**Objective:** To determine the effectiveness of acute-phase pharmacological and psychological treatments of depression in older ambulatory patients by systematically reviewing original research relevant to this topic.

**Methods:** Searches in MEDLINE and PsycINFO and manual reviews of bibliographies located 233 articles. Of these, 40 (37 different studies) met our 8 inclusion criteria: original research, written in English or French, subjects 55 years and older, diagnosis of depression, outpatient or community setting, prospective controlled study design, acute-phase pharmacological or psychological treatment, and outcome measure of depression. Two independent reviewers assessed the methodological quality of each article using a standard form and a quality score was computed. Quantitative data on levels of depression at the end of treatment were abstracted. Results were grouped by specific treatment comparison (type of treatment and type of control group). For comparisons that used the Hamilton Depression Rating Scale, we computed mean posttreatment differences. Effect sizes were computed from the Hamilton Depression Rating Scale or an alternative scale.

**Results:** In studies that compared active drugs with placebo, the heterocyclic drugs significantly reduced the post-treatment Hamilton Depression Rating Scale score (mean difference, −5.78; 95% confidence interval, −8.31 to −3.25); other drugs had smaller effects. In studies that compared active drugs, there were no significant differences overall between different classes of drugs; selective serotonin reuptake inhibitors appeared to be as effective as heterocyclic drugs. Rational psychological treatments performed significantly better than no treatment (mean posttreatment Hamilton Depression Rating Scale difference, −7.25; 95% confidence interval, −10.10 to −4.40) but not significantly better than that for controls who received similar attention. Adjustment for the study quality score did not affect these results.

**Conclusions:** Based on comparisons with untreated controls, heterocyclic antidepressants and rational psychological therapies appear to be the most effective treatments for older ambulatory patients with mild to moderate depression. Based on drug-drug comparisons, selective serotonin reuptake inhibitors appear to be as effective as heterocyclic drugs. However, overall, the magnitude of the treatment effects is modest. Limitations in the quantity and quality of appropriate studies suggest a sober approach to treatment in this population.

*Arch Intern Med.* 1998;158:705-712

---

**RESULTS**

**IDENTIFICATION OF ELIGIBLE STUDIES**

A total of 233 articles were retrieved that met the first 2 inclusion criteria: original studies published in English or French. Of these, 184 articles were excluded for the following reasons: 44 were not controlled trials; 44 were not randomized; 44 were not written in English or French; 10 did not report on anambulatory sample; 10 did not report on both treatments; and 20 did not report on an appropriate outcome measure. The remaining 49 articles were selected for further review. Of these, 17 were excluded because they were not randomized; 11 were excluded because they did not report on an appropriate outcome measure; and 7 were excluded because they did not report on anambulatory sample. The remaining 14 articles were included in the meta-analysis.

**This article is also available on our Web site: www.ama-assn.org/internal.**
METHODS

This research was composed of the following stages: identification of potentially relevant studies, application of inclusion and exclusion criteria, systematic review of studies, computation of a quality score for each study, and quantitative meta-analysis.

IDENTIFICATION OF STUDIES

We identified potentially relevant studies using computerized and manual search strategies. The computerized search included 2 databases: MEDLINE (1981 to March 1995), and PsycINFO (1984 to March 1995). In MEDLINE, the subject headings used were “depressive disorder” or “depression” (as major concepts) combined with “aged” and “clinical trials” (exploded to pick up all types of clinical trials). In PsycINFO, we used the subject headings “major depression” and “aged” and “therapy or treatment.” The PsycINFO search covered a shorter period because the search strategy was less efficient than the MEDLINE strategy. The manual search involved review of the bibliographies of articles and books known to us as well as those from the articles and books identified in the computer search.

INCLUSION CRITERIA

Each article identified in the search was first screened by one of us (E.K.) to determine whether it met our first 2 inclusion criteria: an original study and published in English or French (we had insufficient resources to translate articles published in other languages). Each article meeting these criteria was reviewed by at least 2 of us (J.M., M.C., or E.K.) to determine whether it met the following 6 inclusion criteria.

