PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Implications of the introduction of new criteria for the diagnosis of gestational diabetes: a health outcome and cost of care analysis
AUTHORS	Cade, Thomas; Polyakov, Alexander; Brennecke, Shaun

VERSION 1 – REVIEW

REVIEWER	Diane Farrar Bradford Institute for Health Research, UK
REVIEW RETURNED	17-Apr-2018

GENERAL COMMENTS	The authors have conducted a relatively simple cost-effective analysis of the use of different criteria for the identification of GDM, unfortunately as longer-term outcomes are not included and these are important given the aim of the new IADPSG criteria the results have limited validity, I think longer-term outcomes should be included and if not available (as too soon following application of new criteria) scenario analysis should be conducted Metformin is often used in many countries if glucose control is inadequate with diet I think you should comment on this as insulin costs will be higher than metformin and metformin is used in pregnancy in many countries Please add the numbers of women included in the analysis to tables 2-4 I do not think you should include birth trauma in the analysis as you say it is affected by changes in your coding practice, this should be mentioned in the methods and a the reason for excluding this important outcome stated, including it and suggesting the results are unreliable does not seem appropriate It would be useful to include the outcome results in the GDM population only for comparison with the whole population, so a full picture can be seen, I suspect when less severe GDM is identified and included…….. outcomes will 'seem' to happen less frequently You say the decrease on NND in 2016 (compared to 2014) is unexplained and confined to the non-GDM population however the NNDs in the 2014 cohort could have been in women with GDM not diagnosed with the previous criteria, it is irrelevant that the NNDs are not in the GDM population, identifying more women with GDM may have prevented NNDs through treatment?..... You also say: the change (NND) may have resulted in a small reduction in very large babies but seemed to have no relevant clinical reduction in any other outcome. This may only be because

	you did not report infant adiposity or longer term infant obesity…….. the point of the IADPSG criteria is to reduce the risk of macrosomia through its association with macrosomia…… this requires a comment the NIHR conducted an economic evaluation and found a different approach (to the IADPSG) to be superior (risk factor screening) in their population, but actually that depended on which dataset was used, the are several scenarios in the NIHR guidance report…….. also a large and complex economic analysis that used data from the UK suggests the identification of GDM is not cost effective, based on current evidence…… Farrar D, Simmonds M, Griffin S, Duarte A, Lawlor D A, Sculpher M, Fairley L, Golder S, Tuffnell D, Bland M, Dunne F, Whitelaw D, Wright J, Sheldon T A (2013). The identification and treatment of women with hyperglycaemia in pregnancy: an analysis of individual participant data, systematic reviews and an economic evaluation. HTA 20(86). this work should be mentioned and cited Again as mentioned above you suggest the IADPSG criteria are aimed at uniformity of diagnostic criteria and this is true however the criteria aim to reduce infant obesity through the reduction in macrosomia (theoretically achieved through the lowering of the fasting glucose level) and you do not mention this or the important longer-term outcome that is missing from your analysis, if longer-term outcomes were included in your analysis you may obtain different results Your conclusions overstate what your results are able to suggest, because you do not include longer-term outcomes and these are fundamental to the new criteria's aims…… In your conclusion ….the UK may have saved money by not adopting the IADPSG criteria, but the UK has also 'allowed' women with hyperglycaemia to go untreated (which would be treated in other countries that have adopted the criteria), the offspring of a woman untreated in this way is at higher risk of macrosomia and that infant is at higher risk of developing obesity…… resulting in greater costs for the NHS……

REVIEWER	Alfonso Luis Calle-Pascual Endocrinology and Nutrition Department, HCSC, Madrid, Spain
REVIEW RETURNED	18-Apr-2018

GENERAL COMMENTS	This study shows some benefits of health outcome utilizing the new criteria of diagnosis, but it don't move in economic analysis. Could you explain these think? Changes in health outcome reported are: decrease in birth trauma, Cesarean –S, neonatal death and birth weight >95th percentile. These changes are not associated to economic savings? How consider the care package the cost of several events? We think the expenses are not all accounted. Estimation of economic cost should be include: - laboratory costs (2-step vs. 1-step); bottles of 50 g and 75 g of glucose, respectively; duration of test; How the cost for deliveries is estimated? Cesarean section (greater cost) vs. Vaginal or instrumental delivery without complications (both are similar in relation the cost) Nutritional treatment was effective in attaining glycemic targets in a similar proportion of women (50%) in both cases. Thus, the introduction of IADPSGC did not modify the percentage (50%) of patients needing insulin therapy to achieve glycemic goals. Could the use of the new criteria induce overtreatment?

