Methods[a]

Version 2.0

Produced by the Institute's Steering Committee[b]

Contact:
Institute for Quality and Efficiency in Health Care
(Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; IQWiG)
Dillenburger Str. 27
51105 Cologne
Germany
Tel.: +49 (0)221 – 35685-0
Fax: +49 (0)221 – 35685-1
E-mail: methoden@iqwig.de

---

[a] Translation based on Version 2.0 of "Methoden" (published on 19 December 2006; publication date of translation: 22 February 2007). Please note: This translation is provided as a service by IQWiG to English-language readers. However, solely the German original text is absolutely authoritative and legally binding.

[b] Bastian H, Bender R, Ernst AS, Kaiser T, Kirchner H, Kolominsky-Rabas P, Lange S, Sawicki PT, Weber M.

General comments:

The first draft of Version 1.0 of the Institute's methods paper was produced in autumn 2004 and published online for discussion on 1 November 2004. Following the receipt of comments and expert opinions, a round table was held in February 2005, including the contributors and some members of the Institute's Scientific Advisory Board. The first version of March 2005 (Version 1.0) was subsequently produced. In 2006, the document was revised, and two successive drafts were put forward for discussion: one internal draft (dated 27 April 2006), and a second draft published on the IQWiG website (dated 28 September 2006). This second version was produced after considering the comments submitted on both drafts (Version 2.0 of 19 December 2006).

The methods paper will in future be reviewed annually with regard to any necessary revisions, unless errors in the document or relevant developments necessitate prior updating.

For every document produced by the Institute, the valid version of the Institute's methods at the time of publication applies in each case.

Preamble

With the introduction of the health care reform in 2003 (Health Care Modernisation Act; *Gesundheits-Modernisierungsgesetz, GMG*), legislation determined the establishment of a new Institute, independent of the state, within the German health care system. In June 2004, the Federal Joint Committee[c] *(Gemeinsamer Bundesausschuss, G-BA)* set up this scientific institution in the form of a non-profit and non-government private law foundation that has legal capacity. The sole purpose of the foundation is the creation and maintenance of the Institute for Quality and Efficiency in Health Care *(Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, IQWiG)*. The foundation's bodies include a Foundation Board and a five-member Board of Directors. The Institute is an establishment of the foundation and is under independent scientific management. The Institute's advisory committees are a 30-member Board of Trustees and a Scientific Advisory Board. The Scientific Advisory Board is appointed by the Board of Directors, and comprises 6 to 12 members. Until 2008, the seat of the Institute will be Cologne.

The Institute is responsible for the scientific evaluation of the benefits and harms as well as the quality and efficiency of health care services. This includes the evaluation of clinical practice guidelines, the submission of recommendations on disease management programmes, the evaluation of the benefits and harms of pharmaceuticals, and the publication of health information for patients and consumers.

The Institute addresses fundamental issues relating to the quality and efficiency of services provided by statutory health insurance (SHI), taking into account specific factors such as age, gender, and living conditions. The Institute, as the foundation's professionally independent scientific establishment, is particularly active in the following areas:

- The investigation, evaluation, and presentation of the current state of medical knowledge on diagnostic tests and therapeutic techniques for selected diseases;
- The production of scientific reports, expert opinions and statements on the quality and efficiency of services provided by statutory health insurance;
- The evaluation of evidence-based guidelines for the epidemiologically most important diseases;
- The submission of recommendations on disease management programmes;

[c] The Federal Joint Committee is the decision-making entity of the self-governing body of the German health care system. More information on the Committee's responsibilities is provided under http://www.g-ba.de/cms/front_content.php?idcatart=207&lang=1&client=1

- The evaluation of the benefits and harms of pharmaceuticals;
- The provision of easily understandable general information on the quality and efficiency of health care to the public.

The Institute's responsibility in these areas is to support the Federal Joint Committee in fulfilling its legislative duties by submitting recommendations, and to contribute to continuous improvement in the quality of health care for the public. The Institute's aim is to develop the independent scientific capacity to answer the research questions posed, to evaluate medical issues and concepts relevant to health care, and to assess research requirements relevant to patients' needs. The information compiled is relayed to the Federal Ministry of Health, the Federal Joint Committee, and the public.

The Institute fulfils its duties by producing reports on specific topics requested by the Federal Joint Committee or the Federal Ministry of Health. The Institute also initiates, coordinates, and publishes scientific work in areas where health care knowledge needs to be complemented. To this end, the Institute regularly screens and evaluates literature for innovations related to health care and distributes this information in an understandable form. On the basis of international literature and its own investigations, the Institute can develop proposals for research related to innovative health care, initiate and participate in research projects, and publish the results of these projects.

The Institute's Steering Committee includes the Institute's Management and the Department Heads. This Committee produces and modifies the methods paper and develops and modifies the Institute's working procedures. The methods are published to make the Institute's work transparent, and discussion of the methods is explicitly desired in order to achieve continuous improvement.

Not all steps in an evaluation process can be presented in advance and in detail in every case. Individual procedures are, amongst other things, dependent on the particular research question, the scientific evidence available, and any comments received. This document should therefore be regarded as a guideline when evaluating a medical intervention. The evaluation procedure referring to each commission is developed and presented in the particular report plan (protocol) and preliminary report.

In order to use the available resources meaningfully and efficiently, the Institute considers and, if applicable, makes use of the work conducted previously by other national and international health care institutions.

Table of contents

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. Scientific methods and statistics

1.1 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 strength 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 (possibly after appropriate weighting). For categorical outcome variables, the usual effect and risk measures of 2x2 tables apply [1]. After specification of a primary effect measure for data analysis, the Institute will, if possible, use both absolute measures (e.g. absolute risk reduction or number needed to treat) and also relative measures (e.g. relative risk or odds ratio) for the descriptive presentation of an effect. Chapter 8 of the Cochrane Reviewers' Handbook provides a well-structured summary of the advantages and disadvantages of typical effect measures [2]. Agresti describes the specific aspects to be considered for ordinal data [3,4].

It is mandatory to describe the degree of statistical uncertainty for every effect estimate. The calculation of the standard error and the confidence interval are methods frequently employed for this purpose. 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. give an overview of the most common methods for calculation of confidence intervals [5].

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 provides an up-to-date discussion on exact methods [6].

References

[1] Bender R. Interpretation von Effizienzmaßen der Vierfeldertafel für Diagnostik und Behandlung [Interpretation of efficiency measures of the 2x2 table for diagnosis and treatment]. Med Klin 2001; 96: 116-121. Erratum: Med Klin 2001 96: 181.

[2] Deeks JJ, Higgins JPT, Altman DG. Analysing and presenting results. In: Higgins JPT, Green, S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]; Section 8. In: The Cochrane Library, Issue 3, 2005. Chichester: Wiley; 2005.

[3] Agresti A. Categorical Data Analysis, 2nd Ed. New York: Wiley; 2002.

[4] Agresti A. Modelling ordered categorical data: Recent advances and future challenges. Stat Med 1999; 18: 2191-2207.

[5] Altman DG, Machin D, Bryant TM, Gardner MJ, editors. Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd Ed. London: BMJ Books; 2000.

[6] Agresti A. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. Stat Meth Med Res 2003; 12: 3-21.

1.2 Evaluation of statistical significance

With the help of statistical significance tests, it is possible to test hypotheses formulated a priori with control for type 1 error probability. The convention of speaking of a "statistically significant result" when the *p*-value is lower than the significance level of 0.05 ($p<0.05$) may often be meaningful. It may be necessary to demand a much lower significance level, depending on the research question posed and hypothesis formulated. Conversely, there are situations where a higher significance level is acceptable. The Institute will always explicitly justify such exceptions.

A range of aspects should be considered when interpreting *p*-values. It must be clear in detail which research question and data situation the significance level refers to and how the statistical hypothesis is formulated. In particular, it should be clear whether a one- or two-sided hypothesis applies [1] and whether this hypothesis is to be regarded as part of a multiple hypothesis problem [2]. These two aspects, whether a one- or two-sided hypothesis is to be formulated, and whether adjustments for multiple testing need to be made, are the subject of continual controversy in scientific literature.

Regarding the hypothesis formulation, a two-sided test problem is traditionally assumed. Exceptions include non-inferiority studies (q.v. Section 1.20). The formulation of a one-sided hypothesis problem is, in principle, always possible, but it requires precise justification. In the case of a one-sided hypothesis formulation, the application of one-sided significance tests and the calculation of one-sided confidence limits are appropriate. For better comparability with two-sided

statistical methods, some guidelines for clinical studies demand that the typical significance level should be halved from 5% to 2.5% [3]. The Institute follows the central principle that the hypothesis formulation (one- or two-sided) and the significance level must be specified clearly a priori. In addition, the Institute will justify deviations from the usual specifications (one-sided instead of two-sided hypothesis formulation, significance level unequal to 5%, etc.) or consider the relevant justifications in the primary literature.

If the investigated hypothesis is clearly part of a multiple hypothesis problem, appropriate adjustment for multiple testing is required. Bender and Lange [4] provide an overview of the situations where this case applies and describe the methods available for this purpose. If meaningful and possible, the Institute will apply methods to adjust for multiple testing.

The Institute does not evaluate a statistically non-significant finding as evidence of the absence of an effect (absence or equivalence) [5]. For evidence of equivalence, the Institute will apply appropriate methods for equivalence hypotheses (q.v. Section 1.20).

In principle, Bayesian methods may be regarded as an alternative to statistical significance tests [6,7]. Depending on the question posed, the Institute will, where necessary, also employ Bayesian methods (q.v. Section 1.14).

References

[1] Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994; 309: 248.

[2] Tukey JW. Some thoughts on clinical trials, especially problems of multiplicity. Sc 1977; 198: 679-684.

[3] ICH E9 Expert Working Group. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials. Stat Med 1999; 18: 1905-1942.

[4] Bender R, Lange S. Adjusting for multiple testing – when and how? J Clin Epidemiol 2001; 54: 343-349.

[5] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311: 485.

[6] Spiegelhalter DJ, Freedman LS. Bayesian approaches to randomised trials. J. R. Stat. Soc. A 1994; 157: 357-416.

[7] Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. BMJ 1999; 319: 508-512.

1.3 Evaluation of clinical relevance

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 strength of a possible effect but also by data variability and sample size. When interpreting the relevance of *p*-values, in particular the sample size of the underlying study needs to be taken into account [1]. 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 [2,3]. Consequently, the clinical relevance of a study's results can by no means be derived from a *p*-value alone.

Widely accepted methodological procedures to evaluate the clinical relevance of study results do not yet exist. Only a few guidelines contain details of the definition of clinically relevant or irrelevant differences between groups. A first approach to assess the clinical relevance of study findings is the evaluation of the effect estimate and of the corresponding confidence interval (q.v. Section 1.1) using medical expertise. A formal relevance criterion may be the assessment of the lower confidence limit (in the case of favourable effects) for the effect estimate, or the employment of a statistical test shifting the null hypothesis in order to detect clinically relevant effects [4]. A further option is to formulate a relevance criterion individually, e.g. in terms of a responder definition [5]. Moreover, the individual assessment of affected patients plays an important role; the presentation of patient-relevant endpoints may provide indications for this purpose (q.v. Section 2.1). The Institute will perform the evaluation of clinical relevance in a problem-orientated manner, taking these aspects into account.

References

[1] Royall RM. The effect of sample size on the meaning of significance tests. Am Stat 1986; 40: 313-315.

[2] Feinstein AR. Invidious comparisons and unmet clinical challenges. Am J Med 1992; 92: 117-120.

[3] Hung HMJ, O'Neill RT, Bauer P, Köhne K. The behavior of the *P*-value when the alternative hypothesis is true. Biometrics 1997; 53: 11-22.

[4] Windeler J, Conradt C. Wie können „Signifikanz" und „Relevanz" verbunden werden? [How can "significance" and "relevance" be connected?]. Med Klin 1999; 94: 652-655.

[5] Kieser M, Röhmel J, Friede. Power and sample size determination when assessing the clinical relevance of trial results by 'responder analyses'. Stat Med 2004; 23: 3287-3305.

1.4 Subgroup analyses

Subgroup analyses are discussed very critically in the methodological literature [1,2]. The interpretation of their results is mainly complicated by three factors:

- No characteristic of proof: subgroup analyses are rarely planned a priori, and are rarely a component of the study protocol (or its amendments). If subgroup analyses are conducted "post hoc", the corresponding results cannot be seen as a methodologically correct testing of a hypothesis.
- Multiple testing: if several subgroups are analysed, there is sometimes a rather high probability, depending on the significance level, that the result of a subgroup is statistically significant, even though it is actually a random result. Therefore, as in other situations where a problem of multiple testing exists, the significance level must be adjusted appropriately (q.v. Section 1.2).
- Lack of power: even if a result in a subgroup is not statistically significant, this is not a reliable finding. The sample size of a subgroup is often too small to enable the detection of moderate differences (by means of inferential statistics). 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 [3].

If one or more of the three factors mentioned above are present, the results of subgroup analyses should only be considered in the evaluation with strong reservations and should not dominate the results of the primary analysis, especially if the primary study objective was not achieved.

Furthermore, subgroup analyses are not interpretable if the subgroup-forming characteristic was defined after initiation of treatment (after randomisation), e.g. so-called responder analyses.

The statistical demonstration of different effects between various subgroups should be conducted by means of an appropriate homogeneity or interaction test. The finding that a statistically significant effect was observed in one subgroup, but not in another, cannot be interpreted (by means of inferential statistics) as the existence of a subgroup effect.

Despite the limitations noted above, for some research questions subgroup analyses may represent the best scientific evidence available in the foreseeable future for the evaluation of effects in

subgroups [4], as factors such as ethical considerations may argue against the reproduction of findings of subgroup analyses in a validation study. Rothwell presents an overview of indications for applying subgroup analyses [5]. A possible heterogeneity of an effect in different clearly distinguishable patient populations is an important indication for conducting subgroup analyses [5,6]. If there is information a priori about a possible effect modifier (e.g. age, pathology), it is even necessary to investigate in advance possible heterogeneity with regard to the effect in the various patient populations. If such heterogeneity exists, then the estimated total effect across all patients cannot be interpreted meaningfully [6]. It is therefore important that information on a possible heterogeneity of patient groups is considered appropriately in the study design. It may even be necessary to conduct several studies [7].

The gold standard for a subgroup analysis is an analysis in which the subgroup was specified a priori. This approach includes the use of stratified randomisation on the basis of subgroups, and the employment of an appropriate statistical method (homogeneity test, interaction test) for the particular data analysis [8].

Taking into account the above-named factors, the Institute interprets results of subgroup analyses very cautiously. However, it does not exclude them from the evaluation as a matter of principle.

References

[1] Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116: 78-84.

[2] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355: 1064-1069.

[3] Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57: 229-236.

[4] 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.

[5] Rothwell PM. Treating individuals 2: Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365: 176-186.

[6] Kraemer HC, Frank E, Kupfer DJ. (2006): Moderators of treatment outcomes: Clinical, research, and policy importance. JAMA 2006; 296: 1286-1289.

[7] Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: Statistical and regulatory issues. J Biopharm Stat 2005; 15: 869-882.

[8] Cui L, Hung HM, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. J Biopharm Stat 2002; 12: 347-358.

1.5 Evaluation of study quality

When assessing the overall quality of studies, a range of aspects play a role. In principle, a recognised standardised concept should be followed in a study, from planning to conducting, evaluation, and presentation. This includes a study protocol describing all the important methods and procedures. The usual standards for studies are defined by the basic principles of good clinical practice (GCP) in randomised clinical trials [1,2] and by the guidelines and recommendations to ensure good epidemiological practice (GEP) in epidemiological studies [3]. A key quality criterion in studies is whether the study data were 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. Furthermore, a comparison with the (possibly previously published) study protocol or with the corresponding publication on the study design may be useful.

The following relevant statements were formulated to improve the quality of publications:

- The CONSORT[d] statement on randomised clinical trials [4] and the corresponding explanatory document [5];
- The CONSORT statement on cluster randomised trials [6];
- The QUORUM[e] statement on meta-analyses of randomised trials [7];
- The TREND[f] statement on non-randomised intervention trials [8];
- The STROBE[g] statement for observational studies in epidemiology [9];
- The MOOSE[h] checklist for meta-analysis of observational studies in epidemiology [10];
- The STARD[i] statement on diagnostic studies [11] and the corresponding explanatory document [12].

If a publication fails to conform to these standards, this may be an indication of deficiencies in the relevant study. Additional key papers on this issue describe fundamental aspects of the assessment of the quality of studies [13-15].

[d] Consolidated Standards of Reporting Trials
[e] Quality of Reporting of Meta-analyses
[f] Transparent Reporting of Evaluations with Nonrandomized Designs
[g] Strengthening the Reporting of Observational Studies in Epidemiology
[h] Meta-analysis of Observational Studies in Epidemiology
[i] Standards for Reporting of Diagnostic Accuracy

Various systems, such as the quality index by Chalmers et al. [16], have been developed to support the quality assessment of studies. Moher et al. [17] provide an overview of systems (scales and checklists) to assess the quality of randomised trials. West et al. [18] provide a general overview of systems to assess studies with different designs. However, in practice such systems need to be applied with caution [17], as the application of different systems to the same study pool may lead to varying results with regard to the quality grading of studies and the respective conclusions inferred [19]. Currently, no uniform and generally valid formal system is available for assessing study quality [17,18]. The Institute will therefore perform the evaluation of study quality in a problem-orientated manner, following the sources quoted above [1-15].

The following principles are key aspects in the evaluation of randomised controlled trials (RCTs) by the Institute: adequate concealment, i.e. the unforeseeability and concealment of allocation to groups (e.g. by external randomisation in trials that cannot be [double] blinded); blinded evaluation of outcome measures in trials that cannot be (double) blinded (q.v. Section 1.15); appropriate application of the "intention-to-treat" principle; determination of a clear primary endpoint; and appropriate consideration of possible multiple testing problems (q.v. Section 1.2).

The evaluation of formal criteria provides essential indications for the quality of studies. However, the Institute will always perform an evaluation beyond purely formal aspects, for example, in order to demonstrate errors, contradictions and inconsistencies in publications and assess their relevance in the interpretation of results.

References

[1] Kolman J, Meng P, Scott G. Good Clinical Practice. Standard Operating Procedures for Clinical Researchers. Chichester: Wiley; 1998.

[2] ICH Steering Committee. Official web site for the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).
http://www.ich.org/ [accessed on 20.10.2004].

[3] Arbeitsgruppe Epidemiologische Methoden der Deutschen Arbeitsgemeinschaft für Epidemiologie (DAE). Leitlinien und Empfehlungen zur Sicherung von Guter Epidemiologischer Praxis (GEP). Mit Änderungen nach Evaluation. [Epidemiological Methods Task Group of the German Working Group on Epidemiology. Guidelines and recommendations to ensure Good Epidemiological Practice, GEP; with amendments after evaluation]. April 2004.
http://www.dgepi.de/doc/Empfehlungen.doc [accessed on 17.11.2006].

[4] Moher D, Schulz KF, Altman DG for the CONSORT Group. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001; 134: 657-662.

[5] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne DR et al. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. Ann Intern Med 2001; 134: 663-694.

[6] Campbell MJ, Elbourne DR, Altman DG for the CONSORT Group. CONSORT statement: Extension to cluster randomised trials. BMJ 2004; 328: 702-708.

[7] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF et al. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Lancet 1999; 354: 1896-1900.

[8] Des Jarlais DC, Lyles C, Crepaz N for the TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. Am J Public Health 2004; 94: 361-366.

[9] STROBE statement: Checklist of essential items Version 3 (Sept 2005). http://www.strobe-statement.org/PDF/STROBE-Checklist-Version3.pdf [accessed on 31.10.2005].

[10] Stroup DF, Berlin IA Morton SC. Meta-analysis of observational studies in epidemiology: A proposal for reporting. JAMA 2000; 283: 2008-2012.

[11] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. Ann Intern Med 2003; 138: 40-44.

[12] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Ann Intern Med 2003; 138: W1-12.

[13] Egger M, Davey Smith G, Altman DG, editors. Systematic Reviews in Health Care: Meta-Analysis in Context. London: BMJ Books; 2001.

[14] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 (updated May 2005). In: The Cochrane Library, Issue 3, 2005. Chichester: Wiley; 2005.

[15] Guyatt G, Rennie D, editors. Users' Guide to the Medical Literature. Chicago, IL: AMA Press; 2002.

[16] Chalmers TC, Smith H, Blackburn B, Silverman B, Schroder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981; 2: 31-49.

[17] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 1995; 16: 62-73.

[18] West S, King V, Carey TS, Lohr, K.N., McKoy N., Sutton, S.F., Lux, L. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality. April 2002. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.70996 [accessed on 13.03.2006].

[19] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 262: 1054-1060.

1.6 Determination of the damage potential of medical interventions

The application of a medical intervention, whatever its nature (pharmaceutical, non-pharmaceutical, surgical, diagnostic, preventive, etc.), always carries the risk of adverse effects. The term "adverse effects" signifies all events and effects that are individually perceived or objectively identifiable physical or mental detriments. These detriments may cause a mild to severe, short- or long-term reduction in life expectancy, increase in morbidity, or impairment in quality of life. In this context, an adverse effect is defined as an effect where a causal relationship to the intervention is assumed, whereas an adverse event is defined as an event where a causal relationship to the intervention may or may not exist [1].

The term "damage potential" describes the risk of the occurrence of adverse effects when applying a medical intervention. The description of this damage potential is an essential element of equal value in the evaluation of the benefits and harms of an intervention (see Section 2.1). It ensures an informed, population-related, but also individual weighing of benefits and harms [2]. A prerequisite for this is that, from the data available, the strength of the effects of a medical intervention can be described both for its desired as well as its adverse effects, and be compared with other data, for example, on therapeutic alternatives.

However, within the framework of a systematic review, the description, analysis, and evaluation of the potential damage of a medical intervention is often far more difficult to provide than a description of the benefits (q.v. Section 2.1). This particularly applies to unexpected adverse events [1]. Studies are typically designed to measure the effect of an intervention on a few predefined efficacy endpoints. In such studies, the results for adverse events strongly depend on the underlying methodology of how these events were recorded [3,4]. It should also be noted that studies with the specific objective of detecting rare, serious adverse effects (including the description of a causal relationship to the medical intervention) are considerably underrepresented in medical research [5-7]. Furthermore, in single studies, the quality of reporting of adverse events is poor, which recently led to an amendment to the CONSORT statement for randomised clinical trials [8]. Finally, the systematic evaluation of adverse events of an intervention is also made difficult by the fact that literature database coding in this regard is insufficient; therefore, the specific search for relevant scientific literature often produces an incomplete picture [9].

The consequence of the above-mentioned obstacles is that in many cases, in spite of enormous efforts, the uncertainty of statements on the damage potential of an intervention is greater than that of statements on positive effects [10]. It is necessary to find a meaningful balance, on the one hand, between the completeness of the investigation of adverse effects and, on the other, the amount of

resources required; consequently, it is necessary to limit the investigation and presentation to relevant adverse effects. In particular, adverse effects can be described as relevant that:

- May completely or almost completely eliminate the benefits of an intervention;
- May considerably differ from adverse effects occurring with (an) otherwise equivalent treatment option(s);
- May occur predominantly with (a) treatment option(s) (of several treatment options) that are particularly effective;
- May have a dose-effect relationship;
- May be regarded by patients as especially important;
- May be accompanied by serious morbidity or even increased mortality or may be associated with substantial impairment of quality of life.

In the interests of patient safety and the medical axiom “primum nil nocere”, the Institute observes the following principles when assessing and describing adverse effects:

- The basis for the selection of relevant adverse effects according to the above-mentioned criteria is a compilation of adverse effects and events that are essential in the decision-making for or against the application of the intervention to be evaluated. This compilation is made within the framework of the preliminary literature search for the particular question posed, especially on the basis of data from controlled intervention studies in which the benefit of the intervention was specifically investigated. In addition, it is made on the basis of available epidemiological data (e.g. from cohort or case-control studies), pharmacovigilance and regulatory data, etc. In individual cases, data obtained from animal trials and experiments to test pathophysiological constructs may be useful.
- If indications of the presence of an adverse effect emerge from the above-mentioned data sources, the occurrence of such an effect will be regarded as possible until it can be ruled out with sufficient certainty by the results of specific research. In particular, this applies to serious adverse effects. The hierarchy of evidence corresponds to that of therapeutic studies (q.v. Sections 1.9 and 1.10). “Sufficient safety” can, for example, mean that the corresponding study or studies, in design and planning, were aimed primarily at showing the non-inferiority of the intervention to be evaluated compared with other treatment options (or placebo or no intervention, depending on the research question posed), and that the study or studies include(s) an appropriate definition of non-inferiority.

References

[1] Council of Europe: Committee of Experts on Management of Safety and Quality in Health Care (SP-SQS). Expert Group on Safe Medication Practices. Glossary of terms related to patient and medication safety. http://www.who.int/entity/patientsafety/highlights/COE_patient_and_medication_safety_gl.pdf [accessed on 17.9.2006].

[2] Ziegler DK, Mosier MC, Buenaver M, Okuyemi K. How much information about adverse effects of medication do patients want from physicians? Arch Intern Med 2001; 161: 706-713.

[3] 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: 257-261.

[4] Ioannidis JPA, Mulrow CD, Goodman SN. Adverse events: The more you search, the more you find. Ann Intern Med 2006; 144: 298-300.

[5] Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: Lessons from acute postoperative pain. J Pain Symptom Manage 1999; 18: 427-437.

[6] Ioannidis JPA, Lau J. Completeness of safety reporting in randomized trials – an evaluation of seven medical areas. JAMA 2001; 285: 437-443.

[7] Bonhoeffer J, Zumbrunn B, Heininger U. Reporting of vaccine safety data in publications: Systematic review. Pharmacoepidemiol Drug Saf 2005; 14: 101-106.

