Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.

ABSTRACT

BACKGROUND
Evidence-based medicine is valuable to the extent that the evidence base is complete and unbiased. Selective publication of clinical trials — and the outcomes within those trials — can lead to unrealistic estimates of drug effectiveness and alter the apparent risk–benefit ratio.

METHODS
We obtained reviews from the Food and Drug Administration (FDA) for studies of 12 antidepressant agents involving 12,564 patients. We conducted a systematic literature search to identify matching publications. For trials that were reported in the literature, we compared the published outcomes with the FDA outcomes. We also compared the effect size derived from the published reports with the effect size derived from the entire FDA data set.

RESULTS
Among 74 FDA-registered studies, 31%, accounting for 3449 study participants, were not published. Whether and how the studies were published were associated with the study outcome. A total of 37 studies viewed by the FDA as having positive results were published; 1 study viewed as positive was not published. Studies viewed by the FDA as having negative or questionable results were, with 3 exceptions, either not published (22 studies) or published in a way that, in our opinion, conveyed a positive outcome (11 studies). According to the published literature, it appeared that 94% of the trials conducted were positive. By contrast, the FDA analysis showed that 51% were positive. Separate meta-analyses of the FDA and journal data sets showed that the increase in effect size ranged from 11 to 69% for individual drugs and was 32% overall.

CONCLUSIONS
We cannot determine whether the bias observed resulted from a failure to submit manuscripts on the part of authors and sponsors, from decisions by journal editors and reviewers not to publish, or both. Selective reporting of clinical trial results may have adverse consequences for researchers, study participants, health care professionals, and patients.
MEDICAL DECISIONS ARE BASED ON AN understanding of publicly reported clinical trials. If the evidence base is biased, then decisions based on this evidence may not be the optimal decisions. For example, selective publication of clinical trials, and the outcomes within those trials, can lead to unrealistically estimates of drug effectiveness and alter the apparent risk–benefit ratio.

Attempts to study selective publication are complicated by the unavailability of data from unpublished trials. Researchers have found evidence for selective publication by comparing the results of published trials with information from surveys of authors, registries, institutional review boards, and funding agencies, and even with published methods. Numerous tests are available to detect selective-reporting bias, but none are known to be capable of detecting or ruling out bias reliably.

In the United States, the Food and Drug Administration (FDA) operates a registry and a results database. Drug companies must register with the FDA all trials they intend to use in support of an application for marketing approval or a change in labeling. The FDA uses this information to create a table of all studies. The study protocols in the database must prospectively identify the exact methods that will be used to collect and analyze data. Afterward, in their marketing application, sponsors must report the results obtained using the prespecified methods. These submissions include raw data, which FDA statisticians use in corroborative analyses. This system prevents selective post hoc reporting of favorable trial results and outcomes within those trials.

How accurately does the published literature convey data on drug efficacy to the medical community? To address this question, we compared drug efficacy inferred from the published literature with drug efficacy according to FDA reviews.

METHODS

DATA FROM FDA REVIEWS

We identified the phase 2 and 3 clinical-trial programs for 12 antidepressant agents approved by the FDA between 1987 and 2004 (median, August 1996), involving 12,564 adult patients. For the eight older antidepressants, we obtained hard copies of statistical and medical reviews from colleagues who had procured them through the Freedom of Information Act. Reviews for the four newer antidepressants were available on the FDA Web site. This study was approved by the Research and Development Committee of the Portland Veterans Affairs Medical Center; because of its nature, informed consent from individual patients was not required.

From the FDA reviews of submitted clinical trials, we extracted efficacy data on all randomized, double-blind, placebo-controlled studies of drugs for the short-term treatment of depression. We included data pertaining only to dosages later approved as safe and effective; data pertaining to unapproved dosages were excluded.

We extracted the FDA’s regulatory decisions — that is, whether, for purposes of approval, the studies were judged to be positive or negative with respect to the prespecified primary outcomes (or primary end points). We classified as questionable those studies that the FDA judged to be neither positive nor clearly negative — that is, studies that did not have significant findings on the primary outcome but did have significant findings on several secondary outcomes. Failed studies were also classified as questionable (for more information, see the Methods section of the Supplementary Appendix, available with the full text of this article at www.nejm.org). For fixed-dose studies (studies in which patients are randomly assigned to receive one of two or more dose levels or placebo) with a mix of significant and nonsignificant results for different doses, we used the FDA’s stated overall decisions on the studies. We used double data extraction and entry, as detailed in the Methods section of the Supplementary Appendix.

