Notice of Correction, 5/3/11: Figures 1, 2, and 3 displayed incorrect legends. In Figures 1, 2A, and 3, the legend should have been "△ Trial with mean baseline HDRS score < 20" and "■ Trial with mean baseline HDRS score ≥ 20." In Figure 2B, the legend should have been "■ Trial with mean baseline HDRS score ≥ 20."

Exploratory Analyses of Efficacy Data From Major Depressive Disorder Trials Submitted to the US Food and Drug Administration in Support of New Drug Applications

Ni A. Khin, MD; Yeh-Fong Chen, PhD; Yang Yang, PhD; Peiling Yang, PhD; and Thomas P. Laughren, MD

Objective: There has been concern about a high rate of placebo response and a substantial failure rate in recent clinical trials in major depressive disorder (MDD). This report explores differences in efficacy data from placebo-controlled MDD trials submitted in support of new drug applications (NDAs) over a 25-year period.

Method: We compiled efficacy data from 81 randomized, double-blind clinical trials, with 21,611 evaluable patients, that were submitted to the US Food and Drug Administration as part of NDAs for an antidepressant claim between 1983 and 2008. Trial data were limited to completed, randomized, multicenter, double-blind, placebo-controlled clinical trials in adult patients diagnosed with MDD according to *DSM-III* or *DSM-IV* criteria. The database was further limited to patients who were involved in clinical trials for drugs widely viewed as effective antidepressants and for doses of these drugs also viewed as effective doses. Trials were rated as successful if they showed statistical superiority vs placebo for the investigational drug on change in Hamilton Depression Rating Scale (HDRS) score (last-observation-carried-forward data). (Trials with multiple investigational drug groups were successful if there was superiority in at least 1 drug group after adjustment for multiplicity.) In particular, we explored differences in effect size and success rate of these trials, based on when the studies were conducted, geographic location of the study sites (US vs non-US), trial duration, dosing regimen, study size, and baseline disease characteristics.

Results: Eighty-one percent of MDD patients were enrolled in US sites. Although the observed placebo and drug responses at non-US sites tended to be larger than at US sites, the treatment effect (drug-placebo difference) was similar (mean change from baseline of about −2.5 units in HDRS total score) in US and non-US trials. In both US and non-US trials, the placebo response showed a modest increase over the observation period (1983–2008). Treatment effect clearly diminished over this same period, at a similar rate for both US and non-US trials despite a marked increase in the sample size of the trials. Our analysis showed that 53% of all MDD trials in the last 25 years were successful. US trials had a higher success rate than non-US trials (58% vs 33%). Before 1995, the overall success rate was 55%, compared to 50% for trials in 1995 or later, and, in general, 6-week trials had a higher success rate than 8-week trials (55% vs 42%). It should be noted that the earlier trials were mostly 6 weeks, and the 6-week trials had higher mean baseline HDRS scores than the 8-week trials. Study size did not seem to influence trial success rates. Mean baseline HDRS total scores declined over the 25-year observation period for patients in both US and non-US trials, as did treatment effect in these trials, again, regardless of region. Fixed-dose trials had a numerically slightly greater success rate than flexible-dose trials (57% vs 51%), although on average treatment effect was numerically larger in the flexible-dose trials than in fixed-dose trials (mean of −2.9 vs −2.0 on HDRS units).

Conclusions: Treatment effect has declined over time in MDD trials, and there has been a high failure rate for these trials during the entire period, but the reasons for these findings remain elusive. Baseline disease severity seems to be a more important factor in study outcome than study duration, dosing regimen, sample size, time when studies were conducted, and regions where data were generated. Close attention is needed to a variety of factors in the design and conduct of these studies, including patient population, diagnostic considerations, patient assessment, and clinical practice differences. These considerations become increasingly important as globalization of clinical trials continues to increase.

J Clin Psychiatry 2011;72(4):464–472

Submitted: *April 22, 2010; accepted November 9, 2010 (doi:10.4088/JCP.10m06191).*

Corresponding author: *Ni A. Khin, MD, Division of Psychiatry Products, HFD-130, Food and Drug Administration, 10903 New Hampshire Ave, Bldg 22, Rm. 4110, Silver Spring, MD 20993-0002 (ni.khin@fda.hhs.gov).*

There have been a number of changes in recent years in the design and conduct of placebo-controlled trials intended to support new drug applications (NDAs) submitted to the US Food and Drug Administration (FDA), including a shift in where these trials are conducted. Globalization of clinical trials is rapidly becoming a reality, including trials for psychiatric indications. Although the United States remains in the lead regarding the total number of clinical investigators involved in clinical trials in recent years, growth in the numbers of both clinical investigators and clinical trial sites is observed largely outside the United States, in particular, Asia, Eastern Europe, and Latin America.[1–4] It is anticipated that by the year 2012 about 65% of FDA-regulated trials will be conducted at sites outside the United States.[4]

FDA accepts data generated from foreign sites for NDAs as long as they are from adequate and well-controlled trials that are conducted in compliance with the standards of Good Clinical Practice.[5] There have been concerns, however, about the applicability of foreign data in the US population.[6] Possible ethnic differences have been one concern,[7–10] but perhaps somewhat less so as the US population becomes increasingly heterogeneous. Other concerns persist, however, particularly about the applicability of foreign data to US practice because of possible regional differences in disease characteristics, medical practice, and both placebo response and treatment response.[6,10]

Reprinted with correction(s) to pages 467 and 468.

Clinical Points

- A similar treatment effect (drug-placebo difference) was observed in US and non-US MDD trials.
- A rising placebo response and declining treatment effect over time and a persistently high trial failure rate remain concerns.
- Baseline disease severity is a particularly important factor in study outcome.

Additionally, for clinical trials conducted in psychiatric illnesses, we have observed other changes in the way these trials are designed and conducted. Particularly for trials in major depressive disorder (MDD), trials have tended to have larger sample sizes and have become somewhat longer, and there has been greater diversity in patients entered. With regard to patient characteristics, we have noted a gradual decline in MDD severity at baseline in patients entered into clinical trials over this time period.

It has long been noted that, of the randomized placebo-controlled multicenter trials conducted in support of an antidepressant claim, approximately 50% of these trials have failed.[11,12] It has been suggested that this high failure rate may be a result of a substantial increase in the placebo response in these antidepressant trials, particularly in the last decade.[13,14]

This article provides the results of exploratory analyses to examine differences in effect size and success rate of placebo-controlled MDD trials submitted in support of NDAs over the past 25 years. The focus is on drugs widely accepted as effective and, for these drugs, at doses viewed as effective. Differences are examined with regard to when the trials were conducted, the geographic location of the study sites, sample size per treatment arm, trial duration, dosing regimen, and baseline disease characteristics.

