Do Efficacy and Effectiveness Samples Differ in Antidepressant Treatment Outcome? An Analysis of Eligibility Criteria in Randomized Controlled Trials

Florian Seemüller, MD; Hans-Jürgen Möller, MD; Michael Obermeier, MSc; Mazda Adli, MD; Michael Bauer, MD, PhD; Klaus Kronmüller, MD; Florian Holsboer, MD, PhD; Peter Brieger, MD; Gerd Laux, MD; Wolfram Bender, MD; Isabella Heuser, MD, PhD; Joachim Zeiler, MD; Wolfgang Gaebel, MD; Rebecca Schennach-Wolff, MD; Verena Henkel, MD; and Michael Riedel, MD

Background: Because of strict inclusion and exclusion criteria, results drawn from placebo-controlled randomized antidepressant efficacy trials may not be transferable to real-world patients.

Method: This study was performed from March 2000 to September 2005 as a prospective, multicenter follow-up. Patients were recruited from February 2000 to June 2005. All patients were hospitalized (N = 1,014) and met *DSM-IV* criteria for major depressive episode. Assessments with the 21-item Hamilton Depression Rating Scale were conducted biweekly until discharge. According to the most commonly applied exclusion criteria in randomized controlled antidepressant efficacy trials, patients were retrospectively divided into 2 groups: (1) patients not fulfilling exclusion criteria and therefore eligible for a randomized placebo-controlled trial, referred to as "efficacy sample," and (2) patients fulfilling at least 1 exclusion criterion, not being eligible for inclusion in an efficacy trial ("nonefficacy sample"). The efficacy sample was compared with the nonefficacy sample in terms of sociodemographic and clinical baseline variables and outcome measures, such as remission and response rates, 17-item Hamilton Depression Rating Scale mean scores, time to remission, and time to response.

Results: Significant differences were found, with the efficacy sample being older ($P = .03$) and being more often treated at a university hospital ($P = .02$). The efficacy sample demonstrated superior outcome only in significantly higher mean Global Assessment of Functioning scores at discharge ($P = .03$). There were no differences regarding remission ($P = .68$) and response ($P = .06$) rates, length of hospital stay ($P = .49$), 17-item Hamilton Depression Rating Scale total score at discharge ($P = .13$), or time to response ($P = .39$) or remission ($P = .16$).

Conclusions: Both groups differed significantly in several baseline measures and final Global Assessment of Functioning scores but not in any other outcome measure. Challenging current beliefs, our findings show that results from efficacy antidepressant trials might be more generalizable than previously thought.

J Clin Psychiatry 2010;71(11):1425–1433

Submitted: *February 20, 2009; accepted June 15, 2009.*
Online ahead of print: *August 10, 2010 (doi:10.4088/JCP.09m05166blu).*
Corresponding author: *Florian Seemüller, MD, Department of Psychiatry, Ludwig-Maximilians-Universität, Nussbaumstrasse 7, 80336 Munich, Germany (florian.seemueller@med.uni-muenchen.de).*

Evidence-based medicine was recently ranked as one of the 10 most important medical breakthroughs of the 20th century.[1] Evidence-based medicine aims to apply evidence gained from well-designed scientific trials to medical practice. The highest level of evidence (level A) is mostly defined as evidence obtained from consistent randomized clinical trials. In order to license a new drug in psychiatry, in most indications, placebo-controlled phase 3 studies are demanded by the drug authorities.

The challenge for each physician in everyday clinical practice remains to extrapolate these findings to the individual patient, who usually presents with a more complex illness picture than the strictly defined medical condition investigated in the corresponding phase 3 study. Thus, the question arises as to how generalizable and transferable to everyday clinical practice results drawn from randomized controlled trials with high internal validity, specifically from phase 3 studies, are.

In an earlier attempt, our group retrospectively compared patients with schizophrenia who met inclusion criteria of phase 3 studies and 100 controls matched for diagnosis, sex, and age. We found no significant differences in this case-control study with respect to illness severity (Clinical Global Impressions scale and Global Assessment of Functioning) at trial end.[2]

As for depression, Zimmerman et al[3] recently demonstrated that only 14% of their outpatient population would comply with the strict exclusion criteria usually applied in randomized controlled antidepressant trials. They used criteria that they found to be most consistently used after a careful systematic review of 31 antidepressant efficacy trials.[4] Most recently, Zetin and Hoepner[5] were able to replicate these findings in applying identical criteria to a different cohort of 348 outpatients with depression. Only 9% in this population would have qualified for a randomized controlled trial. In a subsequent study,[3] the authors compared baseline variables of patients who would and who would not qualify for inclusion in a phase 3 trial and found statistically significant differences in the sense that patients not fulfilling the criteria for a phase 3 trial had a longer episode duration, more previous episodes, more prior suicide attempts, and a higher chance of having a cluster B personality disorder. The authors concluded that their findings further support

caution in generalizing results from antidepressant efficacy trials to clinical populations.

But the most pressing question in this context remains unanswered: are there significant differences in outcomes between samples meeting eligibility criteria for phase 3 studies and patients that are usually excluded? Today, there are very limited data with respect to the effect of typical inclusion and exclusion criteria on outcome and specifically on any longitudinal variables such as mean course graphs or survival analysis in a representative population.

We therefore used the exclusion criteria proposed by Zimmerman et al[3] and applied them retrospectively to a large sample (N = 1,014) of naturalistically treated "real-world" inpatients, whom we divided into 2 groups (efficacy sample vs nonefficacy sample). Both groups were compared in several clinical and sociodemographic baseline features as well as outcome variables, which comprised remission and response rates, 17-item Hamilton Depression Rating Scale (HDRS-17)[6] course graphs, and survival curves.

Our main hypotheses were that patients who retrospectively qualified for inclusion in a phase 3 study (efficacy sample) would have better response and remission rates, would show earlier time to response and time to remission, would exhibit a more pronounced decline in the HDRS-17 mean scores, and would have better functional outcome.

METHOD

Study Overview and Organization

This prospective, naturalistic multicenter follow-up was designed to address the issues of treatment resistance, relapse, chronicity, and suicidality in depressive disorders within the framework of psychiatric university and district hospitals. It was part of the German Research Network on Depression and Suicidality and was funded by the German Federal Ministry of Education and Research. The study was planned to be conducted in representative inpatient groups and settings.

