Chapter 2

Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders

Annelieke M. Roest, Peter de Jonge, Craig D. Williams,
Ymkje Anna de Vries, Robert A. Schoevers, Erick H. Turner

JAMA Psychiatry (2015), 72 (5), 500 - 510

Abstract

Importance: Previous studies have shown that the scientific literature has overestimated antidepressant efficacy for depression, but other indications have not been considered.

Objective: To examine reporting biases in clinical trials of antidepressants for anxiety disorders, and to quantify the extent to which these biases inflate estimates of drug efficacy.

Data sources and study selection: Reviews of premarketing trials for 9 second-generation antidepressants were obtained from the Food and Drug Administration (FDA). A systematic search for matching publications was performed using PubMed, EMBASE and Cochrane CENTRAL.

Data extraction and synthesis: Double data extraction was performed for the FDA reviews and the journal articles. Hedges' g was calculated as measure of effect size.

Main outcomes and measures: Reporting bias was classified as study publication bias, outcome reporting bias, or spin. Separate meta-analyses were conducted for the two sources and meta-regression was used to assess the impact of publication status on effect estimates.

Results: Sixteen of 57 (28%) trials were not positive according to the FDA, while only 2 of 45 (4%) published article conclusions were not positive ($p<0.001$). Positive trials were 5 times more likely to be published in agreement with the FDA determination compared to trials determined not-positive (risk ratio=5.20; 95% CI: 1.87 - 14.45; $p<0.001$). We found evidence for study publication bias ($p<0.001$), outcome reporting bias ($p=0.02$), and spin ($p=0.02$). The pooled effect size based on the published literature ($g=0.38$; 95% CI: 0.33 - 0.42; $p<0.001$) was 15% higher than the effect size based on the FDA data ($g=0.33$; 95% CI: 0.29 - 0.38; $p<0.001$), but this difference was not statistically significant ($p=0.18$).

Conclusions and relevance: Various reporting biases were present for trials on the efficacy of FDA-approved second-generation antidepressants for anxiety disorders. Although this did not significantly inflate estimates of drug efficacy, reporting biases led to significant increases in the number of positive findings in the literature.

Introduction

There is strong evidence that significant results from randomized controlled trials are more likely to be published than nonsignificant results [38]. As a consequence, the published literature, including meta-analyses, may overestimate the benefits of treatment while underestimating its harms, thus misinforming clinicians, policy makers, and patients [101].

Different types of reporting biases can be present. Study publication bias occurs when studies with positive results are more likely to be published than studies with negative results [102]. Outcome reporting bias involves publishing outcomes from a study that are "positive" (e.g., statistically significant) without publishing "negative" outcomes or switching the status of primary and secondary outcomes based on results [103]. Finally, spin occurs when treatments are described by investigators as beneficial, even though published results for primary outcomes are nonsignificant [40].

The registry and results database of the Food and Drug Administration (FDA) can be used to assess the degree to which published trial results may overestimate efficacy [19, 104, 105, 106]. Pharmaceutical companies must register all trials they intend to use in support of an application for US marketing approval with the FDA, and information on these trials is compiled in this database. A previous study found that 51% of trials of antidepressants for major depressive disorder were deemed positive by the FDA compared to 94% in the published literature; in addition, a meta-analysis of only published data overestimated the effect of antidepressants by 32% [19]. This was followed by debate and additional research on the efficacy of antidepressants for depression [72, 101, 107, 108].

Antidepressants are widely prescribed for conditions other than depression [109], including anxiety disorders. However, research on reporting biases for these other indications is lacking. Anxiety disorders are common in the general population with an estimated year prevalence of 12% [110]. Second-generation antidepressant drugs, namely selective serotonin reuptake inhibitors (SSRIs) and serotonin norepinephrine reuptake inhibitors (SNRIs), are the primary pharmacological treatments for generalized anxiety disorder (GAD) [13, 111], panic disorder (PD) [13, 112], social anxiety disorder (SAD) [113], post-traumatic stress disorder (PTSD) [114], and obsessive compulsive disorder (OCD) [115].

Several meta-analyses have reported that second-generation antidepressants are superior to placebo in the treatment of GAD [116, 117], PD [118, 119], SAD [120, 121], PTSD [122] and OCD [123]. Some of these meta-analyses suggested the existence of study publication bias based on funnel plot asymmetry [118, 120]. However, such methods cannot prove the existence of publication bias; for that, one must access and analyze unpublished data as well [106]. A recent study examined the efficacy of one SSRI in the treatment of GAD and PD using a complete dataset of trials sponsored by the manufacturer. This study indeed showed that published trials had significantly larger effect sizes than unpublished trials [77].

In the present study, the first objective was to examine reporting bias in the scientific literature on efficacy of second-generation antidepressants that are FDA-approved for the treatment of anxiety disorders. By comparing published articles with the corresponding FDA reviews we examined the presence of study publication bias, outcome reporting bias, and spin. The second objective was to compare the magnitude of the overall effect based on published trial data from premarketing trials with that based on the full cohort of such trials registered with the FDA.

Methods

As in previous studies [19, 104], we began by identifying the inception cohort of premarketing trials for the indications of interest, then conducted a literature search for those trials.

Data from FDA reviews

We identified the phase 2/3 clinical double-blind placebo-controlled trials registered with the FDA and conducted in pursuit of marketing approval of second-generation antidepressants for the treatment of the following five disorders: (1) GAD, (2) PD, (3) SAD, (4) PTSD, and (5) OCD. Nine drugs, approved by the FDA for these indications, were examined: seven SSRIs (paroxetine, paroxetine controlled release [CR], sertraline, fluoxetine, fluvoxamine, fluvoxamine CR, and escitalopram) and two SNRIs (venlafaxine extended release [ER] and duloxetine). We retrieved the FDA Drug Approval Packages (aka FDA reviews) from the FDA's website; if these were not available for download, we requested them from the FDA's Freedom of Information Office (http://www.accessdata.fda.gov/scripts/foi/FOIRequest/requestinfo.cfm).