[8] Ioannidis JPA, Evans SJW, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 2004; 141: 781-788.

[9] 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.

[10] Loke YK, Price D, Herxheimer A on behalf of the Cochrane Adverse Effects Subgroup. Including adverse effects. In: Higgins JPT, Green S, editors: Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]; Appendix 6b. In: The Cochrane Library, Issue 3, 2005. Chichester: Wiley; 2005.

1.7 Evaluation of studies conducted with outdated methods

If an Institute's project includes the evaluation of older studies that do not satisfy current quality standards (q.v. Section 1.8) 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 will, however, only take place if this is clearly justifiable from the research question posed or other circumstances of the evaluation.

1.8 Evaluation of different types of studies

Only the most relevant study designs that play a role in the evaluation of benefits and harms of interventions in medical research are summarised here (q.v. Section 2.1). A distinction is made between observational studies and intervention studies. Observational studies often provide the first information on a topic via case reports or case series. These are susceptible to all kinds of bias, so that evidence on a specific research question can only be inferred to a limited extent (q.v. Section 1.9). The prevalence of diseases can be estimated from population-based cross-sectional studies. Other important epidemiological study types are case-control studies [1], where exposures in cases and controls are assessed retrospectively, and cohort studies [2], where specific groups (cohorts) are observed over a period of time. Cohort studies are prospective in character, although retrospective cohort studies are also conducted where past exposure is recorded (frequently used in studies on occupational or pharmacological epidemiology). In principle, prospective designs are preferable to retrospective designs. However, case-control studies are frequently the only practicable way of gaining information about an association between exposures and rare diseases.

Intervention studies require a control group. In a design with dependent samples without a control group, the effect of an intervention cannot usually be inferred from a sole "before/after" comparison. Exceptions include diseases with a deterministic (or practically deterministic) course (e.g. ketoacidotic diabetic coma). Randomisation and blinding are quality criteria that increase the evidential value of controlled studies (q.v. Sections 1.5 and 1.18). Parallel group studies [3], cross-over studies [4], and cluster randomised studies [5] are common designs in clinical trials. The use of appropriate sequential designs should be considered if interim analyses are planned [6].

The choice of an appropriate design in diagnostic and screening studies depends on their objectives, which may differ substantially (q.v. Sections 2.3 and 2.4).

In the last few years, the relatively new discipline of genetic epidemiology has emerged for the investigation of genetic factors that can cause the development and distribution of diseases [7]. In this field, there is a range of new, specific designs for genetic association and genetic coupling studies.

References

[1] Breslow NE, Day NE. Statistical Methods in Cancer Research Vol. I: The Analysis of Case-Control Studies. Lyon: Int. Agency for Res. on Cancer; 1980.

[2] Breslow NE, Day NE. Statistical Methods in Cancer Research Vol. II: The Design and Analysis of Cohort Studies. Lyon: Int. Agency for Res. on Cancer; 1987.

[3] Pocock SJ. Clinical Trials: A Practical Approach. Chichester: Wiley; 1983.

[4] Jones B, Kenward MG. Design and Analysis of Cross-Over Trials. London: Chapman & Hall; 1989.

[5] Donner A, Klar J. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold; 2000.

[6] Whitehead J. The Design and Analysis of Sequential Clinical Trials. Chichester: Ellis Horwood; 1983.

[7] Khoury MJ, Beaty TH, Cohen BH. Fundamentals of Genetic Epidemiology. New York: Oxford University Press; 1993.

1.9 Ranking of different study types/evidence levels

Within the framework of systematic reviews or the development of guidelines, different approaches exist to allocate specific evidence levels to specific study types. These levels can be used to create a ranking with regard to the validity of evidence from the different study types [1,2]. However, no system currently exists that is generally accepted and universally applicable to all systematic reviews [3]. Due to the complexity of the evaluation of studies (q.v. Section 1.5), no conclusive judgement on quality can be inferred from the hierarchy of evidence. In general, the Institute follows the rough hierarchy of study types described below, which is widely accepted and is also widely consistent with the evidence classification of the Federal Joint Committee [4]. At least within the framework of therapeutic studies, the highest evidence level is allocated to systematic reviews of RCTs. Individual RCTs are ranked next, which in some classifications are further graded into RCTs of higher or lower quality; however, the mixing of the quality of concept and the quality of results has been criticised by some authors [5]. The following levels include non-randomised intervention studies, prospective observational studies, retrospective observational studies, non-experimental studies (e.g. case reports and case series) and, with the lowest evidence level, expert opinions not based on scientific rationale. The Institute will adapt this rough grading system to the particular situation and research question and, if necessary, describe it in more detail [2].

References

[1] Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. JAMA 1995; 274: 1800-1804.

[2] Harbour R, Miller J, for the Scottish Intercollegiate Guidelines Network Grading Review Group. A new system for grading recommendations in evidence based guidelines. BMJ 2001; 323: 334-336.

[3] Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar VSS, Grimme KA. A systematic review of the content of critical appraisal tools. BMC Med Res Methodol 2004; 4: 22.

[4] Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses vom 20. September 2005, veröffentlicht im Bundesanzeiger 2005, S. 16998, in Kraft getreten am 1. Oktober 2005, zuletzt geändert am 18. April 2006, veröffentlicht im Bundesanzeiger 2006, S. 4876, in Kraft getreten am 7. Juli 2006. [Code of Procedure of the Federal Joint Committee dated 20.09.2005; published in the German Federal Gazette 2005 p. 16998; became operative on 01.10.2005. Last changed on 18.04.2006; published in the German Federal Gazette 2006, p. 4876; became operative on 07.07.2006]. http://www.g-ba.de/cms/upload/pdf/richtlinien/2006-07-07-VerfO.pdf [accessed on 30.11.2006].

[5] Windeler J, Ziegler S. EBM-Splitter: Evidenzklassifizierungen. [EBM notes - evidence classifications]. Z ärztl Fortbild Qual sich 2003; 97: 513-514.

1.10 Relationship between study type and research question

The RCT is regarded as the study type of the highest quality. However, this must be seen in connection with the research question posed. This design is usually only required when the objective of the study is to demonstrate the efficacy or effectiveness of an intervention. Alternative study designs may be considered for other research questions; the most important ones are described below. In many cases, a cross-sectional study is sufficient to investigate discrimination ability of diagnostic methods (q.v. Section 2.3). The optimum design to investigate prognostic factors is a prospective cohort study. Case-control studies are used to investigate the association between exposures and very rare diseases. However, if diagnostic tests or prognostic factors are assessed as a strategy together with the consequences resulting from the information gained (e.g. initiation of a therapy), then an RCT is the design of choice (q.v. Sections 2.3, 2.4, and 2.10).

1.11 Evaluation of unpublished or partially published data

In practice, the problem frequently arises that essential data or information is partially or totally missing for the evaluation of publications. This mainly concerns so-called “grey” literature and abstracts, but also full publications. Moreover, it is possible that studies have not (yet) been published at the time of the evaluation of a technology by the Institute.

It is the Institute's aim to conduct an evaluation on the basis of a data set that is as complete as possible (q.v. Section 4.7.3). 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 sponsors of studies. However, depending on the type of product generated (q.v. Section 4), the possibility of requesting unpublished information may be restricted due to time limits.

A common problem is that important information required for the conduct of a meta-analysis (e.g. variances of effect estimates) is missing. However, in many cases the missing data can be calculated or at least estimated from the information available [1-3]. The Institute will apply such procedures as far as possible.

If data are only partly available or if estimated values are used, where appropriate, the robustness of the results will be investigated and discussed with the help of sensitivity analyses, for example, by means of best-case and worst-case scenarios. However, a worst-case scenario can only be used here as evidence of the robustness of a detected effect. It cannot be safely inferred from a worst-case scenario in which a previously detected effect could not be confirmed that such an effect was not demonstrated. In cases where relevant information is largely or completely lacking, it may occur that a publication cannot be evaluated. In such cases, it will merely be noted that further data exist on a particular topic, but are unavailable for evaluation.

References

[1] Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998; 17, 2815-2834.

[2] 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.

[3] Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. Stat Med 2006; 25: 2299-2322.

1.12 Evaluation of the consistency of published data

To assess the evidential value of published results, the Institute will review the consistency of the data (plausibility, completeness). Implausible data are produced, on the one hand, by faulty presentation of results (typing, formatting, or calculation errors), but also by insufficient or incorrect description of the methodology, and even by forged or invented data [1]. Inconsistencies may exist within a publication, and also between publications on the same study.

One problem with many publications is the provision of incomplete information in the methods and results sections. In particular, the description of lost-to-follow-up patients, withdrawals, etc. as well as the way they are considered in the evaluation is often not transparent.

It is therefore necessary to expose potential data inconsistencies. To this end, the Institute, for example, assesses the calculation steps taken, and compares the data presented in text, tables, and graphs. In practice, a common problem of survival-time analyses are inconsistencies between the data on lost-to-follow-up patients and patients at risk in the graphic presentation of survival curves. For certain endpoints (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 to expose forged and invented data [1].

As a rule, if relevant inconsistencies are found in the presentation of results, the aim of the Institute will be to clarify these inconsistencies and/or obtain any missing information, for example, by contacting the particular authors or requesting the complete clinical study report and any additional study documentation. However, it should be considered that firstly, enquiries to the authors often remain unanswered, especially if the publication was produced some time ago, and that secondly, authors' responses may result in further inconsistencies. In individual cases, a weighing of the effort involved and the benefits of such enquiries is therefore meaningful and necessary. If inconsistencies cannot be clarified, the potential influence of these inconsistencies on the effect strengths (magnitude of bias), the uncertainty of results (increase in error probability), and the precision (width of the confidence intervals) will be assessed by the Institute. To this end, sensitivity analyses may be conducted. If the possibility exists that inconsistencies may have a relevant influence on the results, this will be stated and the results will be interpreted with great reservation.

References

[1] Al-Marzouki S, Evans S, Marshall T Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005; 331: 267-270.

1.13 Handling of raw data

For the scientific evaluation of medical services, one of the Institute's principal tasks is to collect and analyse published data from systematic literature searches. For certain research questions, the

Institute may also evaluate raw data provided by external sources (e.g. health insurance funds) that have not previously been analysed. A prerequisite for a meaningful analysis of these data is that the framework within which these data were obtained is clear, and that the plausibility and quality of the data can be reviewed. It is especially important to ensure that essential quality criteria are observed; e.g. for therapeutic studies, the data should have been generated according to GCP standards (q.v. Section 1.5). Furthermore, in most cases the provision of a study protocol is necessary for an appropriate evaluation. Legal aspects of data protection are taken into account when handling raw data (q.v. Section 1.16).

1.14 Evaluation of the uncertainty of results

In principle, every result of an empirical study or a systematic review of empirical studies is uncertain. In this context, one distinguishes between qualitative and quantitative uncertainty of results. Qualitative uncertainty is determined by the study design, from which evidence levels can be inferred (q.v. Section 1.9), as well as by the study quality, which needs to be evaluated depending on the study design (q.v. Section 1.5). In systematic reviews, the quality of the search strategy, as well as possibly the choice of the meta-analytical procedure employed to summarise data, also play a role (q.v. Section 1.21).

In addition to the qualitative uncertainty of results, measurable quantitative uncertainties exist due to statistical principles. The statistical uncertainty of a parameter estimate, which results from the limited sample size, can be quantified and assessed by means of standard errors and confidence intervals. Whenever possible, an appropriate confidence interval should be stated, including information on whether one- or two-sided confidence limits apply and on the confidence level chosen (q.v. Section 1.1).

However, one should not overlook the fact that these calculations are made on the assumption that the statistical method selected is correct and that no other systematic errors and biases exist. The uncertainties that arise because the actual conditions deviate more or less widely from the statistical model chosen remain unconsidered here [1,2]. Formal approaches exist that take these general model uncertainties into account, e.g. Bayesian methods [3] or simulation techniques [4], but they have not been sufficiently developed and investigated to be routinely applied in practice [5-7]. If required, the Institute will, however, consider the employment of these methods. In any case, if necessary, a qualitative assessment of the general uncertainty of results will be performed on the

basis of the current literature on a particular topic. Hill's classic causality criteria [8] are still a valid aid for this purpose.

References

[1] Chatfield C. Model uncertainty, data mining and statistical inference (with discussion). J R Stat Soc A 1995; 158: 419-466.

[2] Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics 1997; 53: 603-618.

[3] Draper D. Assessment and propagation of model uncertainty (with discussion). J R Stat Soc B 1995; 57: 45-97.

[4] Phillips CV. Quantifying and reporting uncertainty from systematic errors. Epidemiology 2003; 14: 459-466.

[5] Hjort NL, Claeskens G. Frequentist model average estimators. J Am Stat Assoc 2003; 98: 879-899.

[6] Wang D, Zhang W, Bakhai A. Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression. Stat Med 2004; 23: 3451-3467.

[7] Augustin N, Sauerbrei W, Schumacher M. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. Stat Modelling 2005; 5, 95-118.

[8] Hill AB. The environment and disease: Association or causation? Proc R Soc Med 1965; 58: 295-300.

1.15 Evaluation of non-blindable techniques

To avoid systematic deviations (bias) of an effect estimate from the true effect (q.v. Section 1.18), controlled studies are, if possible, conducted in a randomised and double-blind way. However, in some situations, blinding of the intervention is not possible. Non-blinded studies can also lead to interpretable results; again, randomisation and the appropriate choice of outcome variables are important instruments to prevent bias. In studies that cannot be (double) blinded, it is crucial to ensure adequate concealment of the allocation of patients to the groups to be compared (q.v. Section 1.5). It is also necessary that the outcome variable is independent of the (non-blinded) treating staff or documented in a blinded manner independent of the treating staff (blinded documentation of outcome measures). If a blinded documentation of outcome measures is not possible, a "hard" objective endpoint should be chosen (e.g. mortality), which practically cannot be influenced (with regard to its dimension and the stringency of its documentation) by the (non-blinded) person who documents it.

1.16 Consideration of legal aspects of data protection/confidentiality

The processing of personal data within the Institute is conducted according to the relevant federal data protection laws.[j] The data protection officer appointed by the Institute is responsible for ensuring compliance with these laws.

The Institute may in future also process personal data (attributed to an *identifiable* individual) obtained from research projects. In exceptional cases, personal data attributed to an *identified* individual may be used. If personal data were originally collected or are being collected by a third party, the corresponding declarations on compliance with legal regulations need to be provided. Furthermore, for each individual case, the fulfilment of the necessary legal requirements (informed consent, patient information, etc.) needs to be assessed thoroughly beforehand.

A further aim is to receive personal data that is primarily attributed to an identified individual in an anonymous or pseudonymous form from third parties and process them. In most cases, it will be sufficient to use data coded this way for research purposes and individual research questions. In particular, possible reservations about transferring data to the Institute should thereby be dispelled.

If data are transferred to the Institute that are not allowed to be published, these data cannot be considered in the Institute's evaluations as this would contradict the obligation for transparency (q.v. Section 4.7.3).

With regard to the confidential handling of data from commercial enterprises, appropriate declarations guaranteeing the Institute's confidentiality will, if necessary, be made to third parties. Besides having the necessary technical infrastructure to ensure data safety, corresponding clauses obliging personnel to observe confidentiality are included in all the Institute's employment contracts. In individual cases, externally appointed persons or institutions must also make corresponding obligations towards the Institute.

References

[1] (German) Federal Data Protection Act of December 20, 1990 (BGBl. I 1990 S.2954), amended by law of September 14, 1994 (BGBl. I S. 2325).

[j] German data protection laws distinguish between personal data attributed to an identified individual (e.g. name, address) and personal data attributed to an identifiable individual (e.g. medical diagnosis). The respective German terms are "personenbezogene Daten" and "personenbeziehbare Daten" (Federal Data Protection Act, [1]).

1.17 Consideration of ethical aspects

The Institute's primary aim is to improve the health care of the population in Germany through its high-quality work. The Institute's main focus is the maximisation of overall as well as individual patient benefits and the strengthening of patient autonomy through health education and information. The methods of evidence-based medicine are seen as essential and valuable tools for this purpose. The Institute applies these tools conscientiously, taking their limitations into account.

Furthermore, the Institute is aware of its standing in the German health care system and especially of its responsibility towards the people and institutions using, performing, financing or developing health care services. Even though health care legislation requires strict separation of scientific evaluation on the one hand, and any decision for or against the inclusion of a medical intervention as a service provided by the statutory health insurance on the other, the Institute is well aware that its work may have a direct or indirect influence on health care. Consequently, the consideration of the possible or probable effects that the Institute's reports have or will have on individuals, population groups or occupational groups, as well as on institutions or commercial enterprises, constitutes an essential element in the Institute's work. The involvement of individual representatives of groups and institutions affected by the Institute's projects will support this approach. The Institute's major responsibility is the public interest, and it therefore pays attention to transparency and independence. It is particularly important for the Institute that the conclusions of its reports and other findings of its work are not influenced by the interests of specific groups. Within the Institute, the prevailing transparency aims to ensure that no surreptitious attempts to influence its work are possible.

The Institute does not ignore questions concerning the fairness of distribution of resources. Having only limited resources means that an increase in investments in one area of the health care system necessarily leads to limitations in other areas. The Institute will convey the message that the decision for or against a medical procedure must derive from a conscientious consideration of generally accepted priorities. In this regard, the Institute sees the consideration of the interests of minorities and disadvantaged population groups as an important responsibility.

Ethical issues also have high priority in the Institute's own research projects. When producing a scientific report, it is necessary to consider the advantages and disadvantages for those affected. Furthermore, where necessary, advice on ethical issues should be sought during the planning and conduct of studies.

1.18 Description of types of bias

Bias is the systematic deviation of the effect estimate (inferred from study data) from the true effect. A range of possible causes may produce bias [1]. The following text only describes the most important types; a detailed overview of various types of bias in different situations is presented by Feinstein [2].

Selection bias is caused by a violation of the random principles for sample procedures. Particularly when comparing two groups, selection bias can lead to systematic differences between groups. If this leads to an unequal distribution of important confounders in the groups, the results of a comparison are usually no longer interpretable. When comparing groups, randomisation is the best method to avoid selection bias, as the groups formed do not differ systematically with regard to known and unknown confounders. However, structural equality can only be guaranteed if the sample sizes are sufficiently large. In small studies, despite randomisation, relevant differences between groups can randomly occur. When comparing groups with structural inequality, the effect of known confounders can be taken into account by employing multi-factorial methods (q.v. Section 1.22). However, the problem of a systematic difference between the study groups due to unknown or insufficiently investigated confounders remains.

Performance bias is a systematic distortion due to the different types of care provided (apart from the intervention to be investigated). Besides the comparability of the study groups with regard to potential prognostic factors, equality of care and the equality of observation of study participants play an important role. A breach of the equality of observation can lead to *detection bias*. Double-blinding is an effective protection against both performance and detection bias, which are summarised as *information bias* in epidemiological studies.

Protocol violations and study withdrawals can cause a systematic distortion of study results *(attrition bias)*. To avoid attrition bias, the "intention-to-treat" principle can be applied, where all randomised study participants are evaluated within the group to which they were assigned, independently of protocol violations.

In diagnostic studies, the assessment of the diagnostic test should be conducted in an appropriate spectrum of patients. If the sample assessed differs systematically from the patient population in which the test is to be applied, this can lead to *spectrum bias* (q.v. also Section 2.3). To avoid this type of bias, the diagnostic test should be assessed in a representative patient population.

When assessing screening programmes, it needs to be considered that earlier diagnosis of a disease often results in only an apparent increase in survival times, due to non-comparable starting points *(lead time bias)*. Increased survival times can also be feigned for diseases if a screening technique

favours the detection of mild or slowly progressing early stages of a disease *(length bias)*. The conduct of a randomised trial to assess the effectiveness of a screening technique can protect against these bias mechanisms (q.v. Section 2.4).

A common problem arising from the estimation of effects is a bias of results by measurement errors and misclassifications in the study data collected [3,4]. In practice, measurement errors can hardly be avoided, and it is known that non-differential measurement errors can also lead to bias in the estimation of an effect. In the case of a simple linear regression model with a random classical measurement error in the explanatory variable, *dilution bias* occurs, i.e. an attenuation of the estimate towards the zero effect. In other models and more complex situations, bias in all directions is possible. Depending on the research question posed, the size of potential measurement errors should be discussed in all studies and, if necessary, methods to adjust bias should be employed.

Missing values present a similar problem. Missing values not due to a random mechanism can also cause bias in a result [5]. 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 employed to account or compensate for bias (q.v. Section 1.11).

Publication bias plays an important role in systematic reviews [6]. As significant results are more frequently published than non-significant ones, a systematic bias of the common effect estimate occurs when published results are summarised. Graphical methods such as the funnel plot [7] and/or statistical methods such as meta-regression are techniques for identifying and considering publication bias [8,9,10] (q.v. Section 1.21).

References

[1] Sackett DL. Bias in analytic research. J Chron Dis 1979; 32: 51-63.

[2] Feinstein AR. Clinical Epidemiology. The Architecture of Clinical Research. Philadelphia: WB Saunders Co.; 1985.

[3] Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu, CM. Measurement Error in Nonlinear Models: A Modern Perspective, 2nd Ed. London: Chapman & Hall; 2006.

[4] Cheng C-L, van Ness JW. Statistical Regression with Measurement Error. London: Arnold; 1999.

[5] Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd Ed. New York: Wiley; 2002.

[6] Begg CB, Berlin JA. Publication bias: A problem in interpreting medical data (with discussion). J R Stat Soc A 1988; 151: 419-463.

[7] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple graphical test. BMJ 1997; 315: 629-634.

[8] Sterne JAC, Egger M, Davey Smith G. Investigating and dealing with publication and other biases in meta-analysis. BMJ 2001; 323: 101-105.

[9] Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. Stat Med 2001; 20: 641-654.

[10] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006; 295: 676-680.

1.19 Evaluation of a difference

Various aspects need to be considered when presenting empirical evidence that certain groups differ with regard to a certain characteristic. It should be noted that the term "evidence" should not be understood in the mathematical sense of the term "proof". 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 minimised in order to provide "statistical evidence" in support of a hypothesis. Significance tests are the typical methods used to provide this type of statistical evidence in medical research (q.v. Section 1.2). This level of argumentation should be distinguished from the evaluation of the clinical relevance of a difference (q.v. Section 1.3). In practice, the combination of both arguments provides a suitable description of a difference on the basis of empirical data.

When employing a significance test to show a difference, there should be an a priori determination of the research question posed and, based on this question, a determination of the outcome variable, the effect measure, and the formulation of the statistical hypothesis. It is necessary to calculate the sample size before the start of study, so that the size of the study population is sufficient to detect a treatment difference. In addition to the above information, a statement on the clinically relevant difference as well as an estimate of the variability of the outcome measure should be provided for simple situations. For more complex designs and/or research questions, further information, for example, on the correlation structure, recruitment scheme, and estimation of drop-out numbers, is required [1,2].

Finally, the description of results should include the following details: a statement on the significance level; a confidence interval for the effect measure chosen (calculated with appropriate methods) (q.v. Section 1.1); a descriptive statement on further effect measures to expound different aspects of the results; and a discussion about the clinical relevance of the results based on the determination of patient-relevant outcomes (q.v. Sections 1.3 and 2.1).

References

[1] Desu MM, Raghavarao D. Sample Size Methodology. Boston: Academic Press; 1990.

[2] Bock J, Toutenburg H. Sample size determination in clinical research. In: Rao CR, Chakraborty R, editors. Handbook of Statistics Vol. 8. Amsterdam: Elsevier; 1991: 515-538.

1.20 Evaluation 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 [1]. To demonstrate "equivalence", methods to test the equivalence hypothesis need to be employed [2]. It is important to understand that showing exact "equivalence", e.g. the statement that the difference in mean values between two groups is exactly zero, is not possible by means of statistical methods. In practice, it is not evidence of exact equivalence that is required, but rather evidence of a difference between two groups that is at most irrelevant. To achieve this objective, it must, of course, first be defined what an irrelevant difference is, i.e. the determination of an equivalence range is necessary.

To draw meaningful conclusions on equivalence, the research question and the resulting outcome variable, effect measure, and statistical hypothesis formulation need to be determined a priori (similar to methods to demonstrate evidence of a difference; q.v. Section 1.19). In addition, the equivalence range must be clearly defined in equivalence studies. 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 being referred to as a "non-inferiority hypothesis" [3-5].

As in superiority studies, it is necessary to calculate the required sample size in equivalence studies before the start of the study. The appropriate method depends on the exact hypothesis and the analysis method chosen [6].

Specifically developed methods should be applied to analyse data from equivalence studies. The "confidence interval inclusion method" is a frequently employed technique. If the confidence interval calculated lies completely within the previously defined equivalence range, then this will be classified as evidence of equivalence. To maintain the level of α=0.05, it is sufficient to calculate a 90% confidence interval [2].

Compared with superiority studies, equivalence studies have specific methodological problems. On the one hand, it is often difficult to provide meaningful definitions of equivalence ranges [7]; on the

other hand, the usual criteria for study designs, such as randomisation and double-blinding, no longer offer sufficient protection from bias [8]. Even without knowledge of the treatment group, it is possible, for example, to shift the treatment difference between groups to zero and hence in the direction of the desired alternative hypothesis. Moreover, "the intention-to-treat" principle should be applied carefully, as its inappropriate application may feign false equivalence [2]. For this reason, particular caution is necessary when assessing equivalence studies.

References

[1] Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311: 485.

[2] Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: The importance of rigorous methods. BMJ 1996; 313: 36-39.

[3] D'Agostino RB, Massaro JM, Sullivan KM. Non-inferiority trials: Design concepts and issues - the encounters of academic consultants in statistics. Stat Med 2003; 22: 169-186.

