Primary care management of major depression in patients aged ≥55 years:
outcome of a randomised clinical trial

Harm WJ van Marwijk, Herman Ader, Marten de Haan, Aartjan Beekman

ABSTRACT

Background
Late-life depression is associated with chronic illness, disability, and a poor prognosis. Primary care management may be in need of improvement.

Aim
To compare the effects of an intervention programme that aims to improve the identification, diagnosis, and treatment of depression in patients aged ≥55 years with the effects of usual care.

Design of study
Cluster randomised controlled trial.

Setting
General practices in the Netherlands.

Method
Trained GPs performed the intervention and their practice assistants conducted the screenings. Patients were screened with the 15-item Geriatric Depression Scale (GDS-15) and given a consultation with the GP who diagnosed depression with the mood module of the Primary Care Evaluation of Mental Disorders (PRIME-MD). Antidepressant treatment was proposed. Primary outcomes were measured with the Montgomery Åsberg Depression Rating Scale (MÅDRS). Trained independent research assistants performed independent evaluations in both arms.

Results
Eighteen practices (23 GPs) were allocated to the intervention and 16 practices (20 GPs) to usual care. From June 2000 to September 2002, 3937 patients were screened; 579 patients had a positive score on the GDS-15, 178 had major depression, of whom 145 participated in the trial. MÅDRS scores for the intervention group dropped from 21.66 at baseline to 9.23 at 6 months, and the usual care group from 20.94 at baseline to 11.45 at 6 months. MÅDRS scores decreased during the year in both arms. For the intervention group, these scores increased between 6 and 12 months.

Conclusion
The programme resulted in lower MÅDRS scores in the intervention group than in the usual care group, but only at the end of the intervention, at 6 months after baseline.

Keywords
disease management programme; general practice; major depression; older adults; screening; usual care.

INTRODUCTION

Late-life depression is associated with chronic illness, disability and, frequently, a poor prognosis.[1] In many countries, studies have been planned and undertaken to improve the identification and treatment of patients with major depression in primary care. Several interventions have been tested in randomised controlled trials, but most have limited effects.[2–4] Screening may improve patient outcomes to some extent when it is coupled with system changes that help to ensure adequate treatment and follow-up,[5] but different reviews come to different conclusions.[6] Collaborative care — that is, structured care in which non-medical specialists (such case managers usually have a nursing background) have a greater role than is usually the case now in augmenting the care — appears to be the best candidate for improving the outcomes of depression in the short and longer term.[3,7]

As a result of comorbidity issues and competing demands in the surgery, countries with a strong primary care service, such as the UK and the Netherlands, have a need for systems that can be operated within the other conflicting demands of primary care, whereas other countries might base policy on a more specialised service.[8] Many studies have looked at younger adults, but few have addressed older adults.

HWJ van Marwijk, *MD, PhD, associate professor of general practice;* **H Ader**, *PhD, biostatistician;* **M de Haan**, *MD, PhD, professor of general practice;* **A Beekman**, *MD, PhD, professor of psychiatry, EMGO Institute, VU University Medical Centre, Amsterdam, The Netherlands.*

Address for correspondence
Harm van Marwijk, Department of General Practice,
EMGO Institute, VU University Medical Centre, Van der
Boechorststraat 7, 1081 BT Amsterdam, The Netherlands.
E-mail: hwj.vanmarwijk@vumc.nl

Submitted: 4 February 2008; **Editor's response:** 15 April 2008; **final acceptance:** 11 June 2008.

©British Journal of General Practice.

This article was originally online first. Cite this article as:
Br J Gen Pract 2008; **58:** 680–687. Advance online publication.
DOI: 10.3399/bjgp08X342165

In one of the first randomised studies that was conducted in the US, a multifaceted intervention increased the identification and treatment of late-life depression, but there was no improvement in the severity of the depression or the disability.[9] A second large US study reported positive results of a collaborative care intervention on many aspects of depression and wellbeing in older people, compared with usual care;[10] however, these results were from a US managed-care setting, which makes it difficult to compare the results with a European primary care setting. A third US study supported the effectiveness of an intervention specifically aimed at reducing suicidal ideation, regardless of the severity of the depression, in older patients in primary care,[11] to the extent that it reduced mortality.[12] In the fourth, and perhaps most relevant, feasibility study in the UK, older patients with major depression in the intervention group were somewhat less likely to suffer from this condition at follow-up compared with those who had had usual care (odds ratio [OR] 0.32, 95% confidence interval [CI] = 0.11 to 0.93, *P* = 0.036).[13]

The aim of the study was to understand how a programme based on disease management concepts for older adults with major depression would compare with usual care in a setting in which usual care and access to it is well developed, as is the case in Dutch primary care. The main aim of the present study was to compare the effects of a disease management programme standardising the identification, diagnosis, and treatment of late-life depression in general practice with the effects of usual care.

METHOD

The background and design of this study have been published previously,[14] as well as the economic results,[15] but a short summary is presented here.

Setting and recruitment

GPs in the eastern region of West-Friesland in the Dutch province of North Holland were invited to participate in a randomised controlled trial comparing the effects of an intervention programme aiming to improve the identification, diagnosis, and treatment of depression in older people, with the effects of usual care. Cluster randomisation was done at practice level and performed by an independent statistician.

