Methodological Issues in Current Antipsychotic Drug Trials

Stefan Leucht1,2, Stephan Heres2, Johannes Hamann2, and John M. Kane3

2Department of Psychiatry and Psychotherapy, Klinikum rechts der Isar der Technischen Universität München, Ismaningerstr. 22, 81675 München, Germany; 3The Zucker Hillside Hospital, Kaufmann Building, 75-59 263rd Street, Glen Oaks, NY 11004

Every year numerous reports on antipsychotic drug trials are being published in neuropsychiatric journals, adding new information to our knowledge in the field. The information however is often hard for the reader to interpret, sometimes contradictory to comparable available studies and leaves more questions open than it actually answers. Although the overall quality of the studies is rather good, there are manifold options for further improvement in the conception, conduct, and reporting of antipsychotic drug trials. In this survey, we address methodological challenges such as the limited generalizability of outcomes due to patient selection and sample size; the vague or even lacking definition of key outcome parameters such as response, remission or relapse, insufficient blinding techniques, the pitfalls of surrogate outcomes and their assessment tools; the varying complex statistical approaches; and the challenge of balancing various ways of reporting outcomes. The authors present practical examples to highlight the current problems and propose a concrete series of suggestions on how to further optimize antipsychotic drug trials in the future.

Key words: schizophrenia/drug trials/methodology/antipsychotics

Introduction

Since decades, psychiatrists, researchers, statisticians, and others debate about how to best design, conduct, and report clinical trials on the treatment in schizophrenia (eg, Kraemer et al1, Bartko et al2, and Stroup et al3), and great efforts have been undertaken to set standards for these issues and to harmonize research for medicine in general (eg, The International Conference on Harmonization [ICH] of Technical Requirements for Registration of Pharmaceuticals for Human Use [www.ICH.org], CONSORT4) or for psychiatric topics (eg, Food and Drug Administration [FDA]-National Institute of Mental Health-Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) workshop on clinical trial design for neurocognitive drugs for schizophrenia5). Likewise, enormous efforts have been invested in clinical trials on the treatment of schizophrenia, yielding new insights into this demanding task.

Evidence taken for granted nowadays, such as, eg, defined dose ranges for the use of antipsychotics was not available during the decades of treatment with the first-generation drugs. Beyond the study of efficacy in the treatment of positive symptoms, increasing value has been placed on negative and depressive symptoms as well as quality of life and the social functioning of the patients, aspects mostly neglected in previous clinical trials. Apart from these advances, there is still need for additional steps to further improve the validity of trial data as well as the quality of the reporting. Study endpoints such as response or relapse indeed lack a commonly accepted definition.6 Statistical approaches vary considerably, leaving room for speculation whether their choice was partly dependent on the desired outcome.7 The influence of the study sponsor is obvious in head-to-head trials of second-generation antipsychotics (SGAs) reporting a prosponsor overall outcome in 90% of the publication abstract.8 The latest landmark study, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) trial, though independently sponsored and carefully designed, stimulated once again the discussion of how clinical trials should be conducted, analyzed, and reported.9 In this context, the current review tries to address a selection of critical aspects, including trial design features, population sample bias, study endpoint choice and definition, outcome assessment and reporting, and other sources of potential bias relevant to clinical trials (table 2).

Results

Participants

Are Current Trial Populations Representative of the Majority of Patients Receiving Routine Care? The most important problem about the selection of participants...
that are currently included in antipsychotic drug trials is whether the results can be generalized to routine care. Many randomized double-blind studies have narrow inclusion criteria and exclude patients with substance abuse, physical diseases, or suicidality, although these problems are frequently encountered in daily practice. Hence, it was shown that up to 90% of patients with a diagnosis of a schizophreniform or schizophrenic disorder generally suitable for a clinical trial do not qualify for study participation because they do not meet the elaborate inclusion and exclusion criteria, a fact that severely impairs the generalizability of the results of the trial.10,11 On the other hand, the inclusion of patients with substance abuse or physical illnesses would of course demand increased sample sizes comprising large enough subgroups so that stratification can illustrate the potential impact on drug efficacy or side effect rates. Admittedly, it is still unclear whether studies with broader inclusion criteria such as CATIE do really solve the problem because the patients must still give informed consent and many will not be willing to do so, eg, because of the double-blind treatment nature of a trial. Many highly symptomatic patients are simply too ill to give informed consent, but for many clinical questions these are precisely the kind of patients who need to be studied in order to inform clinical practice. High levels of agitation as inclusion criterion would, eg, be very important for assessing the effects of intramuscular medication strategies. A number of SGAs are available as intramuscular formulations,12,13 but by definition the double-blind studies examining their effects demanded informed consent and thus excluded in fact those patients for whom these medications are intended: the aggressive patient who is not willing to take medication. Notable exceptions, such as the Tranquilização Rápida-Ensaio Clinic trials on sedation strategies for acutely psychotic patients, exist in which relatives could provide the informed consent.14,15 We feel that within the present ethical and legal context of randomized controlled trials (RCTs) it will be difficult to overcome these problems. Currently, a feasible approach in our eyes may be to consider methodologically less stringent naturalistic studies to help corroborate the results obtained from rigorous RCT. Combining both sources of evidence could enable us to extrapolate patients’ clinical benefit best.

