Relative indices of treatment effect may be constant across different definitions of response in schizophrenia trials

Toshi A. Furukawa [a, b]*

Tatsuo Akechi [b]

Stefan Wagenpfeil [c]

Stefan Leucht [d]

[a] Department of Health Promotion and Human Behavior (Cognitive-Behavioral Medicine), Kyoto University Graduate School of Medicine / School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501 Japan

[b] Department of Psychiatry and Cognitive-Behavioral Medicine, Nagoya City University Graduate School of Medical Sciences, Mizuho-cho, Mizuho-ku, Nagoya 467-8601 Japan

[c] Institute of Medical Statistics and Epidemiology of the Technische Universität München, Klinikum rechts der Isar, Ismaningerstr. 22, Munich 81675 Germany

[d] Department of Psychiatry and Psychotherapy, Technische Universität München, Klinikum rechts der Isar, Ismaningerstr. 22, Munich 81675 Germany

* Corresponding author. Department of Health Promotion and Human Behavior (Cognitive-Behavioral Medicine), Kyoto University Graduate School of Medicine / School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501 JAPAN. Tel: +81-75-753-8291, Fax: +81-75-753-4641, Email: furukawa@kuhp.kyoto-u.ac.jp

ABSTRACT

Background: In randomized controlled trials of antipsychotics, various cutoffs have been used to define response on continuous outcome measures.

Aims: To find a summary effect measure that remains constant across different definitions of response

Method: We conducted secondary analyses of individual patient data from 10 randomized controlled trials of second generation antipsychotics for schizophrenia (n=4278) by applying a meta-analytic approach to produce odds ratios (OR), risk ratios (RR) and risk differences (RD) and their 95% confidence intervals (CI) for different definitions of response, using cutoffs of 10% through 90% reduction on the symptom severity rating scales. Constancy of these indices was examined through visual inspection, by way of I-squared statistics to quantify heterogeneity, and by way of coefficients of variation. If any of these indices were found to remain reasonably constant, we next examined the concordance between the number needed to treat (NNT) predicted from them and the observed NNT.

Results: OR and RR remained reasonably constant across various definitions of response, especially for those using thresholds of 10% through 70% reduction. The NNTs predicted from OR and RR agreed well with the observed NNTs, with ANOVA intraclass correlation coefficients of 0.96 (95% CI: 0.92 to 0.98) and 0.86 (0.72 to 0.93), respectively.

Conclusions: The relative measures of treatment effectiveness remain reasonably constant across different scale-derived definitions of response and, in conjunction with varying control event rates, can give accurate estimates of NNTs for individuals with schizophrenia.

KEYWORDS

Antipsychotic agents, Rating scale, Evidence-based medicine, Number needed to treat, Effect size

1. INTRODUCTION

In psychiatry “hard” outcomes such as death are not readily available or appropriate indices of treatment effectiveness. Instead, continuous outcomes based on rating scales are often employed but it is sometimes not easy to interpret the meaning of these scores (Norman et al., 2001). For example, in a hypothetical drug trial of acute phase treatment of schizophrenia, a statistically significant difference on a certain disease severity measure of 70 vs 80 may be reported for the drug and placebo arms, respectively, at the end of the trial. However, what these 70 or 80, or what this 10-point difference, on this scale means clinically may often not be transparent.

On the other hand, a categorical approach can be more interpretable, for example, if the response or remission rates are reported to be 50% vs 30% in the two arms. Trialists have therefore often included “response” rates defined as a threshold decrease on the continuous outcome (Altman and Royston, 2006). Unfortunately, for many of these continuous outcomes, there usually is no validated or even agreed-upon cutoff to define “response.” In the case of schizophrenia trials, the Brief Psychiatric Rating Scale (BPRS) (Overall and Gorham, 1962) and the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) are the two most frequently used scales but investigators have used various percentage improvements from 20% through 50% to define response (Beasley et al., 1996b; Marder and Meibach, 1994; Peuskens and Link, 1997; Small et al., 1997).

Such lack of consensus in the definition of response poses several related difficulties. First, there is suspicion that the trialists choose their cutoff not because it is clinically appropriate but because it is more likely to result in “statistically significant” differences. In recent reports of schizophrenia trials there is a tendency to use 20% reduction as a cutoff, apparently in the belief that a lower cutoff increases the ability to find statistically significant differences between drugs. However, 20% reduction represents something less than “minimal improvement” (Leucht et al., 2005a; Leucht et al., 2005b). A statistically significant difference in the rates of patients showing borderline or greater improvement (but not necessarilty in moderate or greater improvement) would certainly not be clinically meaningful.

Second, in order to obtain unbiased and generalizable estimates of the true treatment effects, we need comprehensive meta-analyses of relevant trials. However, if "response" is defined variably across different trials addressing a similar clinical question, we cannot be sure if we could safely combine them in a meta-analysis.

These problems would be greatly ameliorated if one could find a measure of effect that remained more or less constant across a range of thresholds. The odds ratio (OR), relative risk (RR) and risk difference (RD) or its inverse, number needed to treat (NNT), are the representative indices of treatment effectiveness for dichotomous outcomes. When the outcome is dichotomous, the results of a trial can be summarized as in the following 2*2 table.

	Response	Non-response
Treatment	a	b
Control	c	d

The more clinically interpretable indices are RR and RD. RR is the ratio of the response rates in the treatment and control arms; it is therefore (a/(a+b))/(c/(c+d)). RD is the difference in the response rates in the treatment and control arms; it is therefore a/(a+b) - c/(c+d). NNT, which is the inverse of RD, shows the number of patients one would need to treat in order to have one more response in the treatment arm that would not have happened if on the control arm. It therefore nicely summarizes the amount of effort that both clinicians and patients need to expend in order to obtain one more response. For example, a treatment that produces a response rate of 50% in comparison with a placebo response rate of 30% would be translated into an NNT of 1/(0.5 - 0.3) = 5. In other words, one would need to treat 5 patients in order to produce one more responder over what would have happened on placebo. On the other hand, OR is intuitively difficult to understand because it is the ratio of the odds of showing response over not showing response in the treatment and control arms; hence it is (a/b)/(c/d) or ad/bc.

