## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | A COST EFFECTIVENESS COMPARISON OF THE NICE 2015 AND WHO 2013 DIAGNOSTIC CRITERIA FOR WOMEN WITH GESTATIONAL DIABETES WITH AND WITHOUT RISK FACTORS |
| **AUTHORS** | Jacklin, Paul; Maresh, Michael; Patterson, Chris; Stanley, Katharine; Dornhorst, Anne; Burman-ROY, Shona; Bilous, Rudy |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER** | Richard Edlin<br>University of Auckland, Auckland, New Zealand<br><br>No relevant conflicts. As a non-relevant issue, part of my time is spent on a trial also considering the cost-effectiveness of two different screening criteria for GDM. |
| **REVIEW RETURNED** | 17-Mar-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | The paper considers two questions, of which only one appears in the title of the paper. The comparison of the NICE 2015 and WHO 2013 criteria is relevant, although is notably not an exhaustive list of GDM criteria available (and may limit international relevance). The second question, of the breadth of screening (e.g. with risk factors, without risk factors or universal) is relevant but is not in the title. The authors correctly consider the cost-effectiveness of screening in those with and without risk factors, although the references to "universal screening" should really be changed to relate instead to "without risk factors", since this is what the paper considers.<br>The data in the study comes from a variety of sources and is claimed to be representative of the UK population. Not enough is really discussed around this, as the results (in terms of cost-effectiveness at 30k per QALY) do vary a little on which dataset is considered, and this could and should be fleshed out more clearly.<br>The authors could have set the paper out to highlight issues more clearly. Whilst the criteria is an issue, this is secondary to which groups should be diagnosed – those who would be diagnosed under any system, or never diagnosed – are less relevant. In this case, there are four (or more) mutually exclusive groups of potential interest - those who will not be diagnosed on any criteria, those who would be diagnosed on NICE criteria only, those diagnosed on WHO criteria only, and those who would be diagnosed on both criteria. Of these, the first group is not interesting --- there is no impact of the decision problem on these people and no reason for them to be included in any analysis. For the rest, I suspect that some of the trickier elements of the discussion (e.g. on Page 24) would have been much easier with these groups separated.<br>Most of the paper is well done. However, I am very puzzled by the statement there are no QALYs for women who aren't screened/treated. This isn't correct, and makes a nonsense of the |

| | |
|---|---|
| | main results of the study – I can see why this statement is made but what's reported should be modified so that this is more carefully (and accurately) stated ... the key is that with the model as provided you can't get a Net Monetary Benefit, and what is provided is only ever an incremental net monetary benefit vs. No Treatment.<br>The presentation of results is also of inconsistent quality. Generally, deterministic results are of lower quality and are likely biased, and the probabilistic costs and QALYs should be displayed. If only one set of results are to be displayed, it needs to be the probabilistic ones. I am concerned that it appears that there are very large differences in the costs and QALYs between the probabilistic and deterministic cases - the net monetary benefits for the deterministic case differ from the reported probabilistic figures by around £70,000-80,000 pounds.<br><br>The limitations of the paper could be broader than discussed, e.g. relating to representativeness, to the lack of head-to-head data, and to what appears to be at best uncertainty, and at worst omission, of incremental costs and QALYs deficits for babies attributable to GDM.<br><br>Substantive comments that should be addressed:<br>Lines 44-45. The statement "compared to £37,669 ... criteria." is misleading. This comparison is rightly between WHO and NICE criteria, although it is written as if it is a comparison between WHO criteria and no treatment.<br>Line 65. Whilst the authors state that the analysis provides "strong evidence", substantial uncertainty around cost-effectiveness exists.<br>Line 206 (table). The relative treatment effects are stated, but in an unhelpful way. It would make more sense to clearly identify the distribution and parameters in a supplementary table, and reserve the table in the paper proper for information that is likely to be of relevance to a wider readership. It is unclear, for instance, how much uncertainty there really is when the RR is given and a standard error of the parameter is provided as a log-RR. An interval showing a 95% credible interval for the parameter would be much more helpful. Note also that the source in many cases is not ACHOIS (2005) and Landon (2009) -- instead, it is the meta-analysis the authors have conducted.<br>Line 220. It is trivial, but the currency should be stated as well as the base year.<br>Line 236. The 2.2 QALYs should be decomposed at least into the three types, as it's too brief at present in the main paper.<br>Line 248. 2000 iterations is really not very many iterations at all ... it may have converged at that level, but I can't tell - and nor could any readers. The authors should check and report.<br>Lines 252-254. It seems that the authors have applied the Cholesky Decomposition to the variance-covariance matrix for parameters in the logistic regression, and then used these to draw parameters for lognormal distributions. As stated, however, the authors state that they sampled from a multivariate normal distribution which, if right, is a poor approach. Can the authors please check and tidy up the wording? Note that in the CHEERS checklist, the authors state that they have provided the distributions that they use, but that do not appear to have provided these parameters and so cannot be said to have met this requirement.<br><br>Line 274 (table). This table is ***critically*** incorrect (as above) |

| | |
|---|---|
| | when it reports no costs and no QALYs for women under "No Treatment".<br>I'm a little bit concerned too that whilst there is a lot of information provided in the supplementary material, some of the issues I'd prefer to see haven't been included. For example in Tables x1 to x6, we have a series of models BUT no indication of fit.<br><br>Comments that should be considered:<br>Line 222. In Table x7, it's stated that a normal distribution is used for costs. It's unlikely to make a difference but this really should be a gamma or lognormal distribution instead.<br>Line 233. I'm unclear why these probabilities are provided without uncertainties, as these seem to be relevant.<br>Line 312. Some form of EVPI would improve the quality of the results.<br>Supplementary Tables x9, x10. Reorder these -- risk factor columns first, then non-risk factors. It's much less clear mixing these up by data source, when the columns deal ultimately with two discrete decision problems that have broadly similar results.<br><br>Minor comments:<br>Line 36. Missing space. Use "four" rather than "4".<br>There are a *lot* of formatting issues with partly blank pages and broken links.<br>Figure 1 appears twice in the paper (as provided) and without relevance or caption on Page 10.<br>Line 264. Table 5-7, rather than "Table 5, Table 6, Table 7"<br>Line 271 (table). Is there a stray "return" in the second row?<br>Supplementary, Table x2 to x6. Are some of these meant to be Model 2?<br>Supplementary, line 185. Table 241? |

| | |
|---|---|
| **REVIEWER** | Trevor Sheldon<br>University of York, UK<br><br>I am a c-author of an HTA report on the same topic area. |
| **REVIEW RETURNED** | 20-Mar-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | This paper aims to compare the cost-effectiveness of two sets of recommendations for the detection of gestational diabetes. Those mainly differ according to the diagnostic thresholds of maternal hyperglycaemia and target populations. The 2013 WHO endorsed IADPSG thresholds are much lower than hitherto used and imply more people being treated. The other is the UK NICE guidelines which were published in 2015, which is more conservative.<br>It is important that practice and policy be informed by economic evaluation as resources are limited and spending has opportunity costs in other areas of health care.<br>The authors conduct a high quality evaluation which reports that NICE guidance is more cost-effective. The authors I believe were involved in the NICE guidance development (which itself was informed by economic analysis) but the analysis is open enough for scrutiny to not be biased by this.<br>Whilst it is generally a well conducted study and the comparison worthy of publication, there are some revisions that would strengthen it. |

