Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression.

OPEN

Molecular Psychiatry (2016) 21, 523–530

www.nature.com/mp

ORIGINAL ARTICLE

Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression

F Hieronymus[1], JF Emilsson[1], S Nilsson[2] and E Eriksson[1]

The recent questioning of the antidepressant effect of selective serotonin reuptake inhibitors (SSRIs) is partly based on the observation that approximately half of company-sponsored trials have failed to reveal a significant difference between active drug and placebo. Most of these have applied the Hamilton depression rating scale to assess symptom severity, the sum score for its 17 items (*HDRS-17-sum*) serving as effect parameter. In this study, we examined whether the negative outcomes of many SSRI trials may be partly caused by the use of this frequently questioned measure of response. We undertook patient-level *post-hoc* analyses of 18 industry-sponsored placebo-controlled trials regarding paroxetine, citalopram, sertraline or fluoxetine, and including in total 6669 adults with major depression, the aim being to assess what the outcome would have been if the single item depressed mood (rated 0–4) had been used as a measure of efficacy. In total, 32 drug-placebo comparisons were reassessed. While 18 out of 32 comparisons (56%) failed to separate active drug from placebo at week 6 with respect to reduction in *HDRS-17-sum*, only 3 out of 32 comparisons (9%) were negative when depressed mood was used as an effect parameter ($P < 0.001$). The observation that 29 out of 32 comparisons detected an antidepressant signal from the tested SSRI suggests the effect of these drugs to be more consistent across trials than previously assumed. Further, the frequent use of the *HDRS-17-sum* as an effect parameter may have distorted the current view on the usefulness of SSRIs and hampered the development of novel antidepressants.

Molecular Psychiatry (2016) **21,** 523–530; doi:10.1038/mp.2015.53; published online 28 April 2015

INTRODUCTION

The future of a potential antidepressant is highly dependent on whether the difference between active drug and placebo with respect to the primary effect parameter reaches significance at the $P < 0.05$-level in at least two independent trials or not.[1] Also, while the existence of negative trials, in addition to those showing a difference, does not prevent a drug from being approved for marketing, the outcome in terms of significance versus non-significance of all trials undertaken may impact how drugs already on the market are being valued; in this vein, the current questioning of the efficacy of the selective serotonin reuptake inhibitors (SSRIs) has to a great extent been spurred by the fact that approximately half of the placebo-controlled SSRI trials conducted by the pharmaceutical companies have failed to show significant superiority of the active drug over placebo.[2] According to critics, the fact that these negative trials have seldom been published has made the scientific community overstate the efficacy of SSRIs,[3,4] some frequently cited debaters even claiming that SSRIs are in fact devoid of any specific, pharmacological antidepressant properties.[2,5]

Needless to say, an effective treatment should be expected to outperform placebo in a vast majority of sufficiently sized trials; the high rate of unsuccessful SSRI trials hence is a matter of legitimate concern. It is therefore important to shed further light on why many trials are negative while others are not, that is, if the negative trials reflect a true inability of SSRIs to generate an antidepressant signal in many cohorts of depressed patients, or if the frequent failures to demonstrate a beneficial effect may be caused by methodological problems.

Constructed in the late 50s,[6] the Hamilton depression rating scale (HDRS) in its most common version (HDRS-17) comprises 17 different symptoms for which a score between 0 and 2 or 0 and 4 may be given. For the majority of antidepressant drug trials, the decrease in the total score of this scale (*HDRS-17-sum*) has served as a primary effect parameter.[7,8] As previously pointed out,[8–11] there are, however, numerous reasons to question the usefulness of this measure.

First, HDRS-17 is multidimensional, indicating that relevant improvement in one domain of symptoms may be masked, due to enhanced variability, by lack of improvement in other possibly less relevant domains.[12–14] Second, the included symptoms differ considerably in terms of burden of illness and many of them correlate poorly with depression severity.[9,15] Third, some items refer to several heterogeneous symptoms, and the different grades for a certain item do not always represent differences in severity but qualitatively distinct phenomena; both these aspects may contribute to the poor inter-rater reliability marring this instrument.[8,16] Fourth, many patients reporting some of the symptoms to be absent already at baseline is bound to reduce the sensitivity of the instrument by enhancing variability, as is the fact that some of the symptoms, such as backaches and headache, are common also in non-depressed subjects, and may therefore be present also after recovery. In line with this, it was recently reported that 25% of

[1]Department of Pharmacology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden and [2]Institute of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden. Correspondence: Professor E Eriksson, Department of Pharmacology, Sahlgrenska Academy, University of Gothenburg, POB 432, Gothenburg SE 405 30, Sweden.
E-mail: elias.eriksson@neuro.gu.se
Received 9 December 2014; revised 10 March 2015; accepted 18 March 2015; published online 28 April 2015

previously depressed subjects scoring 8–12 on HDRS-17 regarded themselves to be in remission.[17]

In spite of these well-established shortcomings of the HDRS-17, the recent debate concerning the efficacy of antidepressants is largely based on the outcome of individual studies or meta-analyses using *HDRS-17-sum* as an effect parameter.[18–22] To explore whether the use of this measure may partly explain why many SSRI trials have been negative, we have re-analyzed eighteen drug company-sponsored depression trials, comprising 32 different comparisons, after replacing *HDRS-17-sum* as an effect parameter with a single item that, unlike many of the other items, is reported by a vast majority of the participants at baseline, that is, depressed mood. To shed additional light on the sensitivity of different measures to detect an antidepressant signal, we also calculated effect sizes, based either on mean effect sizes from each study, or on the pooled population of 6669 subjects, for (i) *HDRS-17-sum*, (ii) various sub-scales and (iii) all 17 items.

MATERIALS AND METHODS

Data acquisition

When designing this study we aimed at including all drug company-sponsored placebo-controlled trials of reasonable size and regarding the treatment of depression in adults that had been conducted for the four major first generation SSRIs, that is, citalopram, fluoxetine, paroxetine and sertraline, at the time of their approval. To this end, we requested patient level-data from all such studies from Lundbeck (Valby, Denmark) (citalopram), GSK (Brentford, UK) (paroxetine), Eli Lilly (Indianapolis, US) (fluoxetine) and Pfizer (New York, US) (sertraline), respectively. While Eli Lilly was unable to provide us with data on fluoxetine trials for practical reasons, the relevant studies not being available in an electronic format, the other three companies granted our request. To preclude any bias in the selection of trials, we confirmed that we had access to all pertinent studies by examining the FDA Approval Packages for citalopram,[23] paroxetine IR,[24] paroxetine CR[25] and sertraline,[26] respectively.

