INTERNATIONAL JOURNAL OF GERIATRIC PSYCHIATRY, VOL. 11: 1089–1095 (1996)

SITE VARIABILITY IN A MULTISITE GERIATRIC DEPRESSION TRIAL*

GARY W. SMALL
Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles (UCLA) School of Medicine; Veterans Affairs Medical Center, West Los Angeles, USA

LON S. SCHNEIDER
Department of Psychiatry and Behavioral Sciences, University of Southern California, Los Angeles, USA

S. H. HAMILTON
Lilly Research Laboratories, Indianapolis, Indiana, USA

ALEXANDER BYSTRITSKY
Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles (UCLA) School of Medicine, Los Angeles, USA

BARNETT S. MEYERS
Department of Psychiatry, Cornell Medical Center, White Plains, New York, USA

CHARLES B. NEMEROFF
Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia, USA

FLUOXETINE COLLABORATIVE STUDY GROUP

SUMMARY

We studied differences in outcome and characteristics among 29 clinical sites of a multisite, double-blind antidepressant trial for geriatric depression. Six hundred and seventy-one outpatients aged 60 years or older (mean $\pm$ SD = 67.7 $\pm$ 5.7) met DSM-III-R criteria for unipolar major depression, had baseline 17-item Hamilton Depression Rating Scale ($HAMD_{17}$) scores ≥ 16 and were randomized to fluoxetine (20 mg daily) or placebo. Effect sizes (ESs, expressed as mean differences between effects divided by the pooled standard deviation of the differences) were calculated for each site using selected outcome measures. ES ranged from 1.84 (favoring fluoxetine) to -0.91 (favoring placebo) for incidence of remitters (endpoint $HAMD_{17}$ total score of ≤ 8). A large, positive ES favoring fluoxetine for remission rates (ES ≥ 0.65) was found at only six sites, moderate ES (0.35–0.64) at eight and small ES (0–0.34) at seven; ES favored placebo (< 0) at eight of 29 sites. Private clinics showed an overall $HAMD_{17}$ ES for change scores more than twice that of university sites. These results suggest that individual practitioners may have vastly different clinical experiences in large, multisite trials for geriatric depression. Interrater reliability, subject selection, recruitment, inadequate or fixed dosing, few patients per site, brief study duration, heterogeneity of geriatric depression, financial incentive and characteristics of individual sites may contribute to response variability.

KEY WORDS—geriatric depression; clinical trial; site variability

Major advantages of multisite, randomized clinical trials include large sample size, reduced time for subject recruitment, a broadening of the representativeness and generalizability of the study sample and the ability to identify possible heterogeneity among sites. Multisite studies also minimize the possibility that idiosyncratic results at a single site are overgeneralized.

Recently, a multisite, double-blind placebo-controlled trial of fluoxetine for geriatric major depression was completed (Tollefson and Holman, 1993; Tollefson *et al.*, 1995). This study is noteworthy as the first large sample size, placebo-controlled trial in geriatric depression. Despite the large number of subjects included ($N = 671$), there

* The views expressed are those of the authors and do not necessarily represent those of the Department of Veterans Affairs.

Address for correspondence: Dr G. W. Small, UCLA Neuropsychiatric Institute, 760 Westwood Plaza, Los Angeles, CA 90024-1759, USA. Tel: 310-825-0291. Fax: 310-206-5287.

CCC 0885-6230/96/121089-07

Received 6 September 1995
Accepted 15 December 1995

was no significant difference between fluoxetine (20 mg/day) and placebo in the 17-item Hamilton Rating Scale for Depression ($HAMD_{17}$) (Hamilton, 1960) mean change scores, but a significantly higher incidence of remission (ie endpoint $HAMD_{17} \leqslant 8$) for fluoxetine compared with placebo (27.6% vs 16.7%, $p = 0.001$). The small magnitude of response to fluoxetine was surprising since fluoxetine is considered an efficacious treatment in the elderly when compared to tricyclic antidepressants in smaller-size studies (Feighner *et al.*, 1988; Klawansky, 1994). Along with nortriptyline, it is recommended for the treatment of late life depression by the Depression Guideline Panel of the Agency for Health Care Policy Research (1993).

