Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

SUPPLEMENTARY APPENDIX

This document provides details that supplement the material in the body of the article but could not be included due to space constraints. The headings used below mostly follow those in the main article.

BACKGROUND

- In order to approve a drug, the FDA requires "substantial evidence" of effectiveness from two positive "adequate and well controlled studies".[1] The language that the FDA uses in labeling reflects this focus on positive trials. For example, the labeling for sertraline states, "The efficacy of Zoloft as a treatment for major depressive disorder was established in two placebo-controlled studies in adult outpatients meeting DSM-III criteria for major depressive disorder."

METHODS

DATA FROM FDA REVIEWS

Data procurement – FDA reviews

- Review documents for many approved drug-indication combinations have been electronically available in the public domain since enactment of the Electronic Freedom of Information Act Amendments of 1996 (eFOIA).[2] Reviews for drugs approved before that are not available electronically. For reasons that are unclear, at the FDA has not, as of this writing, posted the reviews of three antidepressants approved after eFOIA was enacted: bupropion extended-release (Wellbutrin XL®), mirtazapine orally disintegrating (Remeron Soltabs®), and transdermal selegiline (EMSAM®).

- Fluoxetine studies 62-a and 62-b are referred to in the FDA review with a single study number (62). However, the review indicates that these were identically designed but that they were separate and nonoverlapping studies. The first study involved patients with mild depression (but who still met DSM-criteria for major depression), while the other involved patients with moderate depression.

Data extraction – FDA reviews

- The FDA uses statistical superiority (P<.05) to a comparator, usually placebo, to determine whether the study is positive[3] (aka "a win"). Studies are pooled or meta-analyzed only if that is specified in the original protocol and agreed to in advance by the FDA.

- Failed versus negative studies: Active comparator treatment arms are sometimes included in study designs along with study drug and placebo. Because active comparators are approved antidepressants, they are expected to beat (demonstrate statistical superiority to) placebo. When that does not happen and the study drug also does not beat placebo, the FDA "excuses" the study drug for not beating placebo and deems the study inconclusive or "failed".[4] On the other hand, when an active comparator *does* beat

1

placebo (as expected) but the study drug does not, the study is judged negative. When the sponsor elects to omit an active comparator from the design and the study drug does not beat placebo, the study is also deemed negative.[1] However, the validity of distinguishing between negative and failed studies has been questioned.[5]

- If the reviewer's overall judgment regarding a study's outcome was not clearly stated, we used that of the team leader or division director.

- Double data extraction and entry – FDA reviews: This was performed first by ET, AM, and EL. A second extraction and entry of the FDA data was performed by RT and SR, who were blind to the results of the first extraction/entry process. The values obtained in the second process were compared to those obtained in the first. Any discrepancies were resolved by consensus.

DATA FROM JOURNAL ARTICLES

Data procurement – journal articles

- **Literature search**: Literature searches were originally conducted by ET, AM, and EL. Subsequently, an academic reference librarian (AH), who was blind to the results of the original searches, conducted independent searches of Ovid Medline and Cochrane Central Register of Controlled Trials, which identified no new journal articles. A repeat search of reference lists revealed one additional article not indexed in the searched databases.

- We did not consider forms of data disclosure such as conference proceedings (including published conference abstracts), clinical trial registries, book chapters, or newspaper articles.

- We allowed one exception to our rule of including only stand-alone publications of single studies, a paper pooling the results from two identically designed studies of paroxetine controlled-release (Appendix Table A).

- Matching of studies in FDA reviews to journal articles was completed initially by ET, AM, and EL. It was later repeated by RT, who was blind to the results of the original matching process. Any discrepancies between the first and second matching procedures were resolved by consensus.

Data extraction – journal articles

- Double data extraction and entry: Journal data extraction and entry was initially performed by ET. Subsequently, we provided the matched journal articles (clean

---

[1] Unwritten policy learned during the first author's tenure as a reviewer in the division of the FDA handling approval of psychotropic drugs, still applicable according to communications with current employees and apparent from review documents. This may not apply to other review divisions within the FDA.

unmarked copies) to TB and NM, who were blind to the results of the original extraction/entry process. They extracted the results on the apparent primary endpoints, which were entered by NM and RT. The results of this second extraction/entry process were compared with those from the first process. Discrepancies were resolved by consensus.

STATISTICAL ANALYSIS

Continuous data (effect size)

- Using effect size permitted us to combine data from trials using different primary rating scales in a standardized way.

- The meta-analyses described were performed twice, once based on the first data extraction/entry process and again based on the second such process (see above). The overall mean weighted ES values (Figure 3) resulting from the two sets of meta-analyses were within ±.01 of one another. The percentage difference, between the overall weighted mean ES value obtained from the FDA data and the value obtained from the journal data, was unchanged.

- For purposes of the paired (signed-rank) analyses of the g values, we conducted a preliminary "mini-meta-analysis" of the FDA data for paroxetine studies 448 and 449 (Appendix Table A). This provided a single FDA result, which we compared with the single pooled result reported in the journal article.

- As stated in the Methods section of this article, within the FDA dataset, we calculated a mean g value for each drug's published studies and a mean g value for each drug's unpublished studies. However, 2 of the 12 drugs had no unpublished studies; thus a value for $g_{unpublished}$ could not be calculated for these two drugs. We first analyzed the 10 complete pairs of $g_{published}$ and $g_{unpublished}$ values. We followed this with an ancillary analysis, using 12 pairs by imputing $g_{unpublished}=g_{published}$ for these two drugs, assuming the null hypothesis held.

Handling of imprecise P values

- Precise P values were not always available for the calculation of effect size. In cases where they were not available, we looked for data from which to calculate precise P values. We found means and standard deviations (or standard errors or confidence intervals) in the FDA reviews for 9 treatment arms within 6 studies and for 17 treatment arms within 10 journal articles. These instances are shown in Appendix Table A with the superscript "M". For one article[6] whose apparent primary endpoint was based on the proportion of responders, we calculated a precise P value by using the reported number of responders in a replication of the authors' chi-squared test. This instance is shown in Appendix Table A with the superscript "R".

