Eur Arch Psychiatry Clin Neurosci (2008) 258:257–270 DOI 10.1007/s00406-008-0812-0

ORIGINAL ARTICLE

Hans-Jürgen Möller

Do effectiveness ("real world") studies on antipsychotics tell us the real truth?

Received: 5 December 2007 / Accepted: 18 December 2007 / Published online: 11 July 2008

Abstract In recent years, so-called "effectiveness studies" have gained increasing importance in the context of evidence-based medicine. These studies supposedly follow less restrictive methodological standards than phase III studies in terms of patient selection, co-medication and other design issues, and their results should therefore be better generalisable than those of phase III studies. Effectiveness studies, like other types of phase IV studies, can therefore contribute to the knowledge about antipsychotics or other psychopharmaceuticals and supply relevant information in addition to that gained from phase III trials. However, the less restrictive design and inherent methodological problems of phase IV studies mean that their results cannot falsify the results of phase III studies. The greater complexity of phase IV studies, for example the greater variance caused by the different kinds of confounders, means that their results have to be interpreted with great care and especially with a high degree of awareness of problematic design issues, such as insensitive primary outcome criteria, biased randomisation, unblinded treatment conditions, inclusion of chronic refractory patients, etc. Some recently published effectiveness studies on antipsychotic treatment of schizophrenia will be discussed under these methodological aspects. The main conclusions of these trials will be questioned on the basis of their severe methodological pitfalls.

Key words effectiveness studies · real world studies · antipsychotics · SGAs · FGAs

Prof. H.-J. Möller (✉)
Department of Psychiatry
Ludwig-Maximilians-University München
Nussbaumstrasse 7
80336 Munich, Germany
Tel.: +49-89/5160-5501
Fax: +49-89/5160-5522
E-Mail: hans-juergen.moeller@med.uni-muenchen.de

Evidence-based medicine is supposed to form the basis for a rational handling of the available diagnostic and therapeutic possibilities. However, although such a rationalisation is desirable, it can also be associated with many problems [37, 50]. On the one hand there are problems associated with the methods of evidence-based medicine. On the other, the intentions of evidence-based medicine may be too closely related to the aim of generating criteria for an adequate distribution of scarce resources within the healthcare system. Rationalisation may thus possibly become rationing.

In the discussion of the advantages and disadvantages of certain medicines, the phase III "efficacy" studies, such as are performed for drug licensing applications, are being increasingly compared with the results of so-called "effectiveness", "real-world", "pragmatic" or "practical" trials, whereby these terms are used synonymously. Such trials are alleged to follow less restrictive methodological standards, at least in the view of protagonists of "effectiveness" studies (cf., for example, the argumentation of Lieberman [33]). There is a general consensus that the results of phase III studies are not fully generalisable: they have a high internal validity but insufficient external validity. One of the reasons for this is the particularly strict selection of patients according to various clinically relevant characteristics, for example the exclusion of comorbidity. For this reason it has long been a tradition within clinical psychopharmacology to complement such studies with ones more strongly oriented towards everyday clinical practice and conditions, i.e. studies in patients who better represent the patients normally presenting for treatment and under conditions as close as possible to routine care, e.g. phase IV studies (Fig. 1). However, it was thereby always stressed that because of many immanent methodological problems, e.g. distortion of the real situation by lack of double-blind conditions or any blinding, such phase IV studies, for example naturalistic observational studies, only deliver com-

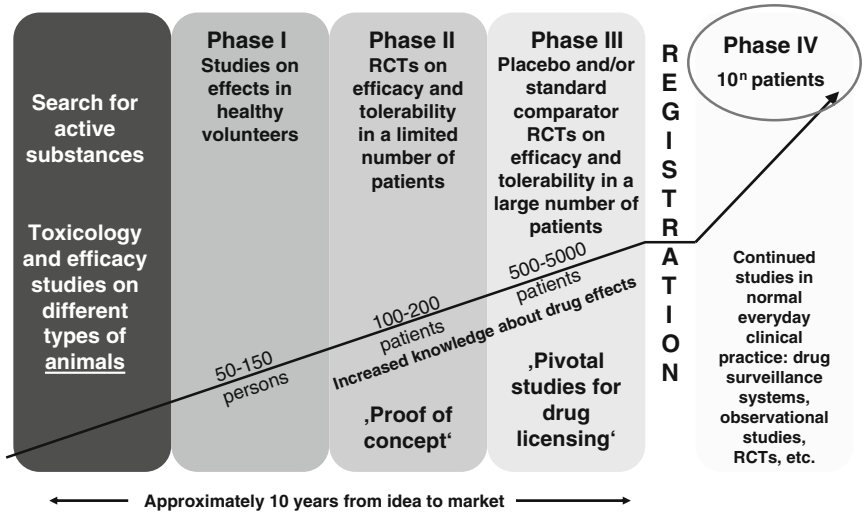

Fig. 1 The four-phases model of clinical psychopharmacology

plementary knowledge and cannot falsify the results of phase III studies. Doubt is currently arising about this basic assumption, which is virtually compulsory from a methodological point of view, because of the increasingly high esteem in which "effectiveness studies" are held. Many people seem inclined to attach a greater importance to the results of these studies than to the methodologically stricter phase III studies. This might result from criticism arising from the increasingly common practice, especially in the USA, to include in phase III studies not "real" patients from care settings but suitable persons found through advertisements. Of course, not this questionable approach but properly performed phase III studies in "real" patients should be advocated.

It is the intention of this paper to answer the question whether effectiveness studies, which are often extremely expensive to perform and mostly sponsored by government or public institutions, are really as advantageous as claimed. The answer will be based on principal methodological considerations and a methodological analysis of some recently published effectiveness trials on antipsychotics which did not confirm the advantages of SGAs over FGAs that have been found in phase III studies.

Some principal methodological considerations of "effectiveness studies"

What are effectiveness studies? "Effectiveness studies are intended to fill the gap between methodologically rigorous randomized clinical trials (RCTs) and naturalistic observational studies. As such, they are hybrids of RCTs and naturalistic or quasi-experimental designs and are termed "practical clinical trials" [77]. They are intentionally designed to evaluate the effectiveness of the treatments under real-world conditions and in representative patient samples" (33, p. 1070). The actual advantage of these "effectiveness" studies remains questionable. Do they do justice to their claim of treating less selective samples of patients than phase III studies? (Table 1)

Some "effectiveness" studies might have a better external validity in terms of sample representativeness. However, even this has to be demonstrated for each study and cannot be seen as being proven a priori. Each study sample has to be shown really to be more representative of the "real world" than typical phase III samples. It is not sufficient to state that certain inclusion/exclusion criteria from phase III studies were not addressed, but the whole selection process from the numbers of principally eligible patients to the number finally included has to be described, together with full details of reasons for non-inclusion. The degree to which the sample represents typical patients can only be assessed on the basis of such information.

Some effectiveness studies appear to have a different kind of selection of patients than phase III trials (Table 2). For example, in the effectiveness study comparing olanzapine and haloperidol in the treatment of schizophrenia [64], of the 4,386 patients assessed for eligibility, only 309 were included in the study (7.0%). This rate is even somewhat lower than the usual rate of 10–15% in phase III studies [19]. Often, patients with milder and more chronic symptoms may be selected than is the case in phase III studies, thus making it more difficult per se to demonstrate drug effects and in particular differences between drugs' effects, because a relevant subgroup of patients might be partially unresponsive to a drug. The data of the CUtLASS study serve as an example here. In this study, the pre-post changes in the PANSS positive score after 52 weeks amounted to only 2.0 in the FGA arm and 1.5 in the SGA arm; these changes are extremely low, even when one takes into account that this study was not an acute treatment study but rather a switch study in partially improved/stabilised patients. A lot of studies have demonstrated much

Table 1 Comparison of characteristics of clinical trials of "efficacy" with trials of "effectiveness" [41]

