Can we inflate effect size and thus increase chances of producing “positive” results if we raise the baseline threshold in schizophrenia trials?

Toshi A. Furukawa [a, *], Stefan Leucht [b]

[a] Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, JAPAN (Tel: +81-75-753-9491, Fax: +81-75-753-4641, Email: furukawa@kuhp.kyoto-u.a.cjp)

[b] Klinik für Psychiatrie und Psychotherapie der TU-München, Klinikum rechts der Isar, Munich, GERMANY (Email: Stefan.Leucht@lrz.tu-muenchen.de)

* Corresponding author.

Abstract

The standardized mean difference (SMD), also referred to simply as effect size, is often used to summarize the results of a clinical trial when the outcome measure is continuous. SMD is calculated by dividing the difference in the mean scores of the experimental and control groups by their standard deviation (SD). One of the major arguments against SMD is that, if the studied sample is chosen to be artificially homogeneous and thus have a small SD, SMD can be overestimated and thus lose generalizability. On the other hand, smaller SDs raise the chances of finding a statistically significant difference. This study examined whether we can increase sample homogeneity and decrease SD by raising the severity threshold to enter a clinical trial in secondary analyses of individual patient data from three large acute phase schizophrenia trials. Raising the baseline threshold on PANSS and BPRS did reduce the SDs at baseline but SMDs at endpoint remained by and large constant. This was so because the SDs at endpoint appeared to bounce back to their natural values. It is concluded that restricting the entry criteria into schizophrenia trials cannot lead to larger SMDs or to smaller sample size necessary to detect an efficacy signal.

Keywords: Clinical trial, Sample size, Effect size, Schizophrenia

1. INTRODUCTION

There is growing and now almost unanimous consensus that all statistical analyses must include appropriate reports of effect size estimates in addition to null hypothesis significance test results. Effect sizes facilitate understanding of the clinical significance and personal importance of an effect (Guyatt et al., 2004) and also allow comparison and synthesis of effects across studies. The practice of reporting an "appropriate" effect size index however is not straightforward as different indexes each have its own advantages and disadvantages. In the case of continuous outcomes, two representative alternatives are "mean difference" (MD) and "standardized mean difference" (SMD). MD is a simple difference in the mean scores between two groups, e.g. an experimental group and its control group, expressed in the original unit of measurement. SMD, by contrast, is MD divided by a standard deviation (SD), usually the pooled SD of the scores of the two groups. SMD is therefore a difference in the two group means in SD units.

SMD enables comparison (and synthesis) of studies using different scales (Higgins and Green, 2011), and is roughly interpretable (Cohen, 1988), while MD is more intuitive and easier to interpret when there is one well-known and dominant scale. More often than not, however, it is the case in psychiatry and related disciplines that several validated instruments purport to measure the same construct though we may have little idea of how to interpret the results (e.g. what 10 points on this scale can mean, or what two-point decrease on that scale may mean) (Furukawa et al., 2008; Leucht et al., 2005). Arguably SMD is more interpretable than MD in the latter case.

There are however other arguments against SMD (Baguley, 2009; Greenland et al., 1986), most notable of which is the problem arising from range restriction (Bobko et al., 2001). If the sample is restricted to a subset of the population of interest, this will influence the variance. If the sample is chosen to be particularly homogeneous and thus to have a small SD, SMD can be inflated. Such an SMD would probably not be generalizable. On the other hand, in clinical trials,

aiming at a larger SMD is often desirable because then type II errors will be reduced and the sample size required to detect an efficacy signal can be smaller. Inability to detect statistically significant difference between an active arm and a placebo arm is a pressing issue in the field of clinical psychopharmacology (Alphs et al., 2012; Kemp et al., 2010). Need to reduce patient heterogeneity not only in etiology and pathophysiology but also in manifestation is sometimes claimed in this context (Hurko and Ryan, 2005).

One obvious way to increase sample homogeneity and to decrease SD is to raise the severity threshold required of a patient to enter a trial. The present study examines the real-world possibility of increasing sample homogeneity and of thus reducing sample variance and its effect on SMD, based on individual patient data from large clinical trials of the acute phase treatment of schizophrenia.

2. METHODS

The study represents post hoc re-analyses of individual patient data from three large randomized controlled trials of the acute phase treatment of schizophrenia (Breier et al., 2005; Colonna et al., 2000; Tollefson et al., 1997). Table 1 shows the characteristics of the three trials including their interventions, sample size and eligibility criteria.

