Article

Can Phase III Trial Results of Antidepressant Medications Be Generalized to Clinical Practice? A STAR*D Report

Stephen R. Wisniewski, Ph.D.

A. John Rush, M.D.

Andrew A. Nierenberg, M.D.

Bradley N. Gaynes, M.D., M.P.H.

Diane Warden, Ph.D., M.B.A.

James F. Luther, M.A.

Patrick J. McGrath, M.D.

Philip W. Lavori, Ph.D.

Michael E. Thase, M.D.

Maurizio Fava, M.D.

Madhukar H. Trivedi, M.D.

Objective: Phase III clinical trials for depression enroll participants with major depressive disorder according to stringent inclusion and exclusion criteria. These patients may not be representative of typical depressed patients seeking treatment. This analysis used data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) project—which used broad inclusion and minimal exclusion criteria—to evaluate whether phase III clinical trials recruit representative depressed outpatients.

Method: Of 2,855 participants, 22.2% met typical entry criteria for phase III clinical trials (efficacy sample) and 77.8% did not (nonefficacy sample). These groups were compared regarding baseline sociodemographic and clinical features and the characteristics and outcomes of acute-phase treatment.

Results: The efficacy sample had a shorter average duration of illness and lower rates of family history of substance abuse, prior suicide attempts, and anxious and atypical symptom features. Despite similar medication dosing and time at exit dose, the efficacy participants tolerated citalopram better. They also had higher rates of response (51.6% versus 39.1%) and remission (34.4% versus 24.7%). These differences persisted even after adjustments for baseline differences.

Conclusions: Phase III trials do not recruit representative treatment-seeking depressed patients. Broader phase III inclusion criteria would increase the generalizability of results to practice, potentially reducing placebo response and remission rates (reducing the risk of failed trials) but at the risk of some increase in adverse events.

(Am J Psychiatry 2009; 166:599–607)

The safety and efficacy of any antidepressant intended for commercial use in the United States must be judged by the Food and Drug Administration (FDA). Approval of an antidepressant requires at least two phase III clinical trials to demonstrate the drug is safe and effective, in comparison to placebo or to a recognized alternative depression treatment. Such trials typically employ stringent inclusion and exclusion criteria that may exclude a substantial portion of the broader population of depressed patients and may limit the generalizability of findings (1). If findings from these phase III trials are not generalizable to the larger population, then the phase III clinical trial design may need to be reconsidered, since such trials may not give an accurate estimate of efficacy in practice, especially if the excluded patients have, on average, poorer outcomes.

We used data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) project (2, 3) to evaluate the generalizability of the results of phase III clinical trials. STAR*D was designed with broad inclusion and minimal exclusion criteria to ensure recruitment of a representative sample of treatment-seeking depressed outpatients who receive treatment in typical clinical settings. The first treatment step (level 1) was a cohort study of the selective serotonin reuptake inhibitor (SSRI) citalopram (4).

This analysis compared participants who would meet typical inclusion and exclusion criteria used in phase III trials (efficacy sample) with those who would not (nonefficacy sample) with regard to baseline sociodemographic and clinical characteristics, treatment characteristics, and treatment outcomes (as measured by depressive symptoms and adverse events). To our knowledge, this study is the first to examine differences in treatment outcomes between efficacy and nonefficacy samples.

Method

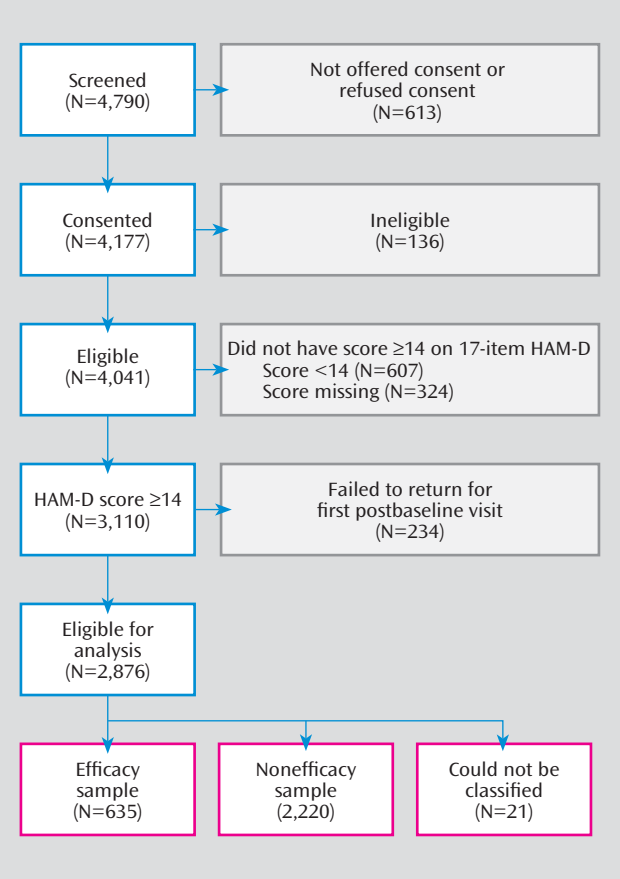
Study Overview and Organization

The rationale and design of STAR*D are detailed elsewhere (2, 3). The purpose of STAR*D was to define prospectively which of several treatments are most effective for outpatients with nonpsychotic major depressive disorder who have an unsatisfactory clinical outcome to an initial and, if necessary, subsequent treatment(s). Between July 2001 and April 2004, STAR*D enrolled participants at 18 primary care and 23 psychiatric specialty care settings across the United States.

Study Sample

The study protocol was approved and monitored by the institutional review boards at the national coordinating center (Dallas), the data coordinating center (Pittsburgh), each clinical site and regional center, and the data safety and monitoring board of the

FIGURE 1. Enrollment of Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)



National Institute of Mental Health (NIMH) (Bethesda, Md.). All risks and benefits associated with STAR*D participation were explained to the participants, who provided written informed consent before study entry.