METHOD

Data Collection

Eighteen antidepressant programs in support of NDAs submitted to FDA between 1983 and 2008 were identified. Trial data from all NDAs, regardless of approval status, were collected initially and were limited to randomized, multicenter, double-blind, placebo-controlled clinical trials of 4 to 12 weeks' duration with 40 or more patients in at least 1 treatment arm. Patients enrolled in these trials were adult patients (age ≥ 18 y) diagnosed with major depressive disorder (MDD) according to *DSM-III* or *DSM-IV* criteria. Trials limited to known drug responders, such as those in maintenance studies using a randomized withdrawal design, were not included.

This search resulted in 86 MDD trials with a total of 23,817 evaluable patients, defined as patients with a baseline and at least 1 postbaseline efficacy assessment. The year of trial conduct was also noted; if the time of trial initiation was not available, the NDA submission date was recorded. Because the question of greatest interest is what trial design and other factors might affect the success of trials for drugs and doses known to work, the analyses focus on data from this pool of trials for antidepressants widely accepted to be effective and, for these drugs, at doses considered to be within the effective range. Thus, the results reported are based on data for 21,611 patients derived from 81 clinical trials in this pool.

Change from baseline in mean total score of a depression rating scale is typically the primary efficacy measure for antidepressant trials. In these MDD trials, the most commonly used rating scales were the Hamilton Depression Rating Scale (HDRS)[15] and the Montgomery-Asberg Depression Rating Scale (MADRS).[16] Both HDRS and MADRS have been evaluated extensively.[17–20] They are highly correlated and have similar sensitivity in detecting antidepressant efficacy in drug trials. Among the MDD trials in our dataset, almost all studies used change from baseline to endpoint in HDRS total score as the prespecified primary endpoint. Two placebo-controlled trials used the MADRS as the primary measure. Both trials, however, collected HDRS scores as well, and results from the 2 rating scales trended in the same direction. We therefore incorporated all trials that used either HDRS or MADRS as the primary efficacy measure in the database and analyzed the dataset using mean change from baseline to endpoint in the total HDRS-17 or -21 scale score, with missing data imputed by the last-observation-carried-forward (LOCF) approach.

Trial data used in this analysis came from studies in which all participating subjects provided informed consent.

Data Analysis

This meta-analysis included short-term, placebo-controlled trials that were judged to be of adequate size and to have appropriate patient populations and entry criteria. Because individual patient-level datasets were not available in our electronic archives for studies submitted prior to 1997 and for some of the studies after 1997, the analyses for this article were based on aggregated data from sponsors' study reports.

Demographic characteristics (age, gender, race), dropout rate, and baseline disease status in terms of mean HDRS total scores were summarized and compared between the US and non-US trials. Most trials were of either 6 or 8 weeks' duration; therefore, comparisons based on trial duration were limited to these 2 durations. Raw mean changes from baseline in HDRS total score at final visit for placebo and drug treatment arms were calculated based on LOCF data for both US and non-US populations.

Mean treatment effect was calculated as the drug-placebo difference, ie, mean change in HDRS total score for the antidepressant group minus mean change for the placebo group.

Each trial was rated as a success or failure based on whether it succeeded in showing statistical superiority for the investigational drug over placebo on change from baseline to endpoint in the HDRS total score based on LOCF data. For trials that included multiple investigational drug

Reprinted with correction(s) to pages 467 and 468.

Table 1. Demographic Features and Baseline Disease Characteristics of MDD Patients

	US Trials (n = 66)	Non-US Trials (n = 15)	Overall (N = 81)
Total patients in ITT population, No. (%)	17,481 (80.7)	4,180 (19.3)	21,661 (100.0)
Age,[a,b] mean (SD), y	41.8 (4.9)	46.9 (9.1)	42.8 (6.2)
Gender,[a] % female, mean (SD)	59.6 (9.1)	67.4 (7.1)	61.1 (9.2)
Race,[c] % Caucasian, mean (SD)	86.7 (8.4)	92.7 (15.7)	87.4 (9.6)
Dropout rate,[a] %, mean (SD)			
All trials	34.5 (10.4)	26.7 (11.4)	33.0 (10.9)
6-Week trials[d]	38.4 (10.0)	31.9 (7.1)	37.7 (9.9)
8-Week trials[e]	30.4 (9.3)	26.9 (11.3)	29.6 (9.7)
Baseline HDRS total score, mean (SD)			
All trials	23.8 (2.7)	25.3 (2.1)	24.1 (2.6)
Trials with mean baseline HDRS score < 20[f]	18.2 (1.0)	…	18.2 (1.0)
Trials with mean baseline HDRS score ≥ 20[g]	24.3 (2.2)	25.3 (2.1)	24.5 (2.2)
6-Week trials[h]	25.2 (2.4)	24.6 (2.4)	25.1 (2.4)
8-Week trials[i]	22.4 (2.3)	25.6 (1.5)	23.1 (2.5)

[a] Data missing from 2 or 3 US trials.
[b] Two US trials and 1 non-US trial were geriatric trials.
[c] Data missing from 8 US and 7 non-US trials.
[d] Twenty-seven US and 4 non-US trials with a duration of 6 weeks.
[e] Twenty-four US and 7 non-US trials with a duration of 8 weeks.
[f] Five US trials with mean baseline HDRS total score < 20 in each treatment group.
[g] Sixty-one US and 15 non-US trials with baseline HDRS score ≥ 20 in each treatment group.
[h] Baseline HDRS total score calculated based on data from 34 US and 4 non-US trials with 6-week treatment duration.
[i] Baseline HDRS total score calculated based on data from 24 US and 7 non-US trials with 8-week treatment duration.
Abbreviations: HDRS = Hamilton Depression Rating Scale, ITT = intent to treat, MDD = major depressive disorder.

groups (different dosages), a trial was rated as successful if there was statistical superiority in at least 1 investigational drug group after adjusting for multiplicity. The multiplicity adjustment was based on the preplanned analysis that in every instance was judged by FDA to adequately control the overall study-wise type I error rate. Mean treatment effect and trial success rate were assessed with regard to where (US vs non-US) and when (prior to 1995 [1983–1994] and afterward [1995–2008)]) the trials were conducted, sample size per arm (< 50, 50–100, > 100), trial duration (6 weeks or 8 weeks), dosing regimen (fixed-dose vs flexible-dose design), and baseline disease severity (HDRS total score).

This analysis is descriptive only, as there were undoubtedly many unidentified interaction/confounding factors and data from individual patients were not available. Statistical modeling based on summary data has the potential to be misleading. Thus, we have limited data presentations in this article to simple plots, tables, and other summary statistics.