Twelve study centers throughout Germany participated in this follow-up, including several university hospitals (Berlin: Campus Charité Mitte and Campus Charité Benjamin Franklin; Düsseldorf: Heinrich Heine University; Halle: Martin Luther University; Heidelberg: University of Heidelberg; and Munich: Max Planck Institute of Psychiatry and Ludwig Maximilian University) and district hospitals (Gabersee/Bavaria: Inn-Salzach Hospital; Haar/Bavaria: Isar-Amper Hospital; and Berlin: Auguste Viktoria Hospital, St Joseph Hospital, and St Hedwig Hospital). The study was conducted from March 2000 to September 2005.

Sample and Data Collection

To allow inclusion of clinically representative populations, only a minimum of inclusion and exclusion criteria was required. Inclusion criteria were as follows: (1) age between 18 and 65 years; (2) signed written informed consent; and (3) hospitalization and fulfillment of *International Classification of Diseases, Tenth Revision (ICD-10)* diagnostic criteria for any major depressive episode (*ICD-10* codes: F31.3x–5x, F32, F33, F34, F38) or for a depressive disorder not otherwise specified (*ICD-10* code: F39) as primary diagnosis. Exclusion criteria were as follows: (1) organic cause of depression, (2) insufficient knowledge of German language, and (3) distance from place of residence to the study center of more than 100 km.

Moreover, the diagnosis of a depressive spectrum disorder according to *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (*DSM-IV*) was confirmed at baseline using the Structured Clinical Interview for *DSM-IV* Axis I Disorders (SCID-I).[7] The Structured Clinical Interview for *DSM-IV* Axis II disorders (SCID-II) was applied to carefully assess comorbid Axis II personality disorders. Among others, the biweekly ratings consisted of the clinician-rated psychopathologic assessments with the 21-item Hamilton Depression Rating Scale (HDRS-21).[8] Assessments at baseline and at discharge additionally consisted of the comprehensive collection of sociodemographic and clinical variables using the systematic basic assessment scale of clinical and sociodemographic variables in psychiatry (Basic Documentation [BADO]).[9] These methods were described in detail in a study protocol approved by the respective local ethics review committee. All patients signed written informed consent statements after detailed description of the study procedure. Patients were included in the analysis if at least 2 assessments were available.

In all, a total of 1,079 patients were enrolled, resulting in a final data set of 1,014 patients with a minimum of 2 complete observations. Of them, 971 patients had complete BADO data.

Definition of the Efficacy Sample

In this analysis, we adopted the criteria by Zimmerman et al[3] mainly for the following 2 reasons: (1) this was the only published collection of criteria based on a systematic review and (2) to make our analysis more easily comparable to results from others (see Discussion).

The criteria used as exclusion criteria were as follows: (1) history of mania, (2) psychotic symptoms, (3) severe suicidality, (4) alcohol or drug abuse, (5) mild depression, (6) comorbid dysthymia, (7) index episode less than 4 weeks or more than 2 years, (8) comorbid borderline personality disorder, or (9) comorbid *ICD-10* disorders other than borderline or substance abuse. See Table 1 for an exact definition of each item. Patients meeting none of the exclusion criteria were assigned to the efficacy sample, whereas patients meeting at least 1 of the 9 exclusion criteria were assigned to the nonefficacy sample.

For testing the differential impact of exclusion criteria, 2 additional groups were defined. Group 1 included patients fulfilling exclusion criteria aimed at excluding patients because of an assumed high placebo response rate (outcome-improving criteria: mild depression, index episode < 4 weeks); group 2 included patients fulfilling exclusion criteria aimed at excluding patients due to assumed low drug response rates (outcome-worsening criteria: bipolar illness, severe suicidality, psychotic features, comorbid alcohol/drug use, comorbid

© Copyright 2010 Physicians Postgraduate Press, Inc.

Table 1. Percentage of Patients Excluded Due to the Criteria of Zimmerman et al[3] (N = 971)[a]

Exclusion Criterion	Scale	n (%)
History of *DSM-IV* manic or hypomanic episodes	*ICD-10* (code F31.x)	64 (6.59)
Experiencing psychotic features during the current episode	*ICD-10* (code F33.3 or F32.3)	78 (8.03)
Significant risk of suicide	HDRS-17 (item 3 = 4)	116 (11.95)
Diagnosis of alcohol or drug abuse	*ICD-10* (code F10.x)	81 (8.34)
Mild depression	HDRS-17 (score < 20)	262 (26.98)
Underlying dysthymic (or cyclothymic) disorder	*ICD-10* (code F34.0 or F34.1)	54 (5.56)
Illness duration of less than 4 weeks	BADO	56 (5.77)
Illness duration of more than 2 years	BADO	70 (7.21)
Comorbid nondepressive, non–substance use Axis I disorders	*ICD-10* (codes F00–F09, F10–19, F20–F29, F40–F48, F50–F59, F70–F79, F80–F89, F90–F98, F99)	138 (14.21)
Borderline personality disorder	*ICD-10* (code F60.3)	18 (1.85)

[a]Numbers indicate absolute numbers of patients fulfilling each exclusion criterion; therefore, multiple selections are possible.
Abbreviations: BADO = Basic Documentation, HDRS = Hamilton Depression Rating Scale.

dysthymia, other Axis I comorbid conditions, episode duration > 2 years, presence of borderline personality disorder).

Statistical Analysis

Discharged patients (patients with at least 1 second HDRS-21 assessment) and dropouts were included in the analysis on an intention-to-treat basis, using the last-observation-carried-forward method for the mean course graphs. Missing follow-up HDRS-21 data between baseline and final visit were imputed with the last and next method.[10] Side effects as rated with the Udvalg for Kliniske Undersøgelser Side Effect Rating Scale (UKU)[11] were described for adverse events only when classified by the clinician as possibly or probably related to medication. For this analysis, an HDRS-17 score was extracted from the HDRS-21 score to achieve better comparability with phase 3 studies.

Response was defined as ≥ 50% HDRS-17 score reduction of the baseline score at final visit. Remission was defined as an HDRS-17 score of 7 or less at final visit. We analyzed both outcome events at the final visit (at discharge) for 2 reasons. First, we implied that clinicians would judge the mental state at that timepoint as stable, in order to allow discharge, and, second, to allow better comparability with results from randomized controlled trials.