GPs were recruited during a meeting of the regional Continuous Medical Education (CME) section covering GPs in the region. Those who did not attend were sent a letter of invitation and were also contacted by telephone.

As the intervention was intended to be implemented by the GPs, randomisation took place at GP level (that is, in the practice building). Cluster randomisation was chosen to prevent the carrying over of results from the intervention group to the usual care group.

How this fits in

Primary care management of late-life depression needs improvement. In a practice randomised controlled trial, the effects of an intervention to improve identification, diagnosis, and treatment of depression in patients aged 55 years or over (18 practices), were compared with the effects of usual care (16 practices). Observed depressive symptoms scores halved during the year in both arms. The programme resulted in somewhat lower depressive symptom levels, but only after 6 months.

Patients

The inclusion criteria for patients were being aged ≥55 years and having a current major depressive episode. Any previous episode of depression that occurred more than 6 months ago was permitted. Exclusion criteria were psychosis, bipolar depression, severe social dysfunctioning, inability to communicate in the Dutch language, alcohol or drug misuse, cognitive impairment, and current use of antidepressants. Those patients who were ineligible were treated routinely by the GP or referred to specialist care.

Intervention group

The practice assistant carried out screening for depression in the waiting room. All patients aged ≥55 years were asked to complete the 15-item version of the Geriatric Depression Scale (GDS-15).[16] The GDS-15 score ranged from 0–15, with higher scores indicating a greater likelihood of depression. Patients with a positive screening result (GDS score of ≥5) were asked to participate in a diagnostic interview administered by the GP. The GP used the mood module of the Primary Care Evaluation of Mental Disorders (PRIME-MD) to assess symptoms of major depression as classified in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM®) IV.[17] The GP diagnosed major depression clinically in patients with at least five depressive symptoms, including mood or activity. Trained research assistants administered the PRIME-MD interview to the control subjects at baseline and at follow-up. If there was a discrepancy between the two assessments and a patient had a positive PRIME-MD according to the interviewer and not according to the GP, the patient was referred back to the GP for diagnostic assessment.

Usual care group

To determine the effectiveness of the disease management programme it was necessary to have a control group that represented current actual practice, that is, a group comprising patients receiving usual care. In the usual care group the practice assistants presented the GDS-15 to all patients aged ≥55 years who visited the practice. The GPs and the practice assistants were unaware of the cut-off score, which was determined at the research centre. The researchers did not inform GPs about which of their patients were included in the study.

A research assistant administered the PRIME-MD diagnostic interview with control subjects who had a positive score on the GDS-15. A positive score indicated that a patient was eligible for participation in the study. Patients were asked not to inform their GPs that they were participating.

Intervention

Treatment offered by GPs was based on the guidelines for the treatment of depression issued by the Dutch College of General Practitioners (NHG).[18] The treatment consisted of education and information, drug therapy, and supportive counselling. Education and information about depression consisted of information on its causes, treatment (in particular, drug therapy), prognoses, and self-help. The patients had to return once every 2 weeks during the first 2 months, and then once a month for a period of 4 months. The total treatment period lasted for 6 months.

Drug therapy consisted of the prescription of 20 mg of paroxetine once daily. This drug was chosen because, when the study started, it was the most commonly prescribed antidepressant in the Netherlands.[19] During supportive counselling the patient and GP together identified one specific problem that bothered the patient. A practical problem was chosen, because it was not usual care to include more reflective forms of psychotherapy. It was planned that, in any encounter, specific attention should be paid to generic counselling. The intensity of the counselling was low. A generic type of activity scheduling (to enhance pleasurable activities) was recommended to the GPs that is probably as effective as problem-solving treatment.[20] At the time of the trial (2002–2003), it was not yet possible to implement problem-solving treatment in Dutch primary care because there were no experienced trainers. The activity-scheduling concept was also used in the IMPACT study, but there it was combined with problem-solving treatment.[10]

Instruments used to measure outcomes

Primary outcome measures were the Montgomery Åsberg Depression Rating Scale (MÅDRS)[21] and the PRIME-MD scores. The MÅDRS, which is a patient interview, measured the severity of symptoms and the effect of treatment over time. Sub-groups of MÅDRS scores (50% symptom reduction and a MÅDRS score of <10) were also assessed. Furthermore, recovery from depression, defined as the absence of a PRIME-MD diagnosis of major depression, was assessed four times: at baseline (after the diagnostic interview), and after 2, 6, and 12 months. These PRIME-MD measures were additional primary outcome measures. The total PRIME-MD score was also analysed. The Clinical Global Impression (CGI) scale was also used as an additional primary outcome;[22] these data are reported in the online version of this article.

Secondary outcome measures were: the SF-36 Mental and Physical Composite Scales,[23] subjective wellbeing, satisfaction, quality of life (visual analogue scale of the EuroQol), and activities of daily life. At baseline and after 12 months the scores for the Mini-Mental State Examination and one Diagnostic Interview Schedule item (are you depressed now?) were also recorded. The scores for the GDS-15 were assessed at all measurement points. A parallel article reports on the cost-effectiveness analysis.[15]

The measurements, which were repeated after 2, 6, and 12 months, were performed by research assistants who had attended one specific training session. The research assistants were blinded with regard to the treatment group to which the patients were allocated.