- If no such data were available, we set the P value equal to the precise P value obtained from the other data source. The purpose of this approach was to create a conservative bias in the direction of the null hypothesis, i.e. of *not* finding a difference between FDA and journal results reporting. As an example, if $P_{sponsor}$ was stated as <.05, but P_{FDA} was

given as .03, we set $P_{sponsor}$=.03 and used that to calculate $g_{sponsor}$. Likewise, if $P_{sponsor}$ was reported as "NS" while P_{FDA} was .24, we set $P_{sponsor}$=.24, also. Each such instance is shown in Appendix Table A with the superscript "F" or "J".

- If the precise P value did *not* lie within the P value range provided by the other source, we set the P value equal to the top of that range, thus bringing the FDA and journal P values as close to equality as reasonably possible. For example, in the case of $P_{sponsor}$<.10 but P_{FDA}=.50, we set $P_{sponsor}$=.10. These instances are shown in Appendix A with the superscript "T".

- Additionally we set three pairs of matching P value ranges equal to the values at the top of their respective ranges. These are shown in Appendix Table A with the superscripts "T" and "J" on the FDA side and "T" and "F" on the journal side. (In one of these cases, the FDA review reported a P value of ".00", which we interpreted as P<.01.)

- Nonsignificant P values: The FDA reviews reported nonsignificant results for 56 treatment arms. From these we obtained 45 precise P values. (Precise values were reported for 41 nonsignificant P values, and we calculated 4 others from data provided in the reviews using the method noted above.) The FDA reported on an additional 11 P values as "NS" but no other data from which we could calculate precise P values. One of these we set equal to the precise (nonsignificant) P value obtained from the corresponding journal publication, following the procedure mentioned above. For the remaining 10 instances (18% of the total number of nonsignificant results), there was no corresponding journal publication. To exclude these nonsignificant P values from the analyses would have created a bias in the proportion of studies found nonsignificant, and it would also have biased the estimates of effect size. Instead, we assigned a precise P value for these 10 nonrecorded nonsignificant P values by transforming the above-mentioned 45 precise P values into their standard normal deviates Z, finding their median (.954), and transforming that back into P =.34. (The upper and lower quantile Z values were 1.254 and 0.628, corresponding to P values of .21 and .53, respectively.) Thus we used P=.34, derived from 45 precise nonsignificant P values, as the precise value for the 10 P values reported only as "NS". These instances are shown in Appendix Table A with the superscript "N".

RESULTS

- Studies of escitalopram and some citalopram studies used the Montgomery-Åsberg Depression Rating Scale (MADRS). All other studies employed the Hamilton Depression Rating Scale, usually the 17-item version.

- For antidepressant studies, the FDA defines intent-to-treat (ITT) patients as all randomized patients who return for at least one on-drug post-baseline visit.

- In enumerating the numbers of patients within the studies, we included only those patients whose data we used in the other analyses. Thus we excluded patients randomized to active comparator treatment arms and to doses that were eventually not approved. Therefore the actual numbers of patients participating were greater than we report here.

STUDY OUTCOME AND PUBLICATION STATUS

Below is the 2x2 table corresponding to the risk ratio (RR) presented in the print version of the article:

	FDA-positive	FDA-nonpositive	Total	
Published & consistent with FDA	37	3	40	RR = 11.7 ($CI_{95\%}$ 6.2 - 22.0)
Not published or published in conflict with FDA	1	33	34	$\chi^2_{(1)} = 59.0$
Total	38	36	74	P<.0001

To check our analyses of categorical data for robustness, we excluded questionable studies. FDA-positive studies were approximately 8 times more likely to be published in a way that agreed with the FDA than FDA-negative studies:

	FDA-positive	FDA-negative	Total	
Published & consistent with FDA	37	3	40	RR = 7.8 ($CI_{95\%}$ 4.3 – 14.1)
Not published or published in conflict with FDA	1	21	22	$\chi^2_{(1)} = 46.3$
Total	38	24	62	P<.0001

Additionally, we ignored whether the publications agreed or conflicted with the FDA and simply compared published to unpublished studies. FDA-positive studies were approximately 3 times more likely to be published than FDA-negative studies:

	FDA-positive	FDA-negative	Total	
Published	37	8	45	RR = 2.9 $(CI_{95\%}\ 2.0 - 4.3)$
Not published	1	16	17	$\chi^2_{(1)} = 30.3$
Total	38	24	62	P<.0001

NUMBER OF PATIENT PARTICIPANTS IN STUDIES

Below is the 2x2 table corresponding to the risk ratio (RR) presented in the print version of this article:

	FDA-positive	FDA-nonpositive	Total	
Published & consistent with FDA	7075	197	7272	RR = 27.1 $(CI_{95\%}\ 25.6 - 28.8)$
Not published or published in conflict with FDA	80	5212	5292	$\chi^2_{(1)} = 11{,}461$
Total	7155	5409	12564	P<.0001

LISTING OF DATA USED IN ANALYSES (APPENDIX TABLES A, C)

Appendix Table A lists the raw data used in the analyses presented here. Each FDA-registered study is shown alongside the corresponding stand-alone journal publication (with reference information), unless the study in question was not published. The calculated effect size and standard error values are shown by study and data source in Appendix Table C.

QUALITATIVE DESCRIPTION OF SELECTIVE REPORTING WITHIN TRIALS (APPENDIX TABLE B)

- Of the 11 publications listed in Appendix Table B (also shown with gray shading in Appendix Table A), 7 highlighted results that did not appear in FDA reviews as either primary or secondary endpoints, suggesting that these analyses were conducted *post hoc*. While the FDA reviews and the journal articles agreed as to the primary rating scale, they differed in how the data derived from these scales were analyzed.