| | |
|---|---|
| | 1) A big omission is that they do not refer to a recent NIHR Health Technology Assessment programme–funded economic analysis (Farrar D et al. The identification and treatment of women with hyperglycaemia in pregnancy: an analysis of individual participant data, systematic reviews, meta-analyses and an economic evaluation. Health Technology Assessment 2016; 20, (86). https://www.journalslibrary.nihr.ac.uk/hta/hta20860#/abstract. This comes to different conclusions - that routinely identifying pregnant women for treatment of GDM to reduce adverse perinatal outcomes is not cost-effective. This report came out in December 2016 so they possibly missed it, but it cannot be ignored I think.<br>Given this, it is not surprising that the paper finds that the "NICE diagnostic criteria for GDM are more cost effective than the WHO criteria " as NICE recommends a more cautious approach with higher thresholds. If something is not particularly cost-effective in general then doing less of it or more targeted delivery is likely to be more cost-effective. I think the authors make a good case that the NICE guidelines lead to more cost effective practice than the WHO supported guidance. The bigger question perhaps, is whether it is sufficiently cost-effective to warrant routine detection of GDM.<br>This manuscript and the published report differ in their estimates. There are a number of reasons why they might differ (e.g. structure of the model, assumptions about costs, risks and effectiveness of treatments, what benefits and outcomes are included etc). This needs careful and thorough discussion in the paper. For example:<br>i. intervention is likely to appear comparatively more cost-effective in this manuscript because they omit a cost for instrumental delivery, the risk of which is increased by treatment. Intervention strategies may also appear more cost-effective in the HAPO (four centres) based on higher baseline risk of GDM and of adverse perinatal outcomes.<br>ii. the regression models underpinning the economic analysis in the NICE guideline did not incorporate fasting blood glucose levels, as this covariate had been dropped in the stepwise selection process. The HTA analysis did not use a stepwise selection process, and its risk models for adverse perinatal outcomes all include both fasting and post-load glucose levels.<br>iii. both analysis estimate risk models on different datasets, and use different model specifications, meaning they will not produce the same results.<br>The paper would be further improved if the following points are taken into account in any revision:<br><br>2) The authors should report the probabilistic ICER. This is now standard practice in preference to the deterministic approach (or at least compare them), as the deterministic ICER will be biased in the presence of a non-linear model. I expect that the ICER will be considerably lower for the deterministic model and so possibly over favouring GDM routine detection. It will not though affect the result that NICE is more cost effective than the WHO-endorsed approach.<br>3) It seems as if the results in this paper are more favourable than the NICE (Table 118 in the NICE guidelines), if so this needs some explanation.<br>4) The referencing is rather limited at times and in other places also out of date. There are many treatment reviews apart from Hovath. HAPO is not the only paper examining glucose and outcomes (line 76) (e.g. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. BMJ 2016; 354 doi: https://doi.org/10.1136/bmj.i4694) |

| | |
|---|---|
| | 5) Rather than refereeing to the WHO criteria, perhaps state these are WHO-endorsed IADPSG criteria<br>6) The authors need to be clear what variables are included in the various data sets used (e.g. Norwich Atlantic Dip etc). Atlantic DIP does not include previous macrosomic infant or previous GDM, both NICE recommended risk factors. So they need to be much more explicit about what is and isn't included in the datasets. I appreciate this is difficult given word limits.<br>7) What assumptions of uptake of the diagnostic were included? I could not see this in the model. Those screening positive but not taking up the diagnostic test add costs with no benefits and it also affects impact overall.<br>8) They assumed that the majority of women would be treated with insulin, whereas this does not necessarily reflect actual routine practice.<br>9) IADPSG criteria aim to reduce the risk of infant obesity through its association with macrosomia. However, there is now some longer term follow up of the Crowther and Landon trials (refs below) which suggest that there is no difference in adiposity between treatment groups at 5 years. This might be worth refereeing to as it further undermines th WHO-endorsed gudelines. Landon et al.. Mild Gestational Diabetes Mellitus and Long-Term Child Health. Diabetes Care. 2015;38(3):445-52. |

**VERSION 1 – AUTHOR RESPONSE**

# Responses to BMJ Open Peer Review

## Peer Reviewer 1

**********
Reviewer(s)' Comments to Author:
**********
Reviewer: 1
Richard Edlin
University of Auckland, Auckland, New Zealand
Please state any competing interests or state 'None declared':
No relevant conflicts. As a non-relevant issue, part of my time is spent on a trial also considering the cost-effectiveness of two different screening criteria for GDM.
**********
The paper considers two questions, of which only one appears in the title of the paper. The comparison of the NICE 2015 and WHO 2013 criteria is relevant, although is notably not an exhaustive list of GDM criteria available (and may limit international relevance). The second question, of the breadth of screening (e.g. with risk factors, without risk factors or universal) is relevant but is not in the title. The authors correctly consider the cost-effectiveness of screening in those with and without risk factors, although the references to "universal screening" should really be changed to relate instead to "without risk factors", since this is what the paper considers.

The comment about the title is a reasonable one and we have amended it to make it explicit that the diagnostic criteria are assessed in populations with and without risk factors for GDM.

However, we think the reviewer is mistaken to make a distinction between 'without risk factors' or 'universal screening'.

The incremental costs and effects of going from risk factor screening to universal screening is measured in a population "without risk factors" (by definition those incremental costs and benefits accrue in the population not previously screened – i.e. those without risk factors).

For the HAPO and Atlantic DiP datasets we first assess

- No screening v Risk factor screening ('with risk factors')
- Risk factor screening v Universal screening ('without risk factors')

For our Norwich dataset we could not divide the population against risk factor and therefore the comparison made was:

- No screening v universal screening

The reviewer is correct that the comparison of NICE 2015 and WHO 2013 diagnostic criteria is not exhaustive and that our analysis in the NICE guideline compared a broader range of alternatives. Frequent references are made to that guideline in the document. We have made a small edit to the text to provide a rationale for choosing NICE 2015/WHO 2013 criteria in that these thresholds have been actually proposed as new diagnostic criteria for GDM by national and international bodies.

The data in the study comes from a variety of sources and is claimed to be representative of the UK population. Not enough is really discussed around this, as the results (in terms of cost-effectiveness at 30k per QALY) do vary a little on which dataset is considered, and this could and should be fleshed out more clearly.

We have added a table to the Supplementary Report to illustrate the ethnic mix within our datasets.

