Effectiveness as an outcome measure for treatment trials in psychiatry

W. Wolfgang Fleischhacker[1], Guy M. Goodwin[2]

[1]Department of Psychiatry and Psychotherapy, Biological Psychiatry Division, University of Innsbruck, 6020 Innsbruck, Austria
[2]University Department of Psychiatry, Warneford Hospital, Oxford, OX3 7JX, UK

There is at present some confusion about the relative value of clinical trials performed to investigate efficacy vs. those designed to investigate effectiveness. This is particularly challenging when studies performed as experiments for regulators by companies are used to shape and inform clinical practice, especially if studies conducted under more real life conditions fail to support predicted benefits. We review the field in relation to the new antipsychotics, in particular. Other indications, including mood disorders, which are also briefly touched upon, have so far received less definitive attention, but are likely to encounter the same difficulties. We conclude that, where the results of efficacy trials are positive and an effectiveness trial is negative, one should not necessarily prefer the effectiveness trial – it may simply have failed. Where efficacy trials and effectiveness trials point to similar conclusions, then the findings are mutually supportive.

Key words: Clinical trials, methodology, schizophrenia, mood disorders, bipolar disorder, depression, antipsychotics, antidepressants, mood stabilizers, efficacy, effectiveness, pragmatic trials

(World Psychiatry 2009;8:23-27)

In recent years, our field has experienced a growing difficulty in translating the results of randomized controlled clinical trials (RCTs) into clinical practice concerning the clinical usefulness of new medications for the treatment of schizophrenia and mood disorders. This difficulty has been accentuated by the fact that meta-analyses and systematic reviews have often delivered discrepant messages. For instance, Leucht et al (1), following a meta-analysis of RCTs comparing first-generation to second-generation antipsychotics, concluded that "risperidone and olanzapine are more effective than haloperidol against global symptomatology and negative symptoms" and that all tested second-generation antipsychotics cause less extrapyramidal symptoms and lead to lower use of anticholinergics. However, analysing more or less the same data set, Geddes et al (2) came to the conclusion that "there is no clear evidence that atypical antipsychotics are more effective or are better tolerated than conventional antipsychotics". Davis et al (3), analysing data from 142 studies, suggested that some second-generation antipsychotics but not others show superior efficacy over the traditional medications, while Tandon and Fleischhacker (4), on the basis of a qualitative review of the available evidence, concluded that "meta-analytic studies of the comparative efficacy of non-clozapine second-generation antipsychotics do not provide undisputed evidence of differential efficacy".

For mood disorders, the controversy has been especially about the drug-placebo difference in efficacy trials of antidepressants (5), with media hype presenting the conclusion that these drugs are no more effective than sugar pills in unipolar depression. The issue in common with the antipsychotic debate has been the extrapolation to everyday practice of studies completed in rather artificial circumstances for regulatory purposes.

Clearly, these publications have provided conclusions which could be read as being mutually contradictory. The field has been therefore challenged to find reasons behind these discrepancies and remedies which would improve the usefulness of clinical trials for everyday practice.

Patient selection has been identified as one of the main culprits for discrepant findings. Clinical trials of antipsychotics in schizophrenia patients have included highly selected patient populations (6-8), not truly representative of the patients these drugs would be used for in ordinary practice. Increasingly large drop-out rates in RCTs, sometimes linked to specific methodologies (9), have called into question analyses which in one way or another must impute results for missing values, and jeopardized simple conclusions – for example that a single treatment is likely to be effective in treating the target condition. The latter is difficult to claim when almost half the patients in the active arm of a three week trial of mania may fail to complete it. Furthermore, it has been questioned whether the traditional outcome criteria, such as improvements on the total score of rating scales measuring psychopathological symptoms, have ecological validity for true patient outcomes (10,11).

The same problems are even more pronounced for trials in depression. Many patients entering RCTs for depression are attracted by advertisement and may be paid to participate. This is most notably true in the United States, where many such trials have been completed. Moreover, inflation of the depression ratings required for entry is widely believed to occur and so confound subsequent effects attributed either to active treatment or placebo (5).

This discontent has brought the concept of effectiveness into play. Effectiveness studies aim to include an unselected or less selected group of patients by using broad inclusion criteria and few reasons for exclusion. Simple trial methodology may be employed to keep

drop-out rates low. Rather than measuring the effects of therapeutic interventions on fairly specific outcomes in psychopathology, effectiveness studies aspire to measure something more tangible. In the case of large scale trials in cardiovascular medicine, the outcome is often death. In psychiatry, death is too rare an outcome to consider, but admission to hospital or drug discontinuation are regarded as clinically relevant outcomes. In slight defiance of the impulse to measure hard outcomes, there is also a parallel desire to find outcomes relative to the patient experience – often subsumed under the clichéd term "quality of life". Moreover, there are pressures to include an economic evaluation of treatment choices. All of this is geared towards producing results which can be translated into everyday clinical practice, but it also sounds deceptively straightforward.

