Eur Arch Psychiatry Clin Neurosci (2010) 260:3–16
DOI 10.1007/s00406-009-0071-8

SPECIAL ISSUE

Principle standards and problems regarding proof of efficacy in clinical psychopharmacology

Hans-Jürgen Möller · Karl Broich

Published online: 4 November 2009

Abstract Proof of efficacy of a psychotropic medicinal product is the key point of clinical psychopharmacology. This especially concerns the licensing of a new compound, but apart from this special case, lots of efficacy questions need to be answered in clinical psychopharmacology, such as, e.g. the question of the efficacy of a combination therapy. The methodology of the scientific proof of efficacy has already had a long tradition and has been developed further in the recent past under different aspects. Especially the double-blind randomised parallel group comparison has been developed as a design of highest methodological standard. However, often designs have their place and justification under certain conditions and in relation to certain questions. Although in the recent past, with the over-emphasis of so-called effectiveness studies, the inherent methodological limitations of these studies have not been addressed properly (Möller in Eur Arch Psychiatry Clin Neurosci 258:257–270, 2008), which in consequence devaluated the scientific merits of the classical double-blind randomised control group study designs in the view of those colleagues, who are not that experienced in study design issues. Therefore, it seems to be timely and necessary to review the principle standards and problems concerning the proof of efficacy in clinical psychopharmacology.

Keywords Proof of efficacy · Clinical psychopharmacology · Methodology

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

K. Broich
Drug Approval Department, BfArM Bundesinstitut für
Arzneimittel und Medizinprodukte, Kurt-Georg-Kiesinger-Allee
3D, 53175 Bonn, Germany

Basic principles of methodological approaches in clinical psychopharmacology to proving efficacy

The rich spectrum methods commonly used in the clinical evaluation of psychotropic medicinal products can be divided according to different aspects, such as

- time aspects: retrospective and prospective procedures
- experimental quality: non-experimental, quasi-experimental and experimental procedures.

The choice of procedure depends on the question being investigated and the way of how the data can be collected. There is no single, generally valid, ideal experimental design. At the most there are optimal/most reliable designs for answering certain questions and settings, whereby, in addition to the actual scientific question, constraints resulting from pragmatic, economic, ethical, and legal problems have to be considered.

To generate hypotheses, non-experimental retrospective or prospective studies are performed aimed at identifying relations that can later be tested in prospective studies designed experimentally [40, 42]. As a matter of principle, prospective and experimental studies have a higher scientific value than retrospective and non-experimental studies, respectively, since their results offer a higher guarantee of unbiased findings (internal validity). Because experimental studies are strongly reductionistic (e.g. they may exclude very severe/psychotic depressions, severe suicidal tendencies, interfering variables such as comorbidity, etc), the generalisability of their results to patients in routine clinical care is limited [50]; this is particularly true for placebo-controlled studies. For this reason, besides the rigorous experimental study design represented by pivotal phase III studies, it is important to perform additional studies with less restrictive methodology in order to obtain a complementary, better generalisable view (external validity).

These methodologically less restrictive studies are often performed as phase IV studies (see below), e.g. after a drug has been licensed, since this phase in particular aims for generalisability of the results. Most of these studies are naturalistic observational trials and a much smaller number are controlled studies with a randomised allocation of the patients to the treatment groups. The randomised, controlled studies in phase IV, however, are normally either not blinded or have different restrictions to their design compared to classical experimental approaches used in phase III studies (see below).

Over the past years, interest has grown in studies that are better aligned with the day-to-day care situation and that has resulted among other things in numerous, sometimes very large so-called 'effectiveness studies' (also referred to as 'real world studies', 'large simple trials', or 'pragmatic trials') in various indications. In recent years such studies have been performed particularly in the USA, mainly with massive financial support from governmental institutions. Although these studies provide interesting information on the circumstances in day-to-day treatment situations, because of their methodological shortcomings they cannot prove that the efficacy and/or tolerability results of the methodologically strict phase III studies are false, but can only give a complementary view [36].

The clinical proof of efficacy of medicinal products is conventionally divided into four phases [2, 59] (Fig. 1):

- The main aim of phase I is to test in healthy subjects the tolerability of a substance that has undergone pharmacological and animal experiments. Phases IIA and IIB evaluate the evidence for therapeutic efficacy in a smaller group of patients. Phase III aims to confirm the results of phase II in larger patient samples. If the results are positive and the risk–benefit ratio acceptable, these confirmatory studies result in licensing. After introduction to the market, effectiveness and tolerability of the drug are checked in phase IV, mainly in naturalistic observational trials. Besides simple naturalistic analyses, which focus on only one drug, these observational trials [27, 28] can simultaneously include other comparative drugs without compromising the naturalistic character of the study [15, 45]. Drug surveillance methods, which are mainly aimed at safety aspects, are also part of the evaluation approaches in phase IV [12–14]. Randomised controlled trials (RCTs) can also be a part of phase IV, as shown by the 'effectiveness studies', for example.

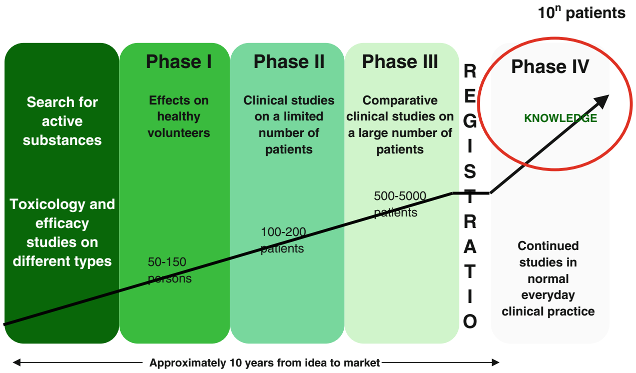
10ⁿ patients
Phase I
Phase II
Phase III
Phase IV
Search for active substances
Toxicology and efficacy studies on different types
Effects on healthy volunteers
Clinical studies on a limited number of patients
Comparative clinical studies on a large number of patients
REGISTRATIO
KNOWLEDGE
50-150 persons
100-200 patients
500-5000 patients
Continued studies in normal everyday clinical practice
Approximately 10 years from idea to market

Fig. 1 The four phases of clinical studies

Table 1 General guidance documents for clinical studies from the International Conference of Harmonisation, and for specific disorders from the European Medicines Agency (EMEA); the current versions can be accessed at the given websites

Guidance documents
ICH (http://www.ich.org/cache/compo/276-254-1.html)
Good clinical practice (E6 (R1))
Studies in support of special populations. Geriatrics (E7)
General consideration of clinical trials (E8)
Statistical principles for clinical trials (E9)
Choice of control group and related issues in clinical trials (E10)
Clinical investigation of medicinal products in the paediatric population (E11)
EMEA (http://www.emea.europa.eu)
Schizophrenia (CPMP/EWP/559/95 + Add.)
Bipolar disorder (CPMP/EWP/567/98)
Depression CPMP/EWP/518/97 Rev. 1)
Panic disorder (CHMP/EWP/4280/02)
Generalised anxiety disorder (CPMP/EWP/4284/02)
Obsessive compulsive disorder (CHMP/EWP/4279/02)
Social anxiety (CHMP/EWP/3635/03)
Post-traumatic stress disorder (CHMP/EWP/358650/06)
Alzheimer's disease (CPMP/EWP/553/95 Rev.1)
Insomnia (CHMP/EWP/310566/07)
ADHD (CHMP/EWP/431734/08)
Smoking and nicotine dependence (CHMP/EWP/369963/05)

Information about the current methodological standards for licensing studies of psychotropic medicinal products is given in the respective guidelines of the regulatory authorities (Table 1), e.g. the respective guidelines of the European regulatory authority, the European Medicines Agency (EMEA) [9] and the respective guidelines of the International Conference of Harmonisation (ICH) [17].