[4] Röhmel J, Hauschke, Koch A, Pigeot. Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren. Nicht-Unterlegenheit in klinischen Studien. [Biometric approaches to demonstrate efficacy in approval procedures. Non-inferiority in clinical trials]. Bundesgesundheitsbl - Gesundheitsforsch - Gesundheitsschutz 2005;48: 562–571.

[5] ICH E9 Expert Working Group. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials. Stat Med 1999; 18: 1905-1942.

[6] Roebruck P, Elze M, Hauschke D, Leverkus F, Kieser M. Literaturübersicht zur Fallzahlplanung für Äquivalenzprobleme [Literature review on sample size planning for equivalence problems]. Inform Biom Epidemiol Med Biol 1997; 28: 51-63.

[7] Lange S, Freitag G. Choice of delta: Requirements and reality - results of a systematic review. Biom J 2005; 47: 12-27.

[8] Senn S. Inherent difficulties with active control equivalence studies. Stat Med 1993; 12: 2367-2375.

1.21 Meta-analyses

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 [1]. The Institute uses the following terms and definitions: a *non-systematic review* is a description and evaluation of study results on a defined topic, without a sufficiently systematic and/or reproducible method for identifying relevant research results on this topic. A quantitative summary of data from several studies is described as a *pooled analysis.* Due to the absence of a systematic approach and the inherent subjective component, reviews and analyses not based on systematic literature

searches are extremely prone to bias. A *systematic review* is based on a comprehensive, systematic literature search and study evaluation in order to minimise 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. A meta-analysis is based in most cases on aggregate study data from publications. An overall effect is hereby calculated from the effect strengths measured in individual studies, taking sample sizes and variances into account. More efficient analysis procedures are possible if individual patient data are available from the studies considered. A *meta-analysis with individual patient data (IPD)* is the analysis of data on the patient level within the framework of a general statistical fixed or random effects model, in which the study is considered as an effect and not as an observation unit. The Institute sees a *prospective meta-analysis* as a statistical summary (planned a priori) of the results of several studies that were jointly planned prospectively. If other studies are available on the particular topic, these must also be considered in the evaluation 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 of the overall effect, including confidence intervals, are presented graphically [2]. On the one hand, fixed effects models are applied, which provide weighted mean values of the effect strengths (for example, 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 subject of controversy and has to be answered individually for each analysis [3,4]. If there is no clear justification for the choice between random and fixed effects models, the Institute will always apply both methods and describe any divergent results.

Before a meta-analysis is conducted, it must first be considered whether the summarisation of the studies investigated is in fact meaningful. On the one hand, the comparability of the studies with regard to the question posed should be given, and on the other hand, the heterogeneity of the studies with regard to their results should be investigated [5]. For this purpose, specific new statistical methods are available, such as the I^2 measure [6]. For this measure, studies exist which allow a rough classification of heterogeneity (for example, low: $I^2 \leq 25\%$; medium: $25\% < I^2 \leq 50\%$; high: $I^2 > 50\ \%$) [7]. If the heterogeneity of the studies is too great, the statistical summarisation of the study results is possibly not meaningful [8]. In such a case the Institute will not conduct a meta-analysis. However, besides statistical measures, issues of content must also 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. It is possible that the choice of a certain measure may lead to greater study heterogeneity, but 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 [9]. 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 should be inferred from them.

In the case of great heterogeneity of the studies, it is necessary to investigate the potential underlying causes. Factors may possibly be detected by way of meta-regression that could explain the heterogeneity of the effect strengths [10,11]. In a meta-regression, the statistical association between the effect strengths of the individual studies and the study characteristics is investigated, so that study characteristics can possibly be identified that have an explanatory value for the different effect strengths; i.e. for the heterogeneity of the studies. 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 randomised studies, only evidence of an observed association can be inferred from a meta-regression, not a causal relationship [10]. Meta-regressions that attempt to show an association between the different effect strengths 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 [12]. 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 individual patient data [10].

References

[1] Egger M, Smith GD, Altman DG, editors. Systematic Reviews in Health Care – Meta-analysis in Context. London: BMJ Books; 2001.

[2] Lewis S, Clarke M. Forest plots: Trying to see the wood and the trees. BMJ 2001; 322: 1479-1480.

[3] Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. Stat Med 2000; 19: 1707-1728.

[4] Villar J, Mackey ME, Carroli G, Donner A. Meta-analysis in systematic reviews of randomized controlled trials in perinatal medicine: Comparison of fixed and random effects models. Stat Med 2001; 20: 3635-3647.

[5] Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. Stat Med 2002; 21: 1503-1511.

[6] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21: 1539-1558.

[7] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327: 557-560.

[8] Deeks JJ, Higgins JPT, Altman DG. Analysing and presenting results. In: Higgins JPT, Green, S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005]; Section 8. In: The Cochrane Library, Issue 3, 2005. Chichester: Wiley; 2005.

[9] 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: 72-76.

[10] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21: 1559-1573.

[11] Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med 2002; 21: 589-624.

[12] Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. Int J Epidemiol 1989; 18: 269-274.

1.22 Adjustment principles and multi-factorial methods

Multi-factorial methods that enable the effect of confounders to be compensated primarily play a key role in non-randomised studies (q.v. Section 1.18) [1]. Studies with several study groups are a further important field of application for these methods [2]. The description of results obtained with multi-factorial methods is unfortunately often insufficient in the medical literature [3,4]. To be able to assess the quality of such an analysis, the description of essential aspects of the statistical model formation is necessary [5,6], as well as details on the quality of the model (goodness of fit) [7].

The most relevant information for this purpose normally is:

- A clear description and an a priori determination of the outcome variables and all explanatory variables,
- Information on the measurement scale and on the coding of all variables,
- Information on the selection of variables and on any interactions,
- Information on how the assumptions of the model were verified,
- Information on the goodness of fit of the model,
- Inclusion of a table with the most relevant results (parameter estimate, standard error, confidence interval) for all explanatory variables.

Depending on the research question posed, this information is of differing relevance. If the issue is a good prediction of the outcome variable within the framework of a prognosis model (see Sections

2.8 and 2.9), a high quality of the model is more important here than in a comparison of groups, where an adjustment for confounders must be made.

Inadequate description of the results obtained with multi-factorial methods is especially critical if, as a result of the (unclearly described) statistical modelling, a shift in effects to the "desired" range occurs that is not recognisable with mono-factorial methods. Detailed comments on the requirements for the use of multi-factorial methods can be found in various reviews and guidelines [1,8,9].

The Institute employs modern methods in its own regression analysis calculations [10]. In this context, results of multi-factorial models that were obtained from a selection process of variables should be interpreted with great caution. If, when choosing a model, such selection processes cannot be avoided, a type of backward elimination will be employed, as this procedure is preferable to the procedure of forward selection [10]. A well-informed and careful pre-selection of the candidate predictor variable is essential in this regard [11]. If required, modern methods such as the lasso technique will also be employed [12]. For the modelling of continuous covariates, the Institute will, if necessary, fall back on flexible modelling approaches, such as regression using fractional polynomials, to enable the appropriate description of non-monotonous associations [13,14].

References

[1] Katz MH. Multivariable analysis: A primer for readers of medical research. N Engl J Med 2003; 138: 644-650.

[2] McAlister A, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: A systematic review. JAMA 2003; 289: 2545-2553.

[3] Bender R, Grouven U. Logistic regression models used in medical research are poorly presented (Letter). BMJ 1996; 313: 628.

[4] Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: A cross-sectional survey. Ann Intern Med 2002; 136: 122-126.

[5] 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: 361-387.

[6] Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med 2000; 19: 1831-1847.

[7] Hosmer DW, Lemeshow S. The importance of assessing the fit of logistic regression models: A case study. Am J Public Health 1991; 81: 1630-1635.

[8] Bender R, Grouven U. Ordinal logistic regression in medical research. J R Coll Physic London 1997; 31: 546-551.

[9] 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: 979-985.

[10] Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

[11] Derksen S, Keselman HJ. Backward, forward, and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Statist Psych 1992; 45: 265-282.

[12] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996; 58: 267-288.

[13] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Appl Stat 1994; 43: 429-467.

[14] Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. J R Stat Soc A 1999; 162: 71-94.

1.23 Evaluation of qualitative studies

Qualitative research methods are used to explore and understand subjective experiences, individual actions, and the social world [1-4]. The Institute aims to use qualitative studies to generate hypotheses and to assist in the interpretation of data as well to gain insight into patients' experiences and views.

Quantitative research works primarily with numbers of different dimensions and is characterised by strong standardisation, although personal and social experiences may also be taken into account. Conversely, in qualitative research the emphasis is on subjective data [1].

The Institute's main task in the evaluation of social science studies is to determine whether the chosen study design, study quality, and reliability of results were appropriate for the research question investigated. A weaker general consensus exists with regard to the validity of criteria for the conduct, evaluation, and synthesis of qualitative studies than for other research fields [1-5]. The Institute will use defined criteria for its assessment of qualitative studies within the framework of systematic reviews and health technology assessments (HTAs) [1].

References

[1] Dixon-Woods M, Agarwal, S., Young B, Sutton A. Integrative approaches to qualitative and quantitative evidence. London: NHS Health Development Agency; 2004.

[2] Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: A review of the literature. Health Technol Assess 1998; 2: iii-ix, 1-274.

[3] McClelland S. Qualitative Research Methods: Their Role in Health Services Research. NHS Management Briefing, National Electronic Library for Health. 00&04, February 2001. http://libraries.nelh.nhs.uk/healthmanagement [accessed on 22.10.2004].

[4] Harden A, Garcia J, Oliver S, Rees R, Shepherd J, Brunton G, Oakley A. Applying systematic review methods to studies of people's views: An example from public health research. J Epidemiol Community Health 2004; 58: 794-800.

[5] Thomas J, Harden A, Oakley A, Oliver S, Sutcliffe K, Rees R, Brunton G, Kavanagh J. Integrating qualitative research with trials in systematic reviews. BMJ 2004; 328: 1010-1012.

1.24 Use of consultation techniques

According to the research questions and tasks assigned, it may be necessary to use a variety of consultation techniques. Against the background of profound but rapid processing, rapid appraisal methods (e.g. focus groups) are usually employed. However, these methods have various strengths and weaknesses; they can vary strongly in their representativeness and validity. When employing these methods, the Institute therefore follows existing quality criteria for their selection, application, and analysis. Ethical aspects are particularly taken into account where techniques are applied that may have detrimental effects on participants (e.g. focus groups).

Consultation techniques are preceded by a literature search for relevant qualitative data. The findings resulting from the consultation are interpreted in the context of the available results of relevant and more detailed qualitative studies.

The Institute may apply the following consultation techniques:

- Key informant interviews [1],
- Focus groups [2-4],
- Group interviews, group meetings and consultations [5-7],
- Surveys and polling (including online polling and feedback mechanisms),
- Occasional use of consensus techniques, e.g. Delphi techniques [8] and participatory evaluation [9].

The Institute may also develop health impact assessments (HIAs) using both qualitative and quantitative methods [10]. The basis for conducting these assessments is a clear and transparent procedure, also with regard to the potential impact of decisions on equity and social justice.

The different techniques for documenting people's views and attitudes vary greatly in their respective reliability and validity. The Institute therefore needs to take care to ensure that the views of disadvantaged groups are adequately considered in this process.

One of the objectives of the Department of Health Information is to promote health and scientific literacy in the population. On the one hand, the aim is to enhance understanding of scientific terminology pertaining to health issues and evidence-based aspects of the health care system, and on the other, to arouse public interest in the Institute's work. The public is to be actively involved in this process. For this purpose, the department may use and further develop the methodology of consultation techniques, consensus building, and public decision-making [11].

Population level techniques, such as online surveys, citizens' juries [11], and public investigations are widely used in resource allocation decisions in the health care system [11,12]. Citizens' juries have been found to be particularly effective in the investigation of complex issues. Some of these techniques may be adapted by the Institute to achieve the aims described above.

References

[1] USAID Center for Development Information and Evaluation. Conducting key informant interviews. Performance Monitoring and Evaluation Tips 1996; No 2. http://www.usaid.gov/pubs/usaid_eval/pdf_docs/pnabs541.pdf [accessed on 22. 10.2004].

[2] USAID Center for Development Information and Evaluation. Conducting focus group interviews. Performance Monitoring and Evaluation Tips 1996; No 10. http://www.usaid.gov/pubs/usaid_eval/pdf_docs/pnaby233.pdf [accessed on 22.10.2004].

[3] Aylward P. Conducting research with focus groups. Staff Development Session. Flinders University of South Australia. http://www.flinders.edu.au/staffdev/courses/research/resources/Focusgroups.pdf [accessed on 22.10.2004].

[4] Dixon-Woods M, Agarwal, S., Young B, Sutton A. Integrative Approaches to Qualitative and Quantitative Evidence. London: NHS Health Development Agency; 2004.

[5] National Resource Centre for Consumer Participation in Health. Feedback, Participation and Diversity: A Literature Review. Canberra: Commonwealth of Australia, 2000. http://www.participateinhealth.org.au [accessed on 22.10.2004].

[6] National Health and Medical Research Council, Consumers Health Forum of Australia. Statement on Consumer and Community Participation in Research. Canberra: Commonwealth of Australia, 2002. http://www.nhmrc.gov.au/publications/_files/r22.pdf [accessed on 17.11.2006].

[7] National Resource Centre for Consumer Participation in Health. Methods and Models of Consumer Participation. Melbourne: National Resource Centre for Consumer Participation in Health, 2004. http://www.participateinhealth.org.au [accessed on 22.10.2004].

[8] Liberati A, Sheldon TA, Banta HD. EUR-ASSESS Project Subgroup report on Methodology. Methodological guidance for the conduct of health technology assessment. Int J Tech Assess Health Care, 1997; 13: 186-219.

[9] USAID Center for Development Information and Evaluation. Conducting a participatory evaluation. Performance Monitoring and Evaluation Tips 1996; No 1. http://www.usaid.gov/pubs/usaid_eval/pdf_docs/pnabs539.pdf [accessed on 22.10.2004].

[10] Barnes R, Scott-Samuel A. Health Impact Assessment: A Ten Minute Guide. Liverpool: International Health Impact Assessment Consortium, 2000. http://www.ihia.org.uk/hiaguide.html [accessed on 25.10.2004].

[11] Ryan M, Scott DA, Reeves C, Bate A, van Teijlingen ER, Russell EM, Napper M, Robb CM. Eliciting public preferences for healthcare: A systematic review of techniques. Health Technol Assess, 2001; 5: 1-186.

[12] Hicks N, Harford J. Summary Report on Consumer Participation in Resource Allocation. Melbourne: National Resource Centre for Consumer Participation in Health; 2000.

1.25 Appraisal (external review)

To ensure the quality of the Institute's reports and its other products, comments from experts, consumers, patients and their relatives are obtained, depending on the research question and product. Specific aspects of the external review process are described in detail in the corresponding sections.

Specific studies investigating the effectiveness of particular external review procedures have only recently been conducted, and only a few meaningful studies are available in this regard [1,2]. In particular, there is a lack of sufficiently valid intervention studies. According to the studies available [3], the relevance of conventional procedures employed in medical journals [2,3], including the evaluation by consumers and patients [4], is not sufficiently clear. The Institute will therefore evaluate and update its reviewing system.

References

[1] Rennie D. Editorial peer review: Its development and rationale. In: Godlee F, Jefferson T, editors. Peer Review in Health Sciences, 2nd Ed. London: BMJ Books, 2003; 1-13.

[2] Fletcher RH, Fletcher SW. The effectiveness of journal peer review. In: Godlee F, Jefferson T, editors. Peer Review in Health Sciences, 2nd Ed. London: BMJ Books, 2003; 62-75.

[3] Jefferson TO, Rudin M, Brodney FS, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. The Cochrane Database of Method Rev 2006, Issue 1. Chichester: Wiley, 2006.

[4] Bastian H. Non-peer review: Consumer involvement in research review. In: Godlee F, Jefferson T, editors. Peer Review in Health Sciences, 2nd Ed. London: BMJ Books, 2003; 248-262.

2. Specific evaluation of medical and health care issues

The extraction and evaluation of data from relevant publications found in literature searches (q.v. Section 4.7) are conducted and documented in a structured manner. Data extraction forms developed by the Institute should be used for the documentation process. If these forms are not used, this must be justified and agreed to by the project manager. Alternative forms will then be developed, agreed to and applied.

2.1 Evaluation of benefits and harms in medicine

General patient-relevant medical benefit

To define a patient-relevant medical benefit, first of all, it is required to distinguish between the terms “necessity” and “benefit”. According to § 27 Social Code Book V, a medical intervention is necessary if it can detect an illness, cure it, prevent its exacerbation, or alleviate its symptoms. The term “necessity” goes beyond the term “benefit”. Evidence of a benefit is thus a necessary but not a sufficient requirement for evidence of a necessity. The evaluation of evidence may result in the finding that the existence of a benefit (or harm) of an intervention is substantiated, the absence of a benefit is substantiated, or that its existence or absence cannot be substantiated. In addition to the evaluation of effectiveness and safety, the Institute’s reports are designed primarily to describe the benefits and harms of all kinds of medical interventions. As the benefit of an intervention should be related to the patient, the respective evaluation is based on the results of studies that have investigated the effects of an intervention on patient-relevant endpoints. In this regard, the intentional as well as the unintentional effects of the intervention are taken into account.

These effects can comprise disease- and treatment-related changes, particularly in the following endpoints:

1. Mortality,
2. Morbidity (complaints and complications),
3. Health-related quality of life,
4. Time and effort related to disease and intervention,
5. Patient satisfaction.

A positive change (for the patient) in these endpoints is defined as a direct patient-relevant medical benefit, a negative change as a direct patient-relevant medical harm. If possible, the relationship between these two parameters is expressed as the benefit-harm ratio.

Diagnostic measures can be of indirect benefit, as they are an essential prerequisite for therapeutic interventions through which it is possible to influence the outcomes listed above under 1 to 5. Diagnostic tests can also enable decision-making processes relevant to the patients.

The utilisation of health care services can also be a patient-relevant factor. However, its direct relevance should be supported by a coexisting effect on treatment satisfaction. It should also be described to what extent a benefit pertaining to the utilisation of health care services is possibly associated with harm in respect of life expectancy, complaints and complications, patient satisfaction and other aspects of health-related quality of life, and the time and effort related to the disease and intervention.

Interventions may also have an impact on those indirectly affected, such as family members and nursing staff. Where applicable, this impact may also be considered in the Institute's reports.

Primarily, endpoints are considered that present, reliably and directly, concrete changes in patients' state of health. In this context, individual affected persons, patient representatives and/or consumer organisations will be involved with regard to the topic-related definition of patient-relevant outcomes. As a precondition for their use in clinical studies, instruments recording the quality of life or other so-called "patient reported outcomes" (PRO) should be suitable on the basis of an appropriate evaluation [1,2]. In addition, valid surrogate endpoints may be considered in the evaluation of benefits and harms of interventions.

Study duration

When selecting studies relevant to the evaluation of the benefits and harms of an intervention, their duration is an essential criterion. When evaluating an intervention to treat an acute illness whose primary objective is, for example, to shorten the duration of illness or alleviate acute symptoms, it is not usually meaningful to demand long-term studies unless late complications are expected. On the other hand, for a complete evaluation of benefits and harms, short-term studies are not usually suitable to assess interventions to treat chronic illnesses. This especially applies if treatment is required for several years, or even lifelong. In such cases, particularly studies that cover a treatment period of several years are meaningful and desirable. As aspects of both the benefit and the damage potential can be distributed differently over time, in long-term studies a meaningful weighing of

benefits and harms is only feasible if the study duration is long enough. Individual aspects of the benefits and harms of an intervention may indeed be investigated in short-term studies.

However, it can be assumed that a patient-relevant benefit will never occur before the intervention becomes effective. With regard to the selection criterion "minimum study duration", the Institute therefore primarily follows generally accepted requirements for providing evidence of the effectiveness of an intervention. For the evaluation of pharmaceuticals, the Institute will in particular revert to the information provided in the corresponding indication-specific guidelines of regulatory authorities (e.g. [3]). As the evaluation of the benefit of an intervention also includes assessing aspects of its damage potential, when determining the minimum study duration, the generally accepted requirements in this regard are also relevant. Furthermore, for long-term interventions, as described above, the Institute will revert to the criterion "long-term treatment" used in relevant guidelines [4]. In individual cases, the Institute may deviate from this approach (and will justify this deviation), for example, if it is necessary due to aspects of content to demand a follow-up over a longer period, or if specific (sub-) questions apply to a shorter period. Such deviations may also be indicated if short-term effects are also a subject of the evaluation (e.g. when assessing newly available/approved interventions and/or technologies where no appropriate treatment alternative exists).

Surrogates of a patient-relevant medical benefit

Surrogate endpoints are frequently used in medical research as a substitute for patient-relevant endpoints, mostly to obtain conclusions on patient-relevant benefits earlier and more simply [5-7]. Most surrogate endpoints are, however, unreliable in this regard and can be misleading when used in an evaluation of an intervention [8,9]. As a rule, in the Institute's evaluations, surrogate endpoints are therefore not taken as evidence of a benefit of an intervention, unless clear evidence exists from intervention studies of a plausible, strong, consistent, and unidirectional association between the change in the surrogate and the change in the patient-relevant endpoint.

A unidirectional association means that a positive or negative change in the surrogate accompanies, consistently and always in the same manner, a change in the patient-relevant endpoint.

The validity of a surrogate is regarded as not proven if no meaningful studies are available describing the association between the modification of this surrogate and the change in the corresponding patient-relevant endpoint. In addition, a surrogate is not seen as valid if it has been demonstrated in studies that an intervention:

- Had an effect on the surrogate endpoint, but not on the patient-relevant endpoint, or

- Had an effect on the patient-relevant endpoint, but not on the surrogate endpoint, or
- Produced inconsistent effects on the surrogate and patient-relevant endpoints.

Surrogate endpoints of unclear or controversial validity may be described in the Institute's reports. However, such endpoints are not suitable for providing evidence of a benefit of an intervention.

Dramatic effects

If the course of a disease is clearly predictable (or at least to a great extent), and no treatment options are available to influence it, then evidence of a benefit of a medical intervention can also be provided by the observation of a change in the disease's (more or less) deterministic course in well-documented case series. If, for example, it is known of a disease that in many cases it 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 survived for a longer period of time, then this so-called "dramatic effect" may be sufficient to provide evidence of a benefit. 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 case series. Possible harms of the intervention should also be taken into account.

Benefits in small populations

There is no convincing argument to justify deviating from the hierarchy of evidence levels in small populations (e.g. patients with rare diseases or subgroups of patients with common diseases). Patients with very rare illnesses also have a right to the best possible information on treatment options [10]. Non-randomised studies require larger sample sizes than randomised studies due to the need of adjustment for confounding factors. However, it may sometimes be impossible, due to the rarity of a disease, to include so many patients that the study has sufficient statistical power. A meta-analytical summary of smaller studies may be particularly meaningful in such cases. Smaller random samples generally result in less precision of an effect estimate, accompanied by wider confidence intervals. For small sample sizes, it may indeed be necessary, due to the relevance of the assumed effect of an intervention, its size, the availability of alternatives, and the frequency and severity of potential therapy-related harms, to accept a higher *p*-value than 5% (e.g. 10%) for evidence of statistical significance [11]. This must, however, be determined a priori and justified comprehensibly. With small sample sizes, it may also more often be necessary to replace rarely

occurring patient-relevant endpoints with surrogate endpoints. However, these surrogate parameters must also be valid for small sample sizes [12].

Benefit in individual cases

A) Subjective weighing of the potential benefits against potential harms

The possibility of weighing individually between the ever-present risk of harm of a medical intervention and the chance of a benefit rests on the assumption that reliable data are available which are applicable to the particular individual case, and which describe the frequency of beneficial and harmful events, together with their respective degree of uncertainty. The frequency of these events should be expressed in the form of absolute risks (for example in percentages) for the relevant medical intervention and the suggested alternatives. The alternatives should, if applicable, describe any further possible medical interventions available, and also describe the possibility of conscious inaction. In this context, the term "inaction" refers only to the therapeutic decision under investigation and does not exclude supportive interventions of a different kind. The evaluation of the particular possible beneficial and harmful events regarding their size and relevance for the individual case is subjective; therefore, the decision based on the weighing of these benefits and harms can never be generally right or wrong.

B) Evidence of a benefit in individual cases

Evidence of an actual benefit in individual cases can only be provided post hoc. As many medical interventions are directed at the primary or secondary prevention of an undesired event, evidence of a benefit for individual cases cannot serve here as the basis for a medical decision. The results of controlled intervention studies are therefore taken as a basis for the estimate of the probability of the occurrence of events of this kind. It is, however, possible that such studies do not cover all events and/or the individual case deviates from the overall result (if, for example, a patient in fact dies during an operation that usually saves lives). As such events cannot be predicted with certainty for individual cases and are irreversible, the medical decision must, in such an individual case, be founded on the normal case.

In other cases, such as reversible events, it can sometimes be reliably shown for individual cases that a specific event is caused in a specific person by a defined medical intervention. The medical intervention can hereby be "tried out", and the continuation of this measure can conform to the result of this intervention (for example in pain therapy). Various confounding factors should thereby be considered, which can impede the interpretation of this "trial" (for example the effects of time and suggestive behaviour).

C) Single patient trials

Single patient trials are common in practical medicine, but are mostly performed unsystematically; their interpretation is therefore quite often unreliable, due to the usually poor control of confounding factors [13].

In systematic single patient trials – the so-called "n of 1" studies, only a single patient participates, and the outcomes must be completely and quickly reversible [13,14]. If such a study is possible and well performed, it allows conclusions to be drawn as to whether a particular patient profits from a particular treatment. Such "n of 1" studies consist of several "trial and control" study periods, which are applied in random succession in a patient [13]. In this type of study design, the treatment periods, rather than the patients, are randomised. In optimum circumstances, the intervention should be blinded and include either an active or placebo control. Unfortunately, the methodological quality of single patient trials is frequently insufficient. The conclusions drawn from such suboptimal study designs show considerable errors, depending on the effect size and the magnitude and frequency of confounding factors [12]. To provide evidence of a benefit for individual cases, the use of a non-blinded or uncontrolled design requires a comprehensible justification.