Since the aim of this project was to examine the sensitivity of the HDRS to detect changes between active drug and placebo, studies that did not employ the HDRS-17 or an extended variant thereof were excluded from further analysis. The outcome parameter in focus of this study being that of statistical significance of individual trials, it was also deemed important to exclude trials for which a falsely negative outcome might be expected due to low statistical power; for this reason, an *a priori* decision was made not to include any trial with a sample size of less than 50 subjects in any treatment arm.

In the FDA report on paroxetine IR, two multi-center trials (GSK/002 and GSK/003) were presented as 10 minor studies, but since they were in fact conducted as two large trials, they are included in this analysis as such. GSK also provided data from five additional studies regarding paroxetine IR or paroxetine CR that were not available at the time of FDA approval but did meet the other inclusion criteria. Likewise, two of the placebo-controlled sertraline trials submitted by Pfizer were not mentioned in the FDA report; these were, however, not post-marketing but post-registration trials, that is, completed between the submission of the new drug application to the FDA and its approval.

In total, eight trials regarding paroxetine immediate release (IR), five regarding paroxetine controlled release (CR), three regarding citalopram and five regarding sertraline were eligible for inclusion. In two paroxetine trials and in one sertraline trial, another SSRI, fluoxetine, had served as an active comparator, which enabled us to include also three comparisons of this drug versus placebo.

Statistical analysis

First, analysis of covariance (ANCOVA) was used to calculate levels of statistical significance and effect sizes (as defined as the estimated marginal mean difference between groups divided by the root mean squared error) for all comparisons of active drug versus placebo when using either *HDRS-17-sum* or depressed mood as an effect parameter. Change from baseline to end point with respect to either *HDRS-17-sum* or depressed mood were dependent variables, treatment and study center were fixed effects and baseline severity as assessed using the corresponding scale was included as a covariate. The treatment–center interaction was assessed in all comparisons regarding individual studies, but excluded from the model if non-significant ($P > 0.05$). The ratios of comparisons reaching statistical significance when using *HDRS-17-sum* or depressed mood, respectively, as an effect parameter, were compared using McNemar's test. While differences between the two effect parameters with respect to mean effect sizes were assessed using a paired *t*-test, the ratio of effect size estimates improving by changing from *HDRS-17-sum* to depressed mood was assessed using Pearson's χ^2-test.

While we used ANCOVA to comply with how treatment groups have usually been compared with respect to *HDRS-17-sum* data in antidepressant trials, we acknowledge that the use of this method for comparing groups with respect to the depressed mood item could be questioned given that this item is assessed by an ordinal scale comprising merely five points. To address the possible influence of this aspect, all analyses regarding depressed mood were repeated using ordinal logistic regression.

Second, we wanted to compare the effect size for depressed mood also with those for other possible effect parameters, that is, all individual HDRS-17 items and all HDRS-17 subscales.[9,12–14,27–29] To this end, effect sizes for these parameters were extracted for all 32 drug-placebo comparison using an ANCOVA model composed of change in the measure in question as a dependent variable, treatment and center as fixed factors, and baseline rating of the relevant measure as a covariate. These effect sizes were then compared using repeated measures ANOVA, the model consisting of the effects sizes for all parameters as the within-cases factor and whether a particular comparison was conducted pre- or post FDA approval as a between-cases factor. Following the ANOVA, paired *t*-tests were calculated for depressed mood versus all other individual items, all subscales and *HDRS-17-sum*, and also for all subscales versus *HDRS-17-sum*. Notably, for trials with more than one active treatment arm, the same placebo group was used for all drug-placebo comparisons, the effect size estimates revealed hence not being independent. To verify that this did not bias the results, all analyses were repeated using study-level effect sizes produced by calculating the arithmetic mean of all effect sizes from the same study.

Third, to obtain also patient level-based measures of the effect sizes for all various items and scales, a pooled analysis was conducted on all cases categorized into two groups, that is, patients receiving an SSRI (regardless of drug or dose) and those receiving placebo. Again an ANCOVA model composed of change in the measure in question as the dependent variable, treatment and center as fixed factors, and baseline rating of the relevant measure as a covariate, was used. As for the analyses of the individual studies, all calculations regarding single items were repeated using ordinal regression. As a sensitivity analysis, we also assessed the effect size for *HDRS-17-sum* and depressed mood (using ANCOVA) in the pooled population after having excluded the paroxetine and sertraline trials that had been conducted after FDA marketing approval.

Finally, the observation of negative effect sizes for three individual HDRS-17 items, that is, weight change (significant), gastrointestinal complaints (non-significant) and sexual functioning (non-significant), prompted us to explore to what extent this finding may be partly explained by these symptoms often emerging as side effects of SSRI treatment; to this end, we used χ^2-tests to compare the number of subjects reporting an increase in symptom rating at end point as compared with baseline in those given SSRI and placebo, respectively.

To comply with the procedures applied by the pharmaceutical companies and endorsed by the authorities, all analyses were based on the intention-to-treat (ITT) population, defined as all randomized patients with at least one post-baseline measurement. For those discontinuing prematurely, the last observation carried forward approach was used. To enable comparisons of different trials, the symptom assessment at week 6 was extracted and used as an end point measure. One trial, LB/85A, was a four-week-trial; therefore, the ratings at week 4 were used. For trial GSK/874, no ratings from week 6 were available; hence, week 8 ratings were used.

Ethics

The Regional Ethical Review Board of Gothenburg, Sweden, has reviewed the study protocol and issued an advisory opinion stating no objection to the conduct of this study.

RESULTS

Baseline characteristics of all included trials are displayed in Table 1.