	Universal screening is performed in all pregnant women or in high risk women? At 24-28 GW? Could you comment the discrepancies with other studies that demonstrate decrease in adverse events and economic saving? (Diabetes Care 2014;37:1–9 | DOI: 10.2337/dc14-0179) Limitation in Care package for cost estimation will be addressed

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

• Considering analysis of longer term outcomes.

This is an excellent point and, we agree, one we are constrained to evaluate at the present. The new criteria, as presented by the IADPSG paper, centre upon avoiding short term adverse outcomes such as macrosomia, hypertensive disorder of pregnancy, caesarean section, and pre-term less than 37 weeks. These are all identifiable at the time of birth. They may indirectly influence various long term outcomes (for instance caesarean section may be associated with childhood allergy, prematurity may be associated with developmental delay and so on), but these associations are contentious with regards to causation and are not universally defined.
The original HAPO trial did indeed examine neonatal adiposity which is the only outcome we are unable to evaluate as it is not routinely measured. This would indeed have the greatest potential impact on a potential long-term evaluation.
The point of the article was to examine the immediate impact on costs of care within a hospital and compare this to immediate adverse outcome avoided. As there were no obvious adverse outcomes avoided: further multivariate analyses (to examine significance and create appropriate models) and economic analyses were not appropriate.
While longer term health outcomes may potentially prove beneficial: the potential cost savings for these are not borne out under the budget of the initial hospital-of-care. We have tried to make this more clear in the final discussion and conclusion and re-emphasise that we hope to add to the debate for further research into long-term outcomes, economic evaluation of these, and lower-risk models of care for GDM.

• Many countries use metformin.

This is true, however in Australia this is somewhat uncommon (particularly in gestational diabetes). We have added a comment in the conclusion stating that greater use of metformin may also invoke an immediate cost-saving.

• Please include the number of women in tables 2-4

This is already done. Tables are reported as number of women with 95% confidence intervals (for continuous variables) and percentage (for discreet). The methods section is amended to make this more clear.

• Excluding birth trauma as an outcome due to a change in coding practices.

We have made this change as requested

• Request to include outcomes just in population of GDM

We have made many comparisons of GDM versus controls, GDM diet versus GDM insulin, GDM before the change versus GDM after the change. We hope to present these in a structured way in a future manuscript with reference to this one. There are simply too many for a single paper and the current goal is to highlight static overall hospital outcomes with regards to the substantial increase in costs of care.
We have, however, included summary tables (Table 5 and 6) comparing the outcomes of women with GDM in 2014 (before the change) with those in 2016 (after the change) and appropriately referenced this in the methods, results and discussion (and abstract).

• Comments about NND being confined to the non-diabetes population being irrelevant

These comments have been modified to reflect a small but unexplained decrease in this outcome. It is unlikely it can be attributable to tightening GDM criteria as it is such a rare outcome and not one upon which the criteria for diagnosis is based. These deaths most commonly occur in extremes of prematurity well before GDM is screened for.

• The economic evaluation by the NHR…..

We thank the reviewer for pointing this out and have added it, and a number of American attempts at economic analyses as suggested by Reviewer 2.

We agree there may be improvements in neonatal fat adiposity, possibly, but this is not indirectly reflected in our population in improvements in neonatal macrosomia >90th% (the criteria upon which HAPO reported) and not routinely measured in hospitals in Australia. Once again, there may be longer term health outcomes that are worthy of future research but the point of our manuscript is an immediate evaluation of the economic impact of the change with relation to immediate, quantifiable health outcomes. We certainly do not presume to imply the HAPO criteria does not have any long term benefits but that this needs quantification and, once again, any economic benefits may not be "passed on" to the hospital of contact that has borne the cost of initial care.

We have tried to modify the last paragraph of the discussion and the conclusion to better reflect this. I hope these comments modify the perhaps "overly strong" conclusions we have seemingly made and make this a manuscript more to promote discussion about costs-of-care, finding lower-risk models-of-care for GDM and quantitatively examining long-term outcomes to assist in justifying the criteria.

Reviewer 2

• We did not find any clinically and statistically significant changes in health outcomes and thus could not do an economic evaluation on them. The small decrease in babies >95th% is debatable because rates >90th% (which the HAPO criteria are based on) did not change. Macrosomic babies who do not require a caesarean section and are not admitted to the special care nursery do not represent an immediate economic burden to the hospital: caesarean section rates and admissions to SCN/NICU did not change.