This large-scale clinical trial provides the first opportunity to examine variability among clinical sites treating elderly depressed outpatients under one protocol. The purpose of the present study is to evaluate variability of outcomes among sites. In an attempt to explain variability, we compared treatment groups according to demographics (gender, race and age) and baseline levels of depression across sites and affiliation types (university and private).

METHODS

Subjects

Subjects were outpatients aged 60 years or older who met *Diagnostic and Statistical Manual of Mental Disorders, Third Edition—Revised* (DSM-III-R) (APA, 1987) criteria for current unipolar major depression, except illness duration must have been at least 1 month. Baseline scores of at least 16 on the $HAMD_{17}$ were required for inclusion. Reasons for exclusion were scores of $\leqslant 25$ on the Mini-Mental State Examination (Folstein *et al.*, 1975), serious suicidal risk, serious or unstable medical comorbidity, other axis I DSM-III-R major psychiatric disorder or presence of psychosis. Persons with a history of non-response to at least two different antidepressant drug classes or electroconvulsive therapy within 12 months were also excluded. Other psychoactive drugs (except limited use of chloral hydrate for insomnia) were not permitted during the 14 days before screening or the 6-week study. The institutional review boards of all participating sites approved the protocol, and all subjects completed written informed consent.

Study design

After a 1-week, single-blind, placebo lead-in, subjects were randomly and blindly assigned to fluoxetine, 20 mg daily ($N = 335$) or placebo ($N = 336$) for the next 6 weeks. To exclude placebo responders from study participation, subjects with $\geqslant 20\%$ improvement on the $HAMD_{17}$ during the 1-week single-blind period were discontinued before random assignment. Of the 24 subjects who discontinued prior to randomization, a maximum of only three patients discontinued at any one site. Fluoxetine or placebo doses could be reduced to every other day for patients with side-effects that made them no longer willing to participate. Only 6.6% of the fluoxetine group and 2.4% of the placebo group were changed to every other day dosing. Nineteen sites made such changes, with a maximum of three changes at any one site. Weekly capsule counts and daily diaries were used to assess compliance. Illness severity was measured weekly using several standardized rating scales, including the $HAMD_{17}$, an observer-rated instrument, and the Geriatric Depression Scale (GDS; Brink *et al.*, 1982), a self-rating instrument. For the observer-rated instrument (ie $HAMD_{17}$), an interrater reliability session was held before beginning the study.

Statistical analysis

The last available postbaseline measurement for a patient was defined as the endpoint measurement; change was defined as the endpoint value minus the baseline value. Remission was defined as an endpoint $HAMD_{17}$ total score of $\leqslant 8$. Patients were included in the analysis if they had a baseline and at least one postbaseline $HAMD_{17}$ total score. Of the 335 patients randomized to fluoxetine, 326 met these criteria, as did 329 of the 336 patients randomized to placebo. In the analysis, patients were clustered by site and by affiliation type.

A general linear models analysis of variance (ANOVA) was used to determine treatment-by-site interactions for the change scores. The model included terms for treatment, site and the treatment-by-site interaction. The results were obtained using PROC GLM (type III sums of squares) in version 6 of SAS (SAS, 1989). Each of the 30 sites was classified as a university-affiliated site (including Veterans Affairs Medical Centers) or an independent (ie private) clinical research organization, and similar analyses performed for the