Clinical trials of "Effectiveness"
More relaxed exclusion criteria, permitting wider range of:
Treatment settings and interventions (including adjunctive treatments)
Levels and/or types of psychopathology
Allows greater emphasis on clinical need to determine treatment (e.g. doses, etc.)
Broader scope of outcome measurement, some of which may include:
Time to discontinuation
Time to hospitalization/rehospitalization
Time to relapse
Quality of life (including ability to work, level of social functioning)
Advantages
Higher external validity
Arguably greater applicability to "real world" practice settings
Capacity to inform policy process
Longer duration
Can enroll large number of patients more easily
Disadvantages
Internal validity limited
Cannot be used to examine effective dose ranges
Cannot make as meaningful clinical comparisons between agents
Sizes of therapeutic effect(s) cannot be calculated
Clinical trials of "Efficacy"
Highly restricted inclusion criteria to reduce confounding, bias
Randomization and blinding, also to reduce bias
Treatment driven by study protocol
Patients remain only in the treatment group originally assigned
Fewer treatment adjustments are allowed
Strict limitations on adjunctive treatment
Measures taken to insure all members of treatment group receive same intervention(s)
Use of well-validated outcome assessment
Advantages
Higher internal validity for clinical effects
Higher internal validity for adverse effects, tolerability
Contextual and human factors controlled for
Considered "best quality" clinical evidence for informing treatment decisions
Disadvantages
Stringent inclusion criteria limit external validity
Outcome measures may not reflect crucial advantages and limitations of interventions being studied
Outcome measures may not address issues most important to patients and families
Often short in duration

Table 2 Special characteristics of samples in effectiveness studies

No exclusion of comorbidity
No restriction of severity
Suicidality allowed (?)
All age groups allowed
No exclusion of patients with certain pre-treatments
Partial non-responders/chronic cases not excluded
Comedication allowed
Etc.

Advantage efficacy, tolerability, compliance can be measured in "real world" conditions
Disadvantage increase of variance ("more noise"), reduced effects (e.g. small pre-post differences in the CUtLASS study)

Table 3 Methodological restrictions of effectiveness studies

More noise, reducted signal detection, increased ß-error problem
Often completely unblinded procedure → bias of doctors and patients influence results.
The so-called "blind rater" is often not really blind
If no statistically significant difference between two active components, difficult to draw conclusions: real result or ß-error problem?
Lack of placebo control, unclear whether real drug effect

higher score reductions in different samples and a range of different timeframes and conditions [4, 5, 7, 10, 32, 36, 73, 82]. Furthermore, in contrast to phase III studies, the "real world" approach allows more comorbidity, comedication, etc., so that a broader range of information may be obtained than from the respective phase III studies (Table 3). However, there is often no differentiated analysis of the influence of these variables. Thus, no advantage is taken of the chance to learn more about these "confounders". On the other hand, the inclusion of such "confounders" (from the perspective of a phase III trial) increases the variance and results in a reduced signal-to-noise ratio, which makes it more difficult to find differences between two groups (ß error problem), even if these factors are adequately considered in the statistical analysis. It might sometimes even be difficult to judge without placebo conditions whether there is a real drug effect, especially if the pre-post difference is unexpectedly low and if there are no differences between two active comparators.

Because inclusion of patients with comorbidity, comedication, etc., means that variance is much higher, a precise power calculation should be per-

formed and published. It should be based not only on the between-group difference and the standard deviation taken from phase III study results but also on the at least hypothetically reduced delta values of such an effectiveness study, taking into account the higher variance and thus the reduced signal/noise ratio.

Given the fact that these pragmatic trials mostly compare two active compounds, it should be accepted on the basis of the traditional methodology of clinical psychopharmacological trials that only proof of superiority in the statistical sense counts, while the failure to demonstrate a statistically significant difference cannot be interpreted as showing that both treatments are comparable. The latter conclusion is not permissible for principal methodological reasons. A different statistical design is required to demonstrate equivalency: the so-called equivalency design. The lack of a placebo group in "effectiveness" trials also has consequences for study design and possible methodological pitfalls [1, 3, 45, 46]. Without a placebo control one cannot be sure that the active drugs are being compared in a drug-sensitive sample. The worst-case scenario is that the drugs show no outcome difference because they are not effective at all in the respective sample. This is not as unlikely as some people might believe. In the field of antidepressants, failed studies—in the sense that in a three-arm study comparing an experimental drug with a standard comparator and placebo not even the standard comparator (internal validitor) differs from placebo—are quite common. In recent years there has even been an increasing number of failed studies, especially in the United States, in the field of antipsychotics, although the antipsychotics generally have a larger effect size. Several factors are relevant in this context such as low inter-rater reliability, especially in huge multicentre trials, inclusion of less responsive patients, more chronic patients with residual symptomatology or comorbid patients, no restriction of permitted comedications, etc.

It should be questioned whether so-called pragmatic primary outcome criteria such as "non-discontinuation", or similar categorical endpoints like "level of caring", really are ideal outcome criteria, given the fact that they can easily be influenced by the investigators (who may be biased by their expectations if they are not blinded) and are of poorer psychometric value than dimensional ones. It can be generally questioned whether "non discontinuation" really reflects only efficacy and tolerability aspects or whether also other parameters beyond drug effects are involved, e.g. the confidence in the therapeutic concept. For example, therapeutic concepts like psychotherapy, herbal drug therapy, etc., might be more acceptable to a subgroup of patients, although they may have a lower level of efficacy. Different aspects of tolerability can have different effects on discontinuation, depending on the specific tolerability problems and on the time patterns of side effects. Thus, one can presume a priori that severe extrapyramidal symptoms occurring right at the start of a study result in an early drop-out, the slow development of weight gain rather a later drop-out, and tardive dyskinesia (TD) or in most cases even metabolic disorder a much later drop-out. This means that a rough measurement like "discontinuation" or "time to discontinuation" causes a biased distortion per se with respect to the individual antipsychotics being evaluated. This becomes even worse if the transition from the pre-treatment antipsychotic to the study antipsychotic is taken into consideration, especially if it is direct, without a sufficiently long wash-out phase. Depending on the pharmacological profile of the respective pre-treatment drug, for example in terms of D_2 potency, anticholinergic or antihistaminergic properties and the related pharmacological profile of the study drug, several problems can appear immediately after transition [6]. These can include reduced antipsychotic efficacy, discontinuation symptoms, hangover of side effects wrongly attributed to the study drug, pharmacodynamic interactions in terms of oversedation, histaminergic or cholinergic rebound phenomena, etc. Thus, there are good and bad combinations of drugs for this transition process. Theoretically, the best transition is one in which the pre-treatment and the study drug are identical. There are also other critical issues that need to be considered in this context [80, 81].

Another measure of global outcome used as a primary outcome criterion in effectiveness studies is "quality of life". There is no doubt that this is an important outcome criterion which reflects the subjective dimension of the patient's experience [24, 27, 55]. The classical approach in quality of life research assesses quality of life using a self-rating scale in order to guarantee the subjective perspective. The SF36 [78, 79] is particularly widely used in psychiatry as well as in other fields of medicine, but there are also several other scales to assess this dimension [9, 60, 61]. The quality of life scale developed by Heinrichs [17] is often used in schizophrenia research, although it is not a self-rating instrument and might rather measure negative symptoms than quality of life in the stricter sense. It has only a modest correlation with self-rating scales of quality of life [59].

There are pros and cons for the use of self-rating scales. They give a complementary view to the observer rating of the same construct/dimension [43, 49]. The correlation between the observer ratings and self-ratings might not be high and may be quite changeable, depending on the psychopathological state in terms of severity and type of symptoms [58]. It is often unclear exactly what self ratings of quality of life reflect: severity of the psychopathological state in the global sense, certain dimensions of the psychopathological state, e.g. depression, current mood more than real depressive symptoms, side effects of drugs, or the psychosocial situation [2, 14, 24, 59, 62,

76]. If such a scale is used as the primary outcome criterion of a study, it is doubtful whether it is sensitive enough to detect inter-group differences of treatment-induced changes, given the high variance of self-rating in general and of self ratings of quality of life in particular. Not many of the studies on antipsychotics that used a quality of life scale as a secondary outcome criterion found significant intergroup differences [24]. Thus, the use of a quality of life scale carries a high risk of not finding significant differences between two drugs, especially if both are active drugs.