Patient Chronicity and Lack of a Definition for an “Acute Exacerbation.” Both the duration of the disease—chronic course vs recent onset—and the acuity of the current episode—acute exacerbation vs symptomatic nonacute state—are issues in clinical trials. The participants in most recent RCTs on schizophrenia were rather chronic, ie, they were in their mid- to late 30s, had been ill for 10–15 years, and had already experienced a number of episodes and hospitalizations.16 On the one hand, such patients may actually be quite “typical” of daily clinical routine. On the other hand, there is an interest in the effects of medication on people with schizophrenia who have just recently been diagnosed. Many differences are evident between recent onset and chronic schizophrenia regarding morphological changes in the brain, consecutive cognitive impairment, level of compliance, or sensitivity to side effects and may result in different outcomes in clinical trials for the respective patient subgroup.17–19 In clinical trials with large enough sample sizes including a considerable number of patients with a recent onset of schizophrenia, the outcome measurements could be stratified for this patient group. Alternatively, studies explicitly recruiting these patients could be conducted.

When it comes to acuity, things become even more complicated. Some studies focus on patients in an acute exacerbation; but there is simply no definition as to what an acute exacerbation is. It seems that all patients with a certain predefined minimum of symptoms meet the criterion irrespective of the literal acuity in terms of the duration or dynamic of the current episode. A definition comprising the duration since onset of the episode and a certain degree of worsening in symptoms would be required to confirm and describe the exacerbation. Currently, participants often enter the trials after having been hospitalized for several weeks as long as they are still symptomatic. Many patients have been pretreated with antipsychotics before they entered the studies after short wash-out periods of a few days. Such a “stabilization” may also in part explain why in a recent meta-analysis the difference in responder rates between drug and placebo was only 18% (Leucht et al20), although the patients were “symptomatic” at the time of inclusion. Partial exceptions are the few first episode studies that have recently been published, but even most of these allowed for some pretreatment prior to inclusion.21,22

Lack of Quality Data on Specific Symptoms and Patient Subgroups. Another issue is the examination of specific symptoms or symptom cluster. For this approach, trials have to define highly selective inclusion criteria to allow an outcome interpretation for the exclusive patient population affected by the symptom cluster. The classical example is the treatment of negative symptoms. Although SGAs are marketed as being especially efficacious for negative symptoms, good evidence to support this hypothesis is scarce.16 Most RCTs are conducted on patients with predominantly positive symptoms. Any superiority in terms of negative symptoms may be secondary to the effects on positive symptoms. For example, a patient suffering from hallucinations and delusions may isolate himself, suggesting negative symptoms, and if his positive symptoms are successfully treated, these negative symptoms will secondarily improve, as well. Or, when a SGA is compared with a first-generation antipsychotic (FGA) given in high doses, the extrapyramidal side effects (EPS) induced by the FGA will look like (secondary) negative symptoms and thus inflate the
apparent superiority of the SGA. Therefore, studies with lower doses of high-potency antipsychotics or with less EPS inducing FGAs have been required, and some have now been published. Amisulpride is still the best-examined SGA as regards negative symptoms because several studies in patients with predominantly negative symptoms have been carried out with this compound. Even these studies were not ideal because some positive symptoms were present and so the amelioration of negative symptoms may as well be partially a result of decreased hallucinations, delusions, or thought disorders. Similar studies of other agents are even more scarce. Besides the sole presence of primary negative symptoms their persistence over time, both comprised in the deficit syndrome, is interesting but the latter syndrome is studied exclusively in antipsychotic drug trials rarely.