Apart from the usual descriptive statistics, Fisher exact test, χ^2 tests, *t* tests, and Wilcoxon tests were applied as appropriate. Kaplan-Meier curves and log rank tests were used to compare time to response and time to remission in both groups.

Treatment

Patients were treated at the discretion of the psychiatrist in charge according to the rules of "good clinical practice"[12] under consideration of the international clinical guidelines for the treatment of depression (American Psychiatric Association, World Federation of Societies of Biological Psychiatry).[13–15] In addition, the medication class, the active medication compounds, the dosage, and the treatment duration were recorded. Furthermore, the duration and type of other biologic treatments like electroconvulsive therapy, sleep deprivation, transcranial magnetic stimulation, and psychotherapy were carefully recorded.

Table 2. Sociodemographic Baseline Variables (N = 971)

Variable	Efficacy Sample, n = 301 (31.0%)[a]	Nonefficacy Sample, n = 670 (69.0%)[a]	Test	*P* Value
Sex			Fisher exact	.32
Male	106 (35.22)	259 (38.66)		
Female	195 (64.78)	411 (61.34)		
Age group				
18–30 y	38 (12.62)	84 (12.54)		
31–50 y	135 (44.85)	350 (52.23)		
> 51 y	128 (42.52)	236 (35.22)		
Age, mean (SD), y	46.30 (12.39)	44.51 (11.62)	*t* test	.03*
Marital status			χ^2	.94
Married	118 (38.06)	275 (41.04)		
Never married	91 (29.35)	182 (27.16)		
Widowed	7 (2.26)	20 (2.99)		
Divorced	31 (10.00)	80 (11.94)		
Not married, living together	22 (7.10)	46 (6.87)		
Married, living separately	23 (7.42)	53 (7.91)		
Other	2 (0.65)	4 (0.60)		
Unknown	7 (2.26)	10 (1.49)		
Employment status (BADO)			χ^2	.32
Employed	174 (56.13)	426 (63.58)		
Unemployed	45 (14.52)	90 (13.43)		
Retired	57 (18.39)	94 (14.03)		
Other	10 (3.23)	27 (4.03)		
Unknown	15 (4.84)	33 (4.93)		
Setting			Fisher exact	.02*
University hospital	232 (77.08)	466 (69.55)		
District hospital	69 (22.92)	204 (30.45)		

[a]Values are expressed as n (%) unless noted otherwise.
*Denotes significance ($P < .05$).
Abbreviation: BADO = Basic Documentation.

Table 3. Clinical Baseline Variables (N = 971)

Variable	Efficacy Sample (n = 301)[a]	Nonefficacy Sample (n = 670)[a]
Age at onset, mean (SD), y[b]	39.03 (13.10)	37.25 (12.56)
Family history[c]		
Affective disorder	13 (4.32)	29 (4.33)
Schizophrenia	97 (32.23)	216 (32.24)
Alcohol/drug abuse or dependence	13 (4.32)	38 (5.67)
Suicide	29 (9.63)	68 (10.15)
Other	43 (14.29)	105 (15.67)

[a]Values are expressed as n (%) unless noted otherwise. [b]*t* test; $P = .06$. [c]χ^2 test; $P = .95$.

RESULTS

Baseline Measures

Of all patients, 68.8% fulfilled at least 1 of the exclusion criteria. Of the total sample, 31.2% fulfilled none and were thus eligible to form the efficacy sample (Table 1). All

Table 4. Comorbid *ICD-10* Axis I Psychiatric Disorders (N = 971)

	Efficacy Sample (n = 301), n (%)	Nonefficacy Sample (n = 670), n (%)
Comorbid disorder		
Mental and behavioral disorders		
Organic, including symptomatic, mental disorders	1 (0.33)	1 (0.15)
Mental and behavioral disorders due to psychoactive substance use	0 (0.00)	105 (15.67)
Schizophrenia, schizotypal, and delusional disorders	0 (0.00)	2 (0.30)
Neurotic, stress-related, and somatoform disorders	0 (0.00)	109 (16.27)
Behavioral syndromes associated with physiologic disturbances and physical factors	0 (0.00)	10 (1.49)
Disorders of adult personality and behavior	27 (8.97)	91 (13.58)
Mental retardation	0 (0.00)	0 (0.00)
Disorders of psychological development	0 (0.00)	0 (0.00)
Behavioral and emotional disorders with onset usually occurring in childhood and adolescence	0 (0.00)	3 (0.45)
Unspecified mental disorder	0 (0.00)	0 (0.00)
No. of comorbid disorders		
0	270 (89.70)	387 (57.76)
1	26 (8.64)	181 (27.01)
2	3 (1.00)	66 (9.85)
3	1 (0.33)	23 (3.43)
≥ 4	1 (0.33)	13 (1.94)

Table 5. *ICD-10* Depression Subtypes and Length of Current Episode and Illness (N = 971)

	Efficacy Sample (n = 301)[a]	Nonefficacy Sample (n = 670)[a]
Diagnosis		
Depression		
Mild (codes F32.0, F33.0)	4 (1.33)	21 (3.13)
Moderate (codes F32.1, F33.1)	77 (25.58)	203 (30.30)
Severe (codes F32.2, F32.3, F33.2, F33.3)	97 (32.23)	137 (20.45)
Recurrent depression (codes F33.0–F33.9)	167 (55.48)	338 (50.45)
Psychotic depression (codes F33.3, F32.3, F31.5)	0 (0.00)	84 (12.54)
Bipolar depression (code F31)	0 (0.00)	64 (9.56)
Bipolar	9 (2.99)	56 (8.36)
Length of current episode		
Less than 1 mo	0 (0.00)	133 (19.85)
Less than 3 mo	113 (37.54)	168 (25.07)
Less than 6 mo	94 (31.23)	130 (19.40)
Less than 1 y	61 (20.27)	92 (13.73)
Less than 2 y	27 (8.97)	53 (7.91)
Less than 5 y	0 (0.00)	31 (4.63)
Less than 10 y	0 (0.00)	15 (2.24)
More than 10 y	0 (0.00)	24 (3.58)
Unknown	2 (0.66)	11 (1.64)
Length of illness, mean (SD), y[b]	7.28 (9.62)	7.10 (8.93)
Unknown	31 (10.30)	63 (9.40)
Number of hospitalizations (≥ 2)[c]	141 (46.84)	343 (51.19)

[a]Values are expressed as n (%) unless noted otherwise. [b]Wilcoxon test; $P = .55$. [c]Fisher exact test; $P = .21$.