To enhance the generalizability of the results, only self-declared outpatients seeking treatment in either primary care or specialty care settings and identified by their clinicians as having major depressive disorder that required treatment were eligible. Advertising for symptomatic volunteers was proscribed. Broadly inclusive selection criteria were used to ensure recruitment of a representative sample. Eligible participants were 18–75 years of age, met the DSM-IV criteria for single-episode or recurrent nonpsychotic major depressive disorder (established by treating clinicians and confirmed by a DSM-IV checklist), scored 14 or higher (moderate severity) on the 17-item version of the Hamilton Depression Rating Scale (HAM-D) (5, 6) (rated by the clinical research coordinators at each site), and had not been found to be treatment resistant in an adequate antidepressant trial during the current major depressive episode. Patients were excluded if they were pregnant, intending to become pregnant, or breastfeeding; had a primary psychiatric disorder requiring a different treatment approach (a bipolar, psychotic, obsessive-compulsive, or eating disorder); had substance abuse or dependence that required inpatient detoxification; were using medications excluded by the study; or had a seizure disorder or other general medical condition that contraindicated medications used in the first two protocol treatment steps. All other psychiatric and medical comorbidities were allowed.

Baseline Measures

At baseline, the clinical research coordinators collected standard sociodemographic information, self-reported psychiatric history, and information on current general medical conditions as evaluated by the Cumulative Illness Rating Scale (7, 8). In addition to administering the initial HAM-D, the clinical research coordinators assessed depressive symptom severity using the 16-item Quick Inventory of Depressive Symptomatology—Clinician-Rated, and the participant completed the Quick Inventory of Depressive Symptomatology—Self-Report (9–12). Participants also completed the Psychiatric Diagnostic Screening Questionnaire (13, 14), which was used to estimate the presence of 11 potential concurrent DSM-IV disorders.

The research outcomes assessors, blinded to treatment and not located at any site, used a telephone interview at baseline to administer the HAM-D and the 30-item clinician-rated Inventory of Depressive Symptomatology (9, 12, 15) to measure core symptoms and associated symptoms of depression. Responses to items on these measures were used to estimate the presence of atypical (16), anxious (17), and melancholic (18) symptom features.

Intervention

Citalopram was selected as a representative SSRI given the relative absence of discontinuation symptoms, demonstrated safety in elderly and medically fragile patients, once-a-day dosing, few dose-adjustment steps, anticipated generic availability, and favorable drug-drug interaction profile (19). The aim of treatment was to achieve symptom remission (defined as a score of 5 or less on the self-rated Quick Inventory of Depressive Symptomatology, which was administered at each treatment visit for the purposes of clinical decision making). The protocol required a fully adequate dose of citalopram for a sufficient time to ensure that the likelihood of reaching remission was maximized and that participants who did not reach remission were truly experiencing inadequate benefit from the medication.

The protocol aimed to provide an optimal dose of citalopram based on dosing recommendations in a treatment manual (www.star-d.org). Citalopram was to be started at 20 mg/day, then raised to 40 mg/day by week 4, and raised to the final dose of 60 mg/day by week 6. Dose adjustments were guided by symptom changes (Quick Inventory of Depressive Symptomatology completed by the clinical research coordinator), side effect burden (according to the Frequency, Intensity, and Burden of Side Effects Rating [FIBSER] [20]), and how long a participant had received a particular dose. The protocol guided physicians to make management decisions at weeks 4, 6, 9, and 12. These were critical decision points at which a decision could be made to modify the dose and/or address side effects or to move to the next treatment level. Still, appropriate flexibility was allowed to minimize side effects, maximize safety, and optimize the chances of therapeutic benefit for each participant. This included initiation of citalopram at a dose below 20 mg/day or a slower dose escalation to the optimal target dose of 60 mg/day. In this way, the study could safely include patients with concomitant general medical disorders, substance abuse or dependence, or other psychiatric disorders and those sensitive to medication side effects.

The protocol recommended treatment visits at weeks 2, 4, 6, 9, and 12 (with an optional week 14 visit if needed). After an optimal trial (as judged by dose and duration), patients with remission could enter the 12-month naturalistic follow-up, as could responders without remission, although all of those without remission were encouraged to enter the subsequent randomized trial (level 2 of STAR*D). Participants could discontinue citalopram before 12 weeks if 1) intolerable side effects required a medication change, 2) an optimal dose increase was not possible because of side effects or participant choice, or 3) significant symptoms (score of 9 or higher on the clinician-rated Quick Inventory of De-

TABLE 1. Features of Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)