• You are right that expenses are not all accounted. We initially intended to do a thorough evaluation of exactly "what was spent" on each woman individually. However, this was not valid for a number of reasons. Firstly, data collection on time spent within the hospital is not accurate and reliant on staff "clicking in and clicking out". This is the biggest area of expense (other than admissions) and our business reporting unit deems "average occasions of service" a more robust figure. This must be used as an overall rather than an individual figure. Secondly, there is no way to retrospectively

examine the costs of those “undiagnosed” with GDM in the past (as the test was changed rather than being simply modified) and this would be where an individualised approach would be most powerful. Thirdly, we have decided that an “intention to treat” (for want of a better phrase) approach is a reasonable way of reporting costs related to an overall public health policy change in a large population, an approach we clarified with local health economists as the most valid. Our costs, therefore, are indeed estimates but are most representative of adopting the change at a hospital level.

• We certainly take the point that “costs of screening” is an important part of an overall economic evaluation. We have taken the path of assessing “costs of care” (i.e. after diagnosis). The difference in cost between a 75g GCT (followed by secondary fasting GTTs in those who screen positive) and a 75g GTT in our hospital was modest.

• The costs for delivery are well documented in our hospital (i.e. there is a known cost for a basic vaginal delivery, complicated vaginal delivery, elective caesarean section, emergency caesarean section, catastrophic outcome etc etc.). We did not use these, however, because we assessed costs of care. We only planned on assessing outcome costs if there was a significant change in one of the immediate clinical outcomes of interest and there was not. This has unfortunately, but necessarily, made our manuscript more concise than originally planned.

• We have included a small analysis of GDM in 2014 versus GDM in 2016 (as also suggested by reviewer one) to address concerns about the later population being possibly a lower risk cohort.

• We use universal screening (not high risk screening) and this was one of the reasons for assessing overall outcomes throughout the hospital. I have added a comment to make this clearer at the end of the first paragraph of the methods section.

• We have addressed discrepancies in economic evaluations in the discussion. The major problem is that different countries all employ different strategies to diagnose GDM. We initially only included a reference to a UK based economic evaluation (reference 17) but at the reviewers’ suggestion have included an American based evaluation (16,17,18) and also another UK based economic evaluation (reference 15).

VERSION 2 – REVIEW

REVIEWER	Diane Farrar Bradford Institute for Health Research, UK
REVIEW RETURNED	10-May-2018

GENERAL COMMENTS	I do not agree with the authors response regarding the consideration of longer-term outcomes – Although the IADPSG criteria do aim to reduce macrosomia (adiposity at birth and other short term outcomes), the main aim of that is to reduce the risk of infant obesity through its association with macrosomia The authors suggest that their main aim is to examine the ‘immediate’ impact of GDM identification and treatment on costs, but this is only half an analysis, given the longer term aim of the IADPSG criteria and there have been other more comprehensive economic analyses than this one that have been conducted that have included longer-term outcome scenarios With this in mind their suggestion that “as there were no obvious adverse outcomes avoided: further multivariate analyses (to examine significance and create appropriate models) and

	economic analyses were not appropriate" is not correct, or certainly isn't when trials have been conducted, treatment of GDM does reduce the risk of macrosomia (that is clear from trials and meta-analysis of trials) and also women are identified with a higher risk of developing type 2 diabetes, if this risk is reduced by future intervention, cost savings may be made, the HTA report 2016 Farrar et al on GDM I previously asked to be reference clearly suggests this I do not believe the authors have a full grasp of the complexities of GDM, their manuscript makes this clear for example they refer to HAPO as a trial several times in their introduction and it is an observational study, they also suggest the aim of HAPO was "to unify disparate international views about the significance of GDM and the best way to diagnose it" this was not the aim of HAPO, their aim is clearly stated in their published manuscript (Metzger 2008) it was to "clarify the risks of adverse outcomes associated with various degrees of maternal glucose intolerance less severe than that in overt diabetes mellitus". It was the aim of the IADPSG (2010) using data from HAPO to suggest glucose thresholds that would best identify women for treatment of hyperglycaemia, these criteria however have not been consistently adopted globally because of a lack of beneficial evidence to support their use, despite the endorsement of the World Health Organization Given my concerns regarding this simple and partial evaluation and the authors seemingly limited understanding of the subject and also because the analysis does not include the use of metformin which is used in the UK, Europe and North America a treatment less costly than insulin, I do not recommend this paper for publication in the BMJ open

REVIEWER	Alfonso Luis Calle-Pascual Endocrinology an Nutrition Dpartment HCSC, Madrid. Spain
REVIEW RETURNED	27-Apr-2018

GENERAL COMMENTS	The authors have addressed many of my comments

REVIEWER	Adjunct Associate Professor John R. Moss School of Public Health, The University of Adelaide, Australia
REVIEW RETURNED	22-Aug-2018

GENERAL COMMENTS	This is a well-conducted and informative audit of hospital data; and the authors make it clear from their comments in the text that they are indeed aware of the main point of scientific reservation, namely the nature of the comparison and hence the extent to which causal inference is possible. After justifiable exclusions, in 2014 the hospital managed 7,010 pregnant women of whom 416 (5.93%) were diagnosed with GDM under the old criteria; whereas, in 2016, there were 7,488 pregnant women of whom 774 (10.3%) were diagnosed with GDM under the new criteria. Hospital-wide, this amounts to a 74% increase in the proportion of women diagnosed with GDM across the two years. The authors claim that this increase in diagnosis occurred "without overall improvements in primary health outcomes"; and that "[b]abies of women with GDM had lower rates of neonatal hypoglycaemia and special care nursery admissions after the change, suggesting a milder spectrum of disease."