- Among the ways in which the methodology of the journal articles differed from that of the FDA, there were differences as to which data were included in and excluded from the analyses. For example, at the patient level, there were deviations from the intent-to-

treat (ITT) principle,[7,8] specifically by invoking an "efficacy subset" of the ITT population meeting additional criteria or by using an observed cases approach, which omits data from patients who drop out due to lack of efficacy or adverse events. At the site level, there were two journal articles that presented positive data from single sites within multicenter studies, whose overall results, according to the FDA, were nonsignificant. (See footnotes to Appendix Table B for details.)

COMPARISONS OF EFFECT SIZE (APPENDIX TABLE D)

The results of the analyses comparing the sets of effect size values are listed in Supplementary Appendix Table D. These are the same as those mentioned in the text of results section of the main article. However, the table additionally includes the result of the ancillary analysis described in the supplementary methods. Like the others, this result was statistically significant (P=0.003).

DISCUSSION

- Regarding the method of handling dropouts, the Committee for Proprietary Medicinal Products (CPMP) of the European Agency for the Evaluation of Medicinal Products (EMEA) states, "There is no universally applicable method of handling missing values, and different approaches may lead to different results. As such it is essential to pre-specify the selected methods in the statistical section of the study protocol."[9]

REFERENCES

1. Food and Drug Administration, Center for Drug Evaluation and Research. Guidance for Industry: Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products; 1998:6, 9. Available at: http://www.fda.gov/cder/guidance/1397fnl.pdf

2. FOIA Update. The Freedom of Information Act 5 USC 552: Electronic Freedom of Information Act Amendments of 1996. Vol XVII. Available at: http://www.usdoj.gov/oip/foia_updates/Vol_XVII_4/page2.htm

3. Temple R. Government viewpoint of clinical trials of cardiovascular drugs. *Med Clin North Am.* Mar 1989;73(2):495-509.

4. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. *Ann Intern Med.* Sep 19 2000;133(6):455-463.

5. Otto MW, Nierenberg AA. Assay sensitivity, failed clinical trials, and the conduct of science. *Psychother Psychosom.* 2002;71(5):241-243.

6. Cohn CK, Robinson DS, Roberts DL, Schwiderski UE, O'Brien K, Ieni JR. Responders to antidepressant drug treatment: a study comparing nefazodone, imipramine, and placebo in patients with major depression. *J Clin Psychiatry.* 1996;57 Suppl 2:15-18.

7. International Conference on Harmonisation (ICH), European Medicines Agency (EMEA). Statistical Principles for Clinical Trials. Topic E9. 1998. Available at: http://www.fda.gov/cder/guidance/iche3.pdf

8. Lachin JM. Statistical considerations in the intent-to-treat principle. *Control Clin Trials.* Jun 2000;21(3):167-189.

9.   Committee for Proprietary Medicinal Products (CPMP). Evaluation of Medicines for Human Use: Points to Consider on Missing Data: European Agency for the Evaluation of Medicinal Products (EMEA); 2001. Available at:
http://www.emea.eu.int/pdfs/human/ewp/177699EN.pdf

Appendix Table A.   Antidepressant study results as analyzed by FDA and as presented in corresponding journal publications.