OR, however, has some strong mathematical properties because OR of non-response is the inverse of OR of response, whereas such a relationship does not hold for RR (Deeks, 2002).

In the following analyses, we examined individual patient data from several clinical trials of schizophrenia to see if any of OR, RR or RD may remain constant across different definitions of response, so that it can be used as the generalizable index of treatment effectiveness.

2. METHODS

2.1. Database

Individual patient data from 10 trials comparing olanzapine vs haloperidol (5 comparisons, baseline n=2974) (Beasley et al., 1997; Beasley et al., 1996b; Keefe et al., 2006; Lieberman et al., 2003; Tollefson et al., 1997), amisulpride vs haloperidol (4 comparisons, baseline n=1198) (Carriere et al., 2000; Colonna et al., 2000; Moller et al., 1997; Puech et al., 1998), and olanzapine vs placebo (2 comparisons, baseline n=502) (Beasley et al., 1996a; Beasley et al., 1996b) that administered either the BPRS or PANSS were reanalyzed *post hoc*. These 10 trials were selected from among the 13 trials that compared olanzapine and 7 trials that compared amisulpride against various other antipsychotics or placebo and that had been provided to us by respective manufacturers, when they compared olanzapine vs haloperidol, amisulpride vs haloperidol or olanzapine vs placebo because these were the only comparisons that resulted in statistically significant differences between the compared arms when meta-analyzed. Working with non-significant ORs, RRs or RDs would not reveal their differential performances. One trial was a three-armed trial among olanzapine, haloperidol and placebo, and contributed to two comparisons. Table 1 summarizes important characteristics of the included studies.

All studies were randomized and all but one (Colonna et al., 2000) were described as double-blind. All amisulpride studies and one olanzapine study (Beasley et al., 1996b) used the original BPRS, and all the other olanzapine studies used PANSS. For the latter studies we calculated the PANSS-derived BPRS

scores because PANSS includes all items of the BPRS. The BPRS is a clinician-rated rating scale designed to measure change in psychopathology and contains 18 items, each of which is rated on a seven-point scale ranging between 1="not present" and 7="extremely severe," resulting in a total score between 18 and 126 (Overall and Gorham, 1962). The PANSS was developed to improve on the BPRS by including all its items and by adding 12 more items to cover broader psychopathology; its score therefore ranges between 30 and 210 (Kay et al., 1987).

For fixed-dose studies, we selected only those arms with optimum doses of second-generation antipsychotic drugs as reported in dose-finding studies (amisulpride 400-800 mg/day, olanzapine 10-20 mg/day and risperidone 4-6 mg/day) (Leucht et al., 2009). We therefore excluded 61 participants from Puech et al (1998) who had received a potentially subtherapeutic 100 mg/day of amisulpride, 175 participants from Beasley et al (1997) who received 5 mg/day or 1 mg/day of olanzapine, 65 participants from Beasley et al (1996b) who were given 5 mg/day of olanzapine and 52 participants from Beasley et al (1996a) who received 1 mg/day of olanzapine. The active comparator, haloperidol, was given in a fixed dose of 15 mg/day or 16 mg/day or in variable dosage ranging between 2 and 30 mg/day: these dosages have been found to show similar effectiveness (Leucht et al., 2009; Waraich et al., 2002).

The mean BPRS total score of the included participants was 54.3 (SD=10.8) at baseline, which would correspond with "markedly ill" range (Leucht et al., 2005a). There were 2895 men and 1383 women. Their mean age was 36.6 (10.5) years, weight 75.5 (16.4) kg and height 171.6 (9.6) cm.

2.2. Statistical analyses

We first calculated the numbers of responders defined as 10% through 90% reduction on the BPRS or PANSS total score at 4 weeks, with missing data supplemented by the last-observation-carried-forward (LOCF) method even if a participant dropped out before the first post-baseline rating. The percentage

reduction was calculated according to the formulae: $B\% = (B_0 - B_{4LOCF}) * 100 / (B_0 - 18)$ for BPRS and $P\% = (P_0 - P_{4LOCF}) * 100 / (P_0 - 30)$ for PANSS, where B_0 and P_0 are BPRS and PANSS scores at baseline and B_{4LOCF} and P_{4LOCF} are respective scores at 4 weeks with LOCF, because 18 and 30 are the minimum scores for BPRS and PANSS, respectively, according to the original rating system (Kay et al., 1987; Overall and Gorham, 1962).

We then ran meta-analyses of response rates defined as 10% through 90% reduction for each comparison in terms of OR, RR and RD, using Review Manager software by the Cochrane Collaboration (2008). Because we are looking for a single index that may remain constant through different definitions of response, we used the Mantel-Haenszel fixed effect model (Mantel and Haenszel, 1959). However, in order to examine robustness of our findings, we repeated the same analyses based on the DerSimonian random effects model (DerSimonian and Laird, 1986).

Constancy of these three summary indices of treatment effectiveness was examined (i) through visual inspection of the obtained indices and their 95% confidence intervals, (ii) I-squared statistics of the hypothetical meta-analyses of the relevant trials adopting different thresholds for 10% through 90%, and (iii) coefficient of variation (CV). CV is defined as SD/|mean|, and its 95% confidence intervals were calculated according to Johnson & Welch (1939). Because heterogeneity may arise from differences in baseline severity of the included patients, study year and/or methodologic rigor of the included trials, we repeated the analyses by excluding trials that had low baseline BPRS scores (Keefe et al., 2006; Lieberman et al., 2003; Tollefson et al., 1997), old trials before year 2000 (Beasley et al., 1997; Beasley et al., 1996a; Beasley et al., 1996b; Moller et al., 1997; Puech et al., 1998; Tollefson et al., 1997) or non-blinded trials (Colonna et al., 2000).