All treatment–center interactions were non-significant ($P > 0.05$) but for study GSK/448, where a significant ($P < 0.001$) interaction

Table 1. Included trials

Protocol	*Group*	n	*% Females*	*% Completers*	*Trial duration*	*Trial period*	*Comments*
GSK/002	PRX IR FLEX	162	49	67	6 Weeks	1985–1987	
	PLA	162	54	54			
GSK/003	PRX IR FLEX	234	53	57	6 Weeks	1985–1986	
	PLA	234	49	47			
GSK/115	FLX FLEX	274	65	76	12 Weeks	1991	
	PRX IR FLEX	272	64	73			
	PLA	111	72	78			
GSK/128	FLX FLEX	343	61	75	12 Weeks	1991	
	PRX IR FLEX	347	61	74			
	PLA	135	71	84			
GSK/251	PRX IR FLEX	120	68	74	8 Weeks	1992–1993	
	PLA	123	64	76			
GSK/448	PRX CR FLEX	102	60	77	12 Weeks	1996–1997	
	PRX IR FLEX	104	64	74			
	PLA	101	66	85			
GSK/449	PRX CR FLEX	108	67	85	12 Weeks	1996–1997	
	PRX IR FLEX	110	74	80			
	PLA	110	60	83			
GSK/487	PRX CR FLEX	103	48	84	12 Weeks	1996–1997	≥ 60 Years
	PRX IR FLEX	103	55	87			
	PLA	107	63	88			
GSK/810	PRX CR 12.5 mg	151	54	89	8 Weeks	2001–2002	≥ 60 Years
	PRX CR 25.0 mg	143	60	83			
	PLA	142	62	86			
GSK/874	PRX CR 12.5 mg	161	59	77[a]	10 Weeks	2003–2004	≥ 60 Years; no evaluation at week 6; week 8 data used.
	PRX CR 25.0 mg	173	60	82[a]			
	PLA	178	64	71[a]			
LB/85A	CIT FLEX	82	35	59[b]	4 Weeks	1984–1985	Four-week trial; week 4 data used.
	PLA	87	33	59[b]			
LB/89303	CIT 20 mg	68	68	75	6 Weeks	1989–1990	
	CIT 40 mg	61	74	80			
	PLA	64	67	72			
LB/91206	CIT 10 mg	125	63	91	6 Weeks	1992–1993	
	CIT 20 mg	128	66	91			
	CIT 40 mg	127	61	91			
	CIT 60 mg	119	51	89			
	PLA	124	57	90			
PZ/103	SER 50 mg	94	52	63	6 Weeks	1984–1985	
	SER 100 mg	93	56	50			
	SER 200 mg	74	64	51			
	PLA	86	44	56			
PZ/104	SER FLEX	142	54	67	8 Weeks	1984–1986	
	PLA	141	42	65			
PZ/109	SER FLEX	104	60	64	8 Weeks	1987–1989	
	PLA	103	66	61			
PZ/111	FLX FLEX	108	67	74	8 Weeks	1989–1990	
	SER FLEX	106	75	64			
	PLA	105	65	65			
PZ/315	SER FLEX	76	66	61	8 Weeks	1984–1986	
	PLA	73	80	63			

Abbreviations: CIT, citalopram; FLEX, flexible dosage; FLX, fluoxetine; PLA, placebo; PRX CR, paroxetine continuous release; PRX IR, paroxetine immediate release; SER, sertraline. % Completers = completers at week 6. [a]Completers at week 8. [b]Completers at week 4.

was found when *HDRS-17-sum* (but not depressed mood) was used as an effect parameter, and for study PZ/315, where a significant interaction ($P = 0.004$) was found for depressed mood (but not for *HDRS-17-sum*). For GSK/448, follow-up analyses revealed this to be primarily due to one center, comprising 18 subjects, that displayed a highly aberrant outcome favouring both paroxetine groups; exclusion of this center rendered the treatment–center interaction non-significant ($P = 0.2$). For PZ/315, sequential exclusion of the two most divergent centers, one favouring sertraline ($n = 10$) and the other favouring placebo ($n = 6$), yielded a non-significant ($P = 0.13$) treatment–center interaction. Subjects from these three centers were omitted from all further analyses, hence yielding an evaluable population of 6669 patients.

Levels of significance and effect sizes for individual comparisons are summarized in Table 2. While 14 out of 32 (44%) comparisons yielded a significant difference when efficacy was measured using *HDRS-17-sum*, 29 out of 32 (91%) comparisons were significant when depressed mood was used as an effect parameter ($P < 0.001$). Likewise, exchanging *HDRS-17-sum* for depressed mood as an effect parameter improved the effect size for 30 out of 32 comparisons ($P < 0.001$), the mean (± s.d.) effect size for depressed mood (0.39 ±0.13) being higher than for *HDRS-17-sum* (0.24 ± 0.13) ($P < 0.001$). Repeating all comparisons in which

Table 2. Effect sizes and levels of significance for all individual drug-placebo comparisons

