PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Two-year effectiveness of a stepped-care depression prevention intervention and predictors of incident depression in primary care patients with diabetes type 2 and/or coronary heart disease and subthreshold depression; data from the Step-Dep cluster randomized controlled trial
AUTHORS	Adriaanse, Marcel; Pols, Alide; Tulder, Maurits; Heymans, Martijn; Bosmans, J; van Dijk, Susan; Van Marwijk, H

VERSION 1 – REVIEW

REVIEWER	Irena Milaniak Andrzej Frycz Modrzewski Krakow Univeristy, Faculty of Medicine and Health Science
REVIEW RETURNED	29-Nov-2017

GENERAL COMMENTS	Below my comments: Design - there is no explantion about stepped-care intervention to prevent MDD and usual care. In what way patients were selected to the group by whom? Potential predictors are mixed with design of the study and outcome measure. The study protocol is not a part of manuscript for publication, so ie etical consideration should be included in the main text

REVIEWER	Norbert Hermanns Research Institute Diabetes (FIDAM)
REVIEW RETURNED	21-Dec-2017

GENERAL COMMENTS	The authors examined the potential of MDD prevention by a nurse-led prevention programme. The intervention was not able to prevent major depressive episodes in people with diabetes and CHD. The odds ratio for prevention of MDD by the intervention was 1.37 (95% CI 0.52; 3.55). Baseline levels of anxiety, depression, the presence of >3 chronic diseases and stressful life events predicted the incidence of MDD significantly. The study addresses an important question of diabetology – the prevention of depression in people with diabetes. The article is well written, the analysis is methodological sound and the limitations of the study are adequately addressed. There are only a few minor points. The consideration of these points might help to improve the article further. 1. It is surprising that the control group showed an initial improvement of depression scores by more than 30% three months after start of the study. How do the authors explain this surprising

	effect of care as usual, which has a similar magnitude than the nurse-led prevention programme? Not a lack in efficacy of the nurse-led prevention programme but the efficacy of “usual care” seems to be responsible for the non-significant finding. 2. Could a selective drop-out be responsible for the surprising improvement of the usual care group (only the improved participant responded to the follow-up assessment)? 3. Given the highly comparable depression scores after 12 months in both groups, why did the authors expect that an extension of the follow-up period would be able to achieve the primary outcome? 4. Was the extension of the follow-up period prespecified in the study protocol?

REVIEWER	Margaret Maxwell University of Stirling, UK
REVIEW RETURNED	22-Dec-2017

GENERAL COMMENTS	This paper has stemmed from a pre-existing RCT and its participants. It has extended the possibilities of the trial dataset in a novel way to answer a different question. The result is the identification of a set of risk factors for depression in those with T2DM and CHD. Although the increased risk of depression in these populations is well known, there is little knowledge to guide clinicians as to which patients may be more vulnerable (outside of generally known risk factors for depression). There is value for primary care clinicians in identifying these risk factors - to enable targeted assessment and intervention. The potential predictor variables were established with reference to existing literature, patient and practice nurse interviews and as readily accessible to primary care clinicians. This increases the practical application of the findings. The limitations of this approach have been recognised and appropriate external validation has been recommended. This makes the paper worthy of publication. The paper is well written and logical, with the author's assumptions made explicit throughout. I would expect that this paper will be of interest to many, even if the findings are based on a relatively small sample, and from an opportunistic population. I have not commented on the statistical methods used - this will require a statistician.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
COMMENT 1.1: Design - there is no explanation about stepped-care intervention to prevent MDD and usual care. In what way patients were selected to the group by whom?
REPLY 1.1: Details on the randomization of general practices is now added to the methods section. The allocation of the patients’ general practice determined whether they participated in the intervention or usual care group.
CHANGE 1.1: Methods, page 6, line 13 now also states: “A statistician blinded to the characteristics of the GP practices performed the randomization of GP practices using a computer generated list of random numbers. Randomization was done at the level of the GP practice, which corresponds to the participating practice nurse, to avoid contamination between the treatment groups, and was stratified for size (less or more than 5000 patients).”