If any of OR, RR and RD appeared constant through ranges of definitions of response and may therefore be given the role of the representative index of effectiveness of one arm over the other regardless of the cutoffs in the continuous scale, we then examined if that summary effect measure can accurately predict the other effect measures according to the known mathematical relationships among OR, RR and RD on

one hand and control event rate (CER; the response rate in the control group) on the other. Given the 2*2 table as shown in the Introduction, by introducing CER = c/(c+d), we can calculate RR and RD from OR using the formulae:

$$RR = OR/(1 - CER + CER*OR)$$

$$RD = CER * (RR - 1).$$

We can also calculate OR and RD from RR using the same formulae. The degree of absolute agreement between the observed values and the predicted values was expressed by way of two-way mixed ANOVA intraclass coefficient (ICC) for absolute agreement.

3. RESULTS

3.1. Visual inspection of the constancy of OR, RR and RD across different definitions of response

Figures 1, 2 and 3 depict the OR, RR and RD corresponding to the various definitions of response using 10% through 90% reduction in the BPRS or PANSS total scores for the comparisons olanzapine vs haloperidol, amisulpride vs haloperidol and olanzapine vs placebo, respectively. Visual inspection of these graphs indicate that both OR and RR appear to remain relatively constant, especially for the ranges of 10% through 70% reduction.

For the extreme ranges of 80% or 90% reduction, there were too few participants to achieve the so-defined response and the calculated OR and RR were all unstable and resulted in wide 95% confidence intervals.

3.2. Numerical examination of constancy of OR, RR and RD

Table 2 lists the I-squared for the hypothetical pooling across all definitions of response, and CV for the OR, RR and RD.

RD produced I-squared statistics which would be interpreted as representing moderate to high heterogeneity (Higgins et al., 2003) and were always greater in absolute value than those for RR, which were then greater than those for OR, although neither of these differences were statistically significant (Wilcoxon signed rank test). Focusing on more homogeneous trials by excluding trials with low baseline BPRS scores, old trials or non-blinded trials resulted in very similar I-squared values (Table 2).

RD also had the greatest CV, while OR had the smallest CV: the differences in CV between RD on one hand and OR or RR on the other were both statistically significant, because their 95% CIs did not overlap.

3.3. Which relative index of treatment effect enables more accurate prediction of treatment effect on the other indices, OR or RR?

It now appears both OR and RR remain relatively constant through ranges of definitions of response on the continuous scale, especially for 10% through 70% where we have non-small control event rates. The next question then is which of these can allow more accurate prediction of treatment effect according to the other indices, when we apply the mathematical formulae as explained in the Methods section. As the observed OR and RR, we took the average of the values for 10% through 70% reduction, as these appeared to remain particularly constant in Figures 1 through 3.

Using this OR and the varying control event rates, the ANOVA ICC (two-way mixed, absolute agreement) between the observed and the predicted RRs, RDs and NNTs was 0.94 (0.87 to 0.97), 0.95 (0.90 to 0.98) and 0.96 (0.92 to 0.98) respectively for the ranges between 10% through 70%.

With regard to the RR, the ANOVA ICC between the observed and the predicted ORs, RDs and NNTs was -0.03 (-0.37 to 0.34), 0.61 (0.31 to 0.80) and 0.86 (0.72 to 0.93) respectively.

3.4. Sensitivity analyses

All the analyses based on random effects model were essentially unchanged. For example, the ANOVA ICCs between the RR, RD and NNT predicted from OR and those actually observed were 0.93 (0.85 to 0.97), 0.95 (0.90 to 0.98) and 0.91 (0.82 to 0.96) respectively for the ranges between 10% through 70%. And the corresponding ICCs between the OR, RD and NNT predicted from RR and those actually observed were -0.01 (-0.37 to 0.35), 0.59 (0.27 to 0.79) and 0.71 (0.47 to 0.86). Once again, assuming constancy of OR enabled excellent prediction of RR, RD and NNT in conjunction with varying control event rates, while assuming constancy of RR enabled satisfactory prediction of NNT.

4. DISCUSSION

Based on individual patient data of 4278 patients with schizophrenia participating in trials of acute phase antipsychotic treatment, we examined empirically whether OR, RR or RD remain constant across different definitions of response on the BPRS and the PANSS. We found that both OR and RR remain relatively constant across plausible ranges of definitions of response and that OR, in particular, was able to predict RR, RD and NNT very accurately using mathematical formulae and estimates of the control event rate.

The greater generalizability of relative measures of treatment effectiveness (such as OR and RR) over absolute ones (such as RD and NNT) is consistent with previous studies. Using a random subset of meta-analyses contained in the Cochrane Library, Furukawa, Guyatt and Griffith (2002) examined the concordance between treatment indices of each RCT included in a meta-analysis and the meta-analyzed results of all the other RCTs. OR and RR showed the highest concordance rates, even when the control

event rate differed substantially, while the concordance for RD was much lower. In other words, OR and RR appeared more generalizable than RD, regardless of the control event rate.

One of the central goals of evidence-based medicine, to individualize group data from clinical research to match each individual patient's values and preferences, seems therefore to have found some empirical ground here (Sackett, 2001). By assuming a constant relative index of treatment effectiveness, either in terms of OR or RR, and by combining it with each patient's expected event rate when given the control intervention, we can estimate individualized NNT using mathematical formulae as explained in the Methods of this paper. In other words, when the relative difference in effect is constant, the absolute difference in effect will be different, depending on the expected control event rate that can vary from patient to patient.