# Responses to BMJ Open Peer Review

| | HAPO (4) centres | Atlantic DiP | Norfolk | UK |
|---|---|---|---|---|
| White | 79% | 93% | 91% | 87% |
| Black | 2% | 2% | 1% | 3% |
| Asian | 13% | 4% | 2% | 7% |
| Other | 6% | 1% | 6% | 3% |

The authors could have set the paper out to highlight issues more clearly. Whilst the criteria is an issue, this is secondary to which groups should be diagnosed – those who would be diagnosed under any system, or never diagnosed – are less relevant. In this case, there are four (or more) mutually exclusive groups of potential interest - those who will not be diagnosed on any criteria, those who would be diagnosed on NICE criteria only, those diagnosed on WHO criteria only, and those who would be diagnosed on both criteria. Of these, the first group is not interesting --- there is no impact of the decision problem on these people and no reason for them to be included in any analysis. For the rest, I suspect that some of the trickier elements of the discussion (e.g. on Page 24) would have been much easier with these groups separated.

Whilst we understand the point the reviewer is making here, it is complicated by a "no screen/no treatment option". Those diagnosed by both NICE 2015/WHO 2013 criteria would not be diagnosed under a no screen/no treat strategy. Given the findings of the recently published HTA to which the other peer reviewer drew our attention, we think it is important to maintain the "no screen/no treat" comparator.

Most of the paper is well done.

No response required

However, I am very puzzled by the statement there are no QALYs for women who aren't screened/treated. This isn't correct, and makes a nonsense of the main results of the study – I can see why this statement is made but what's reported should be modified so that this is more carefully (and accurately) stated ... the key is that with the model as provided you can't get a Net Monetary Benefit, and what is provided is only ever an incremental net monetary benefit vs. No Treatment.

We think that this comment is something that the reviewer has inferred from our tables as we don't make a "statement there are no QALYs for women who aren't screened/treated".

| Diagnostic threshold | Cost | QALY | Incremental cost | Incremental QALY | ICER |
|---|---|---|---|---|---|
| No Treatment | £0 | 0.00 | n/a | n/a | n/a |
| NICE 2015 | £546,349 | 23.68 | £546,349 | 23.68 | £23,073 |
| WHO 2013 | £778,993 | 29.86 | £232,644 | 6.18 | £37,669 |

Nevertheless, we don't accept that this "isn't correct, and makes a nonsense of the main results of the study" as it is simply a matter of presentation and has no bearing on the ICER or the determinant of what is cost-effective for a given threshold.

# Responses to BMJ Open Peer Review

The model calculates a disutility (or a negative QALY) arising from serious perinatal complications. This leaves a situation where no screen/no treat has the greatest negative QALYs. However, in terms of presentation of the results we believe it is far easier for the reader if the results are presented in terms of positive QALYs. Therefore the QALYs are calculated relative to no screen/no treat. Costs are also rebased as the difference from no screen/no treat and, of course, this makes no difference to the Incremental cost, Incremental QALY or ICER. Indeed it is standard practice in health economic evaluation for results to be compared relative to a "do nothing" baseline.

For the probabilistic analysis we use the net monetary benefit approach. Again it is common practice to calculate the incremental net monetary benefit relative to no treatment. As noted by Briggs et al. (2007) the NMB is a rearrangement of the cost-effectiveness decision rule such that an intervention is cost-effective if:

$$NMB = \gamma \times \Delta E - \Delta C > 0$$

Where:
γ willingness to pay
ΔE Incremental effects in QALYs
ΔC Incremental costs in costs

Where there are multiple strategies the most cost-effective strategy is the one with the highest net benefit. If no treatment was the most cost-effective then the other strategies would have a negative net monetary benefit.

The reason that no screen/no treatment has a NMB > 0 is because in the probabilistic analysis the net monetary benefit is calculated relative to the lowest cost and lowest QALY treatment in any particular simulation, which isn't always no screen/no treat.

We have added a footnote to tables where appropriate to make this clear that the no screen/no treat strategy represents the baseline

The presentation of results is also of inconsistent quality. Generally, deterministic results are of lower quality and are likely biased, and the probabilistic costs and QALYs should be displayed. If only one set of results are to be displayed, it needs to be the probabilistic ones.

With the exception of a single sensitivity analysis undertaken to assess the impact of retaining non-significant blood glucose covariates, probabilistic and deterministic results are given for every analysis. We have not sought to prioritise presentation of deterministic results and the probabilistic and deterministic analyses give a consistent result.

I am concerned that it appears that there are very large differences in the costs and QALYs between the probabilistic and deterministic cases - the net monetary benefits for the deterministic case differ from the reported probabilistic figures by around £70,000-80,000 pounds.

| **Diagnostic threshold** | **Cost** | **QALY** | **Incremental cost** | **Incremental QALY** | **ICER** | **NMB** |
|---|---|---|---|---|---|---|
| No Treatment | £0 | 0.00 | n/a | n/a | n/a | |
| NICE 2015 | £546,349 | 23.68 | £546,349 | 23.68 | £23,073 | £164,051 |
| WHO 2013 | £778,993 | 29.86 | £232,644 | 6.18 | £37,669 | £116,807 |

# Responses to BMJ Open Peer Review

| Diagnostic threshold | NMB (CE threshold £30,000 per QALY) |
|---|---|
| No Treatment | £391 |
| NICE 2015 | £233,192 |
| WHO 2013 | £200,384 |

Above, we list 2 tables from our original submission to BMJ Open. The deterministic results are shown in the top table and the probabilistic in the bottom table. We've added the Net Monetary Benefit column in the top table to demonstrate the issue to which the reviewer is referring.

However, we would note that it is recognised that probabilistic sensitivity analyses combining evidence from a wide variety of sources and using distributions assigned to a range of parameters are likely to exhibit some non-linearity. The NICE 2015 NMB of £233,192 in the Table above is derived from an incremental costs relative to no treatment of £528,778 and incremental QALYs of 25.40. We do not consider these differences from the deterministic result to be concerning and both deterministic and probabilistic sensitivity analyses reach the same conclusion as to which strategy is cost-effective at a threshold of £30,000 per QALY.

The limitations of the paper could be broader than discussed, e.g. relating to representativeness, to the lack of head-to-head data, and to what appears to be at best uncertainty, and at worst omission, of incremental costs and QALYs deficits for babies attributable to GDM.

We have extended our discussion, especially with respect to the recently published HTA.

Substantive comments that should be addressed:
Lines 44-45. The statement "compared to £37,669 ... criteria." is misleading. This comparison is rightly between WHO and NICE criteria, although it is written as if it is a comparison between WHO criteria and no treatment.

We have amended our wording to make this clearer.

Line 65. Whilst the authors state that the analysis provides "strong evidence", substantial uncertainty around cost-effectiveness exists.