The double-blind randomised control group design as the best proof of efficacy

The double-blind, randomised, parallel group comparison (also often referred to as the double-blind randomised control group study) is the most important and best method to prove the efficacy of a pharmaceutical agent according to regulatory demands [33, 44]. In such studies, the efficacy of the test substance in patients of the experimental group is compared with the efficacy of a placebo or of a drug licensed for the same indication (standard drug), or with the efficacy of both, in patients of the control group/groups. The patients are allocated randomly either to the experimental or the control group. Both the general efficacy of a pharmaceutical agent and specific features such as dosage or method of administration (peroral, intramuscular, etc) can be evaluated with respect to efficacy and side effects. The required sample size is determined a priori by a statistical calculation that takes into account both the expected difference in treatment effect and the variance of the variables.

Numerous problems arise from different influencing variables (interference factors) that are included as a 'random sampling error' in the final result. The double-blind, randomised, control-group design distributes the sample-related influencing factors randomly between the two groups compared. Even so, especially in small samples some relevant influencing factors may be unevenly distributed, e.g. age, sex, diagnosis, duration of illness, severity of symptoms, and it may be necessary to check the relevance of these factors for the results in an ex post analysis. The smaller the respective differences between the groups at baseline, the more likely it is that a clear result will be obtained.

The results of a study can be influenced by the investigator. Besides intentional manipulation of results of observation or incorrect methods of statistical analysis—e.g. more favourable results of certain analyses are perhaps reported instead of unfavourable results of the statistical analyses specified a priori—one has to consider in particular unintentional, biased observations on the part of the investigator. Reasons for such a systematic falsification include the Rosenthal effect, the halo effect and logical errors. Also from the patients' side different biases of the results are possible, e.g. due to positive or negative expectations, which do in the end influence their behaviour and/or reporting. Thus, for good reasons the double-blind design, in which neither patient nor investigator is informed about the administered drug, is the most indicated approach to avoid these different biases. Even the statistical analysis should be carried out under these double-blind conditions. For these reasons regulatory authorities demand the double-blind design. However, beyond the process of approval, also the methodological less restrictive non-blinded randomised control group study, often referred to as RCT, is widely accepted. Unfortunately, in times of evidence-based medicine (EBM), the difference between blinded and non-blinded randomised control group study is not as much focused on as should be; e.g. several guidelines only demand the results from RCTs for attributing the highest level of evidence, without differentiating or even preferring double-blind studies [43]. This has of course to be regarded as a misunderstanding of the special values of blinded studies and the limitations of non-blinded studies.

Nosological diagnoses should be made on the basis of recognised operationalised diagnostic systems like ICD-10 or DSM-IV and, if possible, confirmed in a standardised semi-structured or fully structured interview [37, 53]. Treatment success should be evaluated with validated scales, such as outlined in the Collegium Internationale Psychiatriae Scalarum (CIPS) [39, 57]. The primary and secondary efficacy criteria must be defined a priori in both the study protocol and the biometrical analysis plans. Besides the evaluation of efficacy, the standardised assessment of side effects is of great importance.

Influencing factors and systematic falsification tendencies can be largely reduced by the use of a double-blind design and standardised evaluation procedures and the careful choice of comparator. However, even small 'peculiarities', e.g. with respect to dose or efficacy criteria, can determine whether Substance A is superior to Substance B or vice versa [16].

In addition to the double-blind or non-blinded randomised parallel group study, the group comparison can also be performed in a sequential way, blinded or non-blinded, following, e.g. an AB, ABA or ABAB design (where A = placebo or another active substance, for example, and B = study drug). This is a well-established procedure, especially for pilot studies in small samples or for studies in patients with rare diseases. This design type can even be performed on an intraindividual level as a single case study [30]. When this design is used, the possibility of carry-over effects must be considered. It needs to be emphasised that these ABAB designs and their variations are only of academic interest and are not accepted by regulatory authorities as proof of efficacy.

Depending on the question at hand (e.g. efficacy, comparison of efficacy, side effects), more economical and practical procedures can be used than the elaborate ones described above. This is particularly the case for exploratory studies of new psychotropic medicinal products, e.g. procedures without a control group or an intraindividual comparison, and single- or non-blinded procedures. Non-experimental studies in the form of retrospective or prospective follow-up observations, e.g. naturalistic phase IV studies, also fall into this category; these studies are useful as a heuristic method to investigate effects and side effects of marketed drugs in patients routinely treated with them.

Univariate and multivariate designs

The fact that the experimentally varied or manipulated independent variables are only a fraction of the total number of variables responsible for changes in the dependent variables makes clinical research with psychotropic medicinal products, and treatment research in psychiatric patients in general difficult. The effects of the remaining influencing variables (interference factors) are not controlled and are included as a ‘random sampling error’ in the final result. The size of this error can be analysed using the control-group procedure. Furthermore, statistical analyses can be used to try to determine the most important factors in the random sampling error; the effect of these factors can then be evaluated in new experiments. In clinical psychopharmaceutical research, the variable of primary interest—the efficacy of the study drug—is normally abstracted from the other influencing variables. In agreement with this approach, univariate experimental studies are preferred, in which the other influencing variables are not varied or manipulated. Correlative associations between certain other influencing variables and treatment results are normally only investigated in secondary analyses of the results of such univariate clinical-psychopharmacological studies.

If several treatment-relevant factors are known from the start, one can attempt to simultaneously estimate the effect of these different factors and the interactions between them by systematically varying several factors in an experiment (Fig. 2). Such a multivariate dependence analysis is much more informative than the univariate dependence analysis described above. However, it requires a significantly larger sample size, especially if one wants to include several factors relevant for the treatment of psychiatric disorders. For example, just 4 independent variables with 2 values or modalities each yield a total of 16 cells. If each cell is filled with only 5 patients, a total of 80 patients are required. In case of 10 patients per cell, already 160 patients are necessary and so forth. Thus, multifactorial approaches in clinical psychopharmaceutical research soon reach the limits of the available number of patients. If the influencing variables are limited to a few at the start of a multivariate study, the problem arises that although the factors considered to be relevant are distributed homogeneously in the individual cells, the factors considered not to be relevant are distributed inhomogeneously, which may have considerable effects on the result.

		Psychotropic drug I		Psychotropic drug II	
		male	female	male	female
Extroverts	Age 20				
	Age 40				
Introverts	Age 20				
	Age 40				

Fig. 2 Multivariate design with four independent variables, each with two severities or modalities: 2 × 2 × 2 × 2 design

Because of the above problems of multivariate dependence analyses, the model of univariate dependence analyses is consistently used in most clinical-psychopharmacological trials. The underlying hypothesis is that all other factors except for the drug variable can be neglected and their relevance assessed in further statistical analyses. The increasingly large placebo effects in clinical studies of antidepressants [18, 20, 54, 55], which are also becoming even more apparent in antipsychotic studies [24], with the result that placebo-verum differences are becoming smaller and smaller (Fig. 3), indicates that this strategy may not be successful in the long term.