The present study has added support that relative indices of treatment may be generalizable even across a range of scale-derived definitions of response. In other words, because the relative effectiveness is constant across different thresholds, the absolute effectiveness can be calculated taking into account the threshold that the patient wishes to achieve. For example, the OR of olanzapine over haloperidol to bring about a response in the acute phase treatment of schizophrenia is approximately 1.5 for various definitions of response of 10% through 70% reductions on the BPRS or PANSS (Cf. Figure 1). However, olanzapine causes more significant weight gain than haloperidol, with an NNH estimated to be around 6 (95%CI: 4-11) (Duggan et al., 2005). A patient who is normo- to underweight now and who does not have any family and other risk factors for obesity may be happy to try olanzapine to achieve a 30% or more decrease in disease severity. For this patient, given an estimate that approximately 40% of the patients would achieve 30% or more reduction on placebo, NNT will be calculated to be 9 and he or she may find this NNT as small as NNH for weight gain to justify treatment with olanzapine. On the other hand, another patient who is already somewhat overweight and has multiple family history of diabetes mellitus and cardiovascular diseases may like 70% or more decrease in the BPRS before he/she selects olanzapine over haloperidol. However, because the control event rate

for 70% reduction could be as low as 6% and the corresponding NNT may be as large as 50, he/she might reason that trying olanzapine may not be worthwhile.

Several caveats are in order before we conclude. First, we do not yet know if the current results would apply to other continuous outcomes in other areas of psychiatry or medicine. In fact there is no mathematical necessity for one measure of effect to be stable across thresholds in all settings, as performance of summary effect measures would be dependent on the underlying distribution of the continuous outcomes. However, some sporadic examples we find in the literature suggest that the present findings may apply in other areas as well. Among patients with active rheumatoid arthritis refractory to tumor necrosis factor α, abatacept was superior to placebo in bringing about ACR 20% responses with an OR of 4.2 (95%CI: 2.6 to 6.9), ACR 50% responses with an OR of 6.5 (2.5 to 17), and ACR 70% responses with an OR of 7.4 (1.7 to 32) (Genovese et al., 2005). For patients with psoriasis, ustekinumab, a human interleukin-12/23 monoclonal antibody, beat placebo in reducing the Psoriasis Area and Severity Index by at least 50% (OR=45, 95%CI: 26 to 75), by at least 75% (OR=63, 30 to 133) and by at least 90% (OR=36, 14 to 89) (Leonardi et al., 2008). Thus in both studies, ORs for different thresholds showed largely overlapping confidence intervals that contained all the reported point estimates. More definitive analyses would require individual patient data from more studies, and the present study was the first to carry out such an examination in the context of second generation antipsychotic trials for people with schizophrenia.

Second, the individual patient data that we had access to did not include important individual patient characteristics such as duration of untreated psychosis, years ill, treatment history or concurrent psychosocial treatments. How these variables could have moderated the present results is hard to predict. However, if the present results are generalizable, we could expect that OR and RR would remain relatively constant within each prognostic stratum as defined by such baseline characteristics. Third, the included studies are limited to trials around 2000. If the present results would apply to more

recent trials, especially in view of the changes in patient selection and given increasing number of failed placebo-controlled trials recently, awaits replication.

Taking all these considerations into account, we think that our results have several important implications for research. First, we can safely combine trials that adopted different definitions of response in a meta-analysis to derive the pooled OR or RR, which can then be applied to wider ranges of response definitions and of patients. Second, in the original report of a clinical trial, however, it will be more informative not only to report the OR or RR but also the control event rates for different definitions of response (Leucht et al., 2007). Third, when we plan an RCT, given the constant OR, mathematics shows that the threshold at which the response rate approaches 50% will provide the greatest statistical power.

The implication of the present study for clinical practices is straightforward. The constant OR across different scale-derived definitions of response, in conjunction with varying control event rates, will give accurate estimates of individualised NNTs in pharmacotherapy of schizophrenia and possibly in other areas of medicine. The RR also remains relatively constant so that it may be used, in conjunction with varying control event rates, to estimate NNTs. Using RR has the advantage of allowing busy clinicians easier calculation than using OR.

Role of funding source

This work required no external funding.

Contributors

TAF conceived the study. TAF, SW and SL undertook the statistical analyses. TAF wrote the first draft of the manuscript, TA, SW and SL provided essential critical comments. All the authors have approved the final manuscript.

Conflict of interest

TAF received research funds and speaking fees from Astellas, Dai-Nippon Sumitomo, Eli Lilly, GlaxoSmithKline, Janssen, Meiji, Otsuka, Pfizer and Schering-Plough. TA received research funds and speaking feeds from Astellas, AstraZeneca, BMS, Daiichi-Sankyo, Dai-Nippon Sumitomo, Eisai, Eli Lilly, GlaxoSmithKline, Janssen, Kyowa-Hakko, Meiji, Otsuka, Pfizer, SanofiAventis, Shionogi and Yakult. SW has no conflict of interest to declare. SL received speaker/consultancy/advisory board honoraria from SanofiAventis, BMS, Eli Lilly, Essex Pharma, AstraZeneca, GlaxoSmithKline, Janssen/Johnson and Johnson, Lundbeck and Pfizer. SanofiAventis and EliLilly supported research projects by SL.

Acknowledgments

We would like to thank David L. Streiner and Gordon H. Guyatt for their very helpful advice and comments on the earlier drafts of this paper. We would also like to thank Eli Lilly and SanofiAventis for letting us use their individual patient database without any influence on the design, conduct or reporting of this study.

REFERENCES

2008. Review Manager (RevMan). The Nordic Cochrane Centre, The Cochrane Collaboration, Copenhagen.

Altman, D.G., Royston, P., 2006. The cost of dichotomising continuous variables. BMJ 332, 1080.

Beasley, C.M., Jr., Hamilton, S.H., Crawford, A.M., Dellva, M.A., Tollefson, G.D., Tran, P.V., Blin, O., Beuzen, J.N., 1997. Olanzapine versus haloperidol: acute phase results of the international double-blind olanzapine trial. Eur. Neuropsychopharmacol. 7, 125-137.