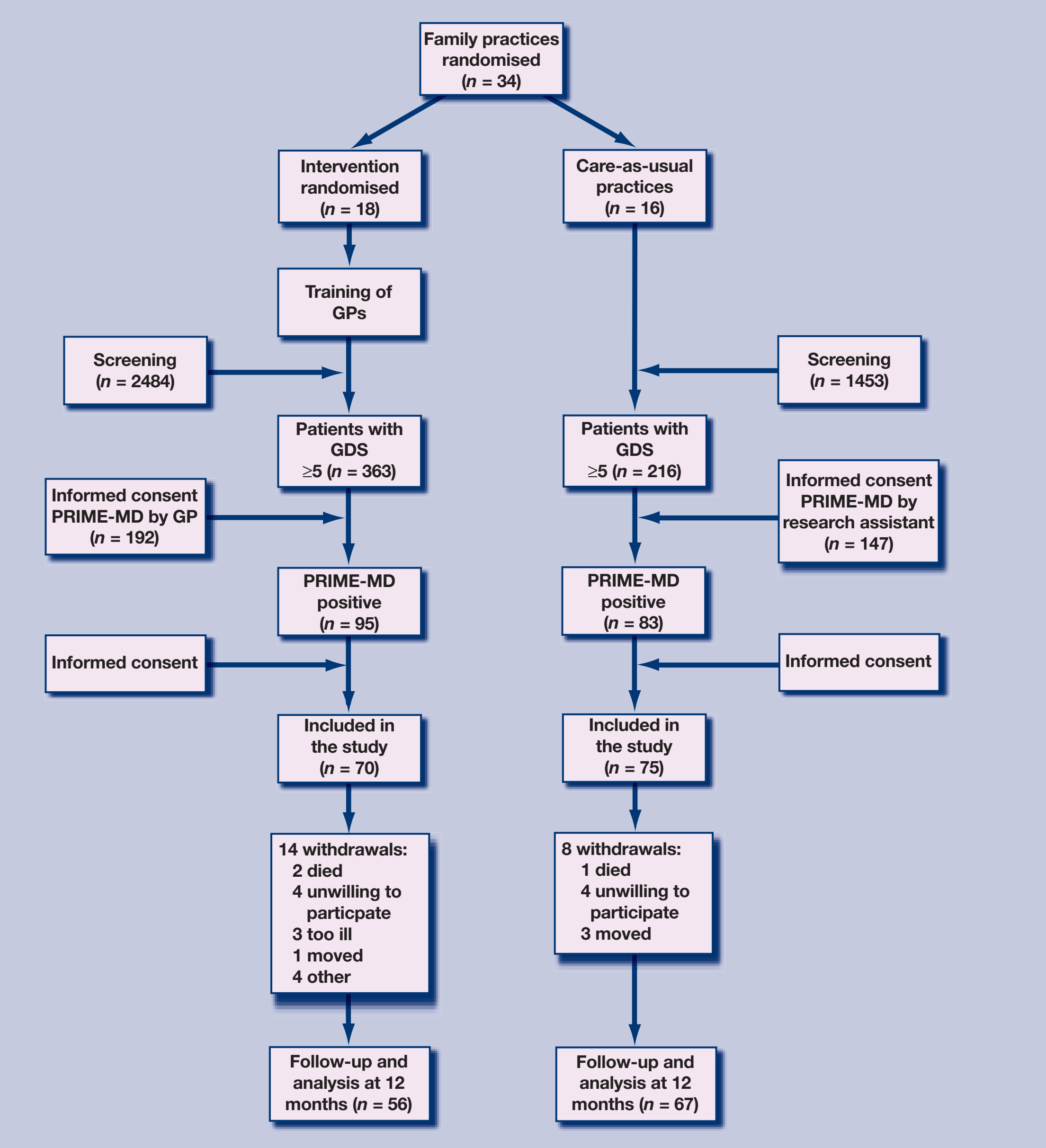
Statistical analysis

The calculation of the power was related to the primary outcome measures. Power analyses showed that, using continuous outcome measures, 70 patients were needed in each group to be able to detect a medium-sized treatment effect (Cohen's δ or 'effect size' of 0.50) in conventional analyses.[24] The principal analysis of the primary outcome measures consisted of an analysis of variances based on intention to treat. Time-trend analyses were performed to determine whether there were differences in the results at any of the measurements points. The differences described apply to those patients who agreed to participate.

To account for the levelled structure of the data and the cluster randomisation procedure, multilevel analysis was performed on the MÅDRS primary outcomes and the additional PRIME-MD primary outcome variable.[25] The levels were GP, patient, and measurement point. All models were specified in such a way that it was possible to determine whether at any measurement point there were differences in change from baseline between the treatment groups.

According to the study protocol, it was intended to base the analysis on matched pairs of GPs, however because of the great variety in GP characteristics and the cluster randomisation procedure, GPs were matched on the basis of propensity scores.[26–30]

Propensity scores were corrected for bias in treatment assignment using a large set of observed confounders and their interactions. In this case, it showed that the randomisation had not been completely successful, possibly due to the small number of practices that were randomised. In particular, some patient characteristics were unequally


Family practices randomised (n = 34)
Intervention randomised (n = 18)
Care-as-usual practices (n = 16)
Training of GPs
Screening (n = 2484)
Screening (n = 1453)
Patients with GDS ≥5 (n = 363)
Patients with GDS ≥5 (n = 216)
Informed consent PRIME-MD by GP (n = 192)
Informed consent PRIME-MD by research assistant (n = 147)
PRIME-MD positive (n = 95)
PRIME-MD positive (n = 83)
Informed consent
Informed consent
Included in the study (n = 70)
Included in the study (n = 75)
14 withdrawals: 2 died 4 unwilling to particpate 3 too ill 1 moved 4 other
8 withdrawals: 1 died 4 unwilling to participate 3 moved
Follow-up and analysis at 12 months (n = 56)
Follow-up and analysis at 12 months (n = 67)

GDS-15 = 15-item version of the Geriatric Depression Scale. PRIME-MD = Primary Care Evaluation of Mental Disorders.

