PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (http://bmjopen.bmj.com/site/about/resources/checklist.pdf) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

<table>
<thead>
<tr>
<th>TITLE (PROVISIONAL)</th>
<th>Lessons learnt from a cluster-randomised trial evaluating the effectiveness of Self-Management Support (SMS) delivered by practice nurses in routine diabetes care</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUTHORS</td>
<td>van Dijk - de Vries, Anneke; van Bokhoven, Loes; Winkens, Bjorn; Terluin, Berend; Knottnerus, André; van der Weijden, Trudy; van Eijk, Jacques</td>
</tr>
</tbody>
</table>

VERSION 1 - REVIEW

<table>
<thead>
<tr>
<th>REVIEWER</th>
<th>Christopher P. Morley, PhD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SUNY Upstate Medical University</td>
</tr>
<tr>
<td></td>
<td>Syracuse, NY</td>
</tr>
<tr>
<td></td>
<td>United States of America</td>
</tr>
<tr>
<td>REVIEW RETURNED</td>
<td>12-Dec-2014</td>
</tr>
</tbody>
</table>

GENERAL COMMENTS

There are a few major issues with this manuscript that I would hope the authors could address in a revision. I suspect there is some utility in reporting the outcomes of this project; however, it is tough to recommend publication in present form. It was also tough to review, because I don't think the manuscript was written clearly enough (to be clear, it is not an English problem, which is excellent; but rather one of organization). I am making the following comments, therefore, with the first comment serving as a caveat to the rest.

1. As a clustered trial with a complex set of inclusion criteria and a complex, social intervention, it is vitally important that everything in the paper be 100% clear. I suspect I have enough of a grasp of what the investigators did at this point, but only after several readings of the manuscript, all of which required lots of cross-checking across sections to make sure I hadn't lost track of a procedure. I don't have a specific list of suggestions, as I think there is a more fundamental issue: I would recommend that the authors have a fresh set of eyes read through this manuscript, and identify all points of confusion; I think several iterative passes through the manuscript to make sure everything is crystal-clear would be of great benefit.

2. The main goal of the study was to evaluate SMS as an intervention; however, there are three main points by which this may have failed. The first, which I will address in this comment, is one that the authors recognize and discuss. That is, PN-led detection seems to have not worked out very well. It seems to me that this, actually, is what the paper should be about. The title implies that the reader will hear about outcomes from a trial of SMS. However, the "outcomes" don't tell us much. In this paper, it is actually the "lessons learnt" that are far more valuable. Structuring the paper so that the discrepancy between the research and PN detection...
processes is front-and-center might actually make a better paper.

3. A second issue may lie in the initial screening sent out to patients. The authors do not address non-respondent bias except to say that they could not assess it in this study. However, it is entirely possible that those left out of the study were those most in distress. Those in most distress, for example, could have had the most trouble or the least interest in completing lots of extra forms. It is unfortunately these potential subjects for whom SMS would have had the most profound effect, and who may ALSO have been the easiest to detect by practice nurses. This seems like a major issue that is largely ignored in the current manuscript.

4. Additionally, the authors note that the ultimate study sample was probably underpowered. There is not much to do about this, except to further note that this paper should not really be about SMS outcomes (as the outcomes are not valid), but rather about the lessons learned.

I have a few other comments and thoughts that the authors may consider. Some may not be workable for this paper, but I’ll state them here in case they are useful in follow-up studies.

- It would be nice to see p-values in Table 1, to explore differences between intervention and control groups. If the study were unbalanced, this might be another reason the SMS appeared to fail. The authors mention there were no differences (except on the PHI index), but it would be nice to see this formally stated with t-tests.

- The PHI seems to have not budged in the intervention group, but does seem to have moved upward (worse) in the control group. Any comment on this?

- The authors make the point that the nurses seem to have been largely compliant with the research protocol. However, if this is a Practice Nurse-led intervention, but, as the authors state, the incentive for participation went to the practice organization, this could mean that the individual nurses were not properly incentivized to participate (and in fact, if the practice, i.e. the GP, is seen as benefiting for their extra work, this could be a DISincentive). This is worth some reflection and discussion, in my view.

- Finally, it seems like the study could have been handled through a cut-off design, using a regression-discontinuity analysis (see: http://www.socialresearchmethods.net/kb/statrd.php for example), rather than through cluster randomization. The patient N would have been higher. I mention this because it might be useful for future consideration. I also mention it in case the data are amenable to reanalysis using a regression-discontinuity model.

REVIEWER

Lamers, Femke
VUMc, Amsterdam, the Netherlands

I was involved in a previous RCT on which the SMS intervention was based. I nevertheless feel capable to provide a unbiased review as I was not in any way involved in the design and execution of this trial.
### GENERAL COMMENTS

This paper describes the results from a cluster-randomized controlled trial within a hybrid effectiveness implementation study to evaluate the effectiveness of a self-management support intervention implemented in the daily routine diabetes care of practice nurses. Forty primary care practice nurses participated in the study and were randomized. All patients with DM type II under the care of the PNs were invited to fill out a screening questionnaire; those with high emotional distress were invited to participate in the trial. PNs in the intervention arm implemented the detection method and SMS in all consultations with all DM patients. The control arm PNs provided usual care. PNs did not know which of their patients participated in the trial. The statistical analysis are appropriate. Loss to follow-up was limited, but major downfall of the study was the low number of patients in the intervention arm reporting high emotional distress in the nurse-led detection, leading to a very low exposure to the SMS intervention. As a result, the analyses of outcomes becomes somewhat of a formality; as one would expect, no differences between arms at 12 month follow-up were detected.

Some thoughts and comments on the study are listed below.

1. I’m not sure if I understand correctly what the implementation part and what the effectiveness part of the study is. From what I understand it is as follows: The content of the SMS intervention was the same as the intervention in the DELTA study, only the place of administration (clinic vs patient’s home) and administrator (practice nurse vs research nurse) was different. This would be the implementation part? The different case finding procedure - using the nurse-led detection - led to the evaluation of detection+SMS in the evaluation part of the study? If I got it all wrong, then some introduction on hybrid effectiveness implementation would be helpful.

2. Some of the differences between research-led identification and nurse-led identification could have been partially caused by waxing and waning of symptoms, as the authors mention in their discussion. But the discrepancy is so large that other factors are likely involved as well. The potential difficulty of people to express their emotional problems could play a much larger role when questions are asked face-to-face rather than in paper-and-pencil form, as is mentioned as well in the discussion. So the big question is: Knowing that patients might be less inclined to mention emotional problems than on a paper questionnaire, shouldn’t the nurse-led identification strategy have been tested before being implemented in this study. In retrospect, was not piloting the procedure a mistake? What recommendations would you have for future implementation studies that are changing the identification procedures? Please expand on this in the discussion.

3. On page 20 (Discussion) the authors mention they considered another screening procedure in the design phase as well. Did the authors consider having DM patients fill out the forms followed by a discussion of responses with the PN as alternative to face-to-face questions? Do the authors feel this is a feasible alternative?