Summarising evaluation

Medical interventions are compared with other clearly defined active or placebo interventions, or with no intervention, with regard to their beneficial and harmful effects on defined patient-relevant endpoints, and described in a summarised form. To this end, exactly one of the five following evaluating conclusions is made for each predefined patient-relevant endpoint on the basis of the analysis of available data:

1) Evidence of a benefit (harm) exists, or
2) Indications are available that a benefit (harm) exists, or
3) No benefit (harm) exists, or
4) Indications are available that no benefit (harm) exists, or
5) No evidence and no indication of a benefit (harm) exist.

An intervention is described as "1)" if scientific evidence of a benefit or harm exists. An intervention is described as "2)" if indications of a benefit or harm exist, but they are not clear and/or consistent. An intervention is described as "3)" if scientific evidence exists that this intervention is not associated with a benefit or harm. An intervention is described as "4)" if indications of a lack of a benefit or harm exist, but they are not clear and/or consistent. An

intervention is described as "5)" if no evidence or indications exist of a benefit or harm, for example, due to insufficient or inconsistent data.

These conclusions, made separately for each patient-relevant endpoint, are then summarised as far as possible in a concluding evaluation in the form of a weighing of benefits and harms. If evidence of a benefit and/or harm exists with regard to Points 1-5, the Institute will present, as far as possible, and on the basis of the available data:

1) The benefit potential,
2) The damage potential,
3) The weighing of the benefit and damage potential.

In this context, the Institute will follow the principle of risk prevention, i.e. if in doubt, assume that a damage potential exists. Furthermore, special features due to age, gender, and living conditions will also be considered. The exact description of the evaluation of the weighing of benefits and harms is topic-specific and should, if this is possible prospectively, be described in the report plan (protocol) and otherwise in the preliminary report.

References

[1] Food and Drug Administration. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Online 2006/19/01. http://www.fda.gov/cder/guidance/5460dft.pdf [accessed on 03.04.2006].

[2] 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 2005/07/27. http://www.emea.eu.int/pdfs/human/ewp/13939104en.pdf [accessed on 03.04.2006].

[3] CPMP/EWP/1080/00 Note for Guidance on Clinical Investigation of Medicinal Products in the Treatment of Diabetes Mellitus (CPMP adopted May 2002). http://www.emea.eu.int/pdfs/human/ewp/108000en.pdf [accessed on 23.11.2006].

[4] ICH E1: The Extent of Population Exposure to Assess Clinical Safety for Drugs Intended for Long-Term Treatment of Non-Life-Threatening Conditions. http://www.ich.org/LOB/media/MEDIA435.pdf [accessed on 23.11.2006].

[5] Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. Stat Med 1989; 8: 431-440.

[6] Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996;125:605-613.

[7] ASCO Special Article. Outcomes of cancer treatment for technology assessment and cancer treatment guidelines. J Clin Oncol 1996; 14: 671-679.

[8] Gøtzsche PC, Liberati A, Torri V, Rossetti L. Beware of surrogate outcome measures. Int J Tech Assess Health Care 1996; 12: 238-246.

[9] Grimes DA, Schulz KF. Surrogate end points in clinical research: Hazardous to your health. Obstet Gynecol 2005; 105: 1114-1118.

[10] Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products. Official Journal of the European Commission of 22.01.2000; L18: 1-5.

[11] Committee for Medicinal Products for Human Use, EMEA. Guideline on the Choice of the Non-inferiority Margin. London, 27 July 2005. http://www.emea.eu.int/pdfs/human/ewp/215899en.pdf [accessed on 27.10.2005].

[12] Committee for Medicinal Products for Human Use, EMEA. Guideline on Clinical Trials in Small Populations. London, 17 March 2005. http://www.emea.eu.int/pdfs/human/ewp/8356105en.pdf [accessed on 23.03.2006].

[13] Schluter PJ, Ware RS. Single patient (n-of-1) trials with binary treatment preference. Stat Med 2005; 24: 2625-2636.

[14] Jull A, Bennett D. Do n-of-1 trials tailor treatment? Lancet 2005; 365: 1992-1994.

2.2 Pharmaceutical and non-pharmaceutical interventions

The objective of the evaluation of a study on a pharmaceutical or non-pharmaceutical intervention is to show with what certainty an effect or the absence of an effect can be derived from the study findings (certainty of results). Moreover, it is necessary to describe whether and to what extent the study results are transferable to local conditions (e.g. the population affected, type of care provided), and which local particularities have or could have an influence on the results or on their interpretation. From this point of view, studies in which actual health care conditions are portrayed as accurately as possible are particularly relevant; however, the criteria on the certainty of results described below should not be disregarded.

Certainty of results is essentially influenced by four components:

- The study design,
- The internal validity (dependent on the study design),
- The consistency of the results of several studies,
- The size of an expected or observed effect.

The criteria for the assessment of internal validity are described in detail in various parts of Section 1 of this methods paper and are applied correspondingly in the evaluation of studies on pharmaceutical and non-pharmaceutical interventions. In this context, different aspects may be of particular relevance, depending on the research question to be investigated.

The study design has considerable influence on the certainty of results insofar as a causal relationship between intervention and effect cannot normally be shown with prospective or retrospective epidemiological studies, whereas an experimental study design is, in principle, suitable for this purpose [1]; at least this is the case if factors influencing the results can be eliminated completely or almost completely. For this reason, an RCT represents the gold standard for the evaluation of pharmaceutical and non-pharmaceutical interventions [2].

To assess effectiveness, the Institute will therefore use non-randomised intervention studies or epidemiological studies only in exceptional cases. These exceptions must be justified. Reasons for exception are, for example, the non-feasibility of an RCT (e.g. if the therapist and/or patient have a strong preference for a specific therapy alternative) or the fact that other study types may also provide sufficient certainty of results for the research question posed. For diseases that, without intervention, would be fatal within a short period, the availability of several consistent case reports can provide sufficient certainty of results as to whether a particular intervention may prevent this otherwise inevitable course [3] (q.v. Section 2.1).

The particular obligation to justify a non-randomised design when testing pharmaceuticals can also be found within the framework of drug approval legislation in the directives relating to the testing of medicinal products (Directive 2001/83/EC, Section 5.2.5 [4]). This is weakened by the conformity evaluation in the EN ISO Norm 14155-2 (Section 4.7 [5]), where RCTs are not presented as the design of choice for medicinal products; however, the choice of design must be justified. The Code of Procedure of the Federal Joint Committee envisages, as far as possible, the preferential consideration of RCTs, independent of the type (pharmaceutical/non-pharmaceutical) of medical intervention to be evaluated (§ 20 [6]).

If a preliminary literature search indicates that, for a specific medical indication, no (or very few) studies of the highest evidence level (RCTs) exist on the intervention or alternative interventions to be evaluated, studies of a lower evidence level may be included in the Institute's reports, with consideration of the effort involved and the benefit in view of any resulting potential uncertainty of results.

As part of the report plan (q.v. Section 4.4), the Institute therefore determines beforehand which study types can, on the basis of the research questions posed, be regarded as feasible and providing sufficient certainty of results (with high internal validity). Studies not complying with these quality standards (q.v. Sections 1.8 and 1.9) are not primarily considered in the evaluation process.

Finally, the transferability of study results must be verified in a separate process that is initially independent of the study design and quality.

References

[1] Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359: 57-61.

[2] Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. The Cochrane Database of Methodology Reviews 2002, Issue 4. Art. No.: MR000012. DOI: 10.1002/14651858.MR000012.

[3] Liberati A, Sheldon TA, Banta HD. EUR-ASSESS Project Subgroup report on Methodology. Methodological guidance for the conduct of health technology assessment. Int J Tech Assess Health Care, 1997; 13: 186-219.

[4] Annex 1 of the Directive 2001/83/EC. Analytical, pharmacotoxicological and clinical standards and protocols in respect of the testing of medicinal products. Official Journal of the European Communities of 27 June 2003; L159: 49-94.

[5] Klinische Prüfung von Medizinprodukten an Menschen - Teil 2: Klinische Prüfpläne. [Clinical testing of medicinal products in human beings – Part 2: Clinical protocols]. (ISO 14155-2:2003); Deutsche Fassung EN ISO 14155-2:2003. Berlin: Beuth; 2003.

[6] Gemeinsamer Bundesausschuss. Verfahrensordnung des Gemeinsamen Bundesausschusses vom 20. September 2005, veröffentlicht im Bundesanzeiger 2005, S. 16998, in Kraft getreten am 1. Oktober 2005, zuletzt geändert am 18. April 2006, veröffentlicht im Bundesanzeiger 2006, S. 4876, in Kraft getreten am 7. Juli 2006. [Code of Procedure of the Federal Joint Committee dated 20.09.2005; published in the German Federal Gazette 2005, p. 16998; became operative on 01.10.2005. Last changed on 18.04.2006, published in the German Federal Gazette 2006, p. 4876; became operative on 07.07.2006]. http://www.g-ba.de/cms/upload/pdf/richtlinien/2006-07-07-VerfO.pdf [accessed on 30.11.2006].

2.3 Diagnostic tests

Diagnostic tests are often still evaluated with methods that do not have the degree of reliability of methods employed to evaluate therapeutic procedures [1,2]. For this reason, it can be assumed that the information for the evaluation of a diagnostic test presented below will only be completely available in exceptional cases (see e.g. [3]). The essential basis for an evaluation is the precise formulation of a research question, as studies on diagnostic tests are conducted with different objectives and, depending on the objectives set, not all information is relevant for an adequate evaluation [4]. Within the framework of the evaluation of benefits and harms of a diagnostic procedure it will thus be necessary to determine exactly in the report plan (see Section 4.4) which information within which type of study design should be investigated for the underlying research question posed. For example, when evaluating minor modifications of a diagnostic test whose benefit has already been proven, in individual cases it may be sufficient only to test whether the intra-test variability is equivalent or better (q.v. Section 2.3.2). If, in contrast, the validity of a new diagnostic principle is to be tested, then comparative randomised trials may be required (q.v. Section 2.3.3).

2.3.1 General aspects

The following information in particular is relevant to evaluate diagnostic tests:

- Clear definition of the disease(s) to be diagnosed/detected or, more generally, the health status to be detected (e.g. physical and mental fitness).
- Information on the prevalence of the disease(s) to be diagnosed/detected in the population and sub-populations to be investigated.
- Unambiguous definition of the "gold standard", i.e. of the method by which the disease to be detected (or the health status to be detected) can be unambiguously identified in a generally accepted manner.
- Exact description of the diagnostic test, including details on the material and human resources needed in training for the test, as well as for its execution and evaluation.
- Description of the risks (potential harms) involved in the application of the diagnostic test, and the acceptance and reasonability both for patients and medical personnel, as well as for the general public (e.g. environmental risks).
- Information on any further consequences to be expected from the particular findings (e.g. further diagnostics, therapy, other non-therapeutic interventions, monitoring, lifestyle, basis for informed decisions) describing to what extent these consequences constitute benefits or harms for the patient.
- Details on alternative diagnostic tests and, if necessary, description of the advantages of the new tests over the conventional ones.

2.3.2 Test quality criteria and test characteristics

The evaluation criteria for diagnostic tests are as follows:

I. Technical prerequisites

- Information on diagnostic accuracy.
- Information on diagnostic sensitivity and specificity.
- Information on reproducibility (reliability).
- Variability:
    - Intra-test/rater variability,

  - Inter-test/rater variability,
  - Intra-patient variability (short and long-term).

- Where appropriate, information on the consistency of measurement values with already established standard tests.
- Information on possible confounders (particularly systematic bias effects).

For studies on technical prerequisites, it should be ensured that the corresponding parameters are (also) determined under everyday clinical conditions and in the situation of application. For example, information on the reproducibility of a diagnostic test is usually insufficient if this test has only been tested in healthy persons.

II. Discrimination ability

- Information on diagnostic sensitivity and specificity or, alternatively, information on likelihood ratios. For quantitative methods: presentation of ROC[k] curves with specification of (an) appropriate cutpoint(s) and the rationale for its/their selection (weighting of sensitivity and specificity).

In principle, there are two types of procedures for studies on discrimination ability. Firstly, the application of the diagnostic test in selected persons with a known disease status; secondly, the application of the test in unselected persons with an unknown disease status [5-7]. The first procedure (with corresponding results) is generally the prerequisite for the conduct of a further (usually more extensive) study according to the second procedure. Studies based on the first procedure typically provide an overoptimistic estimate of discriminatory ability [8-10], in particular if seriously ill patients and clearly healthy persons have specifically been chosen for the comparison [11]; this should be taken into account in the evaluation.

III. Prediction

- Information on predictive values (for quantitative methods: information on predictive values with [a] selected cutpoint[s]).

Furthermore, the following basic methodological principles should be considered in the assessment of studies evaluating discrimination ability and predictability of diagnostic tests:

- The research question and associated study design need to be clearly formulated; this includes sample size planning, which can, for example, be orientated towards the desired precision of

[k] Receiver operating characteristics

the estimate (width of the confidence interval) and/or evidence of exceeding a minimum threshold level for specific parameters (e.g. sensitivity, specificity, predictive values).

- The studies should be conducted in a population of patients or persons to whom the test is to be applied in the future (suitable patient spectrum or prevention of spectrum bias).
- The diagnostic test to be evaluated and the gold standard should be assessed independently of each other and in a blinded manner (mutual blinding).
- Confirmation of diagnosis (gold standard) or the type of confirmation of diagnosis (existence of different gold standards) should not be made dependent on the result of the diagnostic test to be evaluated (danger of verification bias). If confirmation of diagnosis cannot be performed in all patients, the selection of patients should be random. However, this does not solve the problem entirely, especially in situations with low a priori probabilities [12].
- Patients with unclear, non-interpretable, or intermediate test results must be considered appropriately in the evaluation (no exclusion without adequate justification) [13].
- If the diagnostic test to be evaluated is (or is to be) embedded in a diagnostic strategy, an isolated assessment of this test is often not meaningful (problem of the dependence of test quality criteria on the combination of diagnostic tests applied).
- If the diagnostic test to be assessed is a constituent of the gold standard, particular methodological problems requiring detailed discussion and consideration may arise.

Experience shows that the above principles are frequently lacking in published diagnostic studies. A reason for this is that the reporting of diagnostic studies is often insufficient [10,14]. It is therefore necessary to describe exactly the methodological deficits of the individual studies in connection with their results, in order, on the one hand, to take this factor into account and, on the other, to be able to make any kind of statement at all. Caution is also necessary if a (statistical) summary of individual results (in terms of a meta-analysis) is planned.

Similar recommendations for studies on diagnostic tests have been published in analogy to the CONSORT statement on therapeutic studies (publications) to achieve as far as possible a uniform and comprehensive presentation of the topic [15]. Whiting et al. have compiled a checklist for the quality assessment of diagnostic studies in systematic reviews [16]

2.3.3 Evidence of a benefit

The mere fact that an illness can be diagnosed (particularly) well or excluded by a specific diagnostic test does not generally mean that a benefit of applying this test can be inferred (in the sense of an improvement of outcome for patients). In fact, a benefit results from the subsequent (mostly therapeutic) consequence. This interaction between diagnostic information and consequence may (but does not have to be) self-evident [3]. If doubts arise about the existence of such an interaction, then comparative studies are required, or rather, such studies are recommended [17]. A similar approach applies if, by a new diagnostic test, more or different cases of the illness to be diagnosed are detected (compared with an established method) [4,18]. Studies conducted to provide evidence of a benefit of diagnostic tests can be designed as a comparison between patients to whom the diagnostic test is applied and patients to whom this test is not applied. The same requirements as formulated in Section 2.4.4 essentially apply to the evaluation of such studies. One disadvantage of such studies is that the value of the diagnostic information cannot be separated from the resulting consequences, i.e. for a negative outcome it cannot be distinguished whether the diagnostic information is insufficient or whether, for example, the therapy (for those with a pathological test result) is ineffective.

As an alternative to assessing the conduct of the test, the disclosure of the test results can be investigated, i.e. persons for whom the test result is known can be compared with those for whom the result remains blinded [7]. Such a procedure offers the advantage of enabling the evaluation of the natural course of the disease in persons with a positive test result.

In another design option, the diagnostic test to be evaluated is applied to all patients in a therapeutic trial (independent of the study group), and the result remains blinded for all patients throughout the whole trial. In this type of study, it can be assessed whether patients experience a different therapeutic benefit, depending on the result of the diagnostic test [6].

References

[1] Knottnerus JA, van Weel C, Muris JWM. Evidence base of clinical diagnosis. Evaluation of diagnostic procedures. BMJ 2002; 324: 477-480.

[2] Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, Lau J. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142: 1048-1055.

[3] Foerster V, Murtagh J, Lentle BC, Wood RJ, Reed MH, Husereau D, Mensinkai S. CT and MRI for selected clinical disorders: A systematic review of clinical systematic reviews [Technology report no 59]. Ottawa: Canadian Coordinating Office for Health Technology Assessment; 2005.

[4] Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: Assessing new tests against existing diagnostic pathways. BMJ 2006; 332: 1089-1092.

[5] Köbberling J, Trampisch HJ, Windeler J. Memorandum for the Evaluation of Diagnostic Measures. J Clin Chem Clin Biochem 1990; 20: 873-879.

[6] Richter K, Lange S. Methoden der Diagnoseevaluierung [Methods for evaluating diagnoses]. Internist 1997; 38: 325-336.

[7] Sackett DL, Haynes RB. Evidence base of clinical diagnosis. The architecture of diagnostic research. BMJ 2002; 324: 539-541.

[8] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical Evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282: 1061-1066.

[9] Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess 2004; 8: iii, 1-234.

[10] Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis. BMC Med Res Methodol 2005; 5: 20.

[11] Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006; 174: 469-476.

[12] Rückmann A, Windeler J. Selektionsbias bei der Schätzung der Sensitivität von Screeningmaßnahmen [Selection bias in the estimation of the sensitivity of screening procedures]. In: Trampisch HJ, Lange S (eds.). Medizinische Forschung – Ärztliches Handeln [Medical Research – Medical Action]. München: MMV Medizin Verlag; 1995. p. 227-231.

[13] Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate and uninterpretable diagnostic test results. Med Decis Making 1987; 7:107-114.

[14] Smidt N, Rutjes AW, van der Windt DA, Ostelo AW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. Radiology 2005; 235: 347-353.

[15] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy. The STARD initiative. Radiology 2003; 226: 24-28.

[16] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3: 25.

[17] FDA. Guidance for Industry. Developing Medical and Imaging Drug and Biological Products. Part 2: Clinical Indications. http://www.fda.gov/cder/guidance/5742prt2.pdf [accessed on 01.12.2006].

[18] Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials. Ann Intern Med 2006; 144: 850-855.

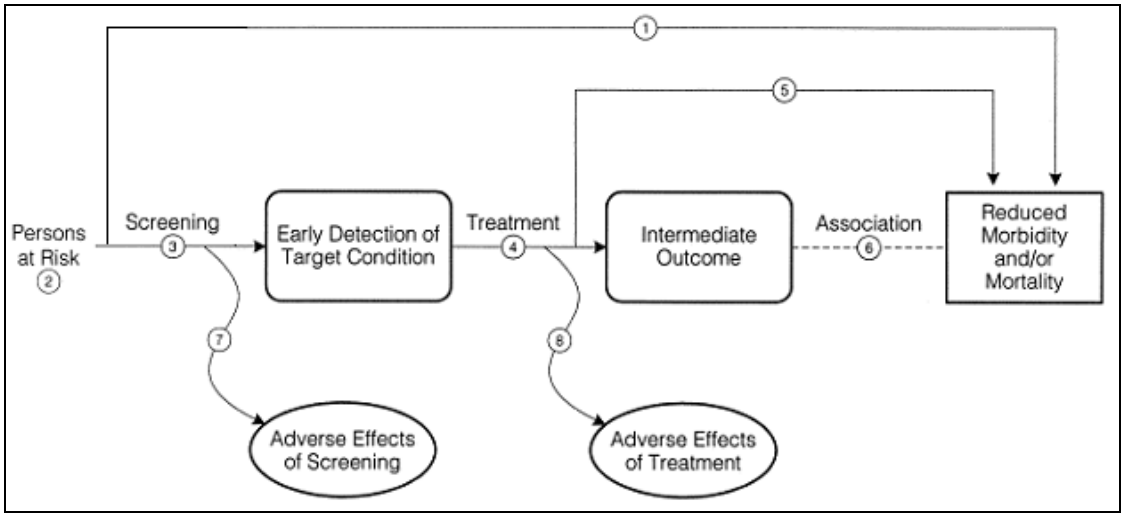
2.4 Screening

Screening programmes are composed of different modules, which can be examined either as a whole or in part. The evaluation of a screening test follows criteria that have already been established and published, for example, by the UK National Screening Committee (UK NSC, [1]) and the US Preventive Services Task Force (US PSTF, [2]).

These evaluation criteria comprise: (1) the disease to be detected; (2) the (diagnostic) screening test to be employed; (3) the type of therapy in the case of a positive (pathological) result or a different consequence derived from a positive result; and (4) the screening programme as a whole.

When positive results are present, the evaluation should discriminate between programmes resulting in therapeutic measures and those resulting in other, non-therapeutic measures. Furthermore, a distinction should be made between (a) situations in which direct evidence for the effectiveness of the screening programme exists and (b) situations in which the evidence is derived indirectly through conclusions by analogy ([a] comparison of persons with regard to a patient-relevant outcome who had or did not have screening within the framework of a study, Arrow 1 in Figure 1; [b] several screening modules are assessed in different studies, e.g. Arrows 3, 4 and 6 in Figure 1).

Figure 1: Screening chain



Modified according to [2].

2.4.1 Target disease

It is to be evaluated whether the disease in question is an important health problem, whereby the evaluation of this aspect can refer to different indicators, e.g. to the frequency, severity or cost of a disease (or to different levels, e.g. population or individual levels).

This requires an exact knowledge of the epidemiology and natural course of the disease. Typical data sources include epidemiological cross-sectional, register, and cohort studies. In exceptional cases, data from case series and economic studies are used.

2.4.2 Screening tests

The general requirements for the assessment of a diagnostic test apply as formulated in Sections 2.3.1 and 2.3.2. An important difference between a diagnostic test and a screening test is the fact that screening tests are (mainly) directed at healthy and symptom-free persons, only few of whom have the target disease. Due to the special ethical implications, higher demands are as a rule to be made on the test quality criteria and the quality of the underlying studies [3]. Moreover, the test should be easy to handle, and it should be assessed whether, in the case of a positive test result, a generally accepted strategy is available for further diagnostic clarification (gold standard) as well as for other available alternatives.

2.4.3 Therapy

For patients with a positive test result and, if appropriate, with subsequent confirmation of diagnosis following further diagnostic tests (gold standard), it needs to be assessed whether an effective treatment or intervention exists (q.v. Section 2.2 for the respective evaluation criteria). In addition, it needs to be assessed whether evidence is available showing that early treatment leads to better results than late treatment. For screening programmes not resulting in immediate therapeutic measures following a positive test result, it should be evaluated whether the information gained from the positive result is associated with a different (non-therapeutic) benefit, e.g. of the kind that allows affected persons to make better informed personal decisions (e.g. prenatal screening for Down syndrome, screening for genetic carriers of incurable diseases). In these cases it may be meaningful to apply decision-analysis methods.

2.4.4 Screening programmes

Ideally, there is available evidence that the screening programme as a whole is effective in reducing morbidity and/or mortality. The criteria formulated in Section 2.2 are used to assess the relevant studies. In particular, it should be considered that in the assessment of screening programmes, evidence from non-randomised studies needs to be assessed critically, as specific bias mechanisms (such as lead time bias or length bias) may occur (q.v. Section 1.18).

If direct evidence for the effectiveness of a screening programme is only available for individual screening modules, then in addition to the assessment of individual modules, an evaluation of their coherence and consistency should be performed. Coherence in this context means that the modules form a comprehensible model; consistency means that different studies or their findings contribute to coherence under different conditions [2].

The screening programme should achieve a net benefit, i.e. the benefit gained from a screening programme should exceed the potential physical or mental damage caused by the screening test or by the subsequent diagnostic measures and/or therapy (q.v. Section 1.6 for the evaluation of adverse effects of an intervention).

If the screening programme or its modules were not assessed in the setting in which the programme is to be implemented, then it needs to be reviewed whether evidence is available showing that the results can be generalised or transferred.

References

[1] UK National Screening Committee. Criteria for Appraising the Viability, Effectiveness and Appropriateness of a Screening Programme. http://www.nsc.nhs.uk/pdfs/criteria.pdf [accessed on 28.10.2004].

[2] Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, Atkins D, Methods Work Group, Third US Preventive Services Task Force. Current methods of the US Preventive Services Task Force: A review of the process. Am J Prev Med 2001; 20 (Suppl 3): 21-35.

[3] Ewart RM. Primum non nocere and the quality of evidence: Rethinking the ethics of screening. J Am Board Fam Pract 2000; 13: 188-196.

2.5 Health economics

2.5.1 Background

According to § 139a (3) Social Code Book V, IQWiG comments on issues of fundamental relevance to the quality and efficiency of services provided within the framework of the statutory health insurance (SHI) system. The responsibilities of the Department of Health Economics therefore include the investigation of health economic questions commissioned to IQWiG by the Federal Joint Committee or Federal Ministry of Health or developed within the framework of IQWiG's general commission. According to current legislation, IQWiG does not conduct cost-benefit assessments for pharmaceuticals.

2.5.2 Aim of health economic evaluations

The aim of health economic evaluations is to provide scientifically founded decision aids for the allocation of health care services in terms of an "evidence-based health policy". The improvement of health care by optimising the use of resources is the goal of these decision processes.