Figure 1. Flowchart of the trial.

distributed over the two treatment groups. As such, all tests were corrected for propensity by including the propensity score as a covariate in the analysis of variance and multilevel models. As all tests were corrected for propensity, adjusted means for the primary and secondary outcomes are also reported. The baseline means are not adjusted. The statistical analyses were performed with SPSS (version 12).

RESULTS

Recruitment

The recruitment procedure yielded a total of 43 GPs (out of a total of 101) from 34 practices (out of a total of 70) who were willing to participate. There were 18 practices (including 23 GPs) allocated to the intervention group and 16 (including 20 GPs) to the usual care group (Figure 1). The inclusion of patients started in June 2000 and ended in September 2002. In total, 3937 patients were screened to achieve the required number of patients for the study. The total number included 145 patients with major depression.

The uptake of the screening and diagnostic procedures was positive. Of the 95 patients with major depression in the intervention group, 70 agreed to participate in the study (74%), and of the 83 patients with major depression in the usual care group, 75 agreed to participate (90%). This may reflect the willingness of the patients and GPs to accept the diagnosis of depression and its subsequent treatment in the intervention group. In the control group no such acceptance was required, but the patients did have to agree to a diagnostic interview.

Table 1. Patient characteristics.

	Intervention, *n* = 70	Usual care, *n* = 75	Total, *n* = 145
Age in years (SD)	66.4 (± 9.2)	65.0 (± 8.3)	65.6 (± 8.7)
Female sex (%)	42 (60)	41 (55)	83 (57)
Married (%)	39 (56)	51 (68)[a]	90 (62)
Living independently (%)	70 (100)	74 (99)	144 (99)
Education: none–low (%)	49 (70)	45 (60)[a]	94 (65)
Occupation: elementary–lower (unskilled) (%)	48 (69)	37 (49)[a]	85 (59)
Sharing household (%)	41 (59)	57 (76)[a]	98 (68)
Number of chronic illnesses (SD)	2.0 ± 1.6	2.3 (± 2.0)	2.2 (± 1.8)
Previous depression (%)	60 (86)	60 (80)	120 (83)
Age at first depression (SD)	44.4 (± 2.6)	38.2 (± 2.5)[a]	41.4 (± 1.8)
Family history of depression (%)	27 (39)	27 (36)	54 (37)

[a]*P<0.05. SD = standard deviation.*

Patient characteristics

The characteristics of the patients who participated are presented in Table 1. The majority of the patients were female. In total, 83% of the patients had experienced a previous episode of depression. There were statistically significant differences in characteristics between the patients in the intervention group and the usual care group, which indicates that the randomisation procedure had not been completely successful. These differences were accounted for in the analyses by using a propensity score as a covariate (note that no covariables that are of obvious substantive interest as a confounder, for example, sex, are included in the propensity score).

Follow-up

After 2 months 141 patients were still in the trial, 126 were still in the trial after 6 months, and 123 patients completed the 1-year follow-up. A total of 22 patients dropped out for various reasons (Figure 1). There were no differences between the two study groups with regard to the proportion who dropped out. The 95% CI for the difference in follow-up of 9% between intervention group (56/70) and usual care group (67/75) was –3 to 22%.

Table 2. Results of multilevel analysis (mean ± standard error) on the primary outcome variables MÅDRS and PRIME-MD.[a]

Outcome		Baseline[b]	2 months	6 months	12 months
MÅDRS	Intervention	21.66 ± 2.86	19.56 ± 3.32	9.23 ± 2.84[c]	10.80 ± 2.85
	Usual care	20.94 ± 2.48	19.58 ± 3.49	11.45 ± 2.52	10.09 ± 2.50
PRIME-MD	Intervention	6.10 ± 0.80	–	2.80 ± 1.04	3.23 ± 1.04
	Usual care	6.33 ± 1.01	–	3.99 ± 1.22	3.74 ± 1.21

[a]*Higher scores on MÅDRS and PRIME-MD indicate more severe depression.* [b]*Baseline scores were not corrected for propensities.* [c]*Statistically significantly different from the usual care group. MÅDRS = Montgomery Åsberg Depression Rating Scale. PRIME-MD = Primary Care Evaluation of Mental Disorders.*

Intervention

During the study period, 95% of patients in the intervention group and 94% of patients receiving usual care had at least one contact with their GP. Forty-six (66%) patients in the intervention group and eight (11%) patients receiving usual care received some form of specific mental health care (antidepressant medication or referral) during follow-up. Twenty-eight (40% patients in the intervention group received antidepressant treatment for at least 6 months, as recommended by Dutch depression guidelines.[20]

Primary outcomes

Table 2 presents the results of the analysis of variance.