Randomised control-group studies sponsored by drug companies are often accused of being heavily biased, even phase III studies used to license a drug. It would indeed be possible to bias the results of a study, especially studies comparing two active compounds, for example when selecting the dose seen to be equivalent. Indeed, such subtle factors can result in the situation that the second generation antipsychotic A from drug company A is found to be superior to the second generation antipsychotic B from company B in a drug trial organised/sponsored by company A, while the opposite result is obtained from a similar active comparator study organised/sponsored by company B. However, contrary to what many experts may have assumed, apart from these kinds of subtle influences a methodologically oriented review of this issue found no serious methodological distortions of recent trials on antipsychotics [18]. From a more realistic point of view this does not appear surprising given the fact that all RCTs used for licensing a drug are under strict methodological control by drug authorities. Protagonists of "effectiveness" studies often underline that their own study or other effectiveness studies are not sponsored by a pharmaceutical company, which seems to imply that they assume that drug companies intentionally bias studies. In this context it should be considered that other sponsors, even public domain/government sponsors, could theoretically cause unintentional or even intentional distortion of the methodology of a drug trial, e.g. with the aim to prove the beneficial aspects of a cheap drug. Because this hypothesis cannot be completely rejected, at least not a priori, the type and interest of a public sponsor (e.g. government institution, insurance company, etc.) should be carefully analysed for potential conflicts of interest and possible bias of design issues such as fair randomisation, selection of compounds, dosing strategies, etc.

In order to demonstrate some of the methodological problems of "effectiveness" studies, four such studies on antipsychotics, published in recent years, will be described below: CATIE [35], CUtLASS [22], and the studies by McCue et al. [39] and Rosenheck et al. [64]. These studies are often cited by critics of the advantages of the modern antipsychotics as evidence that modern antipsychotics are not superior to the classical antipsychotics. It is interesting that all four studies were published in high-ranking journals, although they have considerable methodological shortcomings which mean that the conclusions drawn are not tenable, especially not when they are used to falsify the results of phase III studies.

The CATIE study

There is no doubt that the CATIE study ("Clinical antipsychotic trials of intervention effectiveness"), performed in the USA, is an important study when one considers, for example, the large sample size (N = 1,493 in 57 centres), the complex design with several parallel treatment arms, the 18-month duration of treatment of the first phase, inclusion of sequential treatment phases, etc. (Phase 1 of the study was published in 2005; 35). Also, the double-blind conditions of this study and the sophisticated and comprehensive statistical analysis of the extensive database are appealing. The fact that the study had a public sponsor, the NIMH, and surely cost an enormous amount of money further underlines that its results could be expected from the outset to be of great importance. This was also highlighted by its publication in a high-ranking journal. However, the methodology of the CATIE I study has given rise to much criticism and many concerns [15, 23, 25, 41, 47, 63, 81].

The primary outcome criterion of the CATIE I study was "discontinuation of treatment for any cause"—which is increasingly used as an outcome criterion in effectiveness studies (for example, 75). This "effectiveness measure" generally reflects efficacy in the narrower sense and safety, but in the CATIE study it even includes doctors' and patients' decision-making [72].

As explained above, the "non-discontinuation" criterion is not as unproblematic as it might appear at first glance. This is especially relevant when considering that most of the patients were not treated de novo, but were switched from one drug to another. Depending on the pharmacological profile of both the respective pre-treatment antipsychotic and the respective study antipsychotic, this switch can have severe complications, e.g. cholinergic or histaminergic rebound if the pre-treatment antipsychotic is anticholinergic or antihistaminergic and the study drug not. There is no risk of such problems if both drugs have anticholinergic or antihistaminergic properties. Such reasons might at least partially explain the high drop-out rate in the first weeks of the study as well as the differences found between the groups, even though patients were allowed to continue taking their pretreatment antipsychotic for a maximum of only 4 weeks after randomisation to the study drug. Of special interest in this context is that the percentage of patients in the olanzapine and risperidone groups who had been pre-treated with the

same antipsychotic was much higher than in the other groups, which might have had a positive influence on the outcome of these subgroups [11].

It should be taken into consideration that this study was not an acute treatment study which was subsequently continued as a maintenance study, but that at the time of recruitment the patients were apparently rather in a partially stable, subacute or even partially remitted psychopathological condition. The study was therefore primarily a kind of switch or even maintenance study. Furthermore, the sample consisted of relatively chronic patients, with an average of 14.4 (±10.7) years since the first antipsychotic medication. The proportion of partially treatment refractory patients is not mentioned but has to be seen to be an important confounder in such samples of chronic patients [41]. According to the published recruitment criteria, patients refractory in the stricter sense were excluded. However, the average time since the last acute psychotic episode was 3 months and, since the average PANSS total score was 76 (indicating a relatively high degree of severity) and most of the patients were already being treated with antipsychotics, it has to be concluded that the average therapeutic response was insufficient or, to put it differently, the therapeutic response was insufficient in a relevant subgroup of patients. It is much more difficult to ascertain differences between antipsychotics under substable/stable or partially drug unresponsive conditions than under acute treatment conditions since under the latter allow even a less effective drug to demonstrate "good efficacy" for a certain time. We know that even when stable and particularly remitted patients are switched from an antipsychotic to placebo it usually takes months until relapses occur [20]. Thus the "good efficacy" might be at least partially pseudo-efficacy. Total PANSS scores improved over time in all groups and the mixed model revealed significant variation in treatment effects over time ($P = 0.002$). However, according to a comment by Meltzer and Bobo [41] the improvements in the PANSS scores after 6 weeks were less than half those reported in clinical trials with more acute patients, even in the olanzapine-treated group ([41], p. 18). Apart from this principal selection bias, apparently the sample is an excellent representation of the real world situation: of 1,894 patients screened, 1,493 were randomised and the published reasons for non-inclusion seem acceptable.

As to the primary outcome criterion of time to discontinuation, the studies main finding is that the second generation antipsychotic (SGA) olanzapine was somewhat superior to the other SGA and the first generation antipsychotic (FGA) perphenazine. "The time to the discontinuation of treatment for any cause was longer in the olanzapine group than in the quetiapine group (hazard ratio, 0.63; $P < 0.001$), the risperidone group (hazard ratio, 0.75; $P = 0.002$), or the perphenazine group (hazard ratio, 0.78; $P = 0.021$). However, the difference between the olanzapine group and the perphenazine group was not significant after adjustment for multiple comparisons (required P value, ≤0.017). Within the cohort of 889 patients who underwent randomization after ziprasidone was added to the trial, those receiving olanzapine had a longer interval before discontinuing treatment for any cause than did those in the ziprasidone group (hazard ratio, 0.76; $P = 0.028$). However, this difference was not significant after adjustment for multiple comparisons (required P value, ≤0.013)" ([35], p. 1212). The most general conclusion, which was also widely received by the public and published in several eminent newspapers, was that (in terms of effectiveness) the SGAs are not superior to the FGA perphenazine [35]. Methodologically stringent calculations of several secondary endpoints were performed which demonstrated trends or even statistically significant differences between some antipsychotics in PANSS-related efficacy measures or tolerability measures such as extrapyramidal symptoms (EPS), weight gain, etc. The tolerability-related data in particular more or less confirmed the known tolerability pattern of the individual antipsychotics.

Given the sensitive point of dosing issues mentioned above [18], it seems problematic that one of the drugs (olanzapine) could be administered at a daily dose of up to 30 mg, which is significantly more than the licensed and recommended daily dose (recommended dose in the USA according to the Physicians' Desk Reference: 5–20 mg/day), while all other neuroleptics were administered within the range of the licensed dose. It also seems questionable that perphenazine was intentionally allocated a rather lower dosage range (8–32 mg/day) than is usual with the intention to reduce the risk of EPS ([35], p. 1218). According to the Physicians' Desk Reference, the recommended dose for perphenazine in the USA is 8–64 mg/day. The higher dosing of olanzapine (up to 30 mg/day), which resulted in an average daily dose of 20 mg, may explain the better "effectiveness" (especially in the sense of a lower drop-out rate due to better efficacy) of olanzapine. The fact that olanzapine still had a fairly favourable drop-out rate even at this high dosage (despite weight gain and changes of metabolic parameters) shows that this neuroleptic obviously has a high general tolerability margin, and indicates that adverse events such as weight gain only result in drop-out after a considerable delay.