2.5.3 Methodology

In order to validly interpret health economic evaluation studies, they must show sufficient transparency and methodological quality. The methodological quality of health economic studies is achieved by adherence to indispensable minimum standards. These standards are orientated towards the internationally agreed scientific criteria set by the *US Public Health Service Panel on Cost-Effectiveness in Health and Medicine* [1], the *BMJ Economic Evaluation Working Party* [2], the *Policy on Cost-Effectiveness Analyses of the New England Journal of Medicine* [3], as well as the *German Recommendations on Health Economic Evaluations* [4]. Health economic evaluations not fulfilling these criteria are not taken into account.

In addition, health economic evaluations must orientate their methodological approaches to the context of health care in Germany and take the particular aspects of the German health care system into consideration.

Transferability

The transferability of health economic findings to the general conditions of the German health care system must be given. The transferability of study results from another country to Germany is influenced particularly by the following factors: demography of the country examined; epidemiology of the disease investigated (incidence, prevalence); availability of health care resources (provision of therapeutic services and health care facilities, access routes for patients); variations in clinical practice; reimbursement of service providers (incentive systems); organisational structures; relative prices and costs; and moral concepts of the population investigated [5,6].

The assessment of the transferability of results of studies conducted in other countries requires that the methods applied, the data basis, and the study results have been described in a transparent and comprehensible manner [7,8]. Before the transferability of studies is assessed, the following questions should be answered positively [8]:

1. Is the technology investigated comparable to the technology used in Germany?
2. Is the comparator intervention investigated relevant to Germany?
3. Is the study of acceptable quality, i.e. does it fulfil the international standards required?

If any one of these questions cannot be answered positively, then the transferability of the study results is not given..

Perspective

The costs to be outlined, their investigation, as well as the resultant findings, in each case depend on the adopted perspective [9]. Health economic studies can be conducted from a variety of perspectives, for example, those of society, the payer, the service provider or the patients. According to its remit following § 139a (3) Social Code Book V, IQWiG is active with regard to issues of fundamental relevance to the quality and efficiency of services provided within the framework of the SHI system. The evaluation of costs from the SHI perspective is therefore of particular relevance to IQWiG's legal remit [9].

Beyond the SHI perspective, the Institute will, depending on the research question posed, select the suitable perspective to evaluate cost-efficiency. For example, when evaluating interventions that have an effect on rehabilitation, need for nursing care, or incapacity to work, it can be meaningful to assume the perspective of the social insurance carrier or of society as a whole.

Definition of the decision problem

In an economic evaluation, the aim of the analysis must primarily be specified. This involves the definition of the procedure for which an economic evaluation is to be conducted, as well as the patient groups and indication relevant to this procedure. It must also be described for which region and decision-maker the analysis is to be performed [10].

Comparator standards

Within the framework of health economic evaluations, a comparison between two interventions is sought, whereby one represents the intervention to be evaluated and the other represents the comparator intervention. When selecting the comparator method, it needs to be considered that this corresponds to the health care standard currently applied in practice [11,12], is widely used in the German health care system, and is accessible to recipients of health care services. Where several possible comparable treatment procedures are available, all should be mentioned and at least one included in the evaluation as a comparator intervention [13]. The choice of comparator intervention should be sufficiently justified and exactly described.

Target population

The target population profiting from the intervention investigated should be clearly described. Specific variables that sufficiently characterise the study population must therefore be given, such as age, gender, socio-economic status, previous diseases, risk profile, etc. [10]. The study population must be representative of the target population in Germany.

Data sources and study design

Evaluations using clinical trial data (trial-based economic evaluations) have high priority, whereby their limitations with regard to economic conclusions are taken into account [6]:

- The comparator treatment is not necessarily the relevant comparator treatment for health economic decisions.
- The study populations are subjected to a strict selection procedure by the inclusion and exclusion criteria specified in the study protocol.

- The timeframe is usually designed as short-term, so that long-term health economic effects (e.g. in- and out-patient nursing and rehabilitation measures, recurrent hospital admissions) that follow the acute phase and overlap health care sectors are not sufficiently documented.
- The medical services specified in the study protocol may lead to an increased utilisation of resources, which may considerably differ from health care under day-to-day conditions (protocol-driven costs).

In addition to RCTs, study designs are therefore necessary that ensure the transferability of study results to health care reality [14]. These requirements are fulfilled within the framework of so-called pragmatic trials. In these trials, concepts that minimise bias are also applied; however, fewer restrictions with regard to the recruitment of patients and the course of the trial are imposed. The following requirements should be fulfilled when designing pragmatic trials [6]:

1. Patients showing typical disease severity are included.
2. The comparator intervention corresponds to the health care standard currently applied in practice.
3. The setting and the physicians participating correspond to health care reality.
4. All patients included are treated under routine conditions.
5. A broad spectrum of endpoints is measured (effectiveness, utilisation of resources, quality of life, costs, etc.)

Epidemiological and register studies are suitable for the conduct of such long-term trials.

Modelling

Decision-analytic modelling is not to be regarded as a substitute to obtain reliable evidence for the demonstration of a benefit of an intervention [1,14,15]. However, clinical trials do not always provide all the necessary data for a complete economic evaluation. In these situations, which in each case have to be specifically assessed, modelling may be a useful tool to support the decision-making process [16]. This includes, for example, analysis of the long-term costs of interventions, or the analysis of savings achieved by the prevention of events [17]. In this context, possible errors caused by the extrapolation of data from clinical trials whose timeframe is too short need to be taken into account [1]. Modelling to extrapolate data should only be performed if data of sufficient quality are available for the particular research question posed [14]. The decision to employ a

model to evaluate the cost-efficiency of an intervention must be sufficiently justified. Furthermore, decision-analysis models must meet the following quality standards [2,3,18]:

1. Decision-analysis models must be transparently and comprehensibly described (see "*A suggested checklist for assessing quality in decision analytic models"* [19]).
2. The underlying assumptions of the decision-analysis models must be justified.
3. The data used must correspond to the requirements (see above).
4. The choice of variables for the sensitivity analysis, the range in which the variables are modified, and the results of the sensitivity analysis must be described in a transparent and comprehensible manner.
5. If different models come to different conclusions about the same research question, the reasons for these differences need to be explained by the model developers (cross validation).
6. Decision models should be validated: The mathematical calculations must be reviewed with regard to their correctness and consistency with the model specifications. It must be ensured that the model input data and the outcomes are consistent with the data available.

Utilities

A cost-utility analysis is a type of health economic evaluation in which the benefit of the alternatives to be evaluated is expressed as a product of the health-related quality of life and the years of life with this quality of life. Health-related quality of life is evaluated with the help of utilities, which are included in a scale from 0 (death) to 1 (perfect health). The product from the gain in quality of life and the associated years of life is expressed as the number of quality adjusted life years (QALYs) gained. One QALY is equivalent to a year of life in perfect health.

This approach, which initially appears plausible, is not free of contradiction when assessed in detail, and has so far led to a continuing debate about its theoretical and practical problems. The methods used to measure a person's state of health, as well as to generate utilities, diverge substantially. When evaluating a person's state of health, depending on the evaluation procedure used (Standard Gamble, Time Trade-Off, Visual Analogue Scale) and the reference population, the methods in each case deliver different results. The fundamental issue of the comparability of methods is therefore raised. To date, there is no reliable method to equate the results of the different procedures [20,21]. The problem of comparability is illustrated by the results of a review recently published in Germany [22]. In a systematic review of 18 cost-utility analyses performed in Germany, Schwappach et al. showed that QALYs are not universally applicable and are not

comparable. Furthermore, the utility values applied in the majority of studies do not reflect preferences of patients or of the population in Germany [22].

The underlying assumption of a cost-utility analysis is that, with a given budget, it is rational for society to maximise the sum of individual benefits of interventions (expressed as the aggregation of individual QALYs). Studies show that the distribution of resources on the basis of aggregated individual QALYs is not accepted by the population [23-25]; in particular there is a wish to distribute resources equitably and to reduce inequalities [23,24,26]. The QALY concept is also being criticised for discriminating against elderly, disabled, and chronically ill people [24,27]. This therefore concerns crucial aspects of our social values, such as equity of accessibility and allocation [23].

The use of QALYs as a "virtual" uniform measure to describe results related to health care is to be judged as insufficient and should not be drawn upon as a basis for decision making. Institutions that use QALYs as a measure of cost-benefit analyses are also aware of their methodological shortcomings [28].

Study type

If several alternative procedures exist with regard to a standardised treatment success, a cost-effectiveness analysis should be performed. Cost-minimisation analyses should be conducted if it can be shown that the alternatives investigated lead to the same medical results [13].

Costs

Depending on the research question posed, all costs relevant to the economic evaluation should be identified. The utilisation of resources should be orientated towards clinical practice in Germany [29].

The utilisation of resources (e.g. duration of hospitalisation, utilisation of medicines and medical aids, medical consultations, etc.) should be presented separately from the corresponding costs per unit [2]. The methods of determining costs per unit are to be outlined. The date of pricing and any adaptations due to inflation or currency conversion must also be presented [2]. As health economic evaluations are particularly consulted by decision-makers in health policy, the relevant costs should be presented as average values [2].

Timeframe

The timeframe of health economic studies should be so extensive that all effects and costs arising from the measure investigated are documented. In this context, it should be noted that the acquisition of economically relevant data is possibly only to be expected in a medium- to long-term timeframe.

Endpoints

Health economic evaluations should in principle employ endpoints orientated towards a patient benefit (q.v. Section 2.1).

Presentation of results

The results of health economic evaluations should be presented in such a way that all relevant methodological aspects of the evaluation are transparent and easily understandable [29].

Outlook

Health economic evaluations are based on generally valid scientific principles and methods. In respect of their theoretical development, these methods require continuous further development, as do the standards of practical implementation [30]. To the same degree in which questions are increasingly asked about the efficiency of medical procedures, health economic evaluations in the German health care system will become increasingly important. Therefore, for the health policy dialogue, it is mandatory that the health economic methods and their underlying presumptions have to be understood and accepted.

For health politics in general, as well as for decision makers in particular, the critical appraisal and the assessment of the consequences of health economic evaluations are indispensable. For this purpose, complete transparency of methods is necessary.

Within the shared risk community, the acceptance of health economics as a decision tool will also depend on the degree to which a harmonisation of its methods with social values concerning equity of accessibility and allocation.

It will therefore be a principal task of the department to implement a structured consultation process in respect of new methods of cost-benefit assessment after the passage of the “Act to promote competition” of the SHI *(GKV Wettbewerbsstärkungsgesetz)*. This dialogue will include

all those involved in the health care system and will be conducted in a completely transparent manner.

References

[1] Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. JAMA 1996; 276: 1253-1258.

[2] Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 1996; 313: 275-283.

[3] Kassirer JP, Angell M. The Journal's policy on cost-effectiveness analyses. N Engl J Med 1994; 331: 669-670.

[4] Bestehorn K, Biller M, Brecht JG, Clouth J, Fricke F-U, Glaeske G, Greiner W et al. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation – Revidierte Fassung des Hannoveraner Konsens [German recommendations on health economic evaluations - revised version of the Hanover consensus]. Gesundh ökon Qual manag 1999; 4: A62-A65.

[5] Busse R. Gesundheitsökonomie. Ziele, Methodik und Relevanz [Health economics. Aims, methodology, and relevance]. Bundesgesundhbl Gesundheitsforsch Gesundheitsschutz 2006; 49: 3-10.

[6] Drummond MF, Sculpher MJ, Torrance GW, O'Brien B, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press; 2005.

[7] Boulenger S, Nixon J, Drummond M, Ulmann P, Rice S, de PG. Can economic evaluations be made more transferable? Eur J Health Econ 2005; 6: 334-346.

[8] Welte R, Feenstra T, Jager H, Leidl R. A decision chart for assessing and improving the transferability of economic evaluation results between countries. Pharmacoeconomics 2004; 22: 857-876.

[9] Luce BR, Elixhauser A. Estimating costs in the economic evaluation of medical technologies. Int J Technol Assess Health Care 1990; 6: 57-75.

[10] Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H et al. Generalisability in economic evaluation studies in healthcare: a review and case studies [online]. Last update: 2004.
http://www.ncchta.org/htacd.htm [accessed on 16.03.2006].

[11] Canadian Agency for Drugs and Technologies in Health. Guidelines for the Economic Evaluation of Health Technologies: Canada [online]. Last update: 2006.
http://www.cadth.ca [accessed on 08.11.2006].

[12] National Institute for Clinical Excellence. Guide to the Methods of Technology Appraisal (Reference N0515) [online]. Last update: 07 May 2004.
http://www.nice.org.uk/download.aspx?o=201973 [accessed on 19.09.2006].

[13] Members of the Collège des Économistes de la Santé [the French Health Economists Association]. French Guidelines for the Economic Evaluation of Health Care Technologies [online]. Last update: 2004.
http://www.ces-asso.org/docs/France_Guidelines_HE_Evaluation.PDF
[accessed on 09.02.2006].

[14] Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, Vray M. Modelling in economic evaluation: An unavoidable fact of life. Health Econ 1997; 6: 217-227.

[15] Sheldon TA. Problems of using modelling in the economic evaluation of health care. Health Econ 1996; 5: 1-11.

[16] Commonwealth of Australia Department of Health and Ageing. 1995 Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee: including major submissions involving economic analyses [online]. Last update: 2002. http://www.health.gov.au/internet/wcms/publishing.nsf/Content/health-pbs-general-pubs-pharmpac-gusubpac.htm [accessed on 08.11.2006].

[17] Siebert U. When should decision-analytic modeling be used in the economic evaluation of health care? Eur J Health Econ 2003; 4 143-150.

[18] Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, Luce BR. Principles of good practice for decision analytic modeling in health-care evaluation: Report of the ISPOR Task Force on Good Research Practices - Modeling Studies. Value Health 2003; 6: 9-17.

[19] Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment [online]. Last update: 2004. http://www.hta.nhsweb.nhs.uk/ [accessed on 16.08.2006].

[20] Gafni A, Birch S. Guidelines for the adoption of new technologies: a prescription for uncontrolled growth in expenditures and how to avoid the problem. CMAJ 1993; 148: 913-917.

[21] McGregor M, Caro JJ. QALYs. Are they helpful to decision makers? Pharmacoeconomics 2006; 24: 947-952.

[22] Schwappach DL, Boluarte TA. Wie werden in deutschen Studien qualitätsadjustierte Lebensjahre definiert? Ein systematischer Review deutscher Kosten-Nutzwert-Analysen. [How are quality adjusted life years defined in German studies? A systematic review of German cost-utility analyses]. Dtsch Med Wochenschr 2006; 131: 2004-2009.

[23] Coast J. Is economic evaluation in touch with society's health values? BMJ 2004; 329: 1233-1236.

[24] Dolan P, Shaw R, Tsuchiya A, Williams A. QALY maximisation and people's preferences: a methodological review of the literature. Health Econ 2005; 14: 197-208.

[25] Nord E. Towards cost-value analysis in health care? Health Care Anal 1999; 7 167-175.

[26] Loomes G, McKenzie L. The use of QALYs in health care decision making. Soc Sci Med 1989; 28(4): 299-308.

[27] Hadorn DC. The problem of discrimination in health care priority setting. JAMA 1992; 268: 1454-1459.

[28] National Institute for Health and Clinical Excellence. Social Value Judgements. Principles for the Development of NICE Guidance [online]. Last update: 2005. http://www.nice.org.uk/ [accessed on 27.11.2006].

[29] da Silva EA, Pinto CG, Sampaio C, Pereira JA, Drummond MF, Trindade R. Guidelines for Economic Drug Evaluation Studies [online]. Last update: 1998. http://www.infarmed.pt/portal/page/portal/INFARMED [accessed on 09.02.2006].

[30] Bridges JF. Future challenges for the economic evaluation of healthcare: Patient preferences, risk attitudes and beyond. Pharmacoeconomics 2005; 23: 317-321.

2.6 Clinical practice guidelines and disease management programmes

2.6.1 Background of guideline evaluation

Clinical practice guidelines (CPGs) are seen as key instruments in the improvement and assurance of medical quality in health care [1]. Their objective is to reduce inappropriate differences in patient care and improve care by means of the formulation of concrete recommendations for clinical decision making. Furthermore, in Germany they are used as a basis for decisions on steering procedures in the health care system, e.g. for the formulation of requirements for disease management programmes (DMPs), in acc. with § 137f Social Code Book V. Consequently, CPGs are increasingly influencing decisions affecting the structural level of the German health care system.

Against this background, it should be ensured that CPGs are based on the best available and most up-to-date scientific evidence, and are formulated after due consideration of clinical experience.

However, in many cases the reference to current scientific evidence is lacking [2,3], and CPGs on identical topics in part reveal considerable differences with regard to the content of their recommendations [4,5].

One important reason for this is that the internationally stipulated quality standards for CPG development are not consistently followed [6-8].

2.6.2 Aim of guideline evaluation

The evaluation of CPGs aims to improve care through greater transparency in the health care system. It is therefore particularly important to:

- Discriminate between CPGs of good or bad methodological quality and quality of content;
- Elaborate and review the evidence base on which key recommendations of CPGs are founded;
- Make clear and specific statements on the reasonability and effectiveness of the implementation of different medical recommendations;
- Offer the Federal Joint Committee or its panels a basis for decisions in discussions on DMPs;

- Ensure that only verified (quality-assured) CPGs, where indications of an improvement in outcome exist, are introduced into health care;
- Identify research needs and initiate meaningful projects for the development and implementation of evidence-based recommendations;
- Promote the inclusion of CPGs in total quality management (TQM).

Furthermore, the results of this work provide the users of CPGs (physicians, health facilities, health policy committees, decision-makers in the health care system, and patients) with orientation regarding meaningful and appropriate recommendations on high-priority health care problems.

For the specific evaluation of content of CPGs, the available methodological competence and expertise of external institutes, facilities or organisations are to be used and involved as far as possible [9].

2.6.3 Methods of guideline evaluation

Essential aspects in the assessment and review of the quality of CPGs are the:

- Examination of formal criteria, which essentially reflect the transparency of the developmental process, and assume the precondition that CPGs considering these criteria will reach correct recommendations with a higher probability (comparable to the internal validity of studies) [10,11];
- Exact assessment of content in view of the underlying evidence;
- Evaluation of the appropriateness of recommendations;
- Evaluation of the effects caused by implementing CPGs (outcome evaluation).

In part, the approaches vary substantially according to effort invested, approach used, and evidential value, and are applied according to the research question and the commission. In this context, the methodology applied by the Institute will be reviewed regularly and, if necessary, updated under consideration of current scientific publications as well as national and international experience. Different aspects are usually combined with each other in a step-by-step procedure [9]. The restriction to largely formal and methodological aspects of the evaluation, which can be operationalised well, was critically discussed before the revision of the methods. On the one hand, these restrictions are owed to the current legal framework (Social Code Book V § 139a) which so far (only) envisages a review of CPGs with regard to the underlying evidence. On the other, the methods for an evidence-based (and consented) development and update of guidelines are still

insufficiently applied [12]. For this reason, a formal evaluation cannot be abandoned for the present.

However, what is scientifically correct does not necessarily have to be meaningful, practicable, and appropriate. Individual recommendations and key points in CPGs may be assessed very differently by those affected with regard to their relevance, appropriateness, and practicability.

A review of the "appropriateness of content" (for which there is so far no internationally harmonised procedure, although approaches are noticeable in Germany in the "National Disease Management Guidelines Programme") needs to consider different aspects and also involve other players, such as the Federal Joint Committee, professional representatives, and patients.

Statements on the appropriateness of content of guideline recommendations comprise two key questions:

1. Are the presentation and interpretation of current evidence, from which concrete individual recommendations are generated, appropriate and comprehensible? This also includes the case that deviations occur; however, these must be justified. (The positive answer to this question is a prerequisite for approaching the second question).
2. Is the intervention recommended appropriate?

To answer these questions, the following aspects need to be considered:

- A benefit of an intervention must have been demonstrated, and the weighing of benefits/harms must have come to a positive decision (q.v. Section 2.1).
- The intervention must be relevant to the German health care system.
- The intervention must be available and approved.
- The intervention must be necessary.
- Safe application (by physician/patients) is possible.
- The intervention can be implemented and financed with the resources available.

In a shared risk community, particularly the last point requires a generally agreed decision on the available overall financial frame as well as on the prioritisation of health care services. So far the Federal Joint Committee, at least indirectly and partly, has taken on this function with regard to the inclusion and exclusion of services from the benefits catalogue of statutory health insurance, and forms a consensus between the different interests and stakeholders (including patients). It is conceivable that in future the Federal Joint Committee will also take on a similar function within the framework of the evaluation of the content of CPGs (in respect of their appropriateness), and

will publish corresponding national CPGs after the assessment of the evidence base by IQWiG. Against the background of the assessment of the appropriateness of guidelines, IQWiG will enter a discussion process in 2007, in which expectations with regard to such a procedure, and the aims, methods, and responsibilities need to be structured.

I. Formal assessment

An approach to the question of guideline quality can be made by formal assessment, following methodological criteria [5,6,13,14]. The authors of these publications establish a direct reference to the "validity" of CPGs, but use this term in a very undifferentiated way. In the following text it will not therefore be used in connection with CPGs. Formal CPG evaluation is conducted in a structured manner following the methods of CPG clearing procedures, and referring to CPG evaluation criteria of the German Medical Association *(Bundesärztekammer)* and the Association of Statutory Health Insurance Physicians *(Kassenärztliche Vereinigung)* by means of DELBI,[1] the German Guideline Evaluation Instrument [14,15], which is the German adaptation of the validated AGREE instrument, and includes an additional domain that refers to the applicability of the particular CPG to the German health care system. The formal evaluation is performed by two independent scientists. Where conflicting assessments are made, the issues will be discussed and evaluated once again. If the dissent continues and cannot be solved either by a query to the authors, the matters of dissent will be separately documented.

Several CPGs, often differing greatly in methodology and content, exist worldwide on specific medical issues [4]. The formal assessment of CPGs has an important filter function, which ultimately enables the performance of an evaluation of the content of key statements and specific recommendations of the relevant CPGs (see below).

First, a comprehensive literature search in the relevant databases (CPG and additional literature databases) is conducted in order to identify CPGs currently available on the particular research question (under consideration of the procedures outlined in Section 4.7 [Literature search]). The filtering procedure for CPG evaluation has a multi-step approach. Inclusion and exclusion criteria and the search strategy (search terms, choice of data bases, etc.) are determined and documented in advance, depending on the research question. A first screening step follows in which the hits are selected according to predefined thematic criteria. A short methodological evaluation takes place in

[1] Deutsches Leitlinien-Bewertungs-Instrument *(German Guideline Evaluation Instrument)*

a second step, in which those CPGs are selected that comply with the international minimum standards; an evaluation of content of the resulting documents follows (see below).

II. Comparison of guidelines and evaluation of content of key recommendations

The evaluation of CPG content is of special relevance. The criteria checked so far with conventional instruments ("ÄZQ-Checkliste",[m] AGREE,[n] DELBI) for the identification and interpretation of evidence and for the formulation of CPG recommendations are essentially transparency criteria[o] in which only the description of the process, e.g. for literature searches, can be positively evaluated, without providing an assessment of the completeness and topicality of the search. For the essential key recommendations included in a CPG, their derivation from the underlying evidence must therefore be individually assessed. Besides the evaluation of completeness and topicality of the literature consulted, the assessment of content also includes the evaluation and interpretation of study results. As this procedure involves much time and effort, for pragmatic reasons the assessment of content must be limited to the research questions commissioned by the Federal Joint Committee or to the CPG's key recommendations. Identification of the key recommendations is made within the context of each specific commission in consultation with the Department Heads and external experts concerned. In this context, focussing on selected key issues has several advantages. As CPGs, even if they refer to the same medical topic, are very heterogeneous with regard to their content and recommendations, one can achieve a good comparison between different CPGs by reducing the scope of evaluation to these predefined key issues. Within this framework, a review of the underlying evidence and the resulting recommendations is also possible, provided that there is sufficient transparency in this regard. A comparison of content in respect of the key issue defined is not only possible between CPGs by different publishers, but also between CPGs and other evidence-based sources (systematic reviews, HTA reports, evidence-based reports produced by IQWiG, etc.). Our department also works on individual research questions not addressed in CPGs after consultation with IQWiG's other departments.

A synoptic comparison of the content of CPGs can be helpful in identifying key recommendations. In particular, questions that are the subject of scientific dissent can be identified.

---

[m] Ärztliches Zentrum für Qualität in der Medizin. In: AWMF (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften) 2001, Leitlinien-Manual. [Agency for Quality in Medicine. In: Association of the Scientific Medical Professional Societies, Guideline manual].

[n] Appraisal of Guidelines Research and Evaluation

[o] Efforts are currently being made to initiate methodological improvements to these aspects (e.g. by the GRADE [Grading of Recommendations, Assessment, Development and Evaluation] working group [10]).

Methodologically, the synoptic comparison only facilitates the evaluation process. An evaluation of the underlying evidence is also meaningful for procedures that are consistently recommended. In particular, it should also be assessed whether the outcome parameters included in the CPG are relevant to patients, and whether a weighing of benefits and harms [10] was considered in the formulation of the recommendation.

III. Improvement of outcomes

The central question in the preparation and implementation of CPGs is whether their implementation leads to a measurable improvement in health care [1,16,17]. Strictly speaking, this can only be assessed by a rigorous evaluation of effects [2]. However, due to the financial expenditure and use of human resources, this cannot be realised for every existing CPG. For example, only unsystematic studies with heterogeneous research questions and results are available on the topic "Outcome evaluation after guideline implementation". Furthermore, pilot studies have so far only been conducted for a few guidelines before publication [18,19].