Age

Either all subjects were 55 years or older or at least 20% of the sample was 60 years and older and the results were reported separately for this subgroup. (These cutoff points were based on those used in the studies reviewed.)

Depression

Study subjects met criteria for depression using an accepted diagnostic system or a cutoff point on a depression symptom scale. Diagnostic systems included those based on 3 versions of the Diagnostic and Statistical Manual of Mental Disorders or the Research Diagnostic Criteria.

Setting

Subjects were drawn from community, outpatient, or nursing home settings. Studies were excluded if the source of subjects was unspecified.

METHODOLOGY

We excluded 104 articles, 84 did not meet our age criteria, 42 did not meet the research setting criteria, 11 did not meet the criteria for depression, and 3 did not include an outcome measure of depression. Of the 49 articles remaining, 5 were excluded (1 was a duplicate and 4 were follow-up reports of other studies, with no new relevant data).

Finally, we excluded 3 nonacute-phase drug studies and 4 studies of management interven-
of treatment. The methodological quality of each study was assessed using the following criteria adapted from those proposed by Chalmers and colleagues:11

1. Description of selection criteria: 3, adequate; 2, fair; and 1, not adequate.
2. Description of treatment(s): 3, adequate; 2, fair; and 1, not adequate.
3. Treatment assignment: 3, random; 2, partially random; and 1, nonrandom.
4. Blinding (masking). For pharmacological treatments: 3, explicit statement that drug(s) and/or placebo were identical in appearance; 2, placebo used but no explicit statement about identical appearance; and 1, neither. For other treatments: 3, treatment groups received equal attention and measurement of outcomes was blind to treatment assignment; 2, treatment groups received equal attention but blinding of measurement of outcomes was not done or not stated; and 1, neither.
5. Description of treatment groups at the beginning of study: 3, table provided showing distribution of important prognostic variables by treatment group; 2, partial information provided; and 1, neither.
6. Description of withdrawals: the score was obtained after discussion between the 2 raters based on their independent abstraction of data on number of subjects assigned to and withdrawn from each treatment group: 3, complete agreement; 2, agreement after discussion and rereading the article; and 1, lack of agreement even after further discussion.

Criteria 1, 2, 4, and 5 were rated independently by 2 raters (J.M., M.C., or E.K.) and an average score was calculated. The ratings on criteria 3 and 6 were given following discussion. We summed the ratings of the 6 criteria into a composite quality score with a possible range from 6 to 18, a higher score indicating better quality. For the first 9 articles, we discussed the ratings and developed consensus ratings; thereafter we did not discuss the ratings (except for criteria 3 and 6). Inter-rater agreement for the quality score was assessed with the concordance correlation coefficient.12

The following were abstracted but not included in the quality score: blinding of allocation, prior sample size estimates, and multiple looks at the results because they were not reported explicitly in any of the studies; power analysis was not applicable to all studies.

The quality of pharmacological treatment was determined by summing two 3-point scales: treatment duration (3, ≥ 12 weeks; 2, 8–11 weeks; and 1, < 8 weeks) and adequacy of dosage, determined by consensus among 2 geriatric psychiatrists (3, adequate; 2, probably adequate; and 1, probably inadequate). The score for the quality of the psychological treatment was based only on the treatment duration score.

META-ANALYSIS

Because a single study might contain several comparisons between treatment and/or controls, our analysis was based on comparisons between either a treatment and control or between 2 treatments. These comparisons were grouped by type of treatment and type of control. Two of us (A.B., F.B., or J.M.) abstracted the following data, if available: sample sizes; mean depression rating scale scores at the beginning and end of treatment for each study group; SDs of the posttreatment scores; and the P value for posttreatment comparison between study groups. Discrepancies were resolved by consensus. If a study used multiple depression scales, we selected the most frequently used scale, the Hamilton Depression Rating Scale (HDRS), if available.13 If the HDRS was not used we selected an alternative scale using the following hierarchy, based on the relative frequency with which the scales were used in the studies: Beck Depression Inventory, Zung Depression Rating Scale, Geriatric Depression Scale, and any other depression rating scale. Studies that computed a percentage change on the HDRS but did not include mean HDRS scores were included in the latter.