	In the hospital-wide comparison of maternal and foetal outcomes, it would be difficult to rule out differential selection and/or confounding by unrecognised variables amongst the non-GDM women who made up a substantial majority of each year's total (94.07% and 89.7% of the total for the respective years). Hence, the attribution of cause amongst GDM+ women from hospital-wide data would be a challenge; and in the present paper relies on presumptive evidence. Thus, the third claim under "Strengths and limitations of this study" that assessing the implications of adopting the new criteria on an entire cohort would minimise the risks of selection bias should be qualified. This claim might be helped by presenting a more extensive time series to inform the reader as to the degree of fluctuation in the health outcome variables over perhaps the last five years. The authors also may wish to try a formal quantitative risk of bias analysis along the lines described by Lash et al. (2010) and applied to this particular study design. Although only a minor issue, the justification for labelling a hospital-wide group as representing a "public health policy" perspective is not clear since there is no intervention into the social determinants of health. Given the unconventional design, the choice of statistical tests is open to debate. The authors have used Fisher's exact, Chi-square and t-tests, which provide a P-value. Adopting the analogy of comparing two arms in an RCT would suggest that risk rates and risk differences would have been preferable - because they provide a measure of the strength of association and a confidence interval. Under the circumstances, perhaps neither approach would satisfy the purists. Turning to the GDM comparison between the two years, as the authors explain in other words, it was not possible to either compile a notional cohort of 2014 women diagnosed according to the 2016 criteria nor to stratify the 2016 women into those who would and would not have been diagnosed GDM+ according to the 2014 criteria. We cannot know what proportion of the GDM+ in 2014 would have tested positive for GDM in 2016 and vice versa. Thus a pragmatic comparison is presented, between GDM+ women in 2014 and GDM+ women in 2016, according to the test prevailing during each year. Then it is reasonable to ask what would be the relevant research question: that the 2016 group do proportionately no worse than the smaller 2014 group (non-inferiority) or that the 2016 group do no worse than a hypothetical comparison group not tested? Thinking about it this way indicates that the 2016 group can do worse than the 2014 group yet still perhaps achieve a useful health gain. In Table 5, the point estimate of the proportion of each of the 7 maternal outcomes in women with GDM from 2016 was better than or equal to that from 2014, though only the occurrence of third degree tear was statistically significant. In Table 6, the point estimate of the proportion of 13 out of 16 foetal outcomes from 2016 was better than in 2014, the exceptions being EGA (by 0.2 points), NND (0 versus 1; likely underpowered) and birthweight (by 56 grams) none of which was designated a primary outcome; only the occurrence of hypoglycaemia and admission to SCN were statistically significant. Primary outcomes were designated as those upon which the new criteria were based, namely caesarean section rates, hypertensive disorder of pregnancy, birthweight greater than the 90th percentile, pre-term birth less than 37 weeks;

	and the point estimate in 2016 was better than in 2014 for all 4 of these, but no difference was statistically significant. Testing for noninferiority might help here. Whether these results are due to a reduction in risk due to more appropriate antenatal care being provided for women who would not otherwise have been diagnosed with GDM or to the additional women being at lesser risk is open to debate. The authors plump for the second explanation. In the calculation of the hospital outlays for antenatal care of women with and without GDM, it is the net amount that is relevant not the gross, since the additional proportion of women diagnosed GDM+ in 2016 would have received standard care otherwise. The paper could be improved by the adoption of a full economic framework rather than the limited focus on the costs of antenatal care alone. When resources are scarce, as seems inevitable in a hospital environment, the economic perspective focusses on the incremental cost compared to the incremental outcome – across all relevant categories of resource use and of states of health; it is not obligatory that the intervention be cost-saving. Thus, not only the incremental cost of antenatal care is relevant, but also any increment (plus or minus) in the costs of delivery and post-delivery care for both mother and infant, which are not provided. Furthermore, no mention is made of the value of any life-years gained or suffering averted. Finally, as the authors acknowledge, they have measured only the immediate maternal and infant outcomes and not any long term impact. In summary, this study is a fascinating illustration, in the context of an important resource use question, of the problems in interpreting data that is neither randomised nor based on a single cohort. As such, it is deserving of publication as one contribution to a complicated clinical debate. Regarding recommended revisions, the uncertainties should each be pointed out as explicitly as possible; and an appropriate economic framework should be used if the word “economic” is to appear in the title. Regarding the Checklist Review 2. In the Abstract, the Design, Results and Conclusion need modification as follows:- Design: This is a quasi-experimental study rather than a retrospective cohort study Results: These are subject to substantial uncertainty regarding: a. The claim of no overall improvement in primary health outcomes (because of the potential for unrecognised selection bias or confounding) b. The suggestion that babies of women with GDM had a milder spectrum of the disorder (where it is at least possible that this improvement in outcome reflects better antenatal control due to earlier recognition) • Moreover, the net cost rather than the “gross cost increase” should be reported, because the former is the increase in use of antenatal resources. Conclusion: Is thus subject to similar issues as mentioned for the Results.

