Eur Arch Psychiatry Clin Neurosci (2012) 262:277–289
DOI 10.1007/s00406-011-0273-8

ORIGINAL PAPER

How close is evidence to truth in evidence-based treatment of mental disorders?

Hans-Jürgen Möller

Received: 20 October 2011 / Accepted: 28 October 2011 / Published online: 22 November 2011

Abstract Given the importance of the term 'evidence' in evidence-based medicine (EBM), the meaning of this term is evaluated, going back to the philosophical tradition and current meaning of the terms 'evidence' and 'truth'. Based on this, current problems in the definition of evidence and in the grading of evidence in EBM are described, taking examples from the field of psychiatry and especially pharmacopsychiatry. These problems underline that the use of the term evidence in EBM is inconsistent and inconclusive. This should be fairly stated in all EBM-related publications, especially in EBM-based guidelines, to avoid severe misunderstandings in and outside the field of psychiatry. Although EBM might have increased empirically driven rational decision-making in psychiatry/medicine, the current limitations should be carefully considered.

Keywords Truth · Evidence · Evidence-based medicine (EBM) · Mental disorders

Introduction

Evidence-based medicine (EBM) has become a key concept in modern health care [80, 86, 87] and thus also in psychiatry. In contrast to the traditional everyday use of the term 'evidence' in the sense of intuitive insight/experience, in the context of EBM 'evidence' is understood as the sum of available empirical knowledge on a particular issue.

'Evidence' is the key concept in EBM and indicates whether and to what extent a statement about a certain issue is empirically founded. The evidence grade may have consequences for diagnostic/therapeutic procedures, for example in the form of guidelines or even directives [80] as well as on health care organisation and payment. Considering all this, it has a significant influence on the health care system [82, 89].

If the concept of 'evidence' is transposed to a more philosophical diction, one could assume that it characterises the degree to which the respective empirical data concur with reality or represent reality. If one follows this philosophical diction, it seems reasonable to put the term 'evidence' close to the philosophical term 'truth'. However, one has to consider a priori the limitation that this is a concept of truth within the framework of an empirical approach and that therefore it by no means includes all perspectives of 'truth' from a philosophical point of view.

This article describes in the first part some aspects the terms 'truth' and 'evidence' in the philosophical tradition and how this terminological background is related to the term evidence in EBM. The second part will investigate how close the current concept of evidence in EBM comes to this philosophical tradition and meaning and what, under the aspect, the key problems of the terming of evidence are related to the treatment of mental disorder.

Evidence and the concept of truth in philosophy and modern philosophical theory of science

It seems reasonable to apply the concept of evidence as used in EBM to the traditions and definitions of evidence and truth in the traditional philosophical and modern theory of science. For principal reasons in an article for a psychiatric journal, this can be done only on a quite superficial level and in an abbreviated way. When doing so, one soon

H.-J. Möller (✉)
Department of Psychiatry, Ludwig-Maximilian University Munich, Nussbaumstrasse 7, 80336 Munich, Germany
e-mail: hans-juergen.moeller@med.uni-muenchen.de

realises that there are close links between the concepts of evidence and truth in the context of philosophy and the modern philosophical theory of science. Such conceptions appeared in philosophy as early as Graeco-Roman times and can be followed throughout the history of philosophy within the context of different theoretical approaches and perspectives, resulting in specific implementations. To deal adequately with the associated problematic, one would therefore have to provide a summary of the history of philosophy related to this topic, which would exceed the scope and size limitations of this article. It is important to note that both concepts were dealt with not only in the context of empirical scientific method but also—in particular as concerns the concept of truth—in the context of mathematical proofs or formal logics reasoning and even in the context of metaphysical, for example theological approaches. These last approaches do not appear to be relevant to the concept of evidence in EBM, for which only the definitions of evidence and truth in the context of empirical science are relevant, as philosophically discussed mainly in the modern philosophical theory of empirical sciences [71–73] in connection with the question of how closely a scientific statement is related to the 'reality', meaning how it is supported by empirical data.

In other philosophical contexts, the term evidence describes an insight without methodical mediations [67]. In its Latin form 'evidenzia', M.T. Cicero—in continuation of an analogue conceptualisation among the stoics and epicures—considers the expression analogous with clarity/clearness. In philosophical tradition, its meaning as comprehension without prerequisites or as clear certainty, as formulated by Kant, depends on pre-existing theory of knowledge positions and is accordingly inconsistent. There is variable use of the terminology:

1. In the evaluation of evidence either as the subjective form of acceptance of truth (evidence as the observation of a phenomenon) or as the objective form of establishment of truth (evidence as being shown a phenomenon)
2. In classifications such as metaphysical, logical, psychological and physical evidence, which themselves can be subdivided into subjective or objective evidence.

Contrary to all the above uses is the concept of discursive or conceptual, that is methodically (through proof, explanation, etc.) advancing understanding.

In its dominant meaning, the concept of evidence represents a truth criterion for the 'first' propositions of a theory, the so-called axioms. From a methodological view, in this context an indication of evidence replaces an explanation of ('deductive' and no longer explainable) initial propositions as demanded already by Aristoteles. R. Descartes in particular makes the concept of evidence relevant to knowledge theory in a more general sense. Contrary to the primacy of discursive (logical) procedures, evidence is labelled according to the terminology of concepts of intuition and of clear and distinct opinion. Ideally, chains of evidence should replace chains of logical, independent propositions.

In theoretical reasoning contexts, the problems of a methodically proven concept of evidence are linked—just as is the case with the concept of intuition—to a controllable reference to evidence or to the requirement to differentiate between evidence and apparent evidence. In the particular modern analytical theory of science, for example, Stegmüller referred to this issue [93]. Evidence is used at all times as a postulate for recognising parts of an argument and, in the form of fabricated approval, taken as given. In this respect, evidence either becomes apparent or remains absent but cannot be proven as such. Evidence thus belongs to the pragmatic constituents of every argumentation and understanding. This short review of the philosophical tradition and current philosophical concepts of evidence already offers a basis for understanding that the current use of the term evidence is far away from the tradition and current use of this term in philosophy. The use in everyday's common language and clinical thinking is apparently much closer to the true philosophical terminology. Probably, the term truth in its philosophical use might come closer to the concept of evidence in EBM.

The concept of truth is much more complex than the concept of evidence, both in philosophical traditions and in current philosophy [68]. In its use in everyday and academic language, the concept of truth means as much as correctness or validity, while philosophical usage knows a wealth of terminological definitions that delineate 'true' from 'correct' and 'valid' but also from other terms with similar meanings. Generally speaking, truth is related to knowledge but is normally understood to be limited to theoretical knowledge and learning. Whereby or where and when knowledge is gained is a question of the tools and contexts. One should speak of truth, as distinguished from taken-to-be-true or being convinced, only in relation to such tools that let apparent knowledge be differentiated from genuine (also 'true', 'real') knowledge. The everyday use of the term 'truth' is thus abandoned in favour of a philosophical-reflecting usage without already making decisions for or against a specific theory of truth. Especially in the modern philosophical theory of science, the term truth is predominantly related to the question of how closely a scientific statement/hypothesis is supported by empirical data [15, 16, 93]. Based on this, it seems obvious that the current use of the term evidence in EBM comes much closer to this philosophical meaning of truth than, as already stated before, to the philosophical meaning of

evidence. Even more, the principal question to be answered is whether the use of the term evidence comes close to the philosophical use of the term truth.