Different statistical analyses are used, depending on the type of study. These range from simple comparisons of means through correlation statistics and multivariate analyses, such as analyses of variance and covariance, to statistical analyses of single-case studies. Every statistical test is based on prerequisites that first have to be fulfilled. If these are not fulfilled the test can normally no longer be used. In all statistical analyses based on comparisons of means, one has to consider that there is a considerable loss

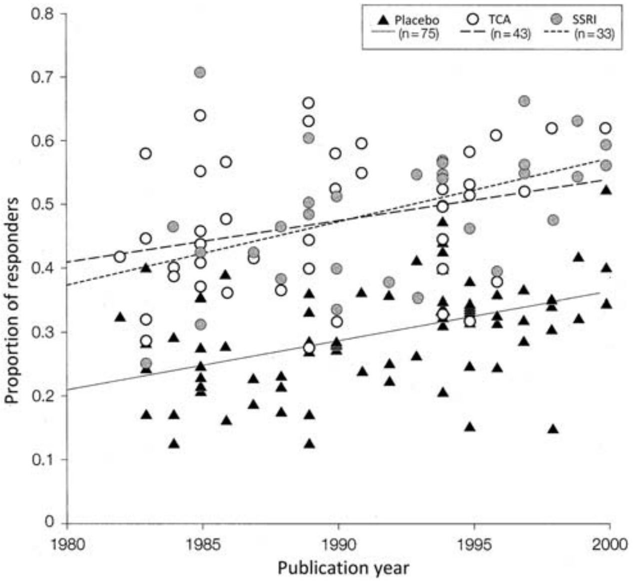

Fig. 3 Increase in the proportion of patients who showed at least 50% improvement of symptoms—measured with the Hamilton Depression Scale—in clinical studies with antidepressants and placebo; shown in relation to publication year modified from [58]

of information when the data are reduced to the mean. Additional statistical analyses should be used to try to compensate for such losses of information. Beside these dimensional analyses, categorical analyses of the frequency of 'responders', 'remitters', etc are performed. Especially the regulatory authorities demand them since they complement results of analyses of means with case-related statements and thus allow a better evaluation of the clinical relevance of differences in mean values.

If different research groups of similar experience achieve identical, statistically significant and clinically relevant results in several studies, the results can be considered to be definite. Problems occur when different studies with similar, acceptable standardised methods produce conflicting results and when the conflict cannot be explained or solved in further studies. If it is not possible to explain contradictory results as being a consequence of different interference factors or influencing variables, the limit of applied research methods has been reached.

Despite general agreement on the basic methodological principles of clinical-psychopharmacological research described here, there are divergences and a wealth of methodical problems in the realization of concrete research projects. This does not refer to the fact that pragmatism, economic constraints and requirements of practicability often force protocol designers to distance themselves from a methodology that is too puristic in face of the relevance of the question being investigated, and that can also suffer from organizational problems while being performed. It goes without saying that scientists should try to obtain the maximum amount of information with the minimum amount of effort. Rather, this refers to the practical problems of research methods that occur independently of economic requirements and pragmatic concessions.

Problems and necessity of placebo-controlled studies

Placebo-controlled studies have the highest informative value [4, 8, 11, 22], although they have been discussed critically [3, 46, 51]. Regulatory authorities consider them to be the clearest proof of efficacy of psychotropic medicinal products in indications like depression, anxiety disorders, schizophrenias and dementia, among others. Of course, the necessary ethical standards have to be adhered to, and only in this case can studies be considered to be ethical [1]. The ethical standards for placebo-controlled studies include special demands on patient selection (e.g. exclusion of severely suicidal patients); special rules for non-continuation of the trial in an individual patient (so-called 'stopping rules'); rules for offering 'rescue medication' in critical situations, etc. and established standard of overall care by a team with experience in conducting clinical trials.

The clear position of important regulatory authorities such as the American FDA and European EMEA, which demand placebo-controlled studies as the best proof of efficacy, is mainly based on the fact that in most psychiatric indications only placebo-controlled studies allow a sufficiently sure statement about the efficacy of a psychopharmaceutical compound to be made while exposing the smallest number of patients possible to the study drug; this approach is the only one that allows false conclusions to be largely avoided in the face of a high placebo response and a series of other special characteristics of clinical-psychopharmacological studies. The alternative of testing against a standard drug—suggested time and again by critics of placebo-controlled studies—is significantly more prone to errors and often results in an over-estimation of the efficacy of the study drug. According to the European drug authority, the best design is to test the study drug not only versus placebo but also versus a standard drug for the respective indication (3-arm design). Such an approach allows efficacy versus placebo to be clearly demonstrated and the side effect profile to be determined. At the same time, one obtains important information about how the efficacy and tolerability of the study drug compare with those of a standard medicinal product. Incidentally, placebo research, i.e. which patients respond to placebo under which conditions, is an interesting area of study [7, 23, 29, 47, 58].

In the following, we will discuss in detail the necessity and problems of placebo studies. As discussed above, the randomised, double-blind, parallel control group design with the smallest possible variance is the best approach to prove efficacy. Thus, it is mandatory from the viewpoint of drug authorities. The question arises which treatment should be given to the control group(s). Non-treatment ('waiting list') is not an option in studies of drugs because blinding would be impossible. The remaining options include a double-blind comparison of different dosages of the experimental drug, versus an active reference drug or versus placebo, i.e. versus preparations identical in every way except that they do not contain the active substance (identical not only in appearance but also with respect to the galenic excipients, weight, taste [even when bitten], etc). Since the dose-efficacy relationship of psychotropic medicinal products is often not sufficiently pronounced, the comparison of distinct dosages does often not lead to clear dosage recommendations. Therefore, the remaining options are comparison to placebo or comparison to an active reference drug (standard treatment).

As for the active comparative studies, there are two principal options: the superiority design or the non-superiority (equivalence) design. Testing for superiority over a standard drug licensed to treat the respective indication is considered to be a feasible alternative to proving efficacy in a placebo-controlled study. However, such a design can seldom be realised because new drugs do not usually have better efficacy, or it is difficult to prove that they do. Thus, judiciously no regulatory authority demands proof of superiority over available treatment options as a requirement for licensing but ‘only’ proof of efficacy per se. Although superior efficacy would of course be welcomed, such a demand would represent a significant obstacle for research and progress, with disadvantages for future patients. It would be impossible to develop drugs with only similar efficacy but better tolerability and safety. Compounds with new mechanisms of action but no increased efficacy could not be submitted for marketing authorisation and the chance would be lost to identify drugs for patient subgroups of certain disorders that require a more specific treatment.

The non-inferiority trial versus a standard treatment is considered to theoretically be a possible alternative to the placebo-controlled design. Such equivalence designs may require smaller sample sizes (depending on the differences considered to be clinically relevant) than tests for superiority over standard drugs; however, they require much larger sample sizes than placebo-controlled studies, which has ethical implications. Of course, this procedure assumes that the chosen reference drug really is effective (and is applied at an effective dose); in this case, effective implies that the reference drug has systematically proven itself to be significantly superior to placebo. However, this is by no means self-evident. For example, one can assume that one in three placebo-controlled studies of an antidepressant generally accepted to be efficacious will fail to show superiority over placebo [38].

Thus, even if the study drug appears to be effective in the equivalence/non-inferiority test, it cannot be ruled out that the study has failed, i.e. that in this study both drugs were (similarly) ineffective, e.g. due to lack of assay sensitivity of the trial. It must also be considered that in the equivalence test, as in every test of active substances, the investigator and patients know the treatment is an active substance; the subsequent expectations of a positive treatment result (‘bias’) can reduce the variances and possible group differences and thus favour the false presumption of either equal efficacy or even efficacy at all (despite actual inefficacy). Finally, the unintentional unblinding through observation of side effects can have similar effects. Thus, in principle, the placebo control is also necessary to prove efficacy in indications for which effective drugs are available, unless other reasons oppose it; in this case, after evaluation of the more important interest, a poorer quality of data has to be accepted.