	3. See 2 (above). In the main text, the authors mention the limitations in the study design. 11. See 2 (above). Expression Strengths and limitations of this study, point 1: “one of the only” is ambiguous Strengths and limitations of this study, point 2: “erstwhile” should be “otherwise” REFERENCE Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer 2010.

VERSION 2 – AUTHOR RESPONSE

Responses to Reviewers

Many thanks to reviewers 2 and 3 for providing courteous and considered responses to help us produce a better manuscript that may be a small part in the ongoing discussion into a complex medical problem.

We hope we have addressed their concerns satisfactorily and have elaborated upon them below.

Reviewer 1

It appears there is an irreconcilable difference of opinion between this reviewer and the authors on both the point of the study and the diagnosis and treatment of GDM.

All three contributors to this manuscript are authors and reviewers (both locally and internationally) and it is exceedingly rare to receive comments that are so undiplomatic and unnecessarily antagonistic.

As there have not been further suggestions toward improving our work toward publication (only a rejection), we have not made any specific changes specific to these latest comments. However, we address them as follows.

Generally

The aim of our article is to contribute to the discussion around adoption of the HAPO/IADPSG criteria over other systems of diagnosis by highlighting the immediate, short-term workforce and cost

implications and assessing immediate, easily quantifiable health-outcomes directly rather than with modelling.

This study is part of a much larger body of work and has not been done in isolation. It is not designed to report anything other than these short-term findings and it is made abundantly clear many times that longer-term analyses are desirable before definitive adjudication on these criteria.

It also is not an economic analysis based on comprehensive modelling data as has been published by others previously. This is acknowledged and cited. It is a simple, “real-life”, quantification of immediate costs which may provide care-givers and policy makers some food for thought.

Multi-variate analyses not being appropriate

We stand by this claim: with any multi-variate analysis two things must be initially considered. Firstly, will a multi-variate analysis (or model) add anything to the findings of a univariate analysis and secondly, are there outcomes of statistical and clinical significance in the univariate analysis. It was the considered opinion of two of the authors (both with either current or pending post-graduate qualifications in Clinical Epidemiology and Medical Statistics) that neither of these criteria were fulfilled in this case.

Further economic analyses quantifying health-outcome savings were not appropriate in this study as there were no apparent, immediate health-outcomes to quantify. Once again, we have stated numerous times that longer-term studies may find previously unquantified savings and would be part of a more comprehensive evaluation.

The authors do not have a clear grasp/have limited knowledge of GDM

One of the authors is the Head of Diabetes at Australia’s largest stand-alone maternity centre, received a Research Higher Degree (at doctoral level) and Early Career Researcher Fellowship on this topic (at Australia’s highest ranked university), and has received praise from the relevant national funding bodies for a proposal for a randomised controlled trial further examining models of care in GDM based on this research. Another has been a Professor of Obstetrics and Perinatal Medicine (including the Chair at Australia’s highest ranked university) for over two decades.

While we acknowledge this reviewer’s academic and publication record and would not be so rude as to aim a similar comment in return, it would perhaps be most diplomatic to say we simply have different views about GDM.

Longer-term outcomes

Relatively few countries have universally adopted the HAPO/IADPSG criteria and longer-term outcomes must be assessed as part of universal (not selective) screening and treatment. Australia adopted this criteria in late 2015 with 2016 being the first calendar year. Thus, infants were being born to mothers with GDM under this system from mid 2016 onward. They are only now reaching two years of age which is an absolute minimum for assessment for meaningful paediatric outcomes. It is also

exceedingly expensive to perform such an analysis properly, there are high lost-to-follow up rates and generally multi-centre collaboration is needed.

One of the goals of publishing these initial findings is to provide the impetus for a discussion of funding for research examining longer-term maternal and neonatal/paediatric outcomes. It cannot be done without providing the initial rationale. Even the HAPO follow-up study will not necessarily provide the full answers required as this was an untreated population: the second piece of the puzzle is examining a treated population (for which there aren't enough children yet) and examining either concomitant controls or using the HAPO follow-up as a "historical" control.