RESULTS

Among the 86 trials in 18 NDAs, 65 trials were conducted solely in the United States, 1 trial was conducted only in Canada, 5 were mixed trials conducted in both the United States and Canada, and 15 trials were conducted solely in foreign countries. For the purposes of this article, the North America region (United States and Canada) is referred to as the United States because the Canadian sites contributed a very limited amount of data (< 3%). This meta-analysis (n = 81) focused on clinical trial data from drugs accepted as effective and at doses viewed as effective for these drugs. Exploratory analyses using the data from all 86 trials regardless of dosing or outcome provided similar results. Two-thirds of trials (n = 53) used a flexible-dose design, and one-third of the trials (n = 28) were fixed-dose trials. Most of the trials (77%) were 6- or 8-week trials, but 3 trials were 4 weeks in duration and 9 trials had up to 12 weeks of double-blind treatment.

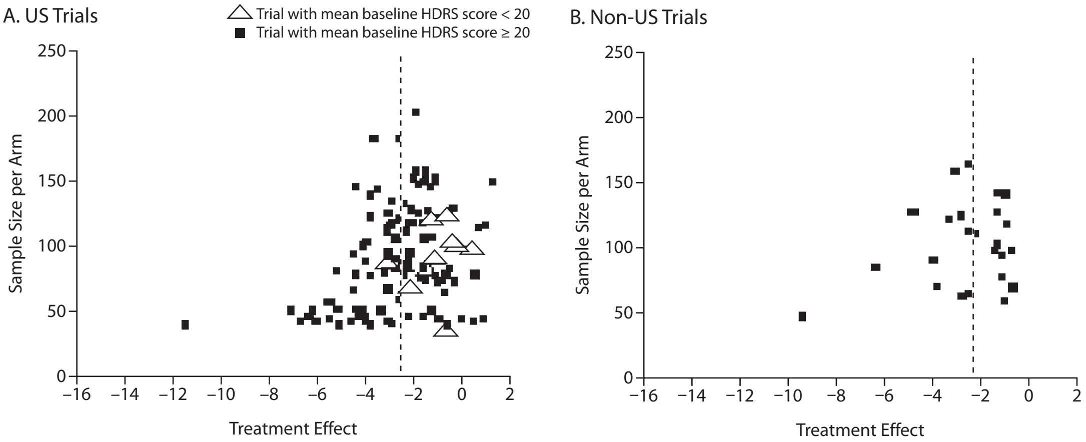
Table 1 lists the numbers of participants for the US and non-US sites, as well as the demographic characteristics, baseline disease status, and dropout rates. Eighty-one percent of the MDD patients were enrolled in US sites. Patients were predominantly Caucasian (87%), were 61% female, and had a mean age of 43 years and an average baseline HDRS total score of about 24. Of the 81 trials, 5 trials had a mean baseline HDRS total score < 20. The mean baseline HDRS total scores for trials with a baseline HDRS score ≥ 20 were 24.3 for US trials and 25.3 for non-US trials. All 5 trials with a mean baseline HDRS total score < 20 were conducted in the United States and were of 8 weeks' duration. This may account for the lower mean baseline HDRS total score observed for 8-week trials in the US as compared to outside the US (22.4 vs 25.6). The average dropout rate was slightly higher in the US trials (34.5%) compared to non-US trials (26.7%). The average dropout rate for trials with 6 weeks' duration was also slightly higher (38%) than for those with 8 weeks' duration (30%).

The mean changes from baseline in HDRS total score for the placebo groups were −8.0 (range, −3.7 to −12.4) for the US trials and −9.5 (range, −4.8 to −13.8) for the non-US trials. The mean changes from baseline in HDRS total score for the antidepressant groups were −10.4 (range, −5.3 to −16.1) for the US trials and −12.5 (range, −6.3 to −15.4) for the non-US trials. Figure 1 provides a comparison of observed treatment effects (ie, drug-placebo difference) for the US and non-US trials. The horizontal axis denotes the estimated change from baseline to endpoint in HDRS total score in each treatment group, and the vertical axis, the total number of evaluable patients per treatment arm in the corresponding trial. In Figure 1A (US trials), the data are separated into trials with a mean baseline HDRS total score ≥ 20 (n = 61) and trials with a mean baseline HDRS total score < 20 (n = 5). No difference in the treatment effect relative to placebo was observed between the US and non-US trials. The mean treatment effect (marked by a vertical line in the figure) was −2.5 with a standard deviation (SD) of 2.0 regardless of the region. The treatment effects from the 5 MDD trials with mean baseline HDRS total score < 20 were smaller than for the other trials (Figure 1A).

It has been suggested that the high failure rate of antidepressant trials has resulted from an increasing placebo response over time.[13,14] Figure 2 was generated in order to explore possible changes in placebo response and effect size

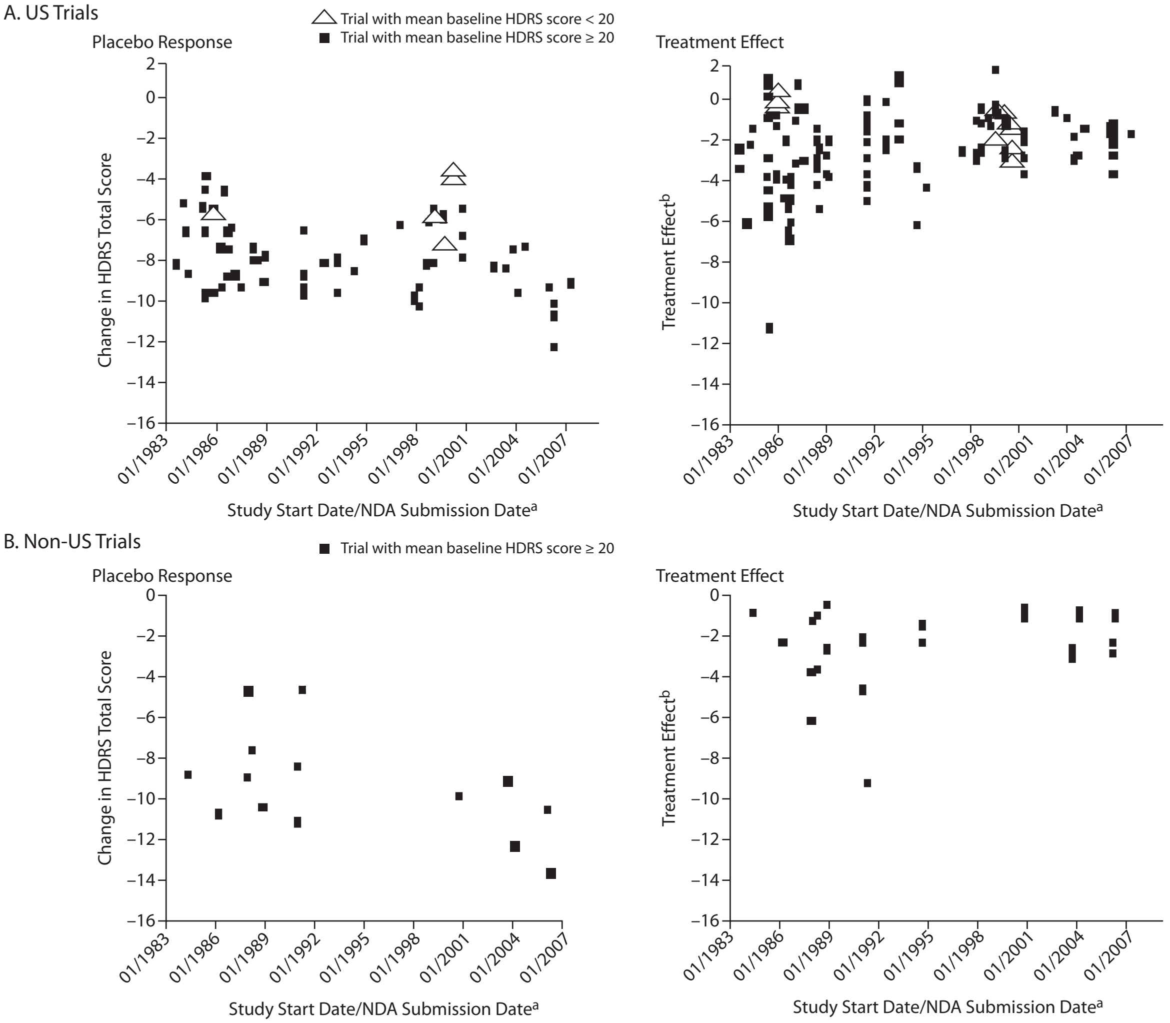
Reprinted with correction(s) to pages 467 and 468.