In the following sections, we will provide some examples of large pragmatic clinical trials in patients suffering from schizophrenia and mood disorders and thereafter discuss the pros and cons of effectiveness studies vis-a-vis traditional RCTs.

EFFECTIVENESS TRIALS IN SCHIZOPHRENIA

Various effectiveness trials in schizophrenia have been performed over the last decade. We focus on studies which have been carried out in large scale samples. Both blinded and open trials are reviewed, provided they have used random treatment allocation. We regard this random allocation, with adequate control and concealment of the allocation process, as the key property allowing fair comparison between two treatments.

Clinical Antipsychotic Trials in Intervention Effectiveness (CATIE)

The CATIE was a clinical trial sponsored by the US National Institute of Mental Health (NIMH) following a bid for a research contract. This large pragmatic trial included three phases. In the first, five new generation antipsychotics were compared to the first generation drug perphenazine. After phase I, patients had the option to switch into two different arms of phase II. One was originally planned to compare clozapine to other new-generation antipsychotics in patients found treatment resistant in phase I, and the other one to include patients who had tolerability problems. Treatment allocation in phase I and II was randomized and double blind, with the exception of the clozapine arm. Following phase II, patients could be switched to open treatment trials of various older or newer antipsychotics. All cause discontinuation was the primary treatment outcome measure (12).

In some way this represents a hybrid methodology, as inclusion criteria and outcome measures followed an effectiveness principle, while the rest of the trial design was that of a traditional RCT. Moreover, this type of staged design may encourage early treatment discontinuation in phase I, as it allows graduation into a second phase of the investigation.

Several papers providing results of phases I and II and more specific treatment outcomes have been published (13-15). With the exception of a significantly lower all cause discontinuation rate with olanzapine, second-generation antipsychotics had no efficacy advantages over perphenazine in any of the analyses published so far. Perphenazine was chosen for pragmatic reasons, to increase the sense of equipoise. A more typical drug such as haloperidol was judged not to be a feasible choice, because of the preconceptions of patients and investigators. It was commented that perphenazine was chosen "because of its lower potency and moderate side-effect profile". Whether it fairly represented the classical antipsychotic group is open to doubt.

Cost Utility of the Latest Antipsychotic Drugs in Schizophrenia Study (CUtLASS)

This study, sponsored by the UK National Health Services, also attempted to compare the effectiveness of newer to older antipsychotics. Clinicians who wanted to enter a patient into this study had to decide at first whether patients had been resistant to previous treatments (in which case they were entered into an arm comparing clozapine to other new-generation antipsychotics) or whether a switch was indicated for other reasons (in which case they were randomly assigned to receive either a first- or a second-generation antipsychotic). Within those two medication groups, clinicians were free to choose the drug of their preference. Quality of life was chosen to be the primary outcome measure.

By the end of this one year open study, clozapine was found to be advantageous over other second-generation drugs in the treatment resistant arm, while there was no advantage of second-generation antipsychotics (46% had been treated with olanzapine) compared to the group of older medications (49% of patients had received sulpiride in this group) (16,17). Pharmacologically speaking, the inclusion and excessive representation of sulpiride in the "first-generation" treatment arm is unhelpful. Sulpiride is pharmacologically very close to amisulpiride, which was included in the second-generation group. This decision in part may have reflected recruitment difficulty, and indeed the trial did not reach its pre-defined recruitment targets. The reason for this was probably a failure of equipoise. The perception of clinicians may have favoured "atypicals" and it was difficult to persuade clinicians to use the older (and more "typical") antipsychotics.

The failure to detect a contrast between first- and second-generation drugs hence becomes questionable. Moreover, the patients entered the study as a consequence of the need to change medication, so potentially selecting patients who were either less responsive (18) or more intolerant of medication (or both).

Comparison of Atypicals in First Episode of psychosis (CAFE)

All cause discontinuation was the primary outcome measure in this dou-

ble-blind clinical trial comparing quetiapine to risperidone and olanzapine (19). Discontinuation rates were high in all three groups, but did not differ from each other. This was also true for changes in scores on the Positive and Negative Syndrome Scale (PANSS). As in the CATIE, olanzapine led to a higher prevalence of weight gain.