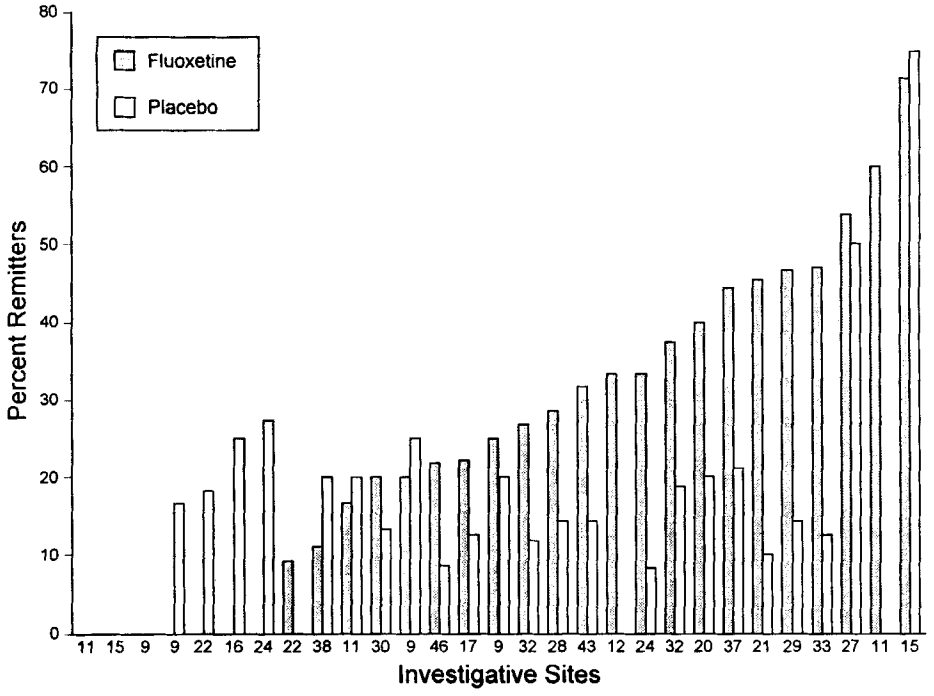

Fig. 1. Remission rates by per cent remitted (ie endpoint $HAMD_{17} \leqslant 8$) for fluoxetine and placebo treatment within each of 29 sites arranged by magnitude of fluoxetine remission rate. (Numbers below horizontal axis denote total sample size at each site.)

sites were also performed for the affiliation types. Breslow-Day tests for homogeneity of odds ratios and chi-square tests were used for the analysis of remission and other categorical data obtained from PROC FREQ in SAS. Main effects were tested at the $\alpha = 0.05$ level. Because the tests for treatment-by-site interactions and homogeneity of odds ratios were less powerful, we tested them at the $\alpha = 0.15$ level.

For descriptive purposes, treatment effect sizes (ESs) expressed as differences in mean change scores and remission rates between treatment groups divided by the pooled standard deviation of the differences were calculated for each site to provide a measure, independent of sample size, conveying magnitude of response (Hedges and Olkin, 1985; Johnson, 1989). ESs were also calculated for each affiliation type.

RESULTS

Of the 671 randomized patients, 534 (79.6%) completed the study (fluoxetine, 78.5%; placebo, 80.7%). Remission rates varied considerably across sites and ranged from 0 to 71% for fluoxetine and 0 to 75% for placebo (Fig. 1). One site had an inadequate sample size (fluoxetine, $N = 2$; placebo, $N = 1$) and was not included. For the other 29 sites, ESs ranged from -0.91 (favoring placebo) to 1.84 (favoring fluoxetine) for incidence of remitters and -0.57 to 1.36 for $HAMD_{17}$ change scores.