Without going into the highly complex history of the truth concept and of truth theories in philosophy, just a few important aspects will be presented here. Kant introduced important differentiations concerning the type of truth of a judgment that went beyond the characteristic distinction typical for Descartes and Leibniz between 'contingent truths' or 'factual truth' and 'necessary truths' or 'sensible truths'. This distinction, which relates to the modality of a judgment, makes use of the type of reasoning used to explain the validity of a judgment, that is it makes use of the way words are used to articulate the research procedures applied to judge the truth or falsity of a representational statement. Word usage plays a communicative role, for example as statements about observations of visual perceptions (factual truths) or as statements about analytical coherences concerning conceptual analyses (sensible truths). Kant replaces this distinction with the distinction—still in use today although referred to in different terms—between 'material' and 'formal' truths. In modern logic, 'material' and 'formal' truths are differentiated from statements such that in the material case truth depends on the components of a statement while in the 'formal' case truth in the sense of freedom from contradiction can only be determined through purely logical operations. Furthermore, Kant distinguishes between 'analytically true' and 'synthetically true' judgments according to the criterion of whether they are true on the basis of the words used in the judgment and are thus a priori (e.g. purely rational constructions in arithmetic or geometry; a priori truths) or whether they require additional experience of meaning and reasoning, that is empirical experiments (a posterior truths). In the Leibniz classification, Kant's synthetic a priori true judgments belong to the 'sensible truths', while in modern analytical scientific theory—unlike the constructive scientific theory that is close to Kant's understanding—they are reconstructed as hypothetical truths, namely as propositions of axiomatic theories whose axioms are always merely assumptions.

In the different truth theories, which were formulated in modern times but refer to historically pre-existing propositions, during the discussion about different truth criteria for the substantiation of truth, one of these criteria is generally chosen as a standard for a definition of truth that leads to a concept of truth. The coherence theory of truth, developed from formal logic, plays an important role in this context. This theory assumes that coherence requirement 'A' is true if and only if 'A' consistently coheres both conceptually and logically and in addition can be embedded in a comprehensive system of propositions in colloquial and scientific language structures. Tarski [94], who worked intensively on the definition of truth in the context of formal language, gives the following concrete example: 'Snow is white' is true if and only if 'Snow is white'. This concept of truth developed from formal logical language was then extended to natural languages in terms of a pragmatic theory of truth that is oriented towards the consensus requirement: 'A' is true if and only if every expert and willing person could have agreed.

Popper introduced the concept of 'truthlikeness', also known as 'approximate truth', into scientific theory as a measure of the agreement of a hypothesis or theory with truth [69, 84]. In this way, Popper wanted to combine the fundamental ideas of his scientific realism—whereby science is the search for truth—with those of his falsifiability concept, which proposed that the theories which are actually at the disposal of a research community and can be accepted only hypothetically can in principle always be false. Popper's idea was to explain scientific progress and rational choice of theories in a non-inductive way as an enlargement of the truthlikeness of theories that replace each other, that is as getting closer to truth.

Despite being etymologically related in many languages, the objective concept of truthlikeness has to be strictly distinguished from the 'epistemic' concept of truth (certainty, reliability, inductive support) of a theory or empirical proposition. The respective differentiations were already made by Locke. According to Popper, scientific theories should have as much empirical content as possible. This content results from a large number of possible empirical counterexamples and corresponds with a high explanatory power and information content of the theory [84]. As a consequence, such a theory cannot be expected a priori to be valid; the a priori probability of its validity is low. The concept of truthlikeness should combine a measure for the degree of correspondence with the truth with a measure of the empirical content. Popper sees the regulative ideal of scientific research in maximum truthlikeness. He defines the logical content of a hypothesis or theory 'A' as the class of its logical consequences in terms of the empirical content. Its 'truth content' lies in the class of its true consequences, and its 'falsity content' lies in the class of its false consequences. It seems obvious that this philosophical approach comes very close to the current meaning of evidence in EBM. This seems to be the fact also for the following correspondence theory of truth.

The correspondence, or adequacy, theory of truth was the truth theory that dominated philosophy for long periods in the past and is still of relevance in the modern philosophical theory of science [93]. This theory assumes that truth is the correspondence between cognitive beliefs/hypotheses and reality. As a matter of principle, this theory understands truth to be a relation between two reference points. Descriptions such as agreement, equivalency,

adequacy, concurrence, etc. are needed to describe this relation. The corresponding relations are also defined differently: to think/to be, subject/object, consciousness/world, knowledge/reality, language/world, claim/fact, etc. Many philosophical problems arise from the correspondence theory; these were widely discussed particularly at the turn of the 20th century, resulting in the development of alternative theories of truth. Problems arise in determining the 'truth-bearer'. What are the objects or entities that should correspond with the facts or reality and that we call *true* in this context? On the other hand, the question of the 'truthmaker' arises, that is what do propositions have to agree with in order to be true? Although correspondence theory supporters mainly agree that *truthmakers* are 'facts', they disagree as to what 'facts' actually are. This discussion underlines that philosophical thinking is a continuous questioning which probably never comes to a satisfying end. Without neglecting the principal need for such a thorough reflexion to the frame of this article, it is probably necessary to understand that the correspondence/adequacy theory of truth and the concept of truthlikeness seem to be of great value in relation to the definition of evidence in EBM.

According to the aims of empirical science, both truth theories/concepts try to relate a scientific statement to the empirical experience, excluding this solely logical or metaphysical deduction/conclusion. Truth is related to the observable world, and it is important to find out by which methods we can make the adequate observation to answer the question what kind of observations are able to support a statement/hypothesis.

Another more principal philosophical question related to this is whether this whole process of developing a hypothesis/theory is primarily and only inductive, that means started by observations and based on this develop a hypothesis which than is confirmed by further observations or whether in contrast to this like Popper suggests a hypothesis can only be falsified but never confirmed by observations [71]. From an EBM point of view, both positions seem to be principally acceptable. However, the evidence definition of EBM is obviously closer to the naive inductive approach (inductive realism) than to the falsification theory (empirical approach) by Popper. Nevertheless, Popper's concept of 'truthlikeness' seems to come close to the evidence grading in EBM, at least in a relatively naïve conceptualisation that does not follow the complicated philosophical constructions defined by Popper.

The problems of the principal concept of evidence in EBM

In EBM, evidence is the result of a critical appraisal of (published) results of scientific studies. The evidence can relate to various areas, such as diagnostics or therapy, for example. According to the codex of EBM, the complex medical knowledge on a certain area is processed and summarised in meta-analyses, systematic reviews, etc., and then the evidence-based findings are translated into recommendations or treatment guidelines approved by specialised committees [2, 18, 19, 37, 47, 49, 63]. A physician should follow these evidence-based recommendations when making diagnostic or therapeutic decisions. This should make clinical care better empirically founded or more rational [24]. As little room as possible should be left for subjective discretion of the physician.

