effects Meta-analysis of Rare Binary Adverse Events. Stat Biopharm Res 2016; 8(1): 49-59. https://dx.doi.org/10.1080/19466315.2015.1096823.

41. Baio G. Bayesian Methods in Health Economics. Boca Raton: CRC Press; 2013.

42. Baker SG. Surrogate endpoints: wishful thinking or reality? J Natl Cancer Inst 2006; 98(8): 502-503. https://dx.doi.org/10.1093/jnci/djj153.

43. Balshem H, Stevens A, Ansari M et al. Finding Grey Literature Evidence and Assessing for Outcome and Analysis Reporting Biases When Comparing Medical Interventions: AHRQ and the Effective Health Care Program [online]. 2013 [Accessed: 04.02.2021]. URL: https://www.ncbi.nlm.nih.gov/books/NBK174882/pdf/Bookshelf_NBK174882.pdf.

44. Banta D. The development of health technology assessment. Health Policy 2003; 63(2): 121-132. https://dx.doi.org/10.1016/s0168-8510(02)00059-3.

45. 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. https://dx.doi.org/10.1001/archinte.119.6.547.

46. Basu A, Maciejewski ML. Choosing a Time Horizon in Cost and Cost-effectiveness Analyses. JAMA 2019; 321(11): 1096-1097. https://dx.doi.org/10.1001/jama.2019.1153.

47. Beauchamp TL. Methods and principles in biomedical ethics. J Med Ethics 2003; 29(5): 269-274. https://dx.doi.org/10.1136/jme.29.5.269.

48. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. New York: Oxford University Press; 2013.

49. Becker C, Leidl R, Stollenwerk B. Entscheidungsanalytische Modellierung in der gesundheitsökonomischen Evaluation. Gesundheitsökonomie & Qualitätsmanagement 2010; 15(5): 260-264. https://dx.doi.org/10.1055/s-0029-1245691.

50. Bell ML, Kenward MG, Fairclough DL et al. Differential dropout and bias in randomised controlled trials: when it matters and when it may not. BMJ 2013; 346: e8668. https://dx.doi.org/10.1136/bmj.e8668.

51. Benchimol EI, Smeeth L, Guttmann A et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015; 12(10): e1001885. https://dx.doi.org/10.1371/journal.pmed.1001885.

52. Bender R. Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung. Med Klin (Munich) 2001; 96(2): 116-121. https://dx.doi.org/10.1007/pl00002179.

53. Bender R, Beckmann L. Limitations of the incidence density ratio as approximation of the hazard ratio. Trials 2019; 20(1): 485. https://dx.doi.org/10.1186/s13063-019-3590-2.

54. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. Pharm Stat 2016; 15(4): 292-296. https://dx.doi.org/10.1002/pst.1740.

55. Bender R, Bunce C, Clarke M et al. Attention should be given to multiplicity issues in systematic reviews. J Clin Epidemiol 2008; 61(9): 857-865. https://dx.doi.org/10.1016/j.jclinepi.2008.03.004.

56. Bender R, Friede T, Koch A et al. Methods for evidence synthesis in the case of very few studies. Res Synth Methods 2018; 9(3): 382-392. https://dx.doi.org/10.1002/jrsm.1297.

57. Bender R, Grouven U. Logistic regression models used in medical research are poorly presented. BMJ 1996; 313: 628. https://dx.doi.org/10.1136/bmj.313.7057.628

58. Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physicians Lond 1997; 31(5): 546-551.

59. Bender R, Grouven U. Möglichkeiten und Grenzen statistischer Regressionsmodelle zur Berechnung von Schwellenwerten für Mindestmengen. Z Arztl Fortbild Qualitatssich 2006; 100(2): 93-98.

60. Bender R, Lange S. Adjusting for multiple testing – when and how? J Clin Epidemiol 2001; 54(4): 343-349. https://dx.doi.org/10.1016/s0895-4356(00)00314-0.

61. Bender R, Schwenke C, Schmoor C et al. Stellenwert von Ergebnissen aus indirekten Vergleichen: gemeinsame Stellungnahme von IQWiG, GMDS und IBS-DR [online]. 2012 [Accessed: 04.02.2021]. URL: https://www.iqwig.de/printprodukte/12-03-07_gemeinsame_stellungnahme_iqwig_gmds_ibs-dr_zum_indirekten_vergleich.pdf.

62. Bennett K, Duda S, Brouwers M et al. Towards high-quality, useful practice guidelines for child and youth mental health disorders: protocol for a systematic review and consensus exercise. BMJ Open 2018; 8(2): e018053. https://dx.doi.org/10.1136/bmjopen-2017-018053.

63. 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. https://dx.doi.org/10.7326/0003-4819-144-4-200602210-00007.

64. Beresniak A, Dupont D. Is there an alternative to quality-adjusted life years for supporting healthcare decision making? Expert Rev Pharmacoecon Outcomes Res 2016; 16(3): 351-357. https://dx.doi.org/10.1080/14737167.2016.1184975.

65. Beynon R, Leeflang MM, McDonald S et al. Search strategies to identify diagnostic accuracy studies in MEDLINE and EMBASE. Cochrane Database Syst Rev 2013; (9): MR000022. https://dx.doi.org/10.1002/14651858.MR000022.pub3.

66. Biesheuvel CJ, Grobbee DE, Moons KG. Distraction from randomization in diagnostic research. Ann Epidemiol 2006; 16(7): 540-544. https://dx.doi.org/10.1016/j.annepidem.2005.10.004.

67. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. Prev Sci 2000; 1(1): 31-49. https://dx.doi.org/10.1023/a:1010024016308.

68. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69(3): 89-95. https://dx.doi.org/10.1067/mcp.2001.113989.

69. 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. S. 515-538.

70. Bock JO, Brettschneider C, Seidl H et al. Standardisierte Bewertungssätze aus gesellschaftlicher Perspektive für die gesundheitsökonomische Evaluation. Baden-Baden: Nomos; 2015.

71. Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. Biom J 2015; 57(4): 633-648. https://dx.doi.org/10.1002/bimj.201400184.

72. Bonell C, Oakley A, Hargreaves J et al. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ 2006; 333(7563): 346-349. https://dx.doi.org/10.1136/bmj.333.7563.346.

73. Bonhoeffer J, Zumbrunn B, Heininger U. Reporting of vaccine safety data in publications: systematic review. Pharmacoepidemiol Drug Saf 2005; 14(2): 101-106. https://dx.doi.org/10.1002/pds.979.

74. Borah R, Brown AW, Capers PL et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017; 7(2): e012545. https://dx.doi.org/10.1136/bmjopen-2016-012545.

75. Borenstein M. In a meta-analysis, the I-squared statistic does not tell us how much the effect size varies. J Clin Epidemiol 2022; 152: 281-284. https://dx.doi.org/10.1016/j.jclinepi.2022.10.003.

76. Borm GF, Donders AR. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol 2009; 62(8): 825-830.e10. https://dx.doi.org/10.1016/j.jclinepi.2008.08.010.

77. Bossuyt PM, Irwig L, Craig J et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006; 332(7549): 1089-1092. https://dx.doi.org/10.1136/bmj.332.7549.1089.

78. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000; 356(9244): 1844-1847. https://dx.doi.org/10.1016/S0140-6736(00)03246-3.

79. Bossuyt PM, Reitsma JB, Bruns DE et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015; 351: h5527. https://dx.doi.org/10.1136/bmj.h5527.

80. Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138(1): 40-44. https://dx.doi.org/10.7326/0003-4819-138-1-200301070-00010.

81. Bossuyt PM, Reitsma JB, Bruns DE et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003; 138(1): W1-12. https://dx.doi.org/10.7326/0003-4819-138-1-200301070-00012-w1.

82. Boutron I, Moher D, Altman DG et al. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008; 148(4): 295-309. https://dx.doi.org/10.7326/0003-4819-148-4-200802190-00008.

83. Boutron I, Moher D, Altman DG et al. Methods and processes of the CONSORT Group: example of an extension for trials assessing nonpharmacologic treatments. Ann Intern Med 2008; 148(4): W60-66. https://dx.doi.org/10.7326/0003-4819-148-4-200802190-00008-w1.

84. Boutron I, Page MJ, Higgins JPT et al. Considering bias and conflicts of interest among the included studies. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 177-204.

85. Bradburn MJ, Deeks JJ, Berlin JA et al. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Stat Med 2007; 26(1): 53-77. https://dx.doi.org/10.1002/sim.2528.

86. Braun S, Prenzler A, Mittendorf T et al. Bewertung von Ressourcenverbräuchen im deutschen Gesundheitswesen aus Sicht der Gesetzlichen Krankenversicherung. Gesundheitswesen 2009; 71(1): 19-23. https://dx.doi.org/10.1055/s-0028-1102930.

87. Brazier J. Measuring and valuing mental health for use in economic evaluation. J Health Serv Res Policy 2008; 13 Suppl 3: 70-75. https://dx.doi.org/10.1258/jhsrp.2008.008015.

88. Brazier J, Ara R, Azzabi I et al. Identification, Review, and Use of Health State Utilities in Cost-Effectiveness Models: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health 2019; 22(3): 267-275. https://dx.doi.org/10.1016/j.jval.2019.01.004.

89. Brazier J, Green C, McCabe C et al. Use of visual analog scales in economic evaluation. Expert Rev Pharmacoecon Outcomes Res 2003; 3(3): 293-302. https://dx.doi.org/10.1586/14737167.3.3.293.

90. Brazier J, Ratcliffe J, Saloman J et al. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2016.

91. Brazier J, Rowen D, Karimi M et al. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. Eur J Health Econ 2018; 19(6): 881-891. https://dx.doi.org/10.1007/s10198-017-0931-5.

92. Breslow NE, Day NE. Statistical Methods in Cancer Research; Volume 1 - The analysis of case-control studies [online]. 1980 [Accessed: 17.02.2021]. URL: https://publications.iarc.fr/_publications/media/download/3468/200a06ebe55fa3d17ed551c5fcdb83f5cf0f3752.pdf.

93. Breslow NE, Day NE. Statistical Methods in Cancer Research; Volume II - The Design and Analysis of Cohort Studies [online]. 1987 [Accessed: 17.02.2021]. URL: https://publications.iarc.fr/_publications/media/download/3494/fb469ed43c52f0c738915cca6a0f31544b9ed7b6.pdf.

94. Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Oxford: Oxford University Press; 2006.

95. Briggs AH, Weinstein MC, Fenwick EA et al. 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. https://dx.doi.org/10.1177/0272989X12458348.

96. Briggs AH, Weinstein MC, Fenwick EA et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. Value Health 2012; 15(6): 835-842. https://dx.doi.org/10.1016/j.jval.2012.04.014.

97. Brockhaus AC, Bender R, Skipka G. The Peto odds ratio viewed as a new effect measure. Stat Med 2014; 33(28): 4861-4874. https://dx.doi.org/10.1002/sim.6301.

98. Brockhaus AC, Grouven U, Bender R. Performance of the Peto odds ratio compared to the usual odds ratio estimator in the case of rare events. Biom J 2016; 58(6): 1428-1444. https://dx.doi.org/10.1002/bimj.201600034.

99. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001; 20(6): 825-840. https://dx.doi.org/10.1002/sim.650.

100. Brok J, Thorlund K, Gluud C et al. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analyses. J Clin Epidemiol 2008; 61(8): 763-769. https://dx.doi.org/10.1016/j.jclinepi.2007.10.007.

101. Brok J, Thorlund K, Wetterslev J et al. 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. https://dx.doi.org/10.1093/ije/dyn188.

102. Brookes ST, Whitely E, Egger M et al. 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. https://dx.doi.org/10.1016/j.jclinepi.2003.08.009.

103. Brouwer WB, Koopmanschap MA. The friction-cost method : replacement for nothing and leisure for free? Pharmacoeconomics 2005; 23(2): 105-111. https://dx.doi.org/10.2165/00019053-200523020-00002.

104. 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. https://dx.doi.org/10.1002/(sici)1099-1050(199709)6:5<511::Aid-hec297>3.0.Co;2-k.

105. Brouwers MC, Kho ME, Browman GP et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ 2010; 182(18): E839-842. https://dx.doi.org/10.1503/cmaj.090449.

106. Brouwers MC, Kho ME, Browman GP et al. Development of the AGREE II, part 1: performance, usefulness and areas for improvement. CMAJ 2010; 182(10): 1045-1052. https://dx.doi.org/10.1503/cmaj.091714.

107. Brouwers MC, Kho ME, Browman GP et al. Development of the AGREE II, part 2: assessment of validity of items and tools to support application. CMAJ 2010; 182(10): E472-478. https://dx.doi.org/10.1503/cmaj.091716.

108. Brown H, Prescott R. Applied Mixed Models in Medicine. Chichester: Wiley; 2006.

109. Brozek JL, Akl EA, Alonso-Coello P et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. Allergy 2009; 64(5): 669-677. https://dx.doi.org/10.1111/j.1398-9995.2009.01973.x.

110. Brozek JL, Akl EA, Compalati E et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines part 3 of 3. The GRADE approach to developing recommendations. Allergy 2011; 66(5): 588-595. https://dx.doi.org/10.1111/j.1398-9995.2010.02530.x.

111. Brozek JL, Akl EA, Jaeschke R et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009; 64(8): 1109-1116. https://dx.doi.org/10.1111/j.1398-9995.2009.02083.x.

112. Brozek JL, Canelo-Aybar C, Akl EA et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. J Clin Epidemiol 2021; 129: 138-150. https://dx.doi.org/10.1016/j.jclinepi.2020.09.018.

113. Brundage M, Blazeby J, Revicki D et al. Patient-reported outcomes in randomized clinical trials: development of ISOQOL reporting standards. Qual Life Res 2013; 22(6): 1161-1175. https://dx.doi.org/10.1007/s11136-012-0252-1.

114. Bucher HC, Guyatt GH, Griffith LE et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997; 50(6): 683-691. https://dx.doi.org/10.1016/s0895-4356(97)00049-8.

115. Büchter RB, Fechtelpeter D, Knelangen M et al. Words or numbers? Communicating risk of adverse effects in written consumer health information: a systematic review and meta-analysis. BMC Med Inform Decis Mak 2014; 14: 76. https://dx.doi.org/10.1186/1472-6947-14-76.

116. Bundesärztekammer, Kassenärztliche Bundesvereinigung, Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. Programm für Nationale VersorgungsLeitlinien; Methodenreport [online]. 2017 [Accessed: 05.08.2021]. URL: https://www.leitlinien.de/methodik/pdf/nvl-methodenreport-5aufl-vers1.pdf.

117. Bundesinstitut für Arzneimittel und Medizinprodukte. PharmNet.Bund-Arzneimittel-Informationssystem [online]. 2020 [Accessed: 04.02.2021]. URL: https://www.pharmnet-bund.de/dynamic/de/arzneimittel-informationssystem/index.html.

118. Bundesministerium der Justiz. Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz (Barrierefreie-Informationstechnik-Verordnung - BITV 2.0) [online]. 2019 [Accessed: 04.02.2021]. URL: http://www.gesetze-im-internet.de/bitv_2_0/BJNR184300011.html.

119. 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]. 2019 [Accessed: 04.02.2021]. URL: http://www.gesetze-im-internet.de/am-nutzenv/AM-NutzenV.pdf.

120. Bundesministerium für Gesundheit. Verordnung über die Verfahrensgrundsätze der Bewertung von Untersuchungs- und Behandlungsmethoden in der vertragsärztlichen Versorgung und im Krankenhaus (Methodenbewertungsverfahrensverordnung - MBVerfV) [online]. 2020 [Accessed: 04.02.2021]. URL: http://www.gesetze-im-internet.de/mbverfv/MBVerfV.pdf.

121. Bundesministerium für Gesundheit, Bundesministerium für Bildung und Forschung. Verordnung über die Voraussetzungen für die Bewertung neuer Untersuchungs- und Behandlungsmethoden mit Medizinprodukten hoher Risikoklasse nach § 137h des Fünften Buches Sozialgesetzbuch (Medizinproduktemethodenbewertungsverordnung - MeMBV) [online]. 2021 [Accessed: 29.06.2021]. URL: http://www.gesetze-im-internet.de/membv/MeMBV.pdf.

122. Bundessozialgericht. Urteil - 06.05.2009 - B 6 A 1/08 R [online]. 2009 [Accessed: 14.08.2023]. URL: https://www.sozialgerichtsbarkeit.de/node/87597/pdf.

123. Bundesverfassungsgericht. Leitsatz zum Beschluss des Ersten Senats vom 6. Dezember 2005 - 1 BvR 347/98 - [online]. 2005 [Accessed: 05.08.2021]. URL: https://www.bundesverfassungsgericht.de/SharedDocs/Downloads/DE/2005/12/rs20051206_1bvr034798.pdf?__blob=publicationFile&v=1.

124. Burton P, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med 1998; 17(11): 1261-1291. https://dx.doi.org/10.1002/(sici)1097-0258(19980615)17:11<1261::aid-sim846>3.0.co;2-z.

125. Burzykowski T, Buyse M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. Pharm Stat 2006; 5(3): 173-186. https://dx.doi.org/10.1002/pst.207.

126. Burzykowski T, Molenberghs G, Buyse M. The Evaluation of Surrogate Endpoints. New York: Springer; 2005.

127. Buyse M, Molenberghs G, Burzykowski T et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000; 1(1): 49-67. https://dx.doi.org/10.1093/biostatistics/1.1.49.

128. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005; 331(7521): 897-900. https://dx.doi.org/10.1136/bmj.331.7521.897.

129. Calvert M, Blazeby J, Altman DG et al. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 2013; 309(8): 814-822. https://dx.doi.org/10.1001/jama.2013.879.

130. Campbell MK, Piaggio G, Elbourne DR et al. Consort 2010 statement: extension to cluster randomised trials. BMJ 2012; 345: e5661. https://dx.doi.org/10.1136/bmj.e5661.

131. Canadian Agency for Drugs and Technologies in Health. Procedures for CADTH Reimbursement Reviews [online]. 2022 [Accessed: 22.09.2022]. URL: https://www.cadth.ca/sites/default/files/Drug_Review_Process/CADTH_Drug_Reimbursement_Review_Procedures.pdf.

132. Canestaro WJ, Hendrix N, Bansal A et al. Favorable and publicly funded studies are more likely to be published: a systematic review and meta-analysis. J Clin Epidemiol 2017; 92: 58-68. https://dx.doi.org/10.1016/j.jclinepi.2017.08.004.

133. Carlson JJ, Brouwer ED, Kim E et al. Alternative Approaches to Quality-Adjusted Life-Year Estimation Within Standard Cost-Effectiveness Models: Literature Review, Feasibility Assessment, and Impact Evaluation. Value Health 2020; 23(12): 1523-1533. https://dx.doi.org/10.1016/j.jval.2020.08.2092.