European First Episode Study in Schizophrenia (EUFEST)

This one year randomized yet unblinded study, conducted in 13 European countries and Israel, studied the effectiveness of the new-generation antipsychotics amisulpride, quetiapine, olanzapine and ziprasidone in comparison to low-dose haloperidol in patients with a first episode of schizophrenia (20). Loss of retention on the drug to which the patients were originally randomized was the primary outcome. All new-generation drugs performed better than haloperidol. In addition, even a low dose of haloperidol produced more extrapyramidal side effects than the newer agents. The PANSS total scores, one of the secondary outcomes, were not different between groups (21). However, the PANSS scores was measured less often than other outcomes.

The findings of the EUFEST contradict the conclusions often claimed for the CATIE and the CUtLASS – that the atypicals show no important advantage over the older compounds. Low-dose haloperidol was less acceptable than second-generation medications and translated into significantly shorter treatment adherence in first-episode patients. The atypicals in both the CATIE and the EUFEST behaved differently in relation to each other, and do not appear to be equivalent at the doses employed. The comparison with perphenazine in the CATIE, and of one heterogeneous group of compounds with another in the CUtLASS, limits the conclusions that can be reached from these studies.

It needs to be clear that naturalistic clinical trials also reflect naturalistic treatment practice, which may not always follow generally accepted evidence and guidelines. For instance, in the CATIE, only about 40% of all patients in phase I received the maximally allowed doses. On the other hand, pragmatic studies which allow researchers more leeway in including patients and modifying treatment are advantageous for improving retention rates, as exemplified by the CUtLASS and the EUFEST. Blinding also has an impact upon discontinuation rates: in general, higher drop-out rates are encountered in double-blind studies, such as the CATIE and the CAFE.

EFFECTIVENESS TRIALS IN MOOD DISORDERS

Sequenced Treatment Alternatives to Relieve Depression (STAR*D)

The STAR*D did not address efficacy of an antidepressant against a comparator in its initial stage. Instead, all participants were treated with a single selective serotonin reuptake inhibitor, citalopram, and outcomes were systematically determined for over 2000 unipolar patients with a major depressive episode.

Overall, remission rates were probably lower than expected, and the side effect burden higher. Thirty percent of subjects obtained remission and the time required was over 8 weeks. Sub-group analysis was useful in suggesting particular efficacy for women with strong personal backgrounds of achievement. The poorest outcomes were in those patients with longer index episodes, more concurrent psychiatric disorders (especially anxiety disorders or drug abuse), more general medical disorders, lower baseline function and quality of life (18).

The original intention of the STAR*D was to compare strategies of treatment after monotherapy with citalopram had been judged insufficient. Unfortunately, a too permissive approach to patient choice resulted in a disappointing rate of true randomization to competing treatments. After failure on citalopram, level 2 options in STAR*D were a switch to another medication (bupropion, sertraline or venlafaxine) or cognitive therapy, or augmentation of citalopram with bupropion, buspirone or cognitive therapy. Only 21 of 1439 patients accepted to be randomized to any of these options. The vast majority had preferences that were allowed in the study design. Thus, comparisons between augmentation and switch strategies were of great clinical interest, but were subverted by allowing patient preference for one of these approaches. About 30% of patients in all groups treated with medication remitted after change in treatment, whatever the type (22,23). The rate for cognitive therapy was substantially lower (but not statistically different because of lack of power) (24). Further steps in the treatment algorithm suffered from falling numbers, and most outcomes were not statistically discriminable one from another.

There are conflicting interpretations of the STAR*D programme. Nihilists will say that we have learned nothing from it. Optimists claim that the treatment strategies showed reasonable overall remission rates if the algorithm was followed. Whether this represented an improvement on real life treatment could not be decided. The strengths of the study were the sample size and some preliminary pharmacogenetic findings.

Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD)

The STEP programme was a major effort, in parallel with the STAR*D, to examine a variety of treatment pathways for bipolar patients. Of the five proposed trials, only that enrolling depressed bipolar patients yielded useful randomized results. The acute depression study addressed in 366 patients the response to adding antidepressants or placebo to ongoing mood stabilizers (in practice this was very liberally interpreted and almost any non-antidepressant co-medication was permitted). The findings were negative, with no evidence for remission (or for switch to mania) occurring preferentially in the antidepressant arm (25).

The result can be interpreted either as a negative study (antidepressants do not

work in bipolar depression) or a failed study (we do not know if antidepressants work in bipolar depression). In the absence of a positive control treatment, the answer remains moot. A lesson from the antidepressant study of STEP-BD is perhaps not to combine the inexactitude of real life with the unreality (placebo) of the regulatory trial.