COMMENT 1.2.: Potential predictors are mixed with design of the study and outcome measure.
REPLY 1.2: Depression severity and anxiety severity are potential predictors and the dichotomized depression severity score is used as the outcome measure. These potential predictors were selected as they have been identified as predictors in multiple other studies [1–7]. Similar approaches (using depression or anxiety severity as both potential predictors and (dichotomized) outcome measures) have been applied in comparable studies [1–4,6,7] and we were able to confirm the results of these other studies.
As we were not entirely sure that we understood what the reviewer meant with comment 1.2., we hope that our response has sufficiently clarified this question. If not, we would value a clarification of the comment and will happily address it in a subsequent version.

COMMENT 1.3: The study protocol is not a part of manuscript for publication, so ie ethical consideration should be included in the main text
REPLY 1.3: A statement on ethical considerations and approval has been added in the methods section.
CHANGE 1.3: Methods, page 6, line 13 now also reads: “The study was performed in accordance with the declaration of Helsinki (2008) and the Dutch Medical Research involving Human Subjects Act (WMO). The protocol was approved by the medical ethics committee of the VU University Medical Centre (NL39261.029.12, registration number 2012/223), and registered in the Dutch Trial Register (NTR3715 http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=3715).”

Reviewer 2:
COMMENT 2.1: It is surprising that the control group showed an initial improvement of depression scores by more than 30% three months after start of the study. How do the authors explain this surprising effect of care as usual, which has a similar magnitude than the nurse-led prevention program? Not a lack in efficacy of the nurse-led prevention program but the efficacy of “usual care” seems to be responsible for the non-significant finding.
REPLY 2.1: This drop in PHQ-9 scores between baseline and three months of follow-up in both groups was to some extent indeed surprising. It exceeded the expectations of spontaneous recovery alone[8]. ‘Subthreshold depression’ is a symptom-diagnosis and therefore fundamentally somewhat uncertain (i.e., it may pick up ‘benign’ context issues such as grief and stress as well). However, this observed improvement is not likely to be caused by a specific treatment either, since most of the participants did not receive any mental treatment during these three months. In the intervention condition, patients were offered watchful waiting during this period. Notifying general practitioners which participants met criteria for subthreshold depression is unlikely to have led to any treatment in this period in the usual care group, because screening for depression alone does not change the management of depression in primary care[9]. Additionally, the Dutch clinical guidelines advice an initial period of watchful waiting for subthreshold depression[10]. Perhaps the decrease in depressive symptoms is partly caused by attention, by regression to the mean or by patients' self-insight into their mental symptoms and problems.
We have added a summary of possible explanations of the statistically non-significant intervention effects at both 12 and 24 months (which justify using the Step-Dep RCT as a cohort to render our prediction model) in the discussion section, as well as a reference to facilitate the reader to find further details on the explanations hypothesized in our publication of the 12-months effects.
CHANGE 2.1: Discussion section on page 13, line 3 now reads: “In a previous publication we have hypothesized the causes for the lack of effect of the Step-Dep intervention as compared to care as usual in preventing incident MDD at 12 months of follow-up[20], which we assume also explain the lack of effect at 24 months of follow-up. In summary, a first explanation could be that subthreshold depression was potentially over-diagnosed in our population, whereas stepped-care may be more effective in patients with more severe symptoms[56]. Secondly, fewer patients than expected were

treated with the more intensive treatment steps. This was partly caused by the fact that a considerable proportion of patients did not want to start one or more of the treatment steps. This may indicate that our program did not sufficiently match their need for care. Furthermore, this was in part due to the low PHQ-9 scores of 6.7 on average at three months after baseline measurements, which made only a relatively small proportion of the patients eligible for more intensive treatment steps. The drop in PHQ-9 scores between baseline and three months of follow-up in both groups exceeded the expectations of spontaneous recovery alone[57]. It is unlikely that either of the groups received any specific treatment during this period. The Step-Dep program entailed an initial period of watchful waiting and Dutch primary care clinical guidelines recommend a similar waiting period before starting treatment for subthreshold depression[58]. Additionally, screening for depression alone does not change the management of depression in primary care[59]. We argue that the decrease in depressive symptoms may partly be caused by attention, regression to the mean, or patients' self-insight into their mental symptoms and problems. Finally, depressive and anxiety symptoms slightly improved over time in both groups, possibly indicating that usual care is already of reasonable quality and, therefore, the room for improvement for new interventions over usual care may be limited."