134. Caro JJ, Briggs AH, Siebert U et al. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1. Value Health 2012; 15(6): 796-803. https://dx.doi.org/10.1016/j.jval.2012.06.012.

135. Caro JJ, Ishak KJ. No head-to-head trial? simulate the missing arms. Pharmacoeconomics 2010; 28(10): 957-967. https://dx.doi.org/10.2165/11537420-000000000-00000.

136. Carrasco-Labra A, Devji T, Qasim A et al. Interpretation of patient-reported outcome measures: an inventory of over 3000 minimally important difference estimates and an assessment of their credibility. Cochrane Database Syst Rev 2018; (9 Suppl 1): 135-136. https://dx.doi.org/10.1002/14651858.CD201801.

137. Carroll RJ, Ruppert D, Stefanski LA et al. Measurement error in nonlinear models; a modern perspective. London: Chapman & Hall; 2006.

138. Carter JL, Coletti RJ, Harris RP. Quantifying and monitoring overdiagnosis in cancer screening: a systematic review of methods. BMJ 2015; 350: g7773. https://dx.doi.org/10.1136/bmj.g7773.

139. Carter SM, Rogers W, Heath I et al. The challenge of overdiagnosis begins with its definition. BMJ 2015; 350: h869. https://dx.doi.org/10.1136/bmj.h869.

140. Celik D, Coban Ö, Kilicoglu Ö. Minimal clinically important difference of commonly used hip-, knee-, foot-, and ankle-specific questionnaires: a systematic review. J Clin Epidemiol 2019; 113: 44-57. https://dx.doi.org/10.1016/j.jclinepi.2019.04.017.

141. Centre for Reviews and Dissemination. PROSPERO; International prospective register of systematic reviews [online]. [Accessed: 04.02.2021]. URL: https://www.crd.york.ac.uk/PROSPERO/.

142. Centre for Reviews and Dissemination. Systematic reviews; CRD's guidance for undertaking reviews in health care [online]. 2009 [Accessed: 09.02.2021]. URL: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf.

143. Chan AW, Hrobjartsson A, Haahr MT et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004; 291(20): 2457-2465. https://dx.doi.org/10.1001/jama.291.20.2457.

144. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol 2010; 63(11): 1205-1215. https://dx.doi.org/10.1016/j.jclinepi.2009.12.011.

145. Cheng CL, Van Ness JW. Statistical regression with measurement error. London: Arnold; 1999.

146. Chi C. Shall we search all trial registers? A comparative study of the sensitivity of five trial registers used by the Cochrane Skin Group [online]. 2012 [Accessed: 04.02.2021]. URL: https://abstracts.cochrane.org/2012-auckland/shall-we-search-all-trial-registers-comparative-study-sensitivity-five-trial-registers.

147. Chugh Y, De Francesco M, Prinja S. Systematic Literature Review of Guidelines on Budget Impact Analysis for Health Technology Assessment. Appl Health Econ Health Policy 2021; 19(6): 825-838. https://dx.doi.org/10.1007/s40258-021-00652-6.

148. Ciani O, Buyse M, Garside R 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. https://dx.doi.org/10.1136/bmj.f457.

149. Clarivate Analytics. EndNote [online]. [Accessed: 04.02.2021]. URL: https://endnote.com/.

150. Claxton K, Paulden M, Gravelle H et al. Discounting and decision making in the economic evaluation of health-care technologies. Health Econ 2011; 20(1): 2-15. https://dx.doi.org/10.1002/hec.1612.

151. Cochrane Effective Practice and Organisation of Care Review Group. Data collection checklist [online]. 2002 [Accessed: 04.02.2021]. URL: http://epoc.cochrane.org/sites/epoc.cochrane.org/files/uploads/datacollectionchecklist.pdf.

152. Collins GS, Reitsma JB, Altman DG et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015; 350: g7594. https://dx.doi.org/10.1136/bmj.g7594.

153. Commission of the European Communities. eEurope 2002: Quality Criteria for Health related Websites [online]. 2002 [Accessed: 04.02.2021]. URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2002:0667:FIN:EN:PDF.

154. 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]. 2005 [Accessed: 04.02.2021]. URL: http://www.who.int/patientsafety/highlights/COE_patient_and_medication_safety_gl.pdf.

155. Concannon TW, Fuster M, Saunders T et al. A systematic review of stakeholder engagement in comparative effectiveness and patient-centered outcomes research. J Gen Intern Med 2014; 29(12): 1692-1701. https://dx.doi.org/10.1007/s11606-014-2878-x.

156. Cooper C, Booth A, Britten N et al. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Syst Rev 2017; 6(1): 234. https://dx.doi.org/10.1186/s13643-017-0625-1.

157. Corbin JM, Strauss AL. Weiterleben lernen; Verlauf und Bewältigung chronischer Krankheiten. Bern: Huber; 2010.

158. Cordoba G, Schwartz L, Woloshin S et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ 2010; 341: c3920. https://dx.doi.org/10.1136/bmj.c3920.

159. Cornell JE, Laine C. The science and art of deduction: complex systematic overviews. Ann Intern Med 2008; 148(10): 786-788. https://dx.doi.org/10.7326/0003-4819-148-10-200805200-00012.

160. Craig P, Dieppe P, Macintyre S et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008; 337: a1655. https://dx.doi.org/10.1136/bmj.a1655.

161. Critical Appraisal Skills Programme. CASP Checklist: 10 questions to help you make sense of a Qualitative research [online]. 2018 [Accessed: 10.08.2021]. URL: https://casp-uk.net/wp-content/uploads/2018/03/CASP-Qualitative-Checklist-2018_fillable_form.pdf.

162. Cui L, Hung HM, Wang SJ et al. Issues related to subgroup analysis in clinical trials. J Biopharm Stat 2002; 12(3): 347-358. https://dx.doi.org/10.1081/bip-120014565.

163. Czypionka T, Reiss M, Ecker M. Lebenserwartung ist nicht genug: Gesundheitsmaße im Vergleich. Health System Watch 2016; 1: 133-148.

164. D'Agostino RB Jr. Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998; (17): 2265-2281. https://dx.doi.org/10.1002/(sici)1097-0258(19981015)17:19<2265::aid-sim918>3.0.co;2-b.

165. D'Agostino RB Jr, D'Agostino RB Sr. Estimating treatment effects using observational data. JAMA 2007; 297(3): 314-316. https://dx.doi.org/10.1001/jama.297.3.314.

166. 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. https://dx.doi.org/10.1002/sim.1425.

167. da Costa BR, Rutjes AW, Johnston BC 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. https://dx.doi.org/10.1093/ije/dys124.

168. Daly J, Willis K, Small R et al. A hierarchy of evidence for assessing qualitative health research. J Clin Epidemiol 2007; 60(1): 43-49. https://dx.doi.org/10.1016/j.jclinepi.2006.03.014.

169. Damarell RA, Tieman JJ, Sladek RM. OvidSP Medline-to-PubMed search filter translation: a methodology for extending search filter range to include PubMed's unique content. BMC Med Res Methodol 2013; 13: 86. https://dx.doi.org/10.1186/1471-2288-13-86.

170. Dans AL, Dans LF, Guyatt GH et al. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. JAMA 1998; 279(7): 545-549. https://dx.doi.org/10.1001/jama.279.7.545.

171. 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. https://dx.doi.org/10.1016/j.jclinepi.2010.09.009.

172. De Angelis CD, Drazen JM, Frizelle FA et al. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. Ann Intern Med 2005; 143(2): 146-148. https://dx.doi.org/10.7326/0003-4819-143-2-200507190-00016.

173. de Vries LM, van Baal PHM, Brouwer WBF. Future Costs in Cost-Effectiveness Analyses: Past, Present, Future. Pharmacoeconomics 2019; 37(2): 119-130. https://dx.doi.org/10.1007/s40273-018-0749-8.

174. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323(7305): 157-162. https://dx.doi.org/10.1136/bmj.323.7305.157.

175. 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. https://dx.doi.org/10.1002/sim.1188.

176. Deeks JJ, Higgins JPT, Altman DG. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 241-284.

177. Dental and Pharmaceutical Benefits Agency. Pricing and reimbursement of medicines [online]. 2020 [Accessed: 13.10.2022]. URL: https://www.tlv.se/in-english/medicines/pricing-and-reimbursement-of-medicines.html.

178. 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. https://dx.doi.org/10.1111/j.2044-8317.1992.tb00992.x.

179. 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. https://dx.doi.org/10.1186/1471-2288-1-7.

180. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7(3): 177-188. https://dx.doi.org/10.1016/0197-2456(86)90046-2.

181. 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. https://dx.doi.org/10.2105/ajph.94.3.361.

182. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ 2019; 367: l5657. https://dx.doi.org/10.1136/bmj.l5657.

183. Desu MM, Raghavarao D. Sample size methodology. Boston: Academic Press; 1990.

184. Deutsche Gesellschaft für Epidemiologie. Leitlinien und Empfehlungen zur Sicherung von Guter Epidemiologischer Praxis (GEP); Langversion [online]. 2018 [Accessed: 04.02.2021]. URL: https://www.dgepi.de/assets/Leitlinien-und-Empfehlungen/Leitlinien_fuer_Gute_Epidemiologische_Praxis_GEP_vom_September_2018.pdf.

185. Deutscher Bundestag. Gesetzentwurf der Fraktionen SPD, CDU/CSU und BÜNDNIS 90/DIE GRÜNEN: Entwurf eines Gesetzes zur Modernisierung der gesetzlichen Krankenversicherung (GKV-Modernisierungsgesetz - GMG) [online]. 2003 [Accessed: 04.02.2021]. URL: http://dipbt.bundestag.de/doc/btd/15/015/1501525.pdf.

186. Deutscher Ethikrat. Nutzen und Kosten im Gesundheitswesen – Zur normativen Funktion ihrer Bewertung; Stellungnahme [online]. 2011 [Accessed: 05.02.2021]. URL: https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/DER_StnAllo-Aufl2_Online.pdf.

187. 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.

188. Deutsches Netzwerk Evidenzbasierte Medizin. Die ‚Gute Praxis Gesundheitsinformation'. Z Evid Fortbild Qual Gesundhwes 2010; 104(1): 66-68. https://dx.doi.org/10.1016/j.zefq.2009.12.018.

189. Deutsches Netzwerk Evidenzbasierte Medizin. Gute Praxis Gesundheitsinformation: ein Positionspapier des Deutschen Netzwerks Evidenzbasierte Medizin e.V. [online]. 2016 [Accessed: 21.11.2019]. URL: https://www.ebm-netzwerk.de/de/medien/pdf/gpgi_2_20160721.pdf.

190. Deville WL, Buntinx F, Bouter LM et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2: 9. https://dx.doi.org/10.1186/1471-2288-2-9.

191. Devji T, Carrasco-Labra A, Lytvyn L et al. A new tool to measure credibility of studies determining minimally important difference estimates. Cochrane Database Syst Rev 2017; (9 Suppl 1): 58. https://dx.doi.org/10.1002/14651858.CD201702.

192. Devji T, Carrasco-Labra A, Qasim A et al. Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. BMJ 2020; 369: m1714. https://dx.doi.org/10.1136/bmj.m1714.

193. Devji T, Guyatt GH, Lytvyn L et al. Application of minimal important differences in degenerative knee disease outcomes: a systematic review and case study to inform BMJ Rapid Recommendations. BMJ Open 2017; 7(5): e015587. https://dx.doi.org/10.1136/bmjopen-2016-015587.

194. Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. Stat Med 2006; 25(13): 2299-2322. https://dx.doi.org/10.1002/sim.2287.

195. Dias S, Welton NJ, Caldwell DM et al. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010; 29(7-8): 932-944. https://dx.doi.org/10.1002/sim.3767.

196. DIPEx. Healthtalk.org [online]. [Accessed: 04.02.2021]. URL: http://www.healthtalk.org/.

197. 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. https://dx.doi.org/10.1016/s0277-9536(00)00247-1.

198. Dixon-Woods M, Agarwal S, Young B et al. Integrative approaches to qualitative and quantitative evidence [online]. 2004 [Accessed: 05.02.2021]. URL: https://www.webarchive.org.uk/wayback/archive/20140616160402mp_/http://nice.org.uk/nicemedia/documents/integrative_approaches.pdf.

199. Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews. Has established a place for itself. BMJ 2001; 323(7316): 765-766. https://dx.doi.org/10.1136/bmj.323.7316.765.

200. 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. https://dx.doi.org/10.1046/j.1365-2753.2001.00257.x.

201. Dixon-Woods M, Sutton A, Shaw R 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. https://dx.doi.org/10.1258/135581907779497486.

202. Djulbegovic B, Kumar A, Soares HP 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. https://dx.doi.org/10.1001/archinte.168.6.632.

203. Dobrescu AI, Nussbaumer-Streit B, Klerings I et al. Restricting evidence syntheses of interventions to English-language publications is a viable methodological shortcut for most medical topics: a systematic review. J Clin Epidemiol 2021; 137: 209-217. https://dx.doi.org/10.1016/j.jclinepi.2021.04.012.

204. Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making 2010; 30(1): 58-67. https://dx.doi.org/10.1177/0272989X07312478.

205. Doganay Erdogan B, Leung YY, Pohl C et al. Minimal Clinically Important Difference as Applied in Rheumatology: An OMERACT Rasch Working Group Systematic Review and Critique. J Rheumatol 2016; 43(1): 194-202. https://dx.doi.org/10.3899/jrheum.141150.

206. Donegan S, Williamson P, D'Alessandro U et al. Assessing key assumptions of network meta-analysis: a review of methods. Res Synth Methods 2013; 4(4): 291-323. https://dx.doi.org/10.1002/jrsm.1085.

207. Donner A, Klar J. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.

208. Draborg E, Gyrd-Hansen D, Poulsen PB et al. International comparison of the definition and the practical application of health technology assessment. Int J Technol Assess Health Care 2005; 21(1): 89-95. https://dx.doi.org/10.1017/s0266462305050117.

209. Droste S, Herrmann-Frank A, Scheibler F et al. Ethical issues in autologous stem cell transplantation (ASCT) in advanced breast cancer: a systematic literature review. BMC Med Ethics 2011; 12: 6. https://dx.doi.org/10.1186/1472-6939-12-6.

210. Drummond MF, Sculpher MJ, Claxton K et al. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2015.

211. Duffy S, de Kock S, Misso K et al. Supplementary searches of PubMed to improve currency of MEDLINE and MEDLINE In-Process searches via Ovid. J Med Libr Assoc 2016; 104(4): 309-312. https://dx.doi.org/10.3163/1536-5050.104.4.011.

212. Dundar Y, Dodd S, Dickson R et al. Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. Health Technol Assess 2006; 10(5): iii-iv, ix-145. https://dx.doi.org/10.3310/hta10050.

213. Dwan K, Altman DG, Arnaiz JA et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008; 3(8): e3081. https://dx.doi.org/10.1371/journal.pone.0003081.

214. Dwan K, Gamble C, Williamson PR et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One 2013; 8(7): e66844. https://dx.doi.org/10.1371/journal.pone.0066844.

215. Eady AM, Wilczynski NL, Haynes RB. PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. J Clin Epidemiol 2008; 61(1): 34-40. https://dx.doi.org/10.1016/j.jclinepi.2006.09.016.

216. Ebbesen M, Jensen TG, Andersen S et al. Ethical perspectives on RNA interference therapeutics. Int J Med Sci 2008; 5(3): 159-168. https://dx.doi.org/10.7150/ijms.5.159.

217. Ebell MH, Siwek J, Weiss BD 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.

218. Ebrahim S, Vercammen K, Sivanand A et al. Minimally Important Differences in Patient or Proxy-Reported Outcome Studies Relevant to Children: A Systematic Review. Pediatrics 2017; 139(3): e20160833. https://dx.doi.org/10.1542/peds.2016-0833.

219. Eddy DM, Hollingworth W, Caro JJ et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. Med Decis Making 2012; 32(5): 733-743. https://dx.doi.org/10.1177/0272989X12454579.

220. Eddy DM, Hollingworth W, Caro JJ et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. Value Health 2012; 15(6): 843-850. https://dx.doi.org/10.1016/j.jval.2012.04.012.

221. Eden J, Levit L, Berg A et al. Finding what works in health care; standards for systematic reviews. Washington: National Academies Press; 2011.

222. Edlin R, McCabe C, Hulme C et al. Cost effectiveness modelling for health technology assessment: a practical course. Heidelberg: Springer; 2015.

223. Edwards JE, McQuay HJ, Moore RA et al. Reporting of adverse effects in clinical trials should be improved: Lessons from acute postoperative pain. J Pain Symptom Manage 1999; 18(6): 427-437. https://dx.doi.org/10.1016/S0885-3924(99)00093-7.

224. Egger M, Davey Smith G, Altman DG. Systematic reviews in health care; meta-analysis in context. London: BMJ Publishing Group; 2001.

225. Egger M, Davey Smith G, Schneider M et al. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315(7109): 629-634. https://dx.doi.org/10.1136/bmj.315.7109.629.

226. Elwyn G, Frosch D, Thomson R et al. Shared decision making: a model for clinical practice. J Gen Intern Med 2012; 27(10): 1361-1367. https://dx.doi.org/10.1007/s11606-012-2077-6.

227. Elwyn G, O'Connor A, Stacey D et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ 2006; 333(7565): 417. https://dx.doi.org/10.1136/bmj.38926.629329.AE.

228. Engelke K, Droste S. Bewertungen der rechtlichen Aspekte von Technologien. In: Perleth M, Busse R, Gerhardus A et al (Ed). Health Technology Assessment; Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 280-296.

229. Engels EA, Schmid CH, Terrin N et al. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000; 19(13): 1707-1728. https://dx.doi.org/10.1002/1097-0258(20000715)19:13<1707::aid-sim491>3.0.co;2-p.

230. Epstein RM, Alper BS, Quill TE. Communicating evidence for participatory decision making. JAMA 2004; 291(19): 2359-2366. https://dx.doi.org/10.1001/jama.291.19.2359.

231. Etzioni R, Gulati R, Mallinger L et al. Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. Ann Intern Med 2013; 158(11): 831-838. https://dx.doi.org/10.7326/0003-4819-158-11-201306040-00008.

232. 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.

233. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EU) Nr. 536/2014 des Europäischen Parlaments und des Rates vom 16. April 2014 über klinische Prüfungen mit Humanarzneimitteln und zur Aufhebung der Richtlinie 2001/20/EG. Amtsblatt der Europäischen Union 2014; (L 158): 1-76.

234. Europäisches Parlament, Rat der Europäischen Union. Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates. Amtsblatt der Europäischen Union 2017; (L117): 1-175.

235. European Commission. Medical Devices - EUDAMED [online]. [Accessed: 05.08.2021]. URL: https://ec.europa.eu/health/md_eudamed/overview_en.