We extracted the results the FDA used to decide whether the trial was positive, i.e. whether it could be used to support marketing approval. Data were extracted preferably from the statistical review, but also from the medical review and administrative correspondence (e.g. memos by team leader). In cases where multiple primary endpoints were identified in a trial, results were extracted for the endpoint that was most consistent with the primary endpoint identified in other trials for the same indication.

In accordance with previous publications [19, 104], the FDA's regulatory decisions were classified as (1) positive (clearly supporting efficacy), or (2) not positive, with the latter including both questionable (neither clearly positive nor clearly negative) and negative trials (not supportive of efficacy).

The questionable category included trials characterized by the FDA as "marginally" or "borderline" positive. These were trials that had non-significant p values for one or more

of the primary endpoints, but were considered by the FDA to be supportive of other positive trials because of significant findings on secondary variables. The questionable category also included "failed" trials (in which neither the study drug nor the active comparator demonstrated statistical superiority to placebo).

For multiple-dose trials, we used the FDA's overall decision on the trial. For purposes of meta-analysis, we extracted data only for approved dosages, thus excluding "subtherapeutic" dosages [19]. Data extraction, classification, and data entry was performed independently by two investigators (AR and CW) with discrepancies resolved by consensus (AR, CW, ET).

Data from journal articles

Having identified the inception cohort of premarketing trials registered with the FDA, we systematically searched for matching publications using PubMed, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL) without language restrictions, with a search cutoff date of December 19, 2012. We searched the title field for the name of the drug and the type of anxiety disorder, and any field for the word "placebo". For example, when searching PubMed for relevant escitalopram trials for GAD, the search syntax was "escitalopram[Title] AND (generalized[Title] OR generalised[Title]) AND anxiety[Title] AND disorder[Title] AND placebo."

Publication matches for trials registered with the FDA were identified using the following information: drug name, name of the active comparator (if applicable), dosage groups, sample sizes, trial duration, and names of investigators. The preferred type of publication was a stand-alone publication, i.e. a full-length article devoted to reporting the results of a single trial. If no stand-alone publication could be found, then pooled analyses were sought in which multiple trials were covered in a single article. Data from journal articles that pooled data from multiple trials that were not identical in design according to the FDA were excluded from this study. Pooled-trials publications were also excluded when one or more of the included trials were published earlier as stand-alone publications and the pooled-trials publication did not present separate results for the included trials. Finally, data published only in abstract form were excluded.

Several steps were taken to minimize the possibility that we missed matching publications. If no publication was found via the electronic database search, PubMed was used to identify the three most recent review articles focusing on the efficacy of the trial drug for the condition treated in the trial. The reference lists for those publications were hand searched. In addition, the drug sponsor's website was searched for bibliographic information on the trials in question.

To assess drug efficacy according to published journal articles, we used the primary endpoint specified in the publication. If a primary endpoint was not specified and if no

endpoint was clearly emphasized, we extracted the drug-placebo comparison reported first in the text of the results section or in the table or figure first cited in the text [104]. If multiple endpoints were identified as primary in a single study, results were extracted for the endpoint reviewed as primary by the FDA. Data extraction and entry was done independently by AR and RS with discrepancies resolved through consensus (AR, RS, YV).

In addition, each article's conclusion was classified as positive or not-positive (including questionable and negative) based on the sentence in the abstract reporting the authors' overall conclusion regarding study outcome. Conclusions were classified independently by AR and PJ, who was blinded to the results of the FDA review.

Statistical analysis

All statistical analyses were performed using STATA 11.0. The binomial probability test was used to assess whether the proportion of positive conclusions in journal articles was significantly different from the proportion of positive trials according to the FDA. In addition, we examined whether not-positive trials (according to the FDA) were more likely to be unpublished, or published in a positive manner, compared with positive trials using Fisher's exact test. The presence of study publication bias (trial results not published), outcome reporting bias (changes in analysis or primary endpoint affecting significance of findings), and spin (abstract conclusion not consistent with published results on primary endpoint) was also compared for positive and not-positive trials.

We conducted two meta-analyses: one using data from the FDA reviews and another one using the corresponding published data [19, 104]. Hedges' g was used as the measure of effect size and was calculated using the following equation in which t represents the t statistic and n_1 and n_2 are the numbers of subjects in the drug and placebo groups, respectively:

$$g = t \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The values for g were adjusted using Hedges' correction for small sample size [19, 104]. The t statistic was calculated from the precise p value and the trial sample size using Microsoft Excel's TINV function, multiplying t by -1 when the study drug was inferior to placebo. If a precise p value was not available because it was reported as a range (e.g. $p<0.05$), the t statistic was calculated from other summary statistics, namely standard deviations, standard errors, and 95% confidence intervals around the mean difference. When the data were presented as dichotomized statistics, Hedges' g was calculated from χ^2 [122]. If none of these data were available, and FDA and journal data were otherwise congruent, data were imputed with data extracted from the other source. Additionally,

for two journal articles Hedges' g was calculated from the F statistic (analysis of variance) [124]. Finally, p values and other efficacy data were not reported for 2 negative FDA trials that were, in one case, not published and, in the other case, published as positive. These p values were imputed with p=0.396, which was derived from 16 nonsignificant but precise P values, according to the method previously described by Turner et al. [19].

For each multiple-dose study, we computed a single study-level effect size using a fixed effects model to pool the values from that trial's multiple treatment arms. When calculating the standard error, each trial's shared placebo n was counted once, rather than redundantly, for each dose group to avoid a spuriously low standard error. A limitation of this method is that it only partially addresses error due to correlation between the comparisons [102]. Calculations of all effect sizes were performed independently by AR and YV.

The random effects pooling method was used to generate summary estimates of Hedges' g. I^2 and confidence intervals around I^2 were calculated to assess heterogeneity [125]. I^2 reflects the proportion of total variance explained by heterogeneity. Meta-regression, using the restricted maximum likelihood method, was conducted to examine the impact of publication status on the effect estimates. In addition, pre-specified subgroup analyses were performed for each anxiety disorder.