Figure 1. Treatment Effect Relative to Placebo (drug-placebo difference) Based on Mean Change From Baseline to Endpoint (LOCF) in HDRS Total Scores in US and Non-US MDD Trials[a]



[a]Dashed lines indicate mean treatment effect.
Abbreviations: HDRS = Hamilton Depression Rating Scale, LOCF = last observation carried forward, MDD = major depressive disorder.

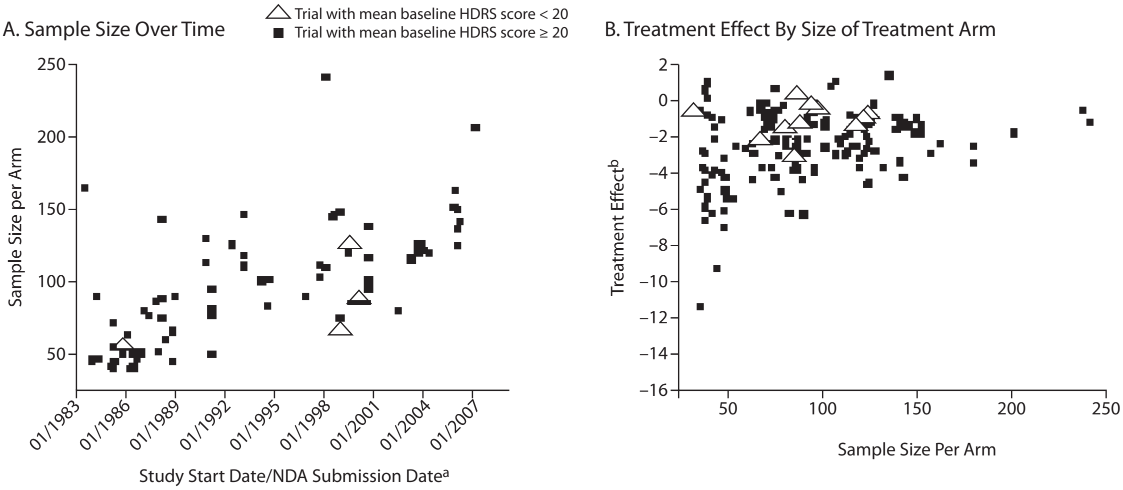
Figure 2. Placebo Responses and Treatment Effects Over Time in US and Non-US MDD Trials



[a]Horizontal axes denote the trial start date, if known, or, if unknown, the NDA submission date. [b]Calculated as the drug-placebo difference, ie, mean change in HDRS total score for the antidepressant group minus mean change for the placebo group.
Abbreviations: HDRS = Hamilton Depression Rating Scale, MDD = major depressive disorder, NDA = new drug application.

Reprinted with correction(s) to pages 467 and 468.

Figure 3. Potential Impact of Sample Size in US and Non-US MDD Trials



[a]Trial start date, if known, or, if unknown, the NDA submission date. [b]Calculated as the drug-placebo difference, ie, mean change in HDRS total score for the antidepressant group minus mean change for the placebo group.
Abbreviations: HDRS = Hamilton Depression Rating Scale, MDD = major depressive disorder, NDA = new drug application.

Table 2. Success Rates of MDD Trials[a]

Study Time Period[b]	US Trials	Non-US Trials	Overall
Entire time span			
All trials[c]	38/66 (58)	5/15 (33)	43/81 (53)
6-Week trials	21/34 (62)	0/4 (0)	21/38 (55)
8-Week trials	10/24 (42)	3/7 (43)	13/31 (42)
1983–1994			
All trials[c]	24/39 (62)	3/10 (30)	27/49 (55)
6-Week trials	18/30 (60)	0/4 (0)	18/34 (53)
8-Week trials	4/7 (57)	1/3 (33)	5/10 (50)
1995–2008			
All trials[c]	14/27 (52)	2/5 (40)	16/32 (50)
6-Week trials	3/4 (75)	…	3/4 (75)
8-Week trials	6/17 (35)	2/4 (50)	8/21 (38)

[a]The numerators in the cells indicate the number of successful trials, and the denominators indicate the total number of trials. Success rates are expressed as percentages in parentheses.
[b]For study conduct time period, the study start date was used. If not available, the new drug application submission date was used.
[c]Included all studies with duration between 4 to 12 weeks, inclusive.
Abbreviation: MDD = major depressive disorder.

over time. The top plots (Figure 2A) and the bottom plots (Figure 2B) display the observed placebo responses and treatment effects over time (1983–2008) in the US and non-US trials, respectively. These plots suggest an increase in placebo response for US and non-US trials over this time period. The plots for observed treatment effect (drug-placebo difference) suggest a diminishing treatment effect size over time in both US and non-US trials, with the typical effect size moving toward 2 HDRS units in contrast to earlier values closer to 3.

Treatment effect was also assessed with regard to sample size per arm. Figure 3A shows a trend for sample size per arm to increase over time. Despite this increase in sample size per arm, there has been a decrease in treatment effect (Figure 3B).