MÅDRS

Baseline values for the MÅDRS were 21.66 ± 2.86 in the intervention group, but a steady decline in the scores for this scale occurred during treatment and follow-up, but scores went up slightly from 9.23 to 10.80 after 6 months. After 2 months the mean scores were 19.56 ± 3.32, at the end of the treatment (after 6 months) the mean scores were 9.23 ± 2.84, and at the end of the follow-up period the scores were 10.80 ± 2.85. In the group of patients receiving usual care the baseline values for the MÅDRS were 20.94 ± 2.48. In this group there was a steady decline in the scores during treatment and follow-up. After 2 months the mean scores were 19.58 ± 3.49, while at the end of the treatment, after 6 months, the mean scores were 11.45 ± 2.52, and at the end of the follow-up phase the scores had further declined to 10.09 ± 2.50. Trend analysis revealed statistically significant differences after 6 months between the MÅDRS scores in the intervention group and the group of patients receiving usual care (effect size δ = 0.28). At 12 months MÅDRS scores were lower for the usual care group than the intervention group.

To gauge the robustness of these findings, two additional analyses were applied to the MÅDRS scores: an analysis of patients with a score of <10 indicating recovery and an analysis of 50% symptom reduction (Appendix 1). Recovery was higher at 6 months in the intervention group than in the control group: 48% versus 27%; risk difference 21% (95% CI = 0.04 to 0.38). At follow-up, a statistically significant difference was found between the intervention group and the usual care group with regard to the 50% symptom reduction (2 months: 31% versus 16%; 6 months: 42% versus 26%).

PRIME-MD

Multilevel analysis showed that the PRIME-MD outcome (major depression) was not influenced by GP level, characteristics, treatment, or age. Baseline values for the number of depressive symptoms were 6.10 ± 0.80 in the intervention group (Table 2). Data on the PRIME-MD at 2 months could not be collected due to an administrative error. Dichotomised, after 6 months, 50% (*n* = 35) of the patients in the intervention group no longer had a diagnosis of major depression (PRIME-MD negative) versus 55% (*n* = 41) of the patients in the usual care group. This is a non-significant difference. After 1 year, 66% (*n* = 37) of the patients in the intervention group no longer had a diagnosis of major depression versus 64% (*n* = 43) in the usual care group. This is also a non-significant difference.

At the end of the treatment (after 6 months) the mean level of depressive symptoms in the intervention group had declined to 2.80 ± 1.04, and at the end of the follow-up the scores were 3.23 ± 1.04. In the usual care group, the baseline number of depressive symptoms was 6.33 ± 1.01. Data collection on the PRIME-MD at 2 months also failed in this group. At the end of the treatment (after 6 months) the mean scores had declined to 3.99 ± 1.22, and at the end of the follow-up, the scores had further declined to 3.74 ± 1.21. Trend analysis revealed no statistically significant differences between the PRIME-MD scores in the intervention group and the usual care group at any measurement point.

No statistically significant differences were shown on the CGI.

Secondary outcomes

There were no statistically significant differences between the intervention group and the usual care group with regard to any of the secondary outcome measures (Appendix 2).

DISCUSSION

Summary of main findings

This study of the effectiveness of an intervention programme to improve the identification, diagnosis, and treatment of late-life depression in primary care shows that on the main outcome, the MÅDRS, at the end of the intervention period, a statistically significant difference was found in favour of the intervention group. It was remarkable that after 6 months and 1 year the patients in both groups had improved significantly on the MÅDRS primary outcome compared with at baseline. The same applies to the number of depressive symptoms (PRIME-MD): the majority of patients in both groups no longer had a diagnosis of major depression. At 12 months MÅDRS scores were lower for the usual care group than the intervention group. The mean decrease in MÅDRS scores was quite substantial, that is, a decrease of more than 50% in both groups.

Strengths and limitations of the study

According to the study protocol, the intention was to base the analysis on matched pairs of GPs. However, because of the cluster randomisation procedure and some variety in characteristics of GPs, it was decided to match according to propensity scores. These scores make it possible to correct for the influence of a large set of observed confounders and their interactions on the treatment assignment. Although there were no differences in depression severity level between the two groups at baseline, the clustered randomisation procedure had not been completely successful for some patient characteristics (such as sex).

It could be expected that this is a group with more than one health problem when they attend the GP, which will have an impact on the identification and treatment of the depression, if only because of competing demands on a GP's time.[8] The required number of 70 per arm was not achieved for the entire follow-up of 1 year.

Comparison with existing literature

Several factors might explain the relatively modest effects of the intervention. The case-finding procedure applied led to the selection of patients with moderate major depression, as demonstrated by the MÅDRS scores (20–21) and by the relatively good prognosis of many patients. There are also are doubts about the effects of antidepressants in primary care.[31] These arguments favour active follow-up and stepped care rather than an immediate prescription of antidepressants, that is, active management should only be commenced when patients do not recover.

In addition, the study asks whether the GPs who participated could have been more motivated to treat depression and whether they were perhaps more familiar with depression guidelines than other GPs in the Netherlands.[18] The data does not support this assertion: a total of only 48% of subjects in the intervention arm received antidepressants for 6 months. Another Dutch study showed that more than 70% of older patients received antidepressants from their GP.[32] The Dutch depression general practice guidelines are highly accessible and are integrated into the medical records systems, including the formularies, that most Dutch GPs use.[21] Most Dutch GPs are, therefore, well aware of the guideline criteria.[16]

When it comes to recognition, the process of the diagnostic interview might have prompted patients to seek help where they might not have done otherwise. However, this was an intended effect of the intervention ('a package').