Trial	*Treatment*	*Baseline mean (s.d.)*		*Endpoint mean (s.d.)*		*Effect size*		P-*value*		
		HDRS-17	*DM*	*HDRS-17*	*DM*	*HDRS-17*	*DM*	*HDRS-17*[a]	*DM*[a]	*DM*[b]
GSK/002	PRX IR FLEX	23.9 (4.2)	2.6 (0.6)	13.9 (8.5)	1.4 (1.1)	0.46	0.56	< **0.001**	< **0.001**	< **0.001**
	PLA	23.6 (3.6)	2.7 (0.6)	17.3 (8.2)	1.9 (1.1)					
GSK/003	PRX IR FLEX	23.5 (3.7)	2.9 (0.5)	14.6 (7.5)	1.8 (1.0)	0.53	0.60	< **0.001**	< **0.001**	< **0.001**
	PLA	23.5 (3.5)	2.9 (0.5)	18.4 (7.7)	2.3 (1.0)					
GSK/115	FLX FLEX	22.5 (3.6)	2.7 (0.7)	14.8 (6.8)	1.6 (1.0)	0.08	0.31	0.50	**0.006**	**0.01**
	PRX IR FLEX	22.5 (3.7)	2.8 (0.7)	15.0 (7.1)	1.6 (1.1)	0.06	0.28	0.58	**0.01**	**0.01**
	PLA	21.8 (4.0)	2.7 (0.6)	14.7 (7.3)	1.9 (1.2)					
GSK/128	FLX FLEX	23.0 (3.7)	2.8 (0.7)	13.4 (7.0)	1.4 (1.1)	0.28	0.40	**0.007**	< **0.001**	< **0.001**
	PRX IR FLEX	23.1 (3.9)	2.9 (0.6)	13.8 (7.1)	1.5 (1.2)	0.22	0.31	**0.03**	**0.003**	**0.003**
	PLA	23.2 (3.7)	2.7 (0.7)	15.5 (7.6)	1.8 (1.2)					
GSK/251	PRX IR FLEX	22.2 (3.5)	2.8 (0.6)	13.7 (6.7)	1.5 (1.1)	0.14	0.32	0.29	**0.01**	**0.01**
	PLA	22.1 (3.3)	2.9 (0.5)	14.5 (7.4)	1.8 (1.1)					
GSK/448	PRX CR FLEX	23.0 (2.6)	2.8 (0.6)	12.5 (6.4)	1.3 (1.0)	0.23	0.52	0.12	< **0.001**	< **0.001**
	PRX IR FLEX	23.4 (2.8)	2.9 (0.6)	14.1 (7.2)	1.6 (1.2)	0.02	0.34	0.90	**0.02**	**0.01**
	PLA	23.4 (2.9)	2.9 (0.6)	14.1 (6.7)	1.9 (1.0)					
GSK/449	PRX CR FLEX	23.8 (3.4)	2.9 (0.6)	12.6 (7.2)	1.4 (1.0)	0.35	0.43	**0.01**	**0.002**	**0.001**
	PRX IR FLEX	23.7 (3.1)	2.9 (0.6)	13.2 (7.2)	1.3 (1.1)	0.25	0.51	0.07	< **0.001**	< **0.001**
	PLA	23.5 (3.1)	2.8 (0.6)	14.8 (6.8)	1.8 (1.1)					
GSK/487	PRX CR FLEX	22.1 (3.5)	2.7 (0.6)	12.2 (6.7)	1.4 (1.0)	0.25	0.34	0.07	**0.01**	**0.01**
	PRX IR FLEX	22.3 (3.1)	2.8 (0.6)	12.1 (6.4)	1.4 (1.0)	0.28	0.30	**0.04**	**0.03**	**0.02**
	PLA	22.1 (3)	2.7 (0.6)	13.6 (6.4)	1.7 (1.1)					
GSK/810	PRX CR 12.5	23.2 (2.9)	2.8 (0.5)	11.7 (6.8)	1.4 (1.0)	0.34	0.29	**0.004**	**0.02**	**0.008**
	PRX CR 25	23.5 (3.3)	2.7 (0.5)	11.1 (7.1)	1.2 (1.0)	0.45	0.50	< **0.001**	< **0.001**	< **0.001**
	PLA	23.8 (3.2)	2.8 (0.5)	14.4 (7.6)	1.7 (1.0)					
GSK/874	PRX CR 12.5	22.6 (3.6)	2.8 (0.5)	13.6 (7.0)	1.6 (1.1)	0.20	0.33	0.07	**0.003**	**0.002**
	PRX CR 25	23.1 (3.9)	2.8 (0.6)	14.1 (6.5)	1.5 (1.0)	0.35	0.46	**0.001**	< **0.001**	< **0.001**
	PLA	22.7 (4)	2.8 (0.6)	17.1 (7.4)	1.8 (1.0)					
LB/85A	CIT FLEX	23.8 (3.2)	2.7 (0.7)	14.4 (7.9)	1.3 (1.0)	0.36	0.51	**0.02**	**0.001**	**0.001**
	PLA	24.0 (3.5)	2.7 (0.7)	14.8 (6.7)	1.8 (1.0)					
LB/89303	CIT 20	24.3 (6.7)	2.8 (0.7)	13.2 (10.6)	1.6 (1.1)	0.02	0.08	0.90	0.67	0.84
	CIT 40	23.0 (6.2)	2.7 (0.7)	9.7 (8.8)	1.1 (1.0)	0.36	0.56	< **0.05**	**0.003**	**0.003**
	PLA	23.7 (6.1)	2.8 (0.7)	13.2 (9.6)	1.7 (1.1)					
LB/91206	CIT 10	22.2 (3.5)	2.9 (0.5)	12.6 (8.1)	1.4 (1.2)	0.19	0.37	0.14	**0.004**	**0.003**
	CIT 20	21.7 (2.9)	2.9 (0.5)	13.0 (7.3)	1.5 (1.2)	0.05	0.30	0.67	**0.02**	**0.006**
	CIT 40	22.1 (3.3)	2.9 (0.5)	11.8 (8.0)	1.3 (1.2)	0.30	0.42	**0.02**	< **0.001**	< **0.001**
	CIT 60	21.8 (3.2)	2.9 (0.5)	11.8 (6.7)	1.2 (1.1)	0.25	0.52	> 0.05	< **0.001**	< **0.001**
	PLA	21.9 (3.4)	2.8 (0.6)	13.6 (6.9)	1.8 (1.1)					
PZ/103	SER 50 mg	24.9 (3.0)	2.9 (0.7)	14.6 (8.8)	1.6 (1.2)	0.30	0.47	< **0.05**	**0.002**	**0.002**
	SER 100 mg	24.7 (2.9)	3.0 (0.7)	15.1 (9.0)	1.7 (1.3)	0.23	0.44	0.12	**0.004**	**0.002**
	SER 200 mg	25.8 (3.4)	3.0 (0.6)	15.9 (9.4)	1.7 (1.2)	0.24	0.46	0.13	**0.004**	**0.003**
	PLA	25.3 (2.9)	3.0 (0.6)	17.4 (8.4)	2.2 (1.2)					
PZ/104	SER FLEX	23.3 (3.7)	2.8 (0.5)	12.5 (8.2)	1.4 (1.1)	0.34	0.38	**0.004**	**0.002**	**0.001**
	PLA	23.4 (3.7)	2.8 (0.6)	15.2 (8.0)	1.8 (1.2)					
PZ/109	SER FLEX	22.0 (3.4)	2.6 (0.6)	13.3 (7.8)	1.4 (1.1)	0.10	0.13	0.48	0.37	0.38
	PLA	21.5 (3.4)	2.7 (0.5)	13.6 (8.0)	1.6 (1.2)					
PZ/111	FLX FLEX	24.4 (2.4)	2.9 (0.5)	13.4 (7.0)	1.4 (1.1)	0.08	0.45	0.56	**0.001**	< **0.001**
	SER FLEX	24.1 (1.9)	2.9 (0.4)	12.8 (6.9)	1.5 (1.1)	0.14	0.43	0.31	**0.002**	**0.001**
	PLA	24.2 (2.2)	2.9 (0.3)	13.8 (7.5)	1.9 (1.0)					
PZ/315	SER FLEX	23.1 (4.2)	2.6 (0.7)	14.5 (8.8)	1.6 (1.0)	0.15	0.10	0.39	0.56	0.39
	PLA	22.2 (4.4)	2.6 (0.8)	15.4 (7.8)	1.8 (1.1)					