Drug Name	Number	Study Number	Dose (mg)	N per FDA Drug	N per FDA Pbo	P Value FDA	P Value Journal	First Author	Year	Journal	PMID	N per publication Drug	N per publication Pbo
bupropion (Wellbutrin SR)	1	203	300	113	117	.04	≤.05 (.04)[F]	Reimherr	1998	Clin Ther	9663366	116	117
	2	205	300 / 400	111 / 111	116	.53 / .30	Not published						
	3	212	300	144	145	.16	Not published						
citalopram (Celexa)	4	85A	20-80	78	82	.0344	<.05 (.0344)[F]	Mendels	1999	Depress Anxiety	10207659	89	91
	5	91206	40 / 60	120 / 110	124	.0025 / .0053	.012	Feighner	1999	J Clin Psychiat	10665628	521	129
	6	86141	10-30	97	50	.316	<.05 (.05)[T]	Nyth	1992	Acta Psychiatr Scand	1529737	60	33
	7	89303	40	61	64	.224	<.05 (.05)[T]	Montgomery	1992	Int Clin Psychopharm	1431024	49	50
	8	89306	40	97	88	.964	Not published						
duloxetine (Cymbalta)	9	HMAT-B	40 / 80	86 / 91	89	.022 / .003	.034 / .002	Goldstein	2004	J Clin Psychopharm	15232330	86 / 91	89
	10	HMAY-A	80 / 120	95 / 93	93	.001 / <.001 (.001)[T]	≤.001 (.002)[M] / ≤.001 (.00001)[M]	Detke	2004	Eur Neuropsychopharm	15589385	95 / 93	93
	11	HMBH-A	60	123	122	<.001 (.001)[T,J]	<.001 (.001)[T,F]	Detke	2002	J Clin Psychiat	12000204	121	115
	12	HMBH-B	60	128	139	.047	.024	Detke	2002	J Psych Res	12393307	128	139
	13	HMAQ-A	20-60	56	57	.146	.009	Goldstein	2002	J Clin Psychiat	11926722	56	57
	14	HMAY-B	80 / 120	93 / 103	99	.253 / .054	.045 / .014	Perahia	2006	Eur Psychiat	16697153	93 / 102	99
	15	HMAQ-B	20-60	81	72	.681	Not published						
	16	HMAT-A	40 / 80	90 / 81	89	.222 / .138	Not published						
escitalopram (Lexapro)	17	99001	10	188	189	<.01 (.007)[M]	.002	Wade	2002	Int Clin Psychopharm	11981349	188	189
	18	99003	10-20	155	154	<.01 (.006)[M]	.002	Lepola	2003	Int Clin Psychopharm	12817155	155	154
	19	SCT-MD-01	10 / 20	118 / 123	119	.0007 / <.0001 (.0001)[T]	<.01 (.0007)[F] / <.01 (.005)[M]	Burke	2002	J Clin Psychiat	12000207	118 / 123	119
	20	SCT-MD-02	10-20	124	125	.251	Not published						
fluoxetine (Prozac)	21	19	>40	22	24	.011	.011	Fabre	1985	Curr Ther Res	n/a	22	26
	22	27	>40	181	163	.012	.012	Stark	1985	J Clin Psychiat	3882682	185	169
	23	62-a	20 / 40 / 60	103 / 99 / 107	56	0.43 / 0.50 / 0.50	NS (.70)[M] / NS (.85)[M] / NS (.76(-))[M]	Dunlop	1990	Psychopharm Bull	2236453	103 / 99 / 97	56
	24	62-b	20 / 40 / 60	97 / 97 / 103	48	.007 / .010 / 0.34	≤.01 (.008)[M] / ≤.01 (.01)[M] / NS (.329)[M]	Wernicke	1987	Psychopharm Bull	3496625	97 / 97 / 103	48
	25	25	40-80	18	24	.50(-)	<.10 (.10)[T]	Rickels	1986	Curr Ther Res	n/a	18	24
mirtazapine (Remeron)	26	003-020/3220	5-35	41	39	.004	Not published						
	27	003-002	5-35	44	44	.0008	≤.001 (.0008)[F]	Claghorn	1995	J Affect Disord	7560544	42	40
	28	003-022/3220	10-35	49	50	0.003	≤.05 (.003)[F]	Bremner	1995	J Clin Psychiat	7592505	47	48
	29	003-023/3220	5-35	49	49	.02	.021	Halikas	1995	Human Psychopharm	not indexed	49	49
	30	003-024/3220	5-35	50	48	.010	<.05 (.010)[F]	Smith	1990	Psychopharm Bull	2236455	47	46
	31	85027	20-60	64	61	.19	.015	Khan MC	1995	Human Psychopharm	not indexed	27	27
	32	84023	15-45	45	45	.347	NS (.136)[M]	Vartiainen	1994	Eur Neuropsychopharm	7919944	37	30
	33	003-021/3220	10-35	45	48	.22	Not published						
	34	003-003	10-35	45	45	.49	Not published						
	35	003-008	15 / 30 / 60	30 / 28 / 30	28	.318(-) / .278(-) / .458(-)	Not published						
nefazodone (Serzone)	36	03AOA-003	100-500	44	45	.03	.01<p≤.05 (.03)[F]	Fontaine	1994	J Clin Psychiat	8071277	44	45
	37	03AOA-004B	300-600	78	75	0.02	≤.05 (.02)[F]	Mendels	1995	J Clin Psychiat	7649971	78	75
	38	CN104-005	100-600	86	91	.00 (.01)[T,J]	<.01 (.01)[T,F]	Rickels	1994	Brit J Psychiat	7952987	86	91
	39	CN104-006	100-600	80	78	.35	Not published / <.01 (.006)[R]	Cohn	1996	J Clin Psychiat	8626358	39	42
	40	030A2-007	300	41	47	.60	Not published						
	41	03AOA-004A	300-600	76	77	.66	Not published						
paroxetine (Paxil)	42	02-001	10-50	51	53	.004	<.01 (.004)[F]	Rickels	1989	Acta Psychiatr Scand	2530761	49	53
	43	02-002	10-50	36	34	.015	<.01 (.01)[T]	Claghorn	1989	Int Clin Psychopharm	1431007	32	27
	44	02-004	10-50	34	32	.001	0.001	Kiev	1992	J Clin Psychiat	1531818	34	32
	45	03-001	10-50	39	37	<.01 (.006)[M]	0.002	Feighner	1989	Acta Psychiatr Scand	2530763	39	37
	46	03-004	10-50	37	37	.04	0.033	Shrivastava	1992	J Clin Psychiat	1531825	33	36
	47	03-005	10-50	40	42	.007	0.007	Peselow	1989	Psychopharm Bull	2532373	40	42
	48	03-006	10-50	39	37	.001	0.002	Fabre	1992	J Clin Psychiat	1531823	38	36
	49	03-002	10-50	40	40	.25	<.05 (.05)[T]	Cohn	1990	Psychopharm Bull	2146697	35	36
	50	03-003	10-50	39	42	.98	Not published						
	51	02-003	10-50	33	33	.311	NS (.311)[F]	Smith	1992	J Clin Psychiat	1531822	33	33
	52	01-001	10-50	24	24	.204	Not published						
	53	07	20	13	12	NS (.34)[N]	Not published						
	54	09	20 / 30 / 40	104 / 99 / 100	51	NS (.34)[N] / NS (.34)[N] / NS (.34)[N]	Not published						
	55	UK-06	30	22	23	NS (.34)[N]	Not published						
	56	UK-09	30	20	21	NS (.11)[J]	NS (.11)[M]	Edwards	1993	Human Psychopharm	n/a	21	20
	57	UK-12	30	19	10	NS (.34)[N]	Not published						
paroxetine (Paxil CR)	58	487	12.5-50	103	107	.007	.007	Rapaport	2003	J Clin Psychiat	14628982	104	109
	59	449	20-62.5	108	110	.004	0.0004	Golden	2002	J Clin Psychiat	12143913	212	211
	60	448	20-62.5	94	93	.254							