In EBM, the most decisive level of scientifically proven evidence—as far as therapeutic aspects are concerned—constitutes randomised, control-group studies. The knowledge obtained from single cases or the cumulation of such cases representing the lowest level of evidence is only relevant as a supplement to such studies or as a substitute if empirical studies of a higher methodological level are lacking. This view agrees with the principal methodological understanding of empirical research [78] but is also judged critically, in part because of the limitations it places on study methods and the associated negative consequences for patient selection and validity in everyday care situations [77]. These problems can be briefly characterised with the poles of internal validity/poor generalisability versus external validity/good generalisability. This discussion shows that the different study methods only deal with excerpts of reality and do not represent the full reality; the differences of studies can lead to different results [92].

Generally, in randomised, controlled, phase III studies in psychiatry, as a rule only about 10% of the patients can be recruited who would in principle be suitable for the study [13, 17, 57, 85]. The recruited patients are selected according to various selection criteria to enter a study (e.g. exclusion of comorbidity, older patients and patients at greater risk), and they are therefore not even representative of the sample of patients with the diagnosis of interest at the respective treatment centre, let alone of the whole group of patients with this diagnosis. The more study-related criteria interfere with patients, the greater the problem of selection. Placebo control is especially problematic in this respect and leads to a particularly high degree of selection. One thinks for example of placebo-controlled studies in patients with mania, which include only those patients who have relatively mild manic symptoms. The same applies to placebo-controlled depression studies in which, for example, patients with severe depressive symptoms or suicidality, or both, are mostly not included.

To avoid the consequence that by favouring study types with too little generalisability EBM will lose touch with clinical reality, other empirical research approaches should

be given greater consideration. A medication that has been evaluated in placebo-controlled studies, with the associated selection problems described above, should be tested in addition in studies with less restrictive methodology, for example randomised, control-group studies versus a standard drug, non-interventive studies, etc., and the results should at least show a tendency to be consistent. In recent years the principal problems related to these methodological problems have been discussed especially in the context of so-called real-world studies or 'effectiveness studies'. Although these studies might have advantages because they can include patients a less restrictive way, it is obvious that at the same time they have lots of methodological problems and come therefore by no means closer to the 'truth' than the classical efficacy studies [75, 78].

It should be noted that traditionally demands have been made that the clinical evaluation of a psychopharmaca should follow a phase model that includes various methodological levels of empirical research and approaches of different methodological stringency. This means that the results of methodologically restrictive randomised control-group studies (RCTs) of phase III studies have to be supplemented with efficacy and particularly tolerability evidence from phase IV studies, which are more closely oriented towards routine care [58, 91]. For the purposes of this phase model of clinical/pharmacological evaluation, the evidence of each testing phase would be viewed as part of the complementary overall evidence [26, 53]. This view is no longer recognisable in evidence grading criteria applied to guidelines because evidence is rated according to the methodologically most demanding study design for the respective therapy (e.g. placebo-controlled studies or randomised control-group studies without placebo arm). There it is not determined whether consistent results from less restrictive study types, which have better generalisability, are also available. An evidence rating that is more relevant to clinical reality should assess whether both studies with high internal validity and low external validity (e.g. randomised control-group studies) and studies with high external validity and low internal validity (e.g. randomised 'real-world' studies, observational studies) are available and in principle have the same results [80].

These principal methodological considerations demonstrate that the principal concept of evidence in EBM might be related to the concepts of 'truth' in the modern philosophy of empirical science in the sense of how far is a statement/hypothesis supported by empirical data. It is difficult to understand why EBM prefers the term 'evidence' to the term 'truth', although the term 'evidence' in the philosophical as well as in the everyday terminology has another meaning (see above). Did the founders of evidence try to avoid the extremely challenging concept of 'truth'? This might be a reason, especially if one considers that in contrast to the sophisticated discussions on how to determine truth or how to come close to the idea of truth, the EBM practises a rather naïve approach, without discussing the different methodological aspects and their different consequences (e.g. study designs).

The point that 'reality' ('truth') is only depict prioritising the design type with high internal but low external validity, while devaluating design with high external and low internal validity is insofar a principal problem from a philosophical standpoint, because due to this a comprehensive and complementary view of the reality is not generated.

Basic considerations on evidence, evidence criteria and evidence grading

In view of EBM, the two most important approaches to determine evidence [50] are systematic reviews and meta-analyses. Systematic reviews use a narrative form to give a critical presentation and qualitative evaluation of the studies available on a certain question. The advantages and disadvantages of individual studies are argumentatively weighed up against one another, and a result of the studies as a whole is presented qualitatively (x has better efficacy than placebo or has the same efficacy as an active drug). This strategy is applied by regulatory authorities, for example, to assess the efficacy of medications; the result of the assessment is an evaluation of whether the new compound has better efficacy than placebo or the same efficacy as a standard treatment. This procedure was used in a modified form to develop some guidelines (e.g. by the World Federation of Biological Psychiatry [8, 11]). For the purposes of EBM, systematic reviews have to fulfil high methodological demands with respect to the completeness and critical evaluation of the studies involved and therefore go beyond other reviews that do not follow such strict requirements.

Meta-analyses quantitatively combine the results of studies performed on a certain question and classed as being methodologically adequate. They calculate an effect size that expresses the quantitative difference between two comparative substances (e.g. placebo vs. active drug, medication A vs. medication B). A comparison of effect sizes would principally require that the studies have been drawn from the same population. However, this requirement is fulfilled only approximately at best, if at all, since the different studies to be combined have different designs and for the most part different framework conditions (e.g. with respect to setting variables, inclusion and exclusion criteria of patients, pre-treatment, concomitant medication). This often leads to comparing apples and oranges, without the necessary differentiation between different

studies and their respective specific conditions. For example, it would not be meaningful at all to combine in one meta-analysis the results from phase III and phase IV studies to determine the relevance of related confounders. Often applied sensitivity analyses cannot solve this problem in a sufficient way.

The results of meta-analyses are gaining increasing importance in the development of EBM-based guidelines [26] and textbooks, perhaps because a quantitative summary of results in effect sizes is easier to convey than the differentiating, qualitative conclusions reached by systematic reviews. Indeed, in comparison with systematic reviews meta-analyses do have the advantage that they can condense the results to quantitative parameters (effect sizes), while reviews only draw somewhat complicated qualitative conclusions. Nevertheless, meta-analyses cannot replace systematic reviews in their narrative form, which have the advantage that they can consider in a differentiated manner the special characteristics of individual studies with respect to study design, patient selection, drug dosing, etc. and can come to more cautiously phrased conclusions. Precisely, this detailed analysis requires a high degree of clinical-psychopharmacological expertise and a detailed presentation, both of which are not always apparent in the often relatively brief systematic reviews introducing a meta-analysis. Both procedures should be viewed as complementary. There is no justification for giving meta-analyses priority over comprehensive narrative systematic reviews. The numerical value of an effect size, which appears so clear and meaningful, is full of ambiguities resulting from fundamental methodological problems of meta-analyses [56, 61]. The apparently so convenient and illustrative value of an effect size can be interpreted only too easily in a naively simplifying or purposely tendentious way because the complex conglomerate of clinical data on which the analyses are based are no longer evident. Over-interpretations of effect sizes or other global statistical measures as the final relevant decision criterion, as can often be read nowadays, are inappropriate in view of different fundamental problems of meta-analyses and have to be critically questioned in every case.

