PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Effectiveness of a peer mediated educational intervention in improving general practitioner diagnostic assessment and management of dementia: a cluster randomised controlled trial
AUTHORS	Pond, Dimity; Mate, Karen; Stocks, Nigel; Gunn, Jane; Disler, Peter; Magin, Parker; Marley, John; Paterson, Nerida; Horton, Graeme; Goode, Susan; Weaver, Natasha; Brodaty, Henry

VERSION 1 – REVIEW

REVIEWER	Louise Robinson Newcastle University, UK
REVIEW RETURNED	19-Jan-2018

GENERAL COMMENTS	Generally this is a well written paper describing a large, double blind, cluster RCT to evaluate the effectiveness of an educational intervention for GPs in Australia to improve the detection and management of dementia. The team are to be commended on successfully undertaking and completely such a large study in a primary care setting. However there are some areas which I feel require more clarification and detail; in addition I am not a statistical or trial expert so would recommend that reviewers with specific expertise in these areas are sought. Abstract: The brief description of the intervention, 'medical detailing intervention', is an unknown term to me; can the authors elaborate on or explain this term? In the outcome measures, can this be structured in to primary and secondary? Background: there are some key global references that would be relevant to include either here and/or in the discussion such as the World Alz Association 2016 report on healthcare for people with dementia which emphasises the need for countries to move to a task-shifted, primary care model of dementia care as need Increases alongside finite/reducing resources. Methods Intervention; I felt the current description whist helpful, focused more on the process of diagnosis and would have liked more information on the education provided on the management of dementia and how this differed, if any from usual care. Would the RACGP guidelines also be available to the control group? How did the educational intervention enhance this raising awareness or use of guidelines if at all?

	Results: similarly in this section, the focus was largely on detection with little data provided in the section on 'evidence-based care' that appeared to be related to management rather than detection yet the former appears in the title and the aim of the study? Discussion: here the authors state that by management they mean 'number of tests and referrals' ie purely linked to the diagnostic process. Clarity is needed through the paper as to whether 'management' refers just to pre-diagnoses or as I assumed post diagnosis care too. Comparison with other studies - ref 28 is highlighted in the introduction as a key study or importance built is not referred to at all here or how this study compares/contrasts with it. I would have liked to have seen more discussion about primary care led models and the future role in dementia care with reference to key global policy.

REVIEWER	Doris S.F. Yu The Chinese University of Hong Kong, Hong Kong
REVIEW RETURNED	12-Feb-2018

GENERAL COMMENTS	This is an interesting study to identify the effects of an intervention to enhance the diagnostic accuracy of GP on dementia. However, the conceptual clarity and study design of the study need more elaboration. Please see the specific comments listed below. Background λ The authors quote several studies on intervention to improve the detection and diagnosis of dementia. However, these studies, which can reflect an up-to-dated knowledge, are not adequately elaborated. [e.g. reference 28, what Is the meaning by delivering the educative intervention in coordinated dementia case management setting; For references 15 and 31, what were the barriers these studies targeted for.]. Without such information, it is difficult to interpret the originality of the current paper. Study design λ As the study intervention was targeted at improving the diagnosis of dementia by GP, it was unclear why the primary outcomes focused on patients' QOL and depression scores. This is less acceptability, especially when the intervention did not have explicit content to improve the dementia management. The intervention may be limited to use case studies to illustrate dementia management and provided the GP with a practice guideline. The dose of the intervention in this aspect is less coherent with the expected changes on patients' QoL and depression outcomes. The control group, indeed, also receive the practice guideline!! λ Carers were a sample subgroup, why the consent is not directly obtained from the carers, but the patients they are taking care of? λ List out what does the cut-off score of the CAMCOG-R <80 indicate for? λ For the intervention, the second audit is less clear. What are the "results of nurse assessment" referring to? Does it mean the reassessment by the GPCOG? If yes, what is the purpose of giving the feedback at 12 months later. λ Why an allocation ration of 1:2 was used? λ There was no information on the psychometric properties of the