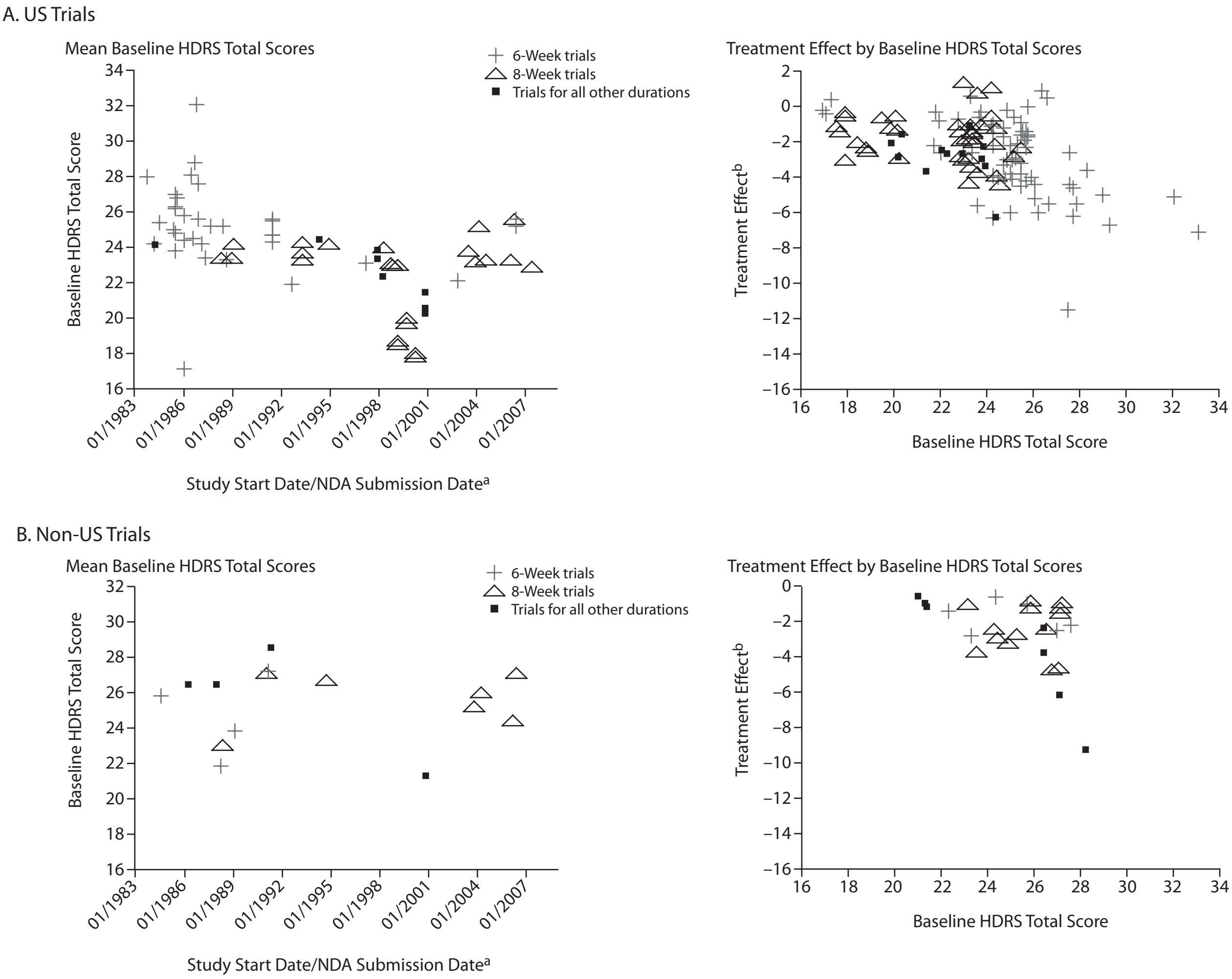
Success rates for trials were also explored. A trial was considered successful if efficacy was demonstrated in at least 1 investigational drug group. In trials in which several doses were compared to placebo, a multiplicity adjustment was applied based on the preplanned procedure that in every instance adequately controlled the overall study-wise type I error rate. The overall trial success rate over the 25 years was 53% (Table 2); it was higher in US trials compared to non-US trials (58% vs 33%). For 8-week trials, the overall success rate was 42%, and roughly the same for US and non-US trials. Six-week trials had a slightly higher overall success rate compared to 8-week trials (55% vs 42%); however, the numbers were too small to provide any basis for comparing US and non-US trials. The results in Table 2 included the 5 MDD trials in the United States with mean baseline HDRS total scores < 20. These trials were of 8 weeks' duration, and only 1 was successful. If these trials were excluded from the analysis, the success rate for 8-week trials would be 50%. Trials conducted prior to 1995 (1983–1994) had a 55% success rate, compared to a 50% success rate for trials conducted from 1995 to 2008. When success rates were calculated based on placebo arm sample sizes of < 50, 50–100, and > 100, the success rates were observed to be 61%, 48%, and 53%, respectively. The average mean treatment effect in studies conducted before 1995 was −3.0 HDRS units (SD = 2.4) as compared to −1.8 HDRS units (SD = 1.0) in trials conducted since 1995.

Two of the panels in Figure 4 display the mean baseline HDRS total score over time in each region. The figure reveals a slightly downward trend in US trials over time. Although there is no clear trend in non-US trials, there were fewer trials. The same figure reveals an increasing treatment effect as the baseline HDRS total score increases, regardless of region.

Two-thirds of the MDD trials (53 of 81) utilized a flexible dosing regimen. The potential impact of dosing regimen on placebo response and treatment effect in our dataset was also explored. As can be seen in Table 3, the observed placebo response was similar for these different dosing strategies; however, the observed treatment effect was larger for flexible-dose studies compared to fixed-dose studies, both

Reprinted with correction(s) to pages 467 and 468.

Figure 4. Mean Baseline HDRS Total Scores for Each MDD Trial and Treatment Effect by Baseline HDRS Total Score



[a]Trial start date if known, or if unknown, the NDA submission date. [b]Calculated as the drug-placebo difference, ie, mean change in HDRS total score for the antidepressant group minus mean change for the placebo group.
Abbreviations: HDRS = Hamilton Depression Rating Scale, MDD = major depressive disorder, NDA = new drug application.