Abbreviations: CIT, citalopram; FLEX, flexible dosage; FLX, fluoxetine; PLA, placebo; PRX CR, paroxetine continuous release; PRX IR, paroxetine immediate release; SER, sertraline. For bold values $P < 0.05$. [a]*P*-value obtained from an ANCOVA-model. [b]*P*-value obtained from an ordinal logistic regression model.

depressed mood was the dependent variable using ordinal logistic regression rather than ANCOVA had only minor impact on levels of significance and did not change the rate of positive trials (Table 2).

For the comparison-level analysis of all different effect parameters, Mauchly's test indicated the assumption of sphericity to be violated ($P < 0.001$); however, the *P*-value referring to the difference with respect to effect sizes (the within-cases factor) remained highly significant regardless of method of correction, the F-value being 47.4 (*P*-value after Greenhouse-Geisser correction: < 0.001). In contrast, the factor indicating if the study had been conducted pre- or post-marketing (the between-cases factor) was non-significant ($P = 0.2$) and therefore excluded from the analysis.

Pairwise comparisons showed the mean effect size for the depressed mood item to be significantly higher not only than the mean effect size for *HDRS-17-sum* ($P < 0.001$), hence confirming the outcome of the previous analysis, but also than the mean effect sizes for all other individual items ($P < 0.001$). All four subscales yielded significantly higher effect sizes than *HDRS-17-sum* ($P < 0.001$) but the effect size for depressed mood was significantly higher than those for all subscales ($P < 0.001$). Using

Table 3. Effect sizes and *P*-values for different measures of efficacy in the pooled population.

Measure of efficacy (scoring range)	*Baseline mean (s.d.)*	*Pooled effect size*	*Pooled analysis* P-*value*[a]	*Pooled analysis* P-*value*[b]
HDRS-17-sum (0–52)	23.1 (3.7)	0.27 (0.32[c])	< **0.001**	
Individual items				
Depressed mood (0–4)	2.8 (0.6)	0.40 (0.44[c])	< **0.001**	< **0.001**
Feelings of guilt (0–4)	1.7 (0.7)	0.26	< **0.001**	< **0.001**
Suicide (0–4)	1.1 (0.9)	0.22	< **0.001**	< **0.001**
Insomnia, early (0–2)	1.2 (0.8)	0.08	**0.005**	**0.002**
Insomnia, middle (0–2)	1.3 (0.8)	0.07	**0.009**	**0.005**
Insomnia, late (0–2)	1.2 (0.8)	0.13	< **0.001**	< **0.001**
Work and activities (0–4)	2.7 (0.6)	0.23	< **0.001**	< **0.001**
Retardation (0–4)	1.1 (0.8)	0.21	< **0.001**	< **0.001**
Agitation (0–4)	1.1 (0.9)	0.08	**0.006**	**0.004**
Anxiety, psychic (0–4)	2.2 (0.7)	0.30	< **0.001**	< **0.001**
Anxiety, somatic (0–4)	1.6 (0.8)	0.06	**0.02**	**0.01**
Somatic symptoms, gastrointestinal (0–2)	0.6 (0.7)	−0.02	0.62	0.66
Somatic symptoms, general (0–2)	1.7 (0.5)	0.16	< **0.001**	< **0.001**
Genital symptoms (0–2)	1.3 (0.8)	−0.01	0.70	0.58
Hypochondriasis (0–4)	0.9 (0.9)	0.12	< **0.001**	< **0.001**
Loss of weight (0–2)	0.3 (0.6)	−0.06	**0.04**	0.07
Insight (0–2)	0.2 (0.4)	0.07	**0.02**	**0.02**
HDRS-17 subscales				
Bech (0–22)	12.3 (1.9)	0.35	< **0.001**	
Maier-Phillips (0–24)	11.7 (2.1)	0.35	< **0.001**	
Santen (0–26)	13.4 (2.3)	0.35	< **0.001**	
Gibbons (0–30)	14.6 (2.5)	0.31	< **0.001**	

Mean baseline ratings, effect sizes and *P*-values (SSRI versus placebo) for *HDRS-17-sum*, all individual items and four different subscales. Bech = items: 1, 2, 7, 8, 10, 13; Maier-Phillips = items: 1, 2, 7, 8, 9, 10; Santen = items: 1, 2, 3, 7, 8, 10, 13; Gibbons = items: 1, 2, 3, 7, 9, 10, 11, 14. Data are based on the pooled population. Bold values refer to $P < 0.05$. [a]*P*-value obtained from an ANCOVA-model. [b]*P*-value obtained from an ordinal logistic regression model. [c]Effect sizes within brackets refer to analyses comprising studies accounted for in the FDA report only, that is, after exclusion of the seven post marketing trials.

study-level ($n = 18$) rather than comparison-level effect sizes ($n = 32$) yielded overall higher *P*-values, but all differences remained significant, depressed mood again outperforming all other individual items (highest $P = 0.004$ for psychic anxiety), all subscales (highest $P = 0.01$ for the Bech subscale) and *HDRS-17-sum* ($P < 0.001$).