COMMENT 2.2: Could a selective drop-out be responsible for the surprising improvement of the usual care group (only the improved participant responded to the follow-up assessment)?
REPLY 2.2: The reviewer poses an interesting explanation for this improvement. However, we have compared the baseline characteristics of patients with missing data to those without. Patients with missing data were more often living alone (61% vs 41%), but no other differences between these groups were found (as stated in Results, page 10, line 17-21) which makes it unlikely to be the cause of the indicated improvements in PHQ-9 scores. Please also see REPLY 2.1 for our hypothesis for this improvement.
CHANGE 2.2: Please see CHANGE 2.1

COMMENT 2.3: Given the highly comparable depression scores after 12 months in both groups, why did the authors expect that an extension of the follow-up period would be able to achieve the primary outcome?
REPLY 2.3: We deemed it possible that the potential health benefits of participating in the stepped-care program would not yet be fully visible after one year and a potential delay in cumulative effects was presumed (as stated in Introduction page 4, line 48-50). The cumulative effects of the treatment could have become more visible after the treatment was finished. Therefore, we chose to perform a follow up study as described in our study protocol[11].

COMMENT 2.4: Was the extension of the follow-up period pre-specified in the study protocol?
REPLY 2.4: Yes, please see the METC protocol attached to the original submission on page 47 line 40-44.

Reviewer: 3
No changes requested.

REFERENCES
1. Bot M, Pouwer F, Ormel J, Slaets JPJ, de Jonge P. Predictors of incident major depression in diabetic outpatients with subthreshold depression. Diabet. Med. 2010;27:1295–301.
2. Katon W, Russo J, Lin E, Heckbert S, Ciechanowski P, Ludman E, et al. Depression and Diabetes: Factors Associated With Major Depression at Five-Year Follow-Up. Psychosomatics. 2011;50:570–9.
3. Pibernik-Okanovic M, Begic D, Peros K, Szabo S, Metelko Z. Psychosocial factors contributing to persistent depressive symptoms in type 2 diabetic patients: a Croatian survey from the European Depression in Diabetes Research Consortium. J. Diabetes Complications. 2008;22:246–53.
4. Ossola P, Paglia F, Pelosi A, De Panfilis C, Conte G, Tonna M, et al. Risk factors for incident depression in patients at first acute coronary syndrome. Psychiatry Res. 2015;228:448–53.

5. Kang H-J, Stewart R, Bae K-Y, Kim S-W, Shin I-S, Hong YJ, et al. Predictors of depressive disorder following acute coronary syndrome: Results from K-DEPACS and EsDEPACS. J. Affect. Disord. 2015;181:1–8.
6. Doyle F, McGee H, Delaney M, Motterlini N, Conroy R. Depressive vulnerabilities predict depression status and trajectories of depression over 1 year in persons with acute coronary syndrome. Gen. Hosp. Psychiatry. 2011;33:224–31.
7. Pedersen SS, Denollet J, van Gestel YRBM, Serruys PW, van Domburg RT. Clustering of psychosocial risk factors enhances the risk of depressive symptoms 12-months post percutaneous coronary intervention. Eur. J. Cardiovasc. Prev. Rehabil. 2008;15:203–9.
8. van 't Veer-Tazelaar PJ, van Marwijk HWJ, van Oppen P, van Hout HPJ, van der Horst HE, Cuijpers P, et al. Stepped-care prevention of anxiety and depression in late life: a randomized controlled trial. Arch. Gen. Psychiatry. 2009;66:297–304.
9. Gilbody S, Sheldon T, House A. Screening and case-finding instruments for depression: A meta-analysis. Cmaj. 2008;178:997–1003.
10. Depressie N. M44 NHG-Standaard Depressie. Huisarts&Wetenschap. 2012;55:252–9.
11. van Dijk SEM, Pols AD, Adriaanse MC, Bosmans JE, Elders PJM, van Marwijk HWJ, et al. Cost-effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: design of a cluster-randomized controlled trial. BMC Psychiatry. 2013;13:128.