In a similar vein, reviews suggest superior effects of SGAs on aggressive behavior. But to the best of our knowledge, only one recent RCT compared clozapine, olanzapine, and haloperidol in high-risk patients with a history of aggressive behavior to confirm the hypothesis. As in negative symptoms, trials on aggressive behavior must define their inclusion criteria based on specific aggression scales to generate evidence for this symptom cluster. In terms of suicide prevention, the Intercept study comparing clozapine and olanzapine is a good example of a trial looking at specific symptoms of the disease in appropriate patients. Most participants in antipsychotic drug trials suffer from paranoid schizophrenia. Whether these results can be applied to other diagnostic subtypes such as catatonic schizophrenia is yet questionable because this subtype is by nature rarely encountered. Large multicenter studies are unlikely to be successfully performed, and so a meta-analytic approach might be helpful in gathering evidence for this patient subgroup.

**Design Aspects**

**Small Study Sizes and Problems of Large Industry-Sponsored Trials.** Thornley and Adams reported that the vast majority of randomized antipsychotic drug trials do not include more than 60 participants (figure 1). These low sample sizes are problematic because on the one hand generalizability is questionable and on the other hand many of the studies may have simply been too low powered to show statistically significant differences. In a recent analysis comparing antipsychotic drugs with placebo, we found an average effect size of only 0.50 (Leucht et al). It is also clear that in head-to-head comparisons of antipsychotic drugs or augmentation strategies in which the adjunct is not efficacious as monotherapy, dramatic differences are unlikely and large sample sizes are necessary to show modest effects.

The mentioned limitations of studies with small sample sizes could possibly be overcome by large industry-sponsored studies for which the necessary resources are available to conduct trials on a multicenter, multinational level. On the other hand, a major problem with these industry studies is that there may be centers that primarily participate for financial reasons, potentially adding more quantity than quality of data to the trial. Indeed, the SDs of rating scale scores in such industry-sponsored studies are often large, reflecting a high variability of patients (see, eg, reviews of the Cochrane schizophrenia group); and this may be in contrast to smaller “academic” studies in well-selected samples with clearly defined characteristics (however, academic studies are surely not “immune” to problems of selection bias, poor rater performance, and misaligned financial incentives). In many studies, there can be problems such as inflating of the baseline ratings to meet a study’s severity inclusion criterion. This may explain why in a recent analysis linking the Brief Psychiatric Rating Scale (BPRS)/Positive and Negative Syndrome Scale (PANSS) total scores to the CGI-severity score the baseline result was slightly different from that of later weeks. Or, in a recent RCT, only 2% percent of the participants reached an at least 20% PANSS reduction in the first 4-week run-in phase using olanzapine or risperidone so that almost all participants qualified for the double-blind core phase comparing aripiprazole with perphenazine. In the latter, 26% reached the same criteria. It is also possible that many ratings are made quickly and without attention to detail under such circumstances. One potential solution to these problems may be to use remote, independent raters connected by telemedicine. Finally, the methodological question whether the meta-analytic approach of combining many small but sound studies yields evidence of more reliable quality than the sole conduction of few large industry-sponsored trials has to remain unanswered.

**Short Wash-out Periods.** Due to ethical problems, the wash-out phases of recent studies have been relatively...
short (eg, the median minimum wash-out duration in 37 placebo controlled antipsychotic drug trials was 3 days (Leucht et al \textsuperscript{20}), whereas there are many examples of studies in the 1960s and 1970s in which the washout often lasted several weeks. The problem of short wash-out phases lies in the carryover effects from previous treatment. Withdrawal effects of previous treatment after short washout may explain why the EPS rates of some SGAs were even lower than those of placebo in the initial registrational studies.\textsuperscript{16} Indeed, in mania studies, where most patients were probably pretreated with antipsychotics, differences in EPS between some SGAs and placebo were observed, while there were no such significant EPS differences in the schizophrenia studies.\textsuperscript{43} Even more problematic is the previous use of depot medication. Usually there must be a gap of at least one injection interval, but due to the long half-life of depots it is unrealistic to expect the drugs to be fully washed out. The key problem remains that the risk of carryover effects cannot be estimated precisely. To our knowledge, very limited data on the persistence of side effects or occurrence of withdrawal effects after the discontinuation of a drug are available. Thus, recommendations on the duration of wash-out phases have to remain speculative. To control for foreseeable interactions, the pharmacokinetic wash-out phase covering 5 half-lives in duration could be considered a basic approach.