Results

FDA reviews

We analyzed 9 second-generation antidepressants for data related to the 5 anxiety disorders. Within those 45 possible drug-indication combinations, 21 are FDA-approved. Of those, we were able to download 9 FDA approval packages through the FDA website; for the remaining 12, we made requests to the FDA Freedom of Information Office. Of these, the FDA Freedom of Information Office fulfilled 11 requests — the FDA informed us that the drug approval package for fluoxetine for panic disorder would not be available for at least 18 months. This left 20 approval packages in this study. These drug approval packages, which were issued between 1994 and 2008, reviewed the results of 57 randomized, placebo-controlled short-term trials.

Journal articles

For the 57 above-mentioned FDA-registered trials, we identified 52 trials published in 48 publications. Three of these articles were excluded from further analyses. As a result, 3 additional trials were judged to be not fully published. Two articles pooled trials which were not identical in design [126, 127] and another pooled-trials article failed to present

separate results for the included trials [128] and included a trial that was previously published as a stand-alone publication [129].

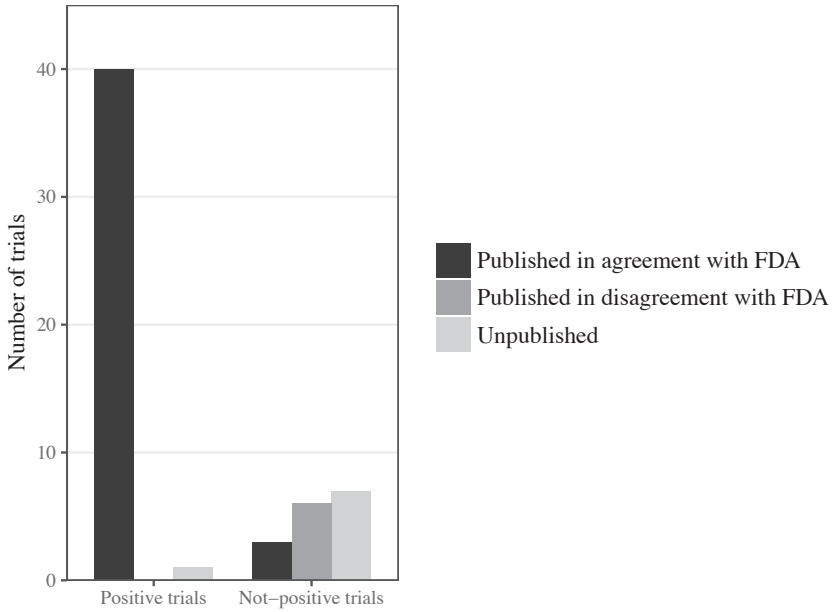

Figure 2.1: *Publication status of positive and not-positive FDA trials*

Trial outcome versus published results

The proportion of positive findings was 72% (41/57) according to the FDA versus 96% (43/45) according to the published literature. This difference was statistically significant (binomial test p<0.001).

Of the 41 positive trials, 40 (98%) were published in agreement with the FDA (Figure 2.1). By contrast, of the 16 not-positive trials, only 3 (19%) were published in agreement with the FDA. This difference was statistically significant (Fisher's exact test p<0.001). Overall, trials that the FDA judged as positive were 5 times more likely to be published in agreement than FDA-not-positive trials (risk ratio: 5.20; 95% CI: 1.87 - 14.45; p<0.001).

Study publication bias

Seven of the 16 (44%) not-positive trials were not published, while only 1 of the 41 (2%) positive trials was not published (Table 2.1). This difference was statistically significant (Fisher's exact test p<0.001).

Table 2.1: *Characteristics of included premarketing trials*

			N				
Disorder	Drug	Trial	Placebo	Drug	Outcome	FDA	Bias
GAD	Escitalopram	MD-05 [130]	128	124	HAM-A	P	-
		MD-06 [130]	138	143	HAM-A	P	-
		MD-07 [131]	153	154	HAM-A	P	-
	Paroxetine	641 [132]	180	385	HAM-A	P	-
		642 [133]	163	161	HAM-A	P	-
		637 [N/A]	183	181	HAM-A	N	SPB
	Duloxetine	HMBR [134]	173	334	HAM-A	P	-
		HMDT [135]	158	161	HAM-A	P	-
		HMDU [136]	158	149	HAM-A	P	-
	Venlafaxine ER	210 [137]	96	253	HAM-A	P	-
		214 [138]	98	174	HAM-A	P	-
PD	Paroxetine	120 [139]	69	72	% 0 attacks	Q	ORB
		108 [140]	60	60	% 0 or 1 attack	P	-
		187 [141]	123	123	% 0 attacks	P	-
		223 [N/A]	68	77	% 0 attacks	Q	SPB
	Paroxetine CR	494 [142]	129	122	% 0 attacks	P	-
		495 [142]	136	123	% 0 attacks	P	-
		497 [142]	130	132	% 0 attacks	N	-
	Sertraline	629 [143]	87	79	# of attacks	P	-
		630 [144]	88	88	# of attacks	P	-
		529 [129]	44	127	# of attacks	Q	ORB
		514 [N/A]	38	112	# of attacks	N	SPB
	Venlafaxine ER	398 [145]	154	315	% 0 attacks	P	-
		399 [146]	157	316	% 0 attacks	P	-
		353 [147]	155	155	% 0 attacks	Q	-
		391 [148]	168	160	% 0 attacks	N	Spin
SAD	Fluvoxamine CR	3107 [149]	125	110	LSAS	P	-
		3108 [150]	148	126	LSAS	P	-
	Paroxetine	502 [151]	145	136	LSAS	P	-
		382 [152]	92	90	LSAS	P	-
		454 [153]	92	268	LSAS	P	-
	Paroxetine CR	790 [154]	184	185	LSAS	P	-
	Sertraline	R-0601 [155]	196	205	LSAS	P	-
		94-004 [156]	69	134	BSPS	P	-
		95-003 [157]	196	191	CGI-L	N	ORB
	Venlafaxine ER	387 [158]	138	133	LSAS	P	-
		393 [159]	135	126	LSAS	P	-
PTSD	Sertraline	641 [160]	82	84	CAPS-2	N	-
		682 [N/A]	94	94	CAPS-2	N	SPB
		640 [161]	104	98	CAPS-2	P	-
		671 [162]	90	93	CAPS-2	P	-
	Paroxetine	651 [163]	167	322	CAPS-2	P	-