VERSION 2 – REVIEW

REVIEWER	Norbert Hermanns Reasearch Institute of the Diabetes Academy Mergentheim, (FIDAM), Germany
REVIEW RETURNED	14-Mar-2018

GENERAL COMMENTS	The authors did an important study analyzing the impact of a step care model on the incidence of depression in type 2 diabetes. The could not observe a significant impact of the step care model on the incidence of depression, but they observed 4 predictors (anxiety and depression level, stressful events and multi-morbidity as predictors for incident depression. The article is well written. The analysis is methodological sound. I do not have major concerns regarding the article. Minor: How do the authors explain that the PHQ9 showed a remarkable drop (by 30%) between baseline and t3, but not the HADS-D (only 12%)? Has the different sensitivity of these instruments implications for the intervention algorithm of the Step care approach? Could the lacking efficiency of the intervention to lower Anxiety and depression 8at least measured by the HADS-D be responsible for the missing effect on incident depression, since depression and anxiety were predictors of incident depression

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2 (Minor)

COMMENT 1: How do the authors explain that the PHQ9 showed a remarkable drop (by 30%) between baseline and t3, but not the HADS-D (only 12%)?

REPLY 1: We can only speculate about this difference in drop between PHQ9 and HADS-D at t3. Currently we have no solid explanation for this difference. We assume that this is a statistical artifact. The PHQ9 is made to align with DSM diagnostic symptoms of depression irrespective of the co-morbid presence of physical conditions while the HADS-D should be robust for physical illnesses and perhaps measures a broader construct (for instance, 'I can laugh and see the funny side of things'). As in our experience, both scales have similar patterns. Additionally, a recent systematic review showed that the PHQ9 has better psychometric properties than the HADS-D in measuring depression in patients with diabetes (Quality of Life Research https://doi.org/10.1007/s11136-018-1782-y).

COMMENT 2: Has the different sensitivity of these instruments implications for the intervention algorithm of the Step care approach?

REPLY 2: We do think that the different sensitivity of these instruments have minimal implications, if at all, fort the intervention algorithm of the Step care approach. In the StepDep effectiveness study (PLoS ONE 12(8): e0181023.) we used the MINI, the PHQ9, the HADS-D and HADS-A to look at the differences in incident major depression and depression and anxiety levels respectively. All instruments used are valid and reliable. We found no statistically significant differences at any time point nor a statistically significant difference in the course of incident MDD or depression and anxiety symptom levels over time between the groups. In other words, the slope of the different outcomes over time were virtually the same.

COMMENT 3: Could the lacking efficiency of the intervention to lower Anxiety and depression, at least measured by the HADS-D, be responsible for the missing effect on incident depression, since depression and anxiety were predictors of incident depression.

REPLY 3: This is an interesting thought. The key reason behind our results we feel is the unexpectedly low incidence of depression. Our focus on depression (where earlier studies encompassed anxiety) might have contributed to this (we would otherwise have seen more cases). We do not think it is likely that the lack in efficiency of the intervention was due to lower anxiety and depression, measured by the HADS-D, is responsible for the missing effect on incident depression. See also REPLY 2.

VERSION 3 – REVIEW

REVIEWER	Norbert Hermanns Research Institute of the Diabetes Academy Bad Mergentheim (FIDAM), Germany
REVIEW RETURNED	08-Jun-2018

GENERAL COMMENTS	The authors addressed all my minor concerns satisfactorily.

VERSION 3 – AUTHOR RESPONSE

Reviewer: 2 (Minor)

COMMENT 1: How do the authors explain that the PHQ9 showed a remarkable drop (by 30%) between baseline and t3, but not the HADS-D (only 12%)?
REPLY 1: We can only speculate about this difference in drop between PHQ9 and HADS-D at t3. Currently we have no solid explanation for this difference. We assume that this is a statistical artifact. The PHQ9 is made to align with DSM diagnostic symptoms of depression irrespective of the co-morbid presence of physical conditions while the HADS-D should be robust for physical illnesses and perhaps measures a broader construct (for instance, 'I can laugh and see the funny side of things'). As in our experience, both scales have similar patterns. Additionally, a recent systematic review showed that the PHQ9 has better psychometric properties than the HADS-D in measuring depression in patients with diabetes (Quality of Life Research https://doi.org/10.1007/s11136-018-1782-y).
CHANGES TO THE TEXT 1:
We addressed this issue in the discussion section on page 14 and 15. We added the following text: "We observed a remarkable drop between baseline and three months in the PHQ-9, but not for the HADS-D. We can only speculate about this difference in drop between PHQ9 and HADS-D at three months. Currently we have no solid explanation for this difference. There is a possibility of a statistical artefact. The PHQ9 is made to align with DSM diagnostic symptoms of depression irrespective of the co-morbid presence of physical conditions while the HADS-D should be robust for physical illnesses and perhaps measures a broader construct (for instance, 'I can laugh and see the funny side of things'). We do think that the different sensitivity of these instruments have minimal implications, if at all, fort the intervention algorithm of the Step care approach. In the StepDep effectiveness study [66] we used the MINI, the PHQ9, the HADS-D and HADS-A to look at the differences in incident major depression and depression and anxiety levels respectively. All instruments used are valid and reliable. We found no statistically significant differences at any time point nor a statistically significant difference in the course of incident MDD or depression and anxiety symptom levels over time between the groups. In other words, the slope of the different outcomes over time were virtually the same."