	outcome measures. Although most of them are well known, more information for GPAQ is needed. Besides, what is the meaning by the GP identification of reversible cause of dementia? What is the conceptual meaning by 'patient enablement"! in fact, what intervention has been delivered directly to the carers, so that their improved outcome can be explained as relating to the study intervention. In the Discussion section, there is one sentence explaining "the improved enablement is related more to communication and empathy characteristics of the GP". The mechanism under which is very unclear! Why and how? λ The CAMCOG-R seems to be used as a golden measure to examine the sensitivity and specificity of the GP diagnosis. This is problematic, as the CAMCOG-R is only a screening tool. How can the authors validate even the cause of dementia? What is the criteria measure for this? Results: λ The results about the positive effect of the study intervention on the dementia diagnostic accuracy by GP need to be interpreted with caution. This is because in this study, the patients were referred to the GP for dementia screening. This is very different from the actual practice that the GP need to suspect the client to have dementia based on initial health assessment, and able to use the relevant tool to do the screening. Discussion: λ For the implication on practice/ policy, there should be more discussion about how the tested intervention can be translated to the actual practice. Not just to repeat the significance of the study intervention

REVIEWER	Seyed-Mohammad Fereshtehnejad Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada
REVIEW RETURNED	12-Apr-2018

GENERAL COMMENTS	1. In the Abstract, section 'Results', line 35: please note that the 95% confidence interval for a statistically significant OR with p-value<0.05, should be either entirely <1 or entirely >1. Here, the OR and P-value for the outcome 'satisfaction with GP communication' are reported as "p=0.024; odds ratio 2.1, 95% CI: 0.27-3.93". A 95% CI of 0.27-3.93 includes both protective (OR<1) and risk factor (OR>1) effects, which is not statistically significant and does not correspond with a p-value of 0.024. Accordingly, the 'Conclusion' that the intervention has improved satisfaction with GP communication is not correct and should be revised. 2. In 'Methods', subsection 'Participants', it is mentioned that 'GPs sent letters of invitation to all patients who met the inclusion criteria'. Please clarify what was the response rate in this study? How did you evaluate the representativeness of the participants who answered positive for participation? Was there any significant difference in the characteristics of the respondents vs. non-respondents? 3. In 'Methods', subsection 'Intervention', line 35: it is stated that the baseline audit of patients was done by the GPs in the intervention group. With this setting, the scorers, GPs in the intervention group, might have been biased during auditing. I wonder why blinding was not performed to prevent scorers' bias. In these trials, it is preferred that the assessments (here auditing) being performed by independent assessors who is blinded to the study arm.