continued

Table 2.1: *Characteristics of included premarketing trials*

Disorder	Drug	Trial	N		Outcome	FDA	Bias
			Placebo	Drug			
		648 [164]	133	136	CAPS-2	P	-
		627 [N/A]	159	154	CAPS-2	Q	SPB
OCD	Fluoxetine	HCEP 1 [165]	47	139	Y-BOCS	P	-
		HCEP 2 [165]	41	122	Y-BOCS	P	-
		E079 [166]	56	158	Y-BOCS	N	Spin
	Fluvoxamine	5529 [N/A]	80	79	Y-BOCS	P	-
		5534 [167]	77	78	Y-BOCS	P	-
	Fluvoxamine CR	3103 [168]	119	113	Y-BOCS	P	-
	Paroxetine	116 [169]	88	166	Y-BOCS	P	-
		118 [N/A]	75	79	Y-BOCS	N	SPB
		136 [170]	99	198	Y-BOCS	P	-
	Sertraline	237/248 [171]	44	43	Y-BOCS	Q	Spin
		371/372 [172]	84	240	Y-BOCS	P	-
		546 [173]	79	85	Y-BOCS	P	-
		495 [N/A]	87	83	Y-BOCS	N	SPB

The FDA column indicates the Food and Drug Administration (FDA) decision (P: positive; N: negative; Q: questionable). Type of bias includes study publication bias (SPB), outcome reporting bias (ORB), and spin; "-" indicates that no bias was present. Other acronyms – BSPS: Brief Social Phobia Scale; CAPS-2: Clinician-Administered PTSD Scale, part 2; CGI-L: Clinical Global Impressions-Liebowitz; GAD: generalized anxiety disorder; HAM-A: Hamilton Rating Scale for Anxiety; LSAS: Liebowitz Social Anxiety Scale; OCD: obsessive compulsive disorder; PD: panic disorder; PTSD: posttraumatic stress disorder; SAD: social anxiety disorder; Y-BOCS: Yale-Brown Obsessive Compulsive Scale

Outcome reporting bias

For 3 of the 16 not-positive trials (19%), results were published with a conclusion that conflicted with that in the FDA review, changing their effects from nonsignificant to statistically significant. By contrast, outcome reporting bias was found in none of the 41 FDA-positive trials (Table 2.1). The difference in proportions was statistically significant (Fisher's exact test p=0.02).

One of the 3 above-mentioned publications (trial 120, paroxetine for PD) presented only observed-cases analyses for the primary outcome [139]; according to the FDA, the primary analysis involved last-observation-carried-forward (LOCF) analyses, the results of which were not statistically significant.

In the article presenting results of trial 529, data from subjects with PD who were randomized to different dosages of sertraline were pooled and compared to the placebo group, yielding a significant result [129]; the FDA review showed that the primary results for each of the individual dosage groups were nonsignificant.

Finally, one article presenting the results of trial 95-003, which compared the effect of sertraline (with and without exposure therapy) to placebo (with and without exposure therapy) in patients with SAD, combined scores on three endpoints (disorder-specific Clinical Global Impression Scale [severity and improvement] and Social Phobia Scale) in response versus non-response categories [157]; the FDA review showed that the primary endpoint was the severity total score of the disorder-specific Clinical Global Impression Scale and that this was nonsignificant.

Spin

Spin was present in an additional 3 out of 16 (19%) of the not-positive trials and not present for positive trials (Fisher's exact test p=0.02) (Table 2.1). Each of these 3 articles [166, 171, 148] reported that the primary endpoint was nonsignificant in the results section but, in the abstract, concluded that the trial was positive. The FDA classified these trials as questionable (trial 237/248: sertraline for OCD) or negative (trial E079: fluoxetine for OCD and trial 391: venlafaxine ER for PD). Conclusions on study drug efficacy for these trials, according to the FDA and the authors of the journal articles, are included in Table 2.3 in the Appendix.

Meta-analysis

The pooled effect size based on the FDA data was 0.33 (95% CI: 0.29 - 0.38; p<0.001). Heterogeneity was moderate (I^2=39%; 95% CI: 15% - 56%). For trials published in agreement with the FDA review results, the pooled effect size (Hedges' g=0.38; 95% CI: 0.34-0.42; p<0.001) was larger than the pooled effect size of trials that were not published or published in disagreement with the FDA conclusion (g=0.17; 95% CI: 0.09 - 0.26; p<0.001). Meta-regression showed this difference to be statistically significant (g=0.21; 95% CI: 0.12 - 0.30; t=4.61; p<0.001).

The pooled effect size based on the published literature was 0.38 (95% CI: 0.33 - 0.42; p<0.001). Heterogeneity was low (I^2=30%; 95% CI: 0% - 51%). This effect size represented a 15% increase in effect size compared with the value based on the FDA data. This difference was not statistically significant by meta-regression (g=0.04; 95% CI: -0.02 - 0.10; t=1.36; p=0.18).

