Response to reviewer comments

<table>
<thead>
<tr>
<th>Reviewer Comment</th>
<th>Authors Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BMJ</strong></td>
<td></td>
</tr>
<tr>
<td>Does it make sense, though, to provide an overall sensitivity and specificity? There must be subpopulations of interest and we worry that these overall estimates can be misleading.</td>
<td>We agree that providing overall sensitivity and specificity could be misleading but we also think readers may wish to know these values. In the revised paper, we have still given these figures but included a caveat that because of high heterogeneity of studies and settings these figures should be treated with caution (pages 2 and 8-9 in the revised manuscript). The reason we did not perform sub-analyses (e.g. for the UK population) was we do not have sufficient data.</td>
</tr>
<tr>
<td>The title should not be a question.</td>
<td>We agree with the reviewer and have amended the title.</td>
</tr>
<tr>
<td>Three searches were carried out in all. Why was “….the prevention of type 2 diabetes in women with a history of gestational diabetes…” (Page 6) search used?</td>
<td>We have included a rationale for searching specifically for a history of gestational diabetes on page 4 of the revised manuscript.</td>
</tr>
<tr>
<td>Table 1. Diagnostic accuracy data. It would be useful here to have the positive predictive values for each study (as well as the sensitivity and specificity values).</td>
<td>We agree with this reviewer that the information would be useful and have now extracted this additional data from the primary studies and included it in the revised manuscript (table 1, page 25). In some instances these had to be calculated from the raw data when absent from the published literature.</td>
</tr>
<tr>
<td>Figures 2 and 3. Labelling the studies by reference test criteria separately would help the interpretation of heterogeneity. Also, labelling the ‘gestational diabetes’ studies would also be beneficial.</td>
<td>We agree with the reviewer’s comments that these figures could be made clearer. We have adjusted figures 2 and 3 to accommodate this suggestion (pages 19,20). We have coloured the points on the graph according to whether the diagnostic criteria followed ADA, WHO or other guidelines as listed in table 1. We have also made it possible to identify which studies were based on at risk populations or populations with a history of gestational diabetes.</td>
</tr>
<tr>
<td><strong>Reviewer 1</strong></td>
<td></td>
</tr>
<tr>
<td>It is not completely clear whether the authors originally aimed to conduct an additional meta-analysis focusing on lifestyle interventions exclusively in women with a history of gestational diabetes, or if these studies were meant to be included in the meta-analysis currently presented that assessed the efficacy of preventive interventions. A better contextualisation in the introduction referring to the latest evidence on GDM and a clearer description in the methods section of this part of the systematic review would strengthen the manuscript.</td>
<td>We agree, and have inserted additional text on page 4</td>
</tr>
</tbody>
</table>
Studies for the meta-analysis assessing the efficacy of lifestyle interventions have been divided in two sub-group based on follow-up length (1-2 yrs/3-6yrs). Could the authors explain why they did not stratify studies in shorter time periods (i.e. every one of two years), like previous studies (see Balk et al, 2015)?

Furthermore, the meta-analysis examining risk reduction at follow-up post intervention includes studies which differ quite substantially in their follow-up period. Authors conclude ‘lifestyle interventions were associated with a mean 36% relative risk reduction in incidence of type 2 diabetes over 1-6 years, attenuating to 20% with follow-up of 2-20 years’.

The latest estimate would be hard to interpret strategically considering the 18 years range. Please explain.

We assessed the intervention trials in two separate forest plots in a way that reflects current UK policy (intervention trials planned as part of local diabetes prevention programmes and NHS DPP last up to 18months). We then did a second analysis to see how much additional benefit is likely to be gained from more longer lasting interventions.

Balk et al’s study had a slightly different focus to ours. They assessed the effectiveness of lifestyle interventions not only by looking (as we did) at the definitive outcome of diabetes incidence but also by studying surrogate markers such as weight loss and changes in glycaemic markers. They placed follow up studies on the same forest plot as the trials of interventions. We have performed separate forest plots to assess the effectiveness of trial interventions once the participants have completed the intervention (i.e. after the end of the period when they would be in regular contact with the study team, attending for regular assessments and so on) and separately assess what happens once participants are away from the intensive environment, back within their everyday lives, without monitoring to see if the risk reductions are maintained.