4. At 4 months there are differences between groups on two autonomy subscales (that disappear at 12 months). What is the authors’ take on this. To what extent could this be a chance finding? Could it have been an effect of attention being given to psychosocial wellbeing? Is it relevant at all, given the relatively small difference?
Minor comments:
1. Abstract: Without reading the Methods of the paper it is not clear what is meant by 'research driven'.
2. Please add a description of the control arm condition in the Methods section.
3. Please provide directions of effects in the Results section (i.e. 'autonomy was higher in the intervention group than in the control groups' rather than 'there was a difference').
4. Consort checklist item 3b; weren't the PNs instructed to use the cut-offs of DFT and DS less rigidly?

Reviewer: Jason Oke
Nuffield Department of Primary Care Health Sciences

Review returned: 23-Feb-2015

General comments:
3. Is the study design appropriate to answer the research question?

No. Although, the study could have worked it didn't because the design did not take into account the inter-individual variation in DFT and/or regression to the mean from the screening phase round which subsequently undermined the findings from this trial. This is however adequately addressed in the title and the whole manuscript.

4. Are the methods described sufficiently to allow the study to be repeated?

A: Incomplete – There is not enough information to replicate this study as it currently stands. For example, there is not enough information about the Daily Functioning Thermometer, we know it is a visual analogue scale, but what does it look like exactly (is the scale linear?), why was favoured over other methods (e.g. likert scales) and has anyone has tested how reproducible it is (see above)? It is important because this is the how the primary outcomes is assessed. In addition, I could not find sufficient detail in this paper as to what SMS consisted of. This trial could not be repeated with the detail given in this manuscript. Could some of this information be given as appendices to the paper?

7. If statistics are used are they appropriate and described fully?

A: Yes, but...... The analysis takes into account clustering and correlation between practice nurses and practices using a logistic three level model. I would normally favour a GEE approach for structured data with a binary outcome as they represent population average effects rather than cluster specific effects. Moreover, the GEE estimates are equal or less than the CS estimates depending on the size of the cluster variance (see Ten-have et a 2004, Deviations from the population-averaged versus cluster-specific relationship for clustered binary data). This is complicated by the fact that the authors have used three levels, and as I do not know if this can be accommodated with GEE I would probably ignore the third level and use GEE or at least check sensitivity to model choice. In short, can the authors assure me or the reader that the effect estimates for the main outcome are not dependent on their choice of model and do not differ significantly from other methods of analyses - e.g. I note that the raw unadjusted OR's are much closer to 1 than...
Also on the sample size: in the manuscript it is stated that 232 patients were required but in the protocol it says it should be at least 331 to take into account loss to follow up. As it happens the loss to follow up is much less than they predicted and the sample size at 4 and 12 months only just fails to reach the required minimum sample size and hence does not weaken the power significantly (which is high to start with ~90%). This slight reduction in numbers should be noted and I think the writing in the manuscript could be clearer so that the reader is not left to work this out for themselves.

15. Is the standard of written English acceptable for publication?

A: Yes. Although, please do not use the term diabetics (page 4, line 11 and 39)

VERSION 1 – AUTHOR RESPONSE

Reviewer 1. Christopher P. Morley, PhD

There are a few major issues with this manuscript that I would hope the authors could address in a revision. I suspect there is some utility in reporting the outcomes of this project; however, it is tough to recommend publication in present form. It was also tough to review, because I don't think the manuscript was written clearly enough (to be clear, it is not an English problem, which is excellent; but rather one of organization). I am making the following comments, therefore, with the first comment serving as a caveat to the rest.

1. As a clustered trial with a complex set of inclusion criteria and a complex, social intervention, it is vitally important that everything in the paper be 100% clear. I suspect I have enough of a grasp of what the investigators did at this point, but only after several readings of the manuscript, all of which required lots of cross-checking across sections to make sure I hadn't lost track of a procedure. I don't have a specific list of suggestions, as I think there is a more fundamental issue: I would recommend that the authors have a fresh set of eyes read through this manuscript, and identify all points of confusion; I think several iterative passes through the manuscript to make sure everything is crystal-clear would be of great benefit.

An independent senior researcher of our department read through the paper with a fresh set of eyes. Based on his feedback, we have tried to make it easier for the reader to understand the content, evaluation and methodological challenges in the SMS-study.

In the introduction, we have put more emphasis on the following issues:

- Our starting point: an evidence-based treatment, and stakeholders who intended to implement this intervention in the routine care setting
- The implementation strategies that were needed for implementation of the original evidence-based intervention
- The content of the adjusted intervention ‘SMS’, which included the treatment phase that was found to be effective in the former trial, added by an integrated detection phase
- The evaluation: why it was needed to perform an effectiveness trial next to an evaluation of the implementation

The bold text in the following paragraph of the introduction has been added (start at p4, line 20):

In collaboration with a regional care group of general practitioners (GPs) in the Netherlands and a health insurer, the SMS (‘Self-Management Support’) implementation project was initiated to realise a shift from biomedically oriented care towards a biopsychosocial approach in diabetes care.[16] The starting point was a nurse-led minimal psychological intervention. A previous randomised trial has shown that this intervention was cost-effective for diabetes patients with minor to moderate depression: 9 months after receiving the intervention, depressive symptoms
were significantly lower, there was a positive effect on patients' quality of life, and patients experienced less anxiety, possessed more self-efficacy skills, demonstrated better glycaemic control and showed more participation in comparison to control patients. [17-19] There was an implementation momentum. The health insurer promised to pay the costs for the extra care, and the care group gave commitment to take care for training facilities and integration of SMS parameters into Electronic Medical Records. Some adjustments to the nurse-led intervention were crucial. Where the original intervention had been delivered at patients' homes by specifically trained nurses from the research team who selected eligible patients by means of an elaborate diagnostic procedure, in SMS it was provided by PNs as structural part of their consultations delivered in the family practice. For the identification of eligible patients a simple detection method became an inherent part of SMS. The eligibility criteria changed from having a mild to moderate depression towards suffering from both emotional distress and interference in their daily functioning due to the burden of diabetes. This focus on daily functioning instead of a diagnosis of depression was supposed to fit the primary care setting. Patients who were detected by the PN received the nurse-led minimal psychological intervention or were referred to the GP to see whether more specialised care would be required, depending on the severity of the symptoms.

SMS was evaluated by means of a type 2 hybrid effectiveness-implementation study design, in which the regional implementation strategies and the effectiveness of SMS were evaluated simultaneously. [20] This paper will focus on the effectiveness part of the SMS project though it is closely connected to the implementation part (see appendix 1).

In the methods, we have tried to explain our methodological choices in more detail.

- Regarding the recruitment and randomisation of practice nurses.
  It was mentioned in a subordinate clause that "we assigned PNs to study arms assuming an allocation ratio of 1:1, after stratification into PNs working alone in a practice, working in a team and working in different practices". To highlight the variety between the PNs in the region in which SMS was implemented, it has been reformulated and replaced (see page 6, line 15): Because the variety of PNs working solo versus in a team, and in one versus more practices, PNs were stratified into working alone in a practice, working in a team and working in different settings. The randomisation was performed by an independent research assistant who used a random number seed computer program to assign PNs to study arms assuming an allocation ratio of 1:1, after stratification into PNs working alone in a practice, working in a team and working in different practices.