Taking the recent meta-analysis on the effect size of modern antidepressants in comparison with placebo as an example, one can easily demonstrate that only based on different statistical methods of meta-analysis the placebo–verum difference as well as the effect sizes can change in a relevant way, although the analyses were based on the same data set [32, 45, 51, 76]. Another example is the fact that the exclusion of only one single study for whatever reason can change the effect size or similar global/abstract measures in a significant way. Due to this too global/abstract approach in the meta-analytic methodologies, apparently the difference of two groups of compounds with the same indication, like the differences between FGAs and SGAs, is described with inconsistent results in different meta-analysis [27, 35, 55]. Similarly, contrasting results of meta-analysis can be found in answering the question whether tricyclic antidepressants or dual action antidepressants can have superior efficacy compared to SSRIs [10, 22, 81].

Having correctly appraised evidence as not being definable in an absolute sense but as being a relative concept, EBM introduced the concept of grading evidence. Evidence grading is based among other things on the assumption that for methodological reasons the use of certain study designs leads to results that are more likely to be reliable (see above). This corresponds with the rules of empirical research methodology [14, 29, 74, 78]. According to these rules, randomised control-group studies have more value than non-randomised or uncontrolled studies and so on.

It must be stressed that, even more than when defining evidence, evidence grading is not a trivial process in which the empirical data level is transposed 1:1 but consists of processes that are full of problems and that far exceed the data level [3, 60]. This is already true for evidence grading but significantly more so for recommendation grading [42]. Because of the problematic of recommendation grading, some guidelines, for example the most recent NICE schizophrenia guidelines, do without a recommendation grading [83].

These considerations underline that again that EBM has in contrast to the sophisticated philosophical consideration to determine the truth following a rather naïve way of describing the reality, that is the truth, by summarising the empirical data. Methodologically spoken is this rather monistic with its preferences for statistical meta-analysis instead of opening the view as far as possible also to other complimentary approaches to summarise the observational data, like the qualitative review. The chance to find a better description of the ‘reality’ (the ‘truth’) by comparing results of both approaches in an open way is insofar totally neglected. In this context, also the rather naïve interpretation of the results of a statistical meta-analysis as if one meta-analysis could determine the ‘real/true’ results is far away from understanding that meta-analysis as such are associated with a lot of methodological problems and can therefore by far not determine ‘truth’ per se. Even meta-analysis based on the identical data set can come to different results. This relativity of the results of meta-analysis should be taken into account much more frequently.

The evidence grading has some similarity with the philosophical concept of truthlikeness. However, all the problems mentioned in the previous paragraph regarding evidence determination in general are consecutively relevant to evidence grading.

Differences in evidence criteria and evidence grading

The set of criteria for the different levels of evidence, which appears to be clearly formulated, is de facto full of risks of inconsistency and does not correspond by far with an operational definition. This becomes clear when one focuses on the respective details, for which there is no space here. The reader is referred to publications that deal with this problem in detail [6, 80].

The principle problem is that there is no uniform, internationally accepted definition of evidence and the derived evidence levels, even though the term 'evidence level' suggests that the definition is unambiguous. Thus, the choice of evidence criteria or evidence levels alone results in very different outcomes for the respective issue. A random selection of some concrete examples, as presented in Maier and Möller (2010), makes this clear.

As described before, EBM as a whole and many different guidelines prefer to base evidence on randomised controlled trials (RCTs), neglecting other methodological approaches. It is unclear whether results of placebo-controlled trials have priority over non-placebo-controlled trials, which would make sense as regards the rules of empirical research in psychopharmacotherapy and would correspond with the demands of regulatory authorities' [78]. The criterion of testing under double-blind conditions is also usually not viewed in a discriminatory way (see below), which has severe consequences. A particularly contentious point is whether the results of important, methodologically outstanding single studies have priority over the results of meta-analyses of all studies related to a specific issue [6, 23, 80]. Most guidelines prefer results from meta-analyses alone or together with results from single studies. Narrative systematic reviews do not appear to play a role in the evidence grading of guidelines, or at least they are not listed in most evidence gradings, although they would deliver important complementary aspects to the statements of meta-analyses. This underlines that the term evidence has not a clearly defined meaning but is filled with different contents. The most common definition is based on the results of RCTs. But especially for evidence grading, further differentiation in terms of methodology would be necessary (see below). Currently, for example evidence level A or I, respectively, could have different meanings in different guidelines or other EBM contexts, from the relevance of evidence and evidence grading in terms of health care, but also allocation of finances and other resources in health care raise severe concerns. Such a volatile definition of evidence is far away from all sophisticated approaches in modern philosophy to determine truth and to access truthlikeness. Due to these limitations of defining evidence and evidence grading, which are not discussed in the contexts of the different guidelines and because most guidelines use arbitrary definitions in a way that readers get the impression that this is the only or even best way to define evidence, EBM currently is in the problematic condition that a lot of inconsistent definitions of evidence confuse the users. This is even a risk situation because, for example, health politicians could be pushed into the direction to make wrong decisions, focussing only on one aspect/definition of evidence. Apparently, there are different evidence/truths on the same issue, only due to different definitions of evidence.

Meta-analytic results versus results of excellent single studies as the basis for the highest level of evidence

Most guidelines define the highest level of evidence by using meta-analyses of RCTs (see above). The prioritisation of meta-analyses is not as unproblematic as it initially appears to be [46, 50, 56, 62, 70, 80].

The basic requirement for performing a meta-analysis and also for the later use of its results is methodological stringency. This is true for the systematic search for studies to be included, for the evaluation of their methods and especially for the clinical evaluation of possible heterogeneity. Especially the different methods of performing meta-analysis and their relevance for the results have to be considered. When developing guidelines, care should be taken when considering meta-analyses of only small RCTs or of RCTs of poor methodological quality and also with respect to the uncritical use and transposition of results of meta-analyses, for example information relating to effect sizes or numbers-needed-to-treat. Especially the results of different meta-analysis on the same issue, reaching inconsistent or even contradicting results, are critical.

It is important to emphasise that the large regulatory authorities such as the North American FDA and European EMEA do not recognise meta-analyses as the primary basis for deciding whether to approve a drug. They base their decisions on the methodological rules of the experimental testing of hypotheses: The results of methodologically adequate clinical trials (in particular phase III studies, mostly placebo-controlled studies) are considered under the aspect whether the hypothesis of efficacy and tolerability is supported by the results of several studies (confirmation or not falsification). The resulting conflicts with the meta-analytical EBM approach are predictable: in an extreme case, an approved substance can be classified on the basis of meta-analyses as lacking efficacy in the context of EBM because—in contrast to the situation with the regulatory submission—not only pivotal phase III studies are evaluated but also other studies with different objectives that were often not primarily performed to prove

efficacy, and vice versa. To put it differently, an EBM with treatment recommendations/guidelines that view meta-analytical results as the highest level of evidence may reach results that are different from and perhaps even contradict those of the regulatory authorities, because a different decision-making logic is being followed.