Effect sizes based on data from the FDA reviews were 0.32 for GAD, 0.28 for PD, 0.27 for PTSD, and 0.39 for both OCD and SAD. For all disorders the pooled effect sizes of trials published in agreement with the FDA review results were larger than the pooled effect sizes of trials that were not published or published in disagreement with the FDA conclusion (Table 2.2). As a result forest plots for all disorders showed fewer nonsignificant trials according to the published literature than according to the FDA, especially for PD

Table 2.2: *Meta-analysis*

	FDA					
	Total	Published in agreement	Not published in agreement	Journal	Overestimate	
Disorder	Hedges' g (95% CI)	Hedges' g (95% CI)	Hedges' g (95% CI)	Hedges' g (95% CI)	%	*p*
GAD	0.32 (0.25-0.39)	0.34 (0.28-0.41)	0.11 (-0.09-0.31)	0.34 (0.27-0.41)	6%	0.65
PD	0.28 (0.20-0.36)	0.33 (0.25-0.41)	0.13 (-0.02-0.29)	0.35 (0.24-0.46)	25%	0.38
SAD	0.39 (0.30-0.49)	0.43 (0.35-0.50)	0.09 (-0.11-0.29)	0.42 (0.35-0.49)	8%	0.56
PTSD	0.27 (0.11-0.44)	0.33 (0.14-0.53)	0.13 (-0.12-0.38)	0.32 (0.14-0.50)	19%	0.76
OCD	0.39 (0.30-0.49)	0.44 (0.33-0.54)	0.30 (0.11-0.48)	0.45 (0.35-0.56)	15%	0.42
Overall	0.33 (0.29-0.38)	0.38 (0.34-0.42)	0.17 (0.09-0.26)	0.38 (0.33-0.42)	15%	0.18

FDA: Food and Drug Administration; GAD: generalized anxiety disorder; OCD: obsessive compulsive disorder; PD: panic disorder; PTSD: posttraumatic stress disorder; SAD: social anxiety disorder.

and OCD (Figures 2.2 and 2.3 [see Figures 2.4, 2.5, and 2.6 in the Appendix for GAD, SAD, and PTSD]).

Effect sizes based on the literature were larger for all disorders as compared with effect sizes based on the FDA reviews, with the smallest increases for GAD (g=0.34, 6% increase) and SAD (g=0.42, 8% increase) and larger increases for OCD (g=0.45, 15% increase), PTSD (g=0.32, 19% increase) and PD (g=0.35, 25% increase). However, the differences in effect estimates based on the journal articles and the FDA reviews were not statistically significant for any of the individual disorders (Table 2.2).

Discussion

This study showed the presence of reporting bias in randomized controlled trials on the efficacy of second-generation antidepressants for anxiety disorders. Trials that the FDA judged to be positive were over 5 times more likely to be published in agreement with the FDA analysis than not-positive trials. As a result, 96% (43/45) of the journal articles were framed positively, while 72% (41/57) of the trials were deemed positive by the FDA. All examined reporting biases were present among the included trials, namely study publication bias, outcome reporting bias, and spin.

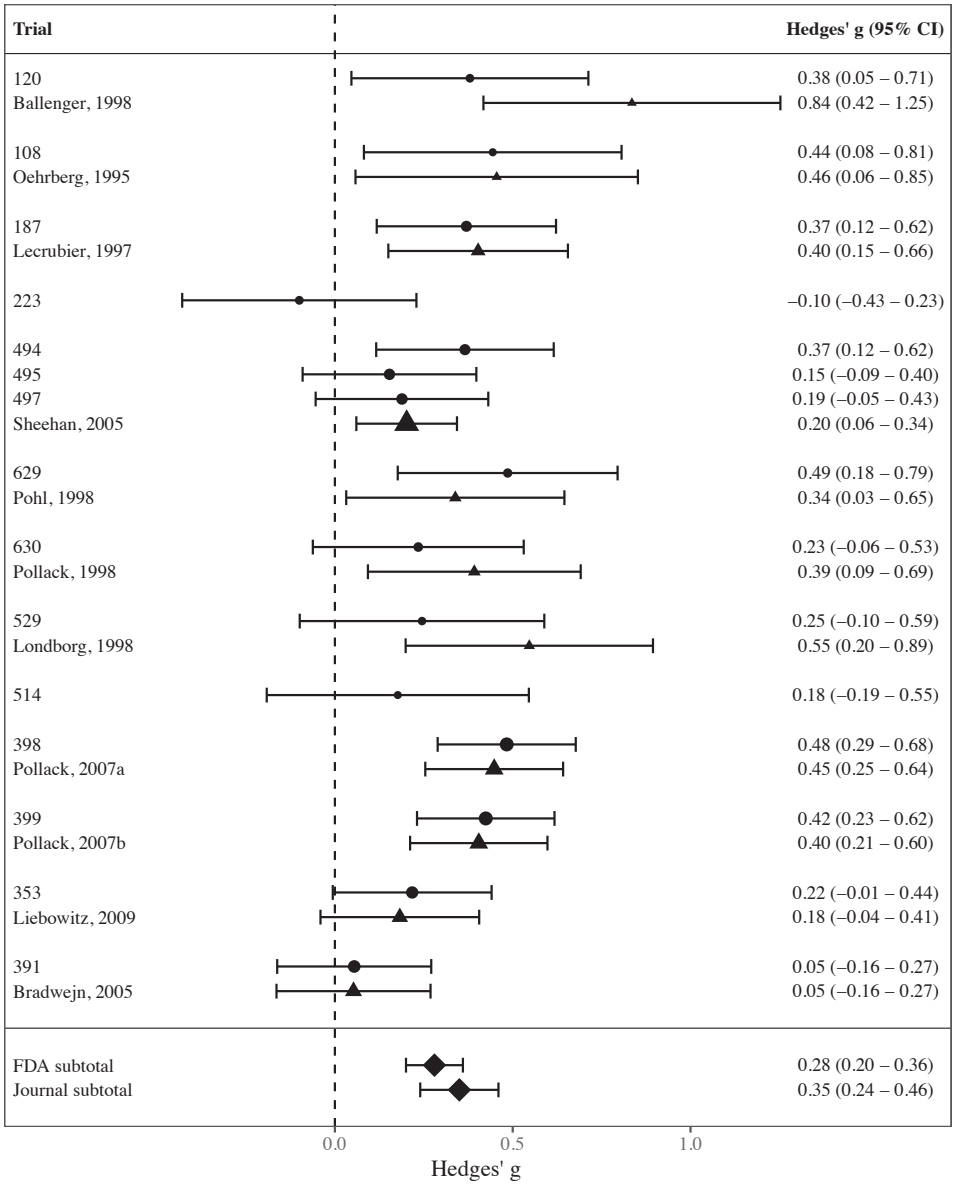

Figure 2.2: *Forest plot for PD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