Table 3 is based on the pooled population and lists effect sizes for (i) *HDRS-17-sum*, (ii) some of the subscales previously proposed[9,12–14,27–29] and (iii) all individual HDRS-17 items. In line with the comparisons of mean effect sizes presented above, all subscales yielded higher effect sizes than the *HDRS-17-sum*; however, for all subscales and for all other individual items, the effect size was lower than it was for the depressed mood item. Exclusion of the seven trials (comprising 12 comparisons) that were not included in the FDA reports since they were completed after the submission of the FDA application yielded somewhat higher effect sizes for both *HDRS-17-sum* and depressed mood, but again the effect size for depressed mood was considerably higher than that for *HDRS-17-sum* (Table 3).

For three items, gastrointestinal complaints, loss of weight and sexual functioning, the effect sizes based on the pooled population were negative (though non-significantly for gastrointestinal complaints and sexual functioning). For all three, an increase in severity at end point as compared with baseline was significantly more common in SSRI-treated patients than in those given placebo, the odds ratios being 1.27 for gastrointestinal symptoms ($P = 0.009$), 1.24 for weight change ($P = .005$) and 1.21 for sexual symptoms ($P = 0.03$).

DISCUSSION

While 18 out of 32 comparisons failed to reveal a significant superiority of the studied SSRI over placebo with respect to reduction in *HDRS-17-sum*, only three out of 32 comparisons were negative with respect to the reduction in depressed mood. The problem of negative SSRI trials, which has been a major argument for the questioning of antidepressants, hence appears to be partly due to the use of an insensitive measure of efficacy.

Our decision to focus on depressed mood, rather than any other item, was motivated by this symptom displaying the largest baseline severity in the pooled population (Table 3), and was further reinforced by the fact that since long it has been attributed particular importance by the FDA when evaluating antidepressant efficacy;[23,25] moreover, it is one of two key symptoms, one of which must be at hand, in the DSM definition of depression. While not claiming that assessing depressed mood only is the optimal way of recording symptom severity, or that other symptoms are irrelevant, we do suggest that a treatment faithfully outperforming placebo in reducing depressed mood can hardly be regarded as ineffective. Also, by demonstrating a consistent reduction in this symptom, our results refute the concern raised by some authors that the superiority of antidepressants sometimes observed also with respect to *HDRS-17-sum* might be due to other influences than a genuine antidepressant effect, such as non-specific sedation.[5]

Many recent attempts to reveal the true efficacy of antidepressants have been based on meta-analyses or patient level-based mega-analyses. While of considerable interest, such approaches are marred by the inevitable problem that individual studies yielding misleading results due to hidden methodological shortcomings, which, given the highly disparate outcome of different trials, are probably not uncommon in this area, may render also the overall outcome misleading. This is why we in this study chose a different approach, that is, to explore the consistency of the ability of SSRIs to generate an antidepressant signal across studies when using a potentially more sensitive measure than the

conventional one. Although it may be argued that reaching a *P*-value of < 0.05 for the primary comparison is an arbitrary definition of efficacy, the actual importance of this criterion for the fate of novel putative antidepressants, and for how established drugs are being valued, is obvious.

Although our primary purpose was to compare the outcome of trials after changing effect parameter from *HDRS-17-sum* to depressed mood, we also compared the mean effect sizes (based on the effect sizes from all individual comparisons) for depressed mood with those for all other individual HDRS-17 items; in addition, the effect sizes for the various uni-dimensional subscales – comprising the depressed mood item and 5-6 additional HDRS-17 symptoms[9,12–14,27–29] – were compared both with the effect size for depressed mood and with that for *HDRS-17-sum*. These analyses revealed all subscales to yield higher effect sizes than did *HDRS-17-sum*; however, the effect size for depressed mood was significantly higher than those for all subscales as well as for all individual items. An analysis of the pooled population, comprising all subjects from all studies, was also undertaken – not because we believe that such an approach would reveal the true magnitude of the actual efficacy of SSRIs, but to obtain a patient level-based assessment of the relative magnitude of different effect sizes. The results obtained by this approach were not markedly different from those based on the means of the effect sizes for the different comparisons, again showing depressed mood to be the parameter resulting in the highest effect size.

Previous reports on effect sizes for the individual items of the HDRS-17 in patients treated with antidepressants are largely in line with our study, depressed mood being one of 3–4 items displaying the largest effect size also in these.[27,30] Minor differences in this regard are, however, notable and likely to be caused both by differences over time with respect to the studied patient populations and by differences in the pharmacological profiles of the tested drugs. For example, when considering that a previous study by Faries *et al*.[27] revealed as high effect size for the items guilt and suicidality as for depressed mood in patients treated with tricylic antidepressants, it should be considered that these symptoms can be assumed to have been fairly common in studies conducted in the tricyclic era, hence making it easier to detect a robust effect of treatment in these days. In contrast, in our population, which is based largely on outpatient studies for which suicidal ideation has been an exclusion criterion, the mean ratings at baseline of both these items, and especially suicidal ideation, were markedly lower than the rating of depressed mood (Table 3). In contrast, the effect size for the symptom early insomnia being higher in the analysis by Faries *et al*.[27] than in the present study may at least partly be explained by SSRIs causing disturbed sleep as a side effect,[31] while tricyclic antidepressants exert a non-specific sedative effect by antagonizing histaminergic H1 receptors.[32]

For three items, that is, weight change, gastrointestinal complaints and sexual functioning, the present analyses revealed negative effect sizes; however, of these effects only that on weight change was statistically significant. The effect sizes of these items being negative is in line with a previous study[30] based on eight comparisons of fluoxetine versus placebo, none of which was included in our analysis as they were conducted by a company producing the non-SSRI antidepressant venlafaxine, fluoxetine merely serving as an active control. For two of the three items for which we observed negative effect sizes, gastrointestinal complaints and weight change, the apparent lack of improvement may be partly explained by the very low rating already at baseline (Table 3) rendering further improvement difficult to detect. For all three items, an important factor may also be that they measure common SSRI side effects that are present also in many patients that have recovered from their illness;[31,33,34] supporting this view, significantly more subjects treated with SSRI than placebo reported aggravation at end point as compared with baseline for these three symptoms. Although the possible negative impact on life quality of these and other side effects must be taken into account when considering the pros and cons of SSRI treatment, the present results underline that including such complaints when assessing efficacy is problematic. Not least for the purpose of future drug development, the importance of separating a drug that is effective but causing side effects from one that is ineffective should be emphasized.