There are also differences between the organisation of primary care in Europe and in the US. The review of the effectiveness of disease management programmes for major depression in primary care contained only one trial that was performed in Europe.[4] This education programme aimed at GPs did not improve the identification or recovery of patients from major depression. Two of three US trials that studied older subjects with major depression in primary care had positive results.[9–11]

Implications for future research and clinical practice

The results of this study support a stepped care or repeated assessment approach that is not too far from current usual care, but requires active follow-up by a nurse or perhaps via the internet. For an example of internet-based self-help and monitoring depression, see http://moodgym.anu.edu.au. Such projects require further large-scale testing within a practice setting.

Funding body

The West-Friesland Study was funded by a grant from the Netherlands Organisation for Health Research and Development (ZonMw). This ia an independent organisation (2200/0019)

Ethical approval

The Medical Ethics Committee of the VU University Medical Centre approved the study protocol (2000/12)

Competing interests

The authors have stated that there are none

Acknowledgements

We would like to thank the participating GPs, patients, and research assistants for their efforts, and especially Paulien Schaap-Hoekstra for all of her hard work. We would also like to thank Dr Dick Bijl for his efforts to make this project possible.

Discuss this article

Contribute and read comments about this article on the Discussion Forum: http://www.rcgp.org.uk/bjgp-discuss

REFERENCES

1. Licht-Strunk E, van der Windt DA, van Marwijk HW, *et al.* The prognosis of depression in older patients in general practice and the community. A systematic review. *Fam Pract* 2007; **24(2):** 168–180.
2. Worrall G, Angel J, Chaulk P, *et al.* Effectiveness of an educational strategy to improve family physicians' detection and management of depression: a randomized controlled trial. *CMAJ* 1999; **161(1):** 37–40.
3. Gilbody S, Bower P, Fletcher J, *et al.* Collaborative care for depression: a cumulative meta-analysis and review of longer-term outcomes. *Arch Intern Med* 2006; **166(21):** 2314–2321.
4. Bijl D, van Marwijk HW, de Haan M, *et al.* Effectiveness of disease management programmes for recognition, diagnosis and treatment of depression in primary care. *Eur J Gen Pract* 2004; **10(1):** 6–12.
5. Pignone MP, Gaynes BN, Rushton JL, *et al.* Screening for depression in adults: a summary of the evidence for the US Preventive Services Task Force. *Ann Intern Med* 2002; **136(10):** 765–776.
6. Gilbody SM, House AO, Sheldon TA. Routinely administered questionnaires for depression and anxiety: systematic review. *BMJ* 2001; **322(7283):** 406–409.
7. Richards DA, Lovell K, Gilbody S, *et al.* Collaborative care for depression in UK primary care: a randomized controlled trial. *Psychol Med* 2008; **38(2):** 279–287.
8. Nutting PA, Rost K, Smith J, *et al.* Competing demands from physical problems: effect on initiating and completing depression care over 6 months. *Arch Fam Med* 2000; **9(10):** 1059–1064.
9. Callahan CM, Hendrie HC, Dittus RS, *et al.* Improving treatment of late life depression in primary care: a randomized clinical trial. *J Am Geriatr Soc* 1994; **42(8):** 839–846.
10. Unützer J, Katon W, Callahan CM, *et al.* Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *JAMA* 2002; **288(22):** 2836–2845.
11. Bruce ML, Ten Have TR, Reynolds CF III, *et al.* Reducing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *JAMA* 2004; **291(9):** 1081–1091.
12. Gallo JJ, Bogner HR, Morales KH, *et al.* The effect of a primary care practice-based depression intervention on mortality in older adults: a randomized trial. *Ann Intern Med* 2007; **146(10):** 689–698.
13. Chew-Graham CA, Lovell K, Roberts C, *et al.* A randomised controlled trial to test the feasibility of a collaborative care model for the management of depression in older people. *Br J Gen Pract* 2007; **57(538):** 364–370.
14. Bijl D, van Marwijk HWJ, Beekman ATF, *et al.* A randomized controlled trial to improve the recognition, diagnosis and treatment of major depression in elderly people in general practice: design, first results and feasibility of the West Friesland Study. *Primary Care Psychiatry* 2003; **8(4):** 135–140.
15. Bosmans J, de Bruijne M, van Hout H, *et al.* Cost-effectiveness of a disease management program for major depression in elderly primary care patients. *J Gen Intern Med* 2006; **21(10):** 1020–1026.
16. Van Marwijk HW, Wallace P, de Bock GH, *et al.* Evaluation of the feasibility, reliability and diagnostic value of shortened versions of the geriatric depression scale. *Br J Gen Pract* 1995; **45(393):** 195–199.
17. Spitzer RL, Williams JB, Kroenke K, *et al.* Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. *JAMA* 1994; **272(22):** 1749–1756.
18. Van Marwijk HWJ, Grundmeijer HGLM, Bijl D, *et al.* NHG-standaard Depressieve stoornis (depressie) (eerste herziening). [NHG Standard depressive disorder (depression) (first revision)]. *Huisarts en Wetenschap* 2003; **46(11):** 614–623 [in Dutch].
19. Van Marwijk HW, Bijl D, Adèr HJ, de Haan M. Antidepressant prescription for depression in general practice in The Netherlands. *Pharm World Sci* 2001; **23(2):** 46–49.
20. Cuijpers P, van Straten A, Warmerdam L. Behavioral activation treatments of depression: a meta-analysis. *Clin Psychol Rev* 2007; **27(3):** 318–326.
21. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. *Br J Psychiatry* 1979; **134:** 382–389.
22. Guy W. Clinical global impression. In: *ECDEU Assessment manual for psychopharmacology, revised*. Rockville, MD: National Institute of Mental Health, 1976.
23. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30(6):** 473–483.
24. Cohen J. *Statistical power for the behavioral sciences*. Hillsdale: Lawrence Erllbaum, 1988.
25. Snijders TAB, Bosker RJ. *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications, 1999.
26. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003; **158(3):** 280–287.
27. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002; **137(8):** 693–695.
28. Rosenbaum P. *Observational studies.* 2nd edn. New York: Springer, 2002.
29. Rubin DB. Assignment to treatment group on the basis of a covariate. *J Educ Statist* 1977; **2:** 1–26.
30. Adèr H, Mellenbergh G, Hand D. *Advising on research methods: A consultant's companion*. Huizen: Johannes van Kessel, 2008.
31. Hermens ML, van Hout HP, Terluin B, *et al.* Clinical effectiveness of usual care with or without antidepressant medication for primary care patients with minor or mild-major depression: a randomized equivalence trial. *BMC Med* 2007; **5:** 36.
32. Volkers AC, Nuyen J, Verhaak PF, Schellevis FG. The problem of diagnosing major depression in elderly primary care patients. *J Affect Disord* 2004; **82(2):** 259–263.
33. Spitzer RL, Williams JB, Kroenke K, *et al.* Validity and utility of the PRIME-MD patient health questionnaire in assessment of 3000 obstetric–gynecologic patients: the PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. *Am J Obstet Gynecol* 2000; **183(3):** 759–769.
34. Wells KB, Stewart A, Hays RD, *et al.* The functioning and well-being of depressed patients. Results from the Medical Outcomes Study. *JAMA* 1989; **262(7):** 914–919.
35. Folstein MF, Folstein SE, McHugh PR. 'Mini-mental state'. A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* 1975; **12(3):** 189–198.
36. Anonymous. EuroQol — a new facility for the measurement of health-related quality of life. *The EuroQol Group. Health Policy* 1990; **16(3):** 199–208.