COMMENT 2: Has the different sensitivity of these instruments implications for the intervention algorithm of the Step care approach?
REPLY 2: We do think that the different sensitivity of these instruments have minimal implications, if at all, fort the intervention algorithm of the Step care approach. In the StepDep effectiveness study (PLoS ONE 12(8): e0181023.) we used the MINI, the PHQ9, the HADS-D and HADS-A to look at the differences in incident major depression and depression and anxiety levels respectively. All instruments used are valid and reliable. We found no statistically significant differences at any time point nor a statistically significant difference in the course of incident MDD or depression and anxiety symptom levels over time between the groups. In other words, the slope of the different outcomes over time were virtually the same.
CHANGES TO THE TEXT 2:
We addressed this issue in the discussion section on page 14 and 15. We added the following text: "We observed a remarkable drop between baseline and three months in the PHQ-9, but not for the HADS-D. We can only speculate about this difference in drop between PHQ9 and HADS-D at three months. Currently we have no solid explanation for this difference. There is a possibility of a statistical artefact. The PHQ9 is made to align with DSM diagnostic symptoms of depression irrespective of the co-morbid presence of physical conditions while the HADS-D should be robust for physical illnesses and perhaps measures a broader construct (for instance, 'I can laugh and see the funny side of things'). We do think that the different sensitivity of these instruments have minimal implications, if at all, fort the intervention algorithm of the Step care approach. In the StepDep effectiveness study [66] we used the MINI, the PHQ9, the HADS-D and HADS-A to look at the differences in incident major depression and depression and anxiety levels respectively. All instruments used are valid and reliable. We found no statistically significant differences at any time point nor a statistically significant difference in the course of incident MDD or depression and anxiety symptom levels over time

between the groups. In other words, the slope of the different outcomes over time were virtually the same."

As a consequence a new reference [66] was added to the reference list on page 23:
66. Pols AD, van Dijk SE, Bosmans JE, Hoekstra T, van Marwijk HWJ, van Tulder MW, Adriaanse MC. Effectiveness of a stepped-care intervention to prevent major depression in patients with type 2 diabetes mellitus and/or coronary heart disease and subthreshold depression: A pragmatic cluster randomized controlled trial. PLoS One. 2017;12(8):e0181023

COMMENT 3: Could the lacking efficiency of the intervention to lower Anxiety and depression, at least measured by the HADS-D, be responsible for the missing effect on incident depression, since depression and anxiety were predictors of incident depression.
REPLY 3: This is an interesting thought. The key reason behind our results we feel is the unexpectedly low incidence of depression. Our focus on depression (where earlier studies encompassed anxiety) might have contributed to this (we would otherwise have seen more cases). We do not think it is likely that the lack in efficiency of the intervention was due to lower anxiety and depression, measured by the HADS-D, is responsible for the missing effect on incident depression.
CHANGES TO THE TEXT 3:
We addressed this issue in reply 1, reply 2 and additional text in the manuscript on pages 14 and 15.

Comment Associate Editor:
COMMENT 4: This wasn't a prespecified outcome of this trial (they need to make that clear). They have good follow up and results are interesting. Our statistician (Hermanns) is happy with the analysis and their response. However I don't see a marked up version of the paper where they have addressed his comments. Other readers may have the same questions. I may have missed it – but can they respond to his questions in the manuscript too.
REPLY 4: We agree that outcomes of the two-year effectiveness of the stepped-care depression intervention and predictors were not pre-specified.
CHANGES TO THE TEXT 4a:
We addressed this issue in the Methods/design section on page 6. We added the following details: "The outcomes of the two-year effectiveness of the Step-Dep study and predictors of incident depression were not pre-specified in designing the study."

CHANGES TO THE TEXT 4b:
We now uploaded a marked up version of the paper where we addressed the comments of reviewer Hermanns in the manuscript to using the track-changes mode in Word.