Drug Name	Number	Study Number	Dose (mg)	N per FDA		P Value		Publication Info				N per publication	
				Drug	Pbo	FDA	Journal	First Author	Year	Journal	PMID	Drug	Pbo
sertraline (Zoloft)	61	104	50-200	142	141	<.01 (.006)^M	≤.001 (.0002)^M	Reihmerr	1990	J Clin Psychiat	2258378	142	141
	62	103	50	90		0.018	≤ .05 (.04)^M	Fabre	1995	Biol Psychiat	8573661	82	76
			100	89	86	**.084**	≤ .05 (.04)^M					74	
			200	82		**.210**	≤.01 (.004)^M					58	
	63	315	50-200	75	73	.46		Not published					
	64	101	50	22		NS (.35)^M		Not published					
			100	19	23	NS (.87-)^M							
			200	17		NS (.21)^M							
			400	12		NS (.64)^M							
	65	310	50	31		NS (.34)^N		Not published					
			100	28	30	NS (.34)^N							
			200	27		NS (.34)^N							
			400	30		NS (.34)^N							
venlafaxine (Effexor)	66	600A-203	75	77		0.004	≤.05 (.004)^F	Rudolph	1998	J Clin Psychiat	9541154	77	92
			150-225	79	92	<.001 (.001)^T	≤.05 (.001)^F					79	
			300-375	75		0.003	≤.05 (.003)^F					75	
	67	600A-206	150-375	46	47	<.01 (.006)^M	<.001 (.00009)^M	Guelfi	1995	J Clin Psychiat	7559370	46	47
	68	600A-301	75-225	64	78	<.001 (.0004)^J	<.05 (.0004)^M	Schweizer	1994	J Clin Psychiat	8071246	64	78
	69	600A-302	75-200	65	75	0.008	≤.05 (.008)^F	Cunningham	1994	J Clin Psychopharm	8195464	65	75
	70	600A-303	75-225	69	79	.493		Not published					
	71	600A-313	75	72	75	**.193**	not reported	Mendels	1993	Psychopharm Bull	8290661	76	78
			200	77		**.142**	≤.05 (.05)					79	
venlafaxine (Effexor XR)	72	208	75-150	85	91	<.001 (.001)^{T,J}	<.001 (.001)^{T,F}	Cunningham	1997	Ann Clin Psychiat	9339881	92	99
	73	209	75-225	91	100	<.001 (.0003)^J	<.001 (.0003)^M	Thase	1997	J Clin Psychiat	9378690	91	100
	74	367	75	82	81	*.37*		Not published					
			150	75		*.14*							

Numbers in second column correspond to those shown in Figure 2a.

Dark blue pattern: unpublished studies. (Colors are same as in Figure 2.)

Medium blue pattern: studies deemed negative or questionable by FDA but presented as positive in journal articles. Details provided in Appendix Table B.

Light blue: studies published in agreement with FDA conclusion.

All P values are two-tailed unless noted otherwise

Parentheses indicate precise P value used in analysis; please see methods.

Bold numbers: statistically nonsignificant results

Italics: "failed" studies (see text)

(-) following p value indicates study drug performed worse than placebo.

PMID = PubMed ID Number

[F] FDA precise P value used as sponsor precise P value

[J] Journal precise P value used as FDA precise P value

[M] Mean used with SD, SE, or CI to calculate precise P value

[R] Responders counts in article used to replicate sponsor's analysis and obtain precise P value

[T] Top of reported P value range used as precise P value

[N] Nonsignificant; precise P value derived as detailed in supplementary methods

Appendix Table B. How apparent primary methods in journal articles differed from protocol-prespecified primary methods per FDA.

Drug name	Number[A]	Study number	Dose group (mg/d)	P value highlighted in journal article	P value per FDA	FDA primary reported in article: Yes	No	Unclear	Word "primary" used in article: Yes	No	Article's apparent primary reported in FDA review: Yes	No	Unclear	Analyzed separate multi-center trials as one	Presented data from only one site from multi-center study	"Efficacy subset" of ITT patients meeting additional criteria	Non-primary method of handling dropouts	Baseline differences not accounted for	Continuous measure rendered binary	Test other than ANOVA or ANCOVA	1-tailed P value highlighted	Highlighted trend level P value as noteworthy	Miscellaneous
citalopram	6	86141	20-80	<.05	.316		X[E]			X		X[Q]				X[X]	X			X[GG]			
citalopram	7	89303	40	<.05	.224		X		X[K]			X[R]					X[Y]						
duloxetine	13	HMAQ-A	40-120	.009	.146		X		X[L]			X[S]					X[BB]						X[KK]
duloxetine	14	HMAY-B	80	.045	.253			X[I]	X[M]		X						X[CC]						
duloxetine	14	HMAY-B	120	.014	.054						X												
fluoxetine	25	25	40-80	<.10	.50(-)[D]	X				X	X[T]						X[DD]			X[HH]	X[JJ]	X	
mirtazapine	31	85027	20-60	.015	.19	X[F]				X	X				X								
nefazodone	39	CN104-006	100-600	<.01	.35	X			X[N]		X					X[W]			X	X		X	
paroxetine	49	03-002	10-50	<.05	.25	X				X	X[P]						X[Z]		X				
paroxetine CR	60	448	20-62.5	.0004	.254	X[G]			X[O]			X		X[V]			X[EE]						X[LL]
sertraline	62	103[B]	100	≤.05	.084			X[J]	X				X[U]				X[AA]						
sertraline	62	103[B]	200	≤.01	.21																		
venlafaxine	71	600A-313	50-75	none[C]	.193	X[H]			X		X						X[FF]			X[II]			
venlafaxine	71	600A-313	150-200	≤.05	.142																		

[A] Number used as column header in this table also used in Figure 2a and Appendix Tables A and C.

[B] Although the two higher dose groups were nonsignificant on the FDA's usual primary outcome (endpoint), the Agency seemed to consider this study weakly supportive of the efficacy of sertraline. "For the HAMD total score and CGI (Clinical Global Impression) items, the LOCF [last observation carried forward] analyses showed only scattered significant comparisons. The OC [observed cases] analyses for the same items were more consistently significant for the higher sertraline doses (ie., not for the 50-mg form). This study did not reveal any evidence for a dose response relationship...Nevertheless, this study provides evidence for the efficacy of sertraline." Elsewhere, in a separate review document by the same FDA statistician, "Study 103 does not provide, in itself, compelling evidence of the efficacy of sertraline..." We therefore classified this study as questionable.