COMMENTARY

What is a propensity score?

The paper by van Marwijk and colleagues[1] illustrates the application of propensity scores to the analysis of a cluster randomised trial. This commentary outlines the role of propensity scores in the analysis of non-randomised studies and randomised trials.

Propensity scores in non-randomised studies. Consider the example of a population-based register of angina patients. Suppose that a researcher wishes to compare the long-term survival of patients who received coronary artery bypass surgery (CABG) with those who did not receive surgery. Patients selected for CABG can be expected to differ from those that did not receive surgery in terms of important prognostic characteristics including the severity of coronary artery disease or the presence of concurrent conditions, such as diabetes. A simple comparison of the survival of patients who either did or did not receive CABG will be biased by these confounding variables. This 'confounding by indication' is almost invariably present in non-randomised studies of healthcare interventions and is difficult to overcome.

Rosenbaum and Rubin[2] proposed the use of propensity scores as a method for allowing for confounding by indication. Propensity may be defined as an individual's probability of being treated with the intervention of interest given the complete set of all information about that individual.[2] The propensity score provides a single metric that summarises all the information from explanatory variables such as disease severity and comorbity; it estimates the probability of a subject receiving the intervention of interest given his or her clinical status.[3] Individual subjects may have the same or similar propensity scores, yet some will have received the intervention of interest and others will not. For example, women of a certain age, with triple vessel disease and the same comorbidities, may have the same propensity for CABG but only some will receive surgery. An assumption of propensity score analysis is that a fair comparison of treatment outcomes can be made between subjects with similar propensity scores who either did or did not receive the treatment of interest. The propensity score may be estimated for each subject from a logistic regression model in which treatment assignment is the dependent variable. An attractive feature of this approach is that explanatory variables are selected on the basis of their ability to predict exposure to the intervention of interest, their possible associations with outcomes need not be considered.

Three methods are commonly employed to include propensity scores in analyses: matching, stratification, and regression adjustment. Matching requires that each treated individual is matched to an untreated individual with the same or similar propensity score. The process of stratification represents a more general extension of matching in which there is more than one treated or untreated individual per stratum. Once matched pairs or strata have been formed, the association of treatment with outcome is estimated by contrasting outcomes between treated and untreated sets of individuals with similar propensity for treatment. Propensity scores are, however, more commonly included in a regression model as an explanatory variable.