Table 3. Placebo Response, Treatment Effect, Dropout Rate, and Trial Success Rate by Dosing Regimen

	Fixed Dosing			Flexible Dosing		
	US Trials	Non-US Trials	Overall	US Trials	Non-US Trials	Overall
No. of trials	22	6	28	44	9	53
Placebo response[a]	−7.7 (2.0)	−9.9 (0.7)	−8.1 (2.0)	−8.0 (1.8)	−9.3 (3.1)	−8.2 (2.1)
No. of drug groups	48	14	62	71	14	85
Treatment effect[b]	−2.0 (1.3)	−2.2 (1.9)	−2.0 (1.5)	−2.9 (2.3)	−2.7 (2.4)	−2.9 (2.3)
Dropout rate, %, mean (SD)	33.2 (9.2)	18.5 (9.7)	30.0 (11.0)	35.0 (11.0)	32.2 (9.2)	34.0 (10.7)
Trial success rate,[c] %	59	50	57	57	22	51

[a]Based on mean change from baseline in Hamilton Depression Rating Scale total score (SD).
[b]Averaged treatment effect (drug-placebo difference) over total number of drug groups (SD).
[c]Trials were rated as successful if they showed statistical superiority vs placebo for the investigational drug on change in Hamilton Depression Rating Scale (HDRS) score (last-observation-carried-forward data). (Trials with multiple investigational drug groups were successful if there was superiority in at least 1 drug group after adjustment for multiplicity.)

overall and within US and non-US studies separately. The mean effect sizes in HDRS units were −2.9 (SD = 2.3) for flexible-dose versus −2.0 (SD = 1.5) for fixed-dose studies. Trial success rates, however, were slightly greater in fixed-dose studies compared to flexible-dose studies (57% vs 51%). Flexible- and fixed-dose trials did not differ meaningfully in dropout rates.

DISCUSSION

The purpose of this meta-analysis was to look for possible differences in treatment effect and success rate of placebo-controlled MDD trials in drugs determined to be effective and at effective doses for these drugs, based on when and where the trials were conducted and with regard

Reprinted with correction(s) to pages 467 and 468.

to trial duration, sample size per arm, and baseline disease characteristics. The potential impact of dosing regimen on treatment effect was also explored.

As drug research has expanded globally, questions have been raised about the applicability of data from diverse foreign sites to the US population. One concern has been about possible differences in response in the placebo group and effect size across different geographic regions. In our meta-analysis, responses in both placebo and drug groups from non-US trials tended to be larger than those observed in the US trials; however, treatment effect (drug-placebo difference) was on average about the same for US and non-US trials. This finding needs to be interpreted with caution as it was based on estimates from the aggregated data rather than individual patient-level data. Furthermore, these estimates were based on the LOCF imputation method, which was the only estimation approach available for all trials. Estimates from LOCF data are likely to be biased when the mechanism of missing data is not completely at random, particularly in the presence of a high dropout rate.[21]

Another concern for antidepressant trials has been an apparent increasing placebo response over time,[13,14] along with a decrease in treatment effect size. High placebo response is considered a major factor contributing to the substantial failure rate observed in MDD trials. We have confirmed this finding; that is, we observed an increasing placebo response for both US and non-US trials conducted over a 25-year period, along with a decline in treatment effect size. As previously reported, conducting larger studies has not generally produced a better outcome for depression trials.[22] Despite a marked increase in the sample size per treatment arm of the trials in our database over time, treatment effect size in depression trials has been diminishing over time, and the trend appears to be similar in US and non-US trials.

Several authors have commented on the relatively modest overall success rate for antidepressant trials.[11–13] The overall trial success rate of all MDD trials over the 25-year period for our current meta-analysis was 53%. When broken out by time of trial conduct, the success rate was 55% for trials conducted during the earlier time period (1983–1994) and 50% for more recently conducted trials (1995–2008). This later time period included 5 MDD trials in the United States with mean baseline HDRS total scores $<$ 20, and only 1 of them was successful. When these MDD trials involving more mildly ill patients were excluded, the overall success rate was 55%.

It has been challenging to try to understand the basis for an apparent decrease in treatment effect in MDD trials over time. One possibility that has been suggested is that the effect of antidepressants is heavily influenced by baseline disease severity, perhaps because of underlying differences in disease. It is a widely held view that enrolling more severely ill patients with MDD should increase the chances of having a successful trial. Previously published results are mixed, however, in showing a clear relationship of baseline depression scores to placebo response or antidepressant effect.[23–27] Our analyses suggest that trials enrolling patients with higher mean baseline HDRS total scores tend to have larger treatment effects regardless of region. In addition, trial duration, patients' average baseline HDRS total scores, time of study conduct, and the trial success rates seem to be correlated. It appears that the earlier trials were mostly 6 weeks, and the 6-week trials had higher mean baseline scores and higher success rates than the 8-week trials. Because our research is based only on the limited study-level efficacy data from 81 MDD trials, we were not able to employ inferential statistical methods to adjust for confounding and interaction effects.

It has been suggested that flexible-dose trials should be favored for MDD because of a greater probability of success with these trials. Khan et al[28,29] reported success rates in 51 MDD trials of 59.6% (34/57 of the antidepressant treatment arms) for those of flexible-dose design compared to only 31.4% (11/35 of the antidepressant treatment arms) for those of fixed-dose design. Khan's group reported that symptom reduction (defined as percent change in mean HDRS total score) was similar for the antidepressant treatment arms in both flexible- and fixed-dose trials, but the magnitude of symptom reduction with placebo was smaller in the flexible-dose trials compared to that observed in the fixed-dose trials. Our analysis found similar placebo responses but some increase in treatment effect size in the flexible-dose trials compared to fixed-dose trials. Unlike Khan et al, we actually found a slightly higher overall success rate for fixed-dose (57%) trials compared to flexible-dose trials (51%). A major reason for these differences between our analyses compared to the Khan et al analyses is the different databases used; we extended the database by including 37 more recent trials that were not part of Khan and colleagues' original database. We calculated success rates using number of trials as the denominator (an approach that we prefer), while Khan's group calculated success rate using total number of treatment arms as the denominator. This difference in approach did not, however, affect the results for each database looked at separately. Specifically, when our approach was applied to the same trials included in Khan's original analysis, our trial success rates were similar to Khan's findings, ie, success rates of 61% for flexible-dose trials and 33% for fixed-dose trials. Similarly, when Khan and colleagues' approach was applied to our 37 more recent studies (not part of their original database), the findings were in favor of the fixed-dose design, with a success rate of 60.5% (23/38 of the drug treatment arms) compared to 34.5% (10/29) for the flexible-dose arms. Using our approach for these more recent trials, the success rate was 70.6% (12/17) for the fixed-dose trials compared to 30% (6/20) for the flexible-dose studies.