- About the selection of patients: the necessity to insert the research-driven self-administered screening and how we performed it. In the former manuscript we stated: "we recruited eligible patients for the trial by means of a self-administered screening questionnaire to select patients in the intervention and control arms in a similar way". Now we have given more explanation about our choices on page 6, start at line 26:

In the selection of patients for the effectiveness trial, the following issues were crucial. First, we wanted to select patients who would receive the complete intervention including detection and treatment. Furthermore, patients from the intervention and control arms should be selected in a similar way, and without interfering in their routine care. We set up a screening procedure by means of written questionnaires sent to patients’ home addresses to identify patients with actual problems of daily functioning and emotional distress. Patients with a clinically established diagnosis of type 2 diabetes mellitus were sent a letter by their GPs to introduce the SMS-project. The enclosed self-administered questionnaire consisted of the screening instruments that intervention PNs also would apply for SMS in their routine practice. It included…(..)

(..) Page 7, Line 16:
The time interval between the research-driven self-administered screening questionnaire and the face-to-face nurse-led detection procedure in the consultation room needed to be as short as possible. However, intervention patients should
complete the baseline measurement before SMS was applied. Because of the logistics, patients received the posted screening questionnaire four to six weeks before a planned diabetes consultation.

- We have changed the text about how the PNs applied the screening instruments in their care practice. As it was confusing that we mentioned the blinding of PNs in the paragraph of study participants, it has been moved to the paragraph about SMS in routine care. We also have added some information about the Daily Functioning Thermometer (DFT). The information about the cut-off scores of the DFT has been moved to the paragraph about the self-administered screening, see reviewer 3, comment 2. The changes can be found at p8, line 6-15:

After the training sessions, PNs started to integrate SMS into their routine care practice. PNs were blinded regarding the outcomes of the recruitment procedure and study participation of their patients. They applied SMS in all their consultations with diabetes patients. PNs were instructed to explore whether patients experienced problems in daily life. They verbally asked patients to rate the Daily Functioning Thermometer between zero (no burden of diabetes in daily life at all) and ten (extreme burden of diabetes in daily life). A score of 4 was chosen to differentiate between patients who could benefit from self-management support (score > 4) and patients who had found a satisfactory way to live with the consequences of their diabetes (score ≤ 4). PNs also verbally administered the three questions of the Distress Screener.

In the results section, we further emphasize that only a few patients who were eligible for SMS according to the research-driven posted questionnaire were detected by the practice nurses during routine consultations. See our response on the next comment of reviewer 1.

2. The main goal of the study was to evaluate SMS as an intervention; however, there are three main points by which this may have failed.

2a) The first, which I will address in this comment, is one that the authors recognize and discuss. That is, PN-led detection seems to have not worked out very well. It seems to me that this, actually, is what the paper should be about. The title implies that the reader will hear about outcomes from a trial of SMS. However, the "outcomes" don't tell us much. In this paper, it is actually the "lessons learnt" that are far more valuable. Structuring the paper so that the discrepancy between the research and PN detection processes is front-and-center might actually make a better paper.

The reviewer recommends to structure the paper in a different way to highlight the lessons learnt about the discrepancy between the screening for research purposes and the nurse-led detection in routine care. We made some adjustments in the abstract, results and title of the manuscript to put the focus of the reader on the discrepancy and lessons learnt in our study rather than on the outcomes of the trial.

In the abstract on page 2, we put more emphasis on the discrepancy between the screening outcomes and detection in routine practice:

Results. Only 16 of the 117 patients in the intervention arm (14%) who were found eligible by a posted research-driven questionnaire were detected by their practice nurses. The self-management support was delivered to only 11 study participants. In the control arm, usual care was provided to 147 patients. Multilevel analyses showed no significant differences in outcomes between intervention and control arm.

The first sentence of the result section (page 15, line 16), in which participants’ exposure to SMS in routine care practice is reported, has been changed. Instead of the neutral statement: ‘The exposure to SMS of intervention patients varied’, we start with a focus on the discrepancy by saying):

“A unexpected outcome was the very low exposure to SMS of the study participants in the intervention arm.”

To make clear that the low detection rates negatively influenced the exposure to the complete intervention, we also added the following sentence (p15, line 26):

“The low detection rates also implied low exposure to the treatment phase of SMS”.


To put the focus on the lessons learnt about the study rather than on the outcomes of SMS for patients’ daily functioning, we have changed the title of the manuscript into:

**Outcomes of a cluster-randomised trial to demonstrate the effectiveness of biopsychosocial Self-Management Support (SMS) integrated into primary diabetes care: lessons learnt from a hybrid effectiveness-implementation study design**

**Lessons learnt from a cluster-randomised trial evaluating the effectiveness of Self-Management Support (SMS) delivered by practice nurses in routine diabetes care**

We discussed whether or not we should change the objective in the abstract. Our study was performed to evaluate the effectiveness of SMS delivered by practice nurses in routine diabetes care. Therefore, our objective was to report these outcomes. However, as the manuscript also is about the lessons learnt, we could also imagine that this need to be stated in the abstract. Then the objective would be: “to report the lessons learnt about evaluating the effectiveness…..”. We would ask the Editor for his preferences in this regard.

This also applied to the end of the introduction (p5, line 16). It ends up with the following statement.

This paper reports the effectiveness of SMS integrated by PNs into routine diabetes consultations regarding patients’ daily functioning, emotional health, quality of life, autonomy and participation, self-efficacy and self-management skills, and blood glucose levels. If needed, this could be either changed or added by “reports the lessons learnt”.

The discussion (has been rewritten with a focus on the problems with the detection of eligible patients. See page 20, line 7-28:

Our study evaluated the effectiveness of detection followed by self-management support as delivered by PNs to diabetes patients with emotional distress and reduced daily functioning. A critical issue was the limited exposure of intervention patients to this whole SMS-approach. The critical issue of the implementation of SMS was the addition of a detection procedure as inherent and integrated part of SMS. PNs were trained to identify eligible patients for the self-management support by applying screening questions during diabetes consultations. The DFT and DS were considered to be simple, patient-centred indicators of patients’ daily functioning and their emotional distress. The 4DSQ would enable the PNs and GPs to get more insight into the presence and severity of the emotional problems. For the effectiveness trial, a self-administered postal screening procedure was inserted into the study design to select patients from both the intervention and control arms similarly. To perform a pragmatic trial, we minimized the interference of the researchers in clinical practice as much as possible. The outcomes of DFT and DS on the research-driven screening were consistent with our assumptions based on the literature that approximately 20% of the diabetic population would be eligible for the treatment phase of SMS. However, these efforts to select patients with emotional problems aimed at increasing the contrast between patients in the intervention and control arms appeared to be useless as the added detection method in routine practice did not function as expected. This was not a problem of PNs’ adherence to the SMS-protocol, as the screening tools were integrated into diabetes consultations of more than 85% of the study participants. The problem was that the majority of study participants that scored above the cut-off value on the postal questionnaire did not meet the required detection criteria when screened face-to-face by the PNs. These patients were therefore not exposed to the self-management support.