The placebo control is also favoured from an ethical point of view because it allows the smallest sample sizes. As a positive side effect, the study results are obtained more quickly, which can lead—in the case of negative results—to the discontinuation of any studies being performed simultaneously and can thus help protect patients.

Nevertheless, the double-blind, randomised, parallel-group comparison with placebo is necessary but not sufficient to prove efficacy.

If proof of superiority over placebo is not obtained in a two-arm, randomised, parallel-group comparison, the study is negative in terms of proving efficacy. But a more cautious interpretation of such a study result would be that the study may have ‘failed’ due to insufficient control of sources of error, like lacking assay sensitivity, e.g. caused by drug unresponsive patients. If a standard drug is available for the indication at hand, this interpretation problem can a priori be avoided by comparing the drug with placebo and an active reference drug in a three-arm, double-blind, randomised, parallel-group study. This is the optimal study design because it combines the advantages and disadvantages of a placebo and an active control (see Table 2). If in such a study the study drug and active reference drug do not differ from placebo, the study has clearly ‘failed’ (‘non-conclusive study’, where clear conclusions regarding the experimental drug cannot be obtained due to apparent lack of assay sensitivity). If only the reference drug is superior to placebo, but not the study drug (experimental drug), the experimental drug is (probably) ineffective (‘negative study’ regarding the experimental drug). Last, only this design allows the absolute effect size to be estimated because the placebo arm acts as an anchorage. From an ethical point of view, such a three-arm design should therefore be favoured, since only this design allows the maximum amount of knowledge to be obtained. Part of the ethical justifiability of a clinical study is that the study should achieve the highest possible informational value. After all, this is the only possibility to make a valid evaluation of the efficacy in comparison to existing treatment options (relative effect size).

Various parties question the ethical justifiability of a comparison with placebo. This criticism can be summarised in the following arguments [11]:

- The patients assigned to placebo are deprived of an effective treatment, which not only impairs their quality of life unacceptably but also puts them at danger of suicidal behaviour.
- In view of the availability of effective treatments, the only interest in new drugs is whether they are more effective than the previous ones; if demands were made for the development of more effective drugs, the development of dispensable ‘me-too’ drugs would be slowed down and innovative advances thus stimulated.
- Placebo-controlled studies are not representative since the recruited collective has been selected through their

Table 2 Advantages and disadvantages of using an active control or placebo in clinical studies

	Advantages	Disadvantages
Placebo-controlled studies	Allow estimation of the assay sensitivity and thus internal validation of the study	Perhaps higher risk from 'non-treatment'
	Allow better evaluation of the clinical relevance	Perhaps more limited generalisability of the results to the general population
	Smaller sample size	
	Lower study costs	
Studies with an active control	Supply data on relative efficacy and tolerability	Risk of false positive studies because assay sensitivity is lacking
	At least theoretically no inactive treatment	Equivalence/non-inferiority not suitable as proof of efficacy
	Fewer drop-outs due to lack of efficacy	Active comparator may not be standard therapy
	May be more acceptable to an ethics commission	More drop-outs due to adverse events
		Tendency to minimise efficacy differences
		Larger sample sizes
		Higher study costs

willingness to participate in such studies. The transfer of the results to the clientele ultimately treated with the licensed drug under everyday conditions is therefore questionable ('efficacy' vs. 'effectiveness') [36]. The results of placebo-controlled studies are therefore not clinically relevant.

- Finally, even if the patient has given informed consent and the ability to do so has been objectified, it cannot be assumed that the consent to participate in a placebo-controlled study can actually be given rationally. The ethical questionability starts with imposing on the patient the decision whether to give consent.

In principle, such points of criticism are legitimate. However, if one considers the arguments given above for the scientific necessity of placebo controls, it becomes clear that the consequence can be just as little a categorical rejection as a categorical endorsement of placebo-controlled studies. Rather, each indication and each planned study requires an individual evaluation of the more important interest, as is explicitly demanded by the rules for conducting clinical studies and the Declaration of Helsinki ("Concern for the interests of the subject must always prevail over the interests of science and society"). It was explicitly stated by this European drug authority that placebo-controlled psychopharmacological studies do principally not offend against the EMEA/CPMP 2002 regulatory authority [9].

Between the outlined extremes, there are several alternatives to choose from. In the interest of guaranteeing effective treatment of patients, the general public is entitled to a proof of efficacy with a method in line with modern science. As a matter of principle, proof of efficacy therefore requires placebo-controlled studies, whereby three- or multiple-arm, placebo- and reference drug-controlled studies represent the optimum in the interest of clear and interpretable study results. This basic principle ceases where the participating patient is at risk of being harmed unacceptably. (Essential) altruism has limits. The risk of harm can be at least reduced by suitable modifications to the study design without questioning the principle of a double-blind, randomised, parallel-group comparison with placebo. For example, based on their metaanalysis of placebo-controlled studies of antidepressants, Kirsch et al. [19] even asked whether antidepressants are superior to placebo in a clinically relevant matter.

Placebo treatment is not completely ineffective: there are large placebo effects in psychopharmacotherapy in particular. Amongst others it is rather the case that simply the expectation of a positive treatment result increases the probability of one. This expectation is probably the decisive mechanism of the placebo effect. In addition, in the framework of controlled studies there are effects of the increased level of care given to patients, among other things. The ethical implications of the fact that placebo is also an effective treatment ('placebo' means 'I shall please') are given little consideration in the discussion of the ethical justifiability of placebo-controlled studies. The patient is not actually being deprived of treatment but (if the study drug is effective) of part of the maximum achievable effect. Placebo treatment has advantages with respect to tolerability and safety. Thus, principally the same is asked of patients treated with placebo as of those treated with study drug. If one supports the development of new drugs, therefore, one inevitably expects altruistic behaviour from patients for the clinical study phase. However, as a matter of principle this altruism is only temporary. Through his or her altruistic behaviour the patient confirms the overall cross-society consensus that there is a need for social and at the same time altruistic behaviour. The altruism demonstrated through participation in placebo-controlled studies is part of and an expression of the principle of social equalization and the solidarity of the general population.

The demands to test only for superior efficacy in treatable disorders, which would make a comparison with placebo dispensable, are opposed by the legitimate ethical demands to choose a study methodology that allows the required sample size to be minimised. In addition, these demands would significantly delay progress in improving the tolerability of drugs and developing drugs with new modes of action (see above), among other things, or make such progress impossible. Only brief mention will be made here of the legal reasons why ‘only’ proof of efficacy and not of superior efficacy is required for licensing of medicinal products.

It is true that the patients participating in placebo-controlled studies represent a selected group (‘selection bias’), so that the question of the generalisability of results of such studies is important. Such studies actually include only 5–10% of the screened patients who have a general indication for treatment. The exclusion of suicidal patients is of particular concern in psychopharmacological studies. But until now there is no positive empirical proof for a lack of generalisability of the results [52]. This is also supported by internal evaluations performed by the regulatory authorities. By way of precaution, however, it should be required that the reasons for exclusion from the study, and all data of scientific relevance for the study, are documented for all patients who were suitable for recruitment but were not included, for whatever reason; the same data should also be recorded for included patients in order to allow their representability to be checked. Moreover, placebo-controlled efficacy trials alone are insufficient anyway: additional studies are required, e.g. versus standard drugs, and sometimes less strict methodological demands should be placed on these studies.