We agree that it is difficult to interpret the strategic implications of these findings due the near-absence of long-term follow-up. This is a significant limitation of most of the primary studies included in this review. This is now discussed briefly on pages 10 and 15 of the revised manuscript and the abstract is amended.

Comparison between screening tests accuracy is extremely interesting. Considering the low sensitivity of both HbA1c and fasting glucose tests, the authors highlight the danger of inaccurately classifying individuals at high risk of type 2 diabetes, especially for those who tested falsely positive. While upstream interventions aimed at the entire population would be ideal, they are not always feasible. Sometimes the screen-and-treat approach can better align with the financial capacity of the health system. I suggest the authors expand the discussion on the ‘pre-diabetes’ label, encouraging, in the future, the usage of alternative terms such as ‘non-diabetic hyperglycaemia’, which are less likely to imply that a type 2 diabetes diagnosis is inevitable.

We agree with this reviewer’s line of reasoning, though we are keen not to extend the word count of our paper too much more. The question of whether to re-label pre-diabetes as ‘non-diabetic hyperglycaemia’ perhaps merits further discussion in the BMJ – whilst we are happy to include such discussion, the editors might consider inviting an accompanying editorial should this paper be accepted. In relation to the affordability of screen-and-treat policies, please note that we have submitted another systematic review, of cost-effectiveness studies, manuscript number BMJ.2016.036366. We understand that this is currently being considered by editors. In that study we discuss the very complex implications of the effectiveness and cost-effectiveness data on diabetes prevention.

Page 10. I might be wrong but it seems like the figure referring to the meta-analysis evaluating the impact of the metformin is figure eight (and not seven).

We thank the reviewer for pointing out this typo, which has now been corrected (page 10).
Page 18. The box reporting that ‘5 conference abstract excluded’ in the flow diagram seems better placed in the upper selection. Please check. I also suggest flipping the two arms of the flow diagram, positioning the information regarding the test accuracy meta-analysis on the left as it comes first in the manuscript.

| We thank the reviewer for these helpful suggestions, which we have followed (figure 1). |

Page 22. Figure 9 is not straightforward to understand. I suggest using footnotes to better explain it.

| We agree, and have now amended the text in the diagram to better explain the data presented (page 24 in the revised manuscript). |

**Reviewer 2**

This review addresses an important topic, in effect asking if the NHS Diabetes Prevention Programme is based on good evidence. The covering letter suggests that it is not, but the Discussion (foot of page 14) in the paper is more favourable.

| We agree with this reviewer that the covering letter and discussion did not quite align. We have now amended the text to reflect a nuanced interpretation of the data that:  
1. Interventions will benefit less than a third of people (page 13 and 14 of discussion)  
2. Quality of the evidence is poor as per risk of bias tool and GRADE (see page 11)  
3. On individual basis interventions will benefit those who are able to engage, however may have limited benefit on a population basis (see page 14). |

One question would be whether the interventions used in the trials could be replicated in routine care, given that some interventions required a lot of resources.

| We discuss this on page 14. In short, there are both practical and resource concerns (the latter being covered more fully in our cost-effectiveness systematic review). As noted above, we wonder if there should be an accompanying editorial by an independent commentator to consider the implications for practice and policy. As we point out in the paper, diabetes prevention programmes are being introduced not just in the UK but in many other high and middle-income countries. An editorial that considered the latest evidence, synthesised in these new reviews, would be timely. |

The stated aims are to review both the diagnostic accuracy of screening tests for “pre-diabetes” and the efficacy of lifestyle interventions and metformin for preventing progression to type 2 diabetes. I don’t like the term “pre-diabetes” because most people don’t go on to diabetes. So the statement on page 5 that “that if individuals do not take action they will develop diabetes” is incorrect. Two UK studies reported that most people with IGT did not develop diabetes. The Ely study reported 10% progressed after 4.5 years, and 57% regressed to normal. The Bedford study reported that only 16% developed diabetes. I prefer “non-diabetic hyperglycaemia”. But pre-diabetes is very commonly used, so this is a