Breier et al and Tellefson et al used the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) while Colonna et al the Brief Psychiatric Rating Scale (BPRS) (Overall and Gorham, 1962) in rating schizophrenic psychopathology. PANSS has 30 items, each rated between 1 and 7, thus producing the overall score between 30 and 210, whereas BPRS has 18 items only, each rated again between 1 and 7, and its total score ranges between 18 and 126.

In order to model a situation whereby increasingly stricter inclusion criteria are required, we restricted the samples to those scoring 50 or higher, 60 or higher, 70 or higher and so on on

PANSS, or 40 or higher, 50 or higher and so on on BPRS when the latter was the primary outcome measure in the study. We then calculated the SD of PANSS or BPRS scores at baseline in order to ascertain that such requirements effectively decreased the SD.

We examined two outcomes at 4 weeks: the endpoint scores based on the last-observation-carried-forward (LOCF) imputation method, and the percent change from baseline to LOCFed endpoint. Although all the three studies followed up the participants longer than 4 weeks, we chose the 4-week time point as our endpoint because we wanted to minimize the possible bias introduced by LOCF due to dropouts that increased substantially after four weeks and because the differences between the experimental and control interventions were already apparent at this time point.

We used SPSS 18.0 (SPSS Inc., 2009) for statistical calculations of the original individual patient dataset. For SMD, we used Hedge's g. We calculated Kendall's tau to examine the correlation between baseline severity threshold and SDs or SMDs.

3. RESULTS

3.1. Requiring higher baseline threshold effectively decreased SD at baseline.

Table 2 shows the SDs at baseline when we restricted the samples to those scoring 50 or higher, 60 or higher, 70 or higher and so on on PANSS in the case of Tollefson et al's and Breier et al's trials, or 40 or higher, 45 or higher, 50 or higher and so on on BPRS in the case of Colonna et al's trial. It is evident that baseline SDs got progressively smaller as we required a higher and higher threshold as an eligibility criterion for entry into the study. Kendall's tau correlation coefficients between the required thresholds and the SDs were all close to -1.0 and highly significant (p=0.002 or smaller).

3.2. But effect size remained stable.

However, as the baseline SDs decreased with higher and higher inclusion thresholds required, SMDs in terms of LOCFed endpoint scores or in terms of LOCFed %change scores remained by and large constant (Table 3, Figures 1 and 2). Kendall's tau correlation coefficients between the baseline thresholds and the SMDs were all small to moderate, and not statistically significant. Figures 2 and 3 demonstrate clearly that 95%CI of these SMDs overlapped with each other.

This happened so apparently because the SDs of endpoint scores tended to increase and those of %change remained largely stable, as the baseline inclusion thresholds were raised and the baseline SDs became smaller (Table 2).

4. DISCUSSION

Raising the entry threshold, an obvious method to increase sample homogeneity, did reduce the SDs at baseline but either increased or did not change SDs at endpoint. To our own great surprise, SMDs at endpoint remained by and large stable.

In other words, although the baseline SDs were effectively reduced by manipulating the entry criterion, it was as if the endpoint SDs bounced back to the natural SDs. This may therefore represent the regression towards the mean of the variability of symptom severity among a group of patients with schizophrenia. For example, among the 18 trials that were included in Cochrane reviews for olanzapine or amisulpride, the SDs of PANSS total scores at endpoint at the end of the acute phase treatment ranged between 19.7 and 30.2 (mean=24.5) and that for BPRS total scores ranged between 4.4 and 15.1 (mean=12.3) (Duggan et al., 2005; Mota et al., 2002). The endpoint SDs observed in various subsamples in our study were approximately comparable to these other studies. The larger SD of endpoint scores among those with severer psychopathology is also understandable because they may show greater heterogeneity of

response to treatment, while the same appears to be cancelled when we take %change as the outcome measure.

The above-described pattern was most conspicuous with the two olanzapine trials, whereas there was a greater fluctuation in estimates of SMDs in the amisulpride trial. It is to note in this connection that the first two trials used the PANSS whereas the last used the BPRS. Whether the greater fluctuation is due to the smaller sample size of the last trial or to the instability of the measuring instrument it used, namely BPRS, is yet to be researched. As noted above, the SDs of endpoint BPRS scores appeared to show greater variability than those of endpoint PANSS scores among the trials included in the Cochrane reviews (Duggan et al., 2005; Mota et al., 2002) too.