Metformin use

This is exceedingly rarely used in GDM in our centre. We cannot analyse a system of treatment that does not occur in our population. It also does not affect the proportion who were managed under dietary measures alone and would be unlikely to significantly change the outcomes in those who were not (and currently are routinely given insulin in our centre). The expenses quoted were mainly due to occasions of service and medical imaging: consumables were a relatively small contributor.

Treatment of GDM does reduce macrosomia

This is unarguable but there is not conclusive evidence that treatment of GDM under HAPO/IADPSG criteria reduces macrosomia rates compared to treatment of GDM under other systems of diagnosis (particularly the two-step system previously used in Australia). That is the point of examining this outcome and presenting it in the way we have.

Referring to HAPO as a trial

This is pure semantics: a trial and a study are relatively interchangeable words. Many epidemiology degrees include subjects as "Clinical Trial Design" which encompass all types of medical studies. Nonetheless, we have changed the term "trial" to "study" with regards to HAPO.

Conclusion

There is clearly a difficulty (either professionally or personally) with us or our work and this particular reviewer. While we support robust debate and peer review, we feel these comments have descended into the personal and are unfair.

We note the favourable comments from the other two reviewers and hope that we can move on to satisfactorily addressing them instead, and leave it in editorial hands to adjudicate further.

Reviewer 2

We once again thank reviewer 2 for their erudite feedback and are glad that we satisfactorily addressed previous concerns.

Reviewer 3

Thank you for a very considered precis, review and critique of our work. We acknowledge that this manuscript is not a comprehensive economic analysis. As we were unfortunately unable to initially demonstrate an overt, short-term health outcome improvement (contrary to our initial hypothesis), we were unable to assign a cost-saving and compare it to the increase in workload expense. We do acknowledge that it may be inappropriate to thus refer to an economic analysis and have thus changed the title to “cost of care analysis”.

We have also acknowledged the potential for unexplained selection bias and confounding inherent in retrospective design more clearly: in the strengths and limitations and in the discussion.

We have addressed the difficulty in labelling changes within a single centre (even with a large cohort) as representative of public health policy also in the strengths.

We have changed the results and discussion to focus on the net (not the gross) increase in costs and have left only a single reference to gross cost as a reference to how the calculation was made.

In both the results and the discussion we have attempted to make it clearer about the inherent difficulties of drawing definitive conclusions regarding “no changes in health outcomes” by moderating language (for example “may be” “approximately” and “appeared to be no change”): more so in the discussion than the results.

The following were inserted into the discussion:

It is important to note that these findings, in a retrospective analysis, may be subject to unrecognised selection bias or confounding and form part of a larger debate into the care for women with GDM.

While these costs are seemingly not redeemed in the short term by marked improved outcomes, better health care is not always defined by more economic models and there may be unquantified health outcomes demonstrable in longer term analysis of women with GDM and their babies treated under this system.

We thank you for drawing our attention to the Lash reference. Indeed, we have commenced work on a proposal for a randomised controlled trial into different models of care for managing GDM and have now included this risk of bias analysis into our initial discussions and planning. It did seem a little unfeasible to use in this current manuscript and thus have moderated our language about selection bias in the strengths and weaknesses.

We certainly agree that this manuscript is a small part of a much larger discussion into this prominent public health problem, specifically which system of diagnosis to use. Past economic analyses have used modelling approaches and future ones would need to examine the long-term outcomes in both mothers and babies in a country (such as Australia) using universal (not selective) screening and treatment. We hope that the points raised within our manuscript may provide incentive for such an analysis. Certainly, it would be possible to then evaluate outcomes such as life-years gained or suffering averted: unfortunately, it is not possible in the short-term without making a large number of potentially problematic assumptions.

With regards to this latter: HAPO/IADPSG criteria are unproven against other criteria (the observational study was based on an untreated cohort) and thus it will only be possible to perform such a study after routine adoption of the criteria. This has been done only recently with the first such offspring being born (in Australia) in mid 2016. As they are only now reaching two years of age (generally the absolute minimum for meaningful paediatric follow-up) data are not yet readily available and would require large numbers and multi-centre collaboration. We do indeed hope to propose and be part of such a study in Australia.

In conclusion, we believe the main recommendations centre around two areas: the conclusions and assumptions made (particularly with regards to the intensity of language) and the methodology of the study itself. With regards to the former, we hope we have adequately modified the language primarily in the abstract and discussion but also in the results and strengths and limitations section. We hope this is sufficient but if it is still unclear or overly presumptuous, we would be happy to revise further if required. With regards to the latter, we have changed the title to reflect the necessary lack of a full economic analysis rather a statement of increase in costs of care. The other methodological suggestions have been noted but may be best employed in a future prospective follow-up study, either by us (hopefully) or others.