Table 6. Baseline Severity and Outcome Variables (N = 971)

	Efficacy Sample (n = 301)[a]	Nonefficacy Sample (n = 670)[a]	Test	*P* Value
Symptom severity				
Baseline				
HDRS-17 score	27.48 (4.79)	23.67 (7.30)		
MADRS score	32.40 (5.70)	28.71 (8.20)		
Unknown, n (%)	32 (10.63)	61 (9.10)		
Final visit				
HDRS-17 score	10.54 (7.89)	9.64 (7.37)	Wilcoxon	.13
MADRS score	12.34 (9.94)	11.57 (9.40)	Wilcoxon	.17
UKU score				
Baseline	18.57 (7.28)	17.75 (8.05)	Wilcoxon	.16
Final visit	8.5 (5.61)	8.67 (6.29)	Wilcoxon	.85
Length of stay in hospital, d	53.55 (37.40)	55.59 (51.95)	*t* test	.49
GAF score				
Baseline	47.87 (10.01)	47.85 (12.09)	Wilcoxon	.68
Final visit	71.37 (10.64)	69.34 (11.60)	Wilcoxon	.03*
Quality of life (SF-36 score)				
Baseline	60.88 (10.69)	62.11 (9.68)	*t* test	.14
Final visit	62.16 (8.81)	63.03 (9.28)	*t* test	.30

[a]Values are expressed as mean (SD) unless noted otherwise.
*Denotes significance ($P < .05$).
Abbreviations: GAF = Global Assessment of Functioning scale, HDRS-17 = 17-item Hamilton Depression Rating Scale, MADRS = Montgomery-Åsberg Depression Rating Scale, SF-36 = Medical Outcomes Study 36-Item Short-Form Health Survey, UKU = Udvalg for Kliniske Undersøgelser Side Effect Rating Scale.

baseline variables were systematically compared. Additionally, criteria that were not part of the exclusion criteria were tested for significant differences (Tables 2–6).

Most of the patients were excluded due to low illness severity, followed in declining order by comorbid *ICD-10* diagnoses, acute and severe suicidality, diagnosis of alcohol or drug abuse, presence of psychotic symptoms, and insufficient length of index episode (Table 1).

No significant differences between the efficacy and nonefficacy samples were found for sex, employment status, marital status (Table 2), family history (Table 3), length of illness, or number of prior hospitalizations (Table 5).

With respect to age, patients within the efficacy sample were significantly older (mean age of 46.3 years vs 44.5 years; Table 2). In addition, patients included in the efficacy sample were hospitalized to a significantly higher proportion in a university setting, and there was also a trend toward a higher age at onset ($P = .06$) in the efficacy sample (Tables 2 and 3).

Outcome Variables

Response and remission rates. Numerically, there were more responders among the efficacy sample (73%) as compared to the nonefficacy sample (69%), although the difference did not reach significance ($P = .11$).

Concerning remission rates, there was also no significant difference between the efficacy (51%) and nonefficacy samples (53%; $P = .68$; Table 7).

HDRS-17 mean scores. Both course graphs showed a similar course, with the most pronounced decline after week 2. As an HDRS-17 score < 20 is part of the exclusion criteria,

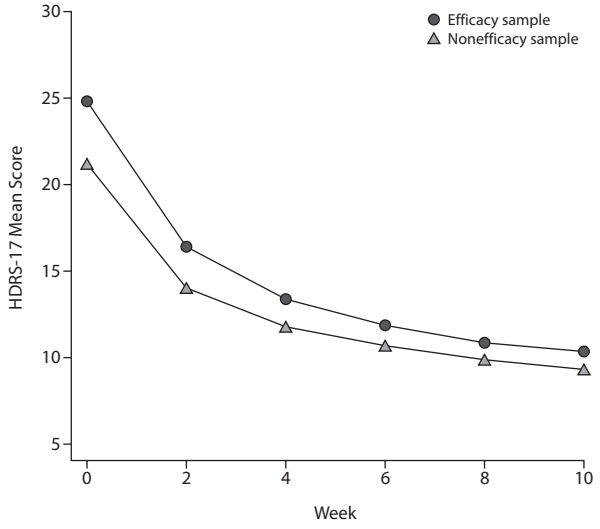
© Copyright 2010 Physicians Postgraduate Press, Inc.

Table 7. Response and Remission Rates and Percentage of HDRS-17 Change in Score From Admission to Discharge in Efficacy Versus Nonefficacy Population (N = 971)

	Efficacy Sample (n = 301)	Nonefficacy Sample (n = 670)
Remission[a]		
Nonremitter, n	143	311
Remitter, n	154	358
Rate, %	51	53
Response[b]		
Nonresponder, n	77	208
Responder, n	220	461
Rate, %	73	69
Total change in HDRS-17 score, %[c]	−60.65	−55.52

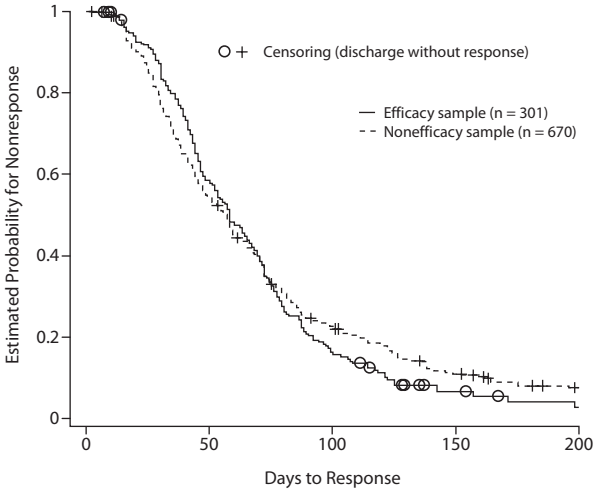
[a]Fisher exact test; $P = .68$. [b]Fisher exact test; $P = .11$. [c]$P = .07$.
Abbreviation: HDRS-17 = 17-item Hamilton Depression Rating Scale.