Bipolar Affective disorder: Lithium/ Anticonvulsant Evaluation (BALANCE)

The BALANCE is a study completing in mid 2008 to compare the combination of lithium plus valproate (as Depakote) with lithium or valproate monotherapy. The question it was designed to address was the superiority of combination treatment over monotherapy in the long-term treatment of bipolar disorder. This was felt to have generic value in bipolar disorder, because combination of different medicines for long-term treatment has become very common, although almost unsupported by independent evidence of benefit. Secondarily, the study was designed to compare lithium with valproate as monotherapy.

The study was initially intended to assess re-admission as the primary outcome, but the size of the sample required would have been very large (over 1000 participants) and, in the absence of adequate funding, recruitment was likely to take too long. In fact, successful placebo controlled studies of lamotrigine (25,26) and the lessons of the failed study of valproate (27) (which had planned to use re-admission as primary end-point also), resulted in a rethink and the adoption of time to intervention for a new mood episode as the primary outcome. Final recruitment numbers were over 400, with 330 successfully randomized. The study outcomes will be analysed later in 2008. Event rates are compatible with adequate assay sensitivity.

Like the EUFEST, the BALANCE was a randomized open study. This conserves the primary advantage of any RCT: random and concealed allocation to different treatments. However, clinician or patient bias could contaminate the study. In practice, a significant run-in on combination therapy helped to protect the study from poor adherence and to some extent mitigated against bias for or against a particular treatment. Nevertheless, treatment could have been driven in part by bias, especially for early interventions. These factors will limit but not invalidate the findings of the study, since the absence of a blind is obviously closer to real clinical practice. In particular, we are not convinced that a single prevailing bias against any one of the study treatments could be detected among participating clinicians or patients.

DISCUSSION

When balancing the merits of efficacy and effectiveness studies, one will have to weigh the advantages of studying well-defined homogeneous patient samples with state-of-the-art double-blind methodology against obtaining data closer to everyday clinical practice. This means recruiting more representative samples and using potentially more relevant outcome measures. But it also means, when unblinded as in normal practice, risking bias from patients and clinicians who determine the outcomes. As open, unblinded studies are always at risk for observer bias, this disadvantage needs to be balanced against the fact that generalizability of results is higher with lower drop-out rates. From a methodological perspective, randomization appears to be a *condicio sine qua non* if one chooses to compromise for an unblinded study.

The definition of relevant outcome measures has also been a source of heated debate. On the one hand, it is argued that all cause discontinuation, even if split into discontinuation due to lack of efficacy, tolerability issues or patient choice, is an unsophisticated and crude outcome measure. On the other, it can be argued that a minor change in PANSS total scores or even more specific factors of a rating scale may only be of marginal clinical relevance.

There is a kind of uncertainty principle at work here. The more rigorous and controlled an experiment, the more confident we become of the treatment effect, but the less a trial corresponds to real life; the closer to real life an effectiveness study becomes, the less it offers confidence of efficacy. In principle, we believe that both kinds of study are desirable, but always together, not as alternatives. Moreover, we are most secure when both types of study indicate similar directions of effect.

CONCLUSIONS

When considering all evidence available to date, we suggest that both the experimental RCT and the more pragmatic effectiveness design have an important place in clinical psychopharmacology. Ideally, drug development, after an exploratory phase I, which more and more includes patients, at risk samples or healthy volunteers in proof of concept studies, will proceed with blinded well-controlled studies with rigorously defined outcomes. Such studies can demonstrate efficacy, but the magnitude of the benefit cannot be simply extrapolated to real life.

Results from these phase II and III studies should then be complemented, perhaps as early as phase IIIb, by larger pragmatic clinical trials. Such trials must be designed to ask the key pragmatic clinical questions in the patient population at large. In the examples we have considered, this could range from head-to-head comparability with earlier generation compounds to use in combination with other drugs or psychological interventions. Very complex designs reduce the acceptability of trials to patients (and investigators). Moreover, all pragmatic studies need to be undertaken before extensive marketing of new compounds has occurred and opinions about them have already hardened in the minds of investigators. We believe that the licensing of new drugs currently seems to demand too much (and increasing) evidence from early stage trials of poor generalizability. A provisional licence harnessed to the implementation of large scale clinical trials might meet some of the needs we perceive for the development of new medicines.

Effectiveness studies need to be planned using key properties of clinical trial methodology, namely randomization and concealment of allocation. They will be assured by statistical planning, clear *a priori* hypotheses and necessary good clinical practice standards. Reporting of adverse events in such trials could provide early indications of unexpected problems with safety. Employing these trial designs earlier in drug development may diffuse some of the controversy around the applicability to ordinary practice of trials completed for drug registration and also allow for a quicker appreciation of a drug's usefulness in meeting real clinical needs.