236. European Medicines Agency. Clinical trials in the European Union [online]. [Accessed: 18.08.2023]. URL: https://euclinicaltrials.eu/.

237. European Medicines Agency. Webauftritt [online]. [Accessed: 01.02.2021]. URL: https://www.ema.europa.eu/en.

238. European Medicines Agency. Points to consider on application with; 1. Meta-analyses; 2. One pivotal study [online]. 2001 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003657.pdf.

239. European Medicines Agency. Guideline on the choice of the non-inferiority margin [online]. 2005 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf.

240. 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]. 2005 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf.

241. European Medicines Agency. Guideline on clinical trials in small populations [online]. 2006 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003615.pdf.

242. European Medicines Agency. Guideline on clinical investigation of medicinal products in the treatment of diabetes mellitus; Draft [online]. 2010 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/02/WC500073570.pdf.

243. European Medicines Agency. Guideline on Missing Data in Confirmatory Clinical Trials [online]. 2010 [Accessed: 01.02.2021]. URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf.

244. European Medicines Agency. Posting of clinical trial summary results in European Clinical Trials Database (EudraCT) to become mandatory for sponsors as of 21 July 2014 [online]. 2014 [Accessed: 01.02.2021]. URL: https://www.ema.europa.eu/en/news/posting-clinical-trial-summary-results-european-clinical-trials-database-eudract-become-mandatory.

245. European Medicines Agency. Guideline on registry-based studies [online]. 2021 [Accessed: 12.10.2022]. URL: https://www.ema.europa.eu/documents/scientific-guideline/guideline-registry-based-studies_en-0.pdf.

246. European Network for Health Technology Assessment. Endpoints used for relative effectiveness assessment: health-related quality of life and utility measures [online]. 2015 [Accessed: 15.08.2022]. URL: https://eunethta.eu/wp-content/uploads/2018/01/Endpoints-used-for-Relative-Effectiveness-Assessment-Health-related-quality-of-life-and-utility-measures_Amended-JA1-Guideline_Final-Nov-2015.pdf.

247. European Network for Health Technology Assessment. Personalised medicine and co-dependent technologies, with a special focus on issues of study designs [online]. 2015 [Accessed: 18.10.2022]. URL: https://www.eunethta.eu/wp-content/uploads/2018/01/2016-03-07_reflection_paper_pm_2nd_draft.pdf.

248. European Network for Health Technology Assessment. Methods for health economic evaluations - A guideline based on current practices in Europe [online]. 2015 [Accessed: 05.09.2022]. URL: https://www.eunethta.eu/wp-content/uploads/2018/01/Methods-for-health-economic-evaluations-A-guideline-based-on-current-practices-in-Europe_Guideline_Final-May-2015.pdf.

249. European Network for Health Technology Assessment. Joint Action on HTA 2012-2015; HTA Core Model Version 3.0 [online]. 2016 [Accessed: 01.02.2021]. URL: https://eunethta.eu/wp-content/uploads/2018/03/HTACoreModel3.0-1.pdf.

250. European Network for Health Technology Assessment. Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness [online]. 2019 [Accessed: 08.08.2023]. URL: https://eunethta.eu/wp-content/uploads/2020/01/EUnetHTA_Guideline_Information_Retrieval_v2-0.pdf.

251. European Network for Health Technology Assessment. REQueST (Registry Evaluation and Quality Standards Tool) [online]. 2019 [Accessed: 13.10.2022]. URL: https://eunethta.eu/wp-content/uploads/2019/10/Registry-Evaluation-and-Quality-Standards-Tool-REQueST-1.xlsm.

252. European Network for Health Technology Assessment. Practical considerations when critically assessing economic evaluations [online]. 2020 [Accessed: 15.08.2022]. URL: https://www.eunethta.eu/wp-content/uploads/2020/03/EUnetHTA-JA3WP6B2-5-Guidance-Critical-Assessment-EE_v1-0.pdf.

253. European Network for Health Technology Assessment. D4.3.2; Direct and indirect comparisons [online]. 2022 [Accessed: 12.10.2022]. URL: https://www.eunethta.eu/wp-content/uploads/2022/08/EUnetHTA-21-Deliverable-D4.3.2-Methodological-Guideline-on-Direct-and-indirect-comparisons-V1.0.pdf.

254. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992; 268(17): 2420-2425. https://dx.doi.org/10.1001/jama.1992.03490170092032.

255. Ewald H, Klerings I, Wagner G et al. Searching two or more databases decreased the risk of missing relevant studies: a metaresearch study. J Clin Epidemiol 2022; 149: 154-164. https://dx.doi.org/10.1016/j.jclinepi.2022.05.022.

256. Eyding D, Lelgemann M, Grouven U 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. https://dx.doi.org/10.1136/bmj.c4737.

257. Facey K, Boivin A, Gracia J et al. Patients' perspectives in health technology assessment: a route to robust evidence and fair deliberation. Int J Technol Assess Health Care 2010; 26(3): 334-340. https://dx.doi.org/10.1017/S0266462310000395.

258. 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. https://dx.doi.org/10.1002/sim.4780091208.

259. Fayers P, Machin D. Quality of life; the assessment, analysis and interpretation of patient-reported outcomes. Chichester: Wiley; 2007.

260. Feenstra T, Corro-Ramos I, Hamerlijnck D et al. Four Aspects Affecting Health Economic Decision Models and Their Validation. Pharmacoeconomics 2022; 40(3): 241-248. https://dx.doi.org/10.1007/s40273-021-01110-w.

261. Feinstein AR. Clinical epidemiology; the architecture of clinical research. Philadelphia: Saunders; 1985.

262. Feinstein AR. Invidious comparisons and unmet clinical challenges. Am J Med 1992; 92(2): 117-120. https://dx.doi.org/10.1016/0002-9343(92)90099-w.

263. 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. https://dx.doi.org/10.1016/j.pec.2006.08.012.

264. Fergusson D, Aaron SD, Guyatt G et al. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002; 325(7365): 652-654. https://dx.doi.org/10.1136/bmj.325.7365.652.

265. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ et al. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012; 344: e686. https://dx.doi.org/10.1136/bmj.e686.

266. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ 2007; 334(7597): 786. https://dx.doi.org/10.1136/bmj.39136.682083.AE.

267. Field MJ, Lohr KN. Clinical practice guidelines; directions for a new program. Washington: National Academy Press; 1990.

268. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 1999; 94(446): 496-509. https://dx.doi.org/10.1080/01621459.1999.10474144.

269. Firth SJ. The Quality Adjusted Life Year: A Total-Utility Perspective. Camb Q Healthc Ethics 2018; 27(2): 284-294. https://dx.doi.org/10.1017/S0963180117000615.

270. Fleiss JL, Tytun A, Ury HK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 1980; 36(2): 343-346. https://dx.doi.org/10.2307/2529990.

271. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. Health Aff (Millwood) 2005; 24(1): 67-78. https://dx.doi.org/10.1377/hlthaff.24.1.67.

272. Fleming TR. Addressing missing data in clinical trials. Ann Intern Med 2011; 154(2): 113-117. https://dx.doi.org/10.7326/0003-4819-154-2-201101180-00010.

273. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996; 125(7): 605-613. https://dx.doi.org/10.7326/0003-4819-125-7-199610010-00011.

274. Fletcher RH, Fletcher SW. Klinische Epidemiologie; Grundlagen und Anwendung. Bern: Huber; 2007.

275. Forsetlund L, Kirkehei I, Harboe I et al. A comparison of two search methods for determining the scope of systematic reviews and health technology assessments. Int J Technol Assess Health Care 2012; 28(1): 59-64. https://dx.doi.org/10.1017/S0266462311000626.

276. 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: 989-991. https://dx.doi.org/10.1136/bmj.322.7292.989.

277. Freemantle N, Calvert M. Weighing the pros and cons for composite outcomes in clinical trials. J Clin Epidemiol 2007; 60(7): 658-659. https://dx.doi.org/10.1016/j.jclinepi.2006.10.024.

278. Friede T, Rover C, Wandel S et al. Meta-analysis of few small studies in orphan diseases. Res Synth Methods 2017; 8(1): 79-91. https://dx.doi.org/10.1002/jrsm.1217.

279. Friede T, Rover C, Wandel S et al. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biom J 2017; 59(4): 658-671. https://dx.doi.org/10.1002/bimj.201500236.

280. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11(2): 88-94. https://dx.doi.org/10.1177/0272989X9101100203.

281. 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. https://dx.doi.org/10.1093/ije/31.1.72.

282. Gartlehner G, Moore CG. Direct versus indirect comparisons: a summary of the evidence. Int J Technol Assess Health Care 2008; 24(2): 170-177. https://dx.doi.org/10.1017/S0266462308080240.

283. Gehanno JF, Rollin L, Le Jean T et al. Precision and recall of search strategies for identifying studies on return-to-work in Medline. J Occup Rehabil 2009; 19(3): 223-230. https://dx.doi.org/10.1007/s10926-009-9177-0.

284. 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: 01.02.2021]. URL: https://www.g-ba.de/downloads/17-98-3627/Anlage%20I_2-Kapitel-VerfO_Erprobungsantrag_Formular_2017-12-21_iK-2018-03-28.pdf.

285. Gemeinsamer Bundesausschuss. Mindestmengenregelungen gemäß § 136b Abs. 1 Satz 1 Nr. 2 SGB V [online]. URL: https://www.g-ba.de/richtlinien/5/.

286. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesauschusses; Anlage II zum 5. Kapitel – Format und Gliederung des Dossiers, einzureichende Unterlagen, Vorgaben für technische Standards [online]. [Accessed: 05.08.2021]. URL: https://www.g-ba.de/richtlinien/anlage/167/.

287. Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses [online]. URL: https://www.g-ba.de/informationen/richtlinien/42/.

288. Gemeinsamer Bundesausschuss. Beauftragung des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen [online]. 2004 [Accessed: 01.02.2021]. URL: https://www.g-ba.de/downloads/39-261-216/2004-12-21-Generalauftrag-IQWiG.pdf.

289. Gemeinsamer Bundesausschuss. Beschluss des Gemeinsamen Bundesausschusses über die Anpassung der Beauftragung des Instituts für Qualität und Wirtschaftlichkeit im Gesundheitswesen zur Erstellung von Patienteninformationen [online]. 2008 [Accessed: 01.02.2021]. URL: https://www.g-ba.de/downloads/39-261-650/2008-03-13-IQWiG-Anpassung-Generalauftrag.pdf.

290. Gemeinsamer Bundesausschuss. Richtlinie des Gemeinsamen Bundesausschusses zur Zusammenführung der Anforderungen an strukturierte Behandlungsprogramme nach § 137f Absatz 2 SGB V (DMP-Anforderungen-Richtlinie/DMP-A-RL) [online]. 2020 [Accessed: 05.08.2021]. URL: https://www.g-ba.de/downloads/62-492-2416/DMP-A-RL_2020-11-20_iK-2021-02-25.pdf.

291. Gerhardt U. Patientenkarrieren. Frankfurt am Main: Suhrkamp; 1986.

292. Gerhardus A, Stich AK. Die Bewertung sozio-kultureller Aspekte im HTA. In: Perleth M, Busse R, Gerhardus A et al (Ed). Health Technology Assessment; Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 312-320.

293. Gesundheit Österreich GmbH. PPRI Report 2018 [online]. 2019 [Accessed: 05.09.2022]. URL: https://jasmin.goeg.at/1068/1/PPRI%20Report2018_2nd_edition_final.pdf.

294. Ghabri S, Stevenson M, Moller J et al. Trusting the Results of Model-Based Economic Analyses: Is there a Pragmatic Validation Solution? Pharmacoeconomics 2019; 37(1): 1-6. https://dx.doi.org/10.1007/s40273-018-0711-9.

295. Glanville J, Kaunelis D, Mensinkai S. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. Int J Technol Assess Health Care 2009; 25(4): 522-529. https://dx.doi.org/10.1017/S0266462309990523.

296. Glanville JM, Duffy S, McCool R et al. Searching ClinicalTrials.gov and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches? J Med Libr Assoc 2014; 102(3): 177-183. https://dx.doi.org/10.3163/1536-5050.102.3.007.

297. Glasziou P, Chalmers I, Rawlins M et al. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007; 334(7589): 349-351. https://dx.doi.org/10.1136/bmj.39070.527986.68.

298. Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328(7430): 39-41. https://dx.doi.org/10.1136/bmj.328.7430.39.

299. Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Stat Med 2002; 21(11): 1503-1511. https://dx.doi.org/10.1002/sim.1183.

300. 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. https://dx.doi.org/10.1186/1472-6963-6-34.

301. Gliklich R, Dreyer N, Leavy M. Registries for evaluating patient outcomes: a user´s guide; 3rd edition; AHRQ pub. no. 13(14)-EHC111 [online]. 2014 [Accessed: 12.10.2022]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/related_files/registries-guide-3rd-edition-vol-2-140430.pdf.

302. Gold MR, Siegel JE, Russell LB et al. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.

303. Goldacre B, Drysdale H, Dale A et al. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. Trials 2019; 20(1): 118. https://dx.doi.org/10.1186/s13063-019-3173-2.

304. Goldacre B, Drysdale H, Marston C et al. COMPare: Qualitative analysis of researchers' responses to critical correspondence on a cohort of 58 misreported trials. Trials 2019; 20(1): 124. https://dx.doi.org/10.1186/s13063-019-3172-3.

305. Gonnermann A, Framke T, Grosshennig A et al. No solution yet for combining two independent studies in the presence of heterogeneity. Stat Med 2015; 34(16): 2476-2480. https://dx.doi.org/10.1002/sim.6473.

306. Gonnermann A, Kottas M, Koch A. Biometrische Entscheidungsunterstützung in Zulassung und Nutzenbewertung am Beispiel der Implikationen von heterogenen Ergebnissen in Untergruppen der Studienpopulation. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2015; 58(3): 274-282. https://dx.doi.org/10.1007/s00103-014-2105-2.

307. Gøtzsche PC, Liberati A, Torri V et al. Beware of surrogate outcome measures. Int J Technol Assess Health Care 1996; 12(2): 238-246. https://dx.doi.org/10.1017/s0266462300009594.

308. Govaerts L, Simoens S, Van Dyck W et al. Shedding Light on Reimbursement Policies of Companion Diagnostics in European Countries. Value Health 2020; 23(5): 606-615. https://dx.doi.org/10.1016/j.jval.2020.01.013.

309. Graf von der Schulenburg JM. Die Entwicklung der Gesundheitsökonomie und ihre methodischen Ansätze. In: Schöffski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. S. 13-21.

310. Graf von der Schulenburg JM, Greiner W, Jost F et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation - dritte und aktualisierte Fassung des Hannoveraner Konsens. Gesundheitsökonomie & Qualitätsmanagement 2007; 12(5): 285-290. https://dx.doi.org/10.1055/s-2007-963505.

311. Graham RM, Mancher M, Miller-Wolman D et al. Clinical practice guidelines we can trust [online]. 2011 [Accessed: 05.02.2021]. URL: http://www.awmf.org/fileadmin/user_upload/Leitlinien/International/IOM_CPG_lang_2011.pdf.

312. Gray JAM. How to get better value healthcare. Oxford: Offox Press; 2007.

313. Greenhalgh T, Hurwitz B. Narrative based medicine: why study narrative? BMJ 1999; 318(7175): 48-50. https://dx.doi.org/10.1136/bmj.318.7175.48.

314. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989; 18(1): 269-274. https://dx.doi.org/10.1093/ije/18.1.269.

315. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359(9300): 57-61. https://dx.doi.org/10.1016/S0140-6736(02)07283-5.

316. Grimes DA, Schulz KF. Uses and abuses of screening tests. Lancet 2002; 359(9309): 881-884. https://dx.doi.org/10.1016/S0140-6736(02)07948-5.

317. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. Obstet Gynecol 2005; 105(5 Pt 1): 1114-1118. https://dx.doi.org/10.1097/01.AOG.0000157445.67309.19.

318. Guddat C, Grouven U, Bender R et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev 2012; 1: 34. https://dx.doi.org/10.1186/2046-4053-1-34.

319. Günster C, Klauber J, Robra BP et al. Versorgungs-Report; Klima und Gesundheit. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2021.

320. Guolo A, Varin C. Random-effects meta-analysis: the number of studies matters. Stat Methods Med Res 2017; 26(3): 1500-1518. https://dx.doi.org/10.1177/0962280215583568.

321. Guyatt G, Rennie D, Meade MO et al. Users' guides to the medical literature; a manual for evidence-based clinical practice. New York: McGraw-Hill Education; 2015.

322. Guyatt GH. Evidence-based medicine. ACP J Club 1991; 114(2): A16. https://dx.doi.org/10.7326/ACPJC-1991-114-2-A16.

323. Guyatt GH, Oxman AD, Kunz R et al. Going from evidence to recommendations. BMJ 2008; 336(7652): 1049-1051. https://dx.doi.org/10.1136/bmj.39493.646875.AE.

324. Guyatt GH, Oxman AD, Kunz R et al. What is "quality of evidence" and why is it important to clinicians? BMJ 2008; 336(7651): 995-998. https://dx.doi.org/10.1136/bmj.39490.551019.BE.

325. Guyatt GH, Oxman AD, Sultan S et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011; 64(12): 1311-1316. https://dx.doi.org/10.1016/j.jclinepi.2011.06.004.

326. Guyatt GH, Oxman AD, Vist G et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol 2011; 64(4): 407-415. https://dx.doi.org/10.1016/j.jclinepi.2010.07.017.

327. Guyatt GH, Oxman AD, Vist GE et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008; 336(7650): 924-926. https://dx.doi.org/10.1136/bmj.39489.470347.AD.

328. Guyatt GH, Sackett DL, Sinclair JC et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA 1995; 274(22): 1800-1804. https://dx.doi.org/10.1001/jama.274.22.1800.

329. Haji Ali Afzali H, Bojke L, Karnon J. Model Structuring for Economic Evaluations of New Health Technologies. Pharmacoeconomics 2018; 36(11): 1309-1319. https://dx.doi.org/10.1007/s40273-018-0693-7.

330. Hall AE, Chowdhury S, Hallowell N et al. Implementing risk-stratified screening for common cancers: a review of potential ethical, legal and social issues. J Public Health (Oxf) 2014; 36(2): 285-291. https://dx.doi.org/10.1093/pubmed/fdt078.

331. Halladay CW, Trikalinos TA, Schmid IT et al. Using data sources beyond PubMed has a modest impact on the results of systematic reviews of therapeutic interventions. J Clin Epidemiol 2015; 68(9): 1076-1084. https://dx.doi.org/10.1016/j.jclinepi.2014.12.017.