	4. Likewise, it is important to clarify who performed the second audit after 12 months. Was it done by blinded independent assessors/interviewers? 5. If several assessors were involved in the study, how inter-raters agreement was evaluated? Is there any data available on inter- and intra-raters agreement? 6. In the 'Methods', subsection 'Sample Size': I wonder why a 7% difference in the WHOQOL-BREF scale score has been used for sample size calculation. Is this considered as a small, medium, or large effect size? Is this a minimum clinically relevant difference on this scale? Please further clarify and/or cite to a relevant reference. 7. In the 'Methods', subsection 'Sample Size': for sample size calculation, while authors assumed 15% for the number of drop-outs prior to study performance, the actual rate of drop-out patients after follow-up was larger (25% waitlist, 20% intervention arm). What was their strategy to deal with lowered statistical power? Was any post-study power analysis performed? 8. In 'Methods', subsection 'Statistical Analyses': authors have used regression-based generalized estimating equations (GEE) models to compare the outcome variables after 12 months follow-up between the two study arms considering the baseline measures as covariates. While this statistical approach is not inherently wrong, a better solution that I recommend is to use mixed effects models for repeated numeric outcomes. With mixed effects models, it is possible to consider outcome trajectory at individual level as a random factor in the model. In other words, mixed effects model has the advantage of taking into account each individual change over time, while the GEE models mostly fit the population average effects. 9. Was any interim analysis planned prior to the trial implementation? This is an important issue, since if the efficacy of the intervention had been shown in the middle of the trial, the entire patients population could have benefited from receiving the intervention. 10. Table 1: in the intervention arm, the sum of 671 (45%) males and 805 (55%) females does not correspond to the total number of 1478 participants on top of the column. A similar problem could be also found in rows 20-21 for the number of males and females in the GPs description in both study arms. Was there any other gender status? Please recheck the numbers and clarify the issue. 11. Table 2 summarizes baseline comparisons in the main outcome variables between the two study groups. While this is still a summary of cross-sectional baseline comparisons, I wonder why the p-values are from GEE models? As instructed in the 'Methods', the GEE models were applied to compare the effect of intervention after follow-up in the change occurred in the outcome variables from baseline. Please clarify this issue. 12. In the 'Results', subsection 'Outcome Measures for Patients', line 32: please refer to the same problem as comment NO.1. The 95% CI reported shows a NON-significant OR since it includes value 1, while the p-value is <0.05. Could it be possible that the range is indeed calculated for a mean difference of a numeric score and not an OR for a categorical scale? Please recheck the calculations and the corresponding scales. 13. It is not clear what strategy the authors followed to deal with the drop-outs in this trial. Was any per-protocol or intention-to-treat analysis performed? How about missing data imputation to increase statistical power while encountering with the drop-outs.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer 1: Louise Robinson, Newcastle University, UK

Generally this is a well written paper describing a large, double blind, cluster RCT to evaluate the effectiveness of an educational intervention for GPs in Australia to improve the detection and management of dementia.

The team are to be commended on successfully undertaking and completely such a large study in a primary care setting. However there are some areas which I feel require more clarification and detail; in addition I am not a statistical or trial expert so would recommend that reviewers with specific expertise in these areas are sought.

Abstract:

The brief description of the intervention, 'medical detailing intervention', is an unknown term to me; can the authors elaborate on or explain this term?

The term 'medical detailing' has been replaced with the more commonly used term 'academic detailing'. A brief background (with relevant references) to academic detailing has been added to the Background section.

In the outcome measures, can this be structured in to primary and secondary?

Outcome measures have now been presented as primary and secondary in both abstract and main text.

Background: there are some key global references that would be relevant to include either here and/or in the discussion such as the World Alz Association 2016 report on healthcare for people with dementia which emphasises the need for countries to move to a task-shifted, primary care model of dementia care as need Increases alongside finite/reducing resources.

Reference to the World Alzheimer's Report 2016, and their recommendations for a shift to primary care based model of dementia diagnosis and care has been included in the Background, as well as other references related to this important policy shift.

Methods

Intervention; I felt the current description whilst helpful, focused more on the process of diagnosis and would have liked more information on the education provided on the management of dementia and how this differed, if any from usual care.

The diagnosis and management aspects of the education were both based on the RACGP guidelines. Some additional details on the management aspects of the intervention have been included in the Methods.

Would the RACGP guidelines also be available to the control group? How did the educational intervention enhance this raising awareness or use of guidelines if at all?

The RACGP guidelines were available to all GPs via the RACGP website, though many may have been unaware of this. The following sentence has been added to the Methods to address the question of raising awareness of the guidelines in the intervention group:

"Intervention GPs were provided with a full copy of the RACGP Dementia Guidelines, as well as an A4-sized summary poster at the conclusion of the academic detailing visit."

“Waitlist GPs were mailed a written summary of their patient’s home assessment and the RACGP Dementia Guidelines after completion of the 12 month audit.”

Results: similarly in this section, the focus was largely on detection with little data provided in the section on 'evidence-based care' that appeared to be related to management rather than detection yet the former appears in the title and the aim of the study?