Imperfect Blinding. We have been told by patients participating in industry-sponsored studies that they opened the medication capsules and found out what medication they were on. Also, it may often be easy to guess the randomly assigned treatment arm if one of the drugs has strong side effects such as EPSs, sedation, or weight gain. Prophylactic antiparkinson medication has been used to avoid the EPS problem, but these medications again have their own side effects.\textsuperscript{44} Although haloperidol was a standard treatment, justifying its use as a comparator in most comparisons with SGA, its high risk for EPS made it easy for the SGA to be better in this regard. Therefore, comparisons with other old compounds with a lower EPS propensity such as sulpiride, thioridazine would be useful, and in this context the choice of a modest dose perphenazine as the comparator in the CATIE trial was a reasonable one.\textsuperscript{9} A meta-analysis suggested that low-potency conventional antipsychotics are not associated with more EPS than SGA as long as the former are used in doses lower than 600 mg/day.\textsuperscript{16} In conclusion, whenever side effect profiles of study drugs are highly diverse, blinded assessment of outcomes by an independent rater (eg, telemedicine) are recommended to maintain a minimum level of blinding.\textsuperscript{4} Additionally probing questions in which participants and investigators must make a guess as to the actual treatment at the end of clinical trials could help detecting systematic blinding problems.

Dosage. An appropriate choice of doses is essential and thus also demanded by guidelines for drug trial conduction (eg, FDA Guideline for Industry E10\textsuperscript{45}). However, choosing fair dosages of, eg, comparator drugs is more difficult than one might imagine. For example, the haloperidol doses used in pivotal studies on SGAs have been criticized as generally being higher than the recommendations in guidelines.\textsuperscript{40} However, the ideal haloperidol dose has not been well established. An important study suggested that neuroleptic threshold doses between 2 and 3 mg/day may be sufficient,\textsuperscript{47} while another study suggested a dose response curve between 5 and 20 mg/day.\textsuperscript{48} Another example is the manufacturer recommending an optimum risperidone dose of 4–6 mg/day and not using higher doses in its own recent trials.\textsuperscript{24,49} Yet, the CATIE study has been criticized for using too low a dose of risperidone (2–6 mg/day). These examples show how difficult it can be to get consensus on what dosage to use. But there are also examples in which the optimum doses were clearly missed, such as a study in which ziprasidone could be given in the full dose range (80–160 mg/day) while olanzapine was restricted to 5–15 (instead of 20) mg/day.\textsuperscript{50} Likewise, in another trial, ziprasidone was given at full dose (80–160 mg/day) while amisulpride was given in low doses (50–200 mg/day).\textsuperscript{51} There will always be discussion about the doses in clinical trials, but a plausible justification in the method section should be mandatory.

Switching of Treatments in Long-Term Studies. Recent cost-effectiveness studies allowed switching to the compound of the comparator group after randomization.\textsuperscript{52–54} Although this strategy is in conformance with the “intention-to-treat” (ITT) principle and reduces the otherwise high dropout rates, it is nevertheless problematic in terms of estimating the efficacy and effectiveness of new antipsychotics. This approach may, eg, make sense for expensive chemotherapies to treat cancer, but in antipsychotic drug trials it is somewhat as in a car race where one driver starts in a Porsche and the other one in a Golf, but the Golf driver soon catches up by changing to the Porsche (of course this comparison is valid only if the atypical antipsychotics really are Porsches and not only Jetta). A skepticism about this approach remains, although some of these studies at least analyzed those patients who did not switch groups in a sensitivity analysis.\textsuperscript{44,54} For further details on the use of the ITT approach and its limitations, please refer to the comprehensive review by Lavori.\textsuperscript{55}

Outcomes

The definition of outcomes is problematic in psychiatric studies and will be discussed in the following. However, apart from the definition itself, the basic problem of numerous clinical trials is the performance of multiple
outcome measures. The determination of a single primary outcome parameter should be viewed as mandatory, as requested in the guidelines of the ICH (www.ICH.org) and adopted by the FDA\textsuperscript{45} and the European medicines evaluation agency (http://www.ema.europa.eu/htms/human/ich/sichefficacy.htm). Additional secondary outcome parameters may yield supplementary data that can be used as basis for future trial planning. The analysis and presentation of these secondary outcomes is often precarious. A potential loss of sensitivity results from the necessary correction for multiple comparisons and so often this correction step is avoided, an approach that is highly debatable.