DATA FROM JOURNAL ARTICLES

Our literature-search strategy consisted of the following steps: a search of articles in PubMed, a search of references listed in review articles, and a search of the Cochrane Central Register of Controlled Trials; contact by telephone or e-mail with the drug sponsor’s medical-information department; and finally, contact by means of a certified letter sent to the sponsor’s medical-information department, including a deadline for responding in writing to our query about whether the study results had been published. If these steps failed to reveal any publications, we concluded that the study results had not been published.
We identified the best match between the FDA-reviewed clinical trials and journal articles on the basis of the following information: drug name, dose groups, sample size, active comparator (if used), duration, and name of principal investigator. We sought published reports on individual studies; articles covering multiple studies were excluded. When the results of a trial were reported in two or more primary publications, we selected the first publication.

Few journal articles used the term "primary efficacy outcome" or a reasonable equivalent. Therefore, we identified the apparent primary efficacy outcome, or the result highlighted most prominently, as the drug–placebo comparison reported first in the text of the results section or in the table or figure first cited in the text. As with the FDA reviews, we used double data extraction and entry (see the Methods section of the Supplementary Appendix for details).

STATISTICAL ANALYSIS

We categorized the trials on the basis of the FDA regulatory decision, whether the trial results were published, and whether the apparent primary outcomes agreed or conflicted with the FDA decision. We calculated risk ratios with exact 95% confidence intervals and Pearson’s chi-square analysis, using Stata software, version 9. We used a similar approach to examine the numbers of patients within the studies. Sample sizes were compared between published and unpublished studies with the use of the Wilcoxon rank-sum test.

For our major outcome indicator, we calculated the effect size for each trial using Hedges’s g — that is, the difference between two means divided by their pooled standard deviation. However, because means and standard deviations (or standard errors) were inconsistently reported in both the FDA reviews and the journal articles, we used the algebraically equivalent computational equation:

\[ g = \frac{t \times \sqrt{\frac{1}{n_{\text{drug}}} + \frac{1}{n_{\text{placebo}}}}} \]

We calculated the t statistic using the precise P value and the combined sample size as arguments in Microsoft Excel’s TINV (inverse T) function, multiplying t by −1 when the study drug was inferior to the placebo. Hedges’s correction for small sample size was applied to all g values.

Precise P values were not always available for the above calculation. Rather, P values were often indicated as being below or above a certain threshold — for example, P<0.05 or “not significant” (i.e., P>0.05). In these cases, we followed the procedure described in the Supplementary Appendix.

For each fixed-dose (multiple-dose) study, we computed a single study-level effect size weighted by the degrees of freedom for each dose group. On the basis of the study-level effect-size values for both fixed-dose and flexible-dose studies, we calculated weighted mean effect-size values for each drug and for all drugs combined, using a random-effects model with the method of Der-Simonian and Laird in Stata.

Within the published studies, we compared the effect-size values derived from the journal articles with the corresponding effect-size values derived from the FDA reviews. Next, within the FDA data set, we compared the effect-size values for the published studies with the effect-size values for the unpublished studies. Finally, we compared the journal-based effect-size values with those derived from the entire FDA data set – that is, both published and unpublished studies.

We made these comparisons at the level of studies and again at the level of the 12 drugs. Because the data were not normally distributed, we used the nonparametric rank-sum test for unpaired data and the signed-rank test for paired data. In these analyses, all the effect-size values were given equal weight.

RESULTS

STUDY OUTCOME AND PUBLICATION STATUS

Of the 74 FDA-registered studies in the analysis we could not find evidence of publication for 23 (31%) (Table 1). The difference between the sample sizes for the published studies (median, 153 patients) and the unpublished studies (median, 146 patients) was neither large nor significant (5% difference between medians; P=0.29 by the rank-sum test).