332. Hamza TH, van Houwelingen HC, Heijenbrok-Kal MH et al. Associating explanatory variables with summary receiver operating characteristic curves in diagnostic meta-analysis. J Clin Epidemiol 2009; 62(12): 1284-1291. https://dx.doi.org/10.1016/j.jclinepi.2009.02.002.

333. Hao Q, Devji T, Zeraatkar D et al. Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation. BMJ Open 2019; 9(2): e028777. https://dx.doi.org/10.1136/bmjopen-2018-028777.

334. Harbord RM, Whiting P, Sterne JA 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. https://dx.doi.org/10.1016/j.jclinepi.2007.09.013.

335. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323(7308): 334-336. https://dx.doi.org/10.1136/bmj.323.7308.334.

336. Harden A, Garcia J, Oliver S 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. https://dx.doi.org/10.1136/jech.2003.014829.

337. Hardt JL, Metzendorf MI, Meerpohl JJ. Surgical trials and trial registers: a cross-sectional study of randomized controlled trials published in journals requiring trial registration in the author instructions. Trials 2013; 14: 407. https://dx.doi.org/10.1186/1745-6215-14-407.

338. Harrell FE Jr. Regression modeling strategies; with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

339. Harrell FE Jr, 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. https://dx.doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

340. Harris RP, Helfand M, Woolf SH 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. https://dx.doi.org/10.1016/s0749-3797(01)00261-6.

341. Hart D. Der regulatorische Rahmen der Nutzenbewertung: Vom Arzneimittelrecht zum HTA-Recht. Medizinrecht 2004; 22(9): 469-481. https://dx.doi.org/10.1007/s00350-004-1243-1.

342. Hart D. Leitlinien und Haftungsrecht - Inkorporation, Rezeption und Wissensbasis in Wissenschaft und Praxis. In: Hart D (Ed). Klinische Leitlinien und Recht. Baden-Baden: Nomos; 2005. S. 81-103.

343. Hartling L, Featherstone R, Nuspl M et al. The contribution of databases to the results of systematic reviews: a cross-sectional study. BMC Med Res Methodol 2016; 16(1): 127. https://dx.doi.org/10.1186/s12874-016-0232-1.

344. Hartung J. An alternative method for meta-analysis. Biom J 1999; 41(8): 901-916. https://dx.doi.org/10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.

345. Hausner E. Problems encountered with ICTRP Search Portal (comment on: "Van Enst WA et al. Identification of additional trials in prospective trial registers for Cochrane systematic reviews. PLoS One 2012; 7(8): e42812") [online]. 2014 [Accessed: 01.02.2021]. URL: https://journals.plos.org/plosone/article/comment?id=info:doi/10.1371/annotation/5aac3b1b-56ed-43bf-a07e-cb38afd31478.

346. Hausner E, Guddat C, Hermanns T et al. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. J Clin Epidemiol 2015; 68(2): 191-199. https://dx.doi.org/10.1016/j.jclinepi.2014.09.016.

347. Hausner E, Metzendorf MI, Richter B et al. Study filters for non-randomized studies of interventions consistently lacked sensitivity upon external validation. BMC Med Res Methodol 2018; 18(1): 171. https://dx.doi.org/10.1186/s12874-018-0625-4.

348. Hausner E, Waffenschmidt S. Value of using different search approaches [online]. 2019 [Accessed: 01.02.2021]. URL: http://vortal.htai.org/?q=node/993.

349. Hausner E, Waffenschmidt S, Kaiser T et al. Routine development of objectively derived search strategies. Syst Rev 2012; 1: 19. https://dx.doi.org/10.1186/2046-4053-1-19.

350. Haute Autorité de Santé. Companion diagnostic test associated with a targeted therapy: definitions and assessment method; INAHTA brief [online]. 2014 [Accessed: 03.11.2022]. URL: https://www.has-sante.fr/plugins/ModuleXitiKLEE/types/FileDocument/doXiti.jsp?id=c_1759172.

351. Haute Autorité de Santé. Dépot d'un dossier en vue d'un avis économique de la commission d'évaluation économique et de santé publique (CEESP) [online]. 2019 [Accessed: 22.09.2022]. URL: https://www.has-sante.fr/jcms/c_1627022/en/depot-d-un-dossier-en-vue-d-un-avis-economique-de-la-commission-d-evaluation-economique-et-de-sante-publique-ceesp.

352. Haute Autorité de Santé. Choix méthodologiques pour l'évaluation économique à la HAS [online]. 2020 [Accessed: 13.10.2022]. URL: https://www.has-sante.fr/plugins/ModuleXitiKLEE/types/FileDocument/doXiti.jsp?id=p_3197550.

353. Haynes RB, Cotoi C, Holland J et al. Second-order peer review of the medical literature for clinical practitioners. JAMA 2006; 295(15): 1801-1808. https://dx.doi.org/10.1001/jama.295.15.1801.

354. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. BMJ Evid Based Med 2002; 7(2): 36-38. https://dx.doi.org/10.1136/ebm.7.2.36.

355. Heintz E, Gerber-Grote A, Ghabri S et al. Is There a European View on Health Economic Evaluations? Results from a Synopsis of Methodological Guidelines Used in the EUnetHTA Partner Countries. Pharmacoeconomics 2016; 34(1): 59-76. https://dx.doi.org/10.1007/s40273-015-0328-1.

356. Helgesson G, Ernstsson O, Astrom M et al. Whom should we ask? A systematic literature review of the arguments regarding the most accurate source of information for valuation of health states. Qual Life Res 2020; 29(6): 1465-1482. https://dx.doi.org/10.1007/s11136-020-02426-4.

357. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol 2016; 183(8): 758-764. https://dx.doi.org/10.1093/aje/kwv254.

358. Herxheimer A, McPherson A, Miller R et al. Database of patients' experiences (DIPEx): a multi-media approach to sharing experiences and information. Lancet 2000; 355(9214): 1540-1543. https://dx.doi.org/10.1016/S0140-6736(00)02174-7.

359. Higgins JP, Ramsay C, Reeves BC 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 Synth Methods 2013; 4(1): 12-25. https://dx.doi.org/10.1002/jrsm.1056.

360. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11): 1539-1558. https://dx.doi.org/10.1002/sim.1186.

361. Higgins JP, Thompson SG, Deeks JJ et al. Measuring inconsistency in meta-analyses. BMJ 2003; 327(7414): 557-560. https://dx.doi.org/10.1136/bmj.327.7414.557.

362. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc 2009; 172(1): 137-159. https://dx.doi.org/10.1111/j.1467-985X.2008.00552.x.

363. Higgins JPT, Savović J, Page MJ et al. Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 205-228.

364. Higgins JPT, Thomas J, Chandler J et al. Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019.

365. Hill AB. Controlled clinical trials. Oxford: Blackwell; 1960.

366. Hingorani AD, Windt DA, Riley RD et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013; 346: e5793. https://dx.doi.org/10.1136/bmj.e5793.

367. Hoaglin DC, Hawkins N, Jansen JP et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value Health 2011; 14(4): 429-437. https://dx.doi.org/10.1016/j.jval.2011.01.011.

368. Hoffmann-Esser W, Siering U, Neugebauer EAM et al. Systematic review of current guideline appraisals performed with the Appraisal of Guidelines for Research & Evaluation II instrument-a third of AGREE II users apply a cut-off for guideline quality. J Clin Epidemiol 2018; 95: 120-127. https://dx.doi.org/10.1016/j.jclinepi.2017.12.009.

369. Hoffmann F, Kaiser T, Apfelbacher C et al. Erratum; Versorgungsnahe Daten zur Evaluation von Interventionseffekten: Teil 2 des Manuals. Gesundheitswesen 2021; 83(6): e40. https://dx.doi.org/10.1055/a-1516-8823.

370. Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. Int J Technol Assess Health Care 2005; 21(3): 312-318. https://dx.doi.org/10.1017/s0266462305050415.

371. Hofmann B, Droste S, Oortwijn W et al. Harmonization of ethics in health technology assessment: a revision of the Socratic approach. Int J Technol Assess Health Care 2014; 30(1): 3-9. https://dx.doi.org/10.1017/S0266462313000688.

372. Hofmann B, Haustein D, Landeweerd L. Smart-Glasses: Exposing and Elucidating the Ethical Issues. Sci Eng Ethics 2017; 23(3): 701-721. https://dx.doi.org/10.1007/s11948-016-9792-z.

373. Hofmann B, Lysdahl KB, Droste S. Evaluation of ethical aspects in health technology assessment: more methods than applications? Expert Rev Pharmacoecon Outcomes Res 2015; 15(1): 5-7. https://dx.doi.org/10.1586/14737167.2015.990886.

374. 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. https://dx.doi.org/10.1111/j.1369-7625.2007.00445.x.

375. Hopewell S, Clarke M, Askie L. Reporting of trials presented in conference abstracts needs to be improved. J Clin Epidemiol 2006; 59(7): 681-684. https://dx.doi.org/10.1016/j.jclinepi.2005.09.016.

376. 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. https://dx.doi.org/10.2105/ajph.81.12.1630.

377. Hoyer A, Hirt S, Kuss O. Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. Res Synth Methods 2018; 9(1): 62-72. https://dx.doi.org/10.1002/jrsm.1273.

378. Hoyer A, Zapf A. Studies for the Evaluation of Diagnostic Tests-Part 28 of a Series on Evaluation of Scientific Publications. Dtsch Arztebl Int 2021; 118(33-34): 555-560. https://dx.doi.org/10.3238/arztebl.m2021.0224.

379. 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. https://dx.doi.org/10.1186/1471-2288-5-13.

380. Hughes D, Charles J, Dawoud D et al. Conducting Economic Evaluations Alongside Randomised Trials: Current Methodological Issues and Novel Approaches. Pharmacoeconomics 2016; 34(5): 447-461. https://dx.doi.org/10.1007/s40273-015-0371-y.

381. Hung HMJ, ONeill RT, Bauer P et al. The behavior of the P-value when the alternative hypothesis is true. Biometrics 1997; 53(1): 11-22.

382. Hutton B, Salanti G, Caldwell DM et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015; 162(11): 777-784. https://dx.doi.org/10.7326/M14-2385.

383. ICH E1 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]. 1994 [Accessed: 18.03.2015]. URL: https://database.ich.org/sites/default/files/E1_Guideline.pdf.

384. ICH E9 Expert Working Group. ICH harmonised tripartite guideline: statistical principles for clinical trials. Stat Med 1999; 18(15): 1905-1942.

385. Iglesias CP, Thompson A, Rogowski WH et al. Reporting Guidelines for the Use of Expert Judgement in Model-Based Economic Evaluations. Pharmacoeconomics 2016; 34(11): 1161-1172. https://dx.doi.org/10.1007/s40273-016-0425-9.

386. Inan H. Measuring the success of your website; a customer-centric approach to website management. Frenchs Forest: Pearson Education Australia; 2002.

387. Institut für Arbeitsmarkt- und Berufsforschung. Aktuelle Ergebnisse [online]. [Accessed: 05.08.2021]. URL: http://www.iab.de/de/befragungen/stellenangebot/aktuelle-ergebnisse.aspx.

388. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Vereinbarung über die vertrauliche Behandlung von Unterlagen [online]. 2005 [Accessed: 17.02.2021]. URL: https://www.iqwig.de/download/IQWiG-VFA-Mustervertrag.pdf.

389. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Entwicklung und Anwendung von Modellen zur Berechnung von Schwellenwerten bei Mindestmengen für die Koronarchirurgie; Vorbericht [online]. 2006 [Accessed: 01.02.2021]. URL: https://www.iqwig.de/download/b05-01b_vorbericht_entwicklung_und_anwendung_von_modellen_zur_berechnung_von_schwellenwerten_bei_mindestmengen_fuer_die_koronarchirurgie.pdf.

390. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Zusammenhang zwischen Menge der erbrachten Leistungen und der Ergebnisqualität für die „Perkutane Transluminale Coronare Angieplastie (PTCA)“; Abschlussbericht [online]. 2006 [Accessed: 01.02.2021]. URL: https://www.iqwig.de/download/q05-01b_abschlussbericht_zusammenhang_menge_erbrachter_leistung_und_ergebnisqualitaet_bei_ptca.pdf.

391. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Bupropion, Mirtazapin und Reboxetin bei der Behandlung von Depression; Abschlussbericht [online]. 2009 [Accessed: 01.02.2021]. URL: https://www.iqwig.de/download/a05-20c_abschlussbericht_bupropion_mirtazapin_und_reboxetin_bei_depressionen.pdf.

392. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Aussagekraft von Surrogatendpunkten in der Onkologie; Rapid Report [online]. 2011 [Accessed: 01.02.2021]. URL: https://www.iqwig.de/download/a10-05_rapid_report_version_1-1_surrogatendpunkte_in_der_onkologie.pdf.

393. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Ticagrelor – Nutzenbewertung gemäß § 35a SGB V; Dossierbewertung [online]. 2011 [Accessed: 01.02.2021]. URL: https://www.iqwig.de/download/a11-02_ticagrelor_nutzenbewertung_35a_sgb_v_.pdf.

394. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Screening auf kritische angeborene Herzfehler mittels Pulsoxymetrie bei Neugeborenen; Abschlussbericht [online]. 2015 [Accessed: 02.08.2022]. URL: https://www.iqwig.de/download/s13-01_abschlussbericht_pulsoxymetrie.pdf.

395. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Allgemeine Methoden; Version 6.1 [online]. 2020 [Accessed: 02.08.2022]. URL: https://www.iqwig.de/methoden/allgemeine-methoden-v6-1.pdf.

396. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Dokumentation und Würdigung der Anhörung zum Entwurf der Allgemeinen Methoden 6.0 [online]. 2020 [Accessed: 09.02.2021]. URL: https://www.iqwig.de/methoden/allgemeine-methoden_dwa-entwurf-fuer-version-6-0_v1-0.pdf.

397. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V; Rapid Report [online]. 2020 [Accessed: 02.08.2022]. URL: https://www.iqwig.de/download/a19-43_versorgungsnahe-daten-zum-zwecke-der-nutzenbewertung_rapid-report_v1-1.pdf.

398. Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen. Datenschutzerklärung; Gesundheitsinformation.de [online]. 2021 [Accessed: 05.08.2021]. URL: https://www.gesundheitsinformation.de/service/datenschutz/.

399. Institute for Clinical and Economic Review. The QALY: rewarding the care that most improves patients' lives [online]. 2018 [Accessed: 12.10.2022]. URL: https://icer.org/wp-content/uploads/2020/12/QALY_evLYG_FINAL.pdf.

400. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Webauftritt [online]. [Accessed: 18.03.2015]. URL: https://www.ich.org/.

401. International Society for Pharmacoeconomics and Outcomes Research. Pharmacoeconomic Guidelines Around the World [online]. 2022 [Accessed: 05.09.2022]. URL: https://www.ispor.org/heor-resources/more-heor-resources/pharmacoeconomic-guidelines.

402. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 2014; 14: 25. https://dx.doi.org/10.1186/1471-2288-14-25.

403. Ioannidis JP, Evans SJ, Gotzsche PC et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004; 141(10): 781-788. https://dx.doi.org/10.7326/0003-4819-141-10-200411160-00009.

404. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001; 285(4): 437-443. https://dx.doi.org/10.1001/jama.285.4.437.

405. Ioannidis JP, Mulrow CD, Goodman SN. Adverse events: the more you search, the more you find. Ann Intern Med 2006; 144(4): 298-300. https://dx.doi.org/10.7326/0003-4819-144-4-200602210-00013.

406. Irmen L, Linner U. Die Repräsentation generisch maskuliner Personenbezeichnungen; eine theoretische Integration bisheriger Befunde. Z Psychol 2005; 213(3): 167-175. https://dx.doi.org/10.1026/0044-3409.213.3.167.

407. Irwig L, Tosteson AN, Gatsonis C et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120(8): 667-676. https://dx.doi.org/10.7326/0003-4819-120-8-199404150-00008.

408. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. Stat Med 2006; 25(15): 2688-2699. https://dx.doi.org/10.1002/sim.2481.

409. Jackson D, Law M, Rücker G et al. The Hartung-Knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? Stat Med 2017; 36(25): 3923-3934. https://dx.doi.org/10.1002/sim.7411.

410. Jackson D, Turner R. Power analysis for random-effects meta-analysis. Res Synth Methods 2017; 8(3): 290-302. https://dx.doi.org/10.1002/jrsm.1240.

411. Jackson N, Waters E. Criteria for the systematic review of health promotion and public health interventions. Health Promot Int 2005; 20(4): 367-374. https://dx.doi.org/10.1093/heapro/dai022.

412. Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. CMAJ 1997; 156(10): 1411-1416.

413. Jadad AR, Enkin MW. Randomized controlled trials; questions, answers and musings. Malden: Blackwell Publishing; 2007.

414. Jain R, Grabner M, Onukwugha E. Sensitivity analysis in cost-effectiveness studies: from guidelines to practice. Pharmacoeconomics 2011; 29(4): 297-314. https://dx.doi.org/10.2165/11584630-000000000-00000.

415. Jansen JP, Fleurence R, Devine B et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health 2011; 14(4): 417-428. https://dx.doi.org/10.1016/j.jval.2011.04.002.

416. Jansen JP, Trikalinos T, Cappelleri JC et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014; 17(2): 157-173. https://dx.doi.org/10.1016/j.jval.2014.01.004.

417. Janssens R, Russo S, van Overbeeke E et al. Patient Preferences in the Medical Product Life Cycle: What do Stakeholders Think? Semi-Structured Qualitative Interviews in Europe and the USA. Patient 2019; 12(5): 513-526. https://dx.doi.org/10.1007/s40271-019-00367-w.

418. Janzen T, Hausner E, Waffenschmidt S. Entwicklung und Evaluation von RCT- und SR-Filtern für die Suche nach nicht verschlagworteten Datensätzen in PubMed [online]. 2013 [Accessed: 01.02.2021]. URL: http://www.egms.de/static/de/meetings/ebm2013/13ebm059.shtml.

419. Jayadevappa R, Cook R, Chhatre S. Minimal important difference to infer changes in health-related quality of life-a systematic review. J Clin Epidemiol 2017; 89: 188-198. https://dx.doi.org/10.1016/j.jclinepi.2017.06.009.

420. Jenkins M. Evaluation of methodological search filters--a review. Health Info Libr J 2004; 21(3): 148-163. https://dx.doi.org/10.1111/j.1471-1842.2004.00511.x.

421. Jenuwine ES, Floyd JA. Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. J Med Libr Assoc 2004; 92(3): 349-353.

422. Johannesson M. Avoiding double-counting in pharmacoeconomic studies. Pharmacoeconomics 1997; 11(5): 385-388. https://dx.doi.org/10.2165/00019053-199711050-00001.