**Lack of “Hard Outcomes” and Difficult-to-Interpret Rating Scales.** An essential problem in schizophrenia treatment research is a lack of hard outcomes. This may be one reason why there is much less debate in other specialties about new treatments, although the magnitude of their effects is much smaller. This is illustrated by the fact that a number needed to treat (NNT) to avoid a vascular event or death by statins is greater than 100,\textsuperscript{56} while the NNT to avoid one relapse by treating participants with an atypical antipsychotic instead of a typical drug was 13. The internists are talking about a hard outcome, death, while we are focusing often on the reduction of the baseline scores of rating scales. Here, the PANSS\textsuperscript{57} and the BPRS\textsuperscript{58} are probably the most frequently used instruments in clinical trials. Both have appropriate psychometric properties, but of course their usefulness is limited by the reliability of raters. Various authors have addressed this problem (eg, Perkins et al\textsuperscript{59}), but actions for improvement of reliability (eg, more intensive rater training, use of multiple raters) have not been implemented fully in research practice. Recent studies have attempted to provide some anchors as to what a given total score on these scales mean in terms of Clinical Global Impression (CGI).\textsuperscript{39,60} Nevertheless, the BPRS does not cover negative symptoms well, and a PANSS interview takes 45 minutes and is thus often difficult to be carefully completed/conducted in busy clinical settings. In the context of the now frequent pragmatic and practical trials (see the article by Stroup and Geddes in this issue), easier-to-rate scales would be useful. One such scale is the CGI Scale\textsuperscript{61} that—although it is used in almost every clinical trial—has never been evaluated well. A recent analysis found that it is as sensitive as the BPRS in detecting between drug differences in clinical trials given the limitations that the ratings were not carried out independently from the BPRS.\textsuperscript{62} Other problems are that it covers only overall symptoms but not specific symptoms and that there are no anchors precisely defining the steps of the scale. Recently, a better anchored version has been developed for schizophrenia, and its psychometric properties have been evaluated.\textsuperscript{63} This version also allows for the assessment of specific symptoms (positive, negative, depressive, and cognitive symptoms), and it has successfully differentiated between the efficacy of various SGAs in a large-scale naturalistic study.\textsuperscript{63} We believe that the new CGI may be an excellent scale for pragmatic trials because it is intuitive and easy to use.

**Lack of Uniformly Applied Definitions of Response and Remission.** Dichotomized measures of response and remission can be understood more intuitively by clinicians than a mean value of a rating scale. Antipsychotic drug trials have used a wide variety of cutoffs in terms of PANSS/BPRS reduction to define response (eg, at least 20%, 30%, 40%, 50%). There is now replicated evidence that the one that was most frequently used in the last decade—at least 20% BPRS/PANSS reduction—hardly means “minimally better” from a CGI point of view.\textsuperscript{39,60,64} Acutely ill, nonrefractory patients with schizophrenia usually do respond well to antipsychotic drugs.\textsuperscript{65,66} We therefore suggest that in such populations, the cutoff at least 50% BPRS/PANSS reduction (reflecting “much improvement”) may be more informative. At least 25% BPRS/PANSS reduction may be an appropriate primary cutoff in refractory populations. Instead of showing only one cutoff, the results should be presented in a table showing the distribution of response according to different cutoffs including the new remission criteria (see table 1).

A problem with all these response measures is that, eg, a patient with an initial PANSS score of 150 and an endpoint score of 90 had a 50% reduction (calculated after subtracting the 30 minimum PANSS points), whereas he is still markedly ill. For this and other reasons, remission criteria in schizophrenia have recently been introduced. According to these criteria, a patient is in symptomatic remission if 8 predefined PANSS items have been only

### Table 1. A Simple Table for the Presentation of Responder and Remission Rates in Schizophrenia

<table>
<thead>
<tr>
<th></th>
<th>Total n</th>
<th>≤0% PANSS/BPRS reduction n (%)</th>
<th>0%–24% PANSS/BPRS reduction n (%)</th>
<th>25%–49% PANSS/BPRS reduction n (%)</th>
<th>50%–74% PANSS/BPRS reduction n (%)</th>
<th>75%–100% PANSS/BPRS reduction n (%)</th>
<th>Remission n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention group</td>
<td>N</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Control group</td>
<td>N</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
</tbody>
</table>