VERSION 3 – REVIEW

REVIEWER	Adjunct Associate Professor John R. Moss University of Adelaide, Australia
REVIEW RETURNED	12-Oct-2018

GENERAL COMMENTS	The resubmission is a substantial improvement, but there are still several outstanding matters to be resolved. The most important is that there are still no data presented for the comparative costs of inpatient care during pregnancy and the immediate postnatal period. Given that the change in screening policy only occurred recently, it is not unreasonable for the authors to have limited their data gathering to the short-term; and they explain in their Discussion that longer term health outcomes and “economic benefits” (the latter being a confusing phrase) may change the conclusions that can be drawn from the short –term. However, the costing is incomplete even in the short-term because inpatient costs for the mother during pregnancy, delivery and the puerperium as well as for the neonate are not considered. The study thus contains an implicit assumption that the mean inpatient costs are the same (after adjustment for inflation) for both screen-positive and screen-negative women and over both 2104 and

	2106. (There might also be differences in ambulatory care costs, but perhaps they can be disregarded for the moment as likely to be of a lesser magnitude.) For an adequate short-run cost analysis, data on the inpatients costs must be presented. Appropriately, the authors have added several statements indicating that the results are subject to uncertainty. However, they have not laid sufficient emphasis on this and on the reasons why this uncertainty exists. Every clinical epidemiological study is subject to uncertainty, but because of its unconventional design this study is subject, all other things being equal, to more uncertainty than a conventional RCT or cohort study. This greater uncertainty arises firstly because, from the data provided, it is impossible to determine whether the screen-positive subgroup in 2014 would all have screened positive in 2016 nor whether all screen-negative patients in 2016 would also have screened negative in 2014. Thus the waters are muddied as to what is being compared with what amongst the screen-positive women. There is also the possibility of unrecognised differential selection between the two whole-of-hospital groups. Stronger statements about these uncertainties are needed on page 9 around lines 3-5 and lines 21-23. Given the unconventional design, the choice of statistical tests is open to debate. A statistical consultant might support my earlier suggestion that risk rates and risk difference (with confidence intervals) would be preferable to P-values or might suggest that the unconventional design does not warrant this. The phrase “public health” is used inappropriately and should be deleted wherever it appears. This is a hospital-based clinical study. It is about clinical policy. The 2014 and 2016 cohorts might be described as “whole-of-hospital” cohorts or “pre-screening” cohorts. “Women with GDM” thus constitute “test-positive” subgroups, and this might be used as an alternative terminology. My mention of the social determinants of health in my original review was in an attempt to explain the modern definition of public health and why this manuscript should not be using the term “public health”. Hence, on page 3 lines 23-24, the words from “however” to the end of the sentence should be deleted as being irrelevant to this manuscript. The use of the term “gross cost” in effect assumes that the 329 extra women with GDM in 2016 would not have received any antenatal care at all. The correct figure to quote is the net cost; and all mention of the gross cost should be deleted as misleading, being likely to suggest a greater increment in cost than can be validly claimed. Other Recommendations p.2 line 46 delete “significantly” because no statistical test has been performed in support of this claim p.3 lines19 “unrecognised” rather than “unexplained” selection bias p.4 line 44 not “an entire cohort” but “two pre-screening cohorts” p. 7 line5 HAPO was not a trial

	p.8 line 7 delete “greater” p.9 line 9 replace “a saving to any potential” with “to identify any overall” because a cost saving is not to be confused with an outcome improvement – they are different measures, one being about resource use and the other about health p.9 line 21 close bracket after “GDM” rather than after “analysed” p.9 lines 21-23 The comparability of the 2014 and 2016 whole-of-hospital pre-screened cohorts is also subject to uncertainty, because of the potential for differential selection. Suggest add after “this was likely to be minimised” the words “although there is potential for unrecognised differential selection”. p.9 line 29 not an “absolute” change p.9 lines 53-54 delete “better health care … economic models” because the meaning is quite unclear and the words add nothing to the main argument p.10 lines 11-13 “There may indeed be economic benefits that can be compared with the initial increase in costs of care …” In economics, costs and cost savings are on one side of the balance and health outcomes (the positive ones sometimes called benefits) are on the other. Do the authors mean to say that long-term cost savings may be possible? p.10 line 28 after “unchanged” add “in the 2016 test-positive subgroup”

VERSION 3 – AUTHOR RESPONSE

Response to Reviewer

Many thanks to reviewer 3 who has clearly examined our manuscript in great detail and with great thought. We hope we have addressed both his general and specific comments below.