Evidence based care relates both to detection and to management: specifically both the GPAQ and the referrals relate to management (see answer below)

Discussion: here the authors state that by management they mean 'number of tests and referrals' ie purely linked to the diagnostic process. Clarity is needed through the paper as to whether 'management' refers just to pre-diagnoses or as I assumed post diagnosis care too.

The question about referrals encompassed both specialist referrals (for diagnosis and management) and referrals to care services, via the Aged Care Assessment Team system, as part of evidence based dementia management as opposed to diagnosis alone. In addition the GPAQ questionnaire measured process issues relating to management, including communication and enablement. We have searched the text and made clarifying changes to our use of the word “management”.

Comparison with other studies - ref 28 is highlighted in the introduction as a key study or importance built is not referred to at all here or how this study compares/contrasts with it.

We have removed this study, in which the intervention was aimed at consumers and not GPs directly. It failed to achieve any improvement in rates of diagnosis.

I would have liked to have seen more discussion about primary care led models and the future role in dementia care with reference to key global policy.

More discussion has been added in the first paragraph, as below:

“Primary care is “more local, more holistic and personalised, and more comprehensive, integrated and continuous” than secondary care (1), and thus better suited to dementia identification and management. Primary care is well placed to include better integration across primary health and social care services, with the World Health Organization (WHO) placing integrated care for the elderly in the centre of a new initiative (7).”

Reviewer 2: Doris S.F. Yu, The Chinese University of Hong Kong, Hong Kong

This is an interesting study to identify the effects of an intervention to enhance the diagnostic accuracy of GP on dementia. However, the conceptual clarity and study design of the study need more elaboration. Please see the specific comments listed below.

Background
The authors quote several studies on intervention to improve the detection and diagnosis of dementia. However, these studies, which can reflect an up-to-dated knowledge, are not adequately elaborated. [e.g. reference 28, what Is the meaning by delivering the educative intervention in coordinated dementia case management setting; For references 15 and 31, what were the barriers these studies targeted for.]. Without such information, it is difficult to interpret the originality of the current paper.

This section of the background has been re-written to provide a clearer description academic detailing as a strategy for GP education. We have emphasised that this type of educational strategy is based on an interactive and personalised approach, and can therefore target different barriers to diagnosis depending on the knowledge, confidence and attitudes of individual GPs, and barriers to management depending on their individual context. The barriers to GP diagnosis of dementia (discussed in some detail in the preceding paragraph)":

"include lack of knowledge [9] and/or confidence [16, 17, 20-23], …..difficult due to slow and fluctuating onset and overlap of symptoms with other diseases, lack of a definitive diagnostic test[22], ……perception of dementia diagnosis as a specialist domain [24]….no treatment to...reverse or halt the progress of these disorders,. We also mention that GPs may not conceptualise social and system support for ongoing cognitive decline as therapeutic; nihilism may also hinder management [25]."

Space precludes a more comprehensive literature review.

Study design
As the study intervention was targeted at improving the diagnosis of dementia by GP, it was unclear why the primary outcomes focused on patients' QOL and depression scores. This is less acceptability, especially when the intervention did not have explicit content to improve the dementia management. The intervention may be limited to use case studies to illustrate dementia management and provided the GP with a practice guideline. The dose of the intervention in this aspect is less coherent with the expected changes on patients' QoL and depression outcomes. The control group, indeed, also receive the practice guideline!!

Adherence to guidelines for detection and management of patients with dementia will be unlikely to result in improvement of cognitive function scores, but will enable access to interventions and support at the most appropriate time thereby maximising quality of life for both the person with dementia and their carer(s), through enablement for the patient and appropriate services to support the carer. The waitlist group received a copy of the RACGP Dementia Guidelines in the mail after 12 months (the reported endpoint).

Carers were a sample subgroup, why the consent is not directly obtained from the carers, but the patients they are taking care of?