Beside the question whether the study had sufficient power to find statistical differences in the main outcome criterion between the groups, especially between the SGA group and the perphenazine group, the decisive problem is that all patients who presented with TD at the start of the study were excluded from randomization to the perphenazine group (the only representative of the FGAs) but not from randomi-

zation to one of the SGAs. As can be seen from the comparison of the numbers of patients in the various treatment arms, this was the case in a relatively large number of patients. There were only about 260 patients in the perphenazine arm while the SGA arms had about 340 patients each. Altogether, 231 patients with TD were excluded from randomisation to the perphenazine group. Of course, such a procedure that favours one treatment arm and thus automatically discriminates against the others represents a severe methodological error. It is difficult to understand why this shortcoming in the study design was overseen or even knowingly accepted. The possible explanatory argument that one cannot treat patients with TD with a classical neuroleptic (and thus harm them further...) does not solve the methodological dilemma as one thus implies in circular argumentation that one actually already knows that perphenazine (similar to other FGAs) bears a higher risk than SGAs of causing extrapyramidal side effects and thus also TD. However, perphenazine was chosen as the representative of the FGAs on the assumption that it may bear a comparatively lower EPS risk ([35], p. 1215). It is known that TD correlates with the general risk of EPS and that the occurrence of early dyskinesia or parkinsonism is a predictor of the later development of TD [42, 57]. If one removes in a randomised, control-group study these risk patients from the treatment arm receiving an FGA, but leaves them in the groups receiving an SGA, one can no longer discern the actual advantages of the SGAs compared to the FGA. Better extrapyramidal tolerability is known to be the main advantage of the SGAs. It is therefore almost contrary to the expectations that even so the study found a certain advantage in this respect for SGAs compared to perphenazine. Of course, this methodological shortcoming means that the EPS risk of perphenazine itself was completely underrated. In this context it should also be considered that not only the motor disturbances of EPS are relevant but these symptoms can also have an impact on psychopathological domains such as negative symptoms, depression and cognitive disturbances. Going beyond the EPS-related problems, it might also be assumed that patients with TD represent a more severely ill group of schizophrenic patients with a poorer prognosis. There are at least some data indicating that this might be the case [16, 34, 52–54]. Finally, it should be underlined that the exclusion of TD patients from randomisation lead in addition to differences in the frequency of drug applications per day to a partial unblinding of the study [15].

The CATIE research group later attempted to defend this error in design [33], although none of their arguments seems convincing. For example, the publication states that in order to control for the subgroup of patients with TD, they were excluded from the statistical analyses comparing SGAs with the FGA perphenazine [35]. However, why were they then included in the study? Why do they appear in the graphical representations of the survival analysis on time to discontinuation? Why do they appear in the table of outcome measures: olanzapine group $N = 336$, etc., while perphenazine group $N = 261$? One would expect a similar number in each group if this randomization failure had been corrected.

Even though the patients with TD were removed from all groups in the statistical analysis, as far as perphenazine is concerned this still does not fully correct another principal methodological problem related to the TD issue. If a special risk group of patients is not included in a drug trial, because one of the drugs is known to be harmful, then the design is biased in favour of this drug. It would be interesting to know why a study sponsored by public/government money had such an unfair design that gave an old/cheap drug a much better chance of success. If the same approach would have been applied concerning the risk of weight gain, antipsychotics with a known risk could potentially be shown not to have this risk. Why were overweight patients or, to go to the extreme, patients with diabetes (about 11% of the total sample!) not excluded from the olanzapine group, for example?

Returning to the implications of the primary outcome criterion "all-cause discontinuation" (ACD), a recently published critical review and reanalysis of the main CATIE results [81] seems of greatest importance: "…, some of the drawbacks of the ACD approach are not well understood in part because of unfamiliarity with the way ACD was assessed and problems with the use of hierarchical criteria to establish the primary reason for medication discontinuation. Using the time until ACD as an endpoint cannot by itself capture the complexity of the trajectory of a patient's response to a new medication. In particular, it is quite plausible that switching medications upon entering CATIE phase 1 would reduce some symptoms, which then would lead to greater desire to make another medication switch. Using ACD criteria, a comparison with CATIE subjects who coincidentally remained on treatment with their pre-switch medication would make it seem that switching was detrimental when in fact it could have been helpful. Another major limitation of the ACD was omitting the recording of the reason for stopping study medication whenever the ACD was considered to be a "patient-decision" discontinuation. This means that patient-initiated discontinuations could never be classified as a tolerability discontinuation. Since the ACD was done by the patients' clinicians, this approach may have underestimated the proportion of side effect discontinuations whenever the patient disagreed with his or her clinician. Moreover, retaining the "patient-decision" discontinuation subgroup in the attributable risk estimates of tolerability discontinuations further minimizes the attributable risk estimate of the role of side effects relative

to other causes of discontinuation. For these assumptions to be valid would require the very optimistic assumption that CATIE clinicians never underestimated tolerability concerns in their patients. Otherwise, this mutually exclusive approach will lead to significant underestimation of the proportion of CATIE discontinuations caused by tolerability problems. It can be argued that excluding the "patient decision" subgroup from the attributable risk estimate of role of tolerability in medication discontinuation is a better approach to mitigate against these biases. A reanalysis using an adjusted *N* of 1,061 evaluable subjects changes the attributable portion of tolerability discontinuation from 14.9 to 38.5%. Regarding specific side effect-medication pairs of interest, the attributable risk of extrapyramidal symptoms as a reason for discontinuing perphenazine increases from 8 to 21% and weight-related discontinuations from olanzapine from 9 to 28%." ([81], p. 12).

Some other methodological problems will be summarised only briefly here as they were addressed by other critical reviews: one of the points mentioned by Glick [15] was the partial unblinding connected with the different schemas of drug administration as well as complications due to the large number of sites involved in the study and some statistical issues; Meltzer and Bobo [41] mentioned the possibility that a high number of patients were already drug refractory before entry into the study and that patients could be reassigned to those antipsychotics which they had taken before without adequate efficacy or tolerability.

As mentioned above, the study has received a lot of publicity, particularly in the general press, where it was portrayed as showing that SGAs are for the most part not better, but much more expensive, than FGAs. This conclusion is not tenable because of the methodological failings described above and elsewhere [25, 47]. However, to end on a more positive note, many other results not only from phase 1 but also phases 2 and 3 are of relevance for clinicians, e.g. on different side effect patterns of individual SGAs, on metabolic issues, on meaningful sequences of antipsychotic treatment in case of partial non-response, on the unique efficacy of clozapine in refractory patients, etc. [41, 74].

The CUtLASS study

The results of the first part of the Cost Utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS I) was published, like CATIE, in a high-ranking journal. This is astonishing given the significant methodological shortcomings that become apparent when the publication is examined more closely, and that cast doubt on the validity of the conclusions drawn from the results. Like the CATIE study, CUtLASS I was not sponsored by the pharmaceutical industry but by a government body (the National Health Service).

CUtLASS I was an effectiveness study performed in the UK [22] involving psychiatric services of five medical schools. Patients ($N = 227$) enrolled required a change in treatment because of inadequate response to or adverse effects from their pre-treatment, so that the study can be classified as a switch study. They were randomly prescribed either FGAs or SGAs (other than clozapine), with the choice of individual drug made by the managing psychiatrist. The study was not blind in the strict sense but used a so-called blind assessment approach. The sample was characterized by symptomatically stabilized, relatively chronic, partially non-responsive patients of community psychiatric services with a mean illness duration of about 14 years. The reasons for referral were as follows: inadequate drug response to pretreatment alone in 44% (FGA arm) and 54% (SGA arm) of patients; adverse effects alone in 30% (FGA arm) and 12% (SGA arm); presence of both reasons in 26% (FGA arm) and 34% (SGA arm).