[C] The P value for low-dose group was not reported in the journal article. Lack of significance on this is evident only from noting that there is no footnoted P value for that dose group in Figure 1a. FDA review shows that, in addition to the LOCF analysis being nonsignificant, the observed cases analysis-presented as significant in journal article-was also nonsignificant (P=.381) for this dose group.

[D] (-) sign indicates that study drug performed numerically worse than placebo.

[E] Methods section states, "The intention-to-treat and endpoint [LOCF] analyses, which are not reported here, yielded results similar to those of the efficacy analysis." The result reported first by the FDA, involving the HAMD and the LOCF method of handling dropouts (P=.316), was not reported in the journal article. The review's conclusion noted "the disappointing results from study 86141".

[F] The introduction of the article states, "Hereby presented results are obtained in one of three participating centres; the results of the multicentre study are going to be presented elsewhere." However, we were unable to find any other publications with mirtazapine in the title and with the principal investigator (PI) of this multicenter study as an author. The FDA review does not present results separately by site, only the results of the entire study, ie. with data from all sites combined.

[G] Article does not mention the results of the two studies, 448 and 449, individually. See footnote V for details.

[H] The article's results section, after reporting significant results with observed cases, provides raw scores for LOCF but does not state whether they were significant or nonsignificant.

[I] The journal article noted a lack of significance using ANCOVA and the LOCF method of handling dropouts, but it did not report the P value from that analysis.

[J] Results section states, "In the intention-to-treat analyses (all patients). . . In the sertraline 100 mg/day group, significant improvement was noted in all parameters except the HAMD Total and CGI Severity scores...in the 200 mg/day group, significant improvement was observed in all but the HAMD Total and CGI Improvement scores." No P values were provided, and the fact that the HAMD Total was the primary efficacy measure was not reported.

[K] Results section states, "The primary efficacy analysis was carried out on a last observation carried forward basis." Other details were unclear. Please see related footnotes for this study.

L The word "primary" was used in the article, but only in reference to the rating scale, not to the statistical analytic technique.

M Methods section of article states, "The a priori primary treatment comparison was the contrast between duloxetine 120 mg/day and placebo on change from baseline at week 8 (visit 8) on the HAMD17 total score...Longitudinal efficacy outcomes were primarily analyzed using a likelihood-based mixed-effects repeated measures approach (MMRM)." However, see footnote BB.

N Methods section states, "Primary efficacy outcome measures specified in the study protocol included the HAMD-17 and the Clinical Global Impressions-Improvement (CGI-I) scale ratings." It did not state how the protocol said that the data obtained from these scales should be analyzed. Results section states, "The primary efficacy measure *used in these analyses* is the number of depressed patients classified as treatment responders." [italics added for emphasis] Thus it was unclear from article's wording alone whether the analyses reported were the same as or different from those prespecified in protocol. (FDA reported nonsignificant result according to primary method, which did not involve treatment responders.)

O "Primary" used in reference to rating scale but not in reference to details of statistical analysis.

P FDA review shows "all efficacy patients" analysis using LOCF technique with P=.04, consistent with the P<.05 result reported in the journal article. However, this gave different results from the P=.25 obtained using the primary intention-to-treat (ITT) analysis in the FDA review (see footnote Z).

Q "Efficacy analysis" result reported on in the journal article could not be found in the FDA review. It was unclear from the article how dropouts were handled in this analysis. It appears that LOCF approach was not used (footnote E), suggesting that an observed cases (OC) approach was used. However, the results highlighted in the journal article do not agree with the OC analysis conducted by FDA. The results section of the journal article states, "The mean total scores decreased at week 2 and onwards in both the citalopram and placebo groups (P<.05)." The FDA's OC analysis found that the difference between citalopram and placebo was nonsignificant at weeks 2 and 4 (P=.374 and .709, respectively) and became significant only at week 6.

R The FDA did not conduct any analyses using the "efficacy subset" definition put forth in journal article. (See also footnote Y.)

S The result obtained using mixed model repeated measures (MMRM) analysis was not reported in the FDA's review of this study, which only reported the result using the LOCF analysis. (See also footnote BB.)

T The FDA review states, "The results for the weekly [OC] analysis were inconsistent. For some variables, fluoxetine was significantly better than placebo for some variables [sic] toward the latter weeks while placebo was occasionally nonsignificantly better than at earlier weeks." The FDA review did not report the P value from this analysis, which was the one highlighted in the journal article. (See footnote DD.)

U The "evaluable patients" analysis reported on in journal article (see footnote AA) was not reported on in FDA review.

V Two similarly designed 20-site studies in the US and Canada, were pooled into a single analysis in the journal article, with an overall positive result. As shown in Appendix Table A, the FDA

review indicates that one of these two studies was negative (P=.254). See also footnote LL.

[W] This study involved two centers, both in Houston, Texas. Center 1 did not show a significant difference for nefazodone (P=.97) or for imipramine (P=.53) vs. placebo. Center 2 showed a trend (P=.09) when analyzed according to protocol (ANOVA on change-from-baseline scores). The FDA reviewer argued that this study could be considered positive by ignoring Center 1 (as a failed sub-study) and by considering that the secondary outcomes provided "strong supportive evidence in favor of nefazodone with both the LOCF and OC results in agreement." The Division Director's memo stated agreement with this. The PI of Center 2 authored the paper shown in Appendix A. Its highlighted analysis was based on percentage of responders, different from the prespecified method noted above, and gave a significant result. The other center was not mentioned.

[X] The FDA's LOCF analysis included 147 patients, while the article's "efficacy evaluation group" included 133.

[Y] Post-baseline depression ratings were conducted at Week 1 and Week 3 according to both FDA review and journal article. FDA review reports definition of ITT was those with any postbaseline data, ie. those returning at Week 1. Journal defined "evaluable patients" as those with data recorded after at least 2 weeks of data. But since the next visit was at Week 3, this analysis required patients to have 3 weeks of data. Thus, patients dropping out between the Week 1 and Week 3 visits were excluded in the journal article but included in the FDA analysis.