A patient’s decision to participate in a placebo-controlled clinical study, even for altruistic reasons, requires them to be capable of giving consent and to receive detailed information about the study. Questioning the autonomy and rationality of this decision with the argument that altruistic decisions cannot be rational per se contradicts the principle of social equalization that is characteristic of a democratic welfare state. The altruistic participation in a clinical study is an example and expression of this principle and therefore it is rational. It is exactly in the interests of protecting personal autonomy to leave the decision to the patient. In this respect the approach in clinical studies is not different to the general approach: in the end every medical decision rests with the patient. This freedom (autonomy) is actually a burden, but this burden is relativised by the fact that the patient can revise his or her decision at any time without having to give a reason or fear any negative consequences.

General requirements and interference factors in clinical studies

The administration of placebo in the control group or in the control phase of intraindividual comparisons and the subsequent ignorance of the patient and investigator about the type of medication (blind conditions or double-blind conditions) are supposed to rule out expectations of the patient or of the patient and investigator and associated auto- or heterosuggestion effects, all of which could falsify the study results. These methodological objectives are generally accepted but their realisability has been questioned [5, 49]. Time and again, investigators or patients manage to differentiate the placebo from the verum through certain phenomena (external appearance of the drug, physicochemical properties, side effects, etc). In such cases, the test result can be decisively influenced by expectations and placebo effects, the consequences of which are very difficult to assess. The reaction to placebo has been found to be a very complex phenomenon that depends on numerous factors, e.g. dosage, appearance and taste of the drug, treatment duration, patient’s personality and the stress character of the situation. The situation is probably similar as far as the investigator’s expectations are concerned (Rosenthal phenomenon) but this phenomenon has not yet been investigated in such detail. To realistically assess the relevance of these factors it may be meaningful to record the patients’ and investigator’s expectations regularly during a study, whereby one must keep in mind that these may change over the course of treatment. Furthermore, if the investigator records such data it may tempt him to focus his attention particularly on such phenomena and effects, which in turn can reinforce this observational error. The investigator’s expectations, e.g. that there will be an improvement the longer the treatment continues, can be misdirected by making videos of the psychopathological findings and showing them in a different order. Some experts demand that placebos should ideally not only look like the verum but also have its side effects in order to make it difficult to identify verum and placebo on the basis of these effects [6]. This procedure is not usually adopted since it also has its disadvantages.

There are no binding rules as to how long a patient should be treated with a study drug in the framework of a clinical psychopharmaceutical study. The following are considered to be meaningful durations for trials of acute (short-term) treatment conditions and are recommended by regulatory authorities: 2–4 weeks for tranquilizers; 6–8 weeks for antidepressants; 6–8 weeks for antipsychotics; and 6–12 months for anti-dementia treatments. Depot preparations require longer study periods. Discrepant results between two different research groups can probably be explained in part by different study durations and methods of drug application. The demand for binding rules about study durations to be specified for an assortment of known psychopharmacological questions, in order to increase the comparability of different studies, has been unsuccessful; one reason is that the adequate study duration is strongly dependent on the pharmacokinetic characteristics of the substance being investigated. One disadvantage of too short study duration can be that specific clinical effects are not recognised because they only occur after a sometimes considerable latent period.

There are also no binding rules on the choice of dose. In the first open studies of a substance, the dose can be freely adjusted whilst taking into account the pharmacokinetic data found in animal experiments and the toxicological threshold values. Controlled studies then aim to prevent dosing differences by using a fixed dose design. However, such a design may favour one of the two preparations being compared since the chosen dose level can deviate by different amounts from the optimal dose of the two drugs. In addition, the determined result can only be applied to the dose level used. Flexible dosing has the advantage that a further-reaching generalisation of the results is made possible by closer similarity to the everyday therapeutic situation. However, the analysis is thereby made more complicated since it is often very difficult to relate a dose increase or reduction to an observed effect. As a compromise, the fix-flexible design is often preferred, which allows the dose to be adjusted within predefined limits and time points.

There are also no set rules about the duration of wash-out periods before treatment or between two treatment phases. It is known from the results of pharmacokinetic investigations that after treatment with antidepressants a placebo period of preferably 7 days, after neuroleptics one of even 30 days would be appropriate. However, considerably shorter wash-out periods, e.g. 3–7 days, are usual in clinical care since a longer treatment-free phase is not considered justifiable because of practical clinical requirements and ethical reasons and because too long wash-out or placebo phases may mean that the spontaneous course of the disorder interferes with the conduct of the study and its results. Also, placebo run-in phases with their high drop out risk should not be included in pivotal phase III licensing studies because they can further reduce the generalisability of the study results to the general population.

If a sedating or sleep-inducing co-medication is necessary, preferably only one drug should be allowed. Administration of the same co-medication is an absolute requirement, particularly if additional biological parameters are being measured and the results will later be compared with those of other studies. Overall, co-medication should be limited as far as possible in order to achieve the best possible differentiation between the two experimental groups and not endanger it through these and other interference factors.

The term *influencing variables* (*interference factors*) refers to variables that influence the evaluation of the effect of variables of action, which in psychopharmacological studies is the effect of the drug. Thus, to be able to clearly analyze the effect of the variables of action, strategies have to be chosen that allow measurable influencing variables to be evaluated separately and that distribute the non-directly measurable influencing variables evenly over the treatments being evaluated.

One can distinguish between four sources of interference factors: patient, therapist, treatment milieu and patient's private milieu [48]. The patient's influencing variables can be further subdivided into personality specific, disorder specific and socioeconomic. A subdivision of the influencing variables into given and changeable variables is of more practical relevance. Given influencing variables include age, sex, disorder, course type of the disorder (e.g. acute, chronic), severity of the disorder (e.g. mild, severe, previously treatment resistant), previous course (e.g. time of first manifestation, average duration of a manifestation), start of the current episode, previous treatment and number of previous treatment attempts. Changeable interference factors include variable duration of treatment, different dosing, co-medication, change of investigator, different measurement methods and measurement criteria.

Particularly noteworthy are influencing variables resulting from the possible systematic falsification of the observation on the part of the investigator:

a. Rosenthal effect: The result of an evaluation is influenced by the investigator's expectations.
b. Halo effect (Thorndike): The result of an evaluation is strongly influenced by knowledge about other characteristics or the overall impression of study subjects.
c. Logical error (Newcomb): The result of an evaluation is influenced by the investigator only including such observations of detail that appear appropriate within the framework of his or her predetermined theoretical and logical concept.

While at randomization the influencing variables should be evenly distributed between the groups being studied in controlled studies—provided that the sample-size is large enough—and therefore can be ignored when evaluating the specific efficacy of a pharmaceutical compound, this is different for the interpretation of the results of uncontrolled pilot studies. Here it is very important to account meticulously for the different influencing variables. For example, a negative result of antidepressant treatment in a sample of previously treatment-refractory depressive patients would be evaluated differently from such a result in a sample of unselected depressive patients.