If results of pilot studies or projects testing CPGs are available, these are to be included in the overall evaluation (e.g. by describing methods, quality indicators, results and consequences).

Furthermore, it is meaningful to compare CPG recommendations with the conventional procedures applied in routine health care. If complex changes are recommended in a CPG, its implementation is more difficult and has to be accompanied by supportive measures and tools [5].

In particular, CPGs from other countries must be assessed as to the transferability of their conclusions to the German health care system and/or to the structural prerequisites needed for their successful implementation. The Institute can also be commissioned by the Federal Joint Committee to evaluate CPGs.

2.6.4 Presentation of quality assessment

Structured reports (Guideline Assessment Reports) are prepared from the results of the evaluation to provide the Federal Joint Committee with a basis for further consultations.

The reports can also serve as a basis for producing topic-related information for physicians and patients, or can be used by professional societies as an aid to revise CPGs.

2.6.5 Submission of recommendations on disease management programmes

The Federal Joint Committee, in accordance with § 91 Social Code Book V, names diagnoses to be included in DMPs following § 137f Social Code Book V, and develops requirements with regard to the content of these programmes. According to § 139a (3) Social Code Book V, it is IQWiG's task to issue recommendations on DMPs. This includes:

- Supporting the Federal Joint Committee in naming new diagnoses to be included in DMPs;
- Revising existing requirements for DMPs;
- Developing new requirements for the content of DMPs.

The concrete possibilities of supporting the Federal Joint Committee and its panels beyond the commissioning of specific research questions within the framework of IQWiG's evaluations are currently being assessed in consultation with the panels responsible.

References

[1] Entwicklung einer Methodik für die Ausarbeitung von Leitlinien für optimale medizinische Praxis. Empfehlung Rec. (2001) 13 des Europarates und erläuterndes Memorandum. [Development of a methodology for guideline development for optimal medical practice. Recommendation Rec. (2001) 13 of the European Council and explanatory memorandum]. Z ärztl Fortb Qual sich 2001; 95 (Suppl. III): 96.

[2] Helou A, Ollenschläger G. Ziele, Möglichkeiten und Grenzen der Qualitätsbewertung von Leitlinien. Ein Hintergrundbericht zum Nutzermanual der Checkliste „Methodische Qualität von Leitlinien". [Aims, possibilities, and limits of the quality evaluation of guidelines. A background report on the user manual for the checklist "Methodological quality of guidelines"]. Z ärztl Fortbild Qual sich 1998; 92: 361-365.

[3] Savoie I, Kazanjian A, Bassett K. Do clinical practice guidelines reflect research evidence? J Health Serv Res Policy 2000; 5: 76-82.

[4] Irani J, Brown CT, van der Meulen J, Emberton M. A review of guidelines on benign prostatic hyperplasia and lower urinary tract symptoms: are all guidelines the same? BJU Int 2003; 92: 937-942.

[5] Burgers JS, Cluzeau FA, Hanna SE, Hunt C, Grol R, and the AGREE Collaboration. Characteristics of high quality guidelines: Evaluation of 86 clinical guidelines developed in ten European countries and Canada. Int J Technol Assess Health Care 2003; 19: 148-157.

[6] Cluzeau F, Littlejohns P, Grimshaw J, Feder G, Moran S. Development and application of a generic methodology to assess the quality of clinical guidelines. Int J Qual Health Care 1999; 11: 21-28.

[7] Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed literature. JAMA 1999; 281: 1900-1905.

[8] On Care Pathways, Bandolier Forum 2003; 7: 1-12.

http://www.jr2.ox.ac.uk/bandolier/Extraforbando/Forum2.pdf [accessed on 17.11.2006].

[9] Fervers B, Burgers JS, Haugh MC, Latreille J, Mlika-Cabanne N, Paquet L et al. Adaptation of clinical guidelines: Literature review and proposition for a framework and procedure. Int J Qual Health Care 2006; 18:167-176.

[10] Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B et al. Grading strength of recommendations and quality of evidence in clinical guidelines: Report from an American College of Chest Physicians Task Force. Chest 2006; 129:174-181.

[11] Grimshaw J, Eccles M, Russell I. Developing clinically valid practice guidelines. J Eval Clin Pract 1995; 1: 37-48.

[12] Ärztliches Zentrum für Qualität in der Medizin. Das Deutsche Leitlinien-Clearingverfahren 1999-2005, Abschlussbericht. [Agency for Quality in Medicine. The German guideline clearing procedure 1999-2005. Final Report]. ÄZQ-Schriftenreihe Band 25; Books on Demand, Norderstedt; 2006.

[13] Field MJ, Lohr KN. Clinical practice guidelines. Directions from a new program. Institute of Medicine, Washington D.C.; 1990.

[14] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften, Ärztliche Zentralstelle Qualitätssicherung. Das Leitlinien-Manual. [Association of the Scientific Medical Professional Societies. Agency for Quality Assurance in Medicine. Guideline Manual]. Z ärztl Fortbild Qual sich 2001; 95 (Suppl. I): 1-84.

[15] Ärztliche Zentralstelle Qualitätssicherung. Checkliste „Methodische Qualität von Leitlinien“. [Agency for Quality Assurance in Medicine. Checklist “Methodological quality of guidelines”]. Dtsch Ärztebl 1998; 95: A2576-A2578, C1838-C1840

[16] Grol R, Dalhuijsen J, Thomas S, Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use of guidelines in general practice: Observational study. BMJ 1998; 317: 858-861.

[17] Worrall G, Chaulk P, Freake D. The effects of clinical practice guidelines on patient outcomes in primary care: A systematic review. CMAJ 1997; 156: 1705-1712.

[18] Ärztliches Zentrum für Qualität in der Medizin. Leitlinien-Clearingbericht „COPD“. [Agency for Quality in Medicine. Guideline clearing report “COPD”]. ÄZQ-Schriftenreihe Band 14; Verlag Videel, Niebüll; 2003.

[19] Ärztliches Zentrum für Qualität in der Medizin. Leitlinien-Clearingbericht „Depression“. [Agency for Quality in Medicine. Guideline clearing report “Depression”]. ÄZQ-Schriftenreihe Band 12; Verlag Videel, Niebüll; 2003.

2.7 Systematic reviews and HTA reports

Systematic reviews and HTA reports provide a summarised presentation of the current status of evidence and are a valuable basis for decisions to be made by physicians and patients, as well as for decision-makers in the healthcare system. The Institute uses systematic reviews and HTA reports for the production of its reports and health information. The prerequisite is that the methodology of these reviews meets the Institute’s requirements. In particular, there must be an explicit literature search and selection strategy. These, as well as the inclusion and exclusion criteria, also need to meet predefined criteria. An essential aspect of the evaluation, in addition to the description of the

review's methodology, is that the quality of the included studies was assessed and reported, and whether this quality assessment influenced any synthesis of individual study results [1,2].

Like any other scientific publication, a systematic review or an HTA report could reach incorrect conclusions, but may also be affected by different types of bias [1]. In principle, considerable differences in quality exist among systematic reviews and HTA reports. For that reason, not every review will justify classification as a high level of evidence.

The number of systematic reviews and HTA reports has increased substantially in recent years; approximately four systematic reviews and HTA reports are currently published per day [3]. This brings with it problems, such as the existence of multiple systematic reviews with contradictory conclusions [4].

In the production of the Institute's reports (q.v. Section 4.4), systematic reviews and HTA reports are primarily used to identify potentially relevant (primary) studies (q.v. Section 4.7.2). Health information produced by the Institute is in large part based on systematic reviews and HTA reports (q.v. Section 3.3.3). However, an IQWiG report will only be based solely or partially on such studies in exceptional circumstances, for example, in especially urgent evaluations or for assessing a partial aspect of a complex issue.

A prerequisite for the inclusion of systematic reviews and HTA reports in the Institute's health information and, if applicable, in its reports, is a methodological assessment following the quality index for systematic reviews by Oxman and Guyatt [5-7]. In addition to this formal assessment, an assessment of content is made (for example, the evaluation of a random sample of primary studies included in the systematic review, or the assessment of individual studies that may dominate the review).

References

[1] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM statement. Lancet 1999; 354: 1896-1900.

[2] Chalmers I, Haynes RB. Reporting, updating, and correcting systematic reviews of the effects of health care. In: Chalmers I, Altman DG, eds. Systematic Reviews. London: BMJ Publishing Group 1995: 86-95.

[3] Bastian H, Glasziou P. Four systematic reviews per day: the epidemiology of research synthesis. Cochrane Colloquium, Melbourne; Oktober 2005.

[4] Jadad AR, Cook DJ, Browman G. A guide to interpreting discordant systematic reviews. CMAJ 1997; 156: 1411-1416.

[5] Jadad AR. Randomised Controlled Trials: A User's Guide. London: BMJ Books; 1998.

[6] Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 44: 1271-1278.

[7] Oxman AD, Guyatt GH, Singer J, Goldsmith CH et al. Agreement among reviewers of review articles. J Clin Epidemiol 1991; 44: 91-98.

2.8 Prognosis

An essential basis for the evaluation of prognostic studies is the precise formulation of a research question, as studies conducted to evaluate prognostic characteristics have different objectives (e.g. evaluation of risk factors, score development or validation). The discrimination between diagnosis and/or screening studies can be difficult and, depending on the study objectives, different evaluation principles to assess prognosis studies are applied.

A prognostic characteristic provides information that should not be an end in itself but should lead to a consequence that constitutes a verifiable benefit for the patient. In this respect, the (general) requirements applying to a prognostic procedure are similar to those applying to a diagnostic test (q.v. Section 2.3). If a prognostic characteristic is to be applied in terms of a screening or prevention programme, then the principles formulated in Sections 2.4 and 2.10 need to be considered in the assessment.

No generally accepted quality criteria exist for evaluating prognosis studies [1,2]. Simon and Altman describe guidelines for the planning and conduct of prognosis studies in oncology [1]. Laupacis et al. suggest a general framework to evaluate prognosis studies [3]. The aspects listed below, which result from the underlying data source and the data analysis applied, should always be considered. As multifactorial regression models often play a central role in prognosis studies, Section 1.22 should also be considered.

The following points are especially relevant:

- Clear formulation of a research question and the study design related to it. This includes sample size planning, which can for example be orientated towards the desired precision of the estimate (width of the confidence interval), and requires an estimate of both the prevalence and incidence of the exposition with regard to the outcome variable concerned.
- Clear description of the target and sample population (e.g. population-, register- or general practitioner-based) and justification of their selection. Description of the selection and of the recruitment procedure for study participants.
- Homogeneity of the population investigated. If the population is heterogeneous, it needs to be considered that a prognostic statement can be made as constantly as possible across the

subgroups causing heterogeneity (e.g. existence of different baseline risks for the outcome variable in question).

- Clear definition of the outcome variable(s) on which the prognostic significance is to be based.
- Clear definition of the prognostic characteristics, including their statistical handling (e.g. dichotomisation or assessment of terziles or quartiles for a quantitative characteristic), and justification of the procedure selected.
- Clear determination and definition of potential confounders and effect modifiers, including their statistical handling.
- In cohort studies, completeness of follow-up, or measures to achieve as complete a follow-up as possible. Estimation of possible selection effects if follow-up is incomplete.
- When assessing prognostic scores, it should be noted that a distinction is made between score development and score validation, e.g. score development within a so-called learning sample and validation in a test sample. Ideally, score development and score validation are carried out in different studies.

Typical study designs for the evaluation of prognostic characteristics in terms of risk factors include cohort studies and case-control studies. In exceptional cases (e.g. when investigating constant characteristics), cross-sectional studies may also play a role. The underlying principles for the evaluation of such studies beyond the aspects mentioned above are described in Section 1.5.

The literature search for the evaluation of prognostic characteristics (within the framework of a systematic review) is more difficult than, for example, for therapeutic studies, and no generally accepted optimum search strategy (yet) exists. Furthermore, it can be assumed that this research field is especially susceptible to publication bias [1,2]. The methodological quality of studies (or their publications) on prognostic characteristics is frequently insufficient [4], so that the extraction of the data needed is difficult or impossible. Therefore, meta-analyses (not however systematic reviews per se) of prognostic studies are often inappropriate and their findings should be applied with reservation [2]. Some important problems of meta-analyses of prognosis studies can be avoided if individual patient data are available [2].

References

[1] Simon R., Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994; 69: 979-985.

[2] Altman D. Systematic reviews of evaluations of prognostic variables. In: Egger M, Davey Smith G, Altman DG, editors. Systematic Reviews in Health Care: Meta-analysis in Context, 2nd Ed. London: BMJ Books; 2001.

[3] Laupacis A, Wells G, Richardson WS, Tugwell P for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. V. How to use an article about prognosis. JAMA 1994; 272: 234-237.

[4] Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA, Kasten LE, McCormack VA. Issues in the reporting of epidemiological studies: A survey of recent practice. BMJ 2004; 329: 883-887.

2.9 Individual risk assessment

Besides using the results of studies investigating single or, in the majority of studies, multiple prognostic characteristics, risk charts (also called risk engines) are being increasingly employed to assess the individual risk of patients (or clinically healthy persons) of experiencing an adverse event. Multi-factorial estimates for the concurrence of numerous risk factors are made in these charts (e.g. the Sheffield Table [1] or Joint British Chart [2]). The basis for these risk charts are mainly multi-factorial regression models, whose results, for easier handling, are presented in tables or points systems [3]. It needs to be considered that risks derived for these risk charts are not "personal" estimates for individuals, but statistical estimates of the average risks of a population with a specific risk profile for a defined period (e.g. ten years). The following factors should be considered when evaluating such instruments:

- What type of population the estimated risks apply to;
- What type of study the underlying data originate from;
- Whether the variables included in the multi-factorial analysis were also analysed together in the underlying studies;
- Whether, and if so, how, a multi-factorial statistical analysis was conducted in the underlying studies (q.v. Section 1.22);
- Whether these instruments were ever validated in subsequent studies (test samples).

References

[1] Wallis EJ, Ramsay LE, Ul Haq I, Ghahramani P, Jackson PR, Rowland-Yeo K, Yeo WW. Coronary and cardiovascular risk estimation for primary prevention: Validation of a new Sheffield table in the 1995 Scottish health survey population. BMJ 2000; 320: 671-676.

[2] British Cardiac Society, British Hyperlipidaemia Association, British Hypertension Society, endorsed by the British Diabetic Association. Joint British recommendations on prevention of coronary heart disease in clinical practice. Heart 1998; 80 (Suppl 2): S1-S29.

[3] Sullivan LM, Massaro JM, D'Agostino RB. Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med 2004, 23: 1631-1660.

2.10 Evaluation of population-based prevention and intervention measures

Prevention is directed at avoiding, reducing the probability of, or delaying damage to health [1]. Whereas primary prevention comprises all measures performed before the occurrence of detectable biological damage to prevent the triggering of contributory causes, secondary prevention comprises measures to detect clinically asymptomatic early stages of diseases and their successful early therapy (q.v. Section 2.4). Tertiary prevention in the narrowest sense describes specific interventions to prevent permanent, especially social, functional deficits after the occurrence of a disease [1].

Measures of primary and secondary prevention are characterised by the fact that, in contrast to curative measures, whole population groups are frequently the focus of the intervention. For a benefit at the population level, besides the efficiency of a programme, the level of participation is decisive. In addition, the question is relevant as to which persons are reached; indications exist, for example, that especially population groups with an increased risk of disease make less use of prevention programmes [2].

One challenge in the evaluation of the benefits and harms of prevention measures consists in the fact that the effect chain between an intervention and the possible prevention of the occurrence of a disease is a complex one, and can possibly span many years, so that to detect any effect on disease events, very long observation periods may be necessary [3]. For this reason, when evaluating preventive measures, surrogate criteria – often called “intermediate outcome parameters” in this context – such as changes in lifestyle (for example, smoking or dietary habits) are frequently resorted to. Sometimes only changes with regard to health-related knowledge or changes in attitude or intention are documented [3]; however, a clear and always consistent association between these surrogate parameters and (favourable) effects in respect of morbidity or even mortality has often not been shown [4]. In these cases, studies with surrogate endpoints cannot be drawn upon alone for an evaluation of the benefits and harms of prevention programmes.

When assessing the effectiveness of prevention programmes or of interventions in general (q.v. Section 1.9), RCTs ensure the greatest internal validity [5]. In particular, comprehensive population-based programmes may, however, entail specific challenges with regard to the suitable study design. Among other things, cluster randomised trials are being increasingly performed because of potential treatment "contamination" between intervention and control groups [6]. In this study design, groups of people (not individuals) are randomly assigned to different conditions. A possible source of systematic bias, which must be considered in the evaluation of these trials, is, for example, when cluster trials recruit their participants after the clusters have been randomly allocated, which may lead to selection effects [7].

It must be assessed for the given situation to what extent the consideration of further study designs may be meaningful in specific cases [5]. For example, the effects of mass-media campaigns are often evaluated within the framework of a so-called "interrupted time-series analyses" [e.g. in 8], and the use of this study design is also advocated for community intervention research [9]. Quality criteria for the evaluation of studies have been developed by the Cochrane Effective Practice and Organisation of Care Review Group [10].

It should be noted, even more so than in clinical interventions, that the effectiveness of any comprehensive prevention programme may depend substantially on (known or unknown) contextual factors [11]. For this reason, a systematic review should preferably provide comprehensive information on the particular setting in which the intervention was investigated to help evaluate the applicability to local health care situations [5,12]. In the synthesis of several studies it should be reviewed in each case whether the studies are in fact sufficiently comparable in their conception to justify meta-analysis (q.v. Section 1.21).

References

[1] Walter U, Schwartz FW. Prävention [Prevention]. In: Schwartz FW, Badura B, Busse R, Leidl R, Raspe H, Siegrist J, eds. Das Public Health Buch [The Public Health Book]. München: Urban & Fischer Verlag; 2003, 189-214.

[2] Laaser U, Hurrelmann K. Gesundheitsförderung und Krankheitsprävention [Health promotion and disease prevention]. In: Hurrelmann K, Laaser U, eds. Handbuch Gesundheitswissenschaften [Health Sciences Manual]. Weinheim: Juventa; 1998, 395-424.

[3] Nutbeam D. Health Promotion Effectiveness - The Questions to be Answered. The Evidence of Health Promotion Effectiveness - Shaping Public Health in a New Europe. Brüssel: IUHPE; 2000, 1-11.

[4] Stacy AW, Bentler PM, Flay BR. Attitudes and health behavior in diverse populations: drunk driving, alcohol use, binge eating, marijuana use, and cigarette use. Health Psychology 1994; 13: 73-78.

[5] Jackson N, Waters E. Guidelines for systematic reviews in health promotion and public health taskforce. Criteria for the systematic review of health promotion and public health interventions. Health Promot Int 2005; 20: 367-374.

[6] Torgerson DJ. Contamination in trials: Is cluster randomisation the answer? BMJ 2001; 322: 355-357.

[7] Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. BMC Med Res Methodol 2005; 5: 10.

[8] Vidanapathirana J, Abramson MJ, Forbes A, Fairley C. Mass media interventions for promoting HIV testing. Cochrane Database Syst Rev 2005 Jul 20; CD004775.

[9] Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. Prev Sci 2000; 1: 31-49.

[10] Cochrane Effective Practice and Organisation of Care (EPOC) Review Group. The Data Collection Checklist [online], 2002 http://www.epoc.uottawa.ca/checklist2002.doc [accessed on 07.02.2006].

[11] Hawe P, Shiell A, Riley T, Gold L. Methods for exploring implementation variation and local context within a cluster randomised community intervention trial. J Epidemiol Community Health 2004; 58: 788-793.

[12] Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De Aguero L, Truman BI, et al. Developing an Evidence-Based Guide to Community Preventive Services - Methods. Am J Prev Med 2000; 18: 35-43.

2.11 Description of the type and size of the placebo effect

The overall effect of an intervention is made up of various components: the actual effect to be ascribed to the intervention alone, the natural course of the illness, and the so-called 'placebo effect' [1], which describes the context in which the treatment takes place. This summarises all the influences accompanying an intervention, for example, the expectations of the patient and treating staff, the suggestive behaviour of those treating, or any effect deriving from the simple fact that a patient is being cared for. Placebo-controlled studies serve to balance the influence of these concomitant effects in both groups, by including a group that receives placebo therapy. The decisive factor in this case is the possibility of blinding patients and treating staff with regard to the intervention. The blinding of treating staff is designed to ensure that the care, attention, and suggestive behaviour going beyond the intervention itself are distributed equally between study groups.

Placebo treatment is not limited to drug interventions alone, but can also be applied in non-drug interventions, partly to at least achieve the blinding of patients. In this respect one also speaks of so-called "sham interventions". The extent of a placebo effect can depend on the type of intervention [2].

In the course of a trial it can happen that in certain interventions, despite original blinding, an unblinding of patients and treating staff occurs, for example, if specific adverse effects of the therapy under investigation exist. A possible unblinding or the omission of a blinding procedure in the first place can lead to a bias of results. The exact extent of this bias will not normally be determinable. To achieve this, an unbiased estimate of the degree of the placebo effect would be necessary, for example, by including a third group that does not receive any treatment at all. However, particularly in situations where unblinding occurs, the placebo effect cannot be estimated without bias, as the difference between the placebo group and the group without treatment is also biased. If, for example, it can be assumed that a possible placebo effect is substantially reduced after unblinding of patients in the placebo group, the difference between the placebo group and the group without treatment would present too small an estimate of the placebo effect.

Despite the problems described, the validity of the results of unblinded trials or studies in which unblinding occurred should be discussed, in case indications of a possible major placebo effect exist.

It has been proposed to assess the success of blinding in end-of-trial tests by comparing how many patients and treating staff correctly guess treatment assignment [3]. Such assessments, however, contain methodological problems that have not yet been satisfactorily solved and are controversially discussed (for example: which procedures should be followed for which null hypotheses? [4-6]). In the case of effective therapies, with effects that can be directly experienced by patients, an unblinding is to a certain extent possible or even probable. This means that in such a situation it is difficult or even impossible to judge whether bias with regard to the therapy effect has occurred through unblinding (for whatever reason), or whether conversely the therapy effect has led to unblinding. Despite these problems, the assessment of the blinding of a trial is to be welcomed, as it provides an indication that the issue of blinding has been taken into appropriate consideration in the planning and conduct of a study.

The interpretation of the results of unblinded trials, or of trials where unblinding (possibly) occurred, must be more cautious than the interpretation of blinded studies, and presupposes that the documentation of endpoints was conducted in a blinded manner (apart from the endpoint “total mortality”).

References

[1] Thompson WG. The Placebo Effect and Health. New York: Prometheus Books, 2005.

[2] Kaptchuk TJ, Stason WB, Davis RB, Legedza AR, Schnyer RN, Kerr CE, Stone DA, Nam BH, Kirsch I, Goldman RH. Sham device *v* inert pill: Randomised controlled trial of two placebo treatments. BMJ 2006; 332: 391-397.

[3] Fergusson D, Cranley Glass K, Waring D, Shapiro S. Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo controlled trials. BMJ 2004; 328: 432-434.

[4] Altman DG, Schulz KF, Moher D. Turning a blind eye: Testing the success of blinding and the CONSORT statement (Letter). BMJ 2004; 328: 1135.

[5] Senn SJ. Turning a blind eye: Authors have blinkered view of blinding (Letter). BMJ 2004; 328: 1135-1136.

[6] Sackett DL. Turning a blind eye: Why we don't test for blindness at the end of our trials. BMJ 2004; 328: 1136.

3. Evidence-based health information for consumers and patients

3.1 Goal

The Institute aims to become an effective, reliable, trusted, and popular provider of evidence-based health information education for consumers and patients. Relevant information from the Institute's reports are communicated to the public in a comprehensive and integrated approach primarily via the IQWiG website http://www.gesundheitsinformation.de. The English version is available under http://www.informedhealthonline.org.

The aim is to provide health care information that is outcome-orientated, objective and commonly used. Furthermore, this information should be adapted to patients' psychological needs and be easily understandable, without the need for specific medical knowledge.

The goal of advancing health and scientific literacy is to:

- Improve understanding of physical, mental and emotional health;
- Improve understanding of medical and scientific information, including the concept of evidence-based medicine;
- Promote health-related behaviour;
- Enable support by relatives and friends;
- Promote the critical use of health care services;
- Support active decision-making about health issues (e.g. participatory doctor-patient relationships) that is responsive to individual needs and values.

The Institute does not provide advice directly to consumers and patients. It is the Institute's intention to enhance independent and responsible informed choices on health issues, giving priority to consumer and patient autonomy [1].

3.2 Information system

Internet-based [2,3] and offline computer-based [4] health information can positively affect consumers' and patients' state of knowledge, choices, and physical, mental, and emotional well-

being. However, information and education interventions can also be ineffective or harmful, and some techniques are more effective than others [3,5-8]. The website www.gesundheitsinformation.de is the first instrument for the dissemination of evidence-based health information produced by the Institute.

The website is being developed into a comprehensive and versatile reference work, aimed at meeting a variety of individual health information needs of consumers and patients at different levels [5,6]. Various information products, which are outlined in the following Section 3.2.1, are being employed for this purpose. These products are interlinked, and supported by definitions, explanations, and supplementary information.

The website includes an electronic newsletter, downloadable texts, and other files on health-related topics as media for health information. Downloadable print versions are also available. In addition, it will be possible to reproduce electronic information on other websites.

3.2.1 Information products

Information products include comprehensive health information topics (named “articles” on the website), information leaflets, short summaries (named “short responses” on the website), and other products.