We accept the point made by the reviewer although there was a greater level of certainty with respect to this result than for other results that were presented. We have amended so that it now refers to "clear evidence"

Line 206 (table). The relative treatment effects are stated, but in an unhelpful way. It would make more sense to clearly identify the distribution and parameters in a supplementary table, and reserve the table in the paper proper for information that is likely to be of relevance to a wider readership. It is unclear, for instance, how much uncertainty there really is when the RR is given and a standard error of the parameter is provided as a log-RR. An interval showing a 95% credible interval for the parameter would be much more helpful.

We think that this table is consistent with economic evaluation papers published in the BMJ – which give the parameters of the probability distribution sampled in the probabilistic sensitivity analysis http://bmjopen.bmj.com/content/5/10/e007925.full.pdf+html (Table 1).

# Responses to BMJ Open Peer Review

The HTA alluded to by the 2nd reviewer presents its table of relative treatment effects in the same way we have reported them in our paper.

Note also that the source in many cases is not ACHOIS (2005) and Landon (2009) -- instead, it is the meta-analysis the authors have conducted.

The table refers to the source of the data. We do state in the main body of the text that treatment effectiveness was estimated from "a random-effects meta-analysis of two studies".

Line 220. It is trivial, but the currency should be stated as well as the base year.

We've amended the text to state the currency.

Line 236. The 2.2 QALYs should be decomposed at least into the three types, as it's too brief at present in the main paper.

We think that we clearly signpost that further detail is available in the supplementary report to the interested reader.

Line 248. 2000 iterations is really not very many iterations at all ... it may have converged at that level, but I can't tell - and nor could any readers. The authors should check and report.

Our model does not use a Markov Chain Monte Carlo (MCMC) method and convergence is not an issue. Nevertheless, we have re-run the analysis for n = 10,000 and it makes negligible difference (as shown below).

**HAPO (4) centres Risk Factors All covariates Backward Elimination of Blood Glucose Coefficients with non-significant coefficients (n = 2,000)**

| Dx Threshold | Mean Net Benefit | Probability cost effective |
|---|---|---|
| No treatment | £486 | 19.8% |
| Fasting 5.6/2hr 7.8 | £230,798 | 53.5% |
| IADPSG 1.75 (incl 1hr) | £178,231 | 26.8% |

**HAPO (4) centres Risk Factors All covariates Backward Elimination of Blood Glucose Coefficients with non-significant coefficients (n = 10,000)**

| Dx Threshold | Mean Net Benefit | Probability cost effective |
|---|---|---|
| No treatment | £958 | 20.7% |
| Fasting 5.6/2hr 7.8 | £233,637 | 52.0% |
| IADPSG 1.75 (incl 1hr) | £180,919 | 27.3% |

Lines 252-254. It seems that the authors have applied the Cholesky Decomposition to the variance-covariance matrix for parameters in the logistic regression, and then used these to draw parameters for lognormal distributions. As stated, however, the authors state that they sampled from a multivariate normal distribution which, if right, is a poor approach. Can the authors please check and tidy up the wording?

# Responses to BMJ Open Peer Review

We have checked and we think our wording is correct

> *"if the probability parameter is estimated from a logistic regression, then the parameters of interest are the coefficients on the log-odds scale and multivariate normality on this scale would be the appropriate assumption"*
> **Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Economics 14(4):pp. 339-347. (Claxton et al. 2005)**

The authors of that book also have a text book "Decision Modelling for Health Economic Evaluation" and state:

> *"A common criticism of probabilistic sensitivity analysis is that parameters are assumed to be independent…….it is possible to correlate parameters if the covariance structure is known……..One example where we clearly do know the covariance relationship of parameters is in a regression framework where we have access to the variance-covariance matrix. In this situation we can employ a technique known as Cholesky decomposition to provide correlated draws from a* ***multivariate normal distribution****"*

We already include a reference to this book when outlining the method.

Note that in the CHEERS checklist, the authors state that they have provided the distributions that they use, but that do not appear to have provided these parameters and so cannot be said to have met this requirement.

The standard errors for the regression coefficients are given in the tables in the supplementary report. The additional information that somebody would need to duplicate our approach is the variance covariance matrix. Other papers have reported these matrices.

http://heart.bmj.com/content/suppl/2016/02/10/heartjnl-2015-308850.DC1/heartjnl-2015-308850supp_appendixC.pdf

https://www.york.ac.uk/che/pdf/tp28.pdf

We have added the variance covariance matrices for the base case analyses. Due to the large number of outcomes and regression models we have not included these for sensitivity analyses. We would be happy to provide all the variance covariance matrices if the editor thought that appropriate. Alternatively, we could note in the paper that these matrices are available from the authors on request.

Line 274 (table). This table is ***critically*** incorrect (as above) when it reports no costs and no QALYs for women under "No Treatment".

As noted previously this arises because "No Treatment" is used as the baseline from which the additional costs of screening/diagnosis/treatment are measured. We have added a footnote to the relevant tables for clarity.

I'm a little bit concerned too that whilst there is a lot of information provided in the supplementary material, some of the issues I'd prefer to see haven't been included. For example in Tables x1 to x6, we have a series of models BUT no indication of fit.

We are grateful to the reviewer for this comment and have now added measures of goodness of fit to the tables and added a discussion of these measures in the manuscript.

Comments that should be considered:

# Responses to BMJ Open Peer Review

Line 222. In Table x7, it's stated that a normal distribution is used for costs. It's unlikely to make a difference but this really should be a gamma or lognormal distribution instead.

The reviewer has failed to appreciate that central limit theorem is involved here. It is true that the 'population' NHS Reference Cost data is usually skewed and that a gamma or lognormal distribution would fit that data better. However, our model inputs are based on the distribution of the mean cost of all submissions to NHS Reference Costs and not the underlying population distribution.

This is explained in more detail in the NICE guideline on p575 (section 9.2.2.4.3). We've added a footnote to Supplementary Report Table x.7 referring to the NICE guideline for more detail.

Line 233. I'm unclear why these probabilities are provided without uncertainties, as these seem to be relevant.

The reviewer raises a legitimate point. Our rationale is that quantifying the uncertainty is more problematic for these variables and because they have relatively small impact on results when varied within plausible ranges. The recently published HTA followed the NICE approach (our approach here) and also treated these variables deterministically.

Line 312. Some form of EVPI would improve the quality of the results.

Whilst recognising a place for the use of EVPI it is not something that is included on checklists used to appraise the quality and reporting of economic evaluations - e.g. Cheers Statement (BMJ, 2013), Drummond (BMJ, 1996), Philips (HTA, 2004).

Whilst, there are uncertainties in our analysis arising from sampling we would argue that in this case, larger uncertainties arise from limitations in current understanding about the pathogenesis and prevention of disease. So, for example, as we discuss, in our paper there is uncertainty about the extent to which a diagnosis of GDM averts or delays onset of Type 2 diabetes mellitus. This was not incorporated into our model as we do not consider that this relationship has been sufficiently well established at this time. There is also conflicting evidence and views on the association between maternal glycaemia and subsequent cardio-metabolic outcomes in human offspring.