Problems of sample composition

Statisticians want samples in confirmatory studies to be as large as necessary to allow clear statements to be made and to avoid the 'error of the small number' (underpowering, so-called beta-error problem). The larger the sample, the smaller the differences that can be recognised in statistical analysis of the data. General rules about the correct size of a study sample do not exist because the size depends on various factors. There are formulae, however, that allow sample size to be estimated on the basis of the variance of the primary efficacy parameter and the anticipated difference in the primary outcome parameter between the compared groups at the end of the treatment. The sample sizes thus calculated, often, especially if research resources are limited, exceed the number of available patients, especially, for example, if small differences between the groups are to be determined. This is explained in the following example. If the expected placebo-verum difference in responder rate is >20% (it is essential that 'response' is defined in the study protocol!), 90–110 patients per treatment arm are required, if the following is specified: $2\alpha = 0.05$; $\beta = 0.20$ (α-error $= 0.05$; β-error $= 0.20$). For continuously distributed random variables, the required sample size can be estimated from the relation 'relevant difference d/standard deviation s'. The following treatment group sizes result from $2\alpha = 0.05$ and $\beta = 0.20$:

$$d = \mathrm{s},\ d = 3\mathrm{s}/4,\ d = 2\mathrm{s}/3,\ d = \mathrm{s}/2,\ d = \mathrm{s}/3,\ d = \mathrm{s}/4$$
$$n = 17,\ n = 29,\ n = 37,\ n = 64,\ n = 143,\ n = 253$$

Multi-centre studies are normally required so that a sufficient number of patients can be recruited [10], which can raise new problems as regarding insufficient comparability of the samples as well as redundant interrater reliability. For example, there may be differences in symptom definitions, estimates of symptom severity, use of diagnostic terms, etc.

The more uniform the samples with regard to diagnosis, duration of illness, age, severity of symptoms, etc, the easier the analysis and the greater the probability of obtaining clear results. However, these advantages of a homogeneous sample are accompanied by a poor generalisability of the results to therapeutic practice because such a sample does not represent the basic population of patients being treated in routine practice (i.e. there is no external validity). For example, in studies of psychopharmaca the age limit is often set at 65 to avoid interference from psycho-organic symptoms. Completely unexpected effects may occur if these drugs are later used in clinical practice, e.g. to treat patients with comorbidities or older patients, without specifically having tested them before in this age group. This problem is taken into account especially in the effectiveness ('real world') studies performed as RCTs [25]. However, to ensure external validity these studies contravene the basic principles of internal validity of the design [34, 36]. In explorative studies a very heterogeneous sample can actually stimulate generation of hypotheses more intensively than one with a narrower range of characteristics (external versus internal validity).

Despite the introduction of internationally accepted, operationalised classification systems for mental disorders (ICD, DSM), a relevant inter- and intrasubject uncertainty factor remains in the diagnosis of psychiatric disorders. This is known from numerous reliability studies on psychiatric diagnostics. The use of operationalised diagnostic criteria and standardised procedures [53] allows the problem of diagnostic classification to be largely but not completely solved in many areas [35]. DSM-IV/DSM-IV-TR diagnoses in particular have gained international acceptance and are preferred by regulatory authorities. Nosologic diagnostics should not only be aimed at the indication under investigation but also cover comorbidity, which is an important influencing factor. In addition to the nosological/syndromatological diagnostics, assessment of the severity of the symptoms is necessary. The predetermined exclusion and inclusion criteria have to be respected to ensure the homogeneity of the study sample [39].

Despite the described improvements of psychiatric diagnostics, considerable inhomogeneities are still possible. It may be possible to reduce these by using case definitions that also cover other levels, such as biochemical factors, psychophysiological factors or personality factors. For example, patient groups with completely different biochemical reactivity may be hiding behind the clinical diagnosis of depression, and this could be responsible in part for differences in response to the distinct psychotropic medicinal products. The situation is similar for personality factors. Patients with the same diagnosis may have different personality characteristics, which may explain differences in response to psychotropic medicinal products and placebo and in the subjective assessment of the effects of psychotropic medicinal products and placebo [41].

Demands have been made that the medication should be the only difference between patient groups being compared in a univariate psychopharmacological study. Other possible influencing factors should be evenly distributed between the treatment groups (structural equality). Different procedures are applied to guarantee structural equality.

Randomization aims to assign patients to the experimental or control group strictly by chance (coin-tossing principle, random number tables, computerised procedures, etc.) and to thus achieve structural equality of the two groups. Each patient has exactly the same chance to be assigned to one or the other group. When assignment is by chance one can expect that the influencing variables other than those being investigated will not falsify the results since they will have similar effects in both groups. However, this is only true for large samples. If samples are small there is a danger that despite random assignment the two groups will differ with respect to various variables, such as psychopathological findings, psychosomatic factors, anamnestic characteristics, etc. This lack of balance has to be considered in the analysis in order to avoid results caused by this imbalance being wrongly attributed to the therapeutic intervention. In such cases, strict random assignment has reached the limits of its possibilities.

Stratification (layering) achieves a priori a balanced distribution of relevant influencing variables over both groups, also when sample sizes are small. Parallelization is used to put patients in whom certain variables are similar into different pairs or blocks so that the differences between the units of observation are small within a block but relatively large between the blocks. The patients of the two blocks are then randomly assigned to the experimental or control group. This procedure allows one to assume that the two groups are structurally equal as far as certain relevant variables are concerned. Although this procedure is feasible if there are two or three known, relevant influencing variables, it reaches its limits if a group has to be parallelised with respect to a large number of influencing variables. In such cases, complicated procedures can take things further, for example the 'minimalization method' proposed by Taves [56] in which randomization is performed on the basis of the difference from a pattern of all given influencing variables.

After completion of a study, homogeneity tests and sensitivity analyses can be used to evaluate whether the criterion of structural equality of both groups was fulfilled, whether there were any centre-dependent effects, etc. When such analyses are performed for small randomised samples, it often becomes apparent that the structural equality of several criteria is unsatisfactory. The syndromatic construction of disorders, different psychopathological characteristics, duration of illness, type of course, psychophysiological reactivity and biochemical parameters, etc, are most probably the reason why different results are often obtained in the different study centres of a multicentre study, even though test conditions are apparently the same and the same study drug is used. Since these days many multi-centre studies are performed in several countries or continents, it is standard practice to foresee the applicable analyses in the protocols of such studies.

The statistical analysis of clinical-psychopharmacological studies is performed according to modern statistical standards and is determined a priori in a biometric analysis plan. It must be confirmed that the preconditions for the test procedures are fulfilled. If the requirements for parametric procedures are not fulfilled, non-parametric methods have to be applied. The statistical analysis evaluates different samples:

- 'Intent-to-treat sample': All patients are analysed who were included in the study, evaluated at least once and received active medication. For statistical analysis, the last recorded value is carried forward to the later evaluations ('last observation carried forward' [LOCF] method).
- 'Observed-case sample'(often referred to as the 'efficacy'sample): All patients are analysed who were in the study during the periods being evaluated and who received medication during this period. The number of cases analysed with this method is smaller than the intent-to-treat sample.

The observed-case analysis (OC analysis) provides information about how good a response is principally possible if a patient has continually taken the medication. This analysis overestimates efficacy so that the regulatory authorities demand an intent-to-treat (ITT) analysis as the decisive analysis. However, the ITT analysis also has its singularities, for example because of the carrying over of the values of drop-out patients. For this reason, additional analysis methods, e.g. the mixed-effects models repeated measures (MMRM) method [21, 26], have been proposed in order to better balance the respective advantages and disadvantages of the two methods.

The ITT analysis is more discerning than the OC analysis and in the view of the regulatory authorities therefore represents the method of choice in confirmatory studies; if both analyses are available, however, the OC analysis provides important additional information. In psychiatric indications, the large number of study drop-outs often causes sizeable problems for the interpretation of study results. The statistical analysis plan must therefore describe how to deal with high drop-out rates and missing data. The LOCF method cannot always be considered a conservative analysis procedure in this context: in dementia disorders, for example, the symptoms of interest continually worsen.