Feature	Total (N=2,855)		Efficacy Sample (N=635)		Nonefficacy Sample (N=2,220)		Analysis		
	Mean	SD	Mean	SD	Mean	SD	χ^2	df	p
Age (years)	40.8	13.0	38.3	11.7	41.5	13.3	30.4	1	<0.0001
Education (years)	13.4	3.2	14.4	3.0	13.2	3.2	75.1	1	<0.0001
Monthly household income (dollars)	2,362	3,039	3,050	3,619	2,163	2,818	53.6	1	<0.0001
Age at illness onset (years)	25.3	14.4	24.9	13.1	25.5	14.7	0.0	1	0.96
Illness duration (years)	15.5	13.2	13.5	12.4	16.1	13.3	23.8	1	<0.0001
Total number of episodes	6.0	11.4	5.5	9.4	6.2	11.9	0.5	1	0.47
	N[a]	%[a]	N[a]	%[a]	N[a]	%[a]	χ^2	df	p
Female gender	1,817	63.6	411	64.7	1,406	63.3	0.4	1	0.53
Race							19.3	2	<0.0001
White	2,166	75.9	517	81.4	1,649	74.3			
Black	499	17.5	74	11.7	425	19.2			
Other	188	6.6	44	6.9	144	6.5			
Hispanic	373	13.1	54	8.5	319	14.4	15.0	1	<0.0001
Employment status							38.2	2	<0.0001
Employed	1,602	56.2	421	66.4	1,181	53.3			
Unemployed	1,088	38.2	195	30.8	893	40.3			
Retired	161	5.6	18	2.8	143	6.5			
Medical insurance							47.7	2	<0.0001
Private	1,414	51.2	378	61.4	1,036	48.3			
Public	383	13.9	41	6.7	342	15.9			
None	964	34.9	197	32.0	767	35.8			
Marital status							10.1	3	0.02
Single	815	28.6	177	27.9	638	28.8			
Married/cohabiting	1,192	41.8	281	44.3	911	41.1			
Divorced/separated	757	26.5	169	26.6	588	26.5			
Widowed	88	3.1	8	1.3	80	3.6			
Early onset (before age 18)	1,071	37.9	241	38.3	830	37.7	0.1	1	0.80
Recurrent depression	2,007	75.7	466	78.6	1,541	74.9	3.4	1	0.07
Suicide attempt	511	17.9	96	15.1	415	18.7	4.3	1	0.04
Family history									
Family history of depression	1,575	55.6	362	57.5	1,213	55.0	1.2	1	0.29
Family history of mood disorder	1,634	57.7	370	58.8	1,264	57.4	0.4	1	0.51
Family history of substance abuse	1,341	47.3	274	43.6	1,067	48.4	4.6	1	0.04
Family history of suicide	100	3.5	19	3.0	81	3.7	0.6	1	0.43
Anxious features	1,516	53.1	300	47.2	1,216	54.8	11.2	1	0.0008
Atypical features	536	18.8	91	14.4	445	20.0	10.5	1	0.002
Melancholic features	668	23.4	157	24.7	511	23.0	0.8	1	0.38
Psychiatric care	1,767	61.9	445	70.1	1,322	59.5	23.2	1	<0.0001

[a] Sums do not always equal the total number of subjects because values were missing for some subjects. Percentages are based on the available data.

pressive Symptomatology) were present after 9 weeks at the maximally tolerated dose. Participants could opt to move to the next treatment level if they had intolerable side effects or if the score on the clinician-rated Quick Inventory of Depressive Symptomatology was higher than 5 after an adequate trial in terms of dose and duration (4).

Intensive efforts to provide consistent, high-quality care are represented by the use of a treatment manual, initial didactic instruction, ongoing support and guidance by the clinical research coordinators, the use of structured evaluation of depressive symptoms and side effects at each visit, and a centralized treatment monitoring and feedback system (www.star-d.org) that provided feedback to clinical research coordinators regarding each participant's fidelity to the treatment recommendations. The clinical research coordinators could then help guide physicians in vigorous dosing when inadequate symptom reduction had occurred despite acceptable side effects (4).

Safety Assessments

In addition to side effects, serious adverse events were monitored with a multitiered approach involving the clinical research coordinators, study clinicians, interactive voice response system, safety officers, regional center directors, and NIMH data safety and monitoring board (3).

Concomitant Medications

Concomitant treatments for current general medical conditions (as part of ongoing clinical care), for associated symptoms of depression (e.g., sleep, anxiety, and agitation), and for citalopram side effects (e.g., sexual dysfunction) were permitted on the basis of clinical judgment. The protocol prohibited the use of stimulants, anticonvulsants, antipsychotics, alprazolam, nonprotocol antidepressants (except trazodone at a dose of 200 mg or less at bedtime for insomnia), and depression-targeted psychotherapies.

Primary Outcome Measures

Phase III trials traditionally assess outcomes 8 weeks after random assignment of treatment. In STAR*D, clinic visits were scheduled at 2, 4, 6, 9, and 12 weeks after enrollment. The week 9 assessment was used to approximate the time frame of the phase III trial. The primary outcome was based on the self-rated Quick Inventory of Depressive Symptomatology, which was administered at baseline and at each treatment visit. Remission was defined as a score of 5 or less (which is equivalent to a score of 7 or less on the 17-item HAM-D) (11) at week 9 or, if the last visit occurred before week 9, the last recorded score. The secondary outcome was response, which was defined as a reduction of at least 50% from the baseline score on the self-rated Quick Inventory of

TABLE 2. Treatment Features for Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)

Feature	Total (N=2,855)		Efficacy Sample (N=635)		Nonefficacy Sample (N=2,220)		Analysis		
	Mean	SD	Mean	SD	Mean	SD	χ^2	df	p
Number of treatment weeks	7.7	2.6	7.9	2.6	7.6	2.6	7.5	1	0.007
Number of postbaseline visits	3.2	1.1	3.3	1.1	3.1	1.1	10.6	1	0.002
Days to first postbaseline visit	16.2	6.5	15.8	5.9	16.3	6.7	1.6	1	0.21
Maximum citalopram dose (mg/day)	41.4	16.3	41.7	16.3	41.4	16.3	0.1	1	0.73
Exit citalopram dose (mg/day)	41.4	16.3	41.7	16.3	41.4	16.3	0.1	1	0.73
Days at exit citalopram dose	5.7	9.7	5.3	9.3	5.8	9.8	0.7	1	0.39
	N[a]	%[a]	N[a]	%[a]	N[a]	%[a]	χ^2	df	p
Categorical treatment duration									
<4 weeks	328	11.5	71	11.2	257	11.6	0.1	1	0.79
<8 weeks	1,006	35.4	201	31.8	805	36.4	4.6	1	0.04
Maximum side effect frequency							5.4	3	0.15
No side effects	479	16.9	94	14.9	385	17.5			
10%–25% of the time	828	29.2	202	32.0	626	28.4			
50%–75% of the time	891	31.4	205	32.4	686	31.2			
90%–100% of the time	636	22.4	131	20.7	505	22.9			
Maximum side effect intensity							9.5	3	0.03
No side effects	470	16.6	93	14.7	377	17.1			
Minimal to mild	823	29.0	188	29.7	635	28.8			
Moderate to marked	1,125	39.7	276	43.7	849	38.6			
Severe to intolerable	416	14.7	75	11.9	341	15.5			
Maximum side effect burden							10.2	3	0.02
No side effects	619	21.8	127	20.1	492	22.3			
Minimal to mild	1,192	42.1	298	47.2	894	40.6			
Moderate to marked	802	28.3	169	26.7	633	28.7			
Severe to intolerable	221	7.8	38	6.0	183	8.3			
Exited treatment level because of side effects	488	17.1	105	16.5	383	17.3	0.2	1	0.68
At least one serious adverse event	115	4.0	15	2.4	100	4.5	5.9	1	0.02
At least one psychiatric serious adverse event	57	2.0	6	0.9	51	2.3	4.6	1	0.04

[a] Sums do not always equal the total number of subjects because values were missing for some subjects. Percentages are based on the available data.