Carers also consented to participate in the study. The following has been added to methods to clarify the process of carer recruitment: **"Eligible carers were provided with an information pack and a letter of invitation. Those who agreed to participate responded by returning a consent form to the local study site."List out what does the cut-off score of the CAMCOG-R <80 indicate for?**

Previously published papers have used a 79/80 cut-off as an indicator of dementia. The following statement has been added to address this concern:
"A cut point of 79/80 differentiates between those having dementia and those not having dementia with 93% sensitivity and 87% specificity (Huppert et al., 1996). For the purposes of this study a CAMCOG-R score of less than 80 was used as an indicator of dementia."

For the intervention, the second audit is less clear. What are the "results of nurse assessment" referring to? Does it mean the reassessment by the GPCOG? If yes, what is the purpose of giving the feedback at 12 months later.

The methods section has been modified to make it clearer that the first and second GP audits are the "before education" and "after education" measures of GP diagnosis rates and management practice:

"GPs completed an audit of their patients prior to the education, in order to obtain a baseline measure of their dementia diagnosis rates and management practices. The educational session that followed was conducted after completion of the baseline audit of patients by the GP, and included (i) instruction in the use…."

and

"GPs were not informed of the outcome of the research nurse assessment until after the 12 month audit, in order to determine the effectiveness of the educational intervention on GP diagnosis and management of patients over the 12 month study period. Following the second audit, GPs were provided with the results of the comprehensive nurse assessments conducted at baseline and 12 months, and offered an opportunity for self-reflection and discussion with their academic detailer."

Why an allocation ration of 1:2 was used?
This allocation ratio was used to address another of the main aims of the study. That is, to determine whether a screening or a case finding approach to dementia results in: i. better outcomes for people with dementia and their carers; ii. a more acceptable process for consumers, support people and GPs. The findings relating to this aspect of the study have published:

Mate, K. E., Magin, P. J., Brodaty, H., Stocks, N. P., Gunn, J., Disler, P. B., . . . Pond, C. D. (2017). An evaluation of the additional benefit of population screening for dementia beyond a passive case-finding approach. International Journal of Geriatric Psychiatry, 32(3), 316-323. doi:10.1002/gps.4466

There was no information on the psychometric properties of the outcome measures. Although most of them are well known, more information for GPAQ is needed. Besides, what is the meaning by the GP identification of reversible cause of dementia? What is the conceptual meaning by 'patient enablement"! in fact, what intervention has been delivered directly to the carers, so that their improved outcome can be explained as relating to the study intervention. In the Discussion section, there is one sentence explaining "the improved enablement is related more to communication and empathy characteristics of the GP". The mechanism under which is very unclear! Why and how?

The GPAQ was chosen as it is brief, acceptable to patients and demonstrated high reliability and validity (Mead et al., BMC Family Practice 2008, 9:13; Roland et al., BMC Family Practice 2013, 14:160). It has been extensively used in the UK (which has a similar health care system to Australia) and has also been used previously in Australia (Potiriadis et al., Med J Aust 2008; 189 (4): 215-219).

The concept of patient enablement has been explained briefly in the Methods section:

"….patient enablement following consultation with their GP (3 questions related to patients ability to understand and cope with their illness or problem)."

***Several examples of reversible causes of dementia have been added:* "….(e.g. depression, vitamin B12 deficiency, hypothyroidism, adverse drug reaction)"**

There was no direct intervention to carers (or people with dementia) provided by the study team, but as the intervention emphasised the importance of the carer, it may be assumed that the intervention affected the GPs' interaction with the carer in a positive way, as indicated by the GPAQ results. As mentioned in the discussion, we did not measure the process by which this occurred.

The CAMCOG-R seems to be used as a golden measure to examine the sensitivity and specificity of the GP diagnosis. This is problematic, as the CAMCOG-R is only a screening tool. How can the authors validate even the cause of dementia? What is the criteria measure for this?