| We accept the reviewer’s comment. We certainly did not mean to imply that pre-diabetes means someone will inevitably go on to develop T2D. Indeed, we share the concern about the term and its ambiguous implications. We have clarified the text on page 4 to ensure that readers do not get the wrong end of the stick. We also discuss progression rates in the discussion on page 12. |
The Abstract states that “In identifying pre-diabetes, HbA1c had a mean sensitivity of 0.50.” This statement, and the similar one on fasting plasma glucose, are meaningless without specifying the levels of HbA1c and FPG used. The sensitivity of HbA1c could be 100% if a low enough level was used. It appears that two cut-offs were used for HbA1c, 5.7% as in the American definition and 6.0% as in the WHO, though in appendix 5, an HbA1c cut-off of 6.4% also appears to be used. But these cut-offs were chosen to define non-diabetic hyperglycaemia, not for screening purposes. In a good screening study, several levels would be used, and paired forest plots used to illustrate the trade-off between sensitivity and specificity. Furthermore, we are not told whether the 50% sensitivity is based on the ADA or WHO thresholds. I would expect a much higher sensitivity with the 5.7% cut-off.

We agree with this line of reasoning. The reviewer is saying that the sensitivity and specificity of the test will vary depending on the cut-off. Different studies used different cut-offs and each set of authors justified their chosen cut-off with reference to a particular standard. The problem (as we discuss in some length in the paper) is that there is more than one standard (ADA, WHO etc). There is confusion for sure, but that confusion is in the data and is not of our making! Our decision to use the sensitivities and specificities reported by authors of primary studies was the only achievable course of action, since applying a ‘standardised’ cut-off to all studies would require re-analysis of individual patient data (which we did not have access to). However, we have clarified the abstract to emphasise that the sensitivities/specificities used in our synthesis are based on different cut-off values and hence should be interpreted with caution.

Methods, page 6. Interventions included lifestyle and metformin. Why were other drugs not included? Liraglutide has been shown to reduce progression to type 2 diabetes (Pi-Sunyer et al NEJM 2015) as has pioglitazone (IRIS trial, Kernan 2016). When considering metformin, age should be taken into account.

Metformin and Lifestyle interventions are currently the NICE recommendations and the only two options available in the UK to local general practitioners. Similar policy restrictions apply in other countries (e.g. US ADA guidelines http://www.ndei.org/ADA-diabetes-management-guidelines-strategies-for-preventing-or-delaying-type-2-diabetes.aspx.html )

For this reason, other pharmaceutical options were not considered in this review.

QUADAS is used to assess the quality of diagnostic studies, but nothing is done with the results. Two options could have been considered – exclusion of studies considered to be of poor quality; or a sensitivity analysis testing inclusion or exclusion or weaker studies. Table 1 shows that some studies were poor. Lin 2014 has three criteria with a high risk of bias rating. Lee 2013 also has three high bias scores. Should the Valdes 2011, Skriver 2010 and Pajunen 2010 studies of HbA1c have been included? Systematic reviews usually included a list of studies considered but excluded.

We thank the author for highlighting the absence of a sensitivity analysis in the final manuscript. We have performed a sensitivity analyses for the two diagnostic accuracy SR’s by excluding studies that were at high risk of bias. An additional sentence explaining the results of this can be found on page 9 and the full results of the sensitivity analysis can be found in the Appendix.

The corresponding summary statistics are presented alongside the primary analyses in figures 2&3.
The 50g GCT test is not included amongst the screening tests. There are studies comparing it with the OGTT and HbA1c (Phillips 2009).

None of the primary studies included in our review used this test in a screening phase, so whilst it may be of potential relevance it was not something we could assess in a systematic review.

There are also problems with the intervention studies. The Cochrane risk of bias was used to assess quality, but the detailed results are given only in the forest plots. No details are given of minimum quality for inclusion and it appears that all trials that provide adequate data were included. Again, a sensitivity analysis could have been used. The Ramachandran trial had the poorest score, but contributed 10% of the weight

We agree with this reviewer that the included studies were of variable methodological quality and this has impacted on our confidence in the meta-analysis results. Our concerns reflected in our risk of bias assessments, which we considered when making our outcome-level GRADE decisions.

In order to provide the most comprehensive synthesis of relevant studies, we chose not to pre-specify a minimum methodological quality for inclusion. To apply a threshold now would be post-hoc and we would risk losing valuable data. However, we accept that it is useful to test the robustness of our findings against the risk of bias and so performed a sensitivity analysis, removing the Ramachandran trial. This had very little impact on the overall effect estimate; the risk ratio was reduced from 0.64 to 0.63.