There are obvious limitations to this small study. Firstly, it used individual patient data from three trials examining olanzapine, amisulpride, ziprasidone and holopridol in the acute phase treatment of schizophrenia only. Whether the observed results apply to other antipsychotics and to other mental disorders is open to future studies. The constant SMDs regardless of baseline symptom severity observed in our study may be all the more noteworthy when we consider a number of recent studies of major depression that claimed that the effects of both drugs and psychotherapies were obvious only among the more severely depressed patients (Driessen et al., 2010; Fournier et al., 2010; Kirsch et al., 2008). Secondly, the current study does not rule out the possibility of other methods to increase sample homogeneity to produce smaller SDs and hence larger SMDs. Aetiological and/or pathophysiological subtyping of schizophrenia may be a better way to increase sample homogeneity than a symptomatological approach although we have not been very successful in either approach up to now.

In conclusion, restricting the entry criteria into schizophrenia trials does not appear to be able to increase SMD or thus to decrease the necessary sample size to detect a statistically significant difference in a clinical trial. SMD may be a more generalizable and reproducible index of

treatment efficacy than some critics would imply.

REFERENCES

Alphs, L., Benedetti, F., Fleischhacker, W.W., Kane, J.M., 2012. Placebo-related effects in clinical trials in schizophrenia: what is driving this phenomenon and what can be done to minimize it? The international journal of neuropsychopharmacology / official scientific journal of the Collegium Internationale Neuropsychopharmacologicum, 1-12.

Baguley, T., 2009. Standardized or simple effect size: what should be reported? Br. J. Psychol. 100(Pt 3), 603-617.

Bobko, P., Roth, P.L., Bobko, C., 2001. Correcting the effect size of d for range restriction and unreliability. Organizational Research Methods 4(1), 46-61.

Breier, A., Berg, P.H., Thakore, J.H., Naber, D., Gattaz, W.F., Cavazzoni, P., Walker, D.J., Roychowdhury, S.M., Kane, J.M., 2005. Olanzapine versus ziprasidone: results of a 28-week double-blind study in patients with schizophrenia. The American journal of psychiatry 162(10), 1879-1887.

Cohen, J., 1988. Statistical Power Analysis in the Behavioral Sciences. Erlbaum, Hillsdale, NJ.

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(1), 13-22.

Driessen, E., Cuijpers, P., Hollon, S.D., Dekker, J.J., 2010. Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. J. Consult. Clin. Psychol. 78(5), 668-680.

Duggan, L., Fenton, M., Rathbone, J., Dardennes, R., El-Dosoky, A., Indran, S., 2005. Olanzapine for schizophrenia. Cochrane Database Syst. Rev.(2), CD001359.

Fournier, J.C., DeRubeis, R.J., Hollon, S.D., Dimidjian, S., Amsterdam, J.D., Shelton, R.C., Fawcett, J., 2010. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 303(1), 47-53.

Furukawa, T.A., Jaeschke, R., Cook, D., Guyatt, G., 2008. Measurement of patients' experience, in: Guyatt, G., Drummond, R., Meade, M.O., Cook, D.J. (Eds.), Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd ed. The McGraw-Hill Companies, Inc., New York, pp. 249-271.

Greenland, S., Schlesselman, J.J., Criqui, M.H., 1986. The fallacy of employing standardized regression coefficients and correlations as measures of effect. Am. J. Epidemiol. 123(2), 203-208.

Guyatt, G., Montori, V., Devereaux, P.J., Schunemann, H., Bhandari, M., 2004. Patients at the center: in our practice, and in our use of language. ACP J. Club 140(1), A11-12.

Higgins, J.P., Green, S., 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated September 2011] Available from www.cochrane-handbook.org.

Hurko, O., Ryan, J.L., 2005. Translational research in central nervous system drug discovery. NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics 2(4), 671-682.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr. Bull. 13(2), 261-276.

Kemp, A.S., Schooler, N.R., Kalali, A.H., Alphs, L., Anand, R., Awad, G., Davidson, M., Dube, S., Ereshefsky, L., Gharabawi, G., Leon, A.C., Lepine, J.P., Potkin, S.G., Vermeulen, A., 2010. What is causing the reduced drug-placebo difference in recent schizophrenia clinical trials and what can be done about it? Schizophr. Bull. 36(3), 504-509.

Kirsch, I., Deacon, B.J., Huedo-Medina, T.B., Scoboria, A., Moore, T.J., Johnson, B.T., 2008. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. PLoS Med. 5(2), e45.