**Note:** PANSS, Positive and Negative Syndrome Scale; BPRS, Brief Psychiatric Rating Scale.
<table>
<thead>
<tr>
<th>Problem</th>
<th>What may be done about it?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study population not representative for the majority of patients in daily clinical practice</td>
<td>Broader inclusion criteria; more pragmatic trials; corroborate the findings in RCTs by naturalistic studies</td>
</tr>
<tr>
<td>Lack of universal definitions for “acute exacerbation,” “relapse,” “response,” or “remission”</td>
<td>Consensus statements on commonly accepted definitions for the respective terms</td>
</tr>
<tr>
<td>Varying thresholds in PANSS/BPRS reduction for the outcome “response”</td>
<td>Use of 50% PANSS/BPRS reduction as threshold for response in acutely ill patients and of 25% in refractory patient populations; ideally display the distribution of patients meeting staged levels of PANSS/BPRS reduction and remission criteria (see table 1)</td>
</tr>
<tr>
<td>Limited evidence on specific aspects of symptomatology (eg, negative symptoms, aggression) and diagnostic subtypes</td>
<td>Careful selection of patients with target symptoms; conduct studies in specific diagnostic subtypes</td>
</tr>
<tr>
<td>Small sample sizes</td>
<td>Multicenter trials</td>
</tr>
<tr>
<td>High dropout rates</td>
<td>Ways of reducing dropout rates must be found; participants should be followed up even after dropping out of the study</td>
</tr>
<tr>
<td>Poor quality of ratings</td>
<td>Assessments by independent raters (not the therapists), eg, using telemedicine</td>
</tr>
<tr>
<td>Carryover effects or withdrawal effects from pretrial medication</td>
<td>Extension of wash-out periods, special rules for preceding depot treatment</td>
</tr>
<tr>
<td>Insufficient blinding</td>
<td>Use of prophylactic anti-parkinson medication; probing questions at study end to detect insufficient blinding</td>
</tr>
<tr>
<td>Dosing of trial drugs</td>
<td>Precursory obtaining of recommendations on dosing by the manufacturer of the study drugs; justify dosage in method section</td>
</tr>
<tr>
<td>Intervention groups can be switched in some long-term trials</td>
<td>Separate analysis of patients not switching the assigned treatment group</td>
</tr>
<tr>
<td>Lack of hard outcomes; difficult to interpret rating scales</td>
<td>Use intuitive, easy-to-use scales like the improved CGI(^2), development of further scales may be warranted</td>
</tr>
<tr>
<td>Little uniformity on the use of rating scales on aspects beyond symptoms</td>
<td>Consensus on relevant domains and measurement instruments must be achieved</td>
</tr>
<tr>
<td>Analysis of the time component of the new remission criteria in the context of high dropout rates</td>
<td>Alternatives to LOCF (inadequate approach, likely to overestimate remission rates) are “worst case” approaches (likely to underestimate remission rates) in addition to completer analyses, but better solutions are needed</td>
</tr>
<tr>
<td>High dropout rates have the potential to distort outcomes</td>
<td>Mixed-effect models appear to be statistically superior to LOCF but are not intuitive; alternative strategies such as the combination of the results of dropout rates and scale-derived efficacy results should be invested; analyses based on the study completers are useful sensitivity analyses</td>
</tr>
<tr>
<td>Quality of reporting of results is heterogeneous</td>
<td>More effective control of the reporting quality by reviewers and editors of scientific journals; special focus on abstract content as most commonly used source of information on a trial; completeness of the reporting of outcomes must be assured; strict application of the CONSORT statement; a description of the randomization method, blinding, and dropouts should be essential components of any publication</td>
</tr>
<tr>
<td>Incomplete side effect reporting</td>
<td>Common standard for the reporting of side effects must be defined; showing only those adverse events that occurred in at least 5% of the cases should be abandoned, all important adverse events should be reported instead, eg, as supplemental material on the web; “dropouts due to adverse event” should be restricted to side effect related dropouts and exclude efficacy-related events. Only then would it be a good measure of overall tolerability</td>
</tr>
</tbody>
</table>
mildly present for at least 6 months. Thus, in contrast to response criteria, the remission criteria indicate a relative absence of symptoms. Their limitation is that they do not reflect the quantity of change. If the study population was on the average only mildly ill at the beginning, many patients will still be in remission at the end of the study. A solution could be to indicate both—remission and responder rates—as in table 1. Such a procedure has been common practice for more than a decade in trials of antidepressants.

Problems in Applying the New Remission Criteria. While the first studies show that the new remission concept is an achievable and clinically valid measure that can differentiate between patients with a good and a poor outcome, the analysis of its time component is difficult, especially in the light of the high dropout rates in recent clinical trials. The last observation carried forward (LOCF) approach does not make sense here. Assume that in a 1-year trial, a patient reaches symptomatic remission and then drops out before reaching the 6-month threshold. Using LOCF, the patient would then be in remission at 1 year, but this would simply ignore the time criterion. Up to now, studies have either not applied the time criterion or used worst case approaches or only shown results based on study completers. In the light of very high dropout rates, these worst case approaches most likely underestimate the true remission rates.