Revised discussion of the use of the Cochrane risk of bias tool can be found on page 10.

The trials had very different lengths of intervention, but much more important is the timing of outcomes measures – in past reviews we have observed that results are often initially good, for example in terms of weight loss, while the intervention continues, but that after it ceases, results deteriorate. So the main outcome, development of type 2 diabetes, should have been assessed a reasonable time (at least 2 years?) after the intervention stopped.

We agree with the reviewer’s line of reasoning. However, as systematic reviewers we can synthesise only the findings of empirical studies that were done, not the ones that ideally should have been done! The limitation of many primary studies – that follow-up was too short to be able to capture the incidence of diabetes over several years – has been acknowledged on page 15, where we also recommend that primary studies of diabetes prevention programmes should be resourced to undertake long-term follow-up.

Figure 7 is mentioned in the text (page 10) as giving the results of the metformin trials, but it should say figure 8. Are all the outcomes reported in Figure 7 well after the intervention has stopped? If so, how long after?

We thank the reviewer for identifying this typo. This has been corrected on page 10.

We agree with the reviewer’s line of reasoning. The majority of trials assessing metformin as an intervention report their outcome after the participants had been on the medication for the pre-determined trial period. The only trial that attempted to assess the potential benefit and long term use of metformin was the long term follow up of the US DPP (US DPPOS). Further discussion has been added to page 10 to clarify this point.

Not all trials were identified. I didn’t see the Melbourne Diabetes Prevention trial by Dunbar

The stated scope of this systematic review (which reflects current UK policy and also policy in other countries such as
et al mentioned anywhere (BMJ Open Diabetes 2015) which is odd, since that trial was one of prevention of diabetes in a “real life” setting. Perhaps that was because screening was by FindRISC not a measure of blood glucose. But it was “screen and treat” so it seems relevant, especially as Barry and colleagues call (foot of page 15) for “pragmatic real-world studies”. the USA) was screening programmes for prediabetes (however defined) and treatment interventions offered to those with prediabetes. As such, the Dunbar study (which we agree attempted to emulate real-world conditions) did not meet our inclusion criteria.

On pages 15 and 16 of the revised paper, we have pointed out that some programmes use risk scores to identify at-risk populations and that research on such programmes could usefully be synthesised.

Table 3 is difficult to use because the trials are in no logical order – alphabetical would have made checking things much easier. It would also have been useful to link different studies from the same trial, such as Lindstrom (not mentioned in table 3, only in figure 7) and Tuomilehto.

We accept this criticism and have amended Table 3 to list the studies alphabetically by first author.

The post-GDM studies might be better in a separate table and analysed separately, though I note (page 11, para 3) that only one GDM trial was included in the meta-analysis.

We agree with the reviewer. The papers relating to populations with a history of gestational diabetes have been placed in separate tables. See now table 3B and 4B (page 33 and page 38).

The Perez-Ferre 2015 trial of Mediterranean lifestyle after GDM appears to have been missed.

Many thanks for identifying this trial, which we have now added to the sample. We have amended the results throughout the paper.

Looking at some of the figures 2 and 3, there is clearly considerable heterogeneity amongst studies. In such circumstances, a strategy to address heterogeneity is required. The methods section (page 5, para 2) says that heterogeneity was assessed using the I2 test. The I2 results are tucked away in an appendix and not reported in main text.

We agree that the I2 statistic is very high for the FPG and IGT comparison but systematic reviews of diagnostic accuracy often combine studies that are quite different with regard to the patient populations and the methodology used and heterogeneity is to be expected. We take a pragmatic view and think the meta-analysis is worth doing but we should be careful and assess whether the heterogeneity among the trials affects the conclusions of the review. However, we agree that we have not included these checks and balances in the manuscript as it stands and we have added sensitivity analysis assessing the effect of studies at high risk of bias and have also done additional analyses that assess the effect of “outliers” by excluding studies with empirical diagnostic odds ratios > 5 and less than 0.2. These analyses have been added to the appendix and discussed on page 8-9.
The I squared test for the lifestyle intervention meta-analysis is displayed on figure 6-8. Statistical heterogeneity was only detected in two of the intervention meta-analyses, at a moderate level in the 3-6 year follow up analysis (I²=45%) and at a substantial in the post intervention follow up analysis (I²=82%). We considered exploring this substantial heterogeneity with subgroup analysis, but with only 5 studies contributing to the analysis subgroup analysis would have been challenging.