Figure 1. HDRS-17 Mean Scores for Efficacy and Nonefficacy Samples[a]



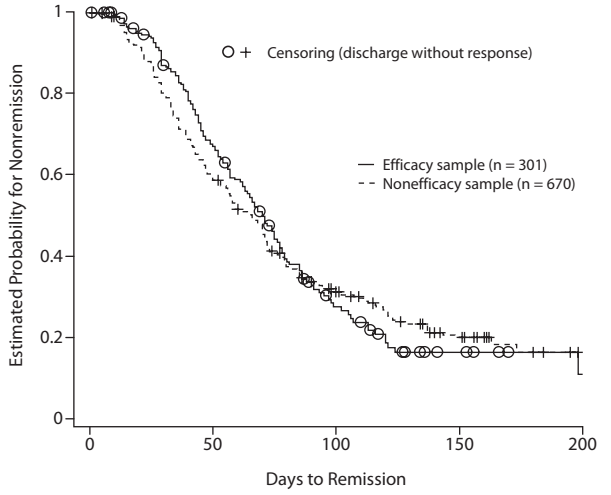
[a]HDRS-17 total score decline from admission to discharge was 60.65% for the efficacy sample and 55.52% for the nonefficacy sample ($P = .07$).
Abbreviation: HDRS-17 = 17-item Hamilton Depression Rating Scale.

Figure 2. Survival Curve for Event Response (≥50% reduction of the initial HDRS-17 score) in Efficacy Versus Nonefficacy Sample



Abbreviation: HDRS-17 = 17-item Hamilton Depression Rating Scale.

Figure 3. Survival Curve for Event Remission (HDRS-17 score ≤ 7 at endpoint) in Efficacy Versus Nonefficacy Sample



Abbreviation: HDRS-17 = 17-item Hamilton Depression Rating Scale.

the line of the nonefficacy sample runs below the efficacy sample. There was a trend for a higher absolute percentage decline in HDRS-17 scores from admission to discharge with an HDRS-17 score reduction of 60.65% for the efficacy sample and a reduction of 55.52% for the nonefficacy sample, which did not reach statistical significance ($P = .07$) (Figure 1, Table 7).

There was also no significant difference in mean HDRS-17 total scores at endpoint ($P = .13$).

Kaplan-Meier survival curves. No significant difference elapsed either for time to remission ($P = .16$) or for time to response ($P = .39$) between groups, tested with the log rank test. The survival curves are displayed in Figures 2 and 3.

Other Variables

The length of hospital stay and outcome variables, such as final HDRS-17 score, final Montgomery-Asberg Depression Rating Scale[16] score, and side effects as measured with the UKU (Table 6), did not differ significantly. Only the final Global Assessment of Functioning scale[17] score was significantly higher at discharge within the efficacy sample (Table 6).

Outcome-Worsening Versus Outcome-Improving Criteria

We additionally tested the hypothesis of whether outcome-improving and outcome-worsening criteria might outweigh each other with respect to remission. Therefore, remission rates for patients excluded due to outcome-improving and outcome-worsening criteria were calculated separately. As expected, patients excluded due to outcome-worsening criteria (N_{worse}) remitted at a lower rate (42%) as compared to

Figure 4. Weighted Sum of Patients Fulfilling Outcome-Improving Criteria Plus Patients Fulfilling Outcome-Worsening Criteria Plus Patients Fulfilling at Least 1 of Both

$$\left(\begin{array}{llllllll} \text{Rate}_{\text{nonefficacy}} & = \text{Rate}_{\text{worse}} \times \frac{N_{\text{worse}}}{N_{\text{nonefficacy}}} & + \text{Rate}_{\text{improve}} \times \frac{N_{\text{improve}}}{N_{\text{nonefficacy}}} & + \text{Rate}_{\text{worse + improve}} \times \frac{N_{\text{worse + improve}}}{N_{\text{nonefficacy}}} \\ 0.51 & = 0.42 \times \frac{301}{646} & + 0.62 \times \frac{195}{646} & + 0.53 \times \frac{150}{646} \end{array}\right)$$

patients excluded due to outcome-improving criteria (62%) ($N_{improve}$). The remaining group of patients with at least 1 improving plus 1 worsening criterion was exactly in between, with 53% ($N_{worse + improve}$). The weighted sum of these 3 percentages results in the final remission rate of 51% for the nonefficacy sample, which was not significantly different from the 52% remission rate of the efficacy sample ($P = .14$) (Figure 4).

DISCUSSION

About one-third of all patients could be classified as part of the efficacy sample. So far, there are 3 publications that have used the criteria of Zimmerman et al[3] to define an efficacy sample and that have looked into percentages of excluded or included patients. Our rate of 31.2% is clearly higher than the 9% found by Zetin and Hoepner.[5] Zimmerman et al[3] found 14% of their outpatients to be eligible for an efficacy sample.

However, Blanco and coworkers[18] applied the same criteria as in this study retrospectively to the large cohort of the National Epidemiologic Survey for Alcohol and Related Conditions (N = 43,093). The cohort included a total of 3,119 subjects with major depressive episode according to *DSM-IV* criteria and a subset of treatment-seeking patients (n = 1,359). The authors found a percentage of 33% to be eligible for an efficacy sample, which is in good accordance with our results.[18] Zimmerman et al[3] did not describe their procedure in detail. Therefore, it should be kept in mind that we, like Blanco and coworkers,[18] transferred the original criteria of Zimmerman et al[3] and used different rating instruments. Zetin and Hoepner[5] used the identical criteria with the identical instruments and also explicitly asked Zimmerman for consultation in preparation of their analysis.

Apart from differences in baseline features that are in good accordance with the results of others, we found significant associations with respect to age and setting, but there were no differences in sex, marital status, or length of hospital stay.[5,19] Regarding outcome, we found numerically but not significantly more responders within the efficacy sample ($P = .11$) and a trend toward a higher percentage decline in HDRS-17 scores from admission to discharge and a significantly higher Global Assessment of Functioning scale score at discharge. All other outcome measures, including remission, time to response, time to remission, and HDRS-17 total score at discharge, revealed no significant difference. During the review process of this article, a similar analysis of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study group was published. In contrast to our analysis, the STAR*D study group found significantly more remitters (34.4% vs 24.7%) and responders (51.6% vs 39.1%) within the efficacy sample. This discrepancy might, in part, be explained by the different exclusion criteria applied. Wisniewski and coworkers[20] used exclusion criteria based on a consensus of the authors that did not include episode duration below 4 weeks but excluded only patients with episode duration longer than 24 months. This might be of importance, as the length of the current episode is one of the best evaluated outcome predictors in major depression.[20–23] In addition, in the STAR*D study, comorbid psychiatric conditions were not treated specifically, which presumably might have worsened the outcome of the nonefficacy group.