It also needs to be defined how often patients should be assessed to ensure that they were in remission for at least 6 months. Daily or weekly assessments are generally not practical. Monthly measurements are more feasible, and a possibility would be for the rating to cover the last month for the remission assessment, although this would be in contrast to the traditional rating guidelines for the PANSS (judging the patient in the last week).

Because the new remission criteria still assess patients on a symptomatic level, a big challenge for the coming years will be to define recovery criteria for schizophrenia. Recovery goes beyond the relief of symptoms and should cover aspects such as quality of life as well as social and vocational functioning. Some suggested criteria are already available, but it may be difficult to weigh the different components against one another in different environments. For example, it may be more difficult for patients to obtain a job in industrialized countries than in a developing country.

Definition of Relapse. There is no uniformly accepted definition for “relapse.” For example, in a recent review comparing the relapse prevention potential of atypical antipsychotics with that of typical antipsychotics and placebo, 11 different criteria were used in 17 studies. Those studies that applied the same criteria were usually organized by the same pharmaceutical company. Some of the criteria such as “hospitalization for psychopathology” are pragmatic and intuitively meaningful, but the problem is that whether a patient is hospitalized or not will depend on the treating psychiatrist, the health system, and many other factors ranging from the psychosocial to the economic. Other criteria are much more sophisticated and complex (eg, increase in CGI rating and >1 increase in 2 BPRS positive items for 3 days, or the same level of deterioration for 24 hours) and requiring hospitalization, or a CGI rating of severely ill for 24 hours but they are less intuitive and much more difficult to apply. Again, other criteria included symptoms such as suicidal ideation that are not specific to schizophrenia or defined relapse as a percentage increase of the BPRS from baseline. A consensus on more universal measure would contribute to the comparability of long-term studies.

Assessment of Aspects Beyond Psychopathology. The assessment of effects other than pure symptoms is even more difficult. For example, multiple tests are available to assess a variety of cognitive domains. There is a hope that the MATRICS initiative will standardize the measurement and reporting of cognitive effects. Quality of life is another complex area. There are schizophrenia-specific scales and generic instruments which all have limitations (for review see Lambert et al). Given these measurement problems and the fact that quality of life is difficult to conceptualize, the future focus may be more on how patients feel under treatment with specific antipsychotic drugs. Scales such as the Drug Attitude Inventory or the “Subjective Well-Being Under Neuroleptics Scale” are available for this purpose. Social functioning has also attracted interest as an outcome in clinical trials. A review identified 87 potentially relevant measures that varied widely in terms of measurement approach, number and types of domains covered, and scoring systems and asked for a consensus.

Problems Related to the Analysis of Antipsychotic Drug Trials

A review of all statistical problems in schizophrenia trials is beyond the scope of this article. But there is an important recent debate on how we should cope with high dropout rates or early discontinuation rates in randomized antipsychotic drug trials. In the CATIE study, the 78% overall discontinuation rate of the initially assigned drug treatment after 18 months only confirmed the results of 4- to 10-week short-term trials in which over 40% often discontinue prematurely. The LOCF method is a simple imputation method using the last observation before the patient discontinues as an endpoint. While LOCF was the preferred approach in the 1990s, the assumption that the dropouts would not have changed if they had stayed in the study can obviously be wrong. In addition, the assumption about the “missingness” characteristics in LOCF is quite tight (for a definition
of the terms “missing completely at random,” “missing not at random,” “missing at random,” see Mallinckrodt et al and Leon et al. The advantage of the recently preferred “mixed-effects models” is that in contrast to LOCF they make use of all the data (and not merely the endpoint data) and that the assumption about the “missingness” is more relaxed. The most important disadvantage of “mixed-effects models” is that there are many possible variations of these statistics, making them relatively nontransparent, less intuitive, and open to manipulation. Combining the results of dropout rates and scale-derived efficacy results is another approach that has recently been suggested and deserves further investigation. It has been suggested to present completer-only analyses as secondary sensitivity analyses. Although the assumptions necessary for completer analyses (completely at random) are usually not met in such trials disqualifying them as primary approaches, still the question of “what happens to those patients who completed a full course of treatment” is of interest compared with the modeled results.