The heterogeneity of the intervention studies is taken into account when performing the GRADE analysis and was an important consideration leading to the downgrading of the strength of the evidence. In fact, we downgraded our confidence in the post intervention follow up analysis twice in view of the substantial heterogeneity. We discuss possible explanations for heterogeneity at the beginning of the results section titled ‘Interventions to prevent diabetes in screen-detected pre-diabetes’.

Additional discussion regarding the heterogeneity of the studies added to page 8 and 11.

I liked figure 9 as a neat illustration of selection bias. Thanks!

Two of the appendices use “IGT as the reference standard”. But IGT is a condition, not a test. It is defined through the OGTT as having normal FPG but raised post-load results. So presumably IGT is short-hand for 2-hour OGTT results. But if so, was that not covered in appendices 4 and 5?

The reviewer is correct and we have amended the text accordingly in the Appendix (pages 5-8).

Other odd results in the appendices are not discussed in the text. Appendix 4 used HbA1c threshold of 6.0% and gets sensitivity of 41%. We would expect that to be much less sensitive, but sensitivity actually increases to 52%. Appendix 5 uses a higher cut-off of 6.4%. We would expect that to be much less sensitive, but sensitivity actually increases to 52%. Appendices 7 and 8 use different cut-offs for FPG, of 6.0 and 6.9 mmol/l, but get almost identical sensitivity and specificity. This is not credible, and presumably reflects differences between studies using different thresholds. In appendix 8, there is one marked outlier which by the size of the circle, must carry more weight than any other study. The I² was 94%.

We agree the comparing average sensitivities this way is counter-intuitive and sensitivity should decrease as you move the threshold of the test in the direction of disease. However, the comparisons involve different studies and different people and some of the difference may be explained by the heterogeneity across studies. Moreover, the pooled sensitivity measure is sensitive to the range of specificities (and FPR’s) of the studies included in the analysis and it is better to compare sensitivity at a fixed FPR. For example, for appendix 5 and 6 it can be seen from eyeballing the graphs that if the sensitivity for a FPR of 0.2, a cut-off of 6.0% has sensitivity of 0.55 and for the higher cut-off the sensitivity at FPR=0.2 is 0.4 which is in the intuitive direction and opposite to the averages. Choosing a FPR of 0.1 for the graph in appendix 8 we get a sensitivity of approximately 0.4 and in appendix 9 a FPR of 0.1 the
<table>
<thead>
<tr>
<th>Cell 1</th>
<th>Cell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perhaps in Discussion, there should be some discussion of studies showing that HbA1c and FPG identify similar numbers of people with, for example, diabetes, but that the populations do not overlap well – the tests are picking up different groups. Mostafa et al in Leicester is one such study.</td>
<td>Many thanks for identifying this; we have added this point on page 12.</td>
</tr>
<tr>
<td>Also for Discussion, is why screening should be done. Most people with type 2 diabetes die of cardiovascular disease. HbA1c is a better predictor of that than FPG. However the ADDITION trial (in diabetes not IGT) should no benefit of intensive intervention to reduce CVD.</td>
<td>We agree with this reviewer that even if an intervention successfully prevents type 2 diabetes it MAY not therefore extend life or prevent diabetes-related morbidity – though in our view, this is a highly controversial and much-debated suggestion. However, type 2 diabetes is costly to monitor and treat, so preventing it is arguably a good thing for those reasons. We respectfully suggest that in view of the complex arguments on both sides, the question of whether it’s worth preventing diabetes might be best discussed in the correspondence columns or in an accompanying editorial.</td>
</tr>
<tr>
<td>Discussion para 2, last sentence. IFG is a condition – the test is FPG.</td>
<td>Many thanks for identifying this. This has been amended.</td>
</tr>
<tr>
<td>Definition of terms IEC diagnostic criteria – diagnostic of what?</td>
<td>Many thanks for identifying this. This has been amended.</td>
</tr>
<tr>
<td>Methods. Most systematic reviews have data extractions checked in more than 20% of studies</td>
<td>We would be happy to check a higher proportion of studies in our sample if this is considered key (though we selected 20% on the basis that several recent large systematic reviews used this proportion).</td>
</tr>
</tbody>
</table>