An example of this is the situation of lamotrigine in the treatment of acute bipolar depression. Lamotrigine was licensed for relapse prevention of bipolar disorder based on sufficient data [12], but not for treatment of acute bipolar depression because most placebo-controlled studies were negative; however, a positive result, even showing only a low effect size, in the meta-analysis by Geddes and co-workers [34] could lead to a positive evidence grading, probably if the low effect size is not taken as a criterion against the highest evidence grading. But this would confirm two separate aspects: The size of the effect on the one side and the way efficacy is proven as evident. There is principally the possibility that a high effect size is proven with a low level of evidence and a low effect size with a high level of evidence.

The treatment guidelines of the World Federation of Biological Psychiatry [5, 6, 8, 9, 11, 30, 31, 38–41] relate to a different system of grading evidence. The crucial difference from the evidence criteria of many other guidelines is that the highest level of evidence is not based on the results of meta-analyses but on the results of important and methodologically outstanding and adequate single studies [79]. In this respect, the set of criteria for evidence principally corresponds with the approach of the regulatory authorities.

Again this specific aspect underlines that evidence and evidence grading is apparently distinct from 'truth' or 'reality'. There are apparently different evidence/truths. This has become more and more of a problem in countries where apart from the drug authority there exist also an EBM institute, for example such as the NICE (National Institute of Clinical Excellence) in the UK. A consequence of these two institutions can be that medications, for which the national or European drug authority, the EMA (European Medical Agency), found evidence in the above-mentioned sense and therefore licensed the medication, are finally taken out from the reimbursement in the context of the governmental health care system based on the 'evidence' judgement of NICE following an EBM meta-analytical approach. This is of course confusing and far from being rational. But these two ways of determining evidence are principally rational however, following a different logic. For the health of the health care system, it would be better to find a way out of this dilemma of 'two truths'.

It is extremely important to achieve an internationally uniform grading of evidence, although this will not be easy in view of the problems described above. International working groups of methodologists, for example GRADE, are endeavouring to achieve a standardised evidence grading [3]. It would be meaningful to take into account in this context not only statistical meta-analysis as methodological background but also the focus on single studies.

Placebo-controlled studies versus studies with a standard drug control as a requirement for the highest level of evidence

Some guidelines, such as the WFSBP guidelines, demand placebo-controlled studies as a prerequisite for the highest level of evidence. In other guidelines, such as the APA Practice Guidelines [1], other randomised, controlled treatment studies (in particular randomised, control-group comparisons of a new substance with a standard drug) are sufficient, and often do not even have to be double blind, for example. The APA Practice Guidelines differentiate only minimally between evidence from randomised, double-blind, control-group studies—which result in the evidence level [A]—and evidence from non-blinded control-group studies—which result in the evidence level [A-]. If studies in which a study drug was tested against placebo are combined in the highest class of evidence with studies in which it was tested against a standard drug, perhaps even without a differentiation being made between blinded and non-blinded studies, it means that study types of different validity are being put on a par with each other. This is not meaningful: it is known that at least in several psychiatric indications, for example depression, studies without a placebo control do not allow valid conclusions to be drawn because of immanent method problems (internal validity). For this reason, the large international regulatory authorities (e.g. FDA, EMEA) demand placebo-controlled studies [4, 33, 78].

On the other hand, a too one-sided and extensive overemphasis of the relevance of placebo-controlled studies is not desirable. Although such studies are essential for proving efficacy in many indications, there is often no guarantee that such study results can be generalised to routine clinical care settings (problem of internal versus external validity). It is known that placebo-controlled studies of new psychoactive drugs performed for regulatory purposes have the problem that they are particularly far distanced from routine clinical care and thus have to be understood rather as 'proof of concept' studies. To avoid these dilemmas, it would be better to establish evidence grading, which carefully takes into account the different methodological characteristics of studies, giving the highest evidence grade to the results of the most sophisticated study approach. However, on the other side, also the outcome of less restrictive 'real-world' studies should be taken

into account to avoid regulations primarily based on extremely selective samples.

Principal differences in grading evidence in psychopharmacotherapy and psychotherapy

This article cannot examine the fundamental problems of efficacy research in psychiatry [74, 90], but only problems that arise when effect sizes or evidence evaluations from psychotherapy research are directly compared with effect sizes or evidence evaluations from clinical psychopharmacology [7, 36, 44, 52, 54, 59, 95].

Now that effect sizes are increasingly being calculated and evidence grading introduced to depict the empirical evaluation of psychotherapy/psychosocial therapy, the possibility exists in principle to compare these parameters with the evidence parameters from the area of psychopharmacotherapy. However, this carries the risk that effect sizes or evidence ratings based on different methods of therapy evaluation will be compared with each other, which is meaningless. A psychotherapeutic method X would probably be assigned the highest evidence level for treating depression, in certain guidelines can reach the highest level of evidence only on the basis of double-blind, randomised and perhaps placebo-controlled studies, although it was tested 'only' in an unblinded randomised control group and this is associated with the problem that it is difficult to establish placebo control conditions. Of course, this is a principal limitation of experimental therapy evaluation in psychotherapy and nobody can be blamed for this [43]. However, it is important to take this into account regarding antidepressant in certain guidelines.

Even taking out the condition placebo-controlled design and rely only on double-blind comparisons of the two active treatment conditions, the consequence is the same: Drug treatment can be evaluated in such a way, while psychotherapy mostly is not elevated following these stringent conditions, with the exception that some psychotherapy studies tried to come closer to this methodological idea by using blinded raters, which is of course also far away from real double-blind conditions. Quite often in psychotherapy research instead of randomised parallel group conditions, the waiting group approach was used as a surrogate of the parallel group design, which probably increased effect sizes in a significant way [6]. This argumentation was based on single studies and the related evidence grading but of course it is similar for the methodological evidence grading. The different methodological basis for evidence grading in psychotherapy and psychopharmacology implies that such a direct equivalence is impossible.

Insofar, the recently published German national depression guideline is totally misleading when it states that the level of evidence is equal to psychotherapy as well as to treatment with antidepressants [28]. To avoid such confusion, it would be necessary to develop a uniform evidence grading system for all therapy procedures in psychiatry, taking all design methodologies into account. Psychotherapeutic procedures per se could not achieve the highest level of evidence in such an evidence system because of the specially restrictive conditions of the evaluation methods used, in which for example placebo controls are difficult to realise and double-blind conditions are impossible.

This applies even more to psychosocial procedures which, because of immanent characteristics, usually cannot even fulfil the requirements of randomised, blinded, control-group studies but apply methodologically less restrictive evaluation procedures. In particular in the area of psychiatric care research, there is only limited evidence even below this threshold [48]. It ranges from missing studies on supported housing [21], three studies on day centres [20], nine studies on acute psychiatric day clinic treatment [64], ten studies on intensive (or clinical) case management [65] to 18 studies on supported employment [25] and 20 studies on assertive community treatment [66].

Different evidence grading systems in different fields of psychiatry should be avoided. Different kinds of 'truths' are confusing for doctors, patients and health care providers, health care payers, etc. To obtain this should be seen as a principal rule of empirical research and beyond this as a principal rule of fairness and transparency for all parties concerned in the system.