[Z] According to the FDA review, the ITT Ns were 40 and 40 for paroxetine and placebo groups, respectively, vs. 35 and 36 for the "all efficacy" analysis. The FDA review also states, "The discrepancy between the ITT and the All Efficacy data derives from the 5 paroxetine and 4 placebo patients excluded."

[AA] Article states, "Evaluable patients were classified as those who took study medication on or after the 11th day of the double-blind phase, who had efficacy assessments on or after study day 11, and who were not major protocol violators." Analyses based on "evaluable patients" are shown in Figure 1 and Table 1, while those based on ITT patients appear later in Table 5.

[BB] The FDA statistical review of duloxetine states that sponsor proposed MMRM as the primary analytic method in the protocol it submitted. However, the review also states, "Based on the letter issued by the Agency to the sponsor (dated January 11, 2002), this reviewer considered the LOCF analysis as the primary statistical analysis to evaluate the efficacy of duloxetine." (We did not find this letter among the FDA review documents.)

[CC] Same as for BB above.

[DD] An observed cases (completers analysis) was apparently used. Though not explicit in the text, Table 1 shows that the highlighted P value is at Week 5, where the Ns show significant attrition compared to the Ns at earlier weeks. Also, the Discussion section states, "Because of the slow onset of action, endpoint [LOCF] analyses were considered not appropriate for this data set."

[EE] Result highlighted in Figure 1 was obtained with random-effects mixed model (REMM) analysis.

[FF] Efficacy findings presented first obtained using observed cases (ie. completers only) analysis. Mean scores obtained using the LOCF method were reported but without associated P values or mention of their lack of significance. However, the LOCF scores were used in a "trend analysis" reported (as significant) subsequently in article.

GG t-test was noted as the analytic method in the FDA review, so we assume that this was according to protocol. This is typically not the case, because it cannot account for possible effects of site or of treatment-by-site interaction.

HH The article presents results of a nonparametric two-sample Wilcoxon rank sum test. The FDA review was not clear on what test was used to analyze the data, but because it presented P values alongside mean change scores, it appears that the FDA used a parametric test (e.g. ANOVA, ANCOVA, or t-test).

II Though this result did not meet our criteria for apparent primary result, a positive dose-response effect was presented in the abstract, in the text of results section, and in Table 2. This finding, obtained using the LOCF method of handling dropouts, was contradicted by the relative performance of the doses shown in Figure 1, which showed results obtained using the observed cases approach (see footnote FF). This figure showed that the 50-75mg dose was outperformed on both the HAMD and MADRS scales by the 25-mg dose (a lower dose not FDA-approved as effective).

JJ According to the journal article, the highlighted result of P<.10 one-tailed (equivalent to two-tailed P<.20) "should be interpreted in the light of the small sample size available." For comparison, the usual criterion for statistical significance is two-tailed P<.05 (equivalent to one-tailed P<.025).

KK The Ns shown in the journal article were larger than the Ns shown in the FDA review and in a report of the same study on the sponsor's website, www.lillytrials.com. The reason for this discrepancy is unclear.

LL The FDA analysis excluded Center 2/4 which had a remarkably high (compared to the other 19 sites) drug-placebo difference. The FDA reviewer stated that this raised the question of unblinding at that site. An audit by the sponsor found that the site had provided patients with a copy of the scale during the rating, contrary to instructions given at the investigator meeting. Including this site, P=.021; excluding this outlier site, P=.254. In the analysis presented in the journal article, data from this site were included.

Appendix Table C. Effect size and standard error for individual FDA-registered studies according to the FDA and to journal publications.

Drug_name	Number	Study Number	g_FDA	SE_FDA	g_journal	SE_journal
bupropion SR	1	203	0.2715427	0.1325076	0.269742	0.1316252
bupropion SR	2	205	0.1103332	0.1146443	.	.
bupropion SR	3	212	0.1652994	0.1178499	.	.
citalopram	4	85A	0.3358892	0.159288	0.3164596	0.1500209
citalopram	5	91206	0.3790019	0.1123227	0.2474634	0.0985828
citalopram	6	86141	0.1742657	0.1743961	0.4269396	0.2190233
citalopram	7	89303	0.2173459	0.1794728	0.395872	0.2030172
citalopram	8	89306	0.0066264	0.1472176	.	.
duloxetine	9	HMAT-B	0.3977942	0.131093	0.3942015	0.1310724
duloxetine	10	HMAY-A	0.4870858	0.1284407	0.5574585	0.1289525
duloxetine	11	HMBH-A	0.4243098	0.1292178	0.4326138	0.1317574
duloxetine	12	HMBH-B	0.2437737	0.1229588	0.2773137	0.1230928
duloxetine	13	HMAQ-A	0.2736024	0.1890455	0.4968784	0.191084
duloxetine	14	HMAY-B	0.2195651	0.1236338	0.3199672	0.124115
duloxetine	15	HMAQ-B	0.0663832	0.1620159	.	.
duloxetine	16	HMAT-A	0.204499	0.1310151	.	.
escitalopram	17	99001	0.2787711	0.1035074	0.3199229	0.1036659
escitalopram	18	99003	0.3140545	0.1144804	0.3537735	0.1146689
escitalopram	19	SCT-MD-01	0.4762382	0.1134438	0.4034205	0.1130488
escitalopram	20	SCT-MD-02	0.1453935	0.1269146	.	.
fluoxetine	21	19	0.7702243	0.3063701	0.7550544	0.3001881
fluoxetine	22	27	0.2721216	0.1084808	0.268133	0.1068863
fluoxetine	23	62 (moderate)	0.3655864	0.1561895	0.3642578	0.156185
fluoxetine	24	62 (mild)	0.1180624	0.1453017	0.0151255	0.1456089
fluoxetine	25	25	-0.2082325	0.3126727	0.5151265	0.3170791
mirtazapine	26	003-020/3220	0.6570727	0.22978	.	.
mirtazapine	27	003-002	0.7346137	0.2204361	0.762831	0.2290119
mirtazapine	28	003-022/3220	0.6072609	0.2056916	0.6203963	0.2101886
mirtazapine	29	003-023/3220	0.4742241	0.2049088	0.470375	0.2048626
mirtazapine	30	003-024/3220	0.5268903	0.2056192	0.5411536	0.2112458
mirtazapine	31	85027	0.2343729	0.1795598	0.6748119	0.280094
mirtazapine	32	84023	0.1976222	0.2113441	0.3665797	0.2477782
mirtazapine	33	003-021/3220	0.254148	0.2083517	.	.
mirtazapine	34	003-003	0.1448935	0.2111012	.	.
mirtazapine	35	003-008	-0.2747475	0.2177361	.	.
nefazodone	36	03AOA-003	0.4637128	0.2149076	0.4637128	0.2149076
nefazodone	37	03AOA-004B	0.3783634	0.1631805	0.3783634	0.1631805
nefazodone	38	CN104-005	0.3899649	0.1518269	0.3899649	0.1518269
nefazodone	39	CN104-006	0.1484453	0.159346	0.6220149	0.2278142
nefazodone	40	030A2-007	0.1114961	0.2138671	.	.
nefazodone	41	03AOA-004A	0.0709174	0.1617454	.	.
paroxetine	42	02-001	0.5734214	0.2002189	0.5795371	0.2023744
paroxetine	43	02-002	0.5902355	0.2444405	0.6871751	0.2691268
paroxetine	44	02-004	0.8395159	0.2572312	0.8395159	0.2572312
paroxetine	45	03-001	0.6427158	0.2354977	0.7278378	0.2371654
paroxetine	46	03-004	0.4812097	0.2359282	0.5187855	0.2451312
paroxetine	47	03-005	0.6058415	0.2260609	0.6058415	0.2260609
paroxetine	48	03-006	0.7784628	0.2382492	0.7381631	0.2405774
paroxetine	49	03-002	0.2566561	0.224549	0.4683938	0.2407052
paroxetine	50	03-003	0.0055391	0.2223752	.	.
paroxetine	51	02-003	0.2484467	0.2471605	0.2484467	0.2471605
paroxetine	52	01-001	0.3658727	0.2911844	.	.