Of note is that Klein and Fink questioned the use of multi-item rating scales for the evaluation of drug treatment in psychiatry already in 1963[35] and that numerous reports since then have demonstrated the insufficient sensitivity marring HDRS-17; that this instrument nevertheless has served as primary effect parameter in a vast majority of antidepressant trials during the past 50 years may seem surprising, not least since it would be in the interest of the pharmaceutical industry to avoid a rating scale not optimizing the chance of detecting a difference between active drug and placebo. It should, however, be considered that the primary purpose of the trials conducted by the pharmaceutical companies has been to obtain marketing approval as rapidly and safely as possible, and that one therefore has been inclined to copy the design from previous studies that have been successful in this regard, for example by using an effect parameter that is known to be acceptable to the FDA, rather than to give priority to innovation. From a scientific perspective, it is, however, imperative that both drug companies and regulatory authorities strive for an improvement with respect to antidepressant trial design, so that future such studies are better suited to reflect the true efficacy of the drugs. Moreover, the present study highlights the importance of extracting as much scientifically relevant information as possible from already conducted studies, rather than to feel restricted by the formal aspects often characterizing evidence based medicine, such as to consider only the outcome parameter named primary in the trial protocol.

It should be emphasized that refraining from using the sum of many disparate items as primary effect parameter in depression trials does not mean that one should not record different symptoms when evaluating the therapeutic profile of an antidepressant drug. Whereas a reasonable conclusion from this and many previous reports would be that *HDRS-17-sum* should no longer be used as a primary effect parameter in antidepressant trials; more research is warranted to shed light on the sensitivity of alternative scales (such as the Montgomery-Åsberg Depression Rating Scale[10] and the various brief variants of the HDRS) on the one hand and various single-item assessments on the other. It should be noted that the inter-rater reliability is often higher for multi-item assessments than for single-item assessments and that Cicchetti and Prusoff[16] have shown this to be the case also for HDRS. For the multi-item approach to be advantageous, all items, however, need to be both reliable and relevant; a high inter-rater reliability hence cannot compensate for poor validity. Of note in this context is that Cicchetti and Prusoff[16] reported depressed mood to be the individudal HDRS-17 item showing the highest inter-rater reliability at end point (0.72) and that its reliability was not markedly lower than that for *HDRS-17-sum* (0.82).

While sufficiently large to justify the use of SSRIs, and larger than when estimated using *HDRS-17-sum*, the effect size obtained when using depressed mood as an effect parameter in the pooled population was merely moderate (0.40 in the total population; 0.44 after exclusion of post-marketing trials). When interpreting these numbers, one should, however, consider (i) that also comparisons with suboptimal dosage were included, (ii) that the analysis is based on symptom rating recorded too early (week 6) for a full effect to be at hand[33] and (iii) that the ITT population comprised subjects dropping out before any marked effect of treatment could be expected. In addition, we were unable to address other methodological factors that may hamper the detection of differences between active drug and placebo, such

as poor compliance, poor rater performance,[16,36,37] artificial over-rating at inclusion[38] and an overly liberal inclusion of participants.[39] Similarly, we were unable to control for a factor suggested to overstate the efficacy of active drugs, that is, the possibility that side effects may make patients and investigators realize that a patient is on active treatment.[2]

While these methodological problems unfortunately make it very difficult to evaluate the actual magnitude of the effect of antidepressants on the basis of modern company-sponsored trials,[40,41] it is nevertheless important that these are designed in a way enabling them to identify drugs generating an antidepressant signal so that novel drugs with antidepressant properties are not mistakenly deemed ineffective. The present data highlight the risk of missing an antidepressant signal by not using a sufficiently sensitive outcome parameter.

Recently mixed-effects repeated-measures models have been used for re-analyses of antidepressant trials to reduce the influence of the drop out factor and non-random missing data.[21,28,42] In this study we refrained from using this strategy, in spite of its obvious virtues, as our aim was to explore to what extent the use of an inadequate primary effect parameter might explain why a large percentage of placebo-controlled trials using SSRIs were deemed negative when analyzed using the conventional statistics mostly applied by the drug companies and requested and approved by the authorities, that is, intention to treatment-based ANCOVA.[23,25]

In conclusion, we suggest that many SSRI trials that have been deemed negative, and often never published, do provide scientifically valid support for the tested drug being antidepressant, the negative outcome being caused by the use of an insensitive measure of improvement. Our observation that, in spite of the many methodological shortcomings marring antidepressant trials, 29 out of 32 comparisons in this analysis (including those with suboptimal dosage) did pick up an antidepressant signal from the tested SSRI suggests the antidepressant effect of these drugs to be highly consistent across trials.

CONFLICT OF INTEREST

Drs Hieronymus, Emilsson, and Nilsson report no potential conflict of interest. Elias Eriksson has been on advisory boards and/or received speaker's honoraria from Eli Lilly and H Lundbeck.

ACKNOWLEDGMENTS

We thank H Lundbeck, GSK and Pfizer for kindly providing us with patient data from the included trials. The study was supported by the Swedish Medical Research Council, Bertil Hållsten's Foundation, Söderberg's Foundation and the Swedish Brain Foundation.