One must differentiate between a priori specified analyses of efficacy and tolerability parameters and ex post analyses. The a priori specified analyses are of greater importance; as a matter of principle ex post analyses can only be seen as having supportive value and generating hypotheses, but never as being confirmatory.

Documentation and evaluation of success

Besides structural equality of the patient groups under investigation, equality of observations is essential, i.e. all patients should be observed and evaluated by the same investigators using the same procedure and at the same times. The psychopathological findings are the main criterion in the evaluation of psychopharmacological efficacy. Besides changes in the psychopathological findings, changes in the physical-neurological findings and of clinically or theoretically relevant biochemical parameters can be recorded, especially in order to detect side effects.

Because simple clinical evaluation of findings proved to be unreliable for psychopharmacological research, evaluation scales were developed to allow quantified documentation of findings [39]. These standardised evaluation procedures can only be used to their full advantage if investigators practice the use of the instruments in interrater reliability training. A collection of important scales for psychopharmacology was published by the CIPS [57]. The recording of findings can be supplemented with psychometric performance tests, e.g. performance tests in the areas of perception, learning, retentiveness and psychomotor skills.

Besides the evaluation of the psychopathological findings by a psychiatrist, other sources and levels of information can be included, e.g. observer evaluation of psychopathological abnormalities by nursing staff or relatives, and completion by patients of self-evaluation scales to assess their state of mental health. The inclusion of several levels and sources of information can widen the documentation base in terms of multi-level or multi-method diagnostics so that a complete picture of the changes occurring during treatment with psychopharmaca can be obtained and subtle differences of effect can be determined. Especially in the recent past the subjective dimension in terms of 'well being' or 'quality of life', judged by self-rating of the patients has become an important complementary outcome criterion.

The analysis relates all findings recorded during the study to the baseline findings. Because the correct recording of baseline values can be affected by certain factors (e.g. carry-over of drug effects, the fact that it is the investigator's first contact with the patient), multiple assessment of the baseline findings is desirable, e.g. by independent assessors not involved in the direct treatment, to obtain as reliable as possible baseline data.

If different measuring instruments are used to assess the same constructs, the correlation of the total score for different severities of mental disorders should remain constant. A lack of constancy of the correlation, e.g. before and after treatment, indicates that the scales are measuring different things at different severities of the condition. This phenomenon is known from analyses of the factor structure of scales that were mapped from test people under and not under the influence of psychopharmaca. Among other things, divergences can occur because a self-evaluation scale of depressivity mainly evaluates subjective psychological experience, but an observer-rated scale additionally assesses objectively observable abnormalities of behaviour. These discrepancies between observer ratings and self-ratings are understandable if one assumes that patients with less severe depression can give considerably more verbal information about their subjective experience of their condition than severely ill patients [31, 32, 39].

It is important that the main efficacy criterion or criteria for the confirmatory testing are specified a priori in the study protocol and biometric analysis plan. This is especially important in assessment approaches with multiple methods because chance significances may otherwise occur during multiple statistical testing, which may then be given as a proof of efficacy, if the statistical problems are not recognised. Any evaluations that are not related to the main efficacy criteria specified a priori must be considered to be merely descriptive; any significant results would therefore need to be confirmed in a study designed for this purpose. An alpha adjustment is required if several main efficacy criteria are used. At least one of the main efficacy criteria should come from the area of observer-rated psychopathology and be as closely related as possible to the area of indications of the substance being tested.

To improve the comparability of different clinical studies of the same substance or substances with the same therapeutic objectives, it would be important that the same measurement scale be used, at least for the main efficacy criterion. In this context—at least as far as European psychiatry is concerned—collections of scales are important in which authorised translations of relevant scales are presented in several European languages [37]. As far as comparability on an international level is concerned, the HAMD (and the successor of the MADRS) and BPRS (and the successor of the PANSS) have established themselves as the 'standard meters' for depression and schizophrenia, respectively [39], despite their weaknesses. These procedures should be supplemented with audiovisual recordings to train the investigators and documentation of the interrater reliability.

Thorough methods of assessment must also include the recording of adverse events. An unstructured recording of adverse events that is only based on spontaneous comments results in underreporting and, also for other methodological reasons such as poor reliability, does not correspond with modern methodological standards. A new drug safety system has been in use in Europe since 2004; it is based on uniform definitions of medical terms, e.g. according to MedDRA, and on compulsory reporting of adverse events in clinical studies (several respective ICH Guidelines on Clinical Safety—E1, E2A-F—are available; the current version can be found at: http://www.ich.org/cache/compo/276-254-1.html). When applying for licensing of a drug, a pharmaceutical company also has to submit a detailed description of the planned pharmacovigilance system and action plan for risk surveillance, based on the results of the studies performed with the drug.

References

1. Adam D, Kasper S, Möller HJ, Singer EA (2005) Placebo-controlled trials in major depression are necessary and ethically justifiable: how to improve the communication between researchers and ethical committees. Eur Arch Psychiatry Clin Neurosci 255:258–260
2. Angst J, Bech P, Boyer P et al (1989) Consensus conference on the methodology of clinical trials of antidepressants. Zürich, March; Report of the Consensus Committee. Pharmacopsychiatry 23:171–175
3. Aspinall RL, Goodman NW (1995) Denial of effective treatment and poor quality of clinical information in placebo controlled trials of ondansetron for postoperative nausea and vomiting: a review of published trials. BMJ 311:844–846
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. Beatty WW (1972) How blind is blind? A simple procedure for estimating observer naivete. Psychol Bull 78:70–71
6. Beckmann H, Schmauss M (1983) Clinical investigations into antidepressive mechanisms. I. Antihistaminic and cholinolytic effects: amitriptyline versus promethazine. Arch Psychiatr Nervenkr 233:59–70
7. Brown WA, Johnson MF, Chen MG (1992) Clinical features of depressed patients who do and do not improve with placebo. Psychiatry Res 41:203–214
8. Carpenter WT Jr, Schooler NR, Kane JM (1997) The rationale and ethics of medication-free research in schizophrenia. Arch Gen Psychiatry 54:401–407
9. EMEA Europena Medicines Agency (EMEA) (2009) http://emea.europa.eu.Anonymous
10. Fischer-Cornelsen K, Ferner U (1977) Evaluation of new drugs: clozapine as an example of an European multicenter study. Drugs under Res 1:404
11. Fritze J, Möller HJ (2001) Design of clinical trials of antidepressants: should a placebo control arm be included? CNS Drugs 15:755–764
12. Grohmann R, Engel RR, Geissler KH, Ruther E (2004) Psychotropic drug use in psychiatric inpatients: recent trends and changes over time-data from the AMSP study. Pharmacopsychiatry 37(Suppl 1):S27–S38
13. Grohmann R, Engel RR, Ruther E, Hippius H (2004) The AMSP drug safety program: methods and global results. Pharmacopsychiatry 37(Suppl 1):S4–S11
14. Grohmann R, Hippius H, Helmchen H, Ruther E, Schmidt LG (2004) The AMUP study for drug surveillance in psychiatry—a summary of inpatient data. Pharmacopsychiatry 37(Suppl 1):S16–S26
15. Haro JM, Novick D, Suarez D, Alonso J, Lepine JP, Ratcliffe M (2006) Remission and relapse in the outpatient care of schizophrenia: three-year results from the schizophrenia outpatient health outcomes study. J Clin Psychopharmacol 26:571–578
16. Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S (2006) Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. Am J Psychiatry 163:185–194
17. ICH International Conference of Harmonization (ICH) (2009) http://ich.org/cache/compo/276-254-1.html.Anonymous
18. Khan A, Warner HA, Brown WA (2000) Symptom reduction and suicide risk in patients treated with placebo in antidepressant clinical trials: an analysis of the Food and Drug Administration database. Arch Gen Psychiatry 57:311–317
19. 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
20. Kirsch I, Moore TJ, Scoboria A, Nicholls SS (2002) The emperor's new drugs: an analysis of antidepressant medication data submitted to the US Food and Drug Administration. Prevention and Treatment. Article 23. Posted 15 Jul 2002. http://www.journals.apa.org/prevention/volume5/pre0050023a.html.5.Anonymous
21. Lane P (2007) Handling drop-out in longitudinal clinical trials: a comparison of the LOCF and MMRM approaches. Pharm Stat 7(2):93–106
22. Laporte JR, Figueras A (1994) Placebo effects in psychiatry. Lancet 344:1206–1209
23. Lavin MR (1991) Placebo effects on mind and body. JAMA 265:1753–1754
24. 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
25. 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
26. Lieberman JA, Tollefson G, Tohen M, Green AI, Gur RE, Kahn R, McEvoy J, Perkins D, Sharma T, Zipursky R et al (2003) Comparative efficacy and safety of atypical and conventional antipsychotic drugs in first-episode psychosis: a randomized, double-blind trial of olanzapine versus haloperidol. Am J Psychiatry 160:1396–1404
27. Linden M (1997) Phase IV research and drug utilization observation studies. Pharmacopsychiatry 30:1–3
28. Linden M, Baier D, Beitinger H, Kohnen R, Osterheider M, Philipp M, Reimitz DE, Schaaf B, Weber HJ (1997) Guidelines for the implementation of drug utilization observation (DUO) studies in psychopharmacological therapy. The "Phase IV Research" Task-Force of the Association for Neuropsychopharmacology and Pharmacopsychiatry (AGNP). Pharmacopsychiatry 30:65–70
29. Miller FG, Rosenstein DL (2006) The nature and power of the placebo effect. J Clin Epidemiol 59:331–335
30. Möller H-J, Steinmeyer EM (1990) Mood curves of neurotic-depressive patients undergoing treatment with antidepressants: time-series analyses of experience with HTAKA model. Pharmacopsychiatry 23:215–221