Leucht, S., Kane, J.M., Kissling, W., Hamann, J., Etschel, E., Engel, R.R., 2005. What does the PANSS mean? Schizophr. Res. 79(2-3), 231-238.

Mota, N.E., Lima, M.S., Soares, B.G., 2002. Amisulpride for schizophrenia. Cochrane Database Syst. Rev.(2), CD001357.

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

SPSS Inc., 2009. SPSS for Windows Version 18.0. SPSS Inc., Chicago.

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(4), 457-465.

Figure legends

Figure 1. Changes in endpoint SMDs, based on the required severity threshold

Figure 2. Changes in %change SMDs, based on the required severity threshold

Table 1. Characteristics of the three included trials

Study	Antipsychotic drugs and daily dosage (mg)	Sample size	Baseline threshold required	Baseline severity mean (Range)
Tollefson et al 1997	Olanzapine 5-20 Haloperidol 5-20	1337 659	36 or higher on BPRS (scored 1-7) and/or intolerant of current antipsychotic therapy excluding haloperidol	90.9 (Range: 30-166) on PANSS
Breier et al 2005	Olanzapine 10-20 Ziprasidone 80-160	276 269	42 or higher on BPRS (scored 1-7) and 4 or higher on at least one positive symptom item of PANSS and 4 or higher CGI Severity	100.9 (Range: 63-168) on PANSS
Colonna et al 2000	Amisulpride 200-800 Haloperidol 5-20	368 118	4 or higher on at least two of the four BPRS positive items (scored 1-7)	56.2 (Range: 30-104) on BPRS

BPRS: Brief Psychiatric Rating Scale

CGI: Clinical Global Impression

PANSS: Positive and Negative Syndrome Scale

Table 2. Setting lower limits: SDs at baseline, and SDs and SMDs at endpoint

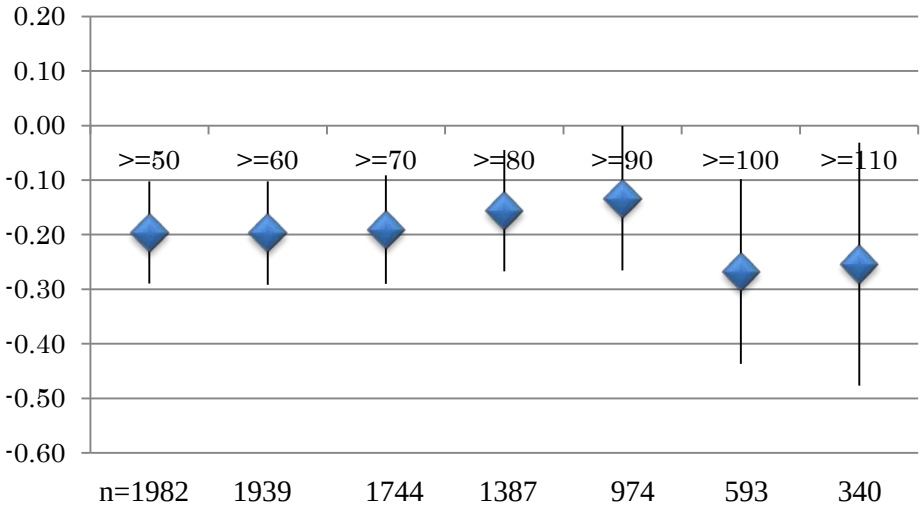
	Entry threshold	SD of baseline scores	SD of endpoint scores	SMD of endpoint scores
Olanzapine vs haloperidol	>=50 (n=1982)	19.2	22.6	-0.20
	>=60 (n=1939)	18.7	22.4	-0.20
	>=70(n=1744)	17.3	22.3	-0.19
	>=80 (n=1387)	15.6	22.6	-0.16
	>=90 (n=974)	14.2	23.4	-0.13
	>=100 (n=593)	12.7	24.3	-0.27
	>=110 (n=340)	11.5	26.0	-0.25
	Kendall's tau	-1.00, p<0.001	0.68, p=0.03	0.00, ns
Olanzapine vs ziprasidone	>=50 (n=545)	20.2	23.6	-0.31
	>=60 (n=545)	20.2	23.6	-0.31
	>=70 (n=536)	19.9	23.7	-0.31
	>=80 (n=473)	18.5	24.0	-0.30
	>=90 (n=372)	17.2	25.0	-0.28
	>=100 (n=244)	15.6	27.0	-0.37
	>=110 (n=156)	13.7	28.0	-0.29
	Kendall's tau	-0.98, p=0.002	0.98, p=0.002	0.31, ns
Amisulpride vs haloperidol	>=40 (n=448)	11.5	13.1	-0.27
	>=45 (n=401)	10.7	14.9	-0.25
	>=50 (n=330)	9.8	15.0	-0.16
	>55 (n=252)	8.7	15.4	-0.30
	>=60 (n=177)	7.7	15.2	-0.09
	>65 (n=125)	6.8	15.1	-0.10
	Kendall's tau	-1.00, p<0.001	0.60, ns	0.47, ns