Beside the primary outcome criterion quality of life, other outcome measures were symptoms, adverse effects, patient satisfaction and costs of care. Treatment duration was one year. The study found no advantage of SGAs over FGAs on Quality of Life Scale scores, symptom scores or discontinuation rates after one year of treatment; costs were similar. After 52 weeks, the average pre-post difference in scores on the PANSS positive subscale was −2 points in the FGA arm and −1.5 points in the SGA arm; changes in the PANSS negative subscale were also small: −3.3 in the FGA arm and −1.8 in the SGA arm. The authors concluded that in people with schizophrenia whose medication is changed for clinical reasons, there is no disadvantage across one year in terms of quality of life, symptoms or associated costs of care in using FGAs rather than non-clozapine SGAs. This is principally a fair conclusion from the results of the study. However, it is astonishing that the prescription of cheaper drugs does not lead to a significant reduction of direct health costs, apparently because the reduction in the budget for antipsychotics is compensated for by higher costs for inpatient treatment. This seems important under health economic aspects because the tendency to prescribe cheaper drugs is mostly driven by the desire to reduce health costs. The prescription of cheaper drugs apparently does not automatically achieve this goal.

Without going into too much detail, it can be stated that many design characteristics of CUtLASS 1 are open to criticism [48]:

- The sample size was quite small ($N = 227$ randomized; $N = 185$ after one year) for comparison of active drug regimes, although in the discussion of the results the authors tried to argue against this assumption.

- The main outcome criterion, quality of life, is generally not very sensitive for outcome differences between two drug treatment groups. In a review of controlled studies comparing SGAs with FGAs, only a few studies showed an advantage for the tested SGA [24]. Quality of life might often be more closely associated with depressive or negative symptoms than with psychotic symptoms [60, 61]. The quality of life scale from Heinrichs et al. [17] was developed primarily as an instrument for rating the schizophrenia deficit syndrome. It is an observer-rated instrument, based on a semi-structured interview, and not a self-rating instrument as is usual for the assessment of quality of life. Thus, quality of life, and especially the scale from Heinrichs et al. cannot be seen to be an ideal primary outcome criterion and its validity as a measure for quality of life in the stricter sense might be questionable.
- The so-called "blind assessments" used in the study cannot be considered to be equivalent to blind or double-blind conditions. The doctor responsible for the patient's care selected the drug from either the FGA or SGA group, and patients were informed which drug they were receiving. Even if the "blind rater" did not belong to the ward they might have been influenced by open or latent hints from the doctors or patients.
- Randomization was only used to assign patients to the FGA or SGA arm. The study thus compared the primary choice of any SGA to the primary choice of any FGA and not specific agents.
- The selection of drugs in the FGA arm was apparently biased (see below!): 58 of 118 patients in the FGA arm (49%) received sulpiride. Sulpiride is a low-potency FGA that is a more selective D_2 antagonist than haloperidol and has strong chemical similarities to the SGA amisulpride.
- Only 59% of patients continued taking their originally assigned medication for the full year: 65% (71/109) of patients in the SGA arm and 54% (64/118) in the FGA arm.
- Adjunctive medication was allowed (although the protocol discouraged antipsychotic polypharmacy) but not taken into account in the analysis as a possible confounder or considered in the interpretation of the results.

It is unclear why the physicians so frequently prescribed sulpiride since a Cochrane Review failed to find substantial evidence that it was superior even to placebo, let alone either FGAs or SGAs [67]. This corresponds to the experience in many European countries in which sulpiride, at low doses, is rather used as a kind of atypical antidepressant than as an antipsychotic. Relatively high dosages are considered necessary for its use as an antipsychotic in patients with positive symptoms [8]. Of interest is that the prescription rate of sulpiride in the UK does not correspond to the high prescribing rate in the CUtLASS study. In 2005, for example, the average defined daily dose (DDD) prescription rate for sulpiride in the UK was 1.25% of the prescribed antipsychotic (IMS/Midas[Sergeant] database). What led to such a large difference between the relatively low prescribing rate of sulpiride in the UK and the over-proportionally high rate in the CUtLASS study? Should one surmise that the doctors were directly or indirectly encouraged to use sulpiride? Why does haloperidol play such a minor role in the CUtLASS study when it is ranked third in the list of prescribing frequencies of FGAs in the UK (IMS/Midas[Sergeant] database)? All these points could lead one to assume that sulpiride was actively employed as a surrogate for the more expensive derivative, amisulpride. Amisulpride is categorized as an atypical neuroleptic on the basis of its clinical characteristics and hypotheses have been proposed as to the pharmacological mechanisms that are relevant for its atypicality [44]. Although sulpiride does not have all the pharmacological characteristics of amisulpride, in the spectrum between typical and atypical neuroleptics it still appears rather to be closer to the atypicals, whereby it has the special property of being a relatively weak potency neuroleptic. The authors of CUtLASS argue that it would have to have remarkably superior efficacy and relative atypicality to negate a real advantage of SGAs, particularly as any such effect would be diluted by the other FGAs [22].

One commentary [65] on the study claimed that some of its design features may be viewed as strengths: the novel design is closer to "real-world practice" than typical monotherapy trials because treatments are always unblinded and numerous drugs are available in real practice; although researchers have focused on differentiating the SGAs from one another, both practice guidelines and physician behaviour suggest that they are treated as a class in clinical practice which justifies evaluating them as a class, as is the case in CUtLASS 1 [65]. However, the fact that doctors were allowed to choose each patient's drug from within the class of FGAs and SGAs was criticised by another commentary [33] since it meant that unequal numbers of patients were assigned to the various drugs within each class; the large range of drugs in each treatment arm complicates the interpretation of the results and makes it impossible to draw any conclusions about cost effectiveness, efficacy or side effects of a specific drug. The non-random selection of drugs combined with the large numbers of drugs and small numbers of patients per group mean that the study only had adequate statistical power to compare groups of drugs rather than individual ones, and the pooling of drugs within classes presumes that all of the various drugs are pharmacologically equivalent, which is very unlikely [33].

A range of sophisticated analytical methods were applied to the CUtLASS data, including multiple imputation to address missing data, and minimally

significant clinical differences were tested to properly support the conclusion that at least in this study FGAs are not inferior to SGAs. In his commentary on the study, Rosenheck writes that these methods might represent an advance for the field in contrast to the potentially biased analytic strategies (most notably, use of last observation carried forward) used in many earlier studies [65]. However, on the other side, this excessive statistical computation of the data could also be seen as too far-reaching an abstraction from reality.

The small changes in the PANSS positive and negative symptoms scales (in a 1-year study!) are either indicative of a drug-insensitive sample or of weak efficacy of the prescribed antipsychotics, meaning that a placebo control would have been necessary to demonstrate a real drug response. For comparison, CUtLASS 2 [32], a similarly designed randomised controlled trial of the effectiveness of prescription of clozapine versus other SGAs in treatment-resistant schizophrenia, enrolled patients whose medication was being changed because of poor clinical response to two or more previous antipsychotic drugs, not because of adverse events as well (as was the case in CUtLASS 1) and found an average pre-post difference of about six points in the PANSS positive subscale and of about five points in the PANSS negative subscale. Although CUtLASS 2 found a statistically significant difference of −4.93 points ($P = 0.013$ between clozapine and a group of novel SGAs in the PANNS total score change, no such difference was found in the first outcome criterion, the Quality of Life score, which again questions the sensitivity of quality of life for finding differences between different drug treatments.