423. John J, Koerber F, Schad M. Differential discounting in the economic evaluation of healthcare programs. Cost Eff Resour Alloc 2019; 17: 29. https://dx.doi.org/10.1186/s12962-019-0196-1.

424. Johnston BC, Ebrahim S, Carrasco-Labra A et al. Minimally important difference estimates and methods: a protocol. BMJ Open 2015; 5(10): e007953. https://dx.doi.org/10.1136/bmjopen-2015-007953.

425. Jones B, Jarvis P, Lewis JA et al. Trials to assess equivalence: the importance of rigorous methods. BMJ 1996; 313(7048): 36-39. https://dx.doi.org/10.1136/bmj.313.7048.36.

426. Jones B, Kenward MG. Design and analysis of cross-over trials. Boca Raton: CRC Press; 2015.

427. Jones CW, Keil LG, Holland WC et al. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. BMC Med 2015; 13: 282. https://dx.doi.org/10.1186/s12916-015-0520-3.

428. Jones HE, Gatsonsis CA, Trikalinos TA et al. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. Stat Med 2019; 38(24): 4789-4803. https://dx.doi.org/10.1002/sim.8301.

429. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001; 323(7303): 42-46. https://dx.doi.org/10.1136/bmj.323.7303.42.

430. Juszczak E, Altman DG, Hopewell S et al. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. JAMA 2019; 321(16): 1610-1620. https://dx.doi.org/10.1001/jama.2019.3087.

431. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. Hoboken: Wiley-Interscience; 2002.

432. Kaltenthaler E, Essat M, Tappenden P et al. Identification and review of cost-effectiveness model parameters: a qualitative study. Int J Technol Assess Health Care 2014; 30(3): 333-340. https://dx.doi.org/10.1017/S0266462314000245.

433. Kapp C, Fujita-Rohwerder N, Sieben W et al. Implementation of a text analysis process using R for search strategy development in systematic reviews: Protocol. Zenodo 2023. https://dx.doi.org/10.5281/zenodo.7762286.

434. Karnon J, Stahl J, Brennan A et al. Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Value Health 2012; 15(6): 821-827. https://dx.doi.org/10.1016/j.jval.2012.04.013.

435. Kass PH, Gold EB. Modern epidemiologic study designs. In: Ahrens W, Pigeot I (Ed). Handbook of epidemiology. Berlin: Springer; 2005. S. 321-344.

436. Kastner M, Wilczynski NL, Walker-Dilks C et al. Age-specific search strategies for Medline. J Med Internet Res 2006; 8(4): e25. https://dx.doi.org/10.2196/jmir.8.4.e25.

437. Katrak P, Bialocerkowski AE, Massy-Westropp N et al. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004; 4: 22. https://dx.doi.org/10.1186/1471-2288-4-22.

438. Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med 2003; 138(8): 644-650. https://dx.doi.org/10.7326/0003-4819-138-8-200304150-00012.

439. Kettunen T, Liimatainen L, Villberg J et al. Developing empowering health counseling measurement. Preliminary results. Patient Educ Couns 2006; 64(1-3): 159-166. https://dx.doi.org/10.1016/j.pec.2005.12.012.

440. Kickbusch IS. Health literacy: addressing the health and education divide. Health Promot Int 2001; 16(3): 289-297. https://dx.doi.org/10.1093/heapro/16.3.289.

441. Kiefer C, Sturtz S, Bender R. Indirect comparisons and network meta-analyses: estimation of effects in the absence of head-to-head trials—part 22 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2015; 112(47): 803-808. https://dx.doi.org/10.3238/arztebl.2015.0803.

442. Kiefer C, Sturtz S, Bender R. A simulation study to compare different estimation approaches for network meta-analysis and corresponding methods to evaluate the consistency assumption. BMC Med Res Methodol 2020; 20(1): 36. https://dx.doi.org/10.1186/s12874-020-0917-3.

443. Kieser M, Hauschke D. Assessment of clinical relevance by considering point estimates and associated confidence intervals. Pharm Stat 2005; 4(2): 101-107. https://dx.doi.org/10.1002/pst.161.

444. 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. https://dx.doi.org/10.1002/sim.1910.

445. Kim DD, Wilkinson CL, Pope EF et al. The influence of time horizon on results of cost-effectiveness analyses. Expert Rev Pharmacoecon Outcomes Res 2017; 17(6): 615-623. https://dx.doi.org/10.1080/14737167.2017.1331432.

446. Kleinbaum DG, Klein M. Survival analysis; a self-learning text. New York: Springer; 2005.

447. Klerings I, Robalino S, Booth A et al. Rapid reviews methods series: Guidance on literature search. BMJ Evid Based Med 2023. https://dx.doi.org/10.1136/bmjebm-2022-112079.

448. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med 2003; 22(17): 2693-2710. https://dx.doi.org/10.1002/sim.1482.

449. Knelangen M, Hausner E, Metzendorf MI et al. Trial registry searches for randomized controlled trials of new drugs required registry-specific adaptation to achieve adequate sensitivity. J Clin Epidemiol 2018; 94: 69-75. https://dx.doi.org/10.1016/j.jclinepi.2017.11.003.

450. Koch A, Ziegler S. Metaanalyse als Werkzeug zum Erkenntnisgewinn. Med Klin (Munich) 2000; 95(2): 109-116. https://dx.doi.org/10.1007/BF03044996.

451. Koch K, Waltering A. IQWiG-Gesundheitsinformation: Pragmatischer Weg zum Themenkatalog. Dtsch Arztebl Ausg A 2016; 113(11): A489-A493.

452. Köhler M, Haag S, Biester K et al. Information on new drugs at market entry: retrospective analysis of health technology assessment reports versus regulatory reports, journal publications, and registry reports. BMJ 2015; 350: h796. https://dx.doi.org/10.1136/bmj.h796.

453. Kohlmann T. Patientenberichtete Studienendpunkte - Stand in Forschung und Praxis. Z Evid Fortbild Qual Gesundhwes 2010; 104(3): 259-263. https://dx.doi.org/10.1016/j.zefq.2010.03.014.

454. Kolman J, Meng P, Scott G. Good clinical practice; standard operating procedures for clinical researchers. Chichester: Wiley; 1998.

455. 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.

456. Koopmanschap MA, Rutten FF, van Ineveld BM et al. The friction cost method for measuring indirect costs of disease. J Health Econ 1995; 14(2): 171-189. https://dx.doi.org/10.1016/0167-6296(94)00044-5.

457. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. JAMA 2006; 296(10): 1286-1289. https://dx.doi.org/10.1001/jama.296.10.1286.

458. Krankheitserfahrungen.de. Webauftritt [online]. [Accessed: 01.02.2021]. URL: https://www.krankheitserfahrungen.de/.

459. Krauth C. Methoden der Kostenbestimmung in der gesundheitsökonomischen Evaluation. Gesundheitsökonomie & Qualitätsmanagement 2010; 15(5): 251-259. https://dx.doi.org/10.1055/s-0029-1245680.

460. Krauth C, Dintsios CM, Brandes I et al. Die Perspektive der Gesetzlichen Krankenversicherung in der gesundheitsökonomischen Evaluation. ZVersWiss 2005; 94: 215–256 https://dx.doi.org/10.1007/BF03353469.

461. Krauth C, Hessel F, Hansmeier T et al. Empirische Bewertungssätze in der gesundheitsökonomischen Evaluation - ein Vorschlag der AG Methoden der gesundheitsökonomischen Evaluation (AG MEG). Gesundheitswesen 2005; 67(10): 736-746. https://dx.doi.org/10.1055/s-2005-858698.

462. Kreis J, Puhan MA, Schünemann HJ et al. Consumer involvement in systematic reviews of comparative effectiveness research. Health Expect 2013; 16(4): 323-337. https://dx.doi.org/10.1111/j.1369-7625.2011.00722.x.

463. Kristensen FB, Sigmund H. Health technology assessment handbook [online]. 2008 [Accessed: 05.02.2021]. URL: https://www.sst.dk/~/media/ECAAC5AA1D6943BEAC96907E03023E22.ashx.

464. 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. https://dx.doi.org/10.1023/b:ejep.0000036568.02655.f8.

465. Krol M, Brouwer W. How to estimate productivity costs in economic evaluations. Pharmacoeconomics 2014; 32(4): 335-344. https://dx.doi.org/10.1007/s40273-014-0132-3.

466. Krol M, Brouwer W, Rutten F. Productivity costs in economic evaluations: past, present, future. Pharmacoeconomics 2013; 31(7): 537-549. https://dx.doi.org/10.1007/s40273-013-0056-3.

467. Krug S. Don't make me think! Web Usability; das intuitive Web. Heidelberg: mitp; 2006.

468. Kulbe A. Grundwissen Psychologie, Soziologie und Pädagogik; Lehrbuch für Pflegeberufe. Stuttgart: Kohlhammer; 2009.

469. Kunz R, Djulbegovic B, Schunemann HJ et al. Misconceptions, challenges, uncertainty, and progress in guideline recommendations. Semin Hematol 2008; 45(3): 167-175. https://dx.doi.org/10.1053/j.seminhematol.2008.04.005.

470. Kuss O. Statistical methods for meta-analyses including information from studies without any events-add nothing to nothing and succeed nevertheless. Stat Med 2015; 34(7): 1097-1116. https://dx.doi.org/10.1002/sim.6383.

471. Kuss O, Blettner M, Borgermann J. Propensity Score: an Alternative Method of Analyzing Treatment Effects. Dtsch Arztebl Int 2016; 113(35-36): 597-603. https://dx.doi.org/10.3238/arztebl.2016.0597.

472. Kwiecien R, Kopp-Schneider A, Blettner M. Concordance analysis: part 16 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(30): 515-521. https://dx.doi.org/10.3238/arztebl.2011.0515.

473. Laaser U, Hurrelmann K. Gesundheitsförderung und Krankheitsprävention. In: Hurrelmann K, Laaser U (Ed). Handbuch Gesundheitswissenschaften. Weinheim: Juventa Verlag; 1998. S. 395-424.

474. Lacny S, Wilson T, Clement F et al. Kaplan-Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. J Clin Epidemiol 2018; 93: 25-35. https://dx.doi.org/10.1016/j.jclinepi.2017.10.006.

475. Lancsar E, Gu Y, Gyrd-Hansen D et al. The relative value of different QALY types. J Health Econ 2020; 70: 102303. https://dx.doi.org/10.1016/j.jhealeco.2020.102303.

476. Langan SM, Schmidt SA, Wing K et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 2018; 363: k3532. https://dx.doi.org/10.1136/bmj.k3532.

477. Lange S, Freitag G. Choice of delta: requirements and reality--results of a systematic review. Biom J 2005; 47(1): 12-27. https://dx.doi.org/10.1002/bimj.200410085.

478. Lapsley P. The patient's journey: travelling through life with a chronic illness. BMJ 2004; 329: 582. https://dx.doi.org/10.1136/bmj.329.7466.582.

479. Lavis JN. How can we support the use of systematic reviews in policymaking? PLoS Med 2009; 6(11): e1000141. https://dx.doi.org/10.1371/journal.pmed.1000141.

480. Law AM. How to build valid and credible simulation models. In: Mason SJ, Hill RR, Mönch L et al (Ed). Proceedings of the 2008 Winter Simulation Conference. Piscataway: IEE; 2008. S. 39-47.

481. Law AM. Simulation modeling and analysis. New York: McGraw-Hill Education; 2015.

482. Lee ET, Wang JW. Statistical methods for survival data analysis. Hoboken: Wiley-Interscience; 2003.

483. Leeflang MM, Deeks JJ, Gatsonis C et al. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149(12): 889-897. https://dx.doi.org/10.7326/0003-4819-149-12-200812160-00008.

484. Lefebvre C, Glanville J, Beale S et al. Assessing the performance of methodological search filters to improve the efficiency of evidence information retrieval: five literature reviews and a qualitative study. Health Technol Assess 2017; 21(69): 1-148. https://dx.doi.org/10.3310/hta21690.

485. Lefebvre C, Glanville J, Briscoe S. Cochrane Handbook for Systematic Reviews of Interventions; Version 6.2; Technical Supplement to Chapter 4: Searching for and selecting studies [online]. 2021 [Accessed: 18.10.2022]. URL: https://training.cochrane.org/handbook/version-6.1/chapter-4-tech-suppl.

486. Lefebvre C, Glanville J, Briscoe S et al. Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 67-107.

487. Leidl R, Reitmeir P. An Experience-Based Value Set for the EQ-5D-5L in Germany. Value Health 2017; 20(8): 1150-1156. https://dx.doi.org/10.1016/j.jval.2017.04.019.

488. Lewin S, Booth A, Glenton C et al. Applying GRADE-CERQual to qualitative evidence synthesis findings: introduction to the series. Implement Sci 2018; 13(Suppl 1): 2. https://dx.doi.org/10.1186/s13012-017-0688-3.

489. Lewin S, Glenton C, Munthe-Kaas H et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 2015; 12(10): e1001895. https://dx.doi.org/10.1371/journal.pmed.1001895.

490. Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. BMJ 2001; 322(7300): 1479-1480. https://dx.doi.org/10.1136/bmj.322.7300.1479.

491. Leys M. Health care policy: qualitative evidence and health technology assessment. Health Policy 2003; 65(3): 217-226. https://dx.doi.org/10.1016/s0168-8510(02)00209-9.

492. 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. https://dx.doi.org/10.1017/s0266462300010369.

493. Lieb K, Klemperer D, Koch K et al. Interessenskonflikt in der Medizin; mit Transparenz Vertrauen stärken. Dtsch Arztebl Ausg A 2011; 108(6): A256-A260.

494. Lijmer JG, Bossuyt PM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009; 62(4): 364-373. https://dx.doi.org/10.1016/j.jclinepi.2008.06.017.

495. Lijmer JG, Mol BW, Heisterkamp S et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11): 1061-1066. https://dx.doi.org/10.1001/jama.282.11.1061.

496. Little RJA, Rubin DB. Statistical analysis with missing data. Hoboken: Wiley; 2002.

497. Lo B, Field MJ. Conflict of interest in medical research, education, and practice. Washington: National Academies Press; 2009.

498. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015; 13(3): 179-187. https://dx.doi.org/10.1097/XEB.0000000000000062.

499. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004; 23(20): 3105-3124. https://dx.doi.org/10.1002/sim.1875.

500. Lu G, Ades AE, Sutton AJ et al. Meta-analysis of mixed treatment comparisons at multiple follow-up times. Stat Med 2007; 26(20): 3681-3699. https://dx.doi.org/10.1002/sim.2831.

501. Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006; 101(474): 447-459. https://dx.doi.org/10.1198/016214505000001302.

502. Lu Y, Dai Z, Chang F et al. Whether and How Disutilities of Adverse Events were Used in the Economic Evaluation of Drug Therapy for Cancer Treatment. Pharmacoeconomics 2023; 41(3): 295-306. https://dx.doi.org/10.1007/s40273-022-01232-9.

503. Lühnen J, Albrecht M, Mühlhauser I et al. Leitlinie evidenzbasierte Gesundheitsinformation [online]. 2017 [Accessed: 01.02.2021]. URL: https://www.ebm-netzwerk.de/de/medien/pdf/leitlinie-evidenzbasierte-gesundheitsinformation-fin.pdf.

504. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002; 21(16): 2313-2324. https://dx.doi.org/10.1002/sim.1201.

505. Lysdahl KB, Mozygemba K, Burns L et al. Guidance for assessing effectiveness, economic aspects, ehtical aspects, socio-cultural aspects and legal aspects in complex technologies [online]. 2016 [Accessed: 05.02.2021]. URL: http://www.integrate-hta.eu/wp-content/uploads/2016/08/IPP_Guidance-INTEGRATE-HTA_Nr.3_FINAL.pdf.

506. Lysdahl KB, Oortwijn W, van der Wilt GJ et al. Ethical analysis in HTA of complex health interventions. BMC Med Ethics 2016; 17: 16. https://dx.doi.org/10.1186/s12910-016-0099-z.

507. Macaskill P, Gatsonis C, Deeks JJ et al. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy; Chapter 10 Analysing and Presenting Results [online]. 2010 [Accessed: 01.02.2021]. URL: https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter%2010%20-%20Version%201.0.pdf.

508. MacDermid JC, Brooks D, Solway S et al. Reliability and validity of the AGREE instrument used by physical therapists in assessment of clinical practice guidelines. BMC Health Serv Res 2005; 5(1): 18. https://dx.doi.org/10.1186/1472-6963-5-18.

509. Malterud K. The art and science of clinical knowledge: evidence beyond measures and numbers. Lancet 2001; 358(9279): 397-400. https://dx.doi.org/10.1016/S0140-6736(01)05548-9.

510. Mandeville KL, Valentic M, Ivankovic D et al. Quality Assurance of Registries for Health Technology Assessment. Int J Technol Assess Health Care 2018; 34(4): 360-367. https://dx.doi.org/10.1017/S0266462318000478.

511. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol 2009; 27(24): 4027-4034. https://dx.doi.org/10.1200/JCO.2009.22.3701.

512. Mandrekar SJ, Sargent DJ. All-comers versus enrichment design strategy in phase II trials. J Thorac Oncol 2011; 6(4): 658-660. https://dx.doi.org/10.1097/JTO.0b013e31820e17cb.

513. Mangiapane S, Velasco Garrido M. Surrogatendpunkte als Parameter der Nutzenbewertung [online]. 2009 [Accessed: 01.02.2021]. URL: https://portal.dimdi.de/de/hta/hta_berichte/hta250_bericht_de.pdf.

514. Marsh K, de Bekker-Grob E, Cook N et al. How to integrate evidence from patient preference studies into health technology assessment: a critical review and recommendations. Int J Technol Assess Health Care 2021; 37(1). https://dx.doi.org/10.1017/S0266462321000490.

515. Marsh K, van Til JA, Molsen-David E et al. Health Preference Research in Europe: A Review of Its Use in Marketing Authorization, Reimbursement, and Pricing Decisions-Report of the ISPOR Stated Preference Research Special Interest Group. Value Health 2020; 23(7): 831-841. https://dx.doi.org/10.1016/j.jval.2019.11.009.

516. Marshall IJ, Noel-Storr A, Kuiper J et al. Machine learning for identifying Randomized Controlled Trials: An evaluation and practitioner's guide. Res Synth Methods 2018; 9(4): 602-614. https://dx.doi.org/10.1002/jrsm.1287.

517. Martin LP, Arias-Gallo J, Perez-Chrzanowska H et al. Transfusion Requirements in Microsurgical Reconstruction in Maxillofacial Surgery: Ethical and Legal Problems of Patients Who Are Jehovah's Witnesses. Craniomaxillofac Trauma Reconstr 2013; 6(1): 31-36. https://dx.doi.org/10.1055/s-0033-1333828.