References

1. Leucht S, Pitschel-Walz G, Abraham D et al. 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 1999;35:51-68.
2. Geddes J, Freemantle N, Harrison P et al. Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. BMJ 2000;321:1371-6.
3. Davis JM, Chen N, Glick ID. A meta-analysis of the efficacy of second generation antipsychotics. Arch Gen Psychiatry 2003;60:553-64.
4. Tandon R, Fleischhacker WW. Comparative efficacy of antipsychotics in the treatment of schizophrenia: a critical assessment. Schizophr Res 2005;79:145-55.
5. Walsh BT, Seidman SN, Sysko R et al. Placebo response in studies of major depression: variable, substantial, and growing. JAMA 2002; 287:1840-7.
6. Robinson D, Woerner MG, Pollack S et al. Subject selection biases in clinical trials: data from a multicenter schizophrenia treatment study. J Clin Psychopharmacol 1996;16:170-6.
7. Hofer A, Hummer M, Huber R et al. Selection bias in clinical trials with antipsychotics. J Clin Psychopharmacol 2000;20:699-702.
8. Hummer M, Fleischhacker WW. Do phase III trials have clinical value? J Clin Psychopharmacol 1999;19:391-2.
9. Kemmler G, Hummer M, Widschwendter C et al. Dropout rates in placebo-controlled and active-control clinical trials of antipsychotic drugs: a meta-analysis. Evid Based Ment Health 2006;9:70.
10. Leucht S, Davis JM, Engel RR et al. Defining "response" in antipsychotic drug trials: recommendations for the use of scale-derived cutoffs. Neuropsychopharmacology 2007;32:1903-10.
11. Fleischhacker WW, Kemmler G. The clinical relevance of percentage improvements on the PANSS score. Neuropsychopharmacology 2007;32:2435-6.
12. Stroup TS, McEvoy JP, Swartz MS et al. The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: schizophrenia trial design and protocol development. Schizophr Bull 2003;29:15-31.
13. Lieberman JA, Stroup TS, McEvoy JP et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 2005;353:1209-23.
14. McEvoy JP, Lieberman JA, Stroup TS et al. 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 2006;163:600-10.
15. Keefe RS, Bilder RM, Davis SM et al. Neurocognitive effects of antipsychotic medications in patients with chronic schizophrenia in the CATIE Trial. Arch Gen Psychiatry 2007;64:633-47.
16. Lewis SW, Barnes TR, Davies L et al. Randomized controlled trial of effect of prescription of clozapine versus other second-generation antipsychotic drugs in resistant schizophrenia. Evid Based Ment Health 2007;10:57.
17. Jones PB, Barnes TR, 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 (CUtLASS 1). Arch Gen Psychiatry 2006;63:1079-87.
18. Trivedi MH, Rush AJ, Wisniewski SR et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 2006;163:28-40.
19. McEvoy JP, Lieberman JA, Perkins DO et al. Efficacy and tolerability of olanzapine, quetiapine, and risperidone in the treatment of early psychosis: a randomized, double-blind 52-week comparison. Am J Psychiatry 2007;164:1050-60.
20. Fleischhacker WW, Keet IP, Kahn RS et al. The European First Episode Schizophrenia Trial (EUFEST): rationale and design of the trial. Schizophr Res 2005;78:147-56.
21. Kahn RS, Fleischhacker WW, Boter H et al. Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial. Lancet 2008;371:1085-97.
22. Rush AJ, Trivedi MH, Wisniewski SR et al. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. N Engl J Med 2006;354:1231-42.
23. Trivedi MH, Fava M, Wisniewski SR et al. Medication augmentation after the failure of SSRIs for depression. N Engl J Med 2006; 354:1243-52.
24. Thase ME, Friedman ES, Biggs MM et al. Cognitive therapy versus medication in augmentation and switch strategies as second-step treatments: a STAR*D report. Am J Psychiatry 2007;164:739-52.
25. Sachs GS, Nierenberg AA, Calabrese JR et al. Effectiveness of adjunctive antidepressant treatment for bipolar depression. N Engl J Med 2007;356:1711-22.
26. Goodwin GM, Bowden CL, Calabrese JR et al. A pooled analysis of 2 placebo-controlled 18-month trials of lamotrigine and lithium maintenance in bipolar I disorder. J Clin Psychiatry 2004;65:432-41.
27. Bowden CL, Calabrese JR, McElroy SL et al. A randomized, placebo-controlled 12-month trial of divalproex and lithium in treatment of outpatients with bipolar I disorder. Arch Gen Psychiatry 2000;57:481-9.