Page 20, line 32:

Differences between the self-administered screening outcomes and the nurse-led detection could be explained by fluctuation in distress symptoms and the phenomenon of regression towards the mean. This may have resulted in study participants scoring less extreme values at nurse-led detection compared to values at research-driven screening. Furthermore, tests may function differently across settings and administration methods. Due to patients’ difficulty expressing emotional problems during consultations and lack of recognition by health professionals of emotional problems experienced by patients with chronic physical illness, the chance of positive screening outcomes seems to have been higher in anonymous research-driven screening than in clinical practice. Nonetheless, we did not expect such a large discrepancy between research-driven screening and nurse-led detection. The outcomes of 4DSQ (data not shown) suggested room for improvement in patients’ emotional functioning since two thirds of the intervention arm had moderate to severe symptoms of distress or somatisation.
months after the baseline measurement. These patients could have benefited from receiving self-management support.

(The comment about the outcomes of the 4DSQ (data not shown) has been deleted to limit the word count.)

2b) A second issue may lie in the initial screening sent out to patients. The authors do not address non-respondent bias except to say that they could not assess it in this study. However, it is entirely possible that those left out of the study were those most in distress. Those in most distress, for example, could have had the most trouble or the least interest in completing lots of extra forms. It is unfortunately these potential subjects for whom SMS would have had the most profound effect, and who may ALSO have been the easiest to detect by practice nurses. This seems like a major issue that is largely ignored in the current manuscript.

The reviewer questioned to what extent the non-respondents on the screening would have had the most emotional problems, and may have been the easiest to be detected by the practice nurse. As practice nurses were informed to apply SMS in all consultations, and didn’t know which patients were selected by the research-driven screening and gave informed consent to participate in the trial, we could use a registration database of outcomes of SMS of all diabetes patients in the intervention practices (anonymous data). The registration data in this database do not suggest a non-respondent bias. Overall, the detection rates were very low. Among 2013 patients, only 81 patients (4%) scored positive (DFT>4 and DS>3) during at least one consultation in the year of follow-up. Based on these data, it did not seem to be a major issue that patients who didn’t want to participate in the research-driven screening were easily found by the practice nurse. We added to the discussion section (p20, line 29):

We have no reason to assume that non-respondents on the research-driven screening were those most in distress, who would have been the easiest to be detected by the PNs, as low detection rates were reported for the whole diabetes population in the intervention practices.

2c) Additionally, the authors note that the ultimate study sample was probably underpowered. There is not much to do about this, except to further note that this paper should not really be about SMS outcomes (as the outcomes are not valid), but rather about the lessons learned.

We believe that our study gives food for reflection about the hybrid effectiveness-implementation design. Implementation of SMS implied that we were faced with less clusters of analysis than we had planned, and an imbalance in the number of patients per cluster. We didn’t want to suggest that we performed an underpowered study. In our case, the total number of patients was acceptable at the end. See also our response to comment 4 of reviewer 3. We made some adjustments to the discussion section (page 22, line 6):

The number of family practices involved in this study was insufficient. Although collaboration with regional stakeholders resulted in organisational and financial benefits for family practices willing to participate, this incentive did not appear to be a decisive factor in the process of considering whether or not to participate. For robust multilevel modelling, it is necessary that both sufficient clusters are recruited and sufficient patients are available per cluster.[44] Our intention was to include 46 PNs but only 40 PNs participated. Since PNs working in a team in which patients were seen alternately needed to be considered as one PN, the final total number of clusters for analysis reduced to 33. Because of the regional approach of the SMS-project, we were confined to the family practices within the region under study. The number of patients within clusters ranged from 1 to 19. The loss to follow-up of patients was less than expected which led to an acceptable total number of study participants. However, the availability of clusters and balance between the clusters need to take into account in an effectiveness trial within an implementation setting.

3. I have a few other comments and thoughts that the authors may consider. Some may not be workable for this paper, but I’ll state them here in case they are useful in follow-up studies.

3a) It would be nice to see p-values in Table 1, to explore differences between intervention and control groups. If the study were unbalanced, this might be another reason the SMS appeared to fail.
The authors mention there were no differences (except on the PHI index), but it would be nice to see this formally stated with t-tests.

The p-values for the health outcomes have been added to table 1 (see p. 13-14). The outcomes do not suggest an unbalanced study.

3b) The PHI seems to have not budged in the intervention group, but does seem to have moved upward (worse) in the control group. Any comment on this?

This need to be rectified. In the original version of the PHI of Petkov et al (2010), a score with 0 indicating high self-management and 8 low self-management was used. However, we have used a reverse scale. On the Dutch PHI-scale, a higher score reflects better self-management skills. This was not correctly stated in table 1 and table 2. It has been changed in the table as well as in the text, see page 15, line 11.

Patients of both groups were comparable for the primary and secondary outcomes at the baseline measurement except for the sum score on the PHI-scale. This reflected better self-management skills of intervention patients regarding knowledge, symptom management and adherence to treatment. Both groups were comparable regarding the subdomain ‘coping’ that included three items about dealing with the effects of diabetes on daily life.

3c) The authors make the point that the nurses seem to have been largely compliant with the research protocol. However, if this is a Practice Nurse-led intervention, but, as the authors state, the incentive for participation went to the practice organization, this could mean that the individual nurses were not properly incentivized to participate (and in fact, if the practice, i.e. the GP, is seen as benefiting for their extra work, this could be a DISincentive). This is worth some reflection and discussion, in my view.

Besides the effect evaluation, we performed an evaluation of the implementation process. The outcomes will be reported in another paper. In our discussions with practice nurses, and in the questionnaires, practice nurses did not mention financial issues. It was not our impression that their motivation for SMS was directly influenced by financial incentives although an indirect influence could have occurred. In practice, it appeared to be difficult to translate the financial support into extra time for PNs for SMS. PNs felt the time constraints in their consultations as a barrier for SMS.

a. Finally, it seems like the study could have been handled through a cut-off design, using a regression-discontinuity analysis (see: http://www.socialresearchmethods.net/kb/statrd.php for example), rather than through cluster randomization. The patient N would have been higher. I mention this because it might be useful for future consideration. I also mention it in case the data are amenable to reanalysis using a regression-discontinuity model.

If we would have used a cut-off design, all patients with mild problems would receive usual care only. Patients with moderate problems would be randomised to receive usual treatment or the self-management support intervention. Patients with severe emotional problems would be allocated to the intervention group. If we aimed just to evaluate the effectiveness of the treatment phase of SMS provided by practice nurses, this design could have been suitable. However, our starting point was actual implementation of SMS in the routine care setting. We needed a cluster-randomised design. As there is no reason to assume discontinuity of our data, we did not perform a reanalysis using a regression-discontinuity model.