The applied exclusion criteria investigated here could, in fact, be useful to enrich patient samples with patients being more likely to respond. This might be of interest when designing efficacy trials, particularly with regard to the demonstration of significantly different drug/placebo response rates. The most important criterion in this respect might be the exclusion of patients with an HDRS-17 score of < 20. This results in a significantly higher mean HDRS-17 total score at baseline of the efficacy sample as compared to the nonefficacy sample. Higher baseline depression scores, in turn, have consistently been found to be associated with higher response rates.[23–28] This finding is well in line with the trend toward a higher percentage of decline in HDRS-17 scores from admission to discharge of the efficacy sample. However, the significance level was not met. Time to response, time to remission, and HDRS-17 scores at discharge did not differ significantly between groups, questioning the clinical meaningfulness of this difference. In addition, assuming that a representative "effectiveness sample" would consist of a combined sample including both patients from the efficacy group and patients from the nonefficacy sample, it would be even more difficult to demonstrate significant differences. Furthermore, since we checked for significance across several baseline and outcome measures (19 in total), a correction for multiple testing would be appropriate. As a consequence, all significant differences would disappear. Due to the exploratory approach of this analysis, we omitted such procedures.

However, it appears necessary to think about what reasons might underlie the equity of outcomes in both groups. First, for a better understanding, the exclusion criteria have to be differentiated further. From a theoretical point of view, they can be divided into exclusion criteria for obvious reasons, eg, due to ethical reasons (acute suicidality). Next, there are reasons that are thought to reduce response to antidepressants (bipolar illness, psychotic features, alcohol/drug use, comorbid dysthymia, other Axis I comorbid conditions, episode duration > 2 years, presence of borderline personality disorder) as opposed to criteria that have been chosen to minimize placebo response rates (sufficient long

© Copyright 2010 Physicians Postgraduate Press, Inc.

Table 8. Treatment Comparison of Efficacy Versus Nonefficacy Sample (N = 971)

Treatment	Nonefficacy Sample (n = 301), %			Efficacy Sample (n = 670), %
	Outcome-Improving Criteria	Outcome-Worsening Criteria	Total Sample	
Tricyclics[a]	31.11	33.33	32.14	32.37
SSRIs[b]	37.78	37.12	38.63	36.93
Other antidepressant[c]	62.78	55.30	59.34	58.12
Antidepressant + tranquilizer[d]	47.78	57.95[e]	50.94	58.09
Antidepressant + neuroleptic[f]	36.67	49.24[e]	42.56	58.54
Antidepressant + lithium	17.78	21.21	19.49	18.67

[a]Tricyclics: nortriptyline, doxepin, imipramine, clomipramine, amitriptyline, trimipramine, desipramine, amitriptylineoxide. [b]SSRIs: citalopram, paroxetine, sertraline, fluoxetine, fluvoxamine, escitalopram. [c]Other antidepressant: venlafaxine, duloxetine, reboxetine, mirtazapine, maprotiline. [d]Tranquilizer: lorazepam, diazepam, alprazolam, oxazepam, bromazepam. [e]Significant difference between patients fulfilling only outcome-worsening criteria and patients fulfilling only outcome-improving criteria on a significance level of $P < .05$. [f]Neuroleptic: clozapine, olanzapine, promethazine, levomepromazine, perazine, risperidone, haloperidol, sertindole, amisulpride, sulpiride, melperone, flupenthixol, quetiapine, zotepine, benperidol, zuclopenthixol, thioridazine, ziprasidone.
Abbreviation: SSRIs = selective serotonin reuptake inhibitors.

illness duration > 4 weeks, sufficient severity of depression, HDRS-17 score > 20). As regards the influence on outcome, suicidality is only poorly studied. The best evidence comes from Serretti and coworkers,[29] who identified suicidality as a negative predictor for outcome in major depression. Therefore, suicidality was added accordingly to outcome-worsening criteria.

Thus, the nonefficacy sample simultaneously contains subjects meeting on the one hand criteria that reduce chances for remission and on the other hand criteria that are connected with notably higher spontaneous remission rates. It might be postulated, consequently, that the negative impact of criteria thought to reduce drug response is outweighed by the positive impact of the features used to minimize placebo response in the criteria proposed by Zimmerman et al.[3] This was demonstrated by separately calculating remission rates for patients excluded due to outcome-worsening and outcome-improving criteria. The simple addition as demonstrated in the weighted sum illustrates how the effect of outcome-worsening criteria was counterbalanced by outcome-improving criteria regarding remission, resulting in the remission rate of 51%, which was not significantly different from the rates of the efficacy group. One reason for this might be that inpatient mental health care in Germany is more easily accessible than in other countries, which is reflected by, eg, a comparably low rate of compulsory admissions (approximately 10%–15% of all psychiatric inpatient admissions) and longer mean inpatient treatment (39 days in the present survey) than in the United States and might therefore explain the high percentage of patients with a low to moderate severity of depression in this sample. Consequently, the negative impact of outcome-worsening criteria was probably more easily counterbalanced in this population by the relatively high number of mildly ill patients who would have been excluded due to outcome-improving criteria.

The unexpected parity between both groups was maybe also caused by the naturalistic design of the present study. The setting of this naturalistic follow-up could have led to directed and specific pharmacotherapeutic and psychotherapeutic interventions, enhancing response and remission in the remaining poor responsive patients of the nonefficacy sample. Psychotic symptoms, for example, might have been treated with a combination of an atypical antipsychotic and an antidepressant, patients with chronic depression might have been treated with an augmentation strategy such as lithium or electroconvulsive therapy, and patients with comorbid personality disorder might have received cognitive-behavioral psychotherapy and/or group therapy, thereby improving the outcome of the nonefficacy sample. This notion is supported by the fact that patients with outcome-worsening criteria received significantly more tranquilizers and neuroleptics than patients fulfilling outcome-improving criteria at baseline (Table 8). On the other hand, it should be considered that, without the specific treatment of outcome-worsening criteria within this naturalistic follow-up, the nonefficacy sample might have shown a significantly worse outcome as compared to the efficacy sample.