A large positive ES indicating clinically meaningful remission was operationalized as $\geqslant 0.65$ and other categories defined as moderate (0.35–0.64), and small (0–0.34), either in favor of fluoxetine (+) or placebo (−), in order to provide descriptive comparisons among outcome measures, as suggested elsewhere (Cohen, 1992). A large positive ES based on remission was found at only six sites, a moderate ES at eight sites and a small ES at seven (Table 1). ESs based on remission favored placebo at eight of the 29. There was a trend towards lack of homogeneity of the odds ratios across the sites (Breslow-Day tests, $p = 0.167$). For $HAMD_{17}$ change scores, the ES distribution shifted in favor of placebo (12 sites) with fewer sites favoring fluoxetine. An ANOVA, using $HAMD_{17}$ change scores, demonstrated no significant treatment-by-site interactions ($F = 0.83$, $p = 0.724$).

Table 1. Frequency of large, moderate, small and placebo-favoring effect sizes according to outcome measure for 29 clinical investigative sites. Effect sites are expressed as 'd' statistics, the difference between the fluoxetine and placebo effect divided by the pooled standard deviation. $HAMD_{17}$ remission is a final score of $\leqslant 8$. Divisions of large, moderate and small effect sizes are adapted from Cohen (1992).

	Effect size					
	Favoring fluoxetine			Favoring placebo		
Outcome measures	Large ($\geqslant 0.65$)	Moderate (0.64–0.35)	Small (0.34–0)	Small (0– –0.34)	Moderate (–0.35– –0.64)	Large (–0.65)
$HAMD_{17}$ remission, *N* (%)	6 (21)	8 (28)	7 (24)	4 (14)	1 (3)	3 (10)
$HAMD_{17}$ change, *N* (%)	4 (14)	8 (28)	5 (17)	10 (34)	2 (7)	0 (0)
Geriatric Depression Scale change, *N* (%)	4 (14)	18 (28)	4 (14)	8 (28)	3 (10)	2 (7)

ES calculated from GDS change scores ranged from –0.86 to 1.63 (Table 1). For several sites, discrepancies between $HAMD_{17}$ and self-rated GDS effect sizes were extreme and in opposite directions. Although there was extreme variability in change scores, there was no significant treatment-by-site interaction on the GDS (ANOVA, $F = 1.10$, $p = 0.335$).

Age and race did not vary across ES classifications but differences were observed for gender. The percentage of women with large ES classification favoring fluoxetine was much greater for the fluoxetine group than for the placebo group (51.9% vs 28.6%). By contrast, the percentage of women with large ES classification favoring placebo was much greater for the placebo group than for the fluoxetine group (73.7% vs 40.9%).

A greater number of patients were enrolled at the 20 university sites compared with the 10 private sites (399 vs 256) but fewer per site (20 vs 26). Baseline $HAMD_{17}$ scores were inconsistent across treatment groups at the two types of sites, and the treatment-by-affiliation interaction (fluoxetine—private vs university: 22.7 ± 4.2 vs 21.8 ± 3.6; placebo—private vs university: 20.0 ± 3.7 vs 22.2 ± 3.9) was significant ($F = 3.09$, $p = 0.079$). No significant differences between treatments in patient age, gender or race were seen at either type of site and heterogeneity was not detected across affiliation types for these demographic variables. Heterogeneity was seen between the affiliation types in the analysis of change in $HAMD_{17}$ ($p = 0.112$). Remission rates were also significantly different: fluoxetine—private vs university: 32.6% vs 24.4% ($p = 0.008$); placebo—private vs university: 18.1% vs 15.8% ($p = 0.033$). Another way of illustrating this is by observing that the pooled ES for change scores from the private sites was more than twice as great as that for the university sites (0.38 vs 0.14).

DISCUSSION

We found variability in magnitude of effect sizes among clinical sites in this multisite trial. Depending on the outcome measure used (change vs remission), four to six investigative sites showed large ESs favoring fluoxetine and eight to 13 sites favored placebo but mostly to a small degree.

The clinical heterogeneity of geriatric depression may have contributed to variability among sites. For example, previous studies indicate differences in subgroups of elderly depressed patients defined according to age at onset of first depressive episode. Late onset patients have a lower frequency of family history for depression, but higher frequencies of deep white matter hyperintensities on magnetic resonance imaging scans, cerebral atrophy and medical comorbidity (Alexopoulos *et al.*, 1993).