Drug_name	Number	Study Number	g_FDA	SE_FDA	g_journal	SE_journal
paroxetine	53	07	0.3772091	0.4041653	.	.
paroxetine	54	09	0.163615	0.15148	.	.
paroxetine	55	UK-06	0.2826948	0.29977	.	.
paroxetine	56	UK-09	0.501093	0.3175503	0.501093	0.3175503
paroxetine	57	UK-12	0.368845	0.3938921	.	.
paroxetine CR	58	487	0.3746463	0.1392549	0.3719997	0.1382665
paroxetine CR	59 & 60	448 & 449	0.286	0.1127551	0.3464183	0.0979737
sertraline	61	104	0.3283027	0.1196923	0.4468546	0.1203734
sertraline	62	103	0.2720957	0.1247663	0.3842469	0.1344884
sertraline	63	315	0.1211724	0.1645668	.	.
sertraline	64	101	0.19708	0.2407849	.	.
sertraline	65	310	0.2473338	0.2053447	.	.
venlafaxine	66	600A-203	0.4756994	0.1247038	0.4756994	0.1247038
venlafaxine	67	600A-206	0.5787581	0.2117928	0.8431283	0.2166139
venlafaxine	68	600A-301	0.6085268	0.1725341	0.6085268	0.1725341
venlafaxine	69	600A-302	0.4535774	0.1716489	0.4535774	0.1716489
venlafaxine	70	600A-303	0.1126633	0.1649074	.	.
venlafaxine	71	600A-313	0.2266648	0.1419875	0.3137843	0.1606124
venlafaxine XR	72	208	0.5027412	0.1532319	0.482147	0.1469203
venlafaxine XR	73	209	0.5315622	0.1474331	0.5315622	0.1474331
venlafaxine XR	74	367	0.1871638	0.1370741	.	.

Number column: Studies are listed in same order as in Appendix Table A.

Period indicates missing data due to study not being published as stand-alone article.

g: Hedges' g

SE: standard error

pub: published version

Appendix Table D. Comparisons of Effect Size

Contrast	Level of Contrast	Data Source FDA Unpublished	FDA Published	Journal articles	Results of Contrast Z-statistic	P Value
Within published studies: Journal version vs. FDA version	Studies		N=50 $\bar{g}$=.40 s=.20	N=50 $\bar{g}$=.47 s=.17	2.99 [d]	0.003
	Drugs		N=12 $\bar{g}$=.37 s=.10	N=12 $\bar{g}$=.42 s=.11	2.51 [d]	0.012
Within FDA: published vs. unpublished studies [a]	Studies	N=23 $\bar{g}$=.18 s=.17	N=51 $\bar{g}$=.40 s=.20		4.63 [e]	<0.0001
	Drugs	N=10 $\bar{g}$=.14 s=.06	N=10 $\bar{g}$=.39 s=.10		2.80 [d]	0.005
		N=12 $\bar{g}$=.17 s=.08	N=12 $\bar{g}$=.37 s=.10		2.95 [c,d]	0.003
All FDA (pub + unpub) results vs. journal results [b]	Studies	N=74 $\bar{g}$=.33 s=.22		N=50 $\bar{g}$=.47 s=.17	3.83 [e]	0.0001
	Drugs	N=12 $\bar{g}$=.31 s=.08		N=12 $\bar{g}$=.42 s=.11	3.06 [d]	0.002

[a] compare with Figure 3, Panel A

[b] compare with Figure 3, Panel B

[c] ancillary analysis with imputation -- please see supplementary methods

[d] signed-rank test

[e] rank-sum test

N = number of studies or drugs analyzed

$\bar{g}$ = Hedges's g (effect size), unweighted mean

s = standard deviation around mean of $\bar{g}$