However, such interventions were usually not initiated within the first weeks after admission and probably do not account for the lack of a difference in outcome during this time period (see Figures 3 and 4). We therefore separately calculated response und remission rates within the first treatment period (defined as treatment time to the first antidepressant change [mean time = efficacy sample: 30.08 days; nonefficacy sample: 28.86 days; $P = .47$]), which revealed equal response rates between the efficacy and nonefficacy samples (efficacy sample: 51.05%; nonefficacy sample: 53.14%; $P = .59$). Surprisingly, significantly more remitters were found in this first treatment period within the nonefficacy sample (efficacy sample: 35.44%; nonefficacy sample: 43.90%; $P = .03$). The latter finding suggests that during the initial treatment phase, the outcome-improving criteria (mild depression, episode duration < 4 weeks) may outbalance the outcome-worsening criteria. Furthermore, it should also be kept in mind that even for the generally known negative predictors for drug response, which are therefore commonly used as exclusion criteria, the findings from the literature are not consistent.[4] For example, for active drinkers with primary depression, antidepressant treatment was found to be superior to placebo.[30,31] Also, bipolar depression has been found to respond equally well to treatment with antidepressants, especially in combination with a mood stabilizer.[32,33]

However, it should be underlined that there are also good reasons not to investigate bipolar together with unipolar subjects in phase 3 trials. For example, patients with bipolar depression should not be treated without mood stabilizers because of treatment-emergent manias, and there is growing

evidence that both diseases might be independent illness entities with a different underlying biologic pathophysiology. We thus also calculated remission (efficacy sample: 48.15%; nonefficacy sample: 46.44%; $P = .67$) and response rates (efficacy sample: 74.07%; nonefficacy sample: 68.76%; $P = .10$) separately for the unipolar sample (without the 64 bipolar subjects) and again found no significant differences between the efficacy and nonefficacy samples.

In summary, 2 major conclusions can be drawn from this analysis. First, data drawn for typical phase 3 studies might be more generalizable at least for a depressive inpatient population under naturalistic treatment conditions according to "good clinical practice" than previously thought.

Second, it might be possible that the negative impact of some exclusion criteria on drug response might be balanced by exclusion criteria that are commonly used to minimize placebo response and by a specific treatment of comorbid conditions.

Drug names: alprazolam (Xanax, Niravam, and others), citalopram (Celexa and others), clomipramine (Anafranil and others), clozapine (Clozaril, FazaClo, and others), desipramine (Norpramin and others), diazepam (Diastat, Valium, and others), doxepin (Zonalon, Silenor, and others), duloxetine (Cymbalta), escitalopram (Lexapro), fluoxetine (Prozac and others), fluvoxamine (Luvox and others), haloperidol (Haldol and others), imipramine (Tofranil and others), lithium (Lithobid and others), lorazepam (Ativan and others), mirtazapine (Remeron and others), nortriptyline (Pamelor, Aventyl, and others), olanzapine (Zyprexa), paroxetine (Paxil, Pexeva, and others), promethazine (Promethegan and others), quetiapine (Seroquel), risperidone (Risperdal and others), sertraline (Zoloft and others), trimipramine (Surmontil and others), venlafaxine (Effexor and others), ziprasidone (Geodon).

Author affiliations: Department of Psychiatry and Psychotherapy, Ludwig Maximilian University Munich, Munich (Drs Seemüller, Möller, Schennach-Wolff, Henkel, and Riedel and Mr Obermeier); Department of Psychiatry and Psychotherapy, Campus Charité Mitte, Berlin (Dr Adli); Department of Psychiatry and Psychotherapy, Carl Gustav Carus University, Technical University, Dresden (Dr Bauer); Department of Psychiatry and Psychotherapy, University of Heidelberg, Heidelberg (Dr Kronmüller); Department of Psychiatry and Psychotherapy, Max Planck Institute of Psychiatry, Munich (Dr Holsboer); Department of Psychiatry and Psychotherapy, Martin Luther University Halle-Wittenberg, Halle (Dr Brieger); Department of Psychiatry and Psychotherapy, Inn-Salzach Hospital, Wasserburg (Dr Laux); Department of Psychiatry and Psychotherapy, Isar-Amper Hospital Munich East, Haar (Dr Bender); Department of Psychiatry and Psychotherapy, Campus Charité Benjamin Franklin, Berlin (Dr Heuser); Department of Psychiatry and Psychotherapy, Auguste Viktoria Hospital, Berlin (Dr Zeiler); and Department of Psychiatry and Psychotherapy, Heinrich Heine University Düsseldorf, Düsseldorf (Dr Gaebel), Germany.

Potential conflicts of interest: **Dr Möller** has received grant/research support from Lundbeck, Pfizer, Sanofi-Aventis, and Servier and is a member of the speakers/advisory boards for and has received honoraria from Servier, Pfizer, Eli Lilly, Janssen, AstraZeneca, Wyeth, Lundbeck, and Sanofi-Aventis. **Dr Bauer** is a consultant for Eli Lilly, GlaxoSmithKline, Servier, Wyeth, and Novartis; has received grant/research support from the Stanley Medical Research Institute, the National Alliance for Research on Schizophrenia and Depression, GlaxoSmithKline, Eli Lilly, and AstraZeneca; and has received speakers' honoraria from AstraZeneca, Eli Lilly, Lundbeck, GlaxoSmithKline, Pfizer, and sanofi aventis. **Dr Brieger** is a member of the speakers/advisory boards for AstraZeneca, Pfizer, Lundbeck, Bristol-Myers Squibb, and Servier. **Dr Heuser** is a member of the advisory board for Bayer Schering. **Dr Gaebel** is or has been a member of the speakers/advisory boards for AstraZeneca, Janssen-Cilag, Eli Lilly, Lundbeck, Servier, and Wyeth and has received grant/research support from Janssen-Cilag, Kendle, Eli Lilly, Lundbeck International Clinical Research, and Pfizer. **Dr Henkel** has received financial or material support from the Swiss Agency for Therapeutic Agents (Swissmedic). **Dr Riedel** has received grant/research support from AstraZeneca and Pfizer and is a member of the speakers/advisory boards of Otsuka, Janssen-Cilag, AstraZeneca, and Pfizer. **Drs Seemüller**, **Adli**, **Kronmüller**, **Holsboer**, **Laux**, **Bender**, **Zeiler**, and **Schennach-Wolff** and **Mr Obermeier** have no personal affiliations or financial relationships with any commercial interest to disclose relative to the article.