In a previous study that examined reporting bias in trials on second-generation antidepressants for major depressive disorder [19], the overall effect size based on the FDA data was 0.31, quite comparable to the effect size of 0.33 found in this study. After conducting two meta-analyses, one based on data from the FDA reviews and the other based on data from the corresponding journal articles, we found that reporting bias inflated the

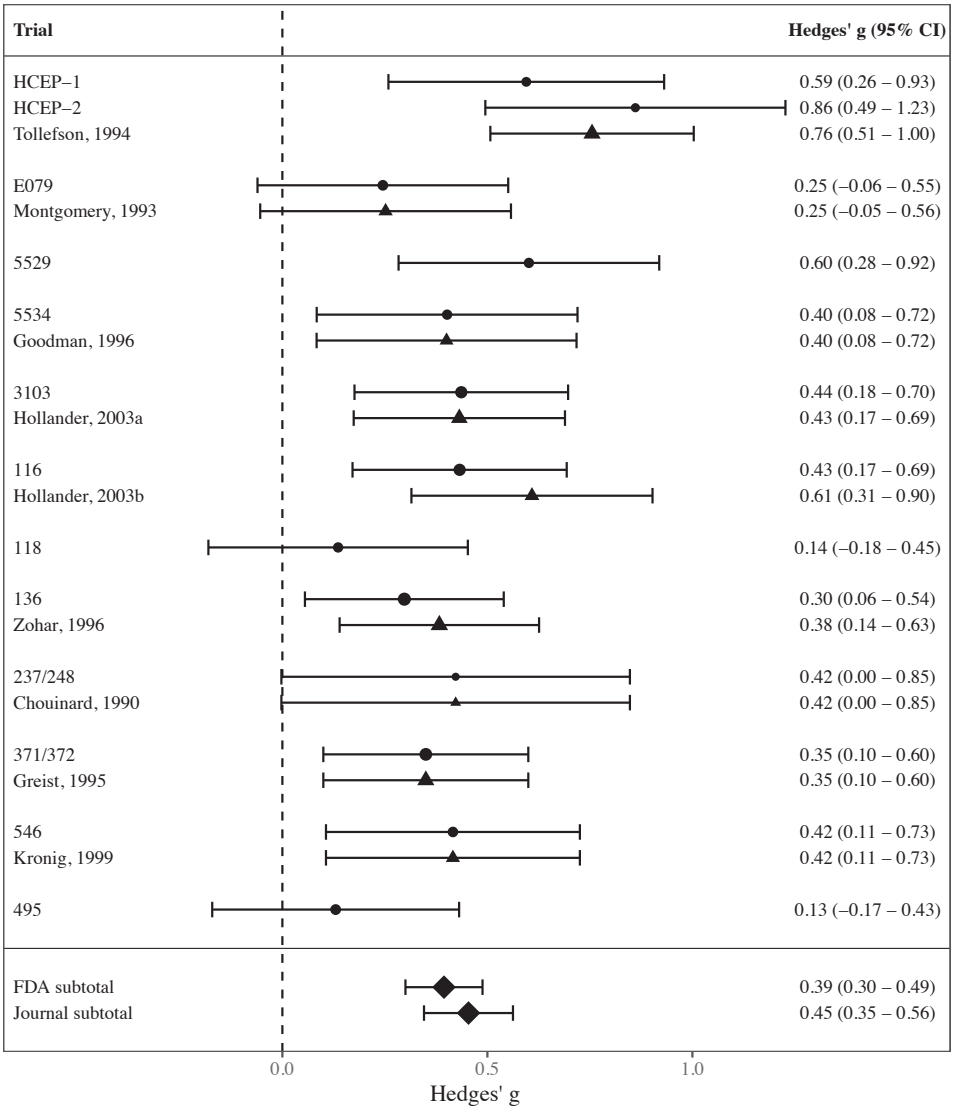

Figure 2.3: *Forest plot for OCD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

apparent effect size by 15%. This increase was not statistically significant, in contrast to the larger inflation factor (32%) found earlier with major depressive disorder [19]. For the individual anxiety disorders, the inflation factors ranged from 6% (GAD) to 25% (PD), indicating the importance of using unbiased data in meta-analyses on the efficacy of second-generation antidepressants for anxiety disorders.

In the main analyses we combined five disorders classified as anxiety disorders in DSM-

IV; however, in DSM-V, OCD is now classified under obsessive-compulsive and related disorders, while PTSD is under trauma- and stressor-related disorders. Therefore, with the recent change in taxonomy, our grouping of these disorders could be viewed as a limitation, although efficacy of the drugs was comparable across disorders.

A clearer limitation of the current study is that the trials for the individual anxiety disorders were few in number, decreasing the power of the subgroup analyses. An additional limitation is that we did not examine biased reporting of harm outcomes, which figures into the overall risk-benefit ratio of a drug, but such an examination would have been beyond the scope of the current study.

Certain data available only in pooled-trials publications were classified as unpublished, which could also be viewed as a limitation. However, a study of antidepressant trials submitted to the Swedish drug regulatory authority showed that positive trials were more likely to be published as stand-alone publications, while negative trials tended to be reported only within pooled-trials publications [27]. Pooled analyses may not follow the predetermined analysis plan and power calculation and can therefore yield different conclusions than the original trials. Pooled analyses are also associated with "salami slicing" (publishing similar results from one study in multiple publications) [174]. Therefore, although these publications may provide new information, especially on subgroups and secondary endpoints, they are susceptible to bias [175]. This bias can be reduced by first publishing the original trial results. Future research could assess the bias that is introduced by pooled-trials publications.

Finally, we did not contact drug sponsors to ask whether specific trials were published in the scientific literature, so there is a small chance that trials could have been misclassified as unpublished. However, considering the extensive literature search methods, it seems unlikely that such trial publications would be discoverable by the typical health care professional.

A strength of this study is that, for 20 of the 21 FDA-approved drug-indication combinations, we were able to include data from all premarketing randomized controlled trials, thereby allowing a reliable assessment of different reporting biases for these trials. However, it is important to note that we could not include data from rejected drug-indication applications because the FDA does not release these reviews [105]. It is likely that the amount of reporting bias that was found would increase if these trials were to be examined as well.