Beasley, C.M., Jr., Sanger, T., Satterlee, W., Tollefson, G., Tran, P., Hamilton, S., 1996a. Olanzapine versus placebo: results of a double-blind, fixed-dose olanzapine trial. Psychopharmacology (Berl.) 124, 159-167.

Beasley, C.M., Jr., Tollefson, G., Tran, P., Satterlee, W., Sanger, T., Hamilton, S., 1996b. Olanzapine versus placebo and haloperidol: acute phase results of the North American double-blind olanzapine trial. Neuropsychopharmacology 14, 111-123.

Carriere, P., Bonhomme, D., Lemperiere, T., 2000. Amisulpride has a superior benefit/risk profile to haloperidol in schizophrenia: results of a multicentre, double-blind study (the Amisulpride Study Group). Eur Psychiatry 15, 321-329.

Colonna, L., Saleem, P., Dondey-Nouvel, L., Rein, W., 2000. Long-term safety and efficacy of amisulpride in subchronic or chronic schizophrenia. Amisulpride Study Group. Int. Clin. Psychopharmacol. 15, 13-22.

Deeks, J.J., 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat. Med. 21, 1575-1600.

DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Control. Clin. Trials 7, 177-188.

Duggan, L., Fenton, M., Rathbone, J., Dardennes, R., El-Dosoky, A., Indran, S., 2005. Olanzapine for schizophrenia. Cochrane Database Syst. Rev., CD001359.

Furukawa, T.A., Guyatt, G.H., Griffith, L.E., 2002. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int. J. Epidemiol. 31, 72-76.

Genovese, M.C., Becker, J.C., Schiff, M., Luggen, M., Sherrer, Y., Kremer, J., Birbara, C., Box, J., Natarajan, K., Nuamah, I., Li, T., Aranda, R., Hagerty, D.T., Dougados, M., 2005. Abatacept for rheumatoid arthritis refractory to tumor necrosis factor alpha inhibition. N. Engl. J. Med. 353, 1114-1123.

Higgins, J.P., Thompson, S.G., Deeks, J.J., Altman, D.G., 2003. Measuring inconsistency in meta-analyses. BMJ 327, 557-560.

Johnson, N.L., Welch, B.L., 1939. Application of the non-central t-distribution. Biometrika 31, 362-389.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13, 261-276.

Keefe, R.S., Young, C.A., Rock, S.L., Purdon, S.E., Gold, J.M., Breier, A., 2006. One-year double-blind study of the neurocognitive efficacy of olanzapine, risperidone, and haloperidol in schizophrenia. Schizophr. Res. 81, 1-15.

Leonardi, C.L., Kimball, A.B., Papp, K.A., Yeilding, N., Guzzo, C., Wang, Y., Li, S., Dooley, L.T., Gordon, K.B., 2008. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients

with psoriasis: 76-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 1). Lancet 371, 1665-1674.

Leucht, S., Corves, C., Arbter, D., Engel, R.R., Li, C., Davis, J.M., 2009. Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. Lancet 373, 31-41.

Leucht, S., Davis, J.M., Engel, R.R., Kane, J.M., Wagenpfeil, S., 2007. Defining 'response' in antipsychotic drug trials: recommendations for the use of scale-derived cutoffs. Neuropsychopharmacology 32, 1903-1910.

Leucht, S., Kane, J.M., Kissling, W., Hamann, J., Etschel, E., Engel, R., 2005a. Clinical implications of Brief Psychiatric Rating Scale scores. Br. J. Psychiatry 187, 366-371.

Leucht, S., Kane, J.M., Kissling, W., Hamann, J., Etschel, E., Engel, R.R., 2005b. What does the PANSS mean? Schizophr. Res. 79, 231-238.

Lieberman, J.A., Tollefson, G., Tohen, M., Green, A.I., Gur, R.E., Kahn, R., McEvoy, J., Perkins, D., Sharma, T., Zipursky, R., Wei, H., Hamer, R.M., 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.

Mantel, N., Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. J. Natl. Cancer Inst. 22, 719-748.

Marder, S.R., Meibach, R.C., 1994. Risperidone in the treatment of schizophrenia. Am. J. Psychiatry 151, 825-835.

Moller, H.J., Boyer, P., Fleurot, O., Rein, W., 1997. Improvement of acute exacerbations of schizophrenia with amisulpride: a comparison with haloperidol. PROD-ASLP Study Group. Psychopharmacology (Berl.) 132, 396-401.

Norman, G.R., Sridhar, F.G., Guyatt, G.H., Walter, S.D., 2001. Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. Med. Care 39, 1039-1047.

Overall, J.E., Gorham, D.R., 1962. The brief psychiatric rating scale. Psychol. Rep. 10, 799-812.

Peuskens, J., Link, C.G., 1997. A comparison of quetiapine and chlorpromazine in the treatment of schizophrenia. Acta Psychiatr. Scand. 96, 265-273.

Puech, A., Fleurot, O., Rein, W., 1998. Amisulpride, and atypical antipsychotic, in the treatment of acute episodes of schizophrenia: a dose-ranging study vs. haloperidol. The Amisulpride Study Group. Acta Psychiatr. Scand. 98, 65-72.

Sackett, D.L., 2001. Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!). CMAJ 165, 1226-1237.

Small, J.G., Hirsch, S.R., Arvanitis, L.A., Miller, B.G., Link, C.G., 1997. Quetiapine in patients with schizophrenia. A high- and low-dose double-blind comparison with placebo. Seroquel Study Group. Arch. Gen. Psychiatry 54, 549-557.

Tollefson, G.D., Beasley, C.M., Jr., Tran, P.V., Street, J.S., Krueger, J.A., Tamura, R.N., Graffeo, K.A., Thieme, M.E., 1997. Olanzapine versus haloperidol in the treatment of schizophrenia and schizoaffective and schizophreniform disorders: results of an international collaborative trial. Am. J. Psychiatry 154, 457-465.