The data in Table 1 are displayed in terms of the study outcome in Figure 1A. The questions of whether the studies were published and, if so, how the results were reported were strongly related to their overall outcomes. The FDA deemed 38 of the 74 studies (51%) positive, and all but 1 of the 38 were published. The remaining 36 studies (49%) were deemed to be either negative (24 studies) or questionable (12). Of these 36 studies, 3 were pub-
lished as not positive, whereas the remaining 33 either were not published (22 studies) or were published, in our opinion, as positive (11) and therefore conflicted with the FDA’s conclusion. Overall, the studies that the FDA judged as positive were approximately 12 times as likely to be published in a way that agreed with the FDA analysis as were studies with nonpositive results according to the FDA (risk ratio, 11.7; 95% confidence interval [CI], 6.2 to 22.0; P<0.001). This association of publication status with study outcome remained significant when we excluded questionable studies and when we examined publication status without regard to whether the published conclusions and the FDA conclusions were in agreement (for details, see the Supplementary Appendix).

Overall, 48 of the 51 published studies were reported to have positive results (94%; binomial 95% CI, 84 to 99). According to the FDA, 38 of the 74 registered studies had positive results (51%; 95% CI, 39 to 63). There was no overlap between these two sets of confidence intervals.

These data are broken down by drug and study number in Figure 2A. For each of the 12 drugs, the results of at least one study either were unpublished or were reported in the literature as positive despite a conflicting judgment by the FDA.

**Number of Study Participants**

As shown in Table 1, a total of 12,564 patients participated in these trials. The data from 3449 patients (27%) were not published. Data from an additional 1843 patients (15%) were reported in journal articles in which the highlighted finding conflicted with the FDA-defined primary outcome. Thus, the percentages for the patients closely mirrored those for the studies (Table 1).

Whether a patient’s data were reported in a way that was in concert with the FDA review was associated with the study outcome (Fig. 1B) (risk ratio, 27.1), which was consistent with the above-reported finding with the studies. Figure 2B shows these same data according to the drug being evaluated.

**Qualitative Description of Selective Reporting within Trials**

The methods reported in 11 journal articles appear to depart from the prespecified methods reflected in the FDA reviews (Table B of the Supplementary Appendix). Although for each of these studies the finding with respect to the protocol-specified primary outcome was nonsignificant, each publication highlighted a positive result as if it were the primary outcome. The nonsignificant results for the prespecified primary outcomes were either subordinated to nonprimary positive results (in two reports) or omitted (in nine). (Study-level methodologic differences are detailed in the footnotes to Table B of the Supplementary Appendix.)

**Effect Size**

The effect-size values derived from the journal reports were often greater than those derived from the FDA reviews. The difference between these two sets of values was significant whether the studies (P=0.003) or the drugs (P=0.012) were used as the units of analysis (see Table D in the Supplementary Appendix).

The effect sizes of the published and unpublished studies reviewed by the FDA are compared in Figure 3A. The overall mean weighted effect-size value was 0.37 (95% CI, 0.33 to 0.41) for published studies and 0.15 (95% CI, 0.08 to 0.22) for unpublished studies. The difference was significant whether the studies (P<0.001) or the drugs (P=0.005) were used as the units of analysis (Table D in the Supplementary Appendix).

The mean effect-size values for all FDA studies, both published and unpublished, are compared with those for all published studies, as shown in Figure 3B. Again, the differences were significant whether the studies (P<0.001) or the drugs (P=0.002) were used as units of analysis (Table D in the Supplementary Appendix).

For each of the 12 drugs, the effect size derived from the journal articles exceeded the effect size derived from the FDA reviews (sign test, P<0.001) (Fig. 3B). The magnitude of the increases in effect size between the FDA reviews and the published reports ranged from 11 to 69%, with a median

<table>
<thead>
<tr>
<th>Publication Status</th>
<th>No. of Studies (%)</th>
<th>No. of Patients in Studies (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published results agree with FDA decision</td>
<td>40 (54)</td>
<td>7,272 (58)</td>
</tr>
<tr>
<td>Published results conflict with FDA decision (published as positive)</td>
<td>11 (15)</td>
<td>1,843 (15)</td>
</tr>
<tr>
<td>Results not published</td>
<td>23 (31)</td>
<td>3,449 (27)</td>
</tr>
<tr>
<td>Total</td>
<td>74 (100)</td>
<td>12,564 (100)</td>
</tr>
</tbody>
</table>
increase of 32%. A 32% increase was also observed in the weighted mean effect size for all drugs combined, from 0.31 (95% CI, 0.27 to 0.35) to 0.41 (95% CI, 0.36 to 0.45).