Supplementary Tables x9, x10. Reorder these -- risk factor columns first, then non-risk factors. It's much less clear mixing these up by data source, when the columns deal ultimately with two discrete decision problems that have broadly similar results.

Again we appreciate the comment but do think that the clarity being referred to is somewhat subjective. Furthermore, as well as grouping by dataset we also wanted to group by regression model and therefore on balance we would rather not make this change.

Minor comments:
Line 36. Missing space. Use "four" rather than "4".

We have corrected this.

There are a *lot* of formatting issues with partly blank pages and broken links.

BMJ Open converts the Word document to a PDF which the reviewer sees. We are not aware how we can fix these formatting issues but presume it will be straightforward for the BMJ Open editors.

Figure 1 appears twice in the paper (as provided) and without relevance or caption on Page 10.

# Responses to BMJ Open Peer Review

Issue with Word to PDF conversion perhaps as our version does not have this Figure repeated on page 10.

Line 264. Table 5-7, rather than "Table 5, Table 6, Table 7"

We agree that's better but have used the latter approach for cross-referencing purposes in Word. We would be happy for BMJ Open to make this change

Line 271 (table). Is there a stray "return" in the second row?

We have corrected this.

Supplementary, Table x2 to x6. Are some of these meant to be Model 2?

We have amended this table to correct this mistake but also to make the ordering of the columns more logical.

Supplementary, line 185. Table 241?

Our Word version says Table x24 which is correct.

# Responses to BMJ Open Peer Review

## Peer Reviewer 2

**********
Reviewer: 2
Trevor Sheldon
University of York, UK
Please state any competing interests or state 'None declared': I am a co-author of an HTA report on the same topic area.
**********
Please leave your comments for the authors below This paper aims to compare the cost-effectiveness of two sets of recommendations for the detection of gestational diabetes. Those mainly differ according to the diagnostic thresholds of maternal hyperglycaemia and target populations. The 2013 WHO endorsed IADPSG thresholds are much lower than hitherto used and imply more people being treated. The other is the UK NICE guidelines which were published in 2015, which is more conservative.
It is important that practice and policy be informed by economic evaluation as resources are limited and spending has opportunity costs in other areas of health care.
The authors conduct a high quality evaluation which reports that NICE guidance is more cost-effective. The authors I believe were involved in the NICE guidance development (which itself was informed by economic analysis) but the analysis is open enough for scrutiny to not be biased by this. Whilst it is generally a well conducted study and the comparison worthy of publication, there are some revisions that would strengthen it.

1) A big omission is that they do not refer to a recent NIHR Health Technology Assessment programme–funded economic analysis (Farrar D et al. The identification and treatment of women with hyperglycaemia in pregnancy: an analysis of individual participant data, systematic reviews, meta-analyses and an economic evaluation. Health Technology Assessment 2016; 20, (86). https://www.journalslibrary.nihr.ac.uk/hta/hta20860#/abstract. This comes to different conclusions - that routinely identifying pregnant women for treatment of GDM to reduce adverse perinatal outcomes is not cost-effective. This report came out in December 2016 so they possibly missed it, but it cannot be ignored I think.

We are grateful to the reviewer for drawing our attention to this paper.

Given this, it is not surprising that the paper finds that the "NICE diagnostic criteria for GDM are more cost effective than the WHO criteria "as NICE recommends a more cautious approach with higher thresholds. If something is not particularly cost-effective in general then doing less of it or more targeted delivery is likely to be more cost-effective. I think the authors make a good case that the NICE guidelines lead to more cost effective practice than the WHO supported guidance. The bigger question perhaps, is whether it is sufficiently cost-effective to warrant routine detection of GDM.
This manuscript and the published report differ in their estimates. There are a number of reasons why they might differ (e.g. structure of the model, assumptions about costs, risks and effectiveness of treatments, what benefits and outcomes are included etc). This needs careful and thorough discussion in the paper.

We agree with the reviewer and have added a discussion of the HTA to our manuscript.

For example:

# Responses to BMJ Open Peer Review

i. intervention is likely to appear comparatively more cost-effective in this manuscript because they omit a cost for instrumental delivery, the risk of which is increased by treatment. Intervention strategies may also appear more cost-effective in the HAPO (four centres) based on higher baseline risk of GDM and of adverse perinatal outcomes.

We agree that instrumental delivery could plausibly increase with treatment, as a reduction in caesarean section would lead to more vaginal births. However, with no treatment there may be more vaginal births and with larger babies, so there is likely to be more need for instrumental deliveries. Hence the effects may cancel themselves out. Reviewing the HTA they used an increased risk of instrumental delivery of 1.37 (Table 31), but this relates to one study on women with an abnormal glucose challenge test and normal GTT (their ref 202) and secondly the confidence intervals are between 0.20- 9.27. We consider that the evidence presented in the HTA would provide a further rationale for not including that outcome within our model. Given the wide confidence intervals around the treatment effect size for instrumental delivery it would be surprising if that was a major cause of any discrepancy between our findings and those of the HTA Report.

ii. the regression models underpinning the economic analysis in the NICE guideline did not incorporate fasting blood glucose levels, as this covariate had been dropped in the stepwise selection process. The HTA analysis did not use a stepwise selection process, and its risk models for adverse perinatal outcomes all include both fasting and post-load glucose levels.

We did include a sensitivity analysis, reported in the supplementary material, which did include fasting blood glucose variables in the regression models.

**Table x23:** Summary of deterministic ICERs for each population without backward elimination of non-significant coefficients

| | All covariates | | Plasma glucose covariates | | | | |
|---|---|---|---|---|---|---|---|
| **Diagnostic threshold** | **HAPO Risk factor (n=3,549)** | **HAPO No Risk factor (n=2,614)** | **HAPO Risk factor (n=3,549)** | **HAPO No Risk factor (n=2,614)** | **Atlantic DiP Risk factor (n=1,988)** | **Atlantic DiP No Risk factor (n=3,302)** | **Norwich (n=12,754)** |
| No Treatment | - | - | - | - | - | - | - |
| NICE 2015 | £22,786 | £46,677 | £24,802 | £39,338 | £22,126 | £37,887 | £31,191 |
| WHO 2013 | £33,876 | £107,247 | £35,852 | £54,288 | £41,652 | £43,106 | £43,694 |

iii. both analysis estimate risk models on different datasets, and use different model specifications, meaning they will not produce the same results.
The paper would be further improved if the following points are taken into account in any revision:

We agree with this point although the same 'decision' would be made in our analyses for the 3 different populations assessed.

2) The authors should report the probabilistic ICER. This is now standard practice in preference to the deterministic approach (or at least compare them), as the deterministic ICER will be biased in the presence of a non-linear model. I expect that the ICER will be considerably lower for the deterministic model and so possibly over favouring GDM routine detection. It will not though affect the result that NICE is more cost effective than the WHO-endorsed approach.

Stinett and Mullahy (1998) made a convincing case for the use of a net benefit framework in order to handle uncertainty in probabilistic sensitivity analysis and we consider that to be standard/common practice in economic evaluation reporting. We have followed such an approach in our paper.