518. Martini P. Methodenlehre der therapeutischen Untersuchung. Berlin: Springer; 1932.

519. Mathes T, Kuss O. A comparison of methods for meta-analysis of a small number of studies with binary outcomes. Res Synth Methods 2018; 9(3): 366-381. https://dx.doi.org/10.1002/jrsm.1296.

520. Matthias K, Gruber S, Pietsch B. Evidenz von Volume-Outcome-Beziehungen und Mindestmengen: Diskussion in der aktuellen Literatur. Gesundheits- und Sozialpolitik 2014; 68(3): 23-30. https://dx.doi.org/10.5771/1611-5821-2014-3-23.

521. Mattke S, Kelley E, Scherer P et al. Health Care Quality Indicators Project: Initial Indicators Report [online]. 2006 [Accessed: 01.02.2021]. URL: http://dx.doi.org/10.1787/481685177056.

522. McAlister FA, Straus SE, Sackett DL et al. Analysis and reporting of factorial trials: a systematic review. JAMA 2003; 289(19): 2545-2553. https://dx.doi.org/10.1001/jama.289.19.2545.

523. McCulloch P, Taylor I, Sasako M et al. Randomised trials in surgery: problems and possible solutions. BMJ 2002; 324(7351): 1448-1451. https://dx.doi.org/10.1136/bmj.324.7351.1448.

524. McGauran N, Wieseler B, Kreis J et al. Reporting bias in medical research - a narrative review. Trials 2010; 11: 37. https://dx.doi.org/10.1186/1745-6215-11-37.

525. McGowan J, Sampson M, Salzwedel DM et al. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E) [online]. 2016 [Accessed: 01.02.2021]. URL: https://www.cadth.ca/sites/default/files/pdf/CP0015_PRESS_Update_Report_2016.pdf.

526. McInnes MDF, Moher D, Thombs BD et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA 2018; 319(4): 388-396. https://dx.doi.org/10.1001/jama.2017.19163.

527. Meerpohl J, Wild C. Evidenzbasierte Gesundheitsversorgung. Z Evid Fortbild Qual Gesundhwes 2016; 110-111: 6-7. https://dx.doi.org/10.1016/j.zefq.2015.12.011.

528. Merlin T, Lehman S, Hiller JE et al. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care 2013; 29(3): 343-350. https://dx.doi.org/10.1017/S0266462313000287.

529. Merlin TL, Hiller JE, Ryan P. Impact of the "Linked Evidence Approach" Method on Policies to Publicly Fund Diagnostic, Staging, and Screening Medical Tests. MDM Policy Pract 2016; 1(1): 2381468316672465. https://dx.doi.org/10.1177/2381468316672465.

530. Moher D, Hopewell S, Schulz KF et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869. https://dx.doi.org/10.1136/bmj.c869.

531. Moher D, Pham B, Lawson ML et al. The inclusion of reports of randomised trials published in languages other than English in systematic reviews. Health Technol Assess 2003; 7(41): 1-90. https://dx.doi.org/10.3310/hta7410.

532. Moher D, Shamseer L, Clarke M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015; 4: 1. https://dx.doi.org/10.1186/2046-4053-4-1.

533. Möhler R. Qualitative Evidenzsynthesen - Methodologien, Methoden und Herausforderungen. QuPuG 2016; 3(2): 70-77.

534. Molenberghs G, Burzykowski T, Alonso A et al. A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials. Stat Methods Med Res 2010; 19(3): 205-236. https://dx.doi.org/10.1177/0962280209105015.

535. 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. https://dx.doi.org/10.1111/j.1532-5415.2008.02122.x.

536. Moons KG, Altman DG, Reitsma JB et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015; 162(1): W1-73. https://dx.doi.org/10.7326/M14-0698.

537. Mozygemba K, Hofmann B, Lysdal KB et al. Guidance to assess socio-cultural aspects [online]. 2016 [Accessed: 17.02.2021]. URL: http://www.integrate-hta.eu/wp-content/uploads/2016/08/IPP_Guidance-INTEGRATE-HTA_Nr.3_FINAL.pdf.

538. Muhlbacher AC, Johnson FR. Giving Patients a Meaningful Voice in European Health Technology Assessments: The Role of Health Preference Research. Patient 2017; 10(4): 527-530. https://dx.doi.org/10.1007/s40271-017-0249-5.

539. Muhlbacher AC, Kaczynski A. Making Good Decisions in Healthcare with Multi-Criteria Decision Analysis: The Use, Current Research and Future Development of MCDA. Appl Health Econ Health Policy 2016; 14(1): 29-40. https://dx.doi.org/10.1007/s40258-015-0203-4.

540. Mullan RJ, Flynn DN, Carlberg B et al. Systematic reviewers commonly contact study authors but do so with limited rigor. J Clin Epidemiol 2009; 62(2): 138-142. https://dx.doi.org/10.1016/j.jclinepi.2008.08.002.

541. 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. https://dx.doi.org/10.7326/0003-4819-136-2-200201150-00009.

542. Munn Z, Porritt K, Lockwood C et al. Establishing confidence in the output of qualitative research synthesis: the ConQual approach. BMC Med Res Methodol 2014; 14: 108. https://dx.doi.org/10.1186/1471-2288-14-108.

543. Murphy E, Dingwall R, Greatbatch D et al. Qualitative research methods in health technology assessment: a review of the literature. Health Technol Assess 1998; 2(16): iii-ix, 1-274.

544. Mustafa RA, Wiercioch W, Cheung A et al. Decision making about healthcare-related tests and diagnostic test strategies. Paper 2: a review of methodological and practical challenges. J Clin Epidemiol 2017; 92: 18-28. https://dx.doi.org/10.1016/j.jclinepi.2017.09.003.

545. National Advisory Committee on Health and Disability. Screening to Improve Health in New Zealand; Criteria to assess screening programmes [online]. 2003 [Accessed: 05.02.2021]. URL: https://www.nsu.govt.nz/system/files/resources/screening_to_improve_health.pdf.

546. National Health and Medical Research Council. Cultural competency in health: a guide for policy, partnerships and participation [online]. 2006 [Accessed: 05.02.2021]. URL: https://www.nhmrc.gov.au/file/2771/download?token=H3v1MEBC.

547. National Institute for Health and Care Excellence. Guidance and advice list: interventional procedures guidance [online]. [Accessed: 01.02.2021]. URL: https://www.nice.org.uk/guidance/indevelopment?type=ipg.

548. National Institute for Health and Care Excellence. Guide to the processes of technology appraisal [online]. 2018 [Accessed: 03.11.2022]. URL: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf.

549. National Institutes of Health, Department of Health and Human Services. Clinical trials registration and results information submission; final rule. Fed Regist 2016; 81(183): 64981-65157.

550. Neumann PJ, Cohen JT. QALYs in 2018-Advantages and Concerns. JAMA 2018; 319(24): 2473-2474. https://dx.doi.org/10.1001/jama.2018.6072.

551. Neumann PJ, Cohen JT, Ollendorf DA. Drug-Pricing Debate Redux - Should Cost-Effectiveness Analysis Be Used Now to Price Pharmaceuticals? N Engl J Med 2021; 385(21): 1923-1924. https://dx.doi.org/10.1056/NEJMp2113323.

552. Newcombe RG, Bender R. Implementing GRADE: calculating the risk difference from the baseline risk and the relative risk. Evid Based Med 2014; 19(1): 6-8. https://dx.doi.org/10.1136/eb-2013-101340.

553. Nielsen J, Loranger H. Web Usability. München: Addison-Wesley; 2008.

554. Nilsen ES, Myrhaug HT, Johansen M et al. Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. Cochrane Database Syst Rev 2006; (3): CD004563. https://dx.doi.org/10.1002/14651858.CD004563.pub2.

555. Nordin A, Taft C, Lundgren-Nilsson A et al. Minimal important differences for fatigue patient reported outcome measures-a systematic review. BMC Med Res Methodol 2016; 16: 62. https://dx.doi.org/10.1186/s12874-016-0167-6.

556. Norwegian Medicines Agency. Guidelines for the submission of documentation for single technology assessment (STA) of pharmaceuticals [online]. 2021 [Accessed: 13.10.2022]. URL: https://legemiddelverket.no/Documents/English/Public%20funding%20and%20pricing/Documentation%20for%20STA/Guidelines%2018.10.2021.pdf.

557. Nothacker M, Muche-Borowski C, Kopp IB. 20 Jahre ärztliche Leitlinien in Deutschland - was haben sie bewirkt? Z Evid Fortbild Qual Gesundhwes 2014; 108(10): 550-559. https://dx.doi.org/10.1016/j.zefq.2014.10.012.

558. Nüesch E, Jüni P. Commentary: Which meta-analyses are conclusive? Int J Epidemiol 2009; 38(1): 298-303. https://dx.doi.org/10.1093/ije/dyn265.

559. Nutbeam D. Health promotion glossary. Health Promot Int 1998; 13(4): 349-364. https://dx.doi.org/10.1093/heapro/13.4.349.

560. O'Mahony JF, Newall AT, van Rosmalen J. Dealing with Time in Health Economic Evaluation: Methodological Issues and Recommendations for Practice. Pharmacoeconomics 2015; 33(12): 1255-1268. https://dx.doi.org/10.1007/s40273-015-0309-4.

561. O'Neill RT. Statistical analyses of adverse event data from clinical trials. Special emphasis on serious events. Drug Inf J 1987; 21(1): 9-20. https://dx.doi.org/10.1177/009286158702100104.

562. O'Neill RT. Assessment of safety. In: Peace KE (Ed). Biopharmaceutical statistics for drug development. New York: Dekker; 1988. S. 543-604.

563. O'Rourke B, Oortwijn W, Schuller T et al. The new definition of health technology assessment: A milestone in international collaboration. Int J Technol Assess Health Care 2020; 36(3): 187-190. https://dx.doi.org/10.1017/S0266462320000215.

564. OCEBM Levels of Evidence Working Group. Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence [online]. 2011 [Accessed: 15.11.2021]. URL: https://www.cebm.ox.ac.uk/files/levels-of-evidence/cebm-levels-of-evidence-2-1.pdf.

565. Odgaard-Jensen J, Vist GE, Timmer A et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev 2011; (4): MR000012. https://dx.doi.org/10.1002/14651858.MR000012.pub3.

566. Organisation for Economic Co-operation and Development. OECD Secretary-General's Report to Ministers 2021 [online]. 2021 [Accessed: 05.08.2021]. URL: https://www.oecd-ilibrary.org/deliver/8cd95b77-en.pdf.

567. Ousmen A, Touraine C, Deliu N et al. Distribution- and anchor-based methods to determine the minimally important difference on patient-reported outcome questionnaires in oncology: a structured review. Health Qual Life Outcomes 2018; 16(1): 228. https://dx.doi.org/10.1186/s12955-018-1055-z.

568. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 44(11): 1271-1278. https://dx.doi.org/10.1016/0895-4356(91)90160-b.

569. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116(1): 78-84. https://dx.doi.org/10.7326/0003-4819-116-1-78.

570. Oxman AD, Guyatt GH, Singer J et al. Agreement among reviewers of review articles. J Clin Epidemiol 1991; 44(1): 91-98. https://dx.doi.org/10.1016/0895-4356(91)90205-N.

571. Page MJ, Higgins JPT, Sterne JAC. Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J et al (Ed). Cochrane handbook for systematic reviews of interventions. Hoboken: Wiley-Blackwell; 2019. S. 349-374.

572. Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021; 372: n71. https://dx.doi.org/10.1136/bmj.n71.

573. Page MJ, Moher D, Bossuyt PM et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021; 372: n160. https://dx.doi.org/10.1136/bmj.n160.

574. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? Health Econ 2006; 15(7): 653-664. https://dx.doi.org/10.1002/hec.1086.

575. 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. https://dx.doi.org/10.1002/(sici)1097-0258(19981230)17:24<2815::aid-sim110>3.0.co;2-8.

576. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980; 302(20): 1109-1117. https://dx.doi.org/10.1056/NEJM198005153022003.

577. Paynter R, Banez LL, Berliner E et al. EPC Methods: An Exploration of the Use of TextMining Software in Systematic Reviews [online]. 2016 [Accessed: 01.02.2021]. URL: https://www.ncbi.nlm.nih.gov/books/NBK362044/pdf/Bookshelf_NBK362044.pdf.

578. Pazzagli L, Linder M, Zhang M et al. Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. Pharmacoepidemiol Drug Saf 2018; 27(2): 148-160. https://dx.doi.org/10.1002/pds.4372.

579. Pearson I, Rothwell B, Olaye A et al. Economic Modeling Considerations for Rare Diseases. Value Health 2018; 21(5): 515-524. https://dx.doi.org/10.1016/j.jval.2018.02.008.

580. Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: a meta-analytical comparison of patient and population health state utilities. Value Health 2010; 13(2): 306-309. https://dx.doi.org/10.1111/j.1524-4733.2009.00610.x.

581. Perleth M, Busse R, Gerhardus A et al. Health Technology Assessment; Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014.

582. Perleth M, Gibis B, Velasco Garrido M et al. Organisationsstrukturen und Qualität. In: Perleth M, Busse R, Gerhardus A et al (Ed). Health Technology Assessment; Konzepte, Methoden, Praxis für Wissenschaft und Entscheidungsfindung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft; 2014. S. 265-280.

583. 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. https://dx.doi.org/10.1016/s0168-8510(00)00138-x.

584. Petitti DB, Teutsch SM, Barton MB et al. Update on the methods of the U.S. Preventive Services Task Force: insufficient evidence. Ann Intern Med 2009; 150(3): 199-205. https://dx.doi.org/10.7326/0003-4819-150-3-200902030-00010.

585. Petkova E, Tarpey T, Huang L et al. Interpreting meta-regression: application to recent controversies in antidepressants' efficacy. Stat Med 2013; 32(17): 2875-2892. https://dx.doi.org/10.1002/sim.5766.

586. Pettitt DA, Raza S, Naughton B et al. The Limitations of QALY: A Literature Review. J Stem Cell Res Ther 2016; 6(4): 1000334. https://dx.doi.org/10.4172/2157-7633.1000334.

587. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee [online]. 2016 [Accessed: 18.10.2022]. URL: https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf.

588. Pharmaceuticals Pricing Board. Preparing a health economic evaluation to be attached to the application for reimbursement status and wholesale price for a medicinal product [online]. 2019 [Accessed: 13.10.2022]. URL: https://www.hila.fi/content/uploads/2020/01/Instructions_TTS_2019.pdf.

589. Piaggio G, Elbourne DR, Pocock SJ et al. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. JAMA 2012; 308(24): 2594-2604. https://dx.doi.org/10.1001/jama.2012.87802.

590. Pieper D, Jülich F, Antoine SL et al. Studies analysing the need for health-related information in Germany - a systematic review. BMC Health Serv Res 2015; 15: 407. https://dx.doi.org/10.1186/s12913-015-1076-9.

591. Pike J, Grosse SD. Friction Cost Estimates of Productivity Costs in Cost-of-Illness Studies in Comparison with Human Capital Estimates: A Review. Appl Health Econ Health Policy 2018; 16(6): 765-778. https://dx.doi.org/10.1007/s40258-018-0416-4.

592. Pirk O, Schöffski O. Primärdatenerhebung. In: Schöffski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. S. 197-242.

593. Pitman R, Fisman D, Zaric GS et al. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5. Med Decis Making 2012; 32(5): 712-721. https://dx.doi.org/10.1177/0272989X12454578.

594. Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. Stat Med 1999; 18(6): 643-654. https://dx.doi.org/10.1002/(sici)1097-0258(19990330)18:6<643::aid-sim76>3.0.co;2-m.

595. Pocock SJ. Clinical trials; a practical approach. Chichester: Wiley; 1983.

596. Pranic S, Marusic A. Changes to registration elements and results in a cohort of Clinicaltrials.gov trials were not reflected in published articles. J Clin Epidemiol 2016; 70: 26-37. https://dx.doi.org/10.1016/j.jclinepi.2015.07.007.

597. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989; 8(4): 431-440. https://dx.doi.org/10.1002/sim.4780080407.

598. Public Health England. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme [online]. 2015 [Accessed: 01.02.2021]. URL: https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme.

599. Pufulete M, Mahadevan K, Johnson TW et al. Confounders and co-interventions identified in non-randomized studies of interventions. J Clin Epidemiol 2022; 148(148): 115-123. https://dx.doi.org/10.1016/j.jclinepi.2022.03.018.

600. Putter H, Schumacher M, van Houwelingen HC. On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine-Gray model revisited. Biom J 2020; 62(3): 790-807. https://dx.doi.org/10.1002/bimj.201800274.

601. Rand LZ, Kesselheim AS. Controversy Over Using Quality-Adjusted Life-Years In Cost-Effectiveness Analyses: A Systematic Literature Review. Health Aff (Millwood) 2021; 40(9): 1402-1410. https://dx.doi.org/10.1377/hlthaff.2021.00343.

602. Raum E, Perleth M. Methoden der Metaanalyse von diagnostischen Genauigkeitsstudien [online]. 2003 [Accessed: 01.02.2021]. URL: http://portal.dimdi.de/de/hta/hta_berichte/hta025_bericht_de.pdf.

603. Reitsma JB, Glas AS, Rutjes AW et al. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005; 58(10): 982-990. https://dx.doi.org/10.1016/j.jclinepi.2005.02.022.

604. Relevo R, Balshem H. Evidence for Comparing Medical Interventions [online]. 2011 [Accessed: 01.02.2021]. URL: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/methods-guidance-finding-evidence_methods.pdf.

605. Rethlefsen ML, Kirtley S, Waffenschmidt S et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Syst Rev 2021; 10(1): 39. https://dx.doi.org/10.1186/s13643-020-01542-z.

606. Revicki D, Hays RD, Cella D et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008; 61(2): 102-109. https://dx.doi.org/10.1016/j.jclinepi.2007.03.012.

607. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011; 342: d549. https://dx.doi.org/10.1136/bmj.d549.

608. Roberts M, Russell LB, Paltiel AD et al. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. Value Health 2012; 15(6): 804-811. https://dx.doi.org/10.1016/j.jval.2012.06.016.

609. Roberts M, Russell LB, Paltiel AD et al. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. Med Decis Making 2012; 32(5): 678-689. https://dx.doi.org/10.1177/0272989X12454941.

610. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11(5): 550-560. https://dx.doi.org/10.1097/00001648-200009000-00011.

611. Robinson KA, Whitlock EP, O'Neil ME et al. Integration of Existing Systematic Reviews. Research White Paper (Prepared by the Scientific Resource Center under Contract No. 290-2012-00004-C) [online]. 2014 [Accessed: 01.02.2021]. URL: https://www.ncbi.nlm.nih.gov/sites/books/NBK216379/pdf/Bookshelf_NBK216379.pdf.