Waraich, P.S., Adams, C.E., Roque, M., Hamill, K.M., Marti, J., 2002. Haloperidol dose for the acute phase of schizophrenia. Cochrane Database Syst. Rev., CD001951.

Table 1. Characteristics of the included studies

Study	Antipsychotic drugs and daily dosage (mg)	Sample size (n)	Duration (weeks)	Mean BPRS at baseline	Selected patient characteristics
Möller et al 1997 (Moller et al., 1997)	Amisulpride 600-800 Haloperidol 15-20	95 96	6	61.7	Inpatients with paranoid, disorganized or undifferentiated schizophrenia (DSM-III-R), BPRS psychotic subscore ≥12 and at least two BPRS psychosis items ≥4
Puech et al 1998 (Puech et al., 1998)	Amisulpride 400-1200 Haloperidol 16	194 64	4	61.3	Inpatients with acute exacerbations of paranoid, disorganized or undifferentiated schizophrenia (DSM-III-R), BPRS psychotic subscore ≥12 and at least two BPRS psychosis items ≥4
Colonna et al 2000 (Colonna et al., 2000)	Amisulpride 200-800 Haloperidol 5-20	368 118	51	56.2	In- or outpatients with acute exacerbations of paranoid, disorganized or undifferentiated schizophrenia (DSM-III-R), at least two BPRS psychosis items ≥4
Carrière et al 2000 (Carriere et al., 2000)	Amisulpride 400-1200 Haloperidol 10-30	97 105	17	65.4	Inpatients with paranoid schizophrenia or schhizophreniform disorder (DSM-IV)
Beasley et al 1997 (Beasley et al., 1997)	Olanzapine 10-15 Haloperidol 15	175 81	6	59.1	Inpatients with acute exacerbations of schizophrenia (DSM-III-R), BPRS total score ≥ 42, CGI-S ≥ 4
Tollefson et al 1997 (Tollefson et al., 1997)	Olanzapine 5-20 Haloperidol 5-20	1337 659	6	51.5	In- and outpatients with schizophrenia, schizophreniform or schizoaffective disorder (DSM-III-R), BPRS total score ≥ 36

Lieberman et al 2003 (Lieberman et al., 2003)	Olanzapine 5-20 Haloperidol 2-20	131 132	12	46.8	In- and outpatients with a first episode of schizophrenia, schizophreniform or schizoaffective disorder (DSM-IV), at least two PANSS psychosis items ≥ 4, CGI-S ≥ 4
Keefe et al 2006 (Keefe et al., 2006)	Olanzapine 5-20 Haloperidol 2-19	159 97	8	48.4	In- and outpatients with schizophrenia or schizoaffective disorder according to DSM-IV, BPRS total score ≥ 36, at least two PANSS psychosis items ≥4
Beasley et al 1996 (Beasley et al., 1996b)	Olanzapine 10-15 Haloperidol 15 Placebo	133 69 68	6	59.9	Inpatients with acute exacerbations of schizophrenia (DSM-III-R), BPRS total score ≥ 42, CGI-S ≥ 4
Beasley et al 1996 (Beasley et al., 1996a)	Olanzapine 10 Placebo	50 50	6	55.2	Inpatients with schizophrenia (residual type excluded) (DSM-III-R), BPRS total score ≥ 42, CGI-S ≥ 4

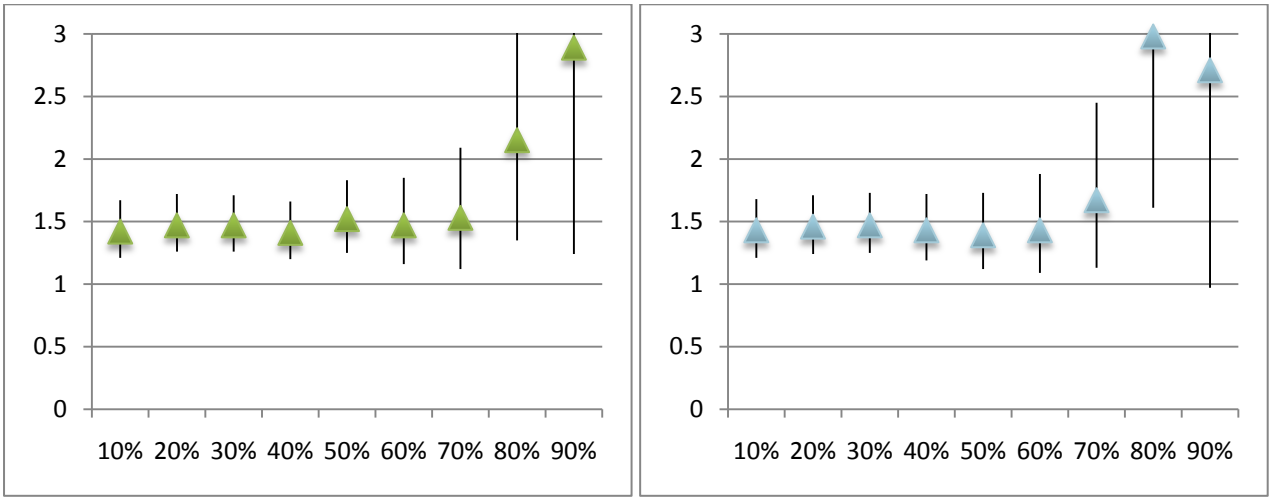
BPRS: Brief Psychiatric Rating Scale, CGI-S: Clinical Global Impression Severity Scale, DSM: Diagnostic and Statistical Manual of Mental Disorders, PANSS: Positive and Negative Syndrome Scale