# Responses to BMJ Open Peer Review

As we note in our response to reviewer 1, there is some evidence of non-linearity but generally the cost-effectiveness estimates are similar regardless of whether a deterministic or probabilistic approach is used (although the ICERs do tend to be slightly lower using a deterministic approach).

3) It seems as if the results in this paper are more favourable than the NICE (Table 118 in the NICE guidelines), if so this needs some explanation.

Table 118 in the NICE guideline presents the results of the deterministic analysis for the HAPO (4) with NICE risk factors. The results have actually changed very little as shown below although the ICERS are different because the NICE guideline compared many more alternative diagnostic strategies.

NICE Table 118 (restricted to comparison made in our BMJ Open submission)

| Strategy | Cost | QALY |
|---|---|---|
| No screening/no treatment | £0 | 0.00 |
| NICE 2015 | £560,419 | 22.11 |
| WHO 2013 | £802,2013 | 27.58 |

Results for same HAPO (4) population presented in BMJ Open submission

| Strategy | Cost | QALY |
|---|---|---|
| No screening/no treatment | £0 | 0.00 |
| NICE 2015 | £546,439 | 23.68 |
| WHO 2013 | £778,993 | 29.86 |

The increase in QALYs has arisen because the analysis presented in this paper includes a maternal health state utility which was estimated from quality of life data collected as part of the ACHOIS study. Table 118 in the NICE guideline does not include this maternal utility (the base case analysis in the guideline) but Table 126 in the NICE guideline does report results of a sensitivity analysis where a maternal health state utility was included, with concomitantly lower ICERs.

The costs in the NICE Guideline were updated for this paper and therefore it is to be expected that there would be differences in cost but as the tables above show these differences are relatively small.

4) The referencing is rather limited at times and in other places also out of date. There are many treatment reviews apart from Hovath. HAPO is not the only paper examining glucose and outcomes (line 76) (e.g. Hyperglycaemia and risk of adverse perinatal outcomes: systematic review and meta-analysis. BMJ 2016; 354 doi: https://doi.org/10.1136/bmj.i4694)

We agree and have added a reference to this recent systematic review in the introduction.

5) Rather than refereeing to the WHO criteria, perhaps state these are WHO-endorsed IADPSG criteria

I think we make it clear in the introduction that the WHO 2013 criteria came from IADPSG

*"New diagnostic thresholds were proposed by the International Association of Diabetes in Pregnancy Study Group (IADPSG) based upon the HAPO study levels of plasma glucose when fasting, and at 1 and 2 hours after an oral 75g glucose load that were associated with covariate adjusted odds ratio of*

# Responses to BMJ Open Peer Review

*1.75 relative to the mean glucose value in the whole HAPO cohort on three offspring outcomes: exceeding the 90th centile for birth weight, for cord serum C-peptide concentration and for percent fetal body fat.* ***These diagnostic criteria have been subsequently adopted by the WHO****"*

6) The authors need to be clear what variables are included in the various data sets used (e.g. Norwich Atlantic Dip etc). Atlantic DIP does not include previous macrosomic infant or previous GDM, both NICE recommended risk factors. So they need to be much more explicit about what is and isn't included in the datasets. I appreciate this is difficult given word limits.

This is something we do address (see below) particularly with respect to HAPO (4).

"The NICE risk factor approach could not be replicated exactly because the patient data used in the model do not include information on previous offspring birth weight, and the HAPO (4) dataset does not provide information on previous GDM. Therefore, the comparison in the model was between universal screening and a subset of NICE risk factors."

"Although the NICE risk factor approach could not be replicated exactly, we felt that the approximation used was acceptable, as the only women who would be omitted from the model risk factor population were multiparous and would have had a large baby previously and/or a past history of GDM. This approximation would over-estimate slightly the benefits of universal screening, as the baseline risk in a group designated as being without NICE risk factors present would be over-stated."

However, we have added some text to indicate that the variables included in the Atlantic DiP study also meant that the population could only approximately be sub-divided into those with and without NICE risk factors. We also added text to explain why the Norwich dataset could not be sub-divided by risk factor.

7) What assumptions of uptake of the diagnostic were included? I could not see this in the model. Those screening positive but not taking up the diagnostic test add costs with no benefits and it also affects impact overall.

We did assume 100% uptake because the view of the Guideline Development Group who developed the original NICE guideline on Diabetes in Pregnancy (2008) was that uptake would be much higher in a group screened on the basis of risk factors. The HTA also assumes higher uptake with risk factor screening compared to universal screening but does not assume 100%. As we do not find universal screening to be cost-effective then relaxing the assumption of 100% uptake would only re-inforce our finding for that screening strategy.

In the light of this comment, we investigated the impact of relaxing the assumption of 100% uptake in groups screened on the basis of risk factors but found that it made negligible differences to the result. The reason for this is that the costs of diagnosis represent a very small proportion (≈3%) of total costs.

Original

| **Dx Threshold** | **Cost** | **QALY** | **ICER** |
|---|---|---|---|
| No treatment | £0 | 0.0000 | n/a |
| Fasting 5.6/2hr 7.8 | £546,349 | 26.7815 | £20,400 |
| IADPSG 1.75 (incl 1hr) | £800,725 | 34.3531 | £33,596 |

| | |
|---|---|
| Patients | 3549 |
| DNA Cost 2 sample | £12.58 |

# Responses to BMJ Open Peer Review

DNA Cost 3 sample £16.66

Uptake 90%

| **Dx Threshold** | **Cost** | **QALY** | **ICER** |
|---|---|---|---|
| No treatment | £0 | 0.0000 | n/a |
| Fasting 5.6/2hr 7.8 | £496,179 | 24.1033 | £20,585 |
| IADPSG 1.75 (incl 1hr) | £726,566 | 30.9178 | £33,809 |

8) They assumed that the majority of women would be treated with insulin, whereas this does not necessarily reflect actual routine practice.

The reviewer makes a reasonable point with respect to current treatment and it is something we were aware of when developing the model. We discuss our assumptions and rationale with respect to hypoglycaemic therapy in Section 9.2.4.4 in the NICE guideline. To provide greater clarity to the reader we have added a sentence in the sub-section on costs to state:
"Costing methodology and assumptions are described in greater detail elsewhere (Ref: NICE Guideline 2015)"

9) IADPSG criteria aim to reduce the risk of infant obesity through its association with macrosomia. However, there is now some longer term follow up of the Crowther and Landon trials (refs below) which suggest that there is no difference in adiposity between treatment groups at 5 years. This might be worth refereeing to as it further undermines the WHO-endorsed gudelines.
Landon et al.. Mild Gestational Diabetes Mellitus and Long-Term Child Health. Diabetes Care. 2015;38(3):445-52.

We have added this to our discussion.