612. Rockwood K, Fay S, Song X et al. Attainment of treatment goals by people with Alzheimer's disease receiving galantamine: a randomized controlled trial. CMAJ 2006; 174(8): 1099-1105. https://dx.doi.org/10.1503/cmaj.051432.

613. Roebruck P, Elze M, Hauschke D et al. Literaturübersicht zur Fallzahlplanung für Äquivalenzprobleme. Inform Biom Epidemiol Med Biol 1997; 28(2): 51-63.

614. Röhmel J, Hauschke D, Koch A et al. Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren; Nicht-Unterlegenheit in klinischen Studien. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2005; 48(5): 562-571. https://dx.doi.org/10.1007/s00103-005-1042-5.

615. Rosenthal JA. Qualitative Descriptors of Strength of Association and Effect Size. J Soc Serv Res 1996; 21(4): 37-59. https://dx.doi.org/10.1300/J079v21n04_02.

616. Roters D. "Recht auf Wissen": haben gesetzlich Krankenversicherte Anspruch auf diagnostische Leistungen ohne medizinische Konsequenzen? Z Evid Fortbild Qual Gesundhwes 2011; 105(7): 526-530. https://dx.doi.org/10.1016/j.zefq.2011.07.037.

617. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365(9454): 176-186. https://dx.doi.org/10.1016/S0140-6736(05)17709-5.

618. Rover C, Bender R, Dias S et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Synth Methods 2021; 12(4): 448-474. https://dx.doi.org/10.1002/jrsm.1475.

619. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Med Res Methodol 2015; 15: 99. https://dx.doi.org/10.1186/s12874-015-0091-1.

620. Royall RM. The Effect of Sample-Size on the Meaning of Significance Tests. Am Stat 1986; 40(4): 313-315. https://dx.doi.org/10.1080/00031305.1986.10475424.

621. Royle P, Bain L, Waugh N. Systematic reviews of epidemiology in diabetes: finding the evidence. BMC Med Res Methodol 2005; 5: 2. https://dx.doi.org/10.1186/1471-2288-5-2.

622. Royle P, Waugh N. Literature searching for clinical and cost-effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. Health Technol Assess 2003; 7(34): iii, ix-x, 1-51. https://dx.doi.org/10.3310/hta7340.

623. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med 2000; 19(14): 1831-1847. https://dx.doi.org/10.1002/1097-0258(20000730)19:14<1831::aid-sim502>3.0.co;2-1.

624. Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modeling. J R Stat Soc Ser C Appl Stat 1994; 43(3): 429-467. https://dx.doi.org/10.2307/2986270.

625. Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods 2012; 3(4): 312-324. https://dx.doi.org/10.1002/jrsm.1058.

626. Rücker G, Schwarzer G, Carpenter J et al. 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. https://dx.doi.org/10.1002/sim.3511.

627. Rucker G, Schwarzer G, Carpenter JR et al. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008; 8: 79. https://dx.doi.org/10.1186/1471-2288-8-79.

628. Runkel B, Bein G, Sieben W et al. Targeted antenatal anti-D prophylaxis for RhD-negative pregnant women: a systematic review. BMC Pregnancy Childbirth 2020; 20(1): 83. https://dx.doi.org/10.1186/s12884-020-2742-4.

629. Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen. Bedarfsgerechtigkeit und Wirtschaftlichkeit; Band III; Über- Unter- und Fehlversorgung; Gutachten 2000/2001; ausführliche Zusammenfassung [online]. 2001 [Accessed: 05.08.2021]. URL: https://www.svr-gesundheit.de/fileadmin/Gutachten/Gutachten_2000_2001/Kurzfassung_Band3.pdf.

630. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32(1-2): 51-63. https://dx.doi.org/10.1016/0021-9681(79)90012-2.

631. Sackett DL, Rosenberg WM, Gray JA et al. Evidence based medicine: what it is and what it isn't. BMJ 1996; 312(7023): 71-72. https://dx.doi.org/10.1136/bmj.312.7023.71.

632. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods 2012; 3(2): 80-97. https://dx.doi.org/10.1002/jrsm.1037.

633. Salanti G, Higgins JP, Ades AE et al. Evaluation of networks of randomized trials. Stat Methods Med Res 2008; 17(3): 279-301. https://dx.doi.org/10.1177/0962280207080643.

634. Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol 2009; 62(8): 857-864. https://dx.doi.org/10.1016/j.jclinepi.2008.10.001.

635. Sampson CJ, Arnold R, Bryan S et al. Transparency in Decision Modelling: What, Why, Who and How? Pharmacoeconomics 2019; 37(11): 1355-1369. https://dx.doi.org/10.1007/s40273-019-00819-z.

636. Sampson M. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003; 56(10): 943-955. https://dx.doi.org/10.1016/s0895-4356(03)00110-0.

637. Sampson M, McGowan J. Errors in search strategies were identified by type and frequency. J Clin Epidemiol 2006; 59(10): 1057-1063. https://dx.doi.org/10.1016/j.jclinepi.2006.01.007.

638. Sampson M, McGowan J. Inquisitio validus Index Medicus: A simple method of validating MEDLINE systematic review searches. Res Synth Methods 2011; 2(2): 103-109. https://dx.doi.org/10.1002/jrsm.40.

639. Sampson M, McGowan J, Cogo E et al. An evidence-based practice guideline for the peer review of electronic search strategies. J Clin Epidemiol 2009; 62(9): 944-952. https://dx.doi.org/10.1016/j.jclinepi.2008.10.012.

640. Sampson M, McGowan J, Lefebvre C et al. PRESS: Peer Review of Electronic Search Strategies [online]. 2008 [Accessed: 05.02.2021]. URL: http://www.cadth.ca/media/pdf/477_PRESS-Peer-Review-Electronic-Search-Strategies_tr_e.pdf.

641. Sampson M, McGowan J, Tetzlaff J et al. No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 2008; 61(8): 748-754. https://dx.doi.org/10.1016/j.jclinepi.2007.10.009.

642. Samson D, Schoelles KM. Chapter 2: medical tests guidance (2) developing the topic and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. J Gen Intern Med 2012; 27 Suppl 1: S11-19. https://dx.doi.org/10.1007/s11606-012-2007-7.

643. Sanders GD, Neumann PJ, Basu A et al. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. JAMA 2016; 316(10): 1093-1103. https://dx.doi.org/10.1001/jama.2016.12195.

644. Sargent DJ, Conley BA, Allegra C et al. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 2005; 23(9): 2020-2027. https://dx.doi.org/10.1200/JCO.2005.01.112.

645. Sargent DJ, Mandrekar SJ. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. Clin Trials 2013; 10(5): 647-652. https://dx.doi.org/10.1177/1740774513497125.

646. SAS Institute. SAS/STAT 9.2 User's Guide; Second Edition [online]. 2009 [Accessed: 01.02.2021]. URL: http://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf.

647. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc Ser A Stat Soc 1999; 162(1): 71-94. https://dx.doi.org/10.1111/1467-985x.00122.

648. Sauerland S, Fujita-Rohwerder N, Zens Y et al. Premarket evaluation of medical devices: a cross-sectional analysis of clinical studies submitted to a German ethics committee. BMJ Open 2019; 9(2): e027041. https://dx.doi.org/10.1136/bmjopen-2018-027041.

649. Sawaya GF, Guirguis-Blake J, LeFevre M et al. 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. https://dx.doi.org/10.7326/0003-4819-147-12-200712180-00007.

650. Sayers A. Tips and tricks in performing a systematic review. Br J Gen Pract 2007; 57(545): 999. https://dx.doi.org/10.3399/096016407782604938.

651. Schad M, John J. Towards a social discount rate for the economic evaluation of health technologies in Germany: an exploratory analysis. Eur J Health Econ 2012; 13(2): 127-144. https://dx.doi.org/10.1007/s10198-010-0292-9.

652. Schlosser RW, Wendt O, Bhavnani S et al. Use of information-seeking strategies for developing systematic reviews and engaging in evidence-based practice: the application of traditional and comprehensive Pearl Growing. A review. Int J Lang Commun Disord 2006; 41(5): 567-582. https://dx.doi.org/10.1080/13682820600742190.

653. Schmoor C, Schumacher M, Finke J et al. Competing risks and multistate models. Clin Cancer Res 2013; 19(1): 12-21. https://dx.doi.org/10.1158/1078-0432.CCR-12-1619.

654. Schöffski O. Grundformen gesundheitsökonomischer Evaluationen. In: Schöfski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. S. 43-70.

655. Schöffski O, Graf von der Schulenburg JM. Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012.

656. Scholz S, Biermann-Stallwitz J, Brettschneider C et al. Standardisierte Kostenberechnungen im deutschen Gesundheitswesen: Bericht der Arbeitsgruppe „Standardkosten" des Ausschusses „ökonomische Evaluation" der dggö. Gesundheitsökonomie & Qualitätsmanagement 2020; 25(1): 52-59. https://dx.doi.org/10.1055/a-1107-0665.

657. Schulz A, Schürmann C, Skipka G et al. Performing Meta-analyses with Very Few Studies. In: Evangelou V, Veroniki AA (Ed). Meta-Research; Methods and Protocols. New York: Humana; 2022. S. 91-102.

658. Schulz KF, Altman DG, Moher D et al. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c332. https://dx.doi.org/10.1136/bmj.c332.

659. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002; 359(9308): 781-785. https://dx.doi.org/10.1016/S0140-6736(02)07882-0.

660. Schünemann H, Brożek J, Guyatt G et al. GRADE handbook [online]. 2013 [Accessed: 27.05.2021]. URL: https://gdt.gradepro.org/app/handbook/handbook.html.

661. 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. https://dx.doi.org/10.1186/1477-7525-4-62.

662. Schünemann HJ, Best D, Vist G et al. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003; 169(7): 677-680.

663. 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. https://dx.doi.org/10.1186/1478-4505-4-21.

664. Schünemann HJ, Mustafa R, Brozek J. Diagnostik und linked evidence - Wie robust muss die Kette sein? Z Evid Fortbild Qual Gesundhwes 2012; 106(3): 153-160. https://dx.doi.org/10.1016/j.zefq.2012.04.001.

665. Schünemann HJ, Oxman AD, Brozek J et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008; 336(7653): 1106-1110. https://dx.doi.org/10.1136/bmj.39500.677199.AE.

666. Schwalm A, Neusser S, Mostardt S et al. Methoden der Kostenberechnung von Arzneimitteln im deutschen Gesundheitssystem: Bericht der Arbeitsgruppe „Standardkosten“ des Ausschusses „ökonomische Evaluation“ der dggö. Gesundheitsökonomie & Qualitätsmanagement 2020; 25(1): 44-51. https://dx.doi.org/10.1055/a-1107-0104.

667. Schwarzer R, Rochau U, Saverno K et al. Systematic overview of cost-effectiveness thresholds in ten countries across four continents. J Comp Eff Res 2015; 4(5): 485-504. https://dx.doi.org/10.2217/cer.15.38.

668. Sculpher MJ, Claxton K, Drummond M et al. Whither trial-based economic evaluation for health care decision making? Health Econ 2006; 15(7): 677-687.

669. Senn S. Inherent difficulties with active control equivalence studies. Stat Med 1993; 12(24): 2367-2375. https://dx.doi.org/10.1002/sim.4780122412.

670. Senn S. The many modes of meta. Drug Inf J 2000; 34(2): 535-549. https://dx.doi.org/10.1177/009286150003400222.

671. Senn S. Cross-over trials in clinical research. Chichester: Wiley; 2002.

672. Senn S. Trying to be precise about vagueness. Stat Med 2007; 26(7): 1417-1430. https://dx.doi.org/10.1002/sim.2639.

673. Shamseer L, Moher D, Clarke M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015; 350: g7647. https://dx.doi.org/10.1136/bmj.g7647.

674. Shea BJ, Bouter LM, Peterson J et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS One 2007; 2(12): e1350. https://dx.doi.org/10.1371/journal.pone.0001350.

675. Shea BJ, Grimshaw JM, Wells GA et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 7: 10. https://dx.doi.org/10.1186/1471-2288-7-10.

676. Shea BJ, Hamel C, Wells GA 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. https://dx.doi.org/10.1016/j.jclinepi.2008.10.009.

677. Shea BJ, Reeves BC, Wells G et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017; 358: j4008. https://dx.doi.org/10.1136/bmj.j4008.

678. Shechter SM, Schaefer AJ, Braithwaite RS et al. Increasing the efficiency of Monte Carlo cohort simulations with variance reduction techniques. Med Decis Making 2006; 26(5): 550-553. https://dx.doi.org/10.1177/0272989X06290489.

679. Shekelle PG, Ortiz E, Rhodes S 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. https://dx.doi.org/10.1001/jama.286.12.1461.

680. Shojania KG, Sampson M, Ansari MT et al. How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 2007; 147(4): 224-233. https://dx.doi.org/10.7326/0003-4819-147-4-200708210-00179.

681. 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: 01.02.2021]. URL: https://portal.dimdi.de/de/hta/hta_berichte/hta099_bericht_de.pdf.

682. Siebert U, Alagoz O, Bayoumi AM et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--3. Value Health 2012; 15(6): 812-820. https://dx.doi.org/10.1016/j.jval.2012.06.014.

683. Siebert U, Jahn B, Mühlberger N et al. Entscheidungsanalyse und Modellierungen. In: Schöffski O, Graf von der Schulenburg JM (Ed). Gesundheitsökonomische Evaluationen. Berlin: Springer; 2012. S. 275-324.

684. Siering U, Eikermann M, Hausner E et al. Appraisal tools for clinical practice guidelines: a systematic review. PLoS One 2013; 8(12): e82915. https://dx.doi.org/10.1371/journal.pone.0082915.

685. Signorovitch JE, Wu EQ, Yu AP et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics 2010; 28(10): 935-945. https://dx.doi.org/10.2165/11538370-000000000-00000.

686. Silvestre MA, 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. https://dx.doi.org/10.1016/j.jclinepi.2010.09.008.

687. Simmonds MC, Higgins JP. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Stat Med 2007; 26(15): 2982-2999. https://dx.doi.org/10.1002/sim.2768.

688. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. Stat Methods Med Res 2016; 25(6): 2858-2877. https://dx.doi.org/10.1177/0962280214534409.

689. 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. https://dx.doi.org/10.1093/jnci/djp335.

690. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG et al. Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol 2014; 67(6): 612-621. https://dx.doi.org/10.1016/j.jclinepi.2013.12.008.

691. Skipka G, Wieseler B, Kaiser T et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. Biom J 2016; 58(1): 43-58. https://dx.doi.org/10.1002/bimj.201300274.

692. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. BMJ 1999; 318(7197): 1548-1551. https://dx.doi.org/10.1136/bmj.318.7197.1548.

693. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med 1995; 14(24): 2685-2699. https://dx.doi.org/10.1002/sim.4780142408.

694. Söletormos G, Duffy MJ, Hayes DF et al. Design of tumor biomarker-monitoring trials: a proposal by the European Group on Tumor Markers. Clin Chem 2013; 59(1): 52-59. https://dx.doi.org/10.1373/clinchem.2011.180778.

695. Song F, Loke YK, Walsh T et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009; 338: b1147. https://dx.doi.org/10.1136/bmj.b1147.

696. Song F, Parekh S, Hooper L et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010; 14(8): iii, ix-xi, 1-193. https://dx.doi.org/10.3310/hta14080.

697. Spacirova Z, Epstein D, Garcia-Mochon L et al. A general framework for classifying costing methods for economic evaluation of health care. Eur J Health Econ 2020; 21(4): 529-542. https://dx.doi.org/10.1007/s10198-019-01157-9.

698. Spacirova Z, Epstein D, Garcia-Mochon L et al. Correction to: A general framework for classifying costing methods for economic evaluation of health care. Eur J Health Econ 2021; 22(5): 847. https://dx.doi.org/10.1007/s10198-021-01313-0.

699. Spencer FA, Iorio A, You J et al. Uncertainties in baseline risk estimates and confidence in treatment effects. BMJ 2012; 345: e7401. https://dx.doi.org/10.1136/bmj.e7401.

700. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian Approaches to Randomized Trials. J R Stat Soc Ser A Stat Soc 1994; 157(3): 357-387. https://dx.doi.org/10.2307/2983527.

701. Spiegelhalter DJ, Myles JP, Jones DR et al. Methods in health service research. An introduction to bayesian methods in health technology assessment. BMJ 1999; 319(7208): 508-512. https://dx.doi.org/10.1136/bmj.319.7208.508.

702. St-Pierre C, Desmeules F, Dionne CE et al. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. Disabil Rehabil 2016; 38(2): 103-122. https://dx.doi.org/10.3109/09638288.2015.1027004.

703. Stamuli E. Health outcomes in economic evaluation: who should value health? Br Med Bull 2011; 97: 197-210. https://dx.doi.org/10.1093/bmb/ldr001.

704. Stansfield C, O'Mara-Eves A, Thomas J. Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. Res Synth Methods 2017; 8(3): 355-365. https://dx.doi.org/10.1002/jrsm.1250.

705. Statens Beredning för Medicinsk och Social Utvärdering. Utvärdering av metoder i hälso- och sjukvården och insatser i socialtjänsten [online]. 2017 [Accessed: 10.08.2021]. URL: http://www.sbu.se/globalassets/ebm/metodbok/sbushandbok.pdf.

706. Statistisches Bundesamt. Verbraucherpreisindizes [online]. [Accessed: 01.02.2021]. URL: https://www.destatis.de/DE/Themen/Wirtschaft/Preise/Verbraucherpreisindex/Publikationen/_publikationen-verbraucherpreisindex.html.

707. Stausberg J, Maier B, Bestehorn K et al. Memorandum Register für die Versorgungsforschung: Update 2019. Gesundheitswesen 2020; 82(3): e39-e66. https://dx.doi.org/10.1055/a-1083-6417.

708. Steiner JF. The use of stories in clinical research and health policy. JAMA 2005; 294(22): 2901-2904. https://dx.doi.org/10.1001/jama.294.22.2901.

709. Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Med Res Methodol 2016; 16(1): 97. https://dx.doi.org/10.1186/s12874-016-0196-1.

710. Sterne JA, Hernan MA, Reeves BC et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355: i4919. https://dx.doi.org/10.1136/bmj.i4919.

711. Sterne JAC, Savovic J, Page MJ et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019; 366: l4898. https://dx.doi.org/10.1136/bmj.l4898.

712. Stewart LA, Clarke M, Rovers M et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA 2015; 313(16): 1657-1665. https://dx.doi.org/10.1001/jama.2015.3656.

713. Stock S, Lauterbach KW, Sauerland S. Gesundheitsökonomie; Lehrbuch für Mediziner und andere Gesundheitsberufe. Berlin: Hogrefe; 2021.