REFERENCES

1 U.S. Food and Drug Administration CfDEaR. Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products, May. 1998. (Accessed 15 August, 2014, at http://www.fda.gov/downloads/Drugs/.../Guidances/ucm078749.pdf.).
2 Kirsch I. *The emperor's new drugs: exploding the antidepressant myth*. Basic Books: New York, NY, 2010.
3 Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine--selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ* 2003; **326**: 1171–1173.
4 Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008; **358**: 252–260.
5 Moncrieff J. Are antidepressants as effective as claimed? No, they are not effective at all. *Can J Psychiatry* 2007; **52**: 96–97.
6 Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry* 1960; **23**: 56–62.
7 Gelenberg AJ, Thase ME, Meyer RE, Goodwin FK, Katz MM, Kraemer HC *et al*. The history and current state of antidepressant clinical trial design: a call to action for proof-of-concept studies. *J Clin Psychiatry* 2008; **69**: 1513–1528.
8 Bagby RM, Ryder AG, Schuller DR, Marshall MB. The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *Am J Psychiatry* 2004; **161**: 2163–2177.
9 Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. *Acta Psychiatr Scand* 1975; **51**: 161–170.
10 Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*1979; **134**: 382–389.
11 Isacsson G, Adler M. Randomized clinical trials underestimate the efficacy of antidepressants in less severe depression. *Acta Psychiatr Scand* 2012; **125**: 453–459.
12 Bech P, Allerup P, Gram LF, Reisby N, Rosenberg R, Jacobsen O *et al*. The Hamilton depression scale. Evaluation of objectivity using logistic models. *Acta Psychiatr Scand* 1981; **63**: 290–299.
13 Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton Depression Rating Scale measure? *J Psychiatr Res* 1993; **27**: 259–273.
14 Maier W, Philipp M. Improving the assessment of severity of depressive states - a reduction of the Hamilton Depression Scale. *Pharmacopsychiatry* 1985; **18**: 114–115.
15 Demyttenaere K, De Fruyt J. Getting what you ask for: on the selectivity of depression rating scales. *Psychother Psychosom* 2003; **72**: 61–70.
16 Cicchetti DV, Prusoff BA. Reliability of depression and associated clinical symptoms. *Arch Gen Psychiatry*1983; **40**: 987–990.
17 Zimmerman M, Martinez J, Attiullah N, Friedman M, Toba C, Boerescu DA. Why do some depressed outpatients who are not in remission according to the hamilton depression rating scale nonetheless consider themselves to be in remission? *Depress Anxiety* 2012; **29**: 891–895.
18 Kirsch I, Deacon BJ, Huedo-Medina TB, Scoboria A, Moore TJ, Johnson BT. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med* 2008; **5**: e45.
19 Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC *et al*. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 2010; **303**: 47–53.
20 Fountoulakis KN, Veroniki AA, Siamouli M, Moller HJ. No role for initial severity on the efficacy of antidepressants: results of a multi-meta-analysis. *Ann Gen Psychiatry* 2013; **12**: 26.
21 Gibbons RD, Hur K, Brown CH, Davis JM, Mann JJ. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry*2012; **69**: 572–579.
22 Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol*2002; **22**: 40–45.
23 U.S. Food and Drug Administration CfDEaR. Celexa NDA 20-822 approval letter, statistical review (part 2). 17 July. 1998. (Accessed August 15, 2014, at http://www.accessdata.fda.gov/drugsatfda_docs/nda/98/020822a_statr_P2.pdf.).
24 U.S. Food and Drug Administration CfDEaR. Paxil NDA 020031 approval letter, statistical review. December 1992.
25 U.S. Food and Drug Administration CfDEaR. Paxil CR NDA 20-982 & 20-936/S008 approval letter, statistical review. 12 February. 2002. (Accessed August 15, 2014, at http://www.accessdata.fda.gov/drugsatfda_docs/nda/2002/20-936S-008_ParoxetineHydrochloride_statr.pdf.).
26 U.S. Food and Drug Administration CfDEaR. Zoloft NDA 019839 approval letter, statistical review. December 1991.
27 Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton Depression Rating Scale. *J Psychiatr Res* 2000; **34**: 3–10.
28 Mallinckrodt CH, Meyers AL, Prakash A, Faries DE, Detke MJ. Simple options for improving signal detection in antidepressant clinical trials. *Psychopharmacol Bull* 2007; **40**: 101–114.
29 Ruhe HG, Dekker JJ, Peen J, Holman R, de Jonghe F. Clinical use of the Hamilton Depression Rating Scale: is increased efficiency possible? A post hoc comparison of Hamilton Depression Rating Scale, Maier and Bech subscales, Clinical Global Impression, and Symptom Checklist-90 scores. *Compr Psychiatry* 2005; **46**: 417–427.
30 Entsuah R, Shaffer M, Zhang J. A critical examination of the sensitivity of unidimensional subscales derived from the Hamilton Depression Rating Scale to antidepressant drug effects. *J Psychiatr Res* 2002; **36**: 437–448.
31 Ferguson JM. SSRI Antidepressant Medications: Adverse Effects and Tolerability. *Prim Care Companion J Clin Psychiatry* 2001; **3**: 22–27.
32 Richelson E. Tricyclic antidepressants and histamine H1 receptors. *Mayo Clinic Proc*1979; **54**: 669–674.

33 Ferguson JM, Feighner JP. Fluoxetine-induced weight loss in overweight non-depressed humans. *Int J Obes* 1987; **11**: 163–170.
34 Montgomery SA, Baldwin DS, Riley A. Antidepressant medications: a review of the evidence for drug-induced sexual dysfunction. *J Affect Disord* 2002; **69**: 119–140.
35 Klein DF, Fink M. Multiple Item Factors as Change Measures in Psychopharmacology. *Psychopharmacologia* 1963; **4**: 43–52.
36 Kobak KA, Leuchter A, DeBrota D, Engelhardt N, Williams JB, Cook IA *et al*. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. *J Clin Psychopharmacol*2010; **30**: 193–197.
37 Engelhardt N, Feiger AD, Cogger KO, Sikich D, DeBrota DJ, Lipsitz JD *et al*. Rating the raters: assessing the quality of Hamilton rating scale for depression clinical interviews in two industry-sponsored clinical drug trials. *J Clin Psychopharmacol*2006; **26**: 71–74.
38 Landin R, DeBrota DJ, DeVries TA, Potter WZ, Demitrack MA. The impact of restrictive entry criterion during the placebo lead-in period. *Biometrics* 2000; **56**: 271–278.
39 Liu KS, Snavely DB, Ball WA, Lines CR, Reines SA, Potter WZ. Is bigger better for depression trials? *J Psychiatr Res* 2008; **42**: 622–630.
40 Parker G. Antidepressants on trial: how valid is the evidence? *Br J Psychiatry*2009; **194**: 1–3.
41 Horder J, Matthews P, Waldmann R. Placebo, prozac and PLoS: significant lessons for psychopharmacology. *J Psychopharmacol*2011; **25**: 1277–1288.
42 Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. *J Biopharm Stat* 2001; **11**: 9–21.