Depressive Symptomatology at the last assessment at or before week 9.

Defining Efficacy and Nonefficacy Samples

The whole of the STAR*D sample was consistent with a study group enrolled in an effectiveness trial. The criteria for inclusion in the efficacy sample were established a priori by consensus of several authors (A.J.R., M.H.T., M.F., A.A.N., P.J.M., B.N.G.) on the basis of their experience in designing and implementing placebo-controlled registration trials. The efficacy sample met all of the following criteria: 1) baseline HAM-D score higher than 19 (assessed by the clinical research coordinator), 2) no more than one concurrent general medical condition (defined as no more than one item of the Cumulative Illness Rating Scale with a score higher than 1), 3) the absence of obsessive-compulsive disorder (according to the Psychiatric Diagnostic Screening Questionnaire), 4) no more than one additional concurrent axis I psychiatric disorder (according to the Psychiatric Diagnostic Screening Questionnaire), and 5) a current episode lasting less than 24 months.

Those who did not meet the criteria for inclusion in the efficacy sample were included in the nonefficacy sample.

Statistical Analysis

Summary statistics are presented as means and standard deviations for continuous variables and as percentages for discrete variables. Student's t tests and Mann-Whitney U tests were used to compare continuous baseline sociodemographic and clinical features, treatment features, side effect rates, and rates of serious adverse events in the two samples. Chi-square tests were used to compare discrete characteristics in the two samples.

Logistic regression models were used to compare remission and response rates, after adjustment for the effect of baseline characteristics that were not equally distributed across the two groups. Times to first remission and first response were defined as the first observed point in the clinic visit data. Log-rank tests were used to compare the cumulative proportions of participants in each sample who reached remission or response. Additional exploratory logistic regression analyses were conducted to determine if there was a differential (moderating) effect of treatment setting (psychiatric care or primary care) on remission based on the severity of depression, as judged by the baseline score on the self-rated Quick Inventory of Depressive Symptomatology.

Statistical significance was defined as a two-sided p value of <0.05. No adjustments were made for multiple comparisons, so the results must be interpreted accordingly.

Results

STAR*D enrolled a total of 4,041 participants, 2,876 of whom made up an analyzable sample (having at least one postbaseline visit and a score of 14 or higher on the HAM-D). Of these, 2,855 could be classified into the efficacy sample (N=635, 22.2%) or the nonefficacy sample (N=2,220, 77.8%) (Figure 1). On average, participants in the efficacy sample were more likely to be younger, more educated, white, non-Hispanic, employed, married, and privately insured and to have a higher income (Table 1). The efficacy group also had a shorter average duration of illness (time from onset of the first episode of major depres-

TABLE 3. Symptom Outcomes for Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)

Outcome Measure	Efficacy Sample (N=635)		Nonefficacy Sample (N=2,220)		Unadjusted Analysis			Adjusted Analysis[b]		
	N[c]	%[c]	N[c]	%[c]	Odds Ratio	95% CI	p	Odds Ratio	95% CI	p
Remission (score ≤5)	218	34.4	546	24.7	1.60	1.32 to 1.94	<0.0001	1.33	1.07 to 1.65	0.01
Response (score reduction of ≥50%)	326	51.6	862	39.1	1.66	1.39 to 1.99	<0.0001	1.37	1.12 to 1.68	0.002
	Mean	SD	Mean	SD	Beta	95% CI	p	Beta	95% CI	p
Exit score	8.6	5.2	10.0	5.6	–1.39	–1.88 to –0.90	<0.0001	–0.68	–1.20 to –0.17	0.01
Percentage change in score	–45.4	33.2	–37.4	33.3	–8.05	–10.99 to –0.12	<0.0001	–4.28	–7.42 to –0.13	0.008

[a] Based on 16-item Quick Inventory of Depressive Symptomatology—Self-Report (9–12).
[b] Adjusted for regional center, clinical setting, age, race, Hispanic ethnicity, education, employment status, income, medical insurance, marital status, illness duration, suicide attempt, family history of substance abuse, and anxious and atypical features.
[c] Values were missing for some subjects; percentages are based on the available data.

sive disorder to study enrollment) and lower rates of prior suicide attempts, family history of substance abuse, and anxious or atypical symptom features. More participants in the efficacy sample were seen in psychiatric specialty care settings.

Participants in the efficacy sample were less likely to have side effects of severe or intolerable intensity, moderate to intolerable side effect burden, serious adverse events, and psychiatric serious adverse events (Table 2). Of note, there were no significant differences between groups in the dosing of citalopram (maximum dose or exit dose) or in the number of days at the exit dose. Participants in the efficacy sample had, on average, more weeks in treatment and more clinic visits, although these differences were not clinically meaningful.