We agree that clinical diagnosis of dementia requires a clinical decision-making process; the CAMCOG-R may aid diagnosis, but it alone can't provide a definitive clinical diagnosis. Geriatric-psychiatric consultations for every patient in this large community-based study (as for many research studies) were impractical but, given the validation studies of the CAMCOG-R, it is quite a robust and widely accepted means of classifying patients for research purposes.

Identification of the "cause" or sub-type of dementia was not relevant to this study.

Results:
The results about the positive effect of the study intervention on the dementia diagnostic accuracy by GP need to be interpreted with caution. This is because in this study, the patients were referred to the GP for dementia screening. This is very different from the actual practice that the GP need to suspect the client to have dementia based on initial health assessment, and able to use the relevant tool to do the screening.

We agree that screening for people aged 75 and over was a component of this study, and is different from the case finding approach outlined in this comment and advocated in guidelines. We have explored whether screening adds a significant benefit over and above case finding in another study based on the same data as this one, and found that the GPCOG in the case finding group had a higher positive predictive value than in the screening group, thus providing a more accurate result, and supporting the guideline recommendations regarding case finding. For details see: Mate, K. E., Magin, P. J., Brodaty, H., Stocks, N. P., Gunn, J., Disler, P. B., . . . Pond, C. D. (2017). An evaluation of the additional benefit of population screening for dementia beyond a passive case-finding approach. International Journal of Geriatric Psychiatry, 32(3), 316-323. doi:10.1002/gps.4466. Therefore we do not think that the guideline-recommended case finding approach would fail to respond to an academic detailing intervention such as the one in our study.

Discussion:
For the implication on practice/ policy, there should be more discussion about how the tested intervention can be translated to the actual practice. Not just to repeat the significance of the study intervention

Academic detailing is widely used by the pharmaceutical industry so it is not beyond the realms of possibility to consider that it might be duplicated as a GP education model as part of public policy (as outlined in our "Implications" section). This is especially so as primary care is now seen more and more as the place where dementia must be first identified, with large implications in terms of cost – savings and reduction in avoidable suffering if this is done.

Reviewer 3: Seyed-Mohammad Fereshtehnejad, Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada

1. In the Abstract, section 'Results', line 35: please note that the 95% confidence interval for a statistically significant OR with p-value<0.05, should be either entirely <1 or entirely >1. Here, the OR and P-value for the outcome 'satisfaction with GP communication' are reported as "p=0.024; odds ratio 2.1, 95% CI: 0.27-3.93". A 95% CI of 0.27-3.93 includes both protective (OR<1) and risk factor (OR>1) effects, which is not statistically significant and does not correspond with a p-value of 0.024.

Accordingly, the 'Conclusion' that the intervention has improved satisfaction with GP communication is not correct and should be revised.

The term "odds ratio" was used in error, rather than "mean difference". This has now been corrected in the Abstract and Results. The interpretation and conclusions made were correct, and therefore have been retained.

An Odds Ratio (treatment effect for a binary outcome) is statistically significant when the Confidence Interval does not include 1. The mean difference (treatment effect for a continuous outcome) is statistically significant when the Confidence Interval does not include 0. The GPAQ is a continuous outcome so the treatment effect 2.1 [0.27, 3.93] means that on average the Intervention group scored 2.1 points higher on GPAQ than the Control group. The 95% CI of 0.27 to 3.93 means that we are 95% confidence that the population treatment effect is as small as 0.27 points or as high as 3.93 points on the GPAQ scale.

2. In 'Methods', subsection 'Participants', it is mentioned that 'GPs sent letters of invitation to all patients who met the inclusion criteria'. Please clarify what was the response rate in this study? How did you evaluate the representativeness of the participants who answered positive for participation? Was there any significant difference in the characteristics of the respondents vs. non-respondents?