A list of the study-level effect-size values used in the above analyses — derived from both the FDA reviews and the published reports — is provided in Table C of the Supplementary Appendix. These effect-size values are based on P values and sample sizes shown in Table A of the Supplementary Appendix, which also lists reference information for the publications consulted.

**Discussion**

We found a bias toward the publication of positive results. Not only were positive results more likely to be published, but studies that were not positive, in our opinion, were often published in a way that conveyed a positive outcome. We analyzed these data in terms of the proportion of positive studies and in terms of the effect size associated with drug treatment. Using both approaches, we found that the efficacy of this drug class is less than would be gleaned from an examination of the published literature alone. According to the published literature, the results of nearly all of the trials of antidepressants were positive. In contrast, FDA analysis of the trial data showed that roughly half of the trials had positive results. The statistical significance of a study’s results was strongly associated with whether and how they were reported, and the association was independent of sample size. The study outcome also affected the chances that the data from a participant would be published. As a result of selec-
tive reporting, the published literature conveyed an effect size nearly one third larger than the effect size derived from the FDA data. Previous studies have examined the risk–benefit ratio for drugs after combining data from regulatory authorities with data published in journals. We built on this approach by comparing study-level data from the FDA with matched data from journal articles. This comparative approach allowed us to quantify the effect of selective publication on apparent drug efficacy. Our findings have several limitations: they are
restricted to antidepressants, to industry-sponsored trials registered with the FDA, and to issues of efficacy (as opposed to “real-world” effectiveness). This study did not account for other factors that may distort the apparent risk–benefit ratio, such as selective publication of safety issues, as has been reported with rofecoxib (Vioxx, Merck) and with the use of selective serotonin-reuptake inhibitors for depression in children. Because we excluded articles covering multiple studies, we probably counted some studies as unpublished that were technically published. The practice of bundling negative and positive studies in a single article has been found to be associated with duplicate or multiple publication, which may also influence the apparent risk–benefit ratio.

There can be many reasons why the results of a study are not published, and we do not know the reasons for nonpublication. Thus, we cannot determine whether the bias resulted from a failure to submit manuscripts on the part of authors and sponsors, decisions by journal editors and reviewers not to publish submitted manuscripts, or both.

We wish to clarify that nonsignificance in a single trial does not necessarily indicate lack of efficacy. Each drug, when subjected to meta-analysis, was shown to be superior to placebo. On the other hand, the true magnitude of each drug’s superiority to placebo was less than a diligent literature review would indicate.

We do not mean to imply that the primary methods agreed on between sponsors and the FDA are necessarily preferable to alternative methods. Nevertheless, when multiple analyses are conducted, the principle of prespecification controls the rate of false positive findings (type I error), and it prevents HARKing, or hypothesizing after the results are known.

It might be argued that some trials did not merit publication because of methodologic flaws, including problems beyond the control of the investigator. However, since the protocols were written according to international guidelines for efficacy studies and were carried out by companies with ample financial and human resources, to be fair to the people who put themselves at risk to participate, a cogent public reason should be given for failure to publish.

Selective reporting deprives researchers of the accurate data they need to estimate effect size realistically. Inflated effect sizes lead to underestimates of the sample size required to achieve statistical significance. Underpowered studies — and selectively reported studies in general — waste resources and the contributions of investigators and study participants, and they hinder the advancement of medical knowledge. By altering the apparent risk–benefit ratio of drugs, selective publication can lead doctors to make inappropriate prescribing decisions that may not be in the best interest of their patients and, thus, the public health.

Dr. Turner reports having served as a medical reviewer for the Food and Drug Administration. No other potential conflict of interest relevant to this article was reported.

We thank Emily Kizer, Marcus Griffith, and Tammy Lewis for clerical assistance; David Wilson, Alex Sutton, Ohidul Siddiqui, and Benjamin Chan for statistical consultation; Linda Ganzini, Thomas B. Barrett, and Daniel Hilfet-Hilliker for their comments on an earlier version of this manuscript; Arifula Khan, Kelly Schwartz, and David Antonuccio for providing access to FDA reviews; Thomas B. Barrett, Norwan Moaleji and Samantha Ruiny for double data extraction and entry; and Andrew Hamilton for literature database searches.
Selective Publication of Antidepressant Trials


Copyright © 2008 Massachusetts Medical Society.