Although the reasons for the different trends in success rate between flexible- and fixed-dose studies for earlier compared to more recent MDD trials are not entirely clear, the placebo response (mean = −9.2; SD = 2.3) was larger in the more recent set of flexible-dose studies compared to the earlier studies (mean = −7.6; SD = 1.9), while the percentage of patients assigned to placebo remained consistent

Reprinted with correction(s) to pages 467 and 468.

around 40% (SD = 8). We also note that, in the earlier fixed-dose studies, about 25% (SD = 6.3) of patients were assigned to placebo with a mean placebo response around −8.5 (SD = 1.5), while for the more recent fixed-dose studies the proportion of placebo patients was 31% (SD = 9.0), with a mean placebo response of −7.9 (SD = 2.2). Another meta-analysis by Papakostas and Fava[30] published in 2009 and based on MEDLINE/PubMed databases noted that a greater probability of receiving placebo, higher baseline severity, and earlier year of publication predicted greater treatment effect. It was noted in the same paper that fixed- versus flexible-dose design and trial duration did not influence treatment effect defined as response rate.

Sponsors have endeavored for years to find ways to avoid high placebo response and high failure rates for MDD trials. One approach has been a placebo run-in period, ie, an attempt to identify and exclude placebo responders. It has been shown, however, that using a placebo run-in phase has not been a successful strategy in lowering the placebo response rate, nor has it increased drug-placebo differences in MDD trials.[31–33] An alternative enrichment design, ie, the sequential parallel comparison design, has been proposed as an approach to minimizing placebo response in psychiatric trials.[34,35] Further evaluation of the strengths and weaknesses of this study design and its implications will be needed. This novel design has not yet been applied in a regulatory setting. Targeting sicker patients by setting higher thresholds for enrolling patients would appear to be the single step that would best increase success, but this approach could be offset by score inflation at study sites, a serious concern in study conduct.[36] Other factors contributing to high failure rates of trials include poor interrater reliability, interview quality, and rater bias.[37,38] Centralized ratings have been proposed as one approach for improving the precision of patient ratings.[38,39] Another approach to improving patient ratings has been the use of self ratings, particularly through computer automated systems.[40,41]

One of the limitations of our exploratory analysis was the fact that individual patient-level datasets were not available in our electronic archives for studies submitted to FDA prior to 1997, and for some of the studies after 1997. Consequently, the meta-analysis for this article was performed entirely based on the aggregated data from study reports submitted by various sponsors. These findings, therefore, must be considered preliminary until patient-level data can be compiled to support more definitive future analyses. Such analyses will be feasible as FDA moves toward a standard of having all industry sponsors submit datasets in a global platform for regulatory submissions using the electronic common technical document (e-CTD) specifications[42] and in the Clinical Data Interchange Standards Consortium (CDISC), such as Study Data Tabulation Model (SDTM) standard and ADaM (Analysis Data Model) standard.[43] Standardization of data structures and terminology will facilitate the conduct by FDA of a more efficient and comprehensive data review. In addition, having standards for electronic submissions will enable data aggregation and the population of cross-study and cross-product databases that will greatly enhance FDA's capability to perform meta-analyses. However, the success of such meta-analyses still depends on having high-quality data.

Global drug development is inevitable, and continued efforts are needed to try to understand differences between findings from US and non-US sites, although, at least for antidepressants, US and non-US sites appear to have similar overall treatment effect sizes. Differences in study results from trials conducted in various geographic regions for other disorders, however, have been reported.[10,44–46] Possible reasons for such discrepancies may include differences in body weight, drug metabolism, and other ethnic and genetic differences, but also might involve differences in regulatory requirements, medical practice, access to care, and exposure to medications prior to the trials. These differences should be taken into account when planning global trials. Another factor that is not addressed in this article is the compliance issues that reflect clinical practice. The consideration of population pharmacokinetic data is beyond the scope of this article.

Although we are reassured by the finding of a comparable treatment effect (drug-placebo difference) in US and non-US MDD trials, the rising placebo response over time and the high overall trial failure rate remain concerns. Great care is needed in designing and conducting multiregional studies, with attention to possible differences in patient population, diagnostic practices, disease severity, and clinical care of patients. Much additional work is needed to improve the design and conduct of MDD trials, and this effort would benefit from cooperation among academicians, industry sponsors, and regulators.

Drug names: Trial data from all new drug applications for antidepressants regardless of approval status were included. For reasons of confidentiality, individual drugs were not reported here.

Author affiliations: Division of Psychiatry Products, Office of Drug Evaluation I, Office of New Drugs (Drs Khin and Laughren); and Division of Biometrics I, Office of Biostatistics, Office of Translational Science (Drs Chen, Y. Yang, and P. Yang), Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, Maryland.

Potential conflicts of interest: None reported.
Funding/support: None reported.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent those of the FDA.

Previous presentations: A preliminary analysis of this work was presented in a special symposium session at the 110th Annual Meeting of the American Society for Clinical Pharmacology and Therapeutics (ASCPT); March 18–21, 2009; National Harbor, Maryland; and at the Joint Korean Society for Clinical Pharmacology and Therapeutics (KSCPT)/ASCPT Meeting; December 1–3, 2009; Seoul, South Korea.
It was also presented as part of the FDA symposium at the 49th Annual New Clinical Drug Evaluation Unit (NCDEU) Meeting; June 29–July 2, 2009; Hollywood Beach, Florida.
Acknowledgments: The authors thank Robert Temple, MD, and H. M. James Hung, PhD, CDER/FDA, for their critical review of the manuscript. Drs Temple and Hung report no potential conflict of interest.

REFERENCES

1. Getz KA. Global clinical trials activity in the details. *Applied Clinical Trials.* 2007;16(9):42–44.
2. Thiers FA, Sinskey AJ, Berndt ER. Trends in the globalization of clinical trials. *Nat Rev Drug Discov.* 2008;7(1):13–14.

Reprinted with correction(s) to pages 467 and 468.