Conclusions

Based on a review of the definition of the terms 'evidence' and 'truth' in the history of philosophy and the current philosophy of science leads the reconstruction is made that apparently the definition of 'evidence' in EBM is much more closer to the meaning of 'truth' in modern philosophy of science than to the philosophical term evidence. However, it was apparently avoided to use this more adequate term in the context of EBM, because it was probably seen as too demanding and prestigious and probably also too much loaded by philosophical expectations. Given the importance and relevant consequences the statement of 'evidence' and 'evidence grading' has in EBM, it would have been adequate to use the prestigious term 'truth' to underline, that we are talking about a high ideal.

The use of the term 'evidence' and especially the definition and grading of evidence in EBM is currently far away from this ideal. There are huge inconsistencies in defining evidence and in grading evidence. The monistic approach of EBM to assess evidence in a specific field

based primarily on meta-analytic condensation of empirical results leads to a one-sided view on methodological level and does not take into account the principal pitfalls and the different results of meta-analyses on the same issue. Also the priorisation of RCT results leads to a narrowed perspective which does not take into account other approaches to assess 'reality' in a complimentary way. Several examples of this primarily taken from the field of pharmacopsychiatry were presented to avoid only abstract discussion of the topic. Although insofar this paper is primarily focussing on psychiatry, the general problems of EBM described here are also relevant to other fields of medicine.

In short, to return to philosophical diction, in EBM the concept of evidence is currently a long way from being equivalent to a concept of truth. This critical aspect should be given greater consideration because of the central role it plays in EBM, among other things in the distribution of resources in health care systems, and because the supposed rationality thus achieved can turn into rationing. This should be also clearly described in all EBM-related publications, especially in guidelines and textbooks communicating EBM gradings, to avoid misunderstandings especially by people from outside the field like among others health care providers and health care politicians.

Finally, it should be demanded that all data from clinical trials need, for a detailed understanding and interpretation, the knowledge of clinically experienced physicians, in the field covered in this paper, for example clinical psychiatrists and clinical psychopharmacologists. They are not abstract figures that can be aggregated and interpreted solely by statisticians or health economists and their co-workers. Therefore, many publications on meta-analyses raise critical questions, although they might be sound from the statistical methodology. In this context, the famous sentence of one of the founders of EBM should be remembered every day: 'Evidence based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research' [88].

Conflict of interest H.-J. Moeller has received grants or is a consultant for and on the speakership bureaus of AstraZeneca, Bristol-Myers Squibb, Eisai, Eli Lilly, GlaxoSmithKline, Janssen Cilag, Lundbeck, Merck, Merz, Novartis, Organon, Pfizer, Sanofi-Aventis, Schering-Plough, Schwabe, Sepracor, Servier and Wyeth.

References

1. American Psychiatric Association APA Guideline Development Process (2006) http://www.psych.org/psych_pract/treatg/pg/prac_guide.cfm.Anonymous
2. Antes G (2004) The evidence base of clinical practice guidelines, health technology assessments and patient information as a basis for clinical decision-making. Z Ärztl Fortbild Qualitätssich 98:180–184
3. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, Guyatt GH, Harbour RT, Haugh MC, Henry D et al (2004) Grading quality of evidence and strength of recommendations. BMJ 19:1490
4. Baldwin D, Broich K, Fritze J, Kasper S, Westenberg H, Möller HJ (2003) Placebo-controlled studies in depression: necessary, ethical and feasible. Eur Arch Psychiatry Clin Neurosci 253: 22–28
5. Bandelow B, Zohar J, Hollander E, Kasper S, Moller HJ, Zohar J, Hollander E, Kasper S, Moller HJ, Bandelow B et al (2008) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the pharmacological treatment of anxiety, obsessive-compulsive and post-traumatic stress disorders—first revision. World J Biol Psychiatry 9:248–312
6. Bandelow B, Zohar J, Kasper S, Möller HJ (2008) How to grade categories of evidence. World J Biol Psychiatry 9:242–247
7. Bateman A, Fonagy P (1999) Effectiveness of partial hospitalization in the treatment of borderline personality disorder: a randomized controlled trial. Am J Psychiatry 156:1563–1569
8. Bauer M, Whybrow PC, Angst J (2002) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, Part 2: Maintenance treatment of major depressive disorder and treatment of chronic depressive disorders and subthreshold depressions. World J Biol Psychiatry 3:69–86
9. Bauer M, Bschor T, Pfennig A, Whybrow PC, Angst J, Versiani M, Möller HJ, WFSBP Task Force on Unipolar Depressive Disorders (2007) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders in primary care. World J Biol Psychiatry 8:67–104
10. Bauer M, Tharmanathan P, Volz HP, Möller HJ, Freemantle N (2009) The effect of venlafaxine compared with other antidepressants and placebo in the treatment of major depression: a meta-analysis. Eur Arch Psychiatry Clin Neurosci 259(3):172–185
11. Bauer M, Whybrow PC, Angst J, Versiani M, Möller HJ, WFSBP Task Force on Treatment Guidelines for Unipolar Depressive Disorders (2002) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of unipolar depressive disorders, Part 1: Acute and continuation treatment of major depressive disorder. World J Biol Psychiatry 3:5–43
12. Beynon S, Soares-Weiser K, Woolacott N, Duffy S, Geddes JR (2008) Pharmacological interventions for the prevention of relapse in bipolar disorder: a systematic review of controlled trials. Br J Psychiatry 192(1):5–11
13. Bottlender R, Rudolf D, Strauss A, Moller HJ (1998) Antidepressant-associated maniform states in acute treatment of patients with bipolar-I depression. Eur Psychiatry 248:296–300
14. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P (2000) Framework for design and evaluation of complex interventions to improve health. BMJ 321:694–696
15. Carnap R (1956) The methodological character of theoretical concepts. In: Feigl HSM (ed) Minnesota studies on the philosophy of science. 1. Minneapolis, pp 38–76
16. Carnap R (1966) Philosophical foundations on physics. Gardner M, New York
17. Carpenter WT (2001) Evidence-based treatment for first-episode Schizophrenia? Am J Psychiatry 158:1771–1773
18. Cartabellotta A, Minella C, Bevilacqua L, Caltagirone P (1998) Evidence-based medicine. 3. Systematic reviews: a tool for