Reviewer 2. F. Lamers

This paper describes the results from a cluster-randomized controlled trial within a hybrid effectiveness implementation study to evaluate the effectiveness of a self-management support intervention implemented in the daily routine diabetes care of practice nurses. Forty primary care practice nurses participated in the study and were randomized. All patients with DM type II under the care of the PNs were invited to fill out a screening questionnaire; those with high emotional distress were invited to participate in the trial. PNs in the intervention arm implemented the detection method
and SMS in all consultations with all DM patients. The control arm PNs provided usual care. PNs did not know which of their patients participated in the trial. The statistical analysis are appropriate.

Loss to follow-up was limited, but major downfall of the study was the low number of patients in the intervention arm reporting high emotional distress in the nurse-led detection, leading to a very low exposure to the SMS intervention. As a result, the analyses of outcomes becomes somewhat of a formality; as one would expect, no differences between arms at 12 month follow-up were detected.

Some thoughts and comments on the study are listed below.

1. I'm not sure if I understand correctly what the implementation part and what the effectiveness part of the study is. From what I understand it is as follows: The content of the SMS intervention was the same as the intervention in the DELTA study, only the place of administration (clinic vs patient’s home) and administrator (practice nurse vs research nurse) was different. This would be the implementation part? The different case finding procedure - using the nurse-led detection- led to the evaluation of detection+SMS in the evaluation part of the study? If I got it all wrong, then some introduction on hybrid effectiveness implementation would be helpful.

Apparently, both the SMS intervention and the hybrid study design need more explanation. The content of SMS was not only the intervention of DELTA at a different place of administration and with a different provider, but also for a different target group (not only for patients with depression, but also patients experiencing problems in daily functioning and emotional distress, adjusted to primary care setting) and with a detection procedure added to the nurse-led intervention. SMS included both a detection phase and a treatment phase delivered by the PN in their routine care (the treatment phase of SMS was dependent of the detection outcomes and could also mean a referral to the GP for further diagnostics and treatment). Furthermore, compared to the former DELTA trial intervention, a greater emphasis of the treatment was on the problem solving technique in order to increase the benefit for patients with low socio-economic status. This has been explained in the study protocol.

As several critical components of the original intervention needed to be changed for implementation purposes, the effectiveness of the intervention could not be taken for granted. Therefore, we decided to evaluate the implementation strategies as well as the effectiveness of SMS simultaneously. The implementation strategies included not only adjustments to the intervention, but also the financial support and registration of SMS activities integrated in Electronic Medical Records. We elaborate more on the implementation strategies and the adjustments to the evidence-based intervention in the introduction (see reviewer 1, comment 1), and we have added a figure in appendix 1 to help the reader to get a clearer picture of the two parts of the hybrid study.

To put more emphasis on the implementation of SMS in the routine care of practice nurses, and the content of SMS (detection and treatment), we have added the following sentence to the abstract (pag2, line 11):

Intervention. Practice nurses in the intervention arm (n=19) were trained to integrate SMS into routine diabetes consultations. SMS included detection of patients with emotional distress and reduced daily functioning, and provision of self-management support through problem solving and reattribution techniques when needed. Practice nurses in the control arm (n=21) provided usual care.

The criteria for watchful waiting were deleted, because the focus of the effectiveness study was on the patients who would receive the detection and treatment phase of SMS. See page 8, line 23:

Patients who experienced problems of daily functioning (DFT > 4) and emotional health problems (DS > 3 combined with moderate scores on at least one subscale of the 4DSQ) were offered consultations for self-management support. Patients registering a ‘severe’ score on at least one of the four subscales of the 4DSQ were referred to the GP for further diagnostics and treatment. PNs opted for watchful waiting in the following two situations: where patients registered only ‘mild’ scores on the 4DSQ, where patients registered ‘moderate’ scores on the 4DSQ but without accompanying evidence that diabetes had become a burden affecting daily functioning (DFT ≤ 4).

The self-administered postal screening questionnaire was a prerequisite for the research-driven effect evaluation with the nurse-led detection as inherent part of the SMS-intervention. The explanation has
been added to the methods, in the paragraph about study participants (p6, line 26). See also our reply on the first comment of reviewer 1.

In the selection of patients for the effectiveness trial, the following issues were crucial. First, we wanted to select patients who would receive the complete intervention including detection and treatment. Furthermore, patients from the intervention and control arms should be selected in a similar way, and without interfering in their routine care. We set up a screening procedure by means of written questionnaires sent to patients’ home addresses to identify patients with actual problems of daily functioning and emotional distress.

2. Some of the differences between research-led identification and nurse-led identification could have been partially caused by waxing and waning of symptoms, as the authors mention in their discussion. But the discrepancy is so large that other factors are likely involved as well. The potential difficulty of people to express their emotional problems could play a much larger role when questions are asked face-to-face rather than in paper-and-pencil form, as is mentioned as well in the discussion. So the big question is: Knowing that patients might be less inclined to mention emotional problems than on a paper questionnaire, shouldn’t the nurse-led identification strategy have been tested before being implemented in this study. In retrospect, was not piloting the procedure a mistake? What recommendations would you have for future implementation studies that are changing the identification procedures? Please expand on this in the discussion.

We believe that the large discrepancy between the written, self-administered anonymous questionnaire, and the nurse-led detection is a combination of factors regarding the expectations and needs of patients, the course of their emotional well-being, the skills of PNs and the setting in which SMS was integrated. In other words, though we were not able to show the effectiveness of SMS, this study has provided a lot of input to reflect on the identification of patients with problems in their emotional and social functioning. It provided lessons, which are valuable for research and practice. As described earlier in this letter in our remarks to the third editorial comment (see page2 of this letter), there were several reasons to leave out a pilot study. There was an implementation momentum, we could use the experiences from the former study, in which the treatment phase was evaluated. PNs would receive a comprehensive training, and they would ask common questions about psychosocial functioning. Furthermore, the outcomes of the screening did not arise questions about the instruments: as expected 20% of the respondents met the detection criteria.

As there can actually be several reasons to leave out a pilot, our recommendation for future implementation studies that are changing their identification procedure would be to plan specific evaluation moments in the follow-up to decide about optimizing the procedure (p22, line 23): The planning of a more flexible design of the effectiveness trial within the hybrid design would have been helpful in overcoming implementation problems. This calls for inclusion of specific evaluation moments to decide about the need for adjustments in the intervention itself, in implementation activities, or with regard to follow-up measurements.

3. On page 20 (Discussion) the authors mention they considered another screening procedure in the design phase as well. Did the authors consider having DM patients fill out the forms followed by a discussion of responses with the PN as alternative to face-to-face questions? Do the authors feel this is a feasible alternative?

In the planning of the study, the use of written forms in the consultations of diabetes patients was not considered to be a sustainable solution for adoption in the routine care practice. For the shift from a
biomedical towards an integrated and patient-centered approach of care, we did not feel that written forms would be a feasible alternative for the detection procedure of the SMS-approach. However, in retrospect, it could have been a feasible approach as one PN suggested this option during a booster session.

In our discussion, however, we referred to the inclusion of the detection phase of SMS in the control arm. We added the word ‘nurse-led’ to clarify our consideration (p21, line 26):

In the planning of the study, we considered avoiding research-driven screening by inserting the nurse-led detection phase of SMS into both study arms.