The remission rates were 34.4% in the efficacy sample and 24.7% in the nonefficacy sample, and the number needed to treat was 10. The response rate was also lower in the nonefficacy group (51.6% versus 39.1%). Even after adjustment for potential baseline confounding characteristics, the efficacy sample had significantly better depression symptom outcomes (Table 3). They also had a shorter time to remission (Figure 2) and time to response (Figure 3). For those who achieved response, the mean time to response was 4.6 weeks (SD=2.4) for the efficacy sample and 4.8 weeks (SD=2.5) for the nonefficacy sample. For those who achieved remission, the mean time to remission was 5.5 weeks (SD=2.5) for the efficacy sample and 5.3 weeks (SD=2.5) for the nonefficacy sample. Serious adverse events were classified by the type of event. The two most prevalent events were psychiatric hospitalizations and general medical hospitalizations. The groups differed in the rate of psychiatric hospitalizations; for the efficacy sample the percentage was 0.3% (two of 635), and for the nonefficacy sample it was 2.5% (56 of 2,220) (χ^2=12.1, df=1, p<0.001). They also differed in the rate of general medical hospitalization; for the efficacy sample the rate was 1.1% (seven of 635, and for the nonefficacy sample it was 2.7% (60 of 2,220) (χ^2=5.1, df=1, p=0.02).

Discussion

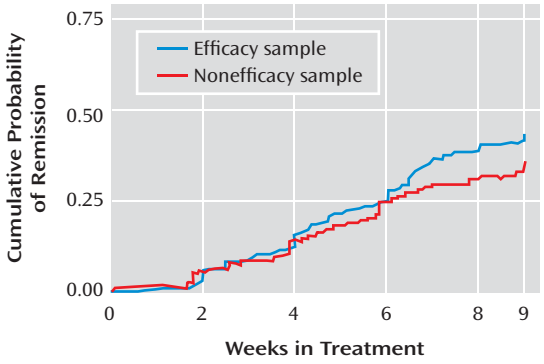
Fewer than one in four (22.2%) of the participants met the criteria for inclusion in the efficacy sample. Such a finding in a group as large and generalizable as the STAR*D sample indicates that a comparably small percentage of depressed patients treated in primary and psychiatric care settings would meet these criteria. Therefore, since the efficacy sample was based on phase III clinical trial criteria, it seems that these criteria would similarly recruit only a small percentage of typical depressed patients into phase III trials.

We found numerous differences in baseline sociodemographic and clinical characteristics between the efficacy and nonefficacy samples and a few differences regarding treatment characteristics. The latter were mostly related to side effects, although both groups received relatively equivalent doses of citalopram. Further, all measures of outcome showed significant but modest differences between the groups, with the efficacy sample having, on average, better outcomes. These differences were consistent in the direction and magnitude of effect when examined separately in primary and psychiatric care settings.

Given these between-group differences, the smaller efficacy sample is clearly not representative of the more inclusive, treatment-seeking population. By inference, a patient sample that meets the inclusion criteria for a phase III clinical trial is not representative of depressed patients seen in typical clinical practice, and phase III trial outcomes may be more optimistic than results obtained in practice.

The issue of the generalizability of randomized clinical trials is a topic that is discussed in the medical literature (21). The concern is that the results of randomized clinical trials are often poorly generalizable to a real-world clinic population, which could lead to the underuse of effective treatments or the overuse of ineffective treatments. This concern arises in both general medicine and psychiatry. Regarding general medicine, Fortin et al. (22) found that in randomized clinical trials targeting a chronic medical condition, most eligible patients had comorbid conditions that precluded eligibility. In psychiatry, Zimmerman et al.

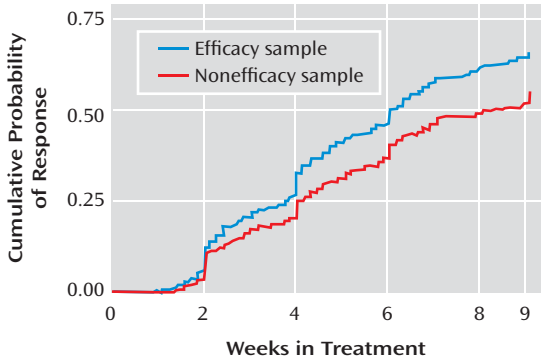
FIGURE 2. Time to Remission[a] for Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)



N at risk					
Efficacy	635	594	486	386	192
Nonefficacy	2,220	2,074	1,739	1,327	677
Total	**2,855**	**2,668**	**2,225**	**1,713**	**869**

[a] Remission was defined as a score of 5 or less on the self-rated Quick Inventory of Depressive Symptomatology.

FIGURE 3. Time to Response[a] for Patients in STAR*D Antidepressant Trial Who Met Typical Trial Criteria (Efficacy Sample) and Those Who Did Not (Nonefficacy Sample)



N at risk					
Efficacy	635	574	412	277	123
Nonefficacy	2,220	2,042	1,547	1,068	496
Total	**2,855**	**2,616**	**1,959**	**1,345**	**619**

[a] Response was defined as a reduction of at least 50% from the baseline score on the self-rated Quick Inventory of Depressive Symptomatology.

(23) found that of 315 patients with major depressive disorder who sought care, only 29 (9.2%) met typical inclusion and exclusion criteria for an efficacy trial. Kessler et al. (24) noted that most real-world patients with major depression would be excluded from randomized, controlled trials because of comorbid conditions. This existing literature, along with our study, highlights the broad public health value of large practical clinical trials and provides a model for how evidence-based psychiatry may be introduced into real-world clinics.

Thus, our results are largely consistent with previous findings that outpatients included in phase III randomized, controlled efficacy trials for major depressive disorder are different from those who would be excluded. These previous studies, however, have only examined baseline characteristics.

To our knowledge, the current study is the first to examine the differences in treatment outcome. Notably, response and remission rates were poorer and the times to response and remission were longer in patients ineligible for efficacy trials. Thus, current efficacy trials suggest a more optimistic outcome than is likely in practice, and the duration of adequate treatment suggested by data from efficacy trials may be too short.