Other studies

The effectiveness study by McCue et al. [39] addresses acute treatment conditions with antipsychotics. In this study, from a screening sample of 584 admissions with a diagnosis of acute schizophrenia, schizoaffective or schizophreniform disorder, a sample of 327 newly admitted patients was randomised to open-label treatment with aripiprazole, haloperidol, olanzapine, quetiapine, risperidone or ziprasidone for a minimum of 3 weeks. The authors report that the study was not supported by pharmaceutical companies. There is no mention at all of a sponsor. Conditions were unblinded, i.e. patients and doctors were aware of the antipsychotic being prescribed. It is interesting that the average dose of haloperidol was quite high (range 4–30, mean 16 mg/day) compared to recommendations from advocates of a low-dose regimen with haloperidol [56], even compared to the suggestions of the APA practice guideline (which recommends a dose range of 5–20 mg/day) [30], which underlines that the haloperidol low-dose approach in particular might often not fit to real world conditions. The average dose of most of the SGAs was more or less in the recommended range. In the first effectiveness measure (improvement in mental status so that the patient no longer required acute in-patient care), haloperidol (89%), olanzapine (92%) and risperidone (88%) were found to be significantly more effective than aripiprazole (64%), quetiapine (64%) and ziprasidone (64%). There were no significant intergroup differences in changes from baseline to endpoint in BPRS ratings, the second measure of effectiveness, with any of the treatments. It should be considered that the main outcome criterion ("improvement in mental status so that the patients no longer require acute inpatient care") may have been voluntarily biased by the study physicians. Interestingly, the statistically significant difference in the main outcome criterion is not reflected in statistically significant differences in the BPRS improvement scores. Prophylactic or intervention treatment with anticholinergics (see below for comments on the problems of prophylactic treatment with anticholinergics!) was allowed, and 47% of haloperidol patients received anticholinergics whereas none of the olanzapine or aripiprazole group did, for example. This explains why there were no differences in the pre-post change of score on the Simpson Angus Scale between treatment groups. Several other comedications were allowed (haloperidol, benzodiazepines, diphenhydramine, antidepressants, mood stabilisers), which might confound efficacy, effectiveness and tolerability measures.

The 12-month effectiveness study by Rosenheck et al. [64], sponsored by the Veterans Administration, compared olanzapine with haloperidol in 309 patients with schizophrenia or schizoaffective disorder using a randomised, double-blind, control-group design. The patients were recruited and treated in 17 US Department of Veterans' Affairs medical centres. A total of 4,386 patients were assessed for eligibility, but only 2,141 (49%) found eligible for further assessment. Of these 2,141 patients, 1,530 (35%) either refused to participate or their clinicians refused to allow them to participate; 7% could not participate for other reasons; and only 309 (7%) of the original number assessed for eligibility provided informed consent and were randomised. These data underline that effectiveness studies do not always manage to avoid being strict in their selection of patients. Among other things, the inclusion criteria stipulated a high symptom score, defined as a BRPS score of ≥36, and also severe social dysfunctioning. Only 8% of randomised patients had been employed in the past three years. The average time of onset of schizophrenia was 20 years earlier. The primary outcome was care costs; secondary outcome parameters included side effects. Flexible dosing of olanzapine and haloperidol was permitted within the range of 5–20 mg/day. The design aspect most open to criticism is the allowance of prophylactic treatment with benztropine (1–4 mg/day), which was given as a fixed combination in the

haloperidol group. This was an unfair design since it gave the haloperidol patients an advantage with respect to extrapyramidal side effects, taking away a priori the main disadvantage of the FGA haloperidol. Given the fact that the better extrapyramidal tolerability profile of the SGA olanzapine is the main and consistently proven advantage of SGAs, there was no chance to find superiority for olanzapine. Of course, driven by the aim to reduce health care costs, it might be acceptable to ask what remains as an advantage of an expensive SGA when cheap drugs like haloperidol can be prescribed in combination with an antidote to extrapyramidal side effects. But such a question is only of interest as long as SGAs are not available as generics, i.e. at a lower price. Some of the SGAs like risperidone and olanzapine will soon be available as generic products, thus radically reducing the enormous price difference between SGAs and FGAs. Apart from this, it has never been medical tradition to prescribe a drug together with an antidote when another drug is available which does not require such a co-administration. In addition, considering the cognitive dysfunction of schizophrenic patients it seems extremely problematic to prescribe an anticholinergic (and, in the case of benztropine, antihistaminergic drug) for a longer period, since anticholinergics have cognitive side effects. For this reason, it is against good clinical practice in the treatment of Alzheimer dementia to prescribe an anticholinergic TCA in case of depressive symptoms. From a clinical standpoint it might be acceptable to give prophylactic benztropine medication in the first few days to reduce the risk of early dyskinesia [38, 69, 70]. Following this line of argumentation, the PORT and APA schizophrenia treatment guidelines support prophylactic treatment with anticholinergics [29, 30]. However, the long-term use has to be questioned in terms of the benefit-risk ratio [28, 68]. Most of the publications on this topic are from the era of FGAs. Given the range of SGAs available, it is principally no longer necessary to administer FGAs with prophylactic anticholinergics, apart from for financial reasons, which will soon no longer be valid (see above). There were no significant differences between groups in study retention; positive, negative or total symptoms of schizophrenia; quality of life; or extrapyramidal symptoms. Olanzapine was associated with reduced akathisia in the intention-to-treat analysis ($P < 0.001$) and with fewer symptoms of TD in a secondary analysis including only observations during blinded treatment with study drug. Small but significant advantages were also observed on measures of memory and motor function [64]. Due to the known negative central side effects of benztropine [66] the discrepancy in memory function found in the study in favour of olanzapine can be seen either as a positive effect of olanzapine or as the result of side effects of benztropine in the haloperidol group.

Of interest is the mean dose of haloperidol: 11.2 mg/day in the first few weeks, 14.3 mg/day in the last six months. This might underline that in the "real world" haloperidol doses are much higher than often recommended by advocates of a low dose haloperidol regimen and more in the range of officially recommended doses for haloperidol in the USA and UK [21].

Summary

To start with the positive issues: effectiveness studies can contribute to our knowledge about the use and effectiveness of antipsychotics. They help us to understand that even novel/expensive drugs have their limitations and that it may not be possible to demonstrate consistently their hypothesised superiority in terms of efficacy, safety, compliance, quality of life, etc., under "real world" conditions in chronic, partially refractory and comorbid patients. In general they can also supply interesting data on dosing issues, sequences of drugs in case of partial response and side effect patterns of individual antipsychotics in "real world" conditions.

The superiority of certain antipsychotics or certain sequential regimens could be demonstrated at least in phases 2 and 3 of the CATIE study and in the CUtLASS II study [32, 40, 71]. The results of these or other effectiveness studies that did not find the expected differences between certain drugs or groups of drugs should be evaluated carefully for methodological problems, which might explain the results that differ from those of phase III studies.

Altogether, the effectiveness studies mentioned above seem to have a lot of methodological pitfalls making it difficult to interpret their results. Given the fact that increased variance due to the inclusion of chronic/poorly responsive/comorbid patients, insensitive or problematic outcome parameters and inadequate sample size increase the risk of a ß-error (failure to detect a difference although there is one), and that unblinded designs can induce different kinds of biases, caution has to be applied when interpreting the results of trials with such problems. As mentioned above, in a careful methodological analysis Heres et al. [18] came to the conclusion that, contrary to the commonly held assumption, most modern clinical trials of SGAs sponsored by the pharmaceutical industry do not have severe or intentional methodological problems. One would hardly have expected studies sponsored by the public domain (and therefore often referred to as "independent trials") to break the traditional rules of experimental design, as was clearly the case in some of the effectiveness studies.

Effectiveness studies, especially those with an inadequate experimental design, are definitely not suitable to cast doubt on the results of the methodologically much stricter Phase III studies. In addition, it is questionable whether some effectiveness studies

really do represent the real-world treatment situation better than classical acute and long-term phase III studies, as some of them obviously also recruit a selective patient sample, although the selection is of a different kind than in phase III studies. Effectiveness studies can therefore give only a complementary and not better picture of reality.

Because of the less restrictive methodology, effectiveness studies are not able to falsificate the results of carefully designed phase III studies. Despite the amount of attention being paid to them, we should not start to doubt earlier findings from phase III studies on antipsychotics [12, 13, 26, 31, 51] but should continue to consider the full array of evidence and use it to guide an evidence-based approach to treatment [74].