Previous reports of fluoxetine's efficacy for geriatric depression (Feighner *et al.*, 1988; Altamura *et al.*, 1989; Falk *et al.*, 1989; La Pia *et al.*, 1992) have demonstrated variable effects, but none of these studies included a placebo comparison group. Patients entering such trials have 100% chance of receiving active drug, and thus both patients' and study physicians' expectations of efficacy may be enhanced compared to expectations during a placebo-controlled trial. Most importantly, however, in these trials considerable symptomatology remained in patients treated with either type of medication.

One reason for the variable fluoxetine effects may have been that the 20 mg dose was inadequate to treat depression in many of the elderly patients. Clinical experience indicates that elderly depressed patients tolerate and respond to higher daily fluoxetine doses. The brief study duration of only 6 weeks also could have contributed to the variability since elderly patients may require up

to 12 weeks for response to antidepressant treatment (Georgotas and McCue, 1989).

Lack of interrater reliability may have contributed to site variability: a single interrater training session may be inadequate for studies of geriatric depression. Concomitant medical illness, emphasis on somatic and cognitive complaints and denial of affective symptoms in old age depression can interfere with accurate diagnosis and symptom severity ratings (Small, 1991; Salzman, 1994). Although the discrepancies between clinician ratings and patient self-ratings further suggested interrater reliability problems at several sites, such problems were not specifically addressed in this study.

Inclusion of subjects with varying severities of depression may also have contributed to outcome variability. Placebo-controlled trials often result in populations of convenience, not necessarily representative of major depression in the community. Outpatient status and the relatively low baseline HAMD scores (inclusion of HAMD scores between 16 and 20) are factors that support the view that, as a group, these depressed patients were not severely depressed.

Comparisons of university and private sites demonstrated greater fluoxetine vs placebo effects for the private sites. Despite the larger sample size for university sites, remission rate differences between treatment groups were more significant for private sites. Varying recruitment methods may account for differences between private and university settings. Private sites might have greater use of advertisements and more aggressive recruitment methods than university sites. Although only 33% of the clinical sites were private affiliates, these private sites recruited 39% of the total multisite sample. Many university sites are tertiary care referral centers where difficult, treatment-resistant patients are referred. Recruitment of such patients might lead to lower ESs. Because geriatric depression is heterogeneous, such variability in recruitment methods could lead to different patient subgroups that have varying responses at different sites.

Varying investigator incentives could also have influenced recruitment methods and subject selection, thus contributing to such ES differences. Some investigators participate in clinical trials in order to recruit subjects for other non-pharmacological, phenomenological or descriptive studies, such as investigations of biological markers for depression. In addition, providing treatment opportunities to patients may diminish attrition rates for longitudinal studies that may offer few other incentives for participation. Funding from industry-sponsored trials may help academic investigators maintain research staffing, particularly during periods when federal funding is limited. Academic investigators may also participate in order to acquire data for publication or to earn a profit for a clinical research organization. A private clinic that is not invested in research might be more inclined to obtain positive patient outcomes. Finally, investigators are rewarded financially according to number of subjects completing a visit or trial. Financial gain may be a major incentive for some investigators. Of course, a variety of other factors not explored in this study may have contributed to variability, including socioeconomic status, urban versus rural residence, personality factors or history of substance abuse.

Multisite clinical trials allow for considerable variability among individual clinical sites. Because the methods used in this trial are typical of the current standard for industry-sponsored trials, these results support the need for greater attention to methodological issues of multisite clinical trials. For example, only one clinician training session at the beginning of a trial may be inadequate to ensure interrater reliability. More frequent training during the course of a trial might also minimize variability in subject selection.