Table 3. Setting upper limits: SDs at baseline, and SDs and SMDs at endpoint

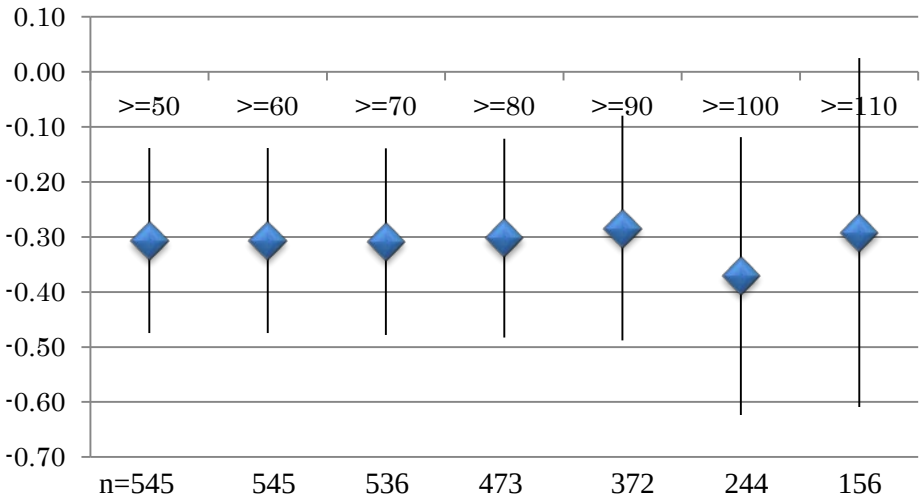
	Entry threshold	SD of baseline scores	SD of endpoint scores	SMD of endpoint scores
Olanzapine vs haloperidol	<=130 (n=1927)	17.3	21.1	-0.18
	<=120 (n=1846)	16.0	20.3	-0.17
	<=110 (n=1670)	13.9	19.2	-0.18
	<=100 (n=1427)	11.9	18.2	-0.16
	<=90 (n=1064)	10.0	16.9	-0.25
	<=80 (n=643)	8.0	15.3	-0.33
	Kendall's tau	-1.00, p<0.001	-1.00, p<0.001	-0.41, ns
Olanzapine vs ziprasidone	<=130 (n=493)	15.2	20.5	-0.27
	<=120 (n=450)	13.0	19.7	-0.27
	<=110 (n=394)	10.8	18.6	-0.30
	<=100 (n=309)	8.7	17.7	-0.24
	<=90 (n=184)	6.4	17.0	-0.31
	<=80 (n=75)	4.1	13.5	-0.45
	Kendall's tau	-1.00, p<0.001	-1.00, p<0.001	-0.56, ns
Amisulpride vs haloperidol	<=80 (n=472)	11.6	14.1	-0.23
	<=75 (n=444)	10.4	13.7	-0.22
	<=70 (n=413)	9.4	13.4	-0.21
	<=65 (n=371)	8.3	12.5	-0.16
	<=60 (n=322)	7.2	11.5	-0.20
	<=55 (n=249)	6.0	10.8	-0.18
	<=50 (n=173)	5.1	10.1	-0.16
	Kendall's tau	-1.00, p<0.001	-1.00, p<0.001	0.78, p=0.02