clinical practice, permanent education and health policy decisions. Italian Group on Evidence-Based Medicine-GIMBE. Recenti Prog Med 89:329–337
19. Cartabellotta A, Montalto G, Notarbartolo A (1998) Evidence-based medicine. How to use biomedical literature to solve clinical problems. Italian Group on Evidence-Based Medicine-GIMBE. Minerva Med 89:105–115
20. Catty J, Burns T, Comas A (2001) Day centres for severe mental illness. Cochrane Database Syst Rev (2):CD001710
21. Chilvers R, Macdonald GM, Hayes AA (2002) Supported housing for people with severe mental disorders. Cochrane Database Syst Rev (2):CD000453
22. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JPT, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H et al (2009) Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 373:746–758
23. Clark W, Mucklow J (1998) Gathering and weighing the evidence. In: Panton R, Chapman S (eds) Medicines management. BMJ Books and Pharmaceutical Press, London, pp 59–74
24. Craig JC, Irwig LM, Stockler MR (2001) Evidence-based medicine: useful tools for decision making. MJA 174:248–253
25. Crowther R, Marshall M, Bond G, Huxley P (2003) Vocational rehabilitation for people with severe mental illness (Cochrane Review) [computer program]. The Cochrane Library, Issue 3. Update Software, Oxford
26. Czekalla J (2006) Kritische Bewertung von Studien und Metaanalysen. Ein Fortbildungsartikel über die wichtigsten Validitätskriterien der Evidence-based Medicine. Psychopharmakotherapie 13:224–230
27. Davis JM, Chen N, Glick ID (2003) A meta-analysis of the efficacy of second-generation antipsychotics. Arch Gen Psychiatry 60:553–564
28. Deutsche Gesellschaft für Psychiatrie PuN DGPPN, BÄK, KBV, AWMF, AkdÄ B, BApK, DAGSHG, DEGAM, DGPM, DGPs et al (eds) (2009) S3-Leitlinien/Nationale VersorgungsLeitlinie Unipolare Depression. Berlin, Düsseldorf. Deutsche Gesellschaft für Psychiatrie, Psychotherapie und Nervenheilkunde
29. Eccles M, Grimshaw J, Campbell M, Ramsay C (2003) Research designs for studies evaluating the effectiveness of change and improvement strategies. Qual Saf Health Care 12:47–52
30. Falkai P, Wobrock T, Lieberman J, Glenthoj B, Gattaz WF, Moller HJ (2006) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, part 2: long-term treatment of schizophrenia. World J Biol Psychiatry 7:5–40
31. Falkai P, Wobrock T, Lieberman J, Glenthoj B, Gattaz WF, Moller HJ, WFSBP Task Force on Treatment Guidelines for Schizophrenia (2005) World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, Part 1: acute treatment of schizophrenia. World J Biol Psychiatry 6:132–191
32. Fountoulakis KN, Möller HJ (2011) Efficacy of antidepressants: a re-analysis and re-interpretation of the Kirsch data. Int J Neuropsychopharmacol 14:405–412
33. Fritze J, Möller HJ (2001) Design of clinical trials of antidepressants. Should a placebo control arm be included? CNS Drugs 15:755–764
34. Geddes JR, Calabrese JR, Goodwin GM (2009) Lamotrigine for treatment of bipolar depression: independent meta-analysis and meta-regression of individual patient data from five randomised trials. Br J Psychiatry 194:4–9
35. Geddes JR, Carney SM, Davies C, Furukawa TA, Kupfer DJ, Frank E, Goodwin GM (2003) Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. Lancet 361:653–661
36. Gerson S, Belin TR, Kaufman A, Mintz J, Jarvik L (1999) Pharmacological and psychological treatments for depressed older patients: a meta-analysis and overview of recent findings. Harv Rev Psychiatry 7:1–28
37. Gonzalez DD (2001) From evidence-based medicine to medicine-based evidence. An Esp Pediatr 55:429–439
38. Grunze H, Kasper S, Goodwin G, Bowden C, Baldwin D, Licht R, Vieta E, Möller HJ (2002) World Federation of Societies of Biological Psychiary (WFSBP) guidelines for biological treatment of bipolar disorders, Part I: Treatment of bipolar depression. World J Biol Psychiatry 3:115–124
39. Grunze H, Kasper S, Goodwin G, Bowden C, Baldwin D, Licht RW, Vieta E, Moller HJ (2003) The World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the biological treatment of bipolar disorders, Part II: Treatment of mania. World J Biol Psychiatry 4:5–13
40. Grunze H, Kasper S, Goodwin G, Bowden C, Moller HJ (2004) The World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the biological treatment of bipolar disorders, Part III: Maintenance treatment. World J Biol Psychiatry 5:120–135
41. Grunze H, Vieta E, Guy M, Goodwin GM, Bowden C, Licht RW, Möller HJ, Kasper S, WFSBP Task Force on Treatment Guidelines for Bipolar Disorders (2009) The World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for the biological treatment of bipolar disorders: update 2009 on the Treatment of Acute Mania. World J Biol Psychiatry 10(2): 85–116
42. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schunemann H (2006) Grading strength of recommendations and quality of evidence in clinical guidelines: report from an American College of Chest Physicians task force. Chest 129:174–181
43. Hegerl U, Hautzinger M, Mergl R, Kohnen R, Schutze M, Scheunemann W, Allgaier AK, Coyne J, Henkel V (2010) Effects of pharmacotherapy and psychotherapy in depressed primary-care patients: a randomized, controlled trial including a patients' choice arm. Int J Neuropsychopharmacol 13:31–44
44. Hegerl U, Plattner A, Moller HJ (2004) Should combined pharmaco- and psychotherapy be offered to depressed patients? A qualitative review of randomized clinical trials from the 1990s. Eur Arch Psychiatry Clin Neurosci 254:99–107
45. Horder J, Matthews P, Waldmann R (2010) Placebo, Prozac and PLoS: significant lessons for psychopharmacology. J Psychopharmacol [Epub ahead of print]
46. Huf W, Kalcher K, Pail G, Friedrich ME, Filzmoser P, Kasper S (2011) Meta-analysis: fact or fiction? How to interpret meta-analyses. World J Biol Psychiatry 12:188–200
47. Jadad AR, Phil D, Cook DJ (1998) Methodology and reports of systematic reviews and meta-analyses-a comparison of cochrane reviews with articles published in paper-based Journals. JAMA 280:278–280
48. Kallert TW (2005) Is mental health services research in need of randomised controlled trials? Psychiatr Prax 32:375–377
49. Kawamura T, Tamakoshi A, Wakai K, Ohno Y (1999) Evidence-based medicine and 'The Cochrane Collaboration'. Nippon Koshu Eisei Zasshi 46:498–506
50. Khan KS, Kunz R, Kleijnen J, Antes G (2004) Systematische Übersichten und Meta-Analysen. Ein Handbuch für Ärzte in Klinik und Praxis sowie Experten im Gesundheitswesen. Springer, Berlin
51. Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT (2008) Initial severity and antidepressant benefits: a meta-analysis of data submitted to the food and drug administration. PLoS Med 5:e45