31. Möller HJ (1991) Outcome criteria in antidepressant drug trials: self-rating versus observer-rating scales. Pharmacopsychiatry 24:71–75
32. Möller HJ (2000) Rating depressed patients: observer- vs self-assessment. Eur Psychiatry 15:160–172
33. Möller HJ (2001) Methodological issues in psychiatry: psychiatry as an empirical science. World J Biol Psychiatry 2:38–47
34. Möller HJ (2005) Are the new antipsychotics no better than the classical neuroleptics? The problematic answer from the CATIE study. Eur Arch Psychiatry Clin Neurosci 255:371–372
35. Möller HJ (2005) Problems associated with the classification and diagnosis of psychiatric disorders. World J Biol Psychiatry 6:45–56
36. Möller HJ (2008) Do effectiveness (“real world”) studies on antipsychotics tell us the real truth? Eur Arch Psychiatry Clin Neurosci 258:257–270
37. Möller HJ (2008) Is there a need for a new psychiatric classification at the current state of knowledge? World J Biol Psychiatry 9:82–85
38. 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
39. Möller HJ (2009) Standardised rating scales in Psychiatry: methodological basis, their possibilities and limitations and descriptions of important rating scales. World J Biol Psychaitry 10:6–26
40. Möller HJ, Bottlender R, Grunze H, Strauss A, Wittmann J (2001) Are antidepressants less effective in the acute treatment of bipolar I compared to unipolar depression? J Affect Disord 67:141–146
41. Möller HJ, Fischer G, von Zerssen D (1987) Prediction of therapeutic response in acute treatment with antidepressants. Results of an empirical study involving 159 endogenous depressive inpatients. Eur Arch Psychiatry Neurol Sci 236:349–357
42. Möller HJ, Grunze H (2000) Have some guidelines for the treatment of acute bipolar depression gone too far in the restriction of antidepressants? Eur Arch Psychiatry Clin Neurosci 250:57–68
43. Möller HJ, Maier W (2009) Evidence-based medicine in psychotherapy: possibilities, problems and limitations. Eur Arch Psychiatry Clin Neurosci (submitted)
44. Nies AS (1990) Principle of therapeutics. In: Goodman Gilman A (ed) The pharmacological basis of therapeutics. Pergamon, New York, pp 62–83
45. Novick D, Haro JM, Suarez D, Lambert M, Lepine JP, Naber D (2007) Symptomatic remission in previously untreated patients with schizophrenia: 2-year results from the SOHO study. Psychopharmacology (Berl) 191:1015–1022
46. Plutchik R, Platman SR, Fieve RR (1969) Three alternatives to the double-blind. Arch Gen Psychiatry 20:428–432
47. Quitkin FM, McGrath PJ, Rabkin JG, Stewart JW, Harrison W, Ross DC, Tricamo E, Fleiss J, Markowitz J, Klein DF (1991) Different types of placebo response in patients receiving antidepressants. Am J Psychiatry 148:197–203
48. Rickels K (1986) Non-specific factors in drug therapy. Thomas, Springfield
49. Rickels K, Lipman RS, Fisher S, Park LC, Uhlenhuth EH (1970) Is a double-blind clinical trial really double-blind? A report of doctors’ medication guesses. Psychopharmacologia 16:329–336
50. Riedel M, Strassnig M, 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
51. Rothman KJ, Michels KB (1994) The continuing unethical use of placebo controls. N Engl J Med 331:394–398
52. Seemüller F, Möller HJ, Obermeier M, Bauer M, Adli M, Kronmüllert K, Holsboer F, Brieger P, Laux G, Bender W et al (2009) Do efficacy and effectiveness samples differ in antidepressant treatment outcome? Am J Psychiatry (in press)
53. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, Hergueta T, Baker R, Dunbar GC (1998) The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 59(Suppl 20):22–33 quiz 34–57, 22–33
54. Sneed JR, Rutherford BR, Rindskopf D, Lane DT, Sackeim HA, Roose SP (2008) Design makes a difference: a meta-analysis of antidepressant response rates in placebo-controlled versus comparator trials in late-life depression. Am J Geriatr Psychiatry 16:65–73
55. Storosum JG, van Zwieten BJ, van den BW, Gersons BP, Broekmans AW (2001) Suicide risk in placebo-controlled studies of major depression. Am J Psychiatry 158(8):1271–1275
56. Taves DR (1974) Minimization: a new method of assigning patients to treatment and control groups. Clin Pharmacol Ther 15:443–453
57. Versavel M, Leonard JP, Herrmann WM (1995) Standard operating procedure for the registration and computer-supported evaluation of pharmaco-EEG data. ‘EEG in Phase I’ of the Collegium Internationale Psychiatriae Scalarum (CIPS). Neuropsychobiology 32:166–170
58. Walsh BT, Seidman SN, Sysko R, Gould M (2002) Placebo response in studies of major depression: variable, substantial, and growing. JAMA 287:1840–1847
59. Wittenborn JR (1977) Guidelines for clinical trials of psychotropic drugs. Pharmakopsychiatr Neuropsychopharmacol 10:205–231