In addition, our estimation of the amount of reporting bias present might also be influenced by the fact that all trials were sponsored by pharmaceutical industries. Yet reporting bias is not restricted to pharmacological treatments sponsored by drug companies [176]. Since reporting bias has been shown for the treatment of depression with psychotherapy [177], it should be worthwhile to systematically assess reporting bias in trials using psychotherapy for anxiety disorders as well.

Spin can result from different, intentional or unintentional, strategies, for example by focusing on secondary endpoints for which significant results were obtained [178]. Journal articles for which spin was identified also often reported "marginally significant results" (p values between 0.05 and 0.10) in the present study. Ideally, interpretation of trial results should not be based solely on a p value indicating whether results are statistically significant or not [40]. In addition to providing p values, future research could consider including Bayes factors as a measure of the strength of the evidence. Bayes factors stem from Bayesian statistics and have the advantage that they can express the strength of the evidence on a continuous scale [179].

Reporting bias significantly increased the number of positive versus negative publications in the literature in the present study. This likely impacts clinician's perceptions of the efficacy of these drugs, which could reasonably be expected to affect prescription behavior. In both Europe and in the US, use of antidepressants has been rising markedly over the last two decades with much of that use appearing to be driven by non-specialists in settings of primary care [180, 181]. Although it should be noted that these studies could not take into account indications for which the drugs were prescribed, a realistic view of the efficacy of these agents is important across all indications. Results of the current study and other studies comparing published results to data from FDA reviews or other registries [19, 77] can perhaps assist clinicians in gaining a more realistic view of the evidence for the efficacy of antidepressants in the (short-term) treatment of affective disorders.

This study adds to the growing body of literature establishing the pervasiveness of reporting bias [34, 176, 182]. It also highlights the need to address this problem using various measures, as recently reviewed [183]. One suggested approach, which would address outcome reporting bias and spin (but not study publication bias), would require peer reviewers to make preliminary decisions based on the strength of the methods in the original trial protocol [176] so that their decisions are not influenced by the statistical significance of study results [184]. Use of study registries, like ClinicalTrials.gov, can also reduce reporting bias in the scientific literature [176], but this registry does not yet function optimally. For example, for the majority of trials subjected to mandatory reporting within one year following trial completion, results were not posted within this timeframe [185].

In summary, although the majority of trials on the efficacy of FDA-approved second-generation antidepressants for anxiety disorders were positive, various reporting biases were present. These reporting biases led to an overly positive representation of significant findings in the scientific literature.

Appendix

Table 2.3: *FDA conclusion vs. journal conclusion for articles with spin*

Drug	Disorder	Trial	FDA	Literature
Venlafaxine ER	PD	391	"The results of study [391]* do not provide adequate evidence of the anti-panic efficacy of venlafaxine ER versus placebo over 10 weeks of treatment"	"Venlafaxine ER seems to be effective and well tolerated in the short-term treatment of PD"
Fluoxetine	OCD	E079	"This trial failed to show significant fluoxetine-placebo differences on any of the scales. A few contrasts were marginal, with p-values between 0.05 and 0.10, but considering the number of scales and number of tests done, I attach no importance to them."	"This study supports the growing evidence for the safety and efficacy of fluoxetine in the treatment of OCD."
Sertraline	OCD	237/ 248	"Although the plots of group means over time appear to show that sertraline beats placebo, differences between group means were inconsistent among the various time points, with significance appearing somewhat early in the study and frequently disappearing towards week 8. It would be difficult to characterize these results as positive, and at best we might call them supportive. Calling it a failed study may be more accurate."	"Results of the Y-BOCS total score, the NIMH score, and the global severity and improvement scores demonstrated a statistically significant superiority of sertraline compared with placebo."

*ER: extended release; FDA: Food and Drug Administration; OCD: obsessive compulsive disorder; PD: panic disorder. FDA conclusions were extracted from the medical review (trial 391) or the statistical review (trials E079 and 237/248). Journal article conclusions were extracted from the article abstract. * Number 353 was changed to 391 since there appears to be a typographical error in the FDA medical review (the conclusion regarding trial 353 is included on another page of the review).*