Our findings could have significant implications for the future design of phase III trials for antidepressant treatment. Perhaps the inclusion criteria for phase III trials could be expanded to generate more generalizable information on the safety and efficacy of antidepressants, but this could come at the cost of a somewhat greater risk of adverse events. The traditional phase III approach assesses treatment efficacy in only a small subset of the population for which the treatment is intended. Therefore, a treatment defined as efficacious in the relatively small study group may be less effective and perhaps not as well tolerated in larger populations. To adequately assess whether this is so, one would have to determine if the efficacy sample has a differential treatment response in a placebo-controlled trial, which is not possible in the current study given the STAR*D design.

In addition, placebo response rates and detectable effect sizes in phase III trials might be reduced by recruiting more representative participants, including patients with concurrent comorbidity and other features (e.g., chronicity), which would increase the efficiency while improving the generalizability of phase III trials. Several studies have found differences among these populations and reduced placebo responsiveness in the presence of such features (25–27).

The present study has several limitations. First, there are no standard inclusion and exclusion criteria for a phase III clinical trial. The characteristics we used to define the efficacy sample in this study were based on an approximation of what is commonly used for a phase III clinical trial. The sensitivity of the current study's criteria was assessed by varying the assumptions to re-create the efficacy and nonefficacy samples by using other criteria and repeating the analyses. Specifically, a more stringent criterion was used that required no prior history of a suicide attempt and no current risk of suicide in addition to the earlier stated criteria for the efficacy sample. As a result of the modification, the size of the efficacy sample decreased from 635 to 522. The association of the sample with outcome remained relatively unchanged. For example, the unadjusted odds ratio for remission changed from 1.60 in the original analysis to 1.64 in the sensitivity analysis, while the adjusted odds ratio changed from 1.33 to 1.26. Thus,

the conclusions derived from the sensitivity sample are identical to those derived from the original analyses.

Another limitation is the use of self-report rather than clinical interviews to assess psychiatric and general medical comorbidities. While this limits the comparability to phase III trials, it does help in generalizing findings to standard clinic practice, where clinicians tend not to use diagnostic instruments (e.g., Structured Clinical Interview for DSM-IV) but instead use self-report. Further, the efficacy sample developed for this study is not fully representative of phase III clinical samples because, unlike most phase III trials, STAR*D proscribed the enrollment of symptomatic participants recruited by advertising. It is likely that the differences in outcomes for the two study groups would be even more pronounced for a phase III trial consisting of participants who are typical symptomatic volunteers. Also, STAR*D's broader inclusion criteria were justified by the enormous safety data available for citalopram, which was administered open-label. Most pivotal clinical trials for registration test compounds that have far less safety information and essentially no information regarding their effect on comorbid medical conditions. In the case of investigational compounds, the lack of demonstrated efficacy and the exiguous safety information would make broad inclusion less justified. The lack of placebo and double-blinding may also affect treatment outcome differences between STAR*D and pivotal clinical trials of investigational drugs.

Despite these limitations, the study also has several strengths. These include a large sample recruited from multiple geographically diverse sites in both primary and psychiatric specialty care settings. Also, measurement-based care (4, 28) with protocol-driven treatment and systematic collection of data on outcomes and adverse events was used in both samples as a method of standardizing the treatment delivery and outcomes assessment. This procedure mimics rather closely the treatment procedure used in efficacy trials.

In summary, we found numerous baseline differences between the efficacy and nonefficacy samples. In addition, patients in the efficacy group had better outcomes even after adjustment for these differences. Thus, inclusion criteria for phase III trials result in samples that are not fully representative of depressed outpatients typically treated in practice. If phase III trials enrolled more representative patients, the results would provide better estimations of the benefit to be expected in practice. One could also speculate that studying more representative groups might also reduce placebo response rates. However, the less well-documented safety profile of investigational antidepressants would have to be considered in broadening phase III trial inclusion criteria.

Received July 10, 2008; revision received Oct. 21, 2008; accepted Dec. 16, 2008 (doi: 10.1176/appi.ajp.2008.08071027). From the Epidemiology Data Center, Graduate School of Public Health, University of Pittsburgh; the Department of Psychiatry, University of Pittsburgh Medical Center; the Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia; the Departments of Clinical Sciences and Psychiatry, University of Texas Southwestern Medical Center, Dallas; the Depression Clinical and Research Program, Massachusetts General Hospital, Harvard Medical School, Boston; the Department of Psychiatry, University of North Carolina School of Medicine, Chapel Hill; the Depression Evaluation Service, New York State Psychiatric Institute and Columbia University, New York; and the Department of Health Research and Policy, Stanford University School of Medicine, Stanford, Calif. Address correspondence and reprint requests to Dr. Wisniewski, Epidemiology Data Center, Graduate School of Public Health, University of Pittsburgh, 127 Parran Hall, 130 DeSoto St., Pittsburgh, PA 15261; wisniew@edc.pitt.edu (e-mail).