**VERSION 2 – REVIEW**

| | |
|---|---|
| **REVIEWER** | Richard Edlin<br>School of Population Health<br>University of Auckland |
| **REVIEW RETURNED** | 05-May-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have signalled in their response that they have taken many of the comments made in my earlier review on board, and I acknowledge this. There are other comments made with which the authors disagree, and where disagreement remains. The tone of some of the authors' comments appears to be somewhat unhelpful. In several cases, the authors provide references to support positions --- but I'd contend that those references support what I have suggested they do in my earlier review, and which they have not always done.<br><br>The authors continue to report that zero QALYs are generated when women are not screened and now state in their report that this is against a baseline of non-screening. If this is the case, then the authors should to revise the tables so that they do not report "Costs" and "QALYs" but incremental costs and incremental QALYs. As originally labelled, the tables were incorrect --- now, it's simply inelegant but probably clear enough.<br><br>The authors also argue against the request to reword their methodology around the use of the Cholesky Decomposition. In doing so they cite Claxton et al. I agree that Claxton et al's approach and wording is correct and wanted the authors' correct their wording accordingly. All that was needed is for the authors to report their methodology accurately, e.g. by making it clear the scale on which the multinormal sampling is taking place. The revised text on Page 14 now makes it clearer that it is the (scaled) parameters from the regression analysis that are drawn -- this wasn't clear in the original document and motivated my comment.<br><br>The authors state that the probabilistic and deterministic results don't differ in a concerning way. This is a subjective judgement. What remains, however, is that the probabilistic results should be those displayed precisely because of the non-linearity. The deterministic results are biased. The authors again provide references for their positions but may misunderstand the implications of those same references.<br><br>The authors define (in their response) an incremental net benefit as being "net benefit". Net monetary benefit should really be reserved for the absolute term lambda*e - c, rather than lambda*delta_e - delta_c. Unfortunately, this is still not done well (and not just by the authors) --- and can lead to misunderstandings and error. As an example of this, the authors clarify that "the reason that no screen/no treamtent has a NMB > 0 is because in the probabilistic analysis the net monetary benefit is calculated relative to the lowest cost and lowest QALY treatment in any particullar simulation, which isn't always no screen/no treat". Here, due to confusion over the net benefit, the authors have not used a consistently-defined baseline |

| | |
|---|---|
| | (in terms of a decision option) in their results. As a result, the different runs of the probabilistic analysis aren't comparable, and it's unclear whether or not they even can be combined. (Even if they can, the lack of a baseline means that whilst overall results may be okay, the specific values provided lack a clear interpretation.) The probabilistic model should be fixed (this is a very minor thing) and re-run to provide results against a consistent baseline of e.g. no screening.<br><br>Whilst I raised the use of EVPI as (only) a possibility, I note the authors' response. They justify not doing something by its non-appearance in CHEERS and the Drummond Checklist, amongst others. The CHEERS checklist covers reporting, not analysis, and the Drummond Checklist is very seriously out of date in terms of methodological guidance. As I recall Prof Drummond's position, it was that the non-appearance of an appropriate methodology in such checklists is not a sufficient reason to not do it. Checklists should never be used to justify poor or outdated methodology in precisely the way that the authors do. The remainder of the argument around not providing EVPI results does not appear to be very relevant --- but again, this was a suggestion rather than something I felt had to be done.<br><br>As a final point --- if the authors contend, as they now appear to on Page 25, that a no identification/no treat policy would not be acceptable to patients or health care providers, then why is this option included in their analysis? An economic analysis ideally considers all relevant outcomes -- and their position would mean that "do nothing" is not then a relevant option for inclusion. |

| | |
|---|---|
| **REVIEWER** | Prof Trevor Sheldon<br>University of York<br><br>I am a co-author of an NIHR HTA report which includes an economic evaluation of GDM detection and treatment approaches and a paper on the same topic which has been submitted for publication. |
| **REVIEW RETURNED** | 12-May-2017 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have taken considerable effort to respond positively to the referees' comments. I will only deal with the response to my comments (Peer reviewer 2) and the overall resulting paper.<br><br>1. The authors have now referred to the HTA report and have adequately dealt with the comparison. They might need to make a stronger case though for what this paper adds. So whilst they say on page 20 line 393-395 that this is one of the most comprehensive assessments (something I agree with), they should perhaps reference the HTA report alongside that as this is also comprehensive and comes to similar conclusions:<br>i. Happy they have addressed the point about instrumental delivery and have made this clear in the discussion.<br>ii. The paper refers to the sensitivity of the deterministic models including fasting blood glucose. In the table x23 this is presented. However the authors fail to then comment of the degree to which this makes difference.<br>iii. Happy this is now dealt with.<br><br>2. The authors should probably lead with the probabilistic results and |

| | |
|---|---|
| | include deterministic for comparison purposes. But I am content that both are presented and that the impact of both approaches are clear.<br><br>3. I am content with this explanation. Should, they not have a sentence on this in the discussion in case people look back at the NICE guidance material? I couldn't see it but might have missed it.<br><br>4. Responded well<br><br>5. Fine (maybe I am being too pedantic in saying they are the IADPSG thresholds and should not always be referred to as the WHO criteria)<br><br>6. This is now clear and acceptable.<br><br>7. The response is acceptable – do they mention this in the discussion. Couldn't see it but might have missed it. In reality uptake rates are not so high and so even if the cost-effectiveness is not that changed the yield would be reduced.<br><br>8. This is OK but it might be worth making this more explicit in a line, though not essential.<br><br>9. Happy with response.<br><br>Overall I am happy with the response to the comments. I think that this is a well written paper and contributes to our understanding in this field. I support publication. |

**VERSION 2 – AUTHOR RESPONSE**

# Responses to BMJ Open Peer Review

## Peer Reviewer 1

---
Reviewers' Comments to Author:
---
Reviewer: 1
Richard Edlin
School of Population Health, University of Auckland
Please state any competing interests or state 'None declared': None
---
Please leave your comments for the authors below

The authors have signalled in their response that they have taken many of the comments made in my earlier review on board, and I acknowledge this. There are other comments made with which the authors disagree, and where disagreement remains. The tone of some of the authors' comments appears to be somewhat unhelpful. In several cases, the authors provide references to support positions --- but I'd contend that those references support what I have suggested they do in my earlier review, and which they have not always done.

The authors continue to report that zero QALYs are generated when women are not screened and now state in their report that this is against a baseline of non-screening. If this is the case, then the authors should to revise the tables so that they do not report "Costs" and "QALYs" but incremental costs and incremental QALYs. As originally labelled, the tables were incorrect --- now, it's simply inelegant but probably clear enough.

We infer from the last sentence that we have addressed this point to the satisfaction of the reviewer. We maintain that it is reasonable to treat the no-screening option as the origin of the cost-effectiveness plane and that it is acceptable to report costs and QALYs relative to non-intervention. In the tables we prefer to note these as incremental to a baseline of no treatment via a footnote so as to avoid confusion with the incremental cost column (measured relative to the next best alternative treatment) as that is the incremental cost used to derive the ICER.