For each comparison we computed our primary measure of effect, the mean posttreatment difference in HDRS scores. To assess potential bias because of the exclusion of studies that did not use the HDRS, we also computed a mean effect size (difference between treatment and control group posttreatment mean depression scale scores divided by the pooled SD and adjusted by a multiplicative factor proposed by Hedges and Olkin13) for each comparison, based either on the HDRS or an alternative scale. If the SDs were not available but the P value of the statistical test comparison was reported, we used it to obtain the pooled SD.13 If the P value was reported as less than or equal to P = .05, we conservatively set it to P = .05. If the result of the test was reported as not significant, we arbitrarily set P = .50. We considered the P values as 2-tailed to derive the pooled SD, unless it was clearly mentioned that a 1-tailed P value was reported.

Within each comparison group, we combined the posttreatment mean differences and effect sizes across studies using the random effects model proposed by DerSimonian and Laird,14 and 95% confidence intervals (CIs) were calculated. Within each group we computed the combined posttreatment mean differences and effect sizes first based on comparisons that used the HDRS and then we combined effect sizes based on all comparisons. In addition, within each comparison group a test of homogeneity of effects across studies was performed,13 and we considered the effects as heterogeneous when the level of the test was less than 0.10. Quality-adjusted combined effect measures were also computed using the quality score as a weighting factor.13 Since the results were similar for the quality-adjusted and the unadjusted effect measures, we present only the unadjusted measures.

If 2 or more comparisons within a group were derived from the same study we used only 1 comparison from the study for the quantitative meta-analysis, selecting the one with the largest effect size. (Comparisons excluded from the meta-analysis are shown as open circles in the figures.)

All statistical analyses were conducted using SAS statistical software (version 6.10 for Windows).17

©1998 American Medical Association. All rights reserved.
funded by pharmaceutical companies or an unknown source, and psychological treatment studies by government sources. Subjects for the drug studies were recruited mainly from outpatient clinics whereas those for the psychological treatments were more likely to be recruited from the community, nursing homes, or mixed sources. Only 4 studies recruited patients from primary care settings. Finally, the drug studies used a formal diagnosis of depression as a selection criterion more often than the psychological studies.

COMPARISONS USING HDRS

A total of 21 drug-placebo comparisons are shown in Figure 1, grouped by drug type. Among 12 comparisons of a heterocyclic drug with placebo, 9 were statistically significant: 4 of imipramine, 2 of nortriptyline, 2 of nomifensine, and 1 of doxepin. There was significant heterogeneity in the study results (P = .03); this could not be explained either by variability in the treatment quality score or the mean HDRS score at baseline. Heterocyclic drug studies as a group (n = 9 after exclusion of 1 comparison in 3 studies with 2 comparisons) had mean posttreatment differences in HDRS scores of −5.78 (95% CI, −8.31 to −3.25).

Regarding antianxiety drugs, buspirone was significantly more effective than placebo but alprazolam was not. Overall, antianxiety drugs were not significantly better than placebo and there was no significant heterogeneity (P = .22). Among other drug-placebo studies, significant benefits were found for fluoxetine, trazodone, and phenelzine (HDRS differences, −2.40, −7.50, and −0.75, respectively).

Seventeen drug-drug comparisons, mainly involving heterocyclic drugs, are shown in Figure 2. The mean differences were smaller than those in the placebo comparisons and none was statistically significant. There was significant heterogeneity among the 5 comparisons of SSRIs with heterocyclic drugs (P = .04). Twelve comparisons between psychological treatments and controls are shown in Figure 3. None of 4 comparisons of emotive treatments with untreated controls was statistically significant, and there was no significant heterogeneity in the 3 comparisons included in the meta-analysis (P = .28); no comparisons of an emotive treatment with an attention control used the HDRS. Five of 6 comparisons of rational treatments with untreated controls were statistically significant. Overall, rational treatments performed significantly better than untreated controls; the mean HDRS difference was −7.25 (95% CI, −10.10 to −4.40) and the P value for homogeneity was .48. However, neither of 2 rational treatments performed significantly better than attention controls.