In this report, we examined variability among individual sites rather than individual patients, yet differences in outcomes were marked. A health services perspective might consider each of these sites as an individual clinical practice. It can then be appreciated how different practitioners may have vastly different clinical experiences using similar treatment interventions. We thus recommend future health services research to clarify the contribution of site variability to treatment effects.

ACKNOWLEDGEMENTS

This work was supported by a grant from Lilly Research Laboratories, Indianopolis, Indiana.

The authors also wish to acknowledge the contributions of the following persons who participated in the conduct of this trial: Faruk Said Abuzzahab, Sr, MD, PhD, Clinical Psychopharmacology Consultants, Minneapolis, MN; George S. Alexopoulos, MD, Cornell Medical Center Westchester Division, White Plains, NY;

Robert J. Bielski, MD, Institute for the Study of Mood Disorders, Okemos, MI; Richard Lewis Borison, MD, PhD, Medical College of Georgia, Augusta, GA; Meryl S. Brod, PhD, University of California, San Francisco, CA; Steven A. Cohen-Cole, MD, Emory University School of Medicine, Atlanta, GA; Cal K. Cohn, MD, The Hauser Clinic & Associates, Houston, TX; John McCall Downs, MD, University of Tennessee at Memphis, TN; Robert L. Dupont, MD, Institute for Behavior and Health, Inc, Rockville, MD; James Mecham Ferguson, MD, Pharmacology Research Institute, Salt Lake City, UT; David G. Folks, MD, University of Alabama at Birmingham, Birmingham, AL; Gary L. Gottlieb, MD, MBA, University of Pennsylvania School of Medicine, Philadelphia, PA; Benjamin Graber, MD, Frank J. Menolascino, MD, University of Nebraska Medical Center, Omaha, NE; Angelos E. Halaris, MD, PhD, Case Western Reserve University, Cleveland, OH; James T. Hartford, MD, Medical Center North, Cincinnati, OH; Marc Hertzman, MD, George Washington University Medical Center, Washington, DC; James W. Jefferson, MD, University of Wisconsin Medical School, Madison, WI; Dilip V. Jeste, MD, University of California, San Diego, CA; Lorrin M. Koran, MD, Stanford University School of Medicine, Stanford, CA; Lawrence W. Lazarus, MD, Rush-Presbyterian-St Luke's Medical Center, Chicago, IL; Bharat Raj Swaroop Nakkra, MD, MRCPsych, St Louis University School of Medicine, St Louis, MO; Gregory F. Oxenkrug, MD, PhD, VA Medical Center, Providence, RI; Stephen A. Rappaport, MD, Methodist Hospital, Indianapolis, IN; Murray H. Rosenthal, DO, Behavioral Medicine Systems, San Diego, CA; Carl Salzman, MD, Harvard Medical School, Boston, MA; Ram Shrivastava, MD, Park Lexington Regional Research, New York, NY; Peter E. Stokes, MD, Cornell University Medical Center, New York, NY; Jaron L. Winston, MD, Austin Diagnostic Clinic, Austin, TX; David W. Wheadon, MD, Ellen J. Schatz, Dena E. Marvel, Raymond Albritton and Gary D. Tollefson, MD, PhD, Eli Lilly and Company, Indianopolis, IN.

Portions of this work were presented at a workshop on treatment of late life depression, sponsored by the NIMH Aging Branch and MacArthur Foundation Research Network on Psychopathology and Development, Alexandria, VA, January 24–25, 1994; the American Psychiatric Association Annual Meeting, Philadelphia, PA, May 25, 1994; and the New Clinical Drug Evaluation Unit (NCDEU) Program, Marco Island, FL, May 31, 1994.

REFERENCES

Alexopoulos, G. S., Young, R. C. and Meyers, B. S. (1993) Geriatric depression: Age of onset and dementia. *Biol. Psychiat.* **34**, 141–145.