Table 2. Numerical examination of constancy of OR, RR and RD

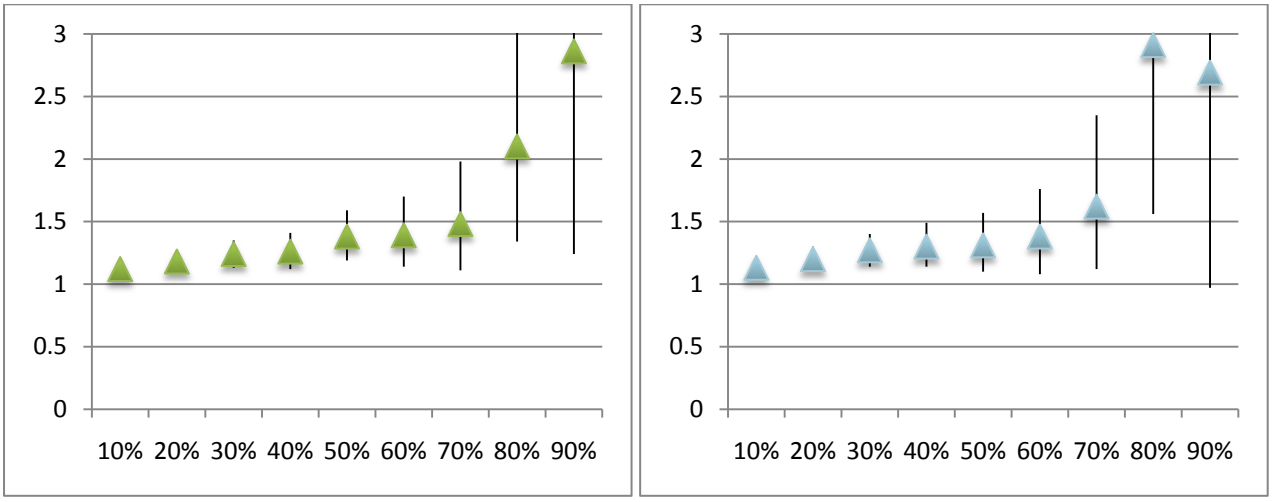
	Comparison	Scale	I-squared				CV
			All	Excl. low BPRS trials	Excl. old trials	Excl. non-blinded trial	
OR	olanzapine vs haloperidol	BPRS	13%	3%	0%	3%	0.29(0.23 to 0.38)
		PANSS	0%	0%	0%	0%	
	amisulpride vs haloperidol	BPRS	0%	0%	0%	0%	
	olanzapine vs placebo	BPRS	0%	0%	0%	0%	
RR	olanzapine vs haloperidol	BPRS	43%	11%	0%	43%	0.33 (0.27 to 0.44)
		PANSS	11%	13%	0%	11%	
	amisulpride vs haloperidol	BPRS	2%	2%	21%	0%	
	olanzapine vs placebo	BPRS	0%	0%	0%	0%	
RD	olanzapine vs haloperidol	BPRS	82%	20%	60%	82%	0.62 (0.55 to 0.71)
		PANSS	86%	76%	47%	86%	
	amisulpride vs haloperidol	BPRS	43%	43%	58%	35%	
	olanzapine vs placebo	BPRS	76%	76%	86%	76%	