Figure 1. Endpoint SMDs according to various lower severity thresholds

1a. Olanzapine vs haloperidol



1b. Olanzapine vs ziprasidone



1c. Amisulpride vs haloperidol

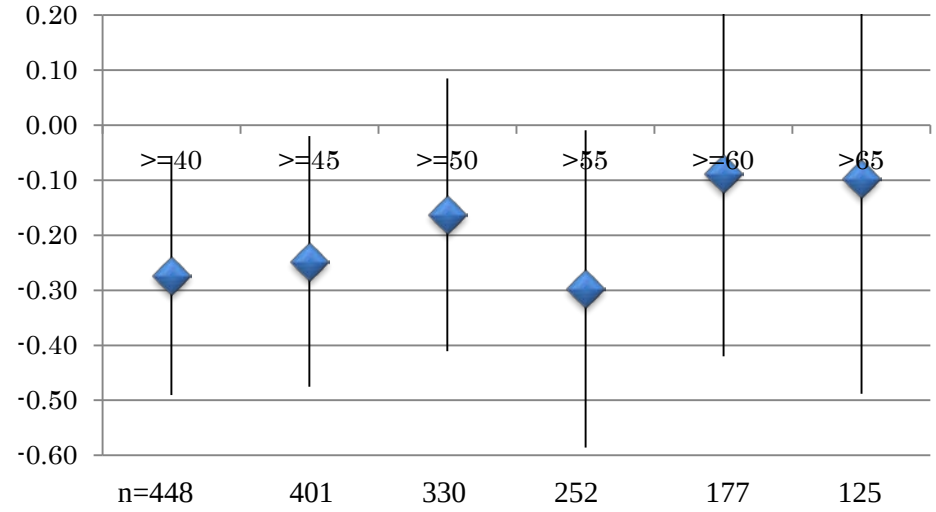
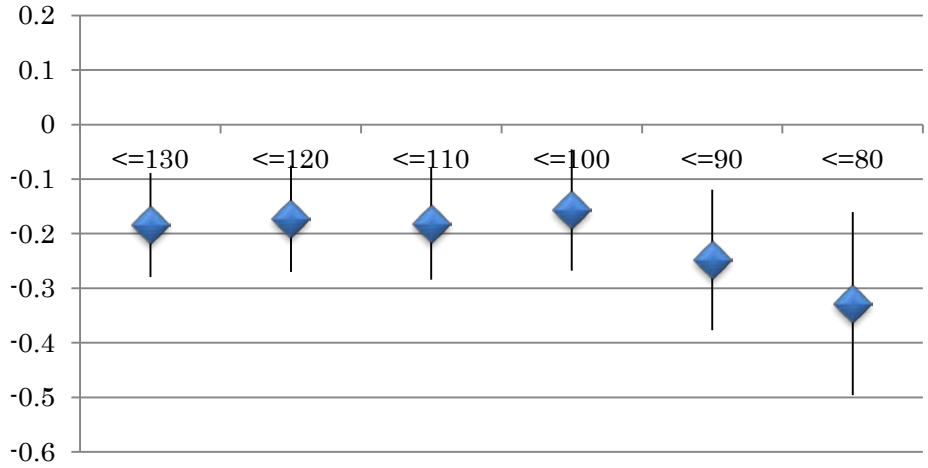
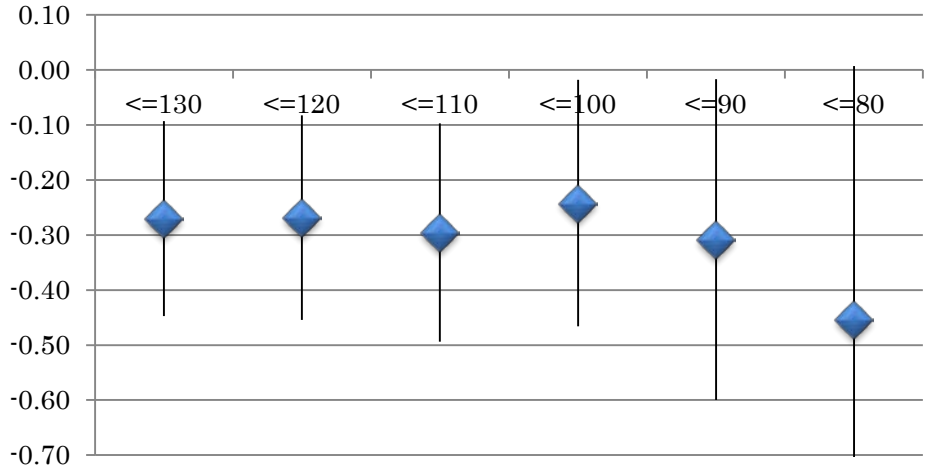


Figure 2. Endpoint SMDs according to various upper severity threshold

2a. Olanzapine vs haloperidol



2b. Olanzapine vs ziprasidone



2c. Amisulpride vs haloperidol