714. Stukel TA, Fisher ES, Wennberg DE et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007; 297(3): 278-285. https://dx.doi.org/10.1001/jama.297.3.278.

715. Sturtz S, Bender R. Unsolved issues of mixed treatment comparison meta-analysis: network size and inconsistency. Res Synth Methods 2012; 3(4): 300-311. https://dx.doi.org/10.1002/jrsm.1057.

716. Suarez-Almazor ME, Belseck E, Homik J et al. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. Control Clin Trials 2000; 21(5): 476-487. https://dx.doi.org/10.1016/s0197-2456(00)00067-2.

717. Sullivan SD, Mauskopf JA, Augustovski F et al. Budget impact analysis-principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. Value Health 2014; 17(1): 5-14. https://dx.doi.org/10.1016/j.jval.2013.08.2291.

718. Sullivan W, Payne K. The appropriate elicitation of expert opinion in economic models: making expert data fit for purpose. Pharmacoeconomics 2011; 29(6): 455-459. https://dx.doi.org/10.2165/11589220-000000000-00000.

719. 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. https://dx.doi.org/10.1016/0895-4356(96)00025-X.

720. Sun X, Briel M, Walter SD et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010; 340: c117. https://dx.doi.org/10.1136/bmj.c117.

721. Sung YT, Wu JS. The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): A new technique for psychological measurement. Behav Res Methods 2018; 50(4): 1694-1715. https://dx.doi.org/10.3758/s13428-018-1041-8.

722. Superchi C, Brion Bouvier F, Gerardi C et al. Study designs for clinical trials applied to personalised medicine: a scoping review. BMJ Open 2022; 12(5): e052926. https://dx.doi.org/10.1136/bmjopen-2021-052926.

723. Sutton A, Ades AE, Cooper N et al. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008; 26(9): 753-767. https://dx.doi.org/10.2165/00019053-200826090-00006.

724. Swart E, Bitzer EM, Gothe H et al. A Consensus German Reporting Standard for Secondary Data Analyses, Version 2 (STROSA-STandardisierte BerichtsROutine fur SekundardatenAnalysen). Gesundheitswesen 2016; 78(S 01): e145-e160. https://dx.doi.org/10.1055/s-0042-108647.

725. Swart E, Bitzer EM, Gothe H et al. STandardisierte BerichtsROutine für Sekundärdaten Analysen (STROSA) – ein konsentierter Berichtsstandard für Deutschland, Version 2. Gesundheitswesen 2016; 78(S 01): e145-e160. https://dx.doi.org/10.1055/s-0042-108647.

726. Swart E, Ihle P, Gothe H et al. Routinedaten im Gesundheitswesen; Handbuch Sekundärdatenanalyse; Grundlagen, Methoden und Perspektiven. Bern: Huber; 2014.

727. Swift TL, Dieppe PA. Using expert patients' narratives as an educational resource. Patient Educ Couns 2005; 57(1): 115-121. https://dx.doi.org/10.1016/j.pec.2004.05.004.

728. Tai FM, Willson ML, Ghersi D. Implications of searching multiple trial registries: how should we search ClinicalTrials.gov and WHO ICTRP? [online]. 2012 [Accessed: 01.02.2021]. URL: https://abstracts.cochrane.org/2012-auckland/implications-searching-multiple-trial-registries-how-should-we-search.

729. Taichman DB, Sahni P, Pinborg A et al. Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. PLoS Med 2017; 14(6): e1002315. https://dx.doi.org/10.1371/journal.pmed.1002315.

730. Tajik P, Zwinderman AH, Mol BW et al. Trial designs for personalizing cancer care: a systematic review and classification. Clin Cancer Res 2013; 19(17): 4578-4588. https://dx.doi.org/10.1158/1078-0432.CCR-12-3722.

731. Takwoingi Y, Guo B, Riley RD et al. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res 2017; 26(4): 1896-1911. https://dx.doi.org/10.1177/0962280215592269.

732. Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013; 158(7): 544-554. https://dx.doi.org/10.7326/0003-4819-158-7-201304020-00006.

733. Tan SS. Microcosting in Economic Evaluations; Issues of accuracy, feasibility, consistency and generalisability [online]. 2009 [Accessed: 01.02.2021]. URL: http://repub.eur.nl/res/pub/17354/091127_Tan,%20Siok%20Swan.pdf.

734. Tan SS, Bouwmans CA, Rutten FF et al. Update of the Dutch Manual for Costing in Economic Evaluations. Int J Technol Assess Health Care 2012; 28(2): 152-158. https://dx.doi.org/10.1017/S0266462312000062.

735. Thomas J, Harden A, Oakley A et al. Integrating qualitative research with trials in systematic reviews. BMJ 2004; 328(7446): 1010-1012. https://dx.doi.org/10.1136/bmj.328.7446.1010.

736. Thomas S. Klinische Relevanz von Therapieeffekten: systematische Sichtung, Klassifizierung und Bewertung methodischer Konzepte. Duisburg/Essen: Universität; 2009.

737. Thompson JC, Quigley JM, Halfpenny NJ et al. Importance and methods of searching for E-publications ahead of print in systematic reviews. Evid Based Med 2016; 21(2): 55-59. https://dx.doi.org/10.1136/ebmed-2015-110374.

738. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21(11): 1559-1573. https://dx.doi.org/10.1002/sim.1187.

739. Thorlund K, Devereaux PJ, Wetterslev J et al. Can trial sequential monitoring boundaries reduce spurious inferences from meta-analyses? Int J Epidemiol 2009; 38(1): 276-286. https://dx.doi.org/10.1093/ije/dyn179.

740. Thurow S. Search engine visibility. Berkley: New Riders; 2008.

741. Tian L, Cai T, Pfeffer MA et al. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2 x 2 tables with all available data but without artificial continuity correction. Biostatistics 2009; 10(2): 275-281. https://dx.doi.org/10.1093/biostatistics/kxn034.

742. Tierney JF, Stewart LA, Ghersi D et al. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 2007; 8: 16. https://dx.doi.org/10.1186/1745-6215-8-16.

743. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? BMJ 2001; 322: 355. https://dx.doi.org/10.1136/bmj.322.7282.355.

744. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? Med Decis Making 2001; 21(4): 329-334. https://dx.doi.org/10.1177/0272989X0102100408.

745. Trevena LJ, Davey HM, Barratt A et al. A systematic review on communicating with patients about evidence. J Eval Clin Pract 2006; 12(1): 13-23. https://dx.doi.org/10.1111/j.1365-2753.2005.00596.x.

746. Tricco AC, Langlois EV, Straus SE. Rapid reviews to strengthen health policy and systems: a practical guide [online]. 2017 [Accessed: 05.02.2021]. URL: https://apps.who.int/iris/bitstream/handle/10665/258698/9789241512763-eng.pdf.

747. Trueman P, Drummond M, Hutton J. Developing guidance for budget impact analysis. Pharmacoeconomics 2001; 19(6): 609-621. https://dx.doi.org/10.2165/00019053-200119060-00001.

748. Turner RM, Omar RZ, Yang M et al. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000; 19(24): 3417-3432. https://dx.doi.org/10.1002/1097-0258(20001230)19:24<3417::aid-sim614>3.0.co;2-l.

749. U.S. Food and Drug Administration. Devices@FDA [online]. [Accessed: 01.02.2021]. URL: http://www.accessdata.fda.gov/scripts/cdrh/devicesatfda/index.cfm.

750. U.S. Food and Drug Administration. Drugs@FDA: FDA-Approved Drugs [online]. [Accessed: 01.02.2021]. URL: https://www.accessdata.fda.gov/scripts/cder/daf/.

751. U.S. Food and Drug Administration. Guidance for Industry; Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims [online]. 2009 [Accessed: 01.02.2021]. URL: https://www.fda.gov/media/77832/download.

752. Unkel S, Amiri M, Benda N et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. Pharm Stat 2019; 18(2): 166-183. https://dx.doi.org/10.1002/pst.1915.

753. van Aert RCM, Wicherts JM, van Assen M. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. PLoS One 2019; 14(4): e0215052. https://dx.doi.org/10.1371/journal.pone.0215052.

754. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21(4): 589-624. https://dx.doi.org/10.1002/sim.1040.

755. Van Tinteren H, Hoekstra OS, Boers M. Do we need randomised trials to evaluate diagnostic procedures? For. Eur J Nucl Med Mol Imaging 2004; 31(1): 129-131. https://dx.doi.org/10.1007/s00259-003-1384-x.

756. van Tulder M, Furlan A, Bombardier C et al. Updated Method Guidelines for Systematic Reviews in the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976) 2003; 28(12): 1290-1299. https://dx.doi.org/10.1097/01.BRS.0000065484.95996.AF.

757. Vandenbroucke JP, von Elm E, Altman DG et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007; 147(8): W163-194. https://dx.doi.org/10.7326/0003-4819-147-8-200710160-00010-w1.

758. Vemer P, Corro Ramos I, van Voorn GA et al. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. Pharmacoeconomics 2016; 34(4): 349-361. https://dx.doi.org/10.1007/s40273-015-0327-2.

759. Vemer P, Krabbe PFM, Feenstra TL et al. Improving model validation in HTA: comments on the guidelines of the ISPOR-SMDM Modeling Good Research Practices. Value Health 2013; 16(8): 1174. https://dx.doi.org/10.1016/j.jval.2013.11.001.

760. Vemer P, van Voom GA, Ramos IC et al. Improving model validation in health technology assessment: comments on guidelines of the ISPOR-SMDM modeling good research practices task force. Value Health 2013; 16(6): 1106-1107. https://dx.doi.org/10.1016/j.jval.2013.06.015.

761. Veroniki AA, Jackson D, Bender R et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Res Synth Methods 2019; 10(1): 23-43. https://dx.doi.org/10.1002/jrsm.1319.

762. Veroniki AA, Jackson D, Viechtbauer W et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res Synth Methods 2016; 7(1): 55-79. https://dx.doi.org/10.1002/jrsm.1164.

763. Veroniki AA, Jackson D, Viechtbauer W et al. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. Cochrane Database Syst Rev 2015; (Suppl 1: Cochrane Methods): 25-27. https://dx.doi.org/10.1002/14651858.CD201501.

764. Vidanapathirana J, Abramson MJ, Forbes A et al. Mass media interventions for promoting HIV testing. Cochrane Database Syst Rev 2005; (3): CD004775. https://dx.doi.org/10.1002/14651858.CD004775.pub2.

765. Villar J, Mackey ME, Carroli G et al. 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. https://dx.doi.org/10.1002/sim.1096.

766. Virtanen H, Leino-Kilpi H, Salantera S. Empowering discourse in patient education. Patient Educ Couns 2007; 66(2): 140-146. https://dx.doi.org/10.1016/j.pec.2006.12.010.

767. Viswanathan M, Ansari MT, Berkman ND 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]. 2012 [Accessed: 01.02.2021]. URL: https://www.ncbi.nlm.nih.gov/books/NBK91433/pdf/Bookshelf_NBK91433.pdf.

768. Vlayen J, Aertgeerts B, Hannes K et al. 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. https://dx.doi.org/10.1093/intqhc/mzi027.

769. von Elm E, Altman DG, Egger M et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007; 335(7624): 806-808. https://dx.doi.org/10.1136/bmj.39335.541782.AD.

770. Waffenschmidt S, Navarro-Ruan T, Hobson N et al. Development and validation of study filters for identifying controlled non-randomized studies in PubMed and Ovid MEDLINE. Res Synth Methods 2020; 11(5): 617-626. https://dx.doi.org/10.1002/jrsm.1425.

771. Wallace BC, Noel-Storr A, Marshall IJ et al. Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach. J Am Med Inform Assoc 2017; 24(6): 1165-1168. https://dx.doi.org/10.1093/jamia/ocx053.

772. Walter U, Schwartz FW. Prävention. In: Schwartz FW, Badura B, Busse R et al (Ed). Das Public Health Buch; Gesundheit und Gesundheitswesen. München: Urban und Fischer; 2003. S. 189-214.

773. Wang SV, Pinheiro S, Hua W et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ 2021; 372: m4856. https://dx.doi.org/10.1136/bmj.m4856.

774. Webster-Clark M, Sturmer T, Wang T et al. Using propensity scores to estimate effects of treatment initiation decisions: State of the science. Stat Med 2021; 40(7): 1718-1735. https://dx.doi.org/10.1002/sim.8866.

775. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med 2006; 25(2): 183-203. https://dx.doi.org/10.1002/sim.2319.

776. West S, King V, Carey TS et al. Systems to Rate the Strength of Scientific Evidence; Summary [online]. 2002 [Accessed: 01.02.2021]. URL: https://archive.ahrq.gov/clinic/epcsums/strengthsum.pdf.

777. Wetterslev J, Thorlund K, Brok J et al. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008; 61(1): 64-75. https://dx.doi.org/10.1016/j.jclinepi.2007.03.013.

778. Wetzel H. Mindestmengen zur Qualitätssicherung: Konzeptionelle und methodische Überlegungen zur Festlegung und Evaluation von Fallzahlgrenzwerten für die klinische Versorgung. Z Arztl Fortbild Qualitatssich 2006; 100(2): 99-106.

779. Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull 2010; 96: 5-21. https://dx.doi.org/10.1093/bmb/ldq033.

780. Whiting P, Rutjes AW, Reitsma JB et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140(3): 189-202. https://dx.doi.org/10.7326/0003-4819-140-3-200402030-00010.

781. Whiting P, Savovic J, Higgins JP et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016; 69: 225-234. https://dx.doi.org/10.1016/j.jclinepi.2015.06.005.

782. Whiting PF, Rutjes AW, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011; 155(8): 529-536. https://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009.

783. Whiting PF, Rutjes AW, Westwood ME et al. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013; 66(10): 1093-1104. https://dx.doi.org/10.1016/j.jclinepi.2013.05.014.

784. Whitlock EP, Lin JS, Chou R et al. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008; 148(10): 776-782. https://dx.doi.org/10.7326/0003-4819-148-10-200805200-00010.

785. Widrig D, Tag B. HTA and its legal issues: a framework for identifying legal issues in health technology assessment. Int J Technol Assess Health Care 2014; 30(6): 587-594. https://dx.doi.org/10.1017/S0266462314000683.

786. Wieseler B, Kerekes MF, Vervoelgyi V et al. Impact of document type on reporting quality of clinical drug trials: a comparison of registry reports, clinical study reports, and journal publications. BMJ 2012; 344: d8141. https://dx.doi.org/10.1136/bmj.d8141.

787. Wigge P. Evidenz-basierte Richtlinien und Leitlinien. Medizinrecht 2000; 18(12): 574-585. https://dx.doi.org/10.1007/s003500000381.

788. Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. Stat Med 2016; 35(15): 2503-2515. https://dx.doi.org/10.1002/sim.6879.

789. Williamson E, Morley R, Lucas A et al. Propensity scores: from naive enthusiasm to intuitive understanding. Stat Methods Med Res 2012; 21(3): 273-293. https://dx.doi.org/10.1177/0962280210394483.

790. Windeler J. Bedeutung randomisierter klinischer Studien mit relevanten Endpunkten für die Nutzenbewertung. In: Begriffsdefinitionen und Einführung; Dokumentation des ersten gemeinsamen Workshops von GFR und IQWiG am 4. September 2007 in Berlin. Bonn: Gesundheitsforschungsrat des Bundesministeriums für Bildung und Forschung; 2007. S. 26-31.

791. Windeler J, Conradt C. Wie können "Signifikanz" und "Relevanz" verbunden werden? Med Klin (Munich) 1999; 94(11): 648-651. https://dx.doi.org/10.1007/BF03045008.

792. Windeler J, Lange S. Nutzenbewertung in besonderen Situationen – Seltene Erkrankungen. Z Evid Fortbild Qual Gesundhwes 2008; 102(1): 25-30. https://dx.doi.org/10.1016/j.zgesun.2007.12.005.

793. Windeler J, Ziegler S. Evidenzklassifizierungen. Z Arztl Fortbild Qualitatssich 2003; 97(6): 513-514.

794. Wolff RF, Moons KGM, Riley RD et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019; 170(1): 51-58. https://dx.doi.org/10.7326/M18-1376.

795. Wong SS, Wilczynski NL, Haynes RB. Comparison of top-performing search strategies for detecting clinically sound treatment studies and systematic reviews in MEDLINE and EMBASE. J Med Libr Assoc 2006; 94(4): 451-455.

796. Wong SS, Wilczynski NL, Haynes RB. Developing optimal search strategies for detecting clinically sound treatment studies in EMBASE. J Med Libr Assoc 2006; 94(1): 41-47.

797. Wong SS, Wilczynski NL, Haynes RB. Optimal CINAHL search strategies for identifying therapy studies and review articles. J Nurs Scholarsh 2006; 38(2): 194-199. https://dx.doi.org/10.1111/j.1547-5069.2006.00100.x.

798. Wood L, Egger M, Gluud LL 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. https://dx.doi.org/10.1136/bmj.39465.451748.AD.

799. Yang B, Mallett S, Takwoingi Y et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. Ann Intern Med 2021; 174(11): 1592-1599. https://dx.doi.org/10.7326/M21-2234.

800. Yang B, Olsen M, Vali Y et al. Study designs for comparative diagnostic test accuracy: A methodological review and classification scheme. J Clin Epidemiol 2021; 138: 128-138. https://dx.doi.org/10.1016/j.jclinepi.2021.04.013.

801. Yang F, Wittes J, Pitt B. Beware of on-treatment safety analyses. Clin Trials 2019; 16(1): 63-70. https://dx.doi.org/10.1177/1740774518812774.

802. Zarin DA, Tse T, Williams RJ et al. Trial Reporting in ClinicalTrials.gov - The Final Rule. N Engl J Med 2016; 375(20): 1998-2004. https://dx.doi.org/10.1056/NEJMsr1611785.

803. Zentner A, Busse R. Internationale Standards der Kosten-Nutzen-Bewertung. Gesundheitsökonomie & Qualitätsmanagement 2006; 11(6): 368-373. https://dx.doi.org/10.1055/s-2006-927192.

804. 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. https://dx.doi.org/10.1111/j.1365-2929.2006.02467.x.

805. Ziegler DK, Mosier MC, Buenaver M et al. How much information about adverse effects of medication do patients want from physicians? Arch Intern Med 2001; 161(5): 706-713. https://dx.doi.org/10.1001/archinte.161.5.706.

806. Zorzela L, Loke YK, Ioannidis JP et al. PRISMA harms checklist: improving harms reporting in systematic reviews. BMJ 2016; 352: i157. https://dx.doi.org/10.1136/bmj.i157.

807. 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. https://dx.doi.org/10.1055/s-0030-1261879.

808. Zwarenstein M, Treweek S, Gagnier JJ et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008; 337: a2390. https://dx.doi.org/10.1136/bmj.a2390.