The response rate to the invitations sent out to patients by their GPs is shown in Figure 1 (2,030 responses from 10,683 invitations, which is approximately 19%). For ethical reasons, the study team did not have any contact with, or information about the patients that declined to be involved in the project. It was therefore not possible to compare the characteristics of respondents vs. non-respondents.

The 19% response is comparable to the modest response rate expected based on other studies in general practice. The prevalence of dementia in the participants recruited (8.5%) was close to the rate expected (Australian Institute of Health and Welfare 2012. Dementia in Australia. Cat. #70. Canberra:AIHW), supporting the external validity of the study.

3. In 'Methods', subsection 'Intervention', line 35: it is stated that the baseline audit of patients was done by the GPs in the intervention group. With this setting, the scorers, GPs in the intervention group, might have been biased during auditing. I wonder why blinding was not performed to prevent scorers' bias. In these trials, it is preferred that the assessments (here auditing) being performed by independent assessors who is blinded to the study arm.

The methods section has been modified to make it clearer that the first and second GP audits are the "before education" and "after education" measures of GP diagnosis rates and management practice in both the intervention and waitlist groups.

4. Likewise, it is important to clarify who performed the second audit after 12 months. Was it done by blinded independent assessors/interviewers?

The GPs completed a second audit after 12 months.

5. If several assessors were involved in the study, how inter-raters agreement was evaluated? Is there any data available on inter- and intra-raters agreement?

Inter-rater and intra-rater were not specifically examined in this study, given that it was a geographically dispersed multi-site study across three states.

The reliability of the GPCOG (including interrater, intraclass correlation and test-retest) have been examined in previous studies, and range from 0.75-0.87 (Brodaty et al., JAGS 2002, 50(3):

530-534). A comparison of the performance of the GPCOG and MMSE from this study has been published previously (Brodaty et al., Dementia and Geriatric Cognitive Disorders, 2016, 42(5-6), 323-330).

6. In the 'Methods', subsection 'Sample Size': I wonder why a 7% difference in the WHOQOL-BREF scale score has been used for sample size calculation. Is this considered as a small, medium, or large effect size? Is this a minimum clinically relevant difference on this scale? Please further clarify and/or cite to a relevant reference.

A minimal clinically important difference has not been established for the WHOQOL-BREF. A large study of "well" and "sick" populations, reported differences of 11.5%, 5.5%, 4% and 1.5% for the physical, psychological, social and environmental domains respectively (Skevington et al., Quality of Life Research, 2004, 13:299-310). Another study estimated a 'small' difference as 3.6, 3.5, 4 and 3.2 points on the physical, psychological, social and environment WHOQOL-BREF domains , respectively (Crocker et al., Journal of Clinical Epidemiology, 2015, 68(5): 584-595). A 7% difference in domain scores was therefore estimated as small to medium effect on quality of life, and used in sample size calculation.

7. In the 'Methods', subsection 'Sample Size': for sample size calculation, while authors assumed 15% for the number of drop-outs prior to study performance, the actual rate of drop-out patients after follow-up was larger (25% waitlist, 20% intervention arm). What was their strategy to deal with lowered statistical power? Was any post-study power analysis performed?

The description of the GP sample size calculation presented in the Method section has been revised, to address error and omissions. The sample size calculation was based on a prevalence of dementia of 10% in the 75+ Australian population. However, we observed a lower rate of dementia in our sample at baseline (43+124 out of 552+1478, which is 8.2%). The study did suffer from a lower sample to do with the dropouts, and this has been added as a limitation in the Discussion.

Post-study power analysis to compute "observed power" is non-informative. The p-value is the observed significance of the test that was performed therefore non-significant p-values always correspond to a test with low observed power by definition. Any calculation of observed power provides no more information than the p-value does; ref below.

Hoenig, J.M. & Heisey, D.M., The abuse of power, The American Statistician 2001 55:1, 19-24.