There were 6 comparisons between different psychological treatments that used the HDRS (Figure 3). Neither the individual comparisons nor the summary comparison indicated significant differences between rational and emotive treatments, and there was no significant heterogeneity in the 4 comparisons included in the summary estimate (P = .80).

Table 1. Selected Characteristics of the 40 Studies* by Type of Treatment

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Drug (n=26)</th>
<th>Psychological (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary care</td>
<td>4 (15)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Medical OPD</td>
<td>1 (4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Psychiatric OPD</td>
<td>13 (50)</td>
<td>3 (22)</td>
</tr>
<tr>
<td>OPD, other or not specified</td>
<td>3 (12)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Community†</td>
<td>2 (8)</td>
<td>7 (50)</td>
</tr>
<tr>
<td>Nursing home</td>
<td>2 (8)</td>
<td>2 (14)</td>
</tr>
<tr>
<td>Mixed</td>
<td>1 (4)</td>
<td>1 (7)</td>
</tr>
<tr>
<td>Depression criteria‡</td>
<td>22 (85)</td>
<td>8 (57)</td>
</tr>
<tr>
<td>Depression rating scale</td>
<td>24 (89)</td>
<td>14 (100)</td>
</tr>
</tbody>
</table>

*Three studies that include both acute-phase drug and psychological treatments appear in both columns. OPD indicates outpatient department.
†Includes home care.
‡One or both.

QUALITY RATINGS

The mean quality score from the 40 ratings was 14.9 (SD, 2.2). In the 32 ratings conducted by 2 independent raters, the concordance correlation coefficient was 0.84 (95% CI, 1.05-0.95). Means (SDs) for each treatment group were 13.9 (2.4) for psychological treatment and 15.5 (1.9) for drug treatment. Means (SDs) were 14.1 (2.7) for studies published before 1985, 15.2 (1.5) for those published from 1985 through 1989, and 15.7 (1.7) for those published in 1990 or later.

DESCRIPTION OF STUDIES

There were 26 studies of pharmacological treatments* and 14 studies of psychological treatments,† (Table 1). The majority of the studies were performed in the United States. Studies of drugs were more likely to have been published during the 1990s, while those of psychological treatments were mostly published before 1990. Sources of funding also differed among the groups: drug studies were more often

---

†References 18, 20, 33-36, 47, 55, 57, 59, 60, 62-64.
COMPARISONS USING OTHER SCALES

There were 16 comparisons derived from 9 studies that used a depression outcome scale other than the HDRS. In comparison groups that included 2 or more treatment comparisons, at least 1 of which did not use the HDRS, 7 additional comparisons were found; none of these was statistically significant and their inclusion had little impact on the effect sizes or homogeneity statistics (Table 2). Among 9 other individual comparisons that did not use the HDRS, not included in these comparison groups, only those of Gerovital H3 (a solution containing 2% procaine hydrochloride with benzoic acid as a preservative and potassium metabisulfite as an antioxidant) vs either imipramine or placebo did not use the HDRS, not included in these comparison groups. Among 9 other individual comparisons that did not use the HDRS, not included in these comparison groups, only those of Gerovital H3 (a solution containing 2% procaine hydrochloride with benzoic acid as a preservative and potassium metabisulfite as an antioxidant) vs either imipramine or placebo showed a significant benefit of treatment.