52. Klein DF (2000) Flawed meta-analyses comparing psychotherapy with pharmacotherapy. Am J Psychiatry 157:1204–1211
53. Koller M, Lorenz W, Abel U (2006) Methodenvielfalt in der klinischen Forschung. MMW-Fortschritte der Medizin 148: 85–91
54. Leichsenring F, Rabung S, Leibing E (2004) The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: a meta-analysis. Arch Gen Psychiatry 61:1208–1216
55. Leucht S, Arbter D, Engel RR, Kissling W, Davis JM (2009) How effective are second-generation antipsychotic drugs? A meta-analysis of placebo-controlled trials. Mol Psychiatry 14(4):429–447
56. Lieberman JA, Greenhouse J, Hamer RM, Krishnan KR, Nemeroff CB, Sheehan DV, Thase ME, Keller MB (2005) Comparing the effects of antidepressants: consensus guidelines for evaluating quantitative reviews of antidepressant efficacy. Neuropsychopharmacology 30:445–460
57. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RS, Davis SM, Davis CE, Lebowitz BD et al (2005) Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 353:1209–1223
58. Linden M (2002) Verhaltenstherapie: theoretische und empirische Grundlagen sowie klinische Anwendungsprinzipien. In: Möller HJ, Laux G, Kapfhammer HP (eds) Psychiatrie and psychotherapie. Springer, Berlin, pp 681–710
59. Linehan MM, Comtois KA, Murray AM, Brown MZ, Gallop RJ, Heard HL, Korslund KE, Tutek DA, Reynolds SK, Lindenboim N (2006) Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. Arch Gen Psychiatry 63:757–766
60. Lohr KN (2004) Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care 16:9–18
61. Maier W, Möller HJ (2005) Metaanalyses—highest level of empirical evidence? Eur Arch Psychiatry Clin Neurosci 255: 369–370
62. Maier W, Möller HJ (2010) Meta-analyses: a method to maximise the evidence from clinical studies? Eur Arch Psychiatry Clin Neurosci 260:17–23
63. Manser R, Walters EH (2001) What is evidence-based medicine and the role of the systematic review: the revolution coming your way. Monaldi Arch Chest Dis 56:33–38
64. Marshall M, Crowther R, Almaraz-Serrano A, Creed F, Sledge W, Kluiter H, Roberts C, Hill E, Wiersma D (2003) Day hospital versus admission for acute psychiatric disorders. Cochrane Database Syst Rev (1):CD004026
65. Marshall M, Gray A, Lockwood A, Green R (2003) Case management for people with severe mental disorders (Cochrane Review). The Cochrane Library, Issue 3. Update Software, Oxford
66. Marshall M, Lockwood A (2003) Assertive communication treatment for people with severe mental disorder (Cochrane Review). The Cochrane Library, Issue 3. Update Software, Oxford
67. Mittelstraß JH (2004) Evidenz. In: Mittelstraß J (ed) Enzyklopädie Philosophie und Wissenschaftstheorie. Sonderausgabe ed. J. B.Metzler, Stuttgart, pp 609–610
68. Mittelstraß JH (2004) Wahrheit. In: Mittelstraß J (ed) Enzyklopädie Philosophie und Wissenschaftstheorie. Sonderausgabe ed. J. B.Metzler, Stuttgart, pp 582–587
69. Mittelstraß JH (2004) Wahrheitsähnlichkeit. In: Mittelstraß J (ed) Enzyklopädie philosophie und wissenschaftstheorie. Sonderausgabe ed. J. B.Metzler, Stuttgart, pp 588–589
70. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. Lancet 354:1896–1900
71. Möller HJ (1976) Methodische Grundprobleme der Psychiatrie. Kohlhammer, Stuttgart
72. Möller HJ (1978) Psychoanalyse. Erklärende Wissenschaft oder Deutungskunst? Zur Grundlagendiskussion in der Psychowissenschaft. Fink, München
73. Möller HJ (2001) Methodological issues in psychiatry: psychiatry as an empirical science. World J Biol Psychiatry 2:38–47
74. Möller HJ (2006) Methodik empirischer Forschung und evidenzbasierter Medizin in der Psychiatrie. In: Möller HJ, Laux G, Kapfhammer HP (eds) Psychiatrie und psychotherapie. 3 Aufl. Springer, Berlin
75. Möller HJ (2008) Do effectiveness ("real world") studies on antipsychotics tell us the real truth? Eur Arch Psychiatry Clin Neurosci 258:257–270
76. Möller HJ (2008) Isn't the efficacy of antidepressants clinically relevant? A critical comment on the results of the metaanalysis by Kirsch et al. 2008. Eur Arch Psychiatry Clin Neurosci 258:451–455
77. Möller HJ (2009) Is evidence sufficient for evidence-based medicine? Eur Arch Psychiatry Clin Neurosci 259(Suppl 2): S167–S172
78. Möller HJ, Broich K (2010) Principle standards and problems regarding proof of efficacy in clinical psychopharmacology. Eur Arch Psychiatry Clin Neurosci 260:3–16
79. Möller HJ, Fuger J, Kasper S (1993) Statistische metaanalyse der Wirksamkeit neuerer Antidepressiva. Antidepressiva und Phasenprophylaktika. Riederer P, Laux G, and Pöldinger W. Neuro-Psychopharmaka.Wien, Springer, pp 252–256
80. Möller HJ, Maier W (2010) Evidence-based medicine in psychotherapy: possibilities, problems and limitations. Eur Arch Psychiatry Clin Neurosci 260:25–39
81. Montgomery SA, Möller HJ (2009) Is the significant superiority of escitalopram compared with other antidepressants clinicly relevant? Int Clin Psychopharmacol 24:111–118
82. Neumann PJ, Tunis SR (2010) Medicare and medical technology–the growing demand for relevant outcomes. N Engl J Med 362:377–379
83. Pilling S, Price K (2006) Developing and implementing clinical guidelines: lessons from the NICE schizophrenia guideline. Epidemiol Psichiatr Soc 15:109–116
84. Popper K (1963) Conjectures and refutations. The growth of scientific knowledge. Routledge & Kegan Paul, London
85. Riedel M, Strassnig N, Müller N, Zwack P, Möller HJ (2005) How representative of everyday clinical populations are schizophrenia patients enrolled in clinical trials? Eur Arch Psychiatry Clin Neurosci 255:143–148
86. Sackett DL (2000) Evidence-Based Medicine: How to practice and teach EBM. Churchill Livingstone, New York
87. Sackett DL, Richardson S, Rosenberg WS, Haynes W, Dt. Ausg: Kunz R, Fritsche L (1999) Evidenzbasierte Medizin. Zuckschwerdt, München
88. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. BMJ 312:71–72
89. Saha S, Coffman DD, Smits AK (2010) Giving teeth to comparative-effectiveness research—the Oregon experience. N Engl J Med 362:e18
90. Schmacke N (2006) Evidenzbasierte Medizin und Psychotherapie: die Frage nach den angemessenen Erkenntnismethoden. Psychother Psychosom Med Psychol 56:202–209
91. Schöchlin C, Möller HJ, Engel RR (2002) Meta-Analysen von Antidepressiva-Studien. In: Riederer P, Laux G, Pöldinger W (eds) Neuro-psychopharmaka, Bd. 3, 2. Aufl. Springer, Wien New York, pp 349–363

92. Seemüller F, Möller HJ, Obermeier M, Adli M, Bauer M, Kronmüller K, Holsboer F, Brieger P, Laux G, Bender W et al (2010) Do efficacy and effectiveness samples differ in antidepressant treatment outcome? An analysis of eligibility criteria in randomized controlled trials. J Clin Psychiatry 71:1426–1433
93. Stegmüller W (1954) Metaphysik, Wissenschaft, Skepsis. Frankfurt, Wien
94. Tarski A (1935) Der Wahrheitsbegriff in den formalisierten Sprachen. Stud Philos 1:262–405
95. Wampold BE, Minami T, Baskin TW, Callen TS (2002) A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. J Affect Disord 68:159–165