Comprehensive health information topics

The comprehensive information topics are published as “articles” on the website together with a summarising information leaflet, and form the main focus of the reference work which is being developed, comprehensively covering a wide variety of health-related issues. Comprehensive information topics are also developed from the Institute’s scientific reports. The comprehensive information topics should consider:

- Education about the disease or condition, including:
  - Anatomy,
  - Physiology,
  - Different types of diseases,
  - Disease aetiology,
  - Recognition of symptoms,

  - Normal course of diseases,
  - Prognosis,
  - Potential complications,
  - Recognition of complications,
  - Recovery,
  - Possible recurrence of disease,
  - Recognition of recurrent disease,
  - Risk groups (including relatives).
- Preventive measures and health promotion, including
  - Nutrition,
  - Physical activity,
  - Screening techniques,
  - Information.
- Diagnostic measures, including complementary diagnostic tests.
- Therapeutic measures, including
  - Pharmaceuticals,
  - Surgery,
  - Other non-pharmaceutical procedures,
- Rehabilitation measures.
- Other health care services.
- Psychosocial aspects, including testimonies of patients suffering from different diseases as well as those of their relatives.

Information leaflets

In addition to comprehensive information topics, information leaflets as an additional basis of the reference work are developed. On the one hand, information leaflets summarise the key statements of the comprehensive information topics.

On the other, they are based on single or multiple short summaries, supplemented with further information. Information leaflets are also developed as a summary of a single or multiple systematic reviews.

Short summaries

In addition to the comprehensive information topics and information leaflets, a much larger number of short summaries are produced ("short responses to scientific questions"). These short summaries complement the reference work and make evidence-based information, currently only available in English, accessible to the general public.

The short summaries consist of easily understandable texts detailing important, interesting and/or current health topics. Short summaries of the Institute's scientific reports are also produced.

Other products

Other products include visual and interactive tools such as diagrams, online calculators (e.g. for cigarette costs), and glossaries (e.g. online dictionaries).

These tools are designed to:

- Promote general understanding of health and medical issues;
- Improve understanding of diseases, e.g. develop knowledge of the normal course of diseases, symptom recognition, possible complications, recovery, and possible recurrence of the disease;
- Increase ability to understand and weigh potential risks;
- Support self-management strategies, e.g. for chronic diseases.

3.2.2 Editorial system

The Department of Health Information, in consultation with the Institute's Management and the affected departments, ensures that the contents of the health information website and other health information products are:

- Evidence-based and consistent with the current state of scientific knowledge;

- Orientated towards consumers' and patients' information needs and adapted to their psychological and emotional needs as far as possible;
- Consistent with all other information products published by the Institute.

Internal and external criticism on contents and quality of the health information provided are referred to the responsible Health Information Editor, who initiates appropriate action and reports regularly to the Steering Committee. Major or urgent problems are reported immediately to the Institute's Management.

Close liaison between the responsible Health Information Editor, the Institute's Management, the Department of Communications and all other departments concerned ensures that all publications are consistent in content. The final decision to publish is the responsibility of the Steering Committee.

3.2.3 Multilingualism

The Institute aims to publish health information in both German and English and keep both versions up to date. A broad international exchange and the best possible quality assurance are only to be achieved through publications in English. This enables the quality of the health information products to profit from feedback from international scientists and reviewers, including authors of systematic reviews.

The Institute also cooperates with external partners to enable at least some of its health information to be translated into the most widely spoken languages in Germany.

It is very difficult to assess the quality of translations according to objective criteria. No specific standard can be defined. Translations can be literal or can aim to capture the intent of the original in the target language. The Institute frequently applies the latter method.

3.3 Development of information products

3.3.1 Selection of topics

The selection of health information topics is designed to reflect public interest as far as possible. It should be balanced, impartial, transparent, clinically relevant, and well-founded.

As the generation and maintenance of comprehensive health information topics involves significant investment of resources, rigorous methods of priority-setting are required [9-11]. Priority-setting for comprehensive health information topics is undertaken according to the following process:

a. Development of filter criteria (main and supplementary criteria):

   The main filter criteria include the quantity and quality of scientific research on a topic, as well as public interest in it. Additional criteria include the size of the target population affected, general information and educational aspects, the current state of knowledge of the population, as well as assumptions about effects on health status and possible risks (at both the individual and population level).

b. Internal and external consultations on filter criteria in the course of 2006.

c. Determination of filter criteria.

d. Evaluation.

On the basis of these filter criteria, the systematic selection of topics is performed by application of the filter criteria to a preferably comprehensive number of topics using a three-stage model:

Stage 1: Application of the main criteria (coarse filter),

Stage 2: Application of supplementary criteria to the comprehensive information topics selected in Stage 1,

Stage 3: Internal and external consultations on the comprehensive information topics selected in Stage 2.

3.3.2 Scope and contents

A range of methods can be used to help identify questions that the health information needs to address. These methods vary in their cost and practicality, as well as in the transferability of their results to other health information issues [12,13]. As far as possible, the Institute uses high-quality data, surveys and studies. These materials may be supplemented by telephone interviews with key informants and/or focus groups (q.v. Section 1.24). The Institute's decisions are always made under the premise that the research questions posed are in the public interest. Special attention is paid to the needs of disadvantaged population groups.

Internal project groups, consisting of members from the Department of Health Information and from other departments, may be formed to work on individual comprehensive health information

topics. Formation of the project group and internal project coordination is the responsibility of the Department of Health Information.

a. In the first project group meeting, a scoping exercise determines the range of subjects to be incorporated in the health information topic.

b. A review of health care information available to consumers and patients on the subject (e.g. diagnostic tests and therapies) is then carried out. A literature search and a search of key websites (the most important websites with information on health issues) are undertaken. Telephone interviews of key informants (e.g. patient representatives and clinical experts) may be undertaken.

c. Consumers' and patients' needs, their current level of knowledge and their potential interest in the subject are analysed as far as possible. Here again, a literature search including a search for websites containing information on patients' experiences [14] and telephone interviews with key informants are carried out. Focus groups may also be formed on some occasions.

d. On this basis, a draft of the health information is developed and discussed in the second project meeting. This draft includes the key questions as well as the outline. In addition to the discussion about the literature search, the further procedure is specified.

3.3.3 Production

In principle, the same evidence-based methods are applied to the production of health information as to those applied in the Institute as a whole (q.v. Sections 1.23 and 2.7). The individual information products are developed according to the draft as follows:

Comprehensive patient information topics

The production of comprehensive patient information topics is carried out as follows:

a. Literature searches for published systematic reviews and published qualitative research. The validity and topicality is subsequently discussed and evaluated. Production of a preliminary version of the comprehensive information topic from the results of the literature search.

b. Internal peer review:
   Multiple checking of the preliminary version. The resulting feedback or criticism is discussed and, if necessary, a further analysis and review is made.

c. External peer review:
   Expert opinions are sought from patient representatives and clinical experts. Lead authors of major systematic reviews are given the opportunity to comment on the Institute's draft (usually the English version). If required, relevant public agencies are consulted.

d. If necessary, the readability and comprehensibility of the comprehensive information topics (German version) are tested by representatives of the target group.

e. The revised version of the comprehensive information topic is prepared in German and English and forwarded to the Institute's Steering Committee. The information is then distributed for external comments (to be provided within a limited period of time), and is then marked up for the Internet as a test version. The usability of this (German) online version can be tested with volunteers, including at least one patient or patient representative [15-17]. The English version is subsequently marked up for the Internet as an (offline) test version.

f. When multiple information products are to be produced from the source material and major parts of their content have been amended, it is also necessary for these products to undergo a quality assurance procedure. This only applies if substantial changes of content are planned.

g. The Steering Committee can either release the final version or alternatively suggest further discussions or revisions.

Short summaries

The production of short summaries is conducted by abstracting short single studies, a single systematic review or an important study, or a few reviews or studies. Short summaries pass through a developmental process, as described in the steps above. The Department of Health Information is responsible for the maintenance and updating of the short summaries.

Patients' narratives

Many patients would like to hear or read about the experiences of other patients with the same disease [18]. Patients' narratives are a common form of communicating information both in

journalism and in the area of patient information. The relevance of patients' narratives in clinical practice and in the health care system is being increasingly acknowledged [19,20].

By documenting narratives of patients and their relatives, individual experiences with diseases and dependency on care in their different dimensions can be made available to other affected or interested persons as a supplement to the evidence-based health information. The patients' narratives should not, however, contradict the evidence-based health information.

The documentation of patients' narratives is conducted according to the following steps:

1. Search for interview partners.
2. Acquisition of the informed consent of the interview partner with regard to the performance and the further use of the interview.
3. Performance of the interview.
4. Acquisition and documentation of the informed consent to publication of the final version.
5. Publication on the website with the consent of the interview partner.

Particular emphasis is placed on comprehensively informing the patient and/or his or her relative before the interview, on informed consent for publication (which is reversible at any time), on a detailed preparation of the interview, on an approach by the interviewer which follows prespecified criteria for interview techniques, as well as on an anonymised interview transcript.

3.3.4 Evidence base and communication standards

In its communication of health information, the Institute's goals are to:

- Communicate respectfully and effectively with the population in Germany so that it trusts the Institute as a reliable and easily understandable source of information;
- Produce health information that is easily understandable, without compromising scientific accuracy;
- Maintain a style of communication that is as neutral and unambiguous as possible;
- Demonstrate sensitivity and respect for consumer and patient knowledge, values and concerns, autonomy, and cultural differences;
- Support patient empowerment;
- Promote health and scientific literacy;

- Help the individual to relate evidence to his or her own personal situation;
- Respect readers' time**.**

The Institute ensures that the website meets disability accessibility standards [18].

The health information should on the one hand not exaggerate what is scientifically known, and on the other, not tell people what they 'should' do. The effects of making people aware of the scientific uncertainty of much health care is largely unknown. The general public is also accustomed to a more directive style of health information, often aimed at directly altering opinions and attitudes. The Institute aims to present information in a variety of ways, in order to enable as many people as possible to gain access to information [7].

As a general rule, health information detailing relative risks should be avoided. However, the inclusion of relative risk data may occasionally be necessary to enable the individual to compare treatments. These data are then not presented on their own, but if possible together with the absolute risk or the number needed to treat (the number of patients needed to be treated before a patient benefits from treatment or suffers harm through it).

There is evidence that the presentation of personalised or individualised risk estimates is an effective form of communicating health information [7,22]. The Department of Health Information can develop or adapt tools with which consumers and patients can estimate their personal risk, providing that reliable data on the development of tools are available.

Patient decision aids (e.g. visualised risk diagrams) have been shown to be an effective means of communicating health information [23]. The Institute may develop or adapt decision aids and incorporate effective elements of these aids into its health information products.

In addition, the Department of Health Information:

- Presents health information in consistent formats;
- Explains the degree of uncertainty associated with this information;
- Indicates to which sub-population(s) the scientific evidence applies;
- Aims to achieve the highest possible standards of web usability (including web navigation);
- Is very careful to clearly distinguish between "absence of evidence" and "evidence of no effect";

- Avoids biasing information in respect of the products of any particular company; the generic names of products are generally used, supported by brand names of products only where relevant;
- Increases the understandability of the health information provided on the website by producing an online dictionary, integrated by way of hyperlinks.

3.4 Publications

The dissemination of health information is discussed with the Institute's Management, the Department Heads affected and the Department of Communications. Proposals are passed on to the Steering Committee. The Institute's Management, the Department Head (Health Information), and the Institute's spokesperson ensure consistency of content for external communication.

Health information is mainly published on the health information website as comprehensive information topics, short summaries or as other products (q.v. Section 3.2).

3.5 Evaluation and updating

There are many instruments and guidelines for the quality evaluation and assurance of health information in the Internet. However, there are no reliable data on the validity of these instruments and guidelines [21-24]. No reliable instrument for the production of health information is therefore available that can be regarded as a reliable indicator of quality [22,24]. There is also no evidence on the cost-effectiveness of production options.

Studies on research with patients, including patients in Germany [25,26], indicate that some issues suggested as being important in the evaluation of health information may not be important to patients. Some recommendations common in instruments and guidelines on the evaluation of information may actually reduce its scientific quality.

Most evaluation instruments focus on information about treatment and do not address the full range of a health issue (including aetiology, prognosis, screening and diagnostic tests). Yet health information on diagnosis and screening involves more complex decision-making and communication issues than information on treatment [6,7,19].

The health information produced by the Institute therefore does not rely on any currently available evaluation instruments, nor will it develop one. The Institute, however, relies on evidence on

specific aspects that could influence patients' decision-making and demonstrably affect the quality of information produced.

The Department of Health Information, in communication of health information, stays up to date with the new scientific evidence on the options of quality and, if necessary, adapts its methods accordingly. Health information products are tested as far as possible with representative population groups**.**

To ensure that information is kept up to date, the literature for individual health information products is coded, enabling monitoring for publications of important new evidence and for changes in Cochrane reviews. The topicality of the health information products is also ensured by the ongoing exchange of information between key personnel at the Cochrane Collaboration, the Centre for Reviews and Dissemination, and *Evidence-Based Medicine*. The currency is also assured by a revision of the health information products, which is carried out at least every two years.

The health information website includes a feedback mechanism for readers. Any form of internal and external feedback may result in immediate revision of health information already published.

Close cooperation between the Department of Health Information and other relevant departments ensures that the health information provided accords with current evidence. The Department of Health Information relies on internal expertise and consultation with key informants, and offers the authors of systematic reviews the opportunity to comment on the Institute's interpretation of their work.

References

[1] Hope T. Evidence-Based Patient Choice. London: King's Fund; 1996.

[2] Bessell TL, McDonald S, Silagy CA, Anderson JN, Hiller JE, Sansom LN. Do internet interventions for consumers cause more harm than good? A systematic review. Health Expect 2002; 5: 28-37.

[3] National Institute of Clinical Studies. The Impact of the Internet on Consumers Health Behaviour. Melbourne: NICS; 2003. http://www.nicsl.com.au/asp/download.asp?media=/data/portal/00000005/content/1574600115387458931 5.pdf [accessed on 17.11.2006].

[4] Lewis D. Computer-based approaches to patient education: A review of the literature. J Am Med Inform Assoc 1999; 6: 272-282.

[5] Coulter A, Entwistle V and Gilbert D. Informing Patients: An Assessment of the Quality of Patient Information Materials. London: King's Fund Publishing; 1998.

[6] Entwistle VA, Watt IS, Davis H, Dickson R, Pickard D, Rosser J. Developing information materials to present the findings of technology assessments to consumers: The experience

of the NHS Centre for Reviews and Dissemination. Int J Tech Assess Health Care 1998; 14: 47-70.

[7] Edwards A, Bastian H. Risk communication – making evidence part of patient choices? In: Edwards A, Elwyn G, editors. Evidence-Based Patient Choice: Inevitable or Impossible? Oxford: Oxford University Press; 2001: 144-160.

[8] Eysenbach G, Jadad AR. Consumer health informatics in the Internet age. In: Edwards A, Elwyn G, editors. Evidence-Based Patient Choice: Inevitable or Impossible? Oxford: Oxford University Press; 2001: 289-307.

[9] Henshall C, Oortwijn W, Stevens A, Granados A, Banta D. Priority setting for health technology assessment. Theoretical considerations and practical approaches. A paper produced for the EUR-ASSESS project. Int J Technol Assess Health Care 1997; 13: 144-185.

[10] Townsend J, Buxton M, Harper G. Prioritisation of health technology assessment. The PATHS model: Methods and case studies. Health Technol Assess 2003; 7: iii, 1-82.

[11] Ghaffar A, de Francisco A, Matlin S. The Combined Approach Matrix: A Priority-setting Tool for Health Research. Geneva: Global Forum for Health Research; 2004. http://www.globalforumhealth.org/filesupld/90.pdf [accessed on 17.11.2006].

[12] Liberati A, Sheldon TA, Banta HD. EUR-ASSESS Project Subgroup report on Methodology. Methodological guidance for the conduct of health technology assessment. Int J Tech Assess Health Care, 1997; 13: 186-219.

[13] Sixma HJ, Kerssens JJ, Campen CV, Peters L. Quality of care from the patients' perspective: from theoretical concept to a new measuring instrument. Health Expec 1998; 1: 82-95.

[14] DiPEX. The Database of Patients' Experiences. http://www.dipex.org [accessed on 18.10.2004].

[15] Krug S. Don't make me think: A common sense approach to web usability. Indiana: New Riders; 2000.

[16] Nielsen J. Designing web usability. Indiana: New Riders; 2000.

[17] Inan H. Measuring the success of your website: A customer-centric approach to website management. Sydney: Pearson Educational Australia; 2002.

[18] Swift TL, Dieppe PA. Using expert patients' narratives as an educational resource. Patient Educ Couns 2005; 57: 115-121.

[19] Greenhalgh T, Hurwitz B. Narrative based medicine: Why study narrative? BMJ 1999, 318: 48-50.

[20] Steiner JF. The use of stories in clinical research and health policy. JAMA 2005; 294: 2901-2904.

[21] W3C. Web Accessibility Initiative. http://www.w3.org [accessed on 26.01.2005]

[22] Edwards A, Unigwe S, Elwyn G, Hood K. Personalised risk communication for informed decision making about entering screening programs (Cochrane Review). In: The Cochrane Library, Issue 3, 2004. Chichester: Wiley; 2004.

[23] O'Connor AM, Stacey D, Entwistle V, Llewellyn-Thomas H, Rovner D, Holmes-Rovner M, Tait V, Tetroe J, Fiset V, Barry M, Jones J. Decision aids for people facing health treatment or screening decisions (Cochrane Review) (last updated October 2003). In: The Cochrane Library, Issue 3, 2004. Chichester: Wiley; 2003.

[24] Jadad AR, Gagliardi A. Rating health information on the internet: navigating to knowledge or to Babel? JAMA 1998; 279: 611-614.

[25] Eysenbach G. Consumer health informatics. BMJ 2000; 320: 1713-1716.

[26] Gagliardi A, Jadad AR. Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination. BMJ 2002; 324: 569-573.

[27] Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health information. Health Expec 2004; 7: 165-175.

[28] van den Brink-Muinen A, Verhaak PF, Bensing JM, Bahrs O, Deveugele M, Gask L, Leiva F, Mead N, Messerli V, Oppizzi L, Peltenburg M, Perez A. Doctor-patient communication in different European health care systems: Relevance and performance from the patients' perspective. Patient Educ Counsel 2000; 39: 115-127.

[29] Eysenbach G, Kohler C. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. BMJ 2002; 324: 573-577.

4. Production of reports

4.1 Products

According to its legal remit, the Institute generates a variety of products in the form of scientific reports and easily understandable health information for consumers and patients (q.v. Preamble). The generation of these products follows the methodology outlined in this document. The scientific reports on the evaluation of benefits and harms of interventions are available in two formats:

- Detailed **reports**.
- **Rapid reports**, which have two aims:
    - To provide information at short notice on relevant (e.g. new) technologies, or
    - To provide information at short notice on current topics; if required, consultation takes place with other responsible public institutions (e.g. the Federal Institute for Drugs and Medical Devices *[Bundesinstitut für Arzneimittel und Medizinprodukte, BfArM]).*

In addition, the Institute received a general commission from the Federal Joint Committee:

"It is the Institute's responsibility, by means of documenting and evaluating the relevant literature, to conduct an ongoing investigation and evaluation of 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 Federal Joint Committee on a regular basis. In this context, the Federal Joint Committee assumes that the Institute, according to the tasks assigned following § 139a (3) Social Code Book V, will not only work on single commissions awarded by the Federal Joint Committee, but will also on its own responsibility take on scientific work and relay necessary information on developments relevant to health care to the Committee, so that the Committee can fulfil its legal tasks. The Institute will also elaborate concrete suggestions for single commissions, which from the point of view of the Institute and against the background of this information are relevant to the German health care system."

Scientific work within the framework of the general commission is published as **working papers**.

The procedures for the production of the different types of reports are adapted to the different objectives of the documents and are described in Sections 4.4 and 4.5.

Working papers are produced in accordance with the procedure defined for the production of rapid reports. For the production of easily understandable information (**health information**) for consumers and patients, the methodological requirements presented in Section 3 apply.

An overview of the different products of the Institute is shown in the table below.

Overview of the Institute's products

Product	Objective	Procedure	Commissioned by
Report	Detailed evaluation of benefits and harms of interventions, including a public written hearing. Among other things, these reports are used by the Federal Joint Committee as decision-making aids for the issuing of directives.	Described in Section 4.4.	Federal Joint Committee, Federal Ministry of Health
Rapid report	Information at short notice in urgent cases; prompt information on current topics.	Described in Section 4.5	Federal Joint Committee, Federal Ministry of Health
Working paper	Information on developments in medicine that are relevant to health care.	Described in Section 4.5	Self-initiative of the Institute
Health information	Easily understandable information for consumers and patients; wide scope of topics.	Described in Section 3	Federal Joint Committee or Federal Ministry of Health / self-initiative of the Institute

4.2 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 scientific research commissions within the framework of the generation of the Institute's products or the review thereof.

The Institute bases the award of scientific research commissions on its own award and procurement guidelines, which follow the provisions of § 22 of the "Regulation on budgeting in social insurance" *(Verordnung zum Haushaltswesen in der Sozialversicherung [SVHV]),* as well as the legal regulations of procurement law. The Institute announces in scientific journals that it regularly

advertises projects to be awarded to external experts on its website (http://www.iqwig.de). The commissions currently to be awarded by IQWiG are listed on the website (http://www.iqwig.de/index.174.en.html). Commissions exceeding the sum of €200 000 are advertised throughout Europe.

The Institute has specified the following suitability criteria, which interested persons or working groups are obliged to fulfil and present:

- Their professional independence: This means that the bidders must disclose any potential conflicts of interests pertaining to the commission (q.v. § 139b [3], Page 2, Social Code Book V). The forms "Disclosure of potential conflicts of interest" and "Formal obligation" are to be completed for this purpose.
- Command of the German language.
- Medical/professional references or experience, in each case, pertaining to the commission.
- Methodological references or experience, in each case, pertaining to the commission, i.e. demonstration in particular of the ability to work following the principles of the prevailing methods of the Institute.

The forms required ("Cover letter", "Disclosure of potential conflicts of interest", and "Formal obligation") can be downloaded from the IQWiG website (http://www.iqwig.de/index.523.en.html).

The documents received are assessed by a committee of the Institute consisting of:

- A member of the Institute's Management,
- A member of the Department of Administration (the Department Head or a lawyer),
- The Department Head from a scientific department not involved in the commission to be awarded,
- The Department Head whose department is concerned with the commission to be awarded.

The professional independence of the potential contractor in respect of the commission to be awarded is assessed by a committee of three, consisting of:

- A member of the Institute's Management,
- A Department Head of a scientific department,
- A Research Associate.

The bidders to be invited for further participation in negotiations are thereby selected. At the time of the invitation, the criteria are stated according to which the bidder expected to perform best is chosen. The evaluation of these criteria can, among other things, be undertaken through the presentation of a commission carried out in a reference project or through the concrete presentation of important sections of the commission to be awarded. In addition, the selection committee forms a personal impression of the bidders selected in an interview introducing the project leader or team concerned. External experts intending to contract out the work agreed on (or part of it) to third parties, are obliged to name these parties. In particular, these persons are subject to the same obligations regarding potential conflicts of interest as the external experts themselves.

Finally, the bidder expected to perform best is selected. The individual steps in this evaluation process are documented.

Strict confidentiality is to be maintained by both sides in respect of the terms of the contract until the commission has been completed and accepted, especially regarding scientific content or (partial) findings. The external experts are additionally obliged to observe any relevant data protection requirements.

4.3 Guarantee of scientific independence

4.3.1 Objectives

The scientific and professional independence of the Institute and the products it is responsible for and publishes are legally founded in § 139a Social Code Book V, as well as in the statutes of the Institute's foundation. The term "independence" can thereby only be approximately applied in reality, as the assessment of scientific independence can vary from one individual to the next. Insofar as the term "independence" can only be a relative objective, the term "transparency" is introduced at the same time to give form, in a fashion comprehensible both internally and externally, to any decision-making processes and their findings, against the background of "relative independence".

4.3.2 Guarantee of external scientific independence

Before any contract is signed between the Institute and an external expert or institution to provide professional advice, to conduct studies, or to produce a scientific report, it must be decided whether

any reservations exist as to potential conflicts of interest. For this purpose, all external experts and institutions must disclose all activities that may potentially influence their scientific independence (conflicts of interest; q.v. Section 4.2). In particular, the following criteria, based on the relevant guidelines of scientific journals, are viewed as conflicts of interest: All financial agreements, employment, advice, fees, reimbursed expert opinions, reimbursed travel expenses, patent applications, and share ownership within the previous three years that could have influenced the work commissioned, as well as all existing personal contacts with other persons or organisations that could influence the commission in question [1]. This list of criteria is also included in the form supplied on the Institute's website (q.v. Section 4.2), which is updated whenever necessary. The downloadable version of the form on the website always applies. The names of all external experts involved in the production of final reports or rapid reports will be published together with these reports, including the disclosure of any reported potential conflicts of interest. This will be presented in such a way that it will be stated whether any potential conflicts of interest concerning criteria listed in the "Form for disclosure of potential conflicts of interest" were reported or not. Other details, e.g. on the amount of any financial remuneration received, will not be published. The procedure for selecting external experts is described in Section 4.2.

4.3.3 Guarantee of internal scientific independence

Internal scientific independence is guaranteed as far as possible by the selection of staff. On being appointed, staff must credibly outline their previous activities and are obliged to cease all (external) assignments likely to call their scientific independence into question where their work for the Institute is concerned. The Institute's scientific staff is prohibited from performing paid external assignments that could in the broadest sense be associated with their professional duties. As a matter of principle, all external assignments must be declared by all members of staff to the Institute's Management or the Department of Administration. External assignments in the broadest sense also include unpaid honorary positions such as positions on boards or in organisations and societies. In individual cases, violations may lead to a reprimand or, in recurrent or serious cases, to dismissal. The Institute's Management, after consultation with the Steering Committee, will decide on a case-by-case basis whether a member of staff must be excluded from a certain activity or project on grounds of suspected bias.