4. At 4 months there are differences between groups on two autonomy subscales (that disappear at 12 months). What is the authors’ take on this. To what extent could this be a chance finding? Could it have been an effect of attention being given to psychosocial wellbeing? Is it relevant at all, given the relatively small difference?

Since we used several secondary outcome measures, the chance to find a statistically significant outcome is high. We did not consider the outcomes on the autonomy subscales as clinical relevant as the effect sizes were small. Both the subscales ‘Indoor activities’ and ‘Family Role’ of the Impact on Participation and Autonomy (IPA) include 7 items. The differences between the intervention and control patients were just 1.1 resp. 1.4 point on a range between 0-28.

On second thoughts, a statistically significant difference on the IPA might have been an effect of the nurse-led detection. The questions are described as: ‘my chances to fulfilling my role / doing this and that as I would like are….very good, good, fair, poor or very poor’. In theory, a discussion about psychosocial functioning may change patients’ perceptions about their own chances. At the start of the year of follow-up, PNs were more compliant to the protocol thus it seems to be logical that the effects disappeared at 12 months. More research is needed to evaluate the relationship between a discussion with a nurse about daily functioning, and patients’ perceptions on their autonomy.

Minor comments of reviewer 2:

1. Abstract: Without reading the Methods of the paper it is not clear what is meant by ‘research driven’.

It has been changed in the abstract to clarify the use of the posted, research-driven screening questionnaires to select study participants.

Participants. Type 2 diabetes patients (n=264) selected by a research-driven self-administered questionnaire aimed to measure emotional distress and diabetes-related reduced daily functioning. (.)

Results. Only 16 of the 117 patients in the intervention arm (14%) who were found eligible by a posted research-driven screening questionnaire were detected by their practice nurses.

Conclusions. SMS in its present form was not effective. The research-driven screening to select trial participants appeared to be inconsistent with nurse-led detection in routine practice.

2. Please add a description of the control arm condition in the Methods section.

The patients in the control arm received usual diabetes care according to the Dutch clinical practice guidelines. This care is normally provided during three quarterly consultations. It has been added in the paragraph about study participants (p6, line 14-15):

Their PNs for diabetes care were randomly assigned to an intervention or control arm. PNs in the intervention arm were trained in SMS. PNs in the control arm provided usual diabetes care, conforming to the Dutch guidelines.

3. Please provide directions of effects in the Results section (i.e. ‘autonomy was higher in the intervention group than in the control groups’ rather than ‘there was a difference’).

It has been changed in the results section, see page 17, line 16:

“For the secondary outcome measures, only the autonomy and participation indoors and family role was higher in the intervention arms than in the control arms (p<0.05) after 4 months.”
We also have made adjustments to the discussion, page 20, line 5:

This study, with low exposure of study participants to the complete intervention, could not demonstrate any effect of SMS on emotional, social or biomedical parameters, except for autonomy and participation with regard to activities indoors and the family role after 4 months. This effect in favour of the patients in the intervention arm had disappeared after 12 months.

4. Consort checklist item 3b; weren’t the PNs instructed to use the cut-offs of DFT and DS less rigidly?

You are right, we have inadvertently not mentioned this in the Consort checklist. It has been added to item 3b of the Consort checklist.

Reviewer 3. Jason Oke

1. Is the study design appropriate to answer the research question? No. Although, the study could have worked it didn’t because the design did not take into account the inter-individual variation in DFT and/or regression to the mean from the screening phase round which subsequently undermined the findings from this trial. This is however adequately addressed in the title and the whole manuscript.

Thank you.

2. Are the methods described sufficiently to allow the study to be repeated? Incomplete – There is not enough information to replicate this study as it currently stands. For example, there is not enough information about the Daily Functioning Thermometer, we know it is a visual analogue scale, but what does it look like exactly (is the scale linear?), why was favoured over other methods (e.g. likert scales) and has anyone has tested how reproducible it is (see above)? It is important because this is the how the primary outcomes is assessed. In addition, I could not find sufficient detail in this paper as to what SMS consisted of. This trial could not be repeated with the detail given in this manuscript. Could some of this information be given as appendices to the paper?

We have added more information about the DFT in the manuscript. We were looking for a very simple and patient friendly measure to ask patients to rate their daily functioning. Most of the validated questionnaires have a focus on specific symptoms like depression or diabetes related-distress rather than on patients’ daily functioning. As the Distress Thermometer is a valid and common used instrument in cancer care, and another study showed the validity of a Visual Analogue Scale indicating the worst to the best possible life as a valid screening tool in adolescents with diabetes, we considered the DFT to be valid and useful for the purpose of our study. As described in the study protocol, a pilot study among 7 diabetes patients confirmed the face validity of the DFT. In comparing Likert scales and visual analogue scales, we found that both methods of presenting response options show comparable responsiveness. The DFT has been explained at page 7 (line 2-14) as follows:

It included …the ‘Daily Functioning Thermometer’ (DFT) which is a visual analogue scale to measure how suffering under the burden of diabetes affects patients’ perceived functioning in everyday life. Patients were asked to indicate a position between zero (no burden at all) to ten (extremely burden) at a continuous vertical 10 cm line. A score of 4 was chosen to differentiate between patients who could benefit from self-management support (score > 4) and patients who had found a satisfactory way to live with the consequences of their diabetes (score ≤ 4). The DFT has been developed for the purpose of this study. It is comparable to the DT, a validated questionnaire in the care for patients with cancer [21], and a validated Visual Analogue Scale regarding the worst (0) and the best possible life (10) for adolescents with type 1 diabetes mellitus [22]. Next to the DFT, patients were asked to complete the three-item Distress Screener (DS) which is a quick-scan instrument for emotional distress and an indicator of potential underlying severe mental health problems. [23]
In the paragraph about outcomes (page 9, line 26) we have deleted the subordinate clause regarding the DFT which may be confusing for the reader.

The effect of SMS on patients’ daily functioning was measured by means of the Daily Functioning Thermometer (DFT), which was also used for the research-led screening, as well as the nurse-led detection in diabetes consultations.

In response to the other reviewers, we have tried to give more details about the SMS intervention to enable the readers to replicate the study.

For a replication of the study, the trainer manual is available on request.

3. If statistics are used are they appropriate and described fully? Yes, but...... The analysis takes into account clustering and correlation between practice nurses and practices using a logistic three level model. I would normally favour a GEE approach for structured data with a binary outcome as they represent population average effects rather than cluster specific effects. Moreover, the GEE estimates are equal or less than the CS estimates depending on the size of the cluster variance (see Ten-have et al 2004, Deviations from the population-averaged versus cluster-specific relationship for clustered binary data). This is complicated by the fact that the authors have used three levels, and as I do not know if this can be accommodated with GEE I would probably ignore the third level and use GEE or at least check sensitivity to model choice. In short, can the authors assure me or the reader that the effect estimates for the main outcome are not dependent on their choice of model and do not differ significantly from other methods of analyses - e.g. I note that the raw unadjusted OR's are much closer to 1 than their model estimates).