The authors also argue against the request to reword their methodology around the use of the Cholesky Decomposition. In doing so they cite Claxton et al. I agree that Claxton et al's approach and wording is correct and wanted the authors' correct their wording accordingly. All that was needed is for the authors to report their methodology accurately, e.g. by making it clear the scale on which the multinormal sampling is taking place. The revised text on Page 14 now makes it clearer that it is the (scaled) parameters from the regression analysis that are drawn -- this wasn't clear in the original document and motivated my comment.

We are pleased that the reviewer considers that the revised text has provided the clarity he sought and we are grateful that his original comment prompted improved reporting of that aspect of our methods.

The authors state that the probabilistic and deterministic results don't differ in a concerning way. This is a subjective judgement. What remains, however, is that the probabilistic results should be those displayed precisely because of the non-linearity. The deterministic results are biased. The authors again provide references for their positions but may misunderstand the implications of those same references.

# Responses to BMJ Open Peer Review

The justification, in our response to the initial reviewer feedback, for saying that the results don't differ in a concerning way, is that using a decision rule based on a cost-effectiveness threshold of £20,000 per QALY or £30,000 per QALY the same decision would be reached using either the probabilistic or deterministic result.

However, we agree with both our reviewers that greater emphasis should be given to the probabilistic result and we have revised the abstract text so that only the probabilistic results are reported.

The authors define (in their response) an incremental net benefit as being "net benefit". Net monetary benefit should really be reserved for the absolute term lambda*e - c, rather than lambda*delta_e - delta_c. Unfortunately, this is still not done well (and not just by the authors) --- and can lead to misunderstandings and error. As an example of this, the authors clarify that "the reason that no screen/no treamtent has a NMB > 0 is because in the probabilistic analysis the net monetary benefit is calculated relative to the lowest cost and lowest QALY treatment in any particullar simulation, which isn't always no screen/no treat". Here, due to confusion over the net benefit, the authors have not used a consistently-defined baseline (in terms of a decision option) in their results. As a result, the different runs of the probabilistic analysis aren't comparable, and it's unclear whether or not they even can be combined. (Even if they can, the lack of a baseline means that whilst overall results may be okay, the specific values provided lack a clear interpretation.) The probabilistic model should be fixed (this is a very minor thing) and re-run to provide results against a consistent baseline of e.g. no screening.

We agree with the reviewer that our results would be better presented against a consistent baseline of no screening and have amended the mode, re-run the analyses and revised the presentation of our probabilistic results accordingly.

Whilst I raised the use of EVPI as (only) a possibility, I note the authors' response. They justify not doing something by its non-appearance in CHEERS and the Drummond Checklist, amongst others. The CHEERS checklist covers reporting, not analysis, and the Drummond Checklist is very seriously out of date in terms of methodological guidance. As I recall Prof Drummond's position, it was that the non-appearance of an appropriate methodology in such checklists is not a sufficient reason to not do it. Checklists should never be used to justify poor or outdated methodology in precisely the way that the authors do. The remainder of the argument around not providing EVPI results does not appear to be very relevant --- but again, this was a suggestion rather than something I felt had to be done.

We note that the reviewer did not think that this was something that had to be done and that was also our understanding from his first set of comments.

Nevertheless we tried to provide a rationale for not including EVPI within our paper. We perhaps did not express our view very well but our main argument did not relate to the content of checklists but rather that EVPI is most useful when the uncertainty about the cost-effectiveness result hinges on parameter uncertainty. It is more problematic to predict the probability of error when uncertainty about cost-effectiveness is also driven by limitations in current understanding of the pathogenesis of disease.

As a final point --- if the authors contend, as they now appear to on Page 25, that a no identification/no treat policy would not be acceptable to patients or health care providers, then why is this option included in their analysis? An economic analysis ideally considers all relevant outcomes -- and their position would mean that "do nothing" is not then a relevant option for inclusion.

# Responses to BMJ Open Peer Review

Whilst the point being made by the reviewer here is a reasonable one, we feel that it is important to retain a "do nothing" option, particularly in the context of the recent UK HTA which also included such an option and reported favourably on its cost-effectiveness.

We would like to thank reviewer 1 for his helpful comments on both versions of the paper reviewed.

---

Reviewer: 2
Prof Trevor Sheldon
University of York
Please state any competing interests or state 'None declared': I am a co-author of an NIHR HTA report which includes an economic evaluation of GDM detection and treatment approaches and a paper on the same topic which has been submitted for publication.

---

Please leave your comments for the authors below

The authors have taken considerable effort to respond positively to the referees' comments. I will only deal with the response to my comments (Peer reviewer 2) and the overall resulting paper.

1.The authors have now referred to the HTA report and have adequately dealt with the comparison. They might need to make a stronger case though for what this paper adds. So whilst they say on page 20 line 393-395 that this is one of the most comprehensive assessments (something I agree with), they should perhaps reference the HTA report alongside that as this is also comprehensive and comes to similar conclusions:

We have revised the text to note the HTA as another comprehensive assessment

i.Happy they have addressed the point about instrumental delivery and have made this clear in the discussion.

No response required

ii.The paper refers to the sensitivity of the deterministic models including fasting blood glucose. In the table x23 this is presented. However the authors fail to then comment of the degree to which this makes difference.

There is a discussion of this sensitivity analysis in the supplementary report including a discussion of the degree to which this makes a difference. This is signposted in the main text at the end of the results section although we have added a note to reference the summary of results (Note this is now Table x30 in the Supplementary Report).

iii.Happy this is now dealt with.

No response required

2.The authors should probably lead with the probabilistic results and include deterministic for comparison purposes. But I am content that both are presented and that the impact of both approaches are clear.

We have revised our paper to lead with the probabilistic results

# Responses to BMJ Open Peer Review

3.I am content with this explanation. Should, they not have a sentence on this in the discussion in case people look back at the NICE guidance material? I couldn't see it but might have missed it.

We have added a sentence in the discussion to explain the small differences between results reported in the paper and in the NICE guideline

4.Responded well
No response required.

5.Fine (maybe I am being too pedantic in saying they are the IADPSG thresholds and should not always be referred to as the WHO criteria)

No response required.

6.This is now clear and acceptable.

No response required.

7.The response is acceptable – do they mention this in the discussion. Couldn't see it but might have missed it. In reality uptake rates are not so high and so even if the cost-effectiveness is not that changed the yield would be reduced.

We did amend the manuscript allude to the difference in the approach to test uptake between our analysis and the HTA.
"Unlike our analysis, the HTA did not assume 100% uptake of the OGTT and that may also have led to a slightly smaller estimate of treatment benefit"
We have elaborated on this in our latest revision.

8.This is OK but it might be worth making this more explicit in a line, though not essential.

No response required.

9.Happy with response.

Overall I am happy with the response to the comments. I think that this is a well written paper and contributes to our understanding in this field. I support publication.

We are pleased that the reviewer is satisfied with our responses and supports publication. We thank him for his helpful comments