Funding/support: The study was performed within the framework of the German Research Network on Depression, which was funded by the German Federal Ministry for Education and Research (BMBF) (grant 01GI0219). The BMBF had no further role in study design; collection, analysis, and interpretation of data; writing of the report; or the decision to submit the report for publication.

REFERENCES

1. Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA*. 1993;270(17):2096–2097.
2. Riedel M, Strassnig M, Müller N, et al. How representative of everyday clinical populations are schizophrenia patients enrolled in clinical trials? *Eur Arch Psychiatry Clin Neurosci*. 2005;255(2):143–148.
3. Zimmerman M, Mattia JI, Posternak MA. Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *Am J Psychiatry*. 2002;159(3):469–473.
4. Posternak MA, Zimmerman M, Keitner GI, et al. A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *Am J Psychiatry*. 2002;159(2):191–200.
5. Zetin M, Hoepner CT. Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *J Clin Psychopharmacol*. 2007;27(3): 295–301.
6. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23(1):56–62.
7. Wittchen HU, Wunderlich U, Gruschwitz S, et al. *Strukturiertes Klinisches Interview für DSM-IV*. Göttingen, Germany: Hogrefe; 1997.
8. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6(4):278–296.
9. Cording C, Gaebel W, Spengler A. Die neue psychiatrische Basisdokumentation. eine empfehlung der DGPPN zur qualitätssicherung im (teil-) stationären bereich. [The new psychiatric basic documentation. a recommendation by the DGPPN for quality assurance in inpatient treatment] *Spektrum Psychiatrie Nervenheilkunde*. 1995;24:3–41.
10. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. *J Clin Epidemiol*. 2003;56(10):968–976.
11. Lingjaerde O, Ahlfors UG, Bech P, et al. The UKU side effect rating scale. A new comprehensive rating scale for psychotropic drugs and a cross-sectional study of side effects in neuroleptic-treated patients. *Acta Psychiatr Scand Suppl*. 1987;76(s334):1–100.
12. Möller HJ. Evidence for beneficial effects of antidepressants on suicidality in depressive patients: a systematic review. *Eur Arch Psychiatry Clin Neurosci*. 2006;256(6):329–343.
13. Bauer M, Bschor T, Pfennig A, et al. WFSBP Task Force on Unipolar Depressive Disorders. World Federation of Societies of Biological Psychiatry (WFSBP) Guidelines for Biological Treatment of Unipolar Depressive Disorders in Primary Care. *World J Biol Psychiatry*. 2007; 8(2):67–104.
14. American Psychiatric Association. Practice Guideline for the Treatment of Patients With Major Depressive Disorder (revision). *Am J Psychiatry*. 2000;157:1–45.
15. Deutsche Gesellschaft für Psychiatrie PuN. *Praxisleitlinien in Psychiatrie und Psychotherapie, 5: Behandlungsleitlinien affektive Erkrankungen*. [*Practice Guideline for the Treatment of Affective Disorders*]. Darmstadt, Germany: Steinkopff; 2000.
16. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry*. 1979;134(4):382–389.
17. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision. Washington, DC: American Psychiatric Association; 2000.
18. Blanco C, Olfson M, Goodwin RD, et al. Generalizability of clinical trial results for major depression to community samples: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry*. 2008;69(8):1276–1280.
19. Zimmerman M, Chelminski I, Posternak MA. Generalizability of antidepressant efficacy trials: differences between depressed psychiatric outpatients who would or would not qualify for an efficacy trial. *Am J Psychiatry*. 2005;162(7):1370–1372.
20. Wisniewski SR, Rush AJ, Nierenberg AA, et al. Can phase III trial results of antidepressant medications be generalized to clinical practice? a STAR*D report. *Am J Psychiatry*. 2009;166(5):599–607.

21. Bagby RM, Ryder AG, Cristi C. Psychosocial and clinical predictors of response to pharmacotherapy for depression. *J Psychiatry Neurosci*. 2002;27(4):250–257.
22. Deykin EY, DiMascio A. Relationship of patient background characteristics to efficacy of pharmacotherapy in depression. *J Nerv Ment Dis*. 1972;155(3):209–215.
23. Paykel ES, Prusoff BA, Klerman GL, et al. Clinical response to amitriptyline among depressed women. *J Nerv Ment Dis*. 1973;156(3):149–165.
24. Rickels K, Gordon PE, Weise CC, et al. Amitriptyline and trimipramine in neurotic depressed outpatients: a collaborative study. *Am J Psychiatry*. 1970;127(2):208–218.
25. Raskin A, Crook TH. The endogenous—neurotic distinction as a predictor of response to antidepressant drugs. *Psychol Med*. 1976;6(1): 59–70.
26. Paykel ES. Depressive typologies and response to amitriptyline. *Br J Psychiatry*. 1972;120(555):147–156.
27. Burt CG, Gordon WF, Holt NF, et al. Amitriptyline in depressive states: a controlled trial. *J Ment Sci*. 1962;108:711–730.
28. 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.
29. Serretti A, Olgiati P, Liebman MN, et al. Clinical prediction of antidepressant response in mood disorders: linear multivariate vs neural network models. *Psychiatry Res*. 2007;152(2–3):223–231.
30. McGrath PJ, Nunes EV, Stewart JW, et al. Imipramine treatment of alcoholics with primary depression: a placebo-controlled clinical trial. *Arch Gen Psychiatry*. 1996;53(3):232–240.
31. Roy-Byrne PP, Pages KP, Russo JE, et al. Nefazodone treatment of major depression in alcohol-dependent patients: a double-blind, placebo-controlled trial. *J Clin Psychopharmacol*. 2000;20(2):129–136.
32. Post RM, Altshuler LL, Leverich GS, et al. Mood switch in bipolar depression: comparison of adjunctive venlafaxine, bupropion and sertraline. *Br J Psychiatry*. 2006;189(2):124–131.
33. Gijsman HJ, Geddes JR, Rendell JM, et al. Antidepressants for bipolar depression: a systematic review of randomized, controlled trials. *Am J Psychiatry*. 2004;161(9):1537–1547.