Dr. Wisniewski has been a consultant for Bristol-Myers Squibb, Case Western University, Cyberonics, ImaRx Therapeutics, and Organon over the last 4 years. In the last 5 years, Dr. Rush has received research support from NIMH and the Stanley Medical Research Institute; during that time he has been an advisory board member, consultant, and/or speaker for Advanced Neuromodulation Systems, AstraZeneca, Best Practice Project Management, Bristol-Myers Squibb, Cyberonics, Forest, Gerson Lehman Group, GlaxoSmithKline, Jazz, Magellan Health Services, Merck, Neuronetics, Novartis, Ono, Organon, Otsuka, Pam Labs, Pfizer, Trancept, The Urban Institute, and Wyeth; he has equity holdings (excluding mutual funds and blinded trusts) in Pfizer; he receives royalties from Guilford Publications and Healthcare Technology Systems; and he receives a stipend as treasurer of the Society of Biological Psychiatry. Dr. Nierenberg has received research support in the past 3 years from Cederroth, Cyberonics, Forest, Medtronics, NIMH, NARSAD, Ortho-McNeil-Janssen, Pam Labs, Pfizer, Shire, and the Stanley Foundation through the Broad Institute; past research support includes Bristol-Myers Squibb, Cederroth, Eli Lilly, Forest, GlaxoSmithKline, Janssen, Lichtwer, Pfizer, and Wyeth; he has received honoraria from the Massachusetts General Hospital (MGH) Psychiatry Academy (2008 academy activities were supported through Independent Medical Education grants from AstraZeneca, Eli Lilly, and Janssen) and has participated in no other speakers bureaus for the past 3 years; past speakers bureaus have included Bristol-Myers Squibb, Cyberonics, Eli Lilly, Forest, GlaxoSmithKline, and Wyeth; he has provided advisory or consulting services in the past 3 years to Abbott, Appliance Computing, Brain Cells, Bristol-Myers Squibb, Eli Lilly, EpiQ, Forest, GlaxoSmithKline, Janssen, Jazz, Merck, Novartis, Pam Labs, Pfizer, PGx Health, Schering-Plough, Sepracor, Shire, Somerset, Takeda, and Targacept; he has received royalties from Cambridge University Press and Belvoir Publishing; he owns stock options in Appliance Computing; he holds copyrights to the Clinical Positive Affect Scale and the MGH Structured Clinical Interview for the Montgomery Asberg Depression Scale exclusively licensed to the MGH Clinical Trials Network and Institute (CTNI). Dr. Gaynes has received grants and research support in the last 12 months from the Agency for Healthcare Research and Quality, Bristol-Myers Squibb, M-3 Information, NIMH, and Novartis; he has served as an advisor for Bristol-Myers Squibb. Dr. Warden currently owns stock in Pfizer and has owned stock in Bristol-Myers Squibb within the last 5 years. Dr. McGrath has received research support in the last 5 years from Eli Lilly, GlaxoSmithKline, Lipha, NARSAD, the National Institute on Alcohol Abuse and Alcoholism, the New York State Department of Mental Hygiene, NIMH, Organon, and the Research Foundation for Mental Hygiene (New York State); he has been on advisory boards or consulted for GlaxoSmithKline, Novartis, Roche, Sanofi-Aventis, and Somerset. Dr. Lavori has received funding in the past 5 years from NIH and the Department of Veterans Affairs; he has received consulting fees from ARCA Discoveries, BMS, Celera Diagnostics, Corcept, Cyberonics, Fibrogen, Forest, Leif Cabrezer, Neuronetics, the Palo Alto Medical Foundation Research Institute, and Pfizer. Dr. Thase has received research funding from Eli Lilly and Sepracor; he has served in an advisory or consulting capacity to, or received speakers honoraria from, AstraZeneca, Bristol-Myers Squibb, Cephalon, Cyberonics, Eli Lilly, Forest, GlaxoSmithKline, Janssen, MedAvante, Neuronetics, Novartis, Organon,

Sanofi-Aventis, Schering-Plough, Sepracor, Shire, Supernus, and Wyeth; he has equity holdings in MedAvante; he has provided expert testimony for Jones Day (Wyeth litigation), Phillips Lytle (GlaxoSmithKline litigation), and Pepper Hamilton (Eli Lilly litigation); he has received royalty, patent, or other income from American Psychiatric Publishing, Guilford Publications, Herald House, and W.W. Norton; his spouse is Senior Medical Director for Advogent (formerly Cardinal Health). Dr. Fava has received research support from Abbott, Alkermes, Aspect Medical Systems, AstraZeneca, Bristol-Myers Squibb, Cephalon, Eli Lilly, Forest, GlaxoSmithKline, J&J Pharmaceuticals, Lichtwer, Lorex, Novartis, Organon, Pam Labs, Pfizer, Pharmavite, Roche, Sanofi-Aventis, Solvay, Synthelabo, and Wyeth; he has provided advisory or consulting services to Abbott, Amarin, Aspect Medical Systems, AstraZeneca, Auspex, Bayer, Best Practice Project Management, Biovail, BrainCells, Bristol-Myers Squibb, Cephalon, CNS Response, Compellis, Cypress, Dov, Eli Lilly, EPIX, Fabre-Kramer, Forest, GlaxoSmithKline, Grunenthal, Janssen, Jazz, J&J Pharmaceuticals, Knoll, Lorex, Lundbeck, MedAvante, Merck, Neuronetics, Novartis, Nutrition 21, Organon, Pam Labs, Pfizer, PharmaStar, Pharmavite, Precision Human Biolaboratory, Roche, Sanofi-Aventis, Sepracor, Solvay, Somaxon, Somerset, Synthelabo, Takeda, Tetragenex, Transcept, Vanda, and Wyeth; he has served on the speakers bureaus of AstraZeneca, Boehringer-Ingelheim, Bristol-Myers Squibb, Cephalon, Eli Lilly, Forest, GlaxoSmithKline, Novartis, Organon, Pfizer, PharmaStar, Primedia, Reed-Elsevier, and Wyeth; he has equity holdings with Compellis; he has applied for patents for SPCD and for a combination of azapirones and bupropion in major depressive disorder; and he receives royalties for the MGH CPFQ, DESS, and SAFER. Dr. Trivedi has received research support from the Agency for Healthcare Research and Quality, Corcept, Cyberonics, Merck, the National Alliance for Research in Schizophrenia and Depression, the National Institute on Drug Abuse, NIMH, Novartis, Pharmacia & Upjohn, Predix, Solvay, and Targacept; and he has received consulting or speaking fees from Abbott, Abdi Brahim, Akzo (Organon), AstraZeneca, Bristol-Myers Squibb, Cephalon, Eli Lilly, Fabre-Kramer, Forest, GlaxoSmithKline, Janssen, Johnson & Johnson, Meade Johnson, Neuronetics, Parke-Davis, Pfizer, Sepracor, VantagePoint, and Wyeth. Mr. Luther reports no competing interests.

Supported by NIMH contract N01 MH-90003 to the University of Texas Southwestern Medical Center at Dallas. Medications were provided at no cost by Bristol-Myers Squibb, Forest, GlaxoSmithKline, King Pharmaceuticals, Organon, Pfizer, and Wyeth.