Propensity scores can only balance the observed patient characteristics between treatment groups.[4] Imbalances may remain even after propensity score adjustment if relevant subject characteristics were not measured or were only measured imprecisely. It is also advisable to check that propensity score groups are balanced with respect to patient characteristics, rather than assuming that such balance exists.[5]

Propensity scores in randomised trials. Randomisation is usually considered the optimal method for addressing problems of confounding by indication. However, imbalances in the distribution of subject characteristics between trial arms is especially likely in cluster randomised trials because clusters represent groups of individuals who may share characteristics that differ from subjects in other clusters. Imbalances are also more likely when the number of clusters in a trial is small. Van Marwijk *et al* have implemented a novel application of propensity scores to control for imbalance of individual subject characteristics between the arms of a cluster randomised trial.[1] In their study, subjects who were allocated to the intervention group were less likely to be married, to share a household, or to have higher levels of education or occupation (Table 1[1]). One approach would have been to use the variables in Table 1 to estimate, for each subject, the predicted probability of receiving the trial intervention given the pattern of observed subject characteristics. This propensity score could then be used to adjust analyses in which outcomes were compared between intervention groups. In the Discussion, the report suggests an analysis in which pairs of GPs are matched for propensity. This suggestion raises a question concerning how the levels of the individual subject and the cluster should be considered in the estimation and application of propensity scores.

Jennifer Nicholas,
Research Assistant, Department of Public Health Sciences, King's College London.

Martin C Gulliford,
Professor of Public Health, Department of Public Health Sciences, King's College London, Capital House, 42 Weston St, London SE1 3QD.
Email: martin.gulliford@kcl.ac.uk

REFERENCE

1. Van Marwijk HWJ, Ader H, de Haan M, Beekman A. Primary care management of major depression in patients aged ≥55 years: outcome of a randomised clinical trial. *Br J Gen Pract* 2008; **58:** 680–687.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70:** 41–55.
3. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79:** 516–524.
4. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002; **137:** 693–695.
5. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007; **26:** 20–36.

DOI: 10.3399/bjgp08X342471

Appendix 1. Additional analyses on the primary outcome measure MÅDRS.

	Baseline		2 months		6 months		12 months	
	Intervention (*n* = 67)	Usual care (*n* = 74)	Intervention (*n* = 67)	Usual care (*n* = 74)	Intervention (*n* = 60)	Usual care (*n* = 66)	Intervention (*n* = 56)	Usual care (*n* = 67)
MÅDRS 50% reduction (%)	–	–	21 (31)[a]	12 (16)	25 (42)	17 (26)	26 (46)	26 (39)
MÅDRS score <10 (%)	5 (7)	2 (3)	18 (27)	16 (22)	29 (48)[a]	18 (27)	26 (46)	29 (43)

[a] *Statistically significant difference compared with usual care.*

Appendix 2. Results on secondary outcome measures (mean).

	Baseline		2 months		6 months		12 months	
	Intervention	Usual care	Intervention	Usual care	Intervention	Usual care	Intervention	Usual care
GDS-155	7.3[a]	7.6	5.5	5.8	4.7	5.2	4.7	4.7
SF-36: physical summary component[16]	60.5	61.2	60.7	63.5	61.4	63.1	60.7	63.6
SF-36: mental summary component[34]	47.0	50.2	54.4	54.6	58.4	57.6	59.2	60.6
Subjective wellbeing	2.7	2.9	–	–	3.4	3.2	3.3	3.2
Satisfaction	2.0	1.8	–	–	2.4	2.3	2.3	2.2
MMSE[35]	25.6	26.6	–	–	–	–	27.3	27.0
ADL	9.5	9.6	11.2	11.5	9.2	9.6	9.4	9.6
DIS (not depressed now, %)	23	33	–	–	–	–	66	81
EuroQol of QoL (VAS)[36]	62.0	62.3	–	–	64.9	65.9	64.2	62.9

[a] *95% confidence interval of the difference: –0.70 to 1.24. ADL = activities of daily living. DIS = diagnostic interview schedule. GDS-15 = 15-item version of the Geriatric Depression Scale. MMSE = Mini Mental State Examination. QoL = quality of life.*

CONSORT Statement 2001: Checklist. Items to include when reporting a randomised trial.

PAPER SECTION and topic	Item	Descriptor	Reported on page #
TITLE and ABSTRACT	1	How participants were allocated to interventions (for example, 'random allocation', 'randomised', or 'randomly assigned')	2
INTRODUCTION			
Background	2	Scientific background and explanation of rationale	4
METHODS			
Participants	3	Eligibility criteria for participants, settings and locations where the data were collected	6
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered	8
Objectives	5	Specific objectives and hypotheses	5
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (for example, multiple observations, training of assessors)	9/10
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules	10
Randomisation: sequence generation	8	Method used to generate the random allocation sequence, including details of any restrictions (for example, blocking, stratification)	6
Randomisation: allocation concealment	9	Method used to implement the random allocation sequence (for example, numbered containers, or central telephone), clarifying whether the sequence was concealed until interventions were assigned	6 (cluster randomisation)
Randomisation: implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups	6 (HA)
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated	2 (cluster randomisation)
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses	10
RESULTS			
Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from study as planned, together with reasons	22
Recruitment	14	Dates defining the periods of recruitment and follow-up	12
Baseline data	15	Baseline demographic and clinical characteristics of each group	20
Numbers analysed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by 'intention-to-treat'. State the results in absolute numbers when feasible (for example, 10/20, not 50%)	10
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (for example, 95% confidence interval)	21
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory	10
Adverse events	19	All important adverse events or side-effects in each intervention group	n/a
DISCUSSION			
Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias, or imprecision and the dangers associated with multiplicity of analyses and outcomes	16
Generalisability	21	Generalisability (external validity) of the trial findings	18
Overall evidence	22	General interpretation of the results in the context of current evidence	18