3. Karlberg JPE. Globalization of sponsored clinical trials. *Nat Rev Drug Discov*. 2008;7(5):458–460.
4. Tufts CSDD Outlook. Boston, MA: Tufts Center for the Study of Drug Development; 2008. http://csdd.tufts.edu/_documents/www/Outlook2008.pdf Accessed May 7, 2010.
5. Food and Drug Administration. Code of Federal Regulations Title 21, Parts 300 to 499. Revised April 1, 2009. Washington, DC: US Government Printing Office; 2009: Ch 1; 312.120: 84–85; 314.106: 140.
6. Temple R. Use of Non-US Data in NDAs. Special Session: Regulatory Considerations of Using Non-US Data in NDAs: Focus on Efficacy, Safety and Clinical Pharmacology. Presented at the Annual Meeting of the American Society for Clinical Psychopharmacology and Therapeutics; March 20, 2009; National Harbor, MD. http://www.ascpt.org/Portals/8/docs/Meetings/2009%20Annual%20Meeting/Friday/Regulatory%20Considerations.pdf. Accessed February 25, 2011.
7. Food and Drug Administration. Guidance for Industry. ICH E-5: Ethnic factors in the acceptability of foreign clinical data. Published February 5, 1998. Updated Questions and Answers, June 2006. http://www.fda.gov/RegulatoryInformation/Guidances/ucm129314.htm. Accessed August 21, 2009.
8. Huang S-M, Temple RT. Is this the drug or dose for you? Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. *Clin Pharmacol Ther*. 2008;84(3):287–294.
9. Yasuda SU, Zhang L, Huang S-M. The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. *Clin Pharmacol Ther*. 2008;84(3):417–423.
10. Committee for Medicinal Products for Human Use (CHMP). European Medicines Agency (EMEA): Reflection paper on the extrapolation of results from clinical studies conducted outside Europe to the EU-population. Reference number: EMEA/CHMP/EWP/692702/2008. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/11/WC500013468.pdf. Updated October 22, 2009. Accessed March 31, 2010.
11. Laughren TP. The scientific and ethical basis for placebo-controlled trials in depression and schizophrenia: an FDA perspective. *Eur Psychiatry*. 2001;16(7):418–423.
12. Khan A, Khan S, Brown WA. Are placebo controls necessary to test new antidepressants and anxiolytics? *Int J Neuropsychopharmacol*. 2002;5(3):193–197.
13. Khan A, Detke M, Khan SR, et al. Placebo response and antidepressant clinical trial outcome. *J Nerv Ment Dis*. 2003;191(4):211–218.
14. Walsh BT, Seidman SN, Sysko R, et al. Placebo response in studies of major depression: variable, substantial, and growing. *JAMA*. 2002;287(14):1840–1847.
15. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56–62.
16. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134(4):382–389.
17. Hedlund JL, Vieweg BW. The Hamilton Rating Scale for Depression: a comprehensive review. *J Oper Psychiatr*. 1979;10:149–165.
18. Kearns NP, Cruickshank CA, McGuigan KJ, et al. A comparison of depression rating scales. *Br J Psychiatry*. 1982;141(1):45–49.
19. Mittmann N, Mitter S, Borden EK, et al. Montgomery-Asberg severity gradations. *Am J Psychiatry*. 1997;154(9):1320–1321.
20. Khan A, Khan SR, Shankles EB, et al. Relative sensitivity of the Montgomery-Asberg Depression Rating Scale, the Hamilton Depression Rating Scale and the Clinical Global Impressions rating scale in antidepressant clinical trials. *Int Clin Psychopharmacol*. 2002;17(6):281–285.
21. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York, NY: Springer-Verlag; 2000.
22. Liu KS, Snavely DB, Ball WA, et al. Is bigger better for depression trials? *J Psychiatr Res*. 2008;42(8):622–630.
23. Wilcox CS, Cohn JB, Linden RD, et al. Predictors of placebo response: a retrospective analysis. *Psychopharmacol Bull*. 1992;28(2):157–162.
24. Khan A, Leventhal RM, Khan SR, et al. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol*. 2002;22(1):40–45.
25. Khan A, Brodhead AE, Kolts RL, et al. Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. *J Psychiatr Res*. 2005;39(2):145–150.

26 Kirsch I, Deacon BJ, Huedo-Medina TB, et al. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS Med*. 2008;5(2):e45, 260–268.

27. Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA*. 2010;303(1):47–53.
28. Khan A, Khan SR, Walens G, et al. Frequency of positive studies among fixed and flexible dose antidepressant clinical trials: an analysis of the Food and Drug Administration summary basis of approval reports. *Neuropsychopharmacology*. 2003;28(3):552–557.
29. Khan A, Kolts RL, Thase ME, et al. Research design features and patient characteristics associated with the outcome of antidepressant clinical trials. *Am J Psychiatry*. 2004;161(11):2045–2049.
30. Papakostas GI, Fava M. Does the probability of receiving placebo influence clinical trial outcome? a meta-regression of double-blind, randomized clinical trials in MDD. *Eur Neuropsychopharmacol*. 2009;19(1):34–40.
31. Trivedi MH, Rush H. Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsychopharmacology*. 1994;11(1):33–43.
32. Faries DE, Heiligenstein JH, Tollefson GD, et al. The double-blind variable placebo lead-in period: results from two antidepressant clinical trials. *J Clin Psychopharmacol*. 2001;21(6):561–568.
33. Lee S, Walker JR, Jakul L, et al. Does elimination of placebo responders in a placebo run-in increase the treatment effect in randomized clinical trials? a meta-analytic evaluation. *Depress Anxiety*. 2004;19(1):10–19.
34. Fava M, Evins AE, Dorer DJ, et al. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. *Psychother Psychosom*. 2003;72(3):115–127.
35. Tamura RN, Huang X. An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clin Trials*. 2007;4(4):309–317.
36. Landin R, DeBrota DJ, DeVries TA, et al. The impact of restrictive entry criterion during the placebo lead-in period. *Biometrics*. 2000;56(1):271–278.
37. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull*. 1998;34(1):9–24.
38. Kobak KA, Kane JM, Thase ME, et al. Why do clinical trials fail? the problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol*. 2007;27(1):1–5.
39. Kobak KA. A comparison of face-to-face and videoconference administration of the Hamilton Depression Rating Scale. *J Telemed Telecare*. 2004;10(4):231–235.
40. Mundt JC, Greist JH, Jefferson JW, et al. Is it easier to find what you are looking for if you think you know what it looks like? *J Clin Psychopharmacol*. 2007;27(2):121–125.
41. Rush AJ, Bernstein IH, Trivedi MH, et al. An evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: a Sequenced Treatment Alternatives to Relieve Depression trial report. *Biol Psychiatry*. 2006;59(6):493–501.
42. Final Guidance for Industry: Providing Regulatory Submissions in Electronic Format–Human Pharmaceutical Applications and Related Submissions Using the eCTD Specifications. June 2008 Revision 2: 1–16. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM072349.pdf. Accessed December 15, 2009.
43. Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) and CDISC Analysis Data Model (ADaM). http://www.cdisc.org/standards. Accessed December 15, 2009.
44. Hjalmarson A, Goldstein S, Fagerberg B, et al; MERIT-HF Study Group. Effects of controlled-release metoprolol on total mortality, hospitalizations, and well-being in patients with heart failure: the Metoprolol CR/XL Randomized Intervention Trial in congestive heart failure (MERIT-HF). *JAMA*. 2000;283(10):1295–1302.
45. Domanski M, Antman EM, McKinlay S, et al. Geographic variability in patient characteristics, treatment and outcome in an International Trial of Magnesium in acute myocardial infarction. *Control Clin Trials*. 2004;25(6):553–562.
46. Blair JEA, Zannad F, Konstam MA, et al; EVEREST Investigators. Continental differences in clinical characteristics, management, and outcomes in patients hospitalized with worsening heart failure results from the EVEREST (Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan) program. *J Am Coll Cardiol*. 2008;52(20):1640–1648.

Reprinted with correction(s) to pages 467 and 468.