Altamura, A. C., DeNovellis, F., Guercetti, G., Invernizzi, G., Percudani, M. and Montgomery, S. A. (1989) Fluoxetine compared with amitriptyline in elderly depression: A controlled clinical trial. *Int. J. Clin. Pharmacol. Res.* **9**, 391–396.

American Psychiatric Association (1987) *Diagnostic and Statistical Manual of Mental Disorders, Third Edition—Revised (DSM-III-R).* American Psychiatric Association Washington, DC.

Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M. and Rose, T. L. (1982) Screening tests for geriatric depression. *Clin. Gerontol.* **1**, 37–43.

Cohen, J. (1992) A power primer. *Psychol. Bull.* **112**, 155–159.

Depression Guideline Panel (1993) *Depression in Primary Care: Volume 2. Treatment of Major Depression. Clinical Practice Guideline, Number 5.* US Department of Health and Human Services, Public Health Service, Rockville, MD. Agency for Health Care Policy Research. AHCPR Publication No. 93–0551.

Falk, W. E., Rosenbaum, J. F., Otto, M. W., Zusky, P. M., Weilburg, J. B. and Nixon, R. A. (1989) Fluoxetine versus trazodone in depressed geriatric patients. *J. Geriatr. Psychiat.* **2**, 208–214.

Feighner, J. P., Boyer, W. F., Meredith, C. H. and Hendrickson (1988) An overview of fluoxetine in geriatric depression. *Brit. J. Psychiat.* **153** (Suppl. 3), 105–108.

Folstein, M., Folstein, S. and McHugh, P. (1975) 'Mini-mental state': A practical method for trading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198.

Georgotas, A. and McCue, R. E. (1989) Relapse of depressed patients after effective continuation therapy. *J. Affect. Disord.* **17**, 159–164.

Hamilton, M. (1960) A rating scale for depression. *J. Neurol. Neurosurg. Psychiat.* **23**, 56–62.

Klawansky (1994) Metaanalysis on the treatment of depression in late-life. In *Diagnosis and Treatment of Depression in the Elderly: Proceedings of the NIH Consensus Development Conference* (L. S. Schneider, C. F. Reynolds, B. Lebowitz and A. Friedhoff, Eds). American Psychiatric Press, Washington, DC.

Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta Analysis.* Academic Press, New York.

Johnson, B. T. (1989) *Software for the Meta-Analytic Review of Research Literatures.* Erlbaum, Hillsdale, NJ.

La Pia, S., Giorgio, D., Cirello, R., Sannino, A., De Simone, L., Paoletti, C. and Colonna, C. V. (1992) Evaluation of the efficacy, tolerability and therapeutic profile of fluoxetine versus mianserin in the treatment of depressive disorders in the elderly. *Curr. Ther. Res.* **52**, 847–858.

Salzman, C. (1994) Pharmacological treatment of depression in elderly patients. In *Diagnosis and Treatment of Depression in the Elderly: Proceedings of the NIH Consensus Development Conference* (L. S. Schneider, C. F. Reynolds, B. Lebowitz and A. Friedhoff, Eds). American Psychiatric Press, Washington, DC.

SAS Institute (1989) *SAS/STAT User's Guide, Version 6, Fourth Edition*. SAS Institute, Cary, NC.

Small, G. W. (1991) Recognition and treatment of depression in the elderly. *J. Clin. Psychiat.* **52**, 11–22.

Tollefson, G. D. and Holman, S. L. (1993) Analysis of the Hamilton Depression Rating Scale factors from a double-blind, placebo-controlled trial of fluoxetine in geriatric major depression. *Int. Clin. Psychopharmacol.* **8**, 253–259.

Tollefson, G. D., Bosomworth, J. C., Heiligenstein, J. H., Potvin, J. H., Holman, S. L. and the Fluoxetine Collaborative Study Group (1995) A double-blind placebo-controlled clinical trial of fluoxetine in geriatric patients with major depression. *Int. Psychogeriat.* **7**, 89–104.