8. In 'Methods', subsection 'Statistical Analyses': authors have used regression-based generalized estimating equations (GEE) models to compare the outcome variables after 12 months follow-up between the two study arms considering the baseline measures as covariates. While this statistical approach is not inherently wrong, a better solution that I recommend is to use mixed effects models for repeated numeric outcomes. With mixed effects models, it is possible to consider outcome trajectory at individual level as a random factor in the model. In other words, mixed effects model has the advantage of taking into account each individual change over time, while the GEE models mostly fit the population average effects.

The claim that mixed models are better for multilevel studies (e.g., patients clustered within GP practice) is not justified; see reference below. In general, mixed models involve assumptions about the distribution of the random effects, and the standard errors are not robust to mis-specification of the model. On the other hand, the GEE provides a robust estimation of the standard errors assuming only that the number of clusters is sufficiently large, i.e., even if the

correlation model is mis-specified. (Here, we have 168 GP practices which is a large number of clusters.)

For linear outcomes, the population-average model estimated via GEE gives equivalent estimates to those of a random intercept mixed model. However, the two approaches differ in the interpretation of the effects. In the GEE, the treatment effect is the change in the outcome associated with being in the intervention group versus control group averaged across all clusters. Whereas in the mixed model, the treatment effect is the change in the outcome associated with being in the intervention group versus control group keeping the random effect (cluster) fixed.

In a cluster RCT our focus is on evaluating the intervention by estimating what the average effect of the treatment would be for a patient in the population, thus our choice of the population-average model estimated via GEE.

Hubbard, A.E. et al, To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health, Epidemiology 2010: 21, 467-474

9. Was any interim analysis planned prior to the trial implementation? This is an important issue, since if the efficacy of the intervention had been shown in the middle of the trial, the entire patients population could have benefited from receiving the intervention.

There was no interim analysis planned prior to the trial implementation. This was a large scale primary care intervention with baseline visits to patients and GPs occurring over several months, making interim analysis unfeasible.

10. Table 1: in the intervention arm, the sum of 671 (45%) males and 805 (55%) females does not correspond to the total number of 1478 participants on top of the column. A similar problem could be also found in rows 20-21 for the number of males and females in the GPs description in both study arms. Was there any other gender status? Please recheck the numbers and clarify the issue.

There were 2 patients and 11 GPs that did not disclose gender. A footnote has been added to Table 1 to this effect.

11. Table 2 summarizes baseline comparisons in the main outcome variables between the two study groups. While this is still a summary of cross-sectional baseline comparisons, I wonder why the p-values are from GEE models? As instructed in the 'Methods', the GEE models were applied to compare the effect of intervention after follow-up in the change occurred in the outcome variables from baseline. Please clarify this issue.

The p-value columns have been removed from the baseline tables (Table 1 and Table 2). As per CONSORT guidelines, testing for differences in baseline characteristics is not necessary when the groups have been randomised.

12. In the 'Results', subsection 'Outcome Measures for Patients', line 32: please refer to the same problem as comment NO.1. The 95% CI reported shows a NON-significant OR since it includes value 1, while the p-value is <0.05. Could it be possible that the range is indeed calculated for a mean difference of a numeric score and not an OR for a categorical scale? Please recheck the calculations and the corresponding scales.

See our response to the first reviewer comment in this section.

13. It is not clear what strategy the authors followed to deal with the drop-outs in this trial. Was any per-protocol or intention-to-treat analysis performed? How about missing data imputation to increase statistical power while encountering with the drop-outs.

Our analysis was intention-to-treat and patient data were analysed according to treatment allocation. No imputation was performed.

VERSION 2 – REVIEW

REVIEWER	Seyed-Mohammad Fereshtehnejad Department of Neurology and Neurosurgery, McGill University, Montreal, QC, Canada
REVIEW RETURNED	17-Jun-2018

GENERAL COMMENTS	The manuscript has been approved after revisions and I believe the statistical methods sound relevant and correct.