We have performed the method of generalized estimating equations (GEE) to check the sensitivity of the model choice. We found similar outcomes. The OR at four months was 0.668 with a p-value of 0.189 (0.366 – 1.220 95%CI). The OR at twelve months was 1.451 with a p-value of 0.216 (0.804 – 2.618 95%CI). As the GEE does not take into account the level of the PN's, we decided not to add this to the manuscript.

4. Also on the sample size: in the manuscript it is stated that 232 patients were required but in the protocol it says it should be at least 331 to take into account loss to follow up. As it happens the loss to follow up is much less than they predicted and the sample size at 4 and 12 months only just fails to reach the required minimum sample size and hence does not weaken the power significantly (which is high to start with ~90%). This slight reduction in numbers should be noted and I think the writing in the manuscript could be clearer so that the reader is not left to work this out for themselves.

Our intention was to invite 10 patients per PN (N=460) to take into account that 20% would not give informed consent and 30% would be lost to follow-up. This would left the required 232 patients. In our case, we could invite 357 patients of 40 PN's for the trial. The percentage of patients who did not give informed consent was 26%, 6% more than we expected. However, the loss to follow-up was only 16% instead of the estimated 30%. We ended up with 223 patients at T12 which is on average 5.5 patients per PN, and almost the number of patients that were required for an alpha of 0.05 and a beta of 0.90.

However, the problem in our study was that the number of PN's was only 40, including teams of PN's of which the patients could receive diabetes care from the PN's alternately. These teams of PN's needed to be considered as one unit of analysis. It left us with 15 units of analysis in the intervention arm (117 patients) and 19 units of analysis in the control arm (147 patients). If our power in the study protocol was calculated based on these 34 units of PN's, we should have included more patients for each of the units. (Based on a power calculation with 15 resp. 19 PN's, we needed at least 7.4 patients per cluster)

In figure 2, we have included the total number of patients who did want to participate, and the loss to follow-up for the intervention and control arms.

The acceptable total number of study participants has been added to the discussion section (page 22, line 16). See also our response to reviewer 1, comment 2c.

Although the loss to follow-up of patients was less than expected which led to an acceptable total number of study participants, the availability of clusters and balance...
between the clusters need to take into account in an effectiveness trial within an implementation setting.

5. Is the standard of written English acceptable for publication? Yes. Although, please do not use the term diabetics (page 4, line 11 and 39)

It has been changed on page 4 in ‘people with diabetes’ (see line 7) and ‘diabetes patients’ (see line 25)

---

**VERSION 2 – REVIEW**

**REVIEWER**
Christopher P. Morley, PhD
SUNY Upstate Medical University
USA

**REVIEW RETURNED**
27-Apr-2015

**GENERAL COMMENTS**

I believe this paper has come a long way from its first iteration. It is still somewhat dense to read, but I am grateful for the figures outlining study and patient recruitment.

I do have a number of suggestions on this version that would go a long way to improving the clarity of the document. All comments use page numbers and lines that refer to the CLEAN (no tracked changes) version of the new manuscript, as labeled at page top (page #) and in the margins (line #).

1. The abstract statement on the primary outcome states as follows: "Primary outcome measure was patients' daily functioning as assessed by a visual analogue measurement of the perceived effect of diabetes on daily functioning." The abstract should state the actual instrument, which, if I understood the paper correctly, is the dichotomized DFT>4 variable.

2. There is a statement in the Methods section, under "Statistical Analysis" (p.11 lines 3-5) that states:

"For the primary outcome, we used the research-led screening outcomes to ensure that all participants of both intervention and control arms had a score DFT>4 at baseline."

I am not sure what this statement means.

3. The authors make reference to a "DT" tool, but the acronym is not spelled out, to my knowledge. (At least, I could not find where it was spelled out, if in fact it was.)

4. There are a number of grammatical and stylistic errors in the paper that remain (or may be new to this version):

- Pg. 9, Line 5: “Treated for SMS...” - I don't think anyone was treated FOR SMS, but perhaps I am misunderstanding something in this sentence.

- Pg. 15, Line 23: "A number of 30 participants..." - I think this means EITHER "A number of patients (n=30)" OR it means "30 patients...", but I don't know what a number of 30 means.

- Pg 17, Lines 14-15: "For the secondary outcome measures, only
the autonomy and participation indoors and family role was higher in the intervention arms than in the control arms (p<0.05) after 4 months.” - I think this might be a better sentence:

"For the secondary outcome measures, only "Autonomy," "Participation Indoors," and "Family Role" were higher in the intervention arms than in the control arms (p<0.05) after 4 months.”

Also pertaining to this line - I think the fact that "Participation Outdoors" was nearly or marginally significant (at p=.06) deserves at least a mention.

- Pg. 22, lines 11-12: “However, the availability of clusters and balance between the clusters need to take into account in an effectiveness trial within an implementation setting.” There is something wrong with this sentence, and the fix depends upon what the authors mean to say.

- Pg 22, lines 25 - 26: I think the “nor” and “neither” are reversed in this sentence. At any rate, it needs another look from the authors.

5. Appendix 1, which I think is supposed to segregate the Implementation from Effectiveness evaluation, does not do a lot of work for this reader. I am not entirely sure what it is supposed to be telling me, in the context of the rest of the paper.

6. I don't see a full description of SMS in the present manuscript. The flow diagram (fig 1) describes a bunch of forms that led to the SMS intervention, but I don't know what actually comprises the SMS intervention. Additionally, the DFT, which is the tool used to assess the primary outcome measure, is simply described as a "visual analogue tool."

I believe there should be a separate whole paragraph (or more, if needed) describing each of these items; perhaps even sub-headings for each. Given that SMS is the intervention under study, and DFT needs to be fully described, including development & evaluation. As a visual tool, this should also possibly serve as an additional figure.

As an overall statement, I would also urge the authors NOT to be so verbose in the reply to reviewers. The response to reviewers should be limited - if that much explanation about what is in the paper is required, then either the paper isn’t clear enough, and/or that explanation belongs IN THE PAPER. The paper should stand on its own, post-publication, without the need for external clarifications, as placed in the response to reviewers.

| REVIEWER | Jason Oke  
University of Oxford, UK. |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>REVIEW RETURNED</td>
<td>27-Apr-2015</td>
</tr>
<tr>
<td>GENERAL COMMENTS</td>
<td>I am satisfied with the authors responses to my initial criticisms of the manuscript,</td>
</tr>
</tbody>
</table>
Reviewer C.P. Morley gave some suggestions to further improve the manuscript. Our point-by-point response to these suggestions will be set out below. The changes have been quoted in the letter. We also refer to the page numbers and lines so that the fragments can easily be traced in the marked copy of the manuscript.

Comment 1.
The abstract statement on the primary outcome states as follows: "Primary outcome measure was patients' daily functioning as assessed by a visual analogue measurement of the perceived effect of diabetes on daily functioning." The abstract should state the actual instrument, which, if I understood the paper correctly, is the dichotomized DFT>4 variable.

Response comment 1: We added more information about the primary outcome to the abstract (p2, line 16). It has been formulated as follows: "Primary outcome measure was a dichotomized score on a Visual Analogue Scale that measured the perceived effect of diabetes on daily functioning."