This review demonstrates that, in comparison with placebo controls, only 2 types of drugs, heterocyclic drugs and SSRIs, are effective in older ambulatory patients in the short term. The evidence on SSRIs is less convincing than for heterocyclic drugs, but when rational therapies were compared with untreated controls, the mean difference in HDRS scores (7 points) was similar to that for heterocyclic antidepressant drugs, but when rational therapies were compared with attention controls, this difference disappeared, suggesting that attention alone may be an effective treatment for many older ambulatory patients with depression. The benefits of emotive therapies may be less than those of heterocyclic drugs or rational psychological therapies, but results of one trial of structured reminiscence therapy were positive and this approach should be replicated. Although many of these psychological approaches might be incorporated into clinical practice and could be an attractive alternative to drugs, it appears that much of the effects of psychological treatment can be attributed to the nonspecific effects of extra attention to the patient. Knowledge of this nonspecific attention effect may be used by the clinician or an adjunctive attention intervention might be implemented by nonprofessionals.


Figure 1. Mean posttreatment Hamilton Depression Rating Scale differences (solid circles) and 95% confidence intervals for drug-placebo comparisons. SSRI indicates selective serotonin reuptake inhibitor; open circles, estimates excluded from the meta-analysis; diamonds, estimates from the meta-analysis.
However, the evidence supporting drug and psychological approaches to the treatment of depression in older, ambulatory patients is far from conclusive, for the reasons outlined below.

**SELECTION OF SUBJECTS**

Differences in sample selection may account for differences in the results of the various treatments that we reviewed. For example, the drug studies were conducted mainly in outpatient settings while the psychological studies more often used community samples; drug studies used formal diagnostic selection criteria more often than psychological studies. Regardless of the setting, studies recruited primarily convenience or volunteer samples; the results may not be generalizable to patients seen in clinical practice, particularly in primary care settings.
TREATMENTS

Many of the treatments studied were short-term, approximately 1 in 4 lasted 4 weeks or less. Only 2 studies included direct comparisons of drugs with psychological treatment and none assessed multifaceted interventions, except for the management studies that were excluded from this review because of their small number and heterogeneity.

OUTCOME MEASURES

The most frequently used scale to measure depressive symptoms, the HDRS, was developed to rate symptoms in patients with serious affective disorders. This scale is heavily weighted for somatic (eg, anorexia, weight loss, or anergy) and melancholic (eg, feelings of guilt, hypochondriasis, or loss of insight) symptoms, and may not be an appropriate measure of the less severe symptoms of depression seen in ambulatory settings. Nevertheless, the effect sizes that included other depression rating scales were of a similar order of magnitude to those of the HDRS. The studies in this review rarely included either posttreatment follow-up measures or measures of physical or cognitive functioning.

CONTROLS

Non-specific effects of treatment may include the placebo effect (due to the expectation that the treatment will be effective), and the Hawthorne effect (due to knowledge that one is being studied). Because a true placebo is not possible for psychological treatments, it is often recommended that an effort be made to control for the increased attention involved in such treatments by using a control group that receives an equivalent amount of attention, but without the specific psychological component used in the experimental group. Our results indicate the importance of choice of control group; larger effects of treatment generally were found when an untreated control group was used.

Other frequent methodological problems in the studies reviewed include failure to describe blinding of randomization or to analyze the results by intention-to-treat.

Our meta-analysis is limited to those original studies published in English or French that were identified in our search strategy, and may not be representative of published studies in all languages. Since unpublished studies were excluded, publication bias is of concern. However, since publication bias might be expected to favor studies with positive results, inclusion of unpublished studies could be expected to further reduce the already modest treatment effects we observed. Our meta-analysis was further limited by the small numbers of published studies available; for some treatments only 1 study was available so that no pooling was possible. Finally, we did not assess adverse events, an important consideration in selecting among effective treatments.

In conclusion, we suggest that clinicians can expect only modest benefits from currently available treatments of depression in older ambulatory patients, taking into account their modest effectiveness and the limitations in the evidence highlighted in this meta-analysis. While awaiting more substantial evidence of effectiveness from more representative study populations, clinicians may wish to be cautious in their use of heterocyclic drugs and SSRIs in this population and consider using psychological or attention interventions, particularly for patients with milder depression.

REFERENCES