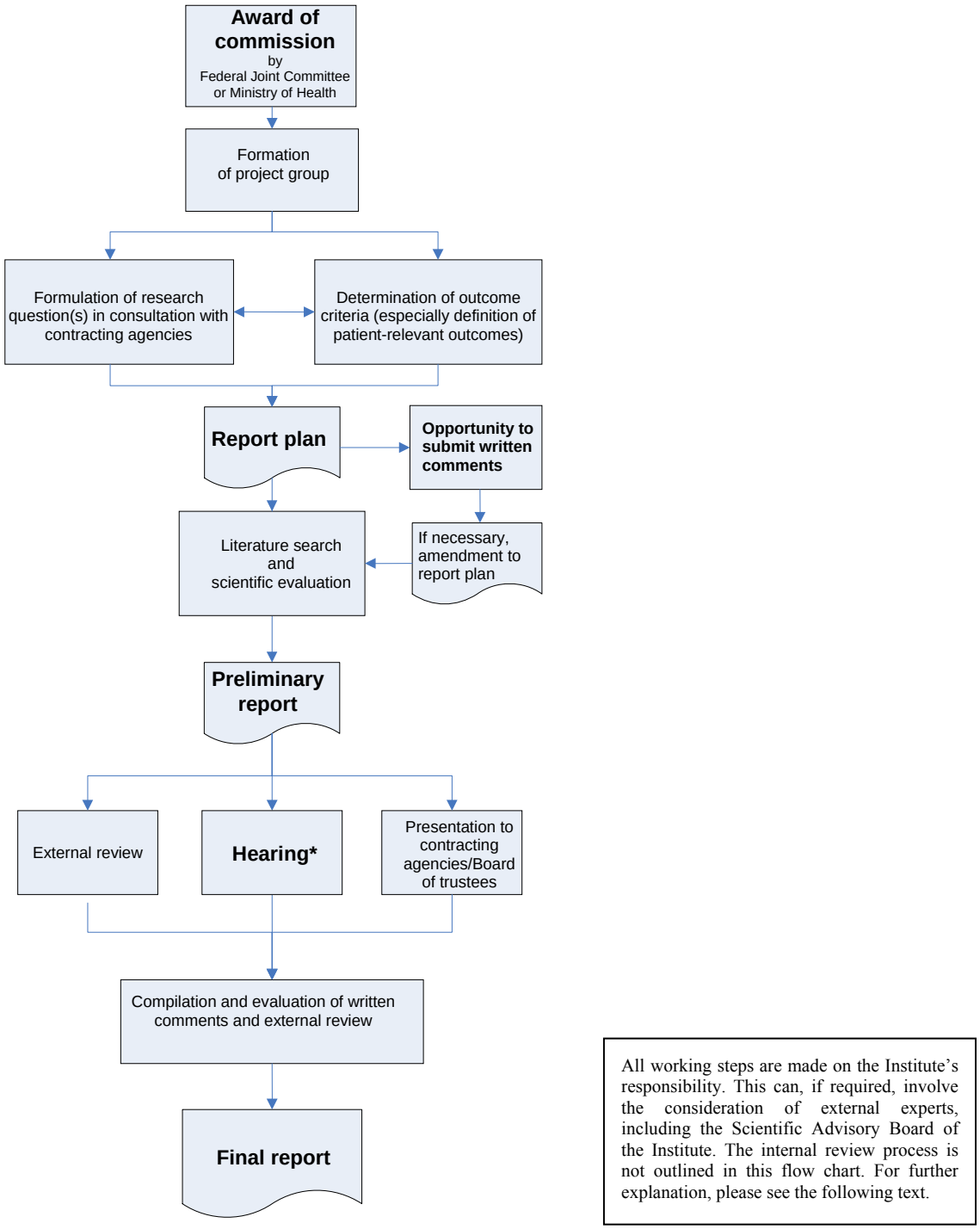
Reference

[1] A. James, R. Horton. The Lancet's policy on conflicts of interest. Lancet 2003; 361: 8-9.

4.4 *Production of reports*

The procedure for report production is outlined in Figure 2.

Figure 2: Flow diagram showing the production of a report



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

After the **award of commission** by the contracting agencies (Federal Joint Committee or Federal Ministry of Health), a project group is formed under management of the department concerned. A project manager is then appointed. The composition of the project group is not fixed at this point, as changes may be necessary due to the subsequent steps taken. As far as necessary, the commission is given concrete shape in consultation with the responsible panels of the contracting agency. This also includes the (rough) definition of the outcome criteria, especially the patient-relevant outcomes. If necessary, this definition will then be refined by the project group, with the inclusion of external expertise (if required). In this context, individual affected persons, patient representatives and/or consumer organisations will regularly be involved with regard to the topic-related definition of patient-relevant outcomes. Subsequently, the report plan will be prepared, which is the basis of both the literature search that follows and the scientific evaluation.

The **report plan**, comparable to the study protocol of a clinical trial, contains the precise scientific research question posed, including the outcome criteria (for example, patient-relevant outcomes), the inclusion and exclusion criteria of the information to be used in the evaluation, as well as a project-specific description of the methodology used in the literature search and in the assessment of this information. In addition, the schedule up to the publication of the final report is outlined. External experts can be involved in the production of the report plan. This plan is prepared under the responsibility of the IQWiG project group and presented to the Steering Committee for internal review. As a rule, on conclusion of the internal reviewing procedure, it is then at the same time:

- Forwarded to the contracting agency and Board of Trustees, also to examine its completeness in respect of the commission originally awarded;
- Published on the Institute's website with the aim of obtaining written comments from interested parties.

For the **hearing** conducted by means of obtaining written comments, the details outlined in the report plan, especially on the inclusion and exclusion criteria for relevant information (e.g. the scientific literature to be included) are of fundamental importance. The deadline for written comments is normally four weeks after publication of the report plan. Further information on the submission of comments can be found on the Institute's website in the corresponding guideline. The conditions stated in the current version of the guideline apply. After evaluation of the written comments received, an amendment to the report plan may be prepared and published. All amendments to the report plan will be described in the preliminary and final reports.

In the **preliminary report**, the results of the literature search and their scientific evaluation are presented. It can, at least in part, be produced by external experts who have shown their suitability according to the selection criteria stated in Section 4.2. The preliminary report is produced on the responsibility of the IQWiG project group and presented to the Steering Committee for internal review. On conclusion of the internal reviewing procedure, it is then as a rule at the same time:

- Forwarded to one or more external experts with proven methodological and/or professional competence;
- Forwarded to the contracting agency and the Board of Trustees, also to examine its completeness in respect of the commission originally awarded;
- Published on the Institute's website with the aim of obtaining written comments from interested parties (written hearing).

The **hearing** is conducted by obtaining written comments. In this context, the details provided in the report plan (including any amendments), especially on the inclusion and exclusion criteria for relevant information (e.g. the scientific literature to be included) are of fundamental importance. The deadline for written comments is at least four weeks after publication of the preliminary report.

In addition, an oral scientific debate including persons submitting comments may optionally be held. The aim of this debate is, if necessary, to clarify aspects presented in the written comments in order to improve the scientific quality of the final report.

Further information on the submission of comments can be found on the Institute's website in the corresponding guideline. The conditions stated in the current version of the guideline apply.

The **final report**, which, building upon the preliminary report contains the evaluation of the scientific findings (considering the results of the written hearing), represents the concluding product of the work on the commission. It is produced under the responsibility of the Institute's project group and presented to the Steering Committee for internal review. Before publication, the final report is forwarded to the contracting agency, the Board of Directors, and the Board of Trustees of the Foundation.

Comments on final reports can be assessed by the Institute with regard to the existence of substantial evidence not considered in a final report as well as the main interpretation of the evidence considered. If appropriate, well-founded information will be provided to the contracting agency as to whether, as a result of such comments, a new commission on the topic is necessary or not. If, due to new information, an **update** of the final report is regarded as necessary, this will be communicated to the contracting agency, which decides on the award of a commission to the Institute with regard to the update of a final report.

The updating process is subject to the same methodological and procedural requirements as the conventional preparation of reports.

4.5 Production of rapid reports and working papers

The procedure for the production of a rapid report is presented in Figure 3. Rapid reports are primarily produced with the aim of providing information at short notice on relevant developments in the health care system, including new technologies, and are not implicitly suitable as a basis for decisions on directives. A shorter preparation period is usually required. The procedure for the production of a rapid report differs from that of a full report in two main points:

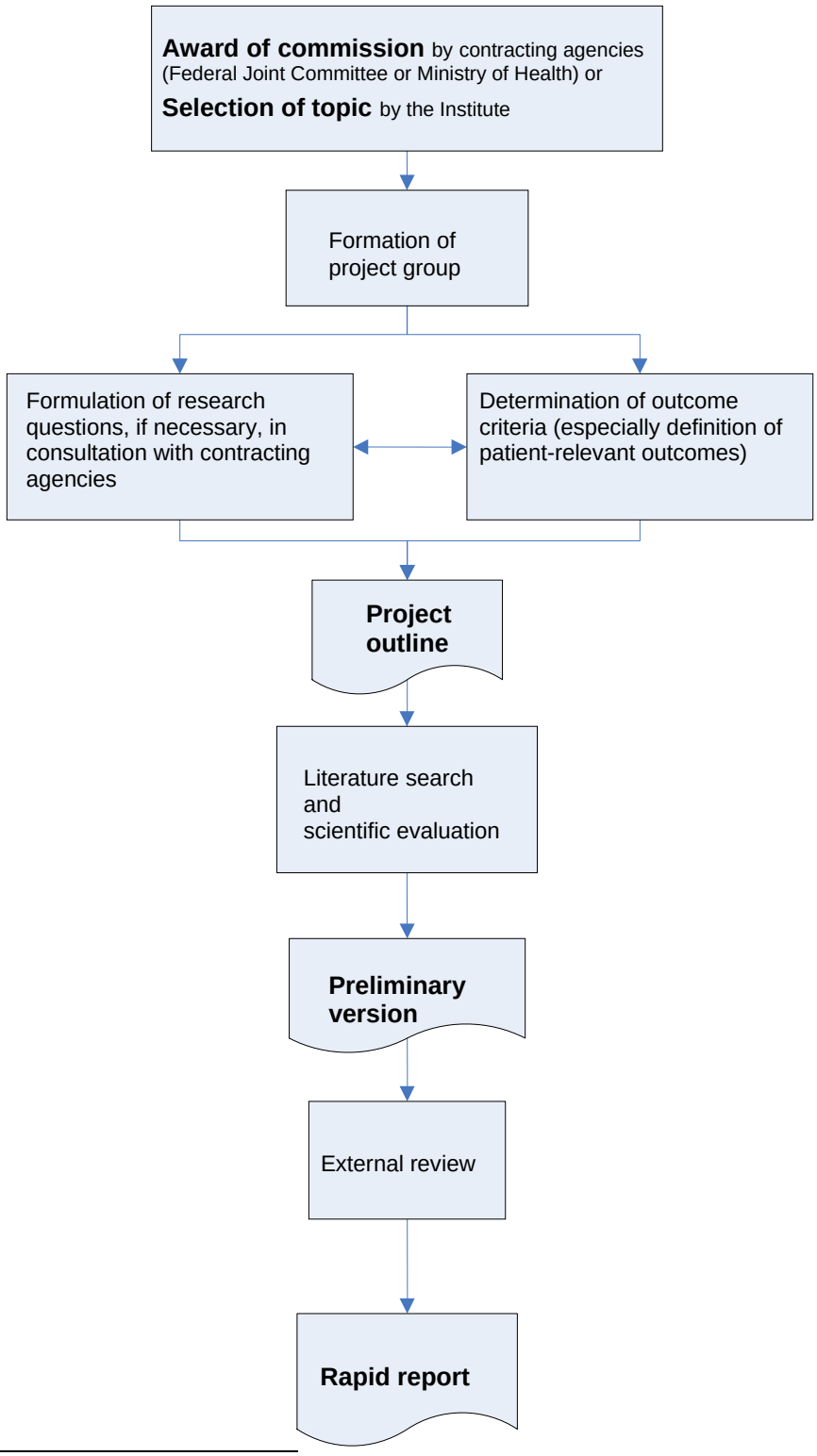
1. Intermediate products are not published;
2. A written hearing is not conducted.

The production of a rapid report can either be commissioned by the Federal Joint Committee or the Ministry of Health, or can be initiated by the Institute itself. In accordance with the report plan generated within the framework of the production of a final report, a project outline for a rapid report is produced in which the most important steps of the literature search and the scientific evaluation are summarised. Depending on the project and the required timeframe, the literature search for a rapid report may be limited, for example, by not considering the inclusion of previously unpublished data. The extent of the literature search is described in the rapid report. A preliminary version of the rapid report is subjected to an external review, and presented to the Steering Committee for internal review. The rapid report is subsequently produced, taking into account the results of the internal and external reviews.

External experts may participate in the production of the project draft and/or the rapid report. Working papers are produced in accordance with the procedure for rapid report production, whereby the performance of an external review is optional.

The schedule for the forwarding and publication of rapid reports and working papers is determined on a case-by-case basis, according to their urgency.

Figure 3: Flow diagram showing the production of a rapid report[p]



[p] All working steps are made on the Institute's responsibility. This can, if required, involve the consideration of external experts, including the Scientific Advisory Board of the Institute. The internal review process is not outlined in this flow chart. For further explanation, please see the preceding text.

4.6 Publication of scientific reports

The Institute's main task is to assess the existing evidence by performing a careful review of the information available, and to utilise its findings with the aim of contributing to the improvement of individual and general health care. The Institute sees itself as having the responsibility to ensure the transparency of its findings, and to publish all findings compiled by the Institute or in the Institute's name. To guarantee the Institute's independence, it must be ruled out that the contracting agencies (Federal Joint Committee or Federal Ministry of Health) or any other interested party has any influence on the content of the reports, as this could lead to conflation of scientific results with political and/or economic aspects and/or interests. At the same time, it must be avoided that the Institute for its part withholds certain findings. All the results produced by the Institute according to its legal responsibilities, including the report plan and a summary of the materials on which the results are based, are therefore published without delay.

If not otherwise agreed, all copyright lies with the Institute.

The schedule for the publication of the preliminary report and the final report is a component of the report plan.

4.7 Literature search

The information, which including its scientific evaluation forms the basis of the Institute's scientific reports, can take a variety of forms, e.g. scientific studies or data collections. The handling of raw data is outlined in Section 1.13. In the following section, the process of a topic-related literature search is described, referring to literature searches conducted both by the Institute's staff and external experts.

4.7.1 General principles of a topic-related literature search

The aim of a topic-related literature search is to identify all relevant publications (i.e. publications that contribute to a greater understanding of the respective research question). The methodology of the systematic literature search therefore follows the general principle that the topic-related literature search concerned must fulfil all criteria which, according to the results of research on this issue, have or could have an important influence on the result (i.e. on the answer to the research question posed). These criteria include in particular:

- The selection of data sources (public databases, private databases, handsearching in selected scientific journals, contacts with experts/industry/patient organisations, etc.);
- Search techniques relating to the selection of study types (RCTs, case reports, etc.);
- Search techniques relating to the medical criteria determined by the research question posed (target population, type of intervention, endpoints, etc.);
- Search techniques relating to formal characteristics of the publication (abstract publications, language, etc.).

Examples are provided by the publications [1-8].

The relevance of these criteria may vary with different research questions. The type of product to be generated (e.g. report, rapid report, working paper) and the resulting timeframe also have an impact on the procedure applied in the literature search. Which criteria are to be applied is assessed on a case-by-case basis. In this context, it is also assessed whether established strategies of international or national working groups (e.g. the Cochrane Collaboration) can be applied.

4.7.2 Search procedure for primary publications

The search for primary publications in bibliographic databases consists of the following nine elements:

1. If necessary, specification of the research question posed;
2. Modification of the research question to a searchable research question;
3. Formulation of a search concept (e.g. language, period);
4. Selection of data sources;
5. Identification of search terms per component of the search concept;
6. Formulation of search strategies;
7. Performance of the search;
8. Storage of the search results in a literature administration programme;
9. Documentation of the search strategies.

In each case, it is reviewed whether consultation with external experts is useful. This may be the case particularly if no specific expertise with regard to the research question posed is available in the Institute.

The search (database query) can be performed by one person.

In this context, it is often useful to initially search for work conducted previously by other working groups, e.g. by searching in specialised databases (Cochrane Database of Systematic Reviews, HTA databases, DARE,[q] etc.). Insofar as relevant reviews can thereby be identified, recourse may be taken to the original search for primary publications performed within the framework of the preparation of these reviews. The prerequisite for this is that the particular search is in line with the Institute's methodology, and that the transferability of the results to the research question posed is assured, particularly with consideration of the inclusion/exclusion criteria in the report plan.

In most cases, even if previous reviews are considered, it will also be necessary to conduct a supplementary search for primary publications, which will then refer to publication periods, languages, or aspects of content which were not considered by the reviews. If no relevant reviews are identified, a search for primary publications is conducted for the whole publication period relevant to the research question posed.

4.7.3 Other data sources for the literature search

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

In addition, depending on the research question, further data sources may be of considerable importance, e.g. study registers or abstract volumes of scientific congresses. In the case of drug evaluations, but also of evaluations of certain (non-drug) medicinal products, publicly accessible drug approval databases or correspondence in this regard are further potential sources of information. Moreover, the manufacturers of the technologies to be assessed are asked as a rule to provide previously unpublished information. The aim of this request is to identify all studies/information relevant to the evaluation, independent of their publication status.

For drug evaluations, this request is usually made in two steps. In the first step, the Institute requests a complete overview from the manufacturer of all studies conducted that included the drug to be evaluated. If appropriate, the Institute defines project-specific inclusion criteria for this overview. From this overview, the Institute identifies studies relevant to the evaluation 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 (q.v. Section 1.11). Previously unpublished information considered in the evaluation will be published in the Institute's reports in order to guarantee transparency. The basis for the incorporation of previously

[q] Database of Abstracts of Reviews of Effects

unpublished information in the evaluation is the conclusion of an agreement on the transfer and publication of study information, which is concluded between the Institute and the manufacturer involved before the submission of data (see sample contract [9]). This agreement specifies the procedure, the requirements for the documents to be submitted, as well as the confidential and non-confidential components of the documents submitted.

The documents provided by the contracting agency for consideration in the evaluation are regarded as a component of the information retrieved. In the subsequent procedure, these documents are handled following the other principles of the literature search and evaluation.

4.7.4 Selection of relevant publications

The selection of relevant publications from the results of the search is usually made in two steps:

- Perusal of the titles and abstracts with the aim of excluding definitely irrelevant publications;
- Perusal of the full papers of the remaining potentially relevant publications.

Both steps are, as a matter of principle, performed by two persons working independently of each other.

The language of publication is usually restricted to those of Western Europe. However, publications written in a different language may also be included if the available information on these publications indicates that additional and relevant information that answers the research question posed is to be expected.

4.7.5 Documentation

All the steps taken in literature search are fully documented. This especially includes:

- The search strategy for the databases selected.
- In addition, if bibliographical database queries are conducted:
    - Documentation of the search;
    - Documentation of the number of results obtained by applying the search strategy;
    - Compilation of the results of the search in a literature administration database (as far as possible, inclusion of complete data sets);
    - Documentation of the publications judged relevant to the research question posed (quotations) after perusal of the primary results;

  - Documentation of the publications not judged relevant after the perusal of full texts, including the reasons for exclusion.

- If personal contact with manufacturers, experts or scientific societies was established:
  - Copies of correspondence.
- The date and period of the search.

References

[1] Pham B, Platt R, McAuley L, Klassen TP, Moher D. Is there a "best" way to detect and minimize publication bias? An empirical evaluation. Eval Health Prof 2001; 24: 109-125.

[2] McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000; 356: 1228-1231.

[3] Moher D, Pham B, Klassen TP et al. What contributions do languages other than English make on the results of meta-analyses? J Clin Epidemiol 2000; 53: 964-972.

[4] Gregoire G, Derderian F, Le Lorier J. Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias? J Clin Epidemiol 1995; 48: 159-163.

[5] Jüni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: Empirical study. Int J Epidemiol 2002; 31: 115-123.

[6] Fergusson D, Laupacis A, Salmi LR, McAlister FA, Huet C. What should be included in meta-analyses? An exploration of methodological issues using the ISPOT meta-analyses. Int J Technol Assess Health Care 2000; 16: 1109-1119.

[7] Sampson M, Barrowman NJ, Moher D et al. Should meta-analysts search Embase in addition to Medline? J Clin Epidemiol 2003; 56: 943-955.

[8] MacLean CH, Morton SC, Ofman JJ, Roth EA, Shekelle PG, Southern California Evidence-Based Practice Center. How useful are unpublished data from the Food and Drug Administration in meta-analysis? J Clin Epidemiol 2003; 56: 44-51.

[9] IQWiG. Vereinbarung über die vertrauliche Behandlung von Unterlagen. Mustervertrag, Entwurf IQWiG/VFA vom 19 Aug. 2005 [Agreement on the confidential handling of documents. Sample contract. IQWiG/German Association of Research-based Pharmaceutical Companies. Version of Aug. 19th, 2005]. Cologne: IQWiG, 2005. http://www.iqwig.de/download/VertraulichkeitIQWiG-VFA_19-8-05_final_2.pdf [accessed on 13.12.2006].

4.7.6 Literature screening

Besides topic-related retrospective searching, early detection and evaluation of current and relevant publications are necessary, based on the systematic monitoring of important scientific data sources.

The term “data sources” includes not only scientific journals, but also the lay press, daily, weekly and monthly press, electronic media, etc.

A publication is especially classified as “relevant” in this regard if:

- The publication is likely to have a considerable influence on the current health care situation;
- The publication is considered to be a milestone study;
- A topic of prime public interest is discussed in the publication;
- The publication refers to a topic high up on the Institute’s internal priority list (q.v. Section 4.9).

The Steering Committee decides which data sources (scientific journals, websites, etc.) are important and should therefore be regularly monitored. The departments evaluate the data sources allocated to them and forward relevant publications to the other departments. A publication may evoke an ad hoc evaluation and an official statement from the Institute. The procedure applied follows defined standards (q.v. Section 4.5).

4.8 Evidence related to the research question posed

It is not the responsibility of the Institute to provide evidence of potential benefits or harms or of the absence of benefits or harms of a medical intervention. It is in fact the responsibility of the Institute to systematically compile and evaluate information that is suited to make interpretable statements on the benefits or harms of a medical intervention. This information has to be generally accessible, or if it is not yet generally accessible, information related to the research question posed has to be provided to the Institute and therefore be made generally accessible. The latter implies that it is in the interest of those institutions and persons who are promoting the application of a medical intervention (or who are rejecting its application) to transfer to the Institute all data available to them that are not yet generally available and are suited to make interpretable statements on the benefits or harms of a medical intervention.

4.9 Priority-setting

A major part of the Institute's resources is used in response to external commissions awarded by the Federal Joint Committee or Federal Ministry of Health; it is their responsibility to set priorities in consultation with the Institute with regard to the start of work on commissions. Moreover, the Institute needs to set internal priorities for processing commissions.

The Institute also needs to set its own priorities with regard to:

- Internal projects of the individual departments;
- Production of health information by the Department of Health Information;
- Commissions awarded to external experts and institutions;
- The Institute's own methodological and scientific work.

Priority-setting may involve particular issues that need to be considered in individual areas. Each department has the option of developing specific priority-setting procedures based on the specific activities of the department. However, overall, priority-setting processes developed by the Steering Committee and individual departments aim to be straightforward and consistent with the general rationale, values, and methods set out below.

Priority-setting is not carried out following a rigid process. Nevertheless, it is essential to aim for fairness and transparency in the distribution of intellectual and financial resources in a public organisation such as the Institute.

Priority-setting for research activities and the systematic evaluation of medical procedures and technologies needs transparent mechanisms, and must be in agreement with the rationale, values, methods, and criteria of the Institute. In this way, it is ensured that the activities and priorities of the Institute can be reviewed and understood externally.

4.9.1 Background of the Institute's priority-setting

Through its activities, the Institute aims to make a beneficial impact on the health of the population in Germany and contribute to the development of public and scientific understanding of health.

In decision-making procedures for the work to be undertaken, the Institute considers the:

- Proportion of the population likely to benefit from the work the Institute chooses to undertake;

- Burden of disease (including cost), currently and in the future, for individuals and society, particularly for disadvantaged groups;
- National health priorities;
- Potential of the Institute to influence clinical practice, clinical decisions, and the health status of the population;
- Resources required by the Institute to perform the activity effectively;
- Potential of a beneficial impact on society or the health care system, considering the principles of equity;
- Potential harms;
- Unique contributions the Institute could make, including the assessment of what others are doing or who else could potentially undertake certain activities;
- Contribution to the quality assurance of the Institute's work;
- Potential contribution to the growth of scientific knowledge.

4.9.2 Processes for priority-setting

Generally, a wide variety of models and processes exist that are used for priority-setting in organisations similar to the Institute and have some features in common. These form the basis of the Institute's model of priority-setting and include:

- Collection of data and opinions (q.v. Section 1.24) to provide a basis for decision-making, and their ongoing documentation;
- Application of the relevant criteria developed by the Institute or its departments;
- Provision of a report of the data situation and prevailing opinions (together with a recommendation) to the Steering Committee, which makes the final decision and documents the reason for making this decision.

4.10 Production times of reports

The evaluation of a method or intervention is theoretically possible at any time. However, IQWiG has the responsibility only to evaluate pharmaceuticals that are licensed to be prescribed, approved according to the German Pharmaceutical Act, and available on the German market.

Furthermore, the time point of evaluation is determined by the timing of the award of the commission to the Institute. There is no general directive that the production of a scientific report by the Institute takes place, at the earliest, following the passage of a certain period of time after approval or establishment of a method or intervention. If, in the case of a report ahead of schedule, there is great uncertainty about results due to a lack of long-term studies, this will be described in accordance with the general working methods.