Reference

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. Awad AG, Voruganti LN (2004) New antipsychotics, compliance, quality of life, and subjective tolerability—are patients better off? Can J Psychiatry 49:297–302
3. 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
4. Barak Y, Shamir E, Mirecki I, Weizman R, Aizenberg D (2004) Switching elderly chronic psychotic patients to olanzapine. Int J Neuropsychopharmacol 7:165–169
5. Bender S, Dittmann-Balcar A, Schall U, Wolstein J, Klimke A, Riedel M, Vorbach EU, Kuhn KU, Lambert M, Dittmann RW et al (2006) Influence of atypical neuroleptics on executive functioning in patients with schizophrenia: a randomized, double-blind comparison of olanzapine vs. clozapine. Int J Neuropsychopharmacol 9:135–145
6. Buckley PF (2007) Receptor-binding profiles of antipsychotics: clinical strategies when switching between agents. J Clin Psychiatry 68(suppl 6):5–9
7. Buckley PF, Harvey PD, Bowie CR, Loebel A (2007) The relationship between symptomatic remission and neuropsychological improvement in schizophrenia patients switched to treatment with ziprasidone. Schizophr Res 94:99–106
8. Caley CF, Weber SS (1995) Sulpride: an antipsychotic with selective dopaminergic antagonist properties. Ann Pharmacother 29:152–160
9. Cramer JA, Rosenheck R, Xu W, Thomas J, Henderson W, Charney DS (2000) Quality of life in schizophrenia: a comparison of instruments. Department of Veterans Affairs Cooperative Study Group on clozapine in refractory schizophrenia. Schizophr Bull 26:659–666
10. Dossenbach MR, Folnegovic-Smalc V, Hotujac L, Uglesic B, Tollefson GD, Grundy SL, Friedel P, Jakovljevic MM (2004) Double-blind, randomized comparison of olanzapine versus fluphenazine in the long-term treatment of schizophrenia. Prog Neuropsychopharmacol Biol Psychiatry 28:311–318
11. Essock SM, Covell NH, Davis SM, Stroup TS, Rosenheck RA, Lieberman JA (2006) Effectiveness of switching antipsychotic medications. Am J Psychiatry 163:2090–2095
12. Falkai P, Wobrock T, Lieberman J, Glenthoj B, Gattaz WF, Moller HJ (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
13. 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
14. Franz M, Lis S, Pluddemann K, Gallhofer B (1997) Conventional versus atypical neuroleptics: subjective quality of life in schizophrenic patients. Br J Psychiatry 170:422–425
15. Glick ID (2006) Understanding the results of CATIE in the context of the field. CNS Spectr 11:40–47
16. Hegerl U, Juckel G, Muller-Schubert A, Pietzcker A, Gaebel W (1995) Schizophrenics with small P300: a subgroup with a neurodevelopmental disturbance and a high risk for tardive dyskinesia? Acta Psychiatr Scand 91:120–125
17. Heinrichs DW, Hanlon TE, Carpenter WT Jr (1984) The Quality of life scale: an instrument for rating the schizophrenic deficit syndrome. Schizophr Bull 10:388–398
18. 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
19. Hofer A, Hummer M, Huber R, Kurz M, Walch T, Fleischhacker WW (2000) Selection bias in clinical trials with antipsychotics. J Clin Psychopharmacol 20:699–702
20. Hogarty GE, Goldberg SC (1973) Drug and sociotherapy in the aftercare of schizophrenic patients. One-year relapse rates. Arch Gen Psychiatry 28:54–64
21. Hugenholtz GW, Heerdink ER, Stolker JJ, Meijer WE, Egberts AC, Nolen WA (2006) Haloperidol dose when used as active comparator in randomized controlled trials with atypical antipsychotics in schizophrenia: comparison with officially recommended doses. J Clin Psychiatry 67:897–903
22. Jones PB, Barnes TR, Davies L, Dunn G, Lloyd H, Hayhurst KP, Murray RM, Markwick A, Lewis SW (2006) Randomized controlled trial of the effect on quality of life of second- vs first-generation antipsychotic drugs in schizophrenia: cost utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS 1). Arch Gen Psychiatry 63:1079–1087
23. Kane JM (2006) Commentary on the clinical antipsychotic trials of intervention effectiveness (CATIE). J Clin Psychiatry 67:831–832
24. Karow A, Naber D (2002) Subjective well-being and quality of life under atypical antipsychotic treatment. Psychopharmacology (Berl) 162:3–10
25. Kasper S, Winkler D (2006) Addressing the limitations of the CATIE study. World J Biol Psychiatry 7:126–127
26. Kupfer DJ, Sartorius N (2003) The usefulness and use of second-generation antipsychotic medication—an update. Curr Opin Psychiatry 16(suppl 1):S1–S44
27. Lambert M, Naber D (2004) Current issues in schizophrenia: overview of patient acceptability, functioning capacity and quality of life. CNS Drugs 18(suppl 2):5–17
28. Lavin MR, Rifkin A (1991) Prophylactic antiparkinson drug use: II. Withdrawal after long-term maintenance therapy. J Clin Pharmacol 31:769–777
29. Lehman AF, Kreyenbuhl J, Buchanan RW, Dickerson FB, Dixon LB, Goldberg R, Green-Paden LD, Tenhula WN, Boerescu D, Tek C et al (2004) The Schizophrenia Patient Outcomes Research Team (PORT): updated treatment recommendations 2003. Schizophr Bull 30:193–217
30. Lehman AF, Lieberman JA, Dixon LB, McGlashan TH, Miller AL, Perkins DO, Kreyenbuhl J (2004) Practice guideline for the treatment of patients with schizophrenia, 2nd edn. Am J Psychiatry 161:1–56
31. Leucht S, Pitschel-Walz G, Abraham D, Kissling W (1999) Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials. Schizophr Res 35:51–68
32. Lewis SW, Barnes TR, Davies L, Murray RM, Dunn G, Hayhurst KP, Markwick A, Lloyd H, Jones PB (2006) Randomized con-