The content of this publication does not necessarily reflect the views or policies of, nor does mention of trade names, commercial products, or organizations imply endorsement by, the U.S. government.

The authors thank Jon Kilner, M.S., M.A., for editorial support.

ClinicalTrials.gov identifier: NCT00021528.

References

1. Rush AJ, Bose A: Escitalopram in clinical practice: results of an open-label trial in a naturalistic setting. Depress Anxiety 2005; 21:26–32
2. Fava M, Rush AJ, Trivedi MH, Nierenberg AA, Thase ME, Sackeim HA, Quitkin FM, Wisniewski S, Lavori PW, Rosenbaum JF, Kupfer DJ: Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. Psychiatr Clin North Am 2003; 26:457–494, x
3. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart JW, McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G: Sequenced Treatment Alternatives to Relieve Depression (STAR*D): rationale and design. Control Clin Trials 2004; 25:119–142
4. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, Norquist G, Howland RH, Lebowitz B, McGrath PJ, Shores-Wilson K, Biggs MM, Balasubramani GK, Fava M, STAR*D Study Team: Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 2006; 163:28–40
5. Hamilton M: A rating scale for depression. J Neurol Neurosurg Psychiatry 1960; 23:56–62
6. Hamilton M: Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967; 6:278–296
7. Linn BS, Linn MW, Gurel L: Cumulative Illness Rating Scale. J Am Geriatr Soc 1968; 16:622–626
8. Miller MD, Paradis CF, Houck PR, Mazumdar S, Stack JA, Rifai AH, Mulsant B, Reynolds CF III: Rating chronic medical illness burden in geropsychiatric practice and research: application of the Cumulative Illness Rating Scale. Psychiatry Res 1992; 41: 237–248
9. Rush AJ, Carmody TJ, Reimitz PE: The Inventory of Depressive Symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms. Int J Methods Psychiatr Res 2000; 9:45–59
10. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB: The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003; 54:573–583; correction, 54:585
11. Rush AJ, Bernstein IH, Trivedi MH, Carmody TJ, Wisniewski S, Mundt JC, Shores-Wilson K, Biggs MM, Woo A, Nierenberg AA, Fava M: 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; 59:493–501
12. Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, Crismon ML, Shores-Wilson K, Toprac MG, Dennehy EB, Witte B, Kashner TM: The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. Psychol Med 2004; 34:73–82
13. Zimmerman M, Mattia JI: A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. Arch Gen Psychiatry 2001; 58:787–794
14. Zimmerman M, Mattia JI: The Psychiatric Diagnostic Screening Questionnaire: development, reliability and validity. Compr Psychiatry 2001; 42:175–189
15. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH: The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996; 26:477–486
16. Novick JS, Stewart JW, Wisniewski SR, Cook IA, Manev R, Nierenberg AA, Rosenbaum JF, Shores-Wilson K, Balasubramani GK, Biggs MM, Zisook S, Rush AJ: Clinical and demographic features of atypical depression in outpatients with major depressive disorder: preliminary findings from STAR*D. J Clin Psychiatry 2005; 66:1002–1011
17. Fava M, Alpert JE, Carmin CN, Wisniewski SR, Trivedi MH, Biggs MM, Shores-Wilson K, Morgan D, Schwartz T, Balasubramani GK, Rush AJ: Clinical correlates and symptom patterns of anxious depression among patients with major depressive disorder in STAR*D. Psychol Med 2004; 34:1299–1308
18. Khan AY, Carrithers J, Preskorn SH, Lear R, Wisniewski SR, Rush AJ, Stegman D, Kelley C, Kreiner K, Nierenberg AA, Fava M: Clinical and demographic factors associated with DSM-IV melancholic depression. Ann Clin Psychiatry 2006; 18:91–98
19. Rush AJ, Bose A, Heydorn WE: Naturalistic study of the early psychiatric use of citalopram in the United States. Depress Anxiety 2002; 16:121–127
20. Wisniewski SR, Rush AJ, Balasubramani GK, Trivedi MH, Nierenberg AA: Self-rated global measure of the frequency, intensity, and burden of side effects. J Psychiatr Pract 2006; 12:71–79

21. Rothwell PM: External validity of randomised controlled trials: "to whom do the results of this trial apply?" Lancet 2005; 365: 82–93
22. Fortin M, Dionne J, Pinho G, Gignac J, Almirall J, Lapointe L: Randomized controlled trials: do they have external validity for patients with multiple comorbidities? Ann Fam Med 2006; 4:104–108
23. 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: 469–473
24. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, Rush AJ, Walters EE, Wang PS: The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). JAMA 2003; 289:3095–3105
25. Quitkin FM, Rabkin JG, Stewart JW, McGrath PJ, Harrison W, Ross DC, Tricamo E, Fleiss J, Markowitz J, Klein DF: Heterogeneity of clinical response during placebo treatment. Am J Psychiatry 1991; 148:193–196
26. Quitkin FM, McGrath PJ, Rabkin JG, Stewart JW, Harrison W, Ross DC, Tricamo E, Fleiss J, Markowitz J, Klein DF: Different types of placebo response in patients receiving antidepressants. Am J Psychiatry 1991; 148:197–203
27. Rapaport MH, Zisook S, Frevert T, Seymour S, Kelsoe JR, Judd LL: A comparison of descriptive variables for clinical patients and symptomatic volunteers with depressive disorders. J Clin Psychopharmacol 1996; 16:242–246
28. Trivedi MH, Rush AJ, Gaynes BN, Stewart JW, Wisniewski SR, Warden D, Ritz L, Luther JF, Stegman D, DeVeaugh-Geiss J, Howland R: Maximizing the adequacy of medication treatment in controlled trials and clinical practice: STAR*D measurement-based care. Neuropsychopharmacology 2007; 32:2479–2489