Comment 2.
There is a statement in the Methods section, under "Statistical Analysis" (p.11 lines 3-5) that states: "For the primary outcome, we used the research-led screening outcomes to ensure that all participants of both intervention and control arms had a score DFT>4 at baseline." I am not sure what this statement means.

Response comment 2: The primary outcome was a dichotomized score to compare the number of patients who had improved in their daily functioning (DFT ≤ 4) with patients who still perceived a burden of diabetes on their daily functioning (DFT > 4). Patients were eligible for trial participation because of their perceived burden of diabetes. Therefore, to make a DFT>4 the starting point for all participants, we used the research-led screening outcomes instead of the baseline measurement. It has been clarified in the manuscript as follows (page 11, line 11): "To ensure that all participants of both intervention and control arms started with a score DFT>4, indicating a perceived burden of diabetes, we used the dichotomized research-led screening outcomes at baseline. Consequently, the model only included 4- and 12-month follow-up as the baseline value was the same for all participants."

Comment 3.
The authors make reference to a "DT" tool, but the acronym is not spelled out, to my knowledge. (At least, I could not find where it was spelled out, if in fact it was.)

Response comment 3: The DT refers to the Distress Thermometer. We have added the full name of the tool to the manuscript (page 7, line 6).

Comment 4.
There are a number of grammatical and stylistic errors in the paper that remain (or may be new to this version):

a) Pg. 9, Line 5: "Treated for SMS..." - I don't think anyone was treated FOR SMS, but perhaps I am misunderstanding something in this sentence.
We referred to patients' exposure to SMS. We have changed the sentence (p9, line 15) into: “These data were used to evaluate whether patients were exposed to SMS.

b) Pg. 15, Line 23: "A number of 30 participants..." - I think this means EITHER "A number of patients (n=30)" OR it means "30 patients...", but I don't know what a number of 30 means.
We have corrected this error as follows (p15, line 23): From 30 participants who scored DS>3 and
were therefore eligible for further diagnostics, ten patients did not receive the 4DSQ because i) they were already undergoing psychological treatment, or ii) they were not interested, or iii) a specific stressor was clearly causing the distress.

c) Pg 17, Lines 14-15: "For the secondary outcome measures, only the autonomy and participation indoors and family role was higher in the intervention arms than in the control arms (p<0.05) after 4 months." - I think this might be a better sentence: "For the secondary outcome measures, only "Autonomy," "Participation Indoors," and "Family Role" were higher in the intervention arms than in the control arms (p<0.05) after 4 months." Also pertaining to this line - I think the fact that "Participation Outdoors" was nearly or marginally significant (at p=.06) deserves at least a mention.

We have revised the paragraph about the secondary outcome measures as follows (p17, line 14): For the secondary outcome measures, the intervention group perceived significantly more participation than the control group regarding ‘autonomy indoors’ and ‘family role’ (p<0.05) after 4 months. The same applied, with nearly statistically significant differences, for ‘autonomy outdoors’ (p=0.06) and ‘social relationships’ (p=0.05). However, these effects disappeared at 12-month follow-up. As illustrated in table 4, we found no other statistically significant differences for the secondary outcome measures at 4 and 12 months.

d) Pg. 22, lines 11-12: "However, the availability of clusters and balance between the clusters need to take into account in an effectiveness trial within an implementation setting." There is something wrong with this sentence, and the fix depends upon what the authors mean to say. This is what we meant to say: "However, the availability of clusters as well as the balance between clusters need attention when designing a cluster-randomised trial that will be performed in an implementation setting." It has been corrected in the manuscript (p22, line 11-12).

e) Pg 22, lines 25 - 26: I think the "nor" and "neither" are reversed in this sentence. At any rate, it needs another look from the authors.

The words nor and neither were reversed. We have corrected this error. (p22, line 26)

Comment 5.
Appendix 1, which I think is supposed to segregate the Implementation from Effectiveness evaluation, does not do a lot of work for this reader. I am not entirely sure what it is supposed to be telling me, in the context of the rest of the paper.

Response comment 5: The figure in appendix 1 was provided to illustrate the relation of the effectiveness study to the implementation activities. It was meant to be helpful for the reader to grasp the differences between the implementation activities and the effect study. As the reviewers do not feel that the appendix is of added value, we would suggest deletion of the appendix. On page 5, line 13, the reference to the appendix has been deleted.

Comment 6.
I don't see a full description of SMS in the present manuscript. The flow diagram (fig 1) describes a bunch of forms that led to the SMS intervention, but I don't know what actually comprises the SMS intervention. Additionally, the DFT, which is the tool used to assess the primary outcome measure, is simply described as a "visual analogue tool." I believe there should be a separate whole paragraph (or more, if needed) describing each of these items; perhaps even sub-headings for each. Given that SMS is the intervention under study, and DFT needs to be fully described, including development & evaluation. As a visual tool, this should also possibly serve as an additional figure.

Response comment 6: The SMS-intervention was a self-management approach, including a detection
and follow-up phase. Thus the intervention was not limited to the self-management support provided during extra consultation by the practice nurses. We have tried to further clarify this by means of the following sentences: ‘SMS included a detection and follow-up phase.’ ‘The detection phase of SMS started by..’ and ‘The follow-up phase of SMS was based on the outcomes of the DFT and 4DSQ’. Furthermore, we have added more information about how the practice nurses used the Daily Functioning Thermometer in their routine practice. As the DFT was just a 10 cm vertical line, ranging from 0-10 with indications added at 0 (no burden) and 10 (extreme burden), we don’t feel that it needs to be added as a figure. We have made some changes in the paragraph about the SMS intervention. See page 8, starting from line 4:

"SMS included a detection and follow-up phase. The flow chart of SMS is presented in figure 1. The detection phase of SMS started by exploring patients’ experienced problems in daily life. Then PNs applied the Daily Functioning Thermometer by showing patients a vertical line, ranging from 0 to 10 with 0 indicating no burden at all and ten indicating extreme burden, and asking patients to rate their perceived burden of diabetes for their daily life. PNs also verbally administered the three questions of the Distress Screener. Patients with score DS > 3, which indicates high risk for underlying mental health problems, were asked to complete the Four-Dimensional Symptom Questionnaire (4DSQ) immediately after the consultation. This self-report instrument is widely used in Dutch primary care to distinguish non-specific distress from depression, anxiety and somatisation. It differentiates between mild, moderate or severe symptoms. The completed 4DSQ was returned to the PN, who computed the sum scores on each domain. The follow-up phase of SMS was based on the outcomes of the DFT and 4DSQ. Patients registering a ‘severe’ score on at least one of the four subscales of the 4DSQ were referred to the GP for further diagnostics and treatment. Patients who experienced problems of daily functioning (DFT > 4) and emotional health problems (moderate scores on at least one subscale of the 4DSQ) were offered consultations for self-management support. These extra consultations delivered by PNs were aimed to support patients in their day-to-day management of diabetes and its emotional and social consequences."

Furthermore, we have made some changes to the legend of figure 1. ‘SMS-approach’ has been changed into ‘SMS’. The abbreviations have been explained (See p24, line 1).