But even the best statistical method can probably not fully account for a situation in which, eg, 50% of the initial population does not reach the study’s end. Ways of reducing these rates must be found. Furthermore, those participants who dropped out of the double-blind phase of a trial still need to be followed up and their results presented. It is essential to know what happens to these people, but at least the industry-sponsored studies usually make little effort to gather follow-up data.

Unsatisfactory Reporting

Selective Reporting of Outcomes. In a world in which, according to estimates, 2 million articles are published in 10 000 scientific medical journals, the reporting of the results of RCTs must be of consistent high quality. In a recent trial showing no difference between ziprasidone and amisulpride in terms of overall and negative symptoms, in another study comparing quetiapine with haloperidol, the PANSS total score and the PANSS positive score were presented, but not the PANSS negative score, although the readers would be interested in negative symptoms in a study of an atypical antipsychotic. Information on EPSs is not very helpful if the amount of antiparkinson medication used is not stated. If in a study the significant superiority of aripiprazole compared with perphenazine in terms of participants with at least 20% quality of life score improvement is highlighted, whereas the mean quality-of-life score tended to favor perphenazine (not significant), the inexperienced reader will not realize the distinction.

Lack of Reporting of Essential Factors (Randomization and Blinding Method, Measures of Variability). Essential aspects of trial reporting are far less than satisfactory. It had been established long before the CONSORT statement that mean values must be presented together with a measure of variability such as the SD. Very often this is not the case. For example, almost all pivotal aripiprazole studies have been reported without SDs, even in high ranking journals such as the Archives of General Psychiatry. A correct randomization method and good blinding are key quality indicators of RCTs, but details on the methods applied are frequently not given. Thus, it is often unclear whether a study stated to be randomized used an appropriate approach or rather quasi randomization such as allocation according to the week or alternate allocation, methods which are imperfect. Journal editors need to make sure that the CONSORT statement is applied.

Side Effect Reporting. Papanikolaou et al found that the side effect reporting in psychiatric trials is poor. Reports usually reveal only those adverse events that occurred in at least 5% or sometimes even 10% of the participants. By virtue of this selection, rare, but sometimes particularly serious side effects, such as seizures or agranulocytosis, are not reported. Although the individual studies themselves may be too small to find statistically significant differences between groups, meta-analyses could combine the results of different studies and increase statistical power. Also death—the hardest possible outcome—should always be indicated, but we doubt whether this is always the case. The outcome dropout due to adverse events could be a very good measure of overall tolerability. Unfortunately, efficacy-related events such as exacerbation of psychosis are also included in this category of adverse events, making it a mixture of tolerability and efficacy. Only tolerability-related adverse events should be listed in this category to make dropout due to adverse events a consistent and useful side effect measure. Overall, side effect reporting needs to be more comprehensive and standardized. Most journals now allow for online supplements where detailed information could be stored for retrieval without inflating article length.
Conclusion

Modern clinical trials in schizophrenia use sophisticated methodology, but this selective review showed that there is also a lot of room for improvement (table 2). While some of the problems could be solved easily, joined efforts of investigators in clinical trials, sponsors, authors, peer reviewers, and journal editors together with organizations such as the “International Society for CNS Clinical Trials and Methodology” (http://joincpa.com/) will be necessary to address the more complex ones.

Acknowledgments

No source of funding or any grant was used to finance this study. Stefan Leucht has received honoraria and/or research support from Abbott, AstraZeneca, Bristol-Myers Squibb, Sanofi-Aventis, Eli Lilly, Janssen-Cilag, Johnson & Johnson, Pfizer Inc., and Lundbeck. Stephan Heres has received honoraria from Abbott, AstraZeneca, Eli Lilly, Janssen-Cilag, Sanofi-Aventis, Pharmastar, and Johnson & Johnson. Dr Heres has accepted travel or hospitality payment from Janssen-Cilag, Sanofi-Aventis, AstraZeneca, Bristol-Myers-Squibb, Eli Lilly, Janssen, Johnson & Johnson PRD, Otsuka, Pfizer Inc., Wyeth, Lundbeck, Vanda, Astra-Zeneca, and PGxHealth.

References


44. Jones PB, Barnes TRE, Davies L, et al. 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 (CuiLASS 1). *Arch Gen Psychiatry*. 2006;63:1079–1086.


83. Davis JM, Chen N, Glick ID. Issues that may determine the outcome of antipsychotic trials: industry sponsorship and extrapyramidal side effect [published online ahead of print 2007]. Neuropsychopharmacology.