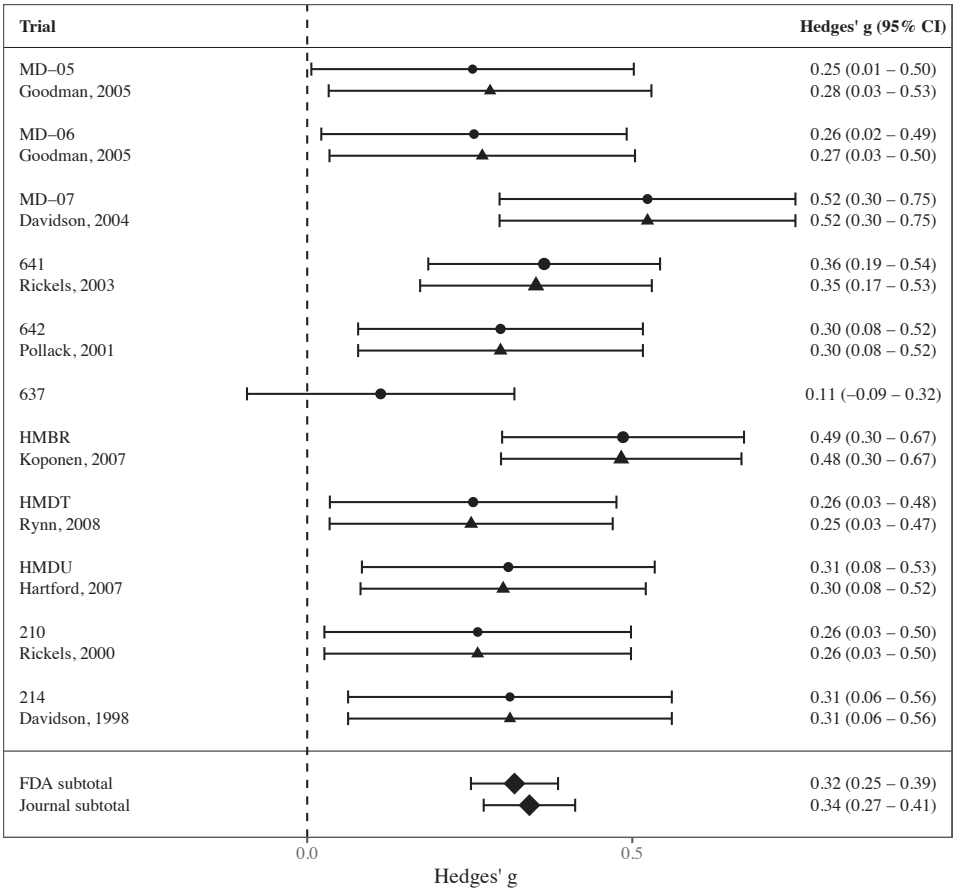

Figure 2.4: *Forest plot for GAD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

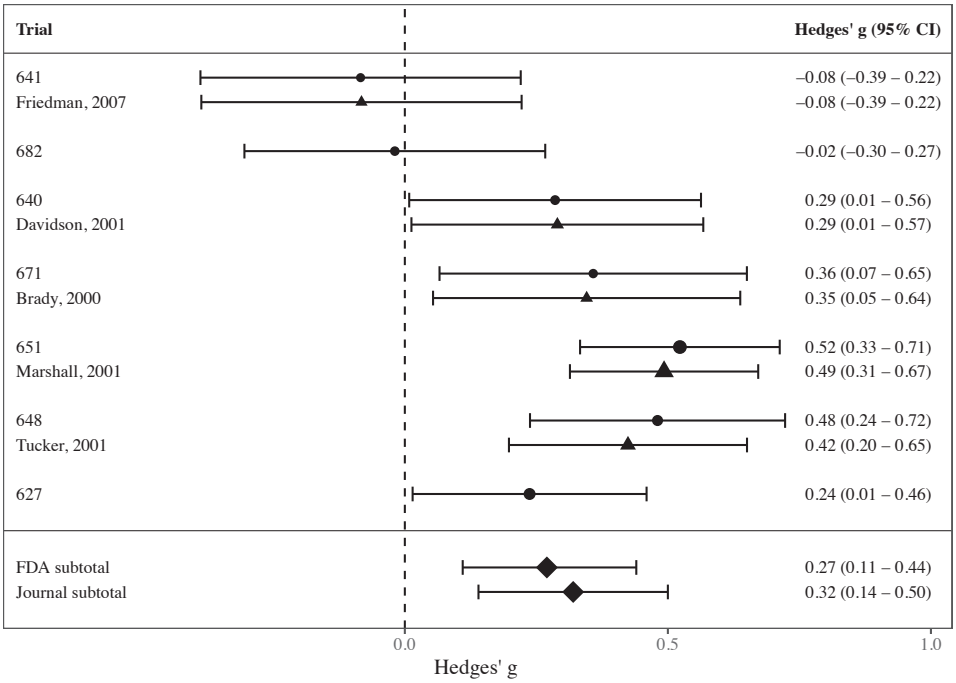

Figure 2.5: *Forest plot for PTSD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

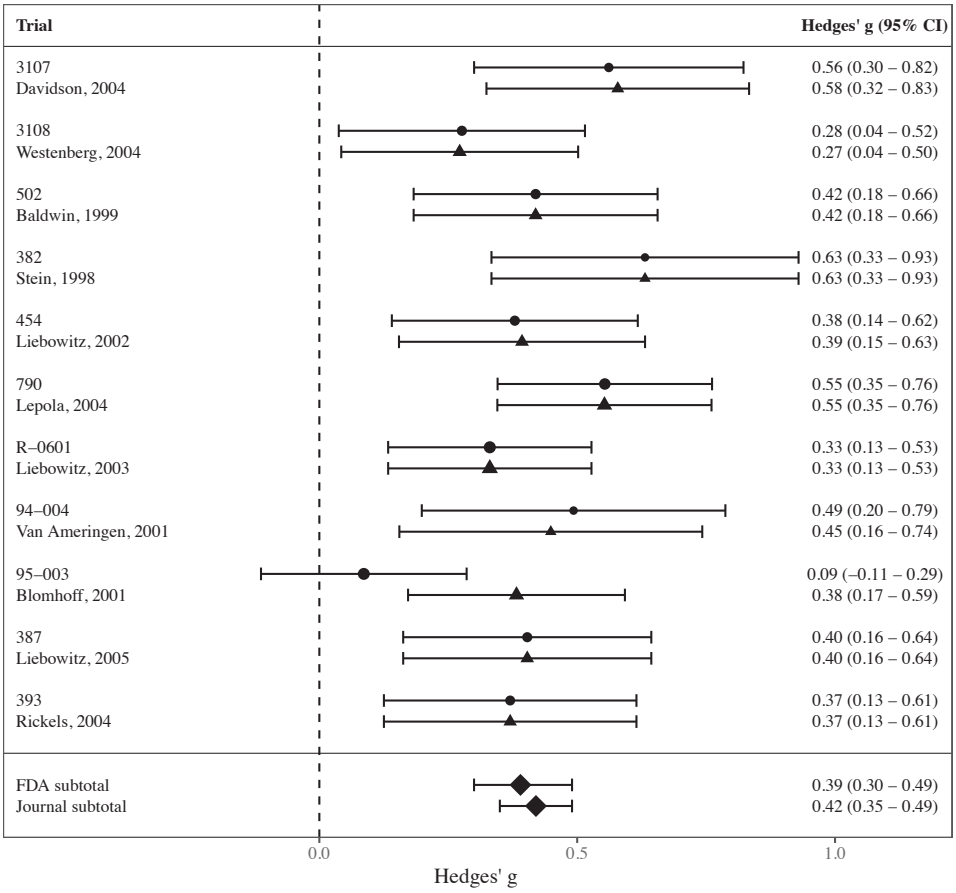

Figure 2.6: *Forest plot for SAD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*