Figure 1. Olanzapine vs haloperidol

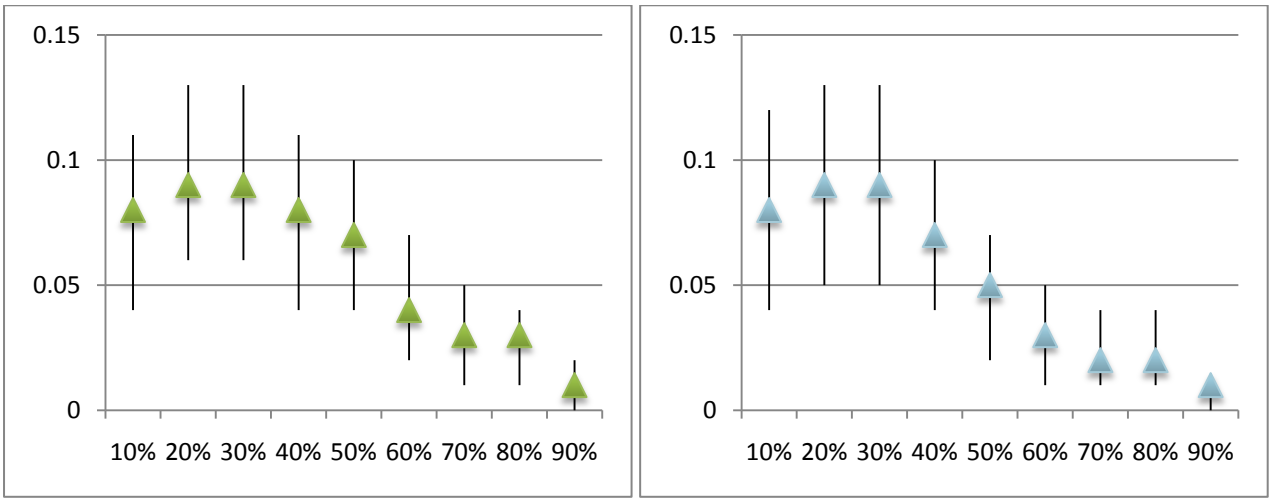
OR



RR



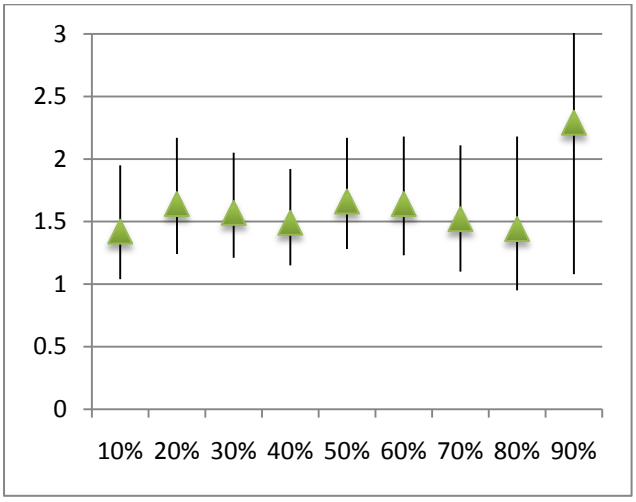
RD



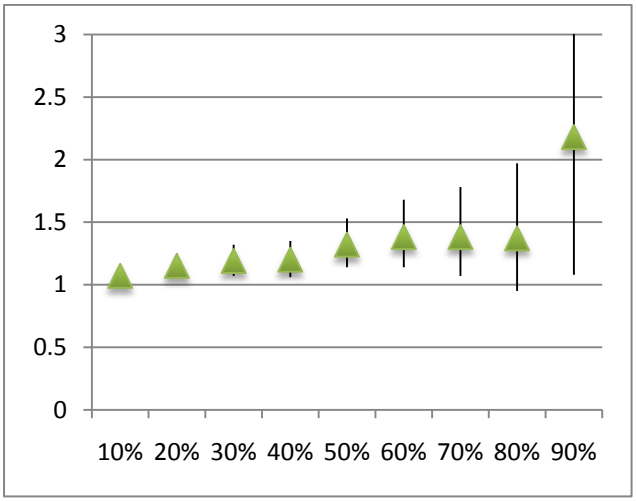
The left column (green triangles) is based on BPRS, and the right column (blue triangles) represents PANSS.

Figure 2. Amisulpride vs haloperidol

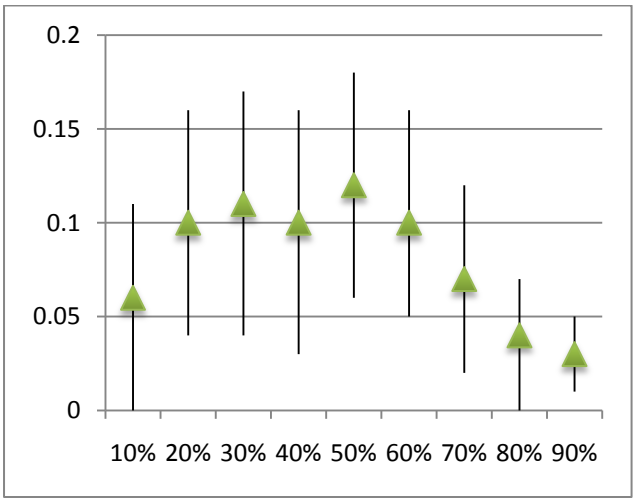
OR



RR



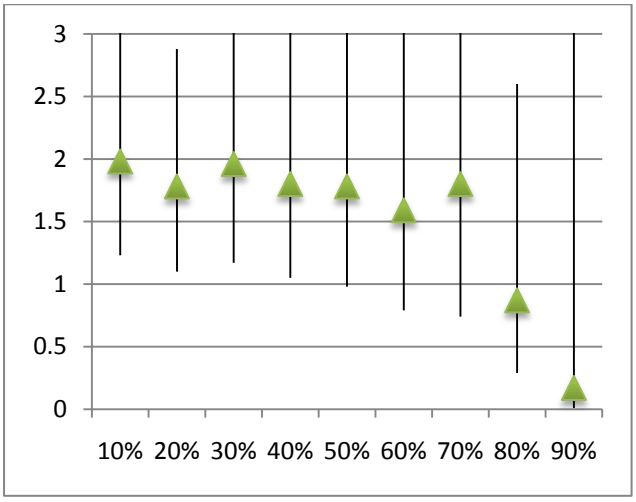
RD



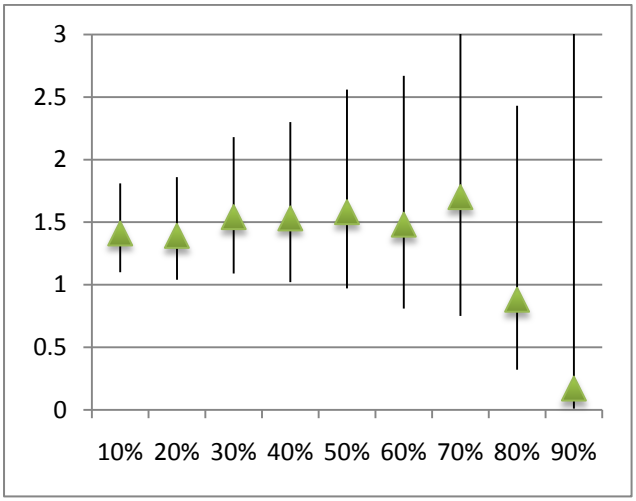
Based on BPRS

Figure 3. Olanzapine vs placebo

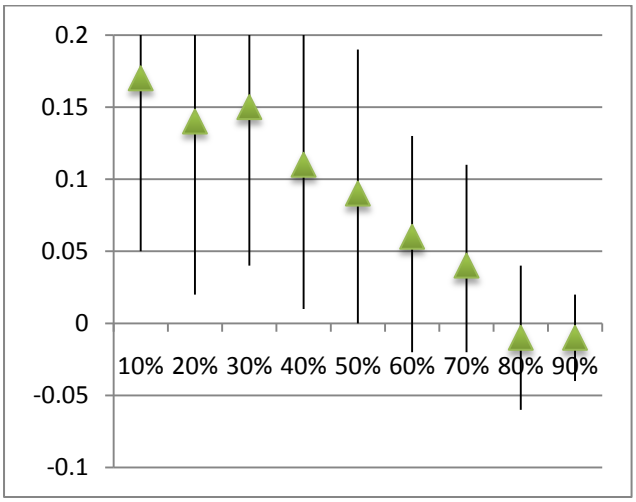
OR



RR



RD



Based on BPRS.