trolled trial of effect of prescription of clozapine versus other second-generation antipsychotic drugs in resistant schizophrenia. Schizophr Bull 32:715–723
33. Lieberman JA (2006) Comparative effectiveness of antipsychotic drugs. A commentary on: Cost Utility Of The Latest Antipsychotic Drugs In Schizophrenia Study (CUtLASS 1) and Clinical Antipsychotic Trials Of Intervention Effectiveness (CATIE). Arch Gen Psychiatry 63:1069–1072
34. Lieberman JA, Alvir J, Geisler S, Ramos-Lorenzi J, Woerner M, Novacenko H, Cooper T, Kane JM (1994) Methylphenidate response, psychopathology and tardive dyskinesia as predictors of relapse in schizophrenia. Neuropsychopharmacology 11:107–118
35. 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
36. 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
37. Maier W, Moller HJ (2007) Meta-analyses: a tool for maximizing therapy study evidence?. Nervenarzt 78:1028–1036
38. Manos N, Lavrentiadis G, Gkiouzepas J (1986) Evaluation of the need for prophylactic antiparkinsonian medication in psychotic patients treated with neuroleptics. J Clin Psychiatry 47:114–116
39. McCue RE, Waheed R, Urcuyo L, Orendain G, Joseph MD, Charles R, Hasan SM (2006) Comparative effectiveness of second-generation antipsychotics and haloperidol in acute schizophrenia. Br J Psychiatry 189: 433–440
40. McEvoy JP, Lieberman JA, Stroup TS, Davis SM, Meltzer HY, Rosenheck RA, Swartz MS, Perkins DO, Keefe RS, Davis CE et al (2006) Effectiveness of clozapine versus olanzapine, quetiapine, and risperidone in patients with chronic schizophrenia who did not respond to prior atypical antipsychotic treatment. Am J Psychiatry 163:600–610
41. Meltzer HY, Bobo WV (2006) Interpreting the efficacy findings in the CATIE study: what clinicians should know. CNS Spectr 11:14–24
42. Miller DD, McEvoy JP, Davis SM, Caroff SN, Saltz BL, Chakos MH, Swartz MS, Keefe RS, Rosenheck RA, Stroup TS et al (2005) Clinical correlates of tardive dyskinesia in schizophrenia: baseline data from the CATIE schizophrenia trial. Schizophr Res 80:33–43
43. Möller HJ (2000) Rating depressed patients: observer- vs self-assessment. Eur Psychiatry 15:160–172
44. Möller HJ (2003) Amisulpride: limbic specificity and the mechanism of antipsychotic atypicality. Prog Neuropsychopharmacol Biol Psychiatry 27:1101–1111
45. Möller HJ (2003) Plazebo-kontrollierte Studien zum Wirkungsnachweis von Antidepressiva sind notwendig! Psychopharmakotherapie 10:85–86
46. Möller HJ (2004) Sind placebokontrollierte Studien zum Wirksamkeitsbeweis von Antidepressiva notwendig? Nervenarzt 5:421–424
47. 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
48. Möller HJ (2007) ''Effectiveness''-Studien: methodische Mängel und negative Konsequenzen für die Versorgung. Psychopharmakotherapie 14:45–46
49. Möller HJ (2008) Standardised rating scales in psychiatry: methodological basis, possibilities, limitations and descriptions of important rating scales. World J Biol Psychiatry (in press)
50. Moller HJ, Maier W (2007) Problems of evidence-based medicine in psychopharmacotherapy: problems of evidence grading and of the evidence basis for complex clinical decision making. Nervenarzt 78:1014–1027
51. Mudge MAC, Davey PJ, Coleman KA, Montgomery W, Croker VS, Mullen K, Castle DJ (2005) A comparison of olanzapine versus risperidone for the treatment of schizophrenia: a meta-analysis of randomised clinical trials. Int J Psychiatry Clin Prac 9:3–15
52. Mulert C, Juckel G, Giegling I, Pogarell O, Leicht G, Karch S, Mavrogiorgou P, Moller HJ, Hegerl U, Rujescu D (2006) A Ser9Gly polymorphism in the dopamine D3 receptor gene (DRD3) and event-related P300 potentials. Neuropsychopharmacology 31:1335–1344
53. Murray RM, O'Callaghan E, Castle DJ, Lewis SW (1992) A neurodevelopmental approach to the classification of schizophrenia. Schizophr Bull 18:319–332
54. Murray RM, Van Os J (1998) Predictors of outcome in schizophrenia. J Clin Psychopharmacol 18:2S–4S
55. Naber D (2005) Subjective effects of antipsychotic treatment. Acta Psychiatr Scand 111:81–83
56. Oosthuizen P, Emsley R, Jadri TH, Keyter N (2004) A randomized, controlled comparison of the efficacy and tolerability of low and high doses of haloperidol in the treatment of first-episode psychosis. Int J Neuropsychopharmacol 7:125–131
57. Oosthuizen PP, Emsley RA, Maritz JS, Turner JA, Keyter N (2003) Incidence of tardive dyskinesia in first-episode psychosis patients treated with low-dose haloperidol. J Clin Psychiatry 64:1075–1080
58. Paykel ES, Norton KRW (1986) Self-report and clinical interview in the assessment of depression. In: Sartorius N, Ban TA (eds) Assessment of depression. Springer, New York, pp 356–366
59. Phillips GA, Van Brunt DL, Roychowdhury SM, Xu W, Naber D (2006) The relationship between quality of life and clinical efficacy from a randomized trial comparing olanzapine and ziprasidone. J Clin Psychiatry 67:1397–1403
60. Pukrop R, Möller HJ, Steinmeyer EM (2000) Quality of life in psychiatry: a systematic contribution to construct validation and the development of the integrative assessment tool "modular system for quality of life". Eur Arch Psychiatry Clin Neurosci 250:120–132
61. Pukrop R, Schlaak V, Möller-Leimkühler AM, Albus M, Czernik A, Klosterkötter J, Möller HJ (2003) Reliability and validity of quality of life assessed by the Short-Form 36 and the modular system of quality of life in patients with schizophrenia and patients with depression. Psych Res 119:63–79
62. Pyne JM, Sullivan G, Kaplan R, Williams DK (2003) Comparing the sensitivity of generic effectiveness measures with symptom improvement in persons with schizophrenia. Med Care 41:208–217
63. Ragins M (2005) Should the CATIE study be a wake-up call? Psychiatr Serv 56:1489
64. Rosenheck R, Perlick D, Bingham S, Liu-Mares W, Collins J, Warren S, Leslie D, Allan E, Campbell EC, Caroff S et al (2003) Effectiveness and cost of olanzapine and haloperidol in the treatment of schizophrenia: a randomized controlled trial. JAMA 290:2693–2702
65. Rosenheck RA (2006) Outcomes, costs, and policy caution. A commentary on: the Cost Utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS 1). Arch Gen Psychiatry 63:1074–1076
66. Schatzberg AF, Nemeroff CB (2000) The American psychiatric publishing textbook of psychopharmacology, 3rd edn. American Psychiatric Publishing, Washington DC
67. Soares BG, Fenton M, Chue P (2000) Sulpiride for schizophrenia. Cochrane Database Syst Rev: CD001162
68. Soares KV, McGrath JJ (2000) Anticholinergic medication for neuroleptic-induced tardive dyskinesia. Cochrane Database Syst Rev: CD000204
69. Sramek JJ, Simpson GM, Morrison RL, Heiser JF (1986) Anticholinergic agents for prophylaxis of neuroleptic-induced dystonic reactions: a prospective study. J Clin Psychiatry 47:305–309
70. Stern TA, Anderson WH (1979) Benztropine prophylaxis of dystonic reactions. Psychopharmacology (Berl) 61:261–262
71. Stroup TS, Lieberman JA, McEvoy JP, Swartz MS, Davis SM, Rosenheck RA, Perkins DO, Keefe RS, Davis CE, Severe J et al (2006) Effectiveness of olanzapine, quetiapine, risperidone, and ziprasidone in patients with chronic schizophrenia following discontinuation of a previous atypical antipsychotic. Am J Psychiatry 163:611–622

72. Stroup TS, McEvoy JP, Swartz MS, Byerly MJ, Glick ID, Canive JM, McGee MF, Simpson GM, Stevens MC, Lieberman JA (2003) The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. Schizophr Bull 29:15–31
73. Strous RD, Kupchik M, Roitman S, Schwartz S, Gonen N, Mester R, Weizman A, Spivak B (2006) Comparison between risperidone, olanzapine, and clozapine in the management of chronic schizophrenia: a naturalistic prospective 12-week observational study. Hum Psychopharmacol 21:235–243
74. Tandon R, Möller HJ, Belmaker RH, Gattaz WF, Lopez-Ibor JJ, Okasha A, Burrows G, Stein DJ, Olie JP, Nutt D, et al. (2008) World Psychiatric Association Pharmacopsychiatry section statement on comparative effectiveness of antipsychotics in the treatment of schizophrenia. Schizophr Res (in press)
75. Tiihonen J, Walhbeck K, Lonnqvist J, Klaukka T, Ioannidis JP, Volavka J, Haukka J (2006) Effectiveness of antipsychotic treatments in a nationwide cohort of patients in community care after first hospitalisation due to schizophrenia and schizoaffective disorder: observational follow-up study. BMJ 333:224
76. Tunis SL, Johnstone BM, Gibson PJ, Loosbrock DL, Dulisse BK (1999) Changes in perceived health and functioning as a cost-effectiveness measure for olanzapine versus haloperidol treatment of schizophrenia. J Clin Psychiatry 60 (suppl 19):38–45
77. Tunis SR, Stryer DB, Clancy CM (2003) Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 290:1624–1632
78. Ware JE, Jr., Sherbourne CD (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 30:473–483
79. Ware JE, Snow KK, Kosinski M, Gandek B (1993) SF-36 Health Survey Manual and Interpretation Guide. The Health Institute, New England Medical Center, Boston
80. Weiden PJ (2006) Switching antipsychotics: an updated review with a focus on quetiapine. J Psychopharmacol 20:104–118
81. Weiden PJ (2007) Discontinuing and switching antipsychotic medications: understanding the CATIE schizophrenia trial. J Clin Psychiatry 68(suppl 1):12–9
82. Weiden PJ, Simpson GM, Potkin SG, O'Sullivan RL (2003) Effectiveness of switching to ziprasidone for stable but symptomatic outpatients with schizophrenia. J Clin Psychiatry 64: 580–588