The most important is that there are still no data presented for the comparative costs of inpatient care during pregnancy and the immediate postnatal period

We have added statements to the cost-of-care description in both the methods and discussion elaborating upon this. After meeting with the Institutional Business Performance Reporting Unit and a Health Economist, it became clear that the vast majority of inpatient care was taken up by “bed-days”. These were in the ward for the mother and in SCN or NICU for the infant. On further analysis of maternity patients at our hospital, other than a few rare events, the vast majority of changes in expected bed-days for the mother came with mode of delivery (a mother who had a caesarean section stays for 3 days, whereas those who delivered vaginally stay 1). We had planned to allocate costs for inpatient care if there was a difference in either caesar rates or admissions to SCN or NICU, and explore this further (indeed we hypothesised that it was likely there would have been a modest change). As there were no differences, we omitted this step.

It was part of an unfortunate circumstance (one which has also repeated itself in smaller sub-groups we have analysed and plan to publish subsequently) where no clinically and statistically significant

changes were found. A large part of our proposed modelling was thus rendered unnecessary (and changed our methodology from a short-term economic analysis to a cost-of-care report).

Indeed, without a change in bed-days the costs of inpatient care for a women with GDM versus a woman without are the same. The woman with GDM must test her BGLs bi-daily but this is with her own glucometer that was already included in the initial cost-of-care package.

We hope we have added comments to clarify this including stating in the discussion that inpatient costs of care at an individual level would be much more desirable in a prospective study with smaller cohorts.

Appropriately, the authors have added several statements indicating that the results are subject to uncertainty. However, they have not laid sufficient emphasis on this and on the reasons why this uncertainty exists.

Thank you for drawing out attention to inadequate language modification. We have further changed the strengths and limitations, discussion and conclusion as suggested and with more temperance in how we suggest our findings are interpreted.

Given the unconventional design, the choice of statistical tests is open to debate.
We have indeed consulted (two) biostatisticians during our project planning and analysis. There are two reasons we have included a simpler table with p-values. Firstly, we have chosen a large cohort with most outcomes as proportions (binary in the vast majority) rather than means: confidence intervals (which were routinely examined) were always narrow and odds ratios always crossed one. Thus, to make it a “cleaner” table we used the p-values instead.

Secondly, we had planned a more comprehensive statistical evaluation of those findings which reached some level of significance (clinical and significant) in the initial univariate analysis. Findings of significance were to be presented in a more extensive way and examined in greater detail using an appropriate multi-variate model. As none did, we were somewhat stymied with where further to go other than to present these findings which were contrary to our initial hypothesis.

The phrase “public health” is used inappropriately and should be deleted wherever it appears

We accept this criticism and have changed as suggested. The only reference to public health is now in phrases such as “Our findings may assist in decision making regarding public health policy” but not referring to our own findings specifically as public health outcomes.

All mention of the gross cost should be deleted as misleading

We apologise for missing references to this in the abstract and discussion after previously extensively deleting. We have further deleted any reference to gross cost, with the exception of the results section (where it used to demonstrate how the calculation of nett cost-by assigning the excess women a low-risk antenatal program-was done).

Other Recommendations

All specific recommendations were changed as directed.

VERSION 4 – REVIEW

REVIEWER	John Robert Moss University of Adelaide, Australia
REVIEW RETURNED	27-Oct-2018

GENERAL COMMENTS	This manuscript is now, in my opinion, acceptable for publication, except for the following two points below: Page 4, lines 26–27. As suggested before, the words “however it is noted that no social determinants of health were defined or analysed as part of the study” are irrelevant to this manuscript and should be deleted. Page 6, lines 26-33. This is no more than a rough and ready comparison of resource use that does not draw upon data in the hospital accounting system; and moreover the cost comparison suffers from the same problems as for the health outcomes. The authors appear to be relying on there being no substantial differences in costs between the 2014 and 2016 pre-screening cohorts. But there are differences between the two screen-positive groups (Tables 5 & 6) which might have increased costs at that level of analysis. Some words to cover this would be helpful.

VERSION 4 – AUTHOR RESPONSE

Response to Reviewer

Many thanks to reviewer 3 once again with two further small comments that we hope we have satisfactorily addressed.

As suggested before, the words “however it is noted that no social determinants of health were defined or analysed as part of the study” are irrelevant to this manuscript and should be deleted.

This line has been removed as suggested.

This is no more than a rough and ready comparison of resource use that does not draw upon data in the hospital accounting system; and moreover the cost comparison suffers from the same problems as for the health outcomes. The authors appear to be relying on there being no substantial differences in costs between the 2014 and 2016 pre-screening cohorts. But there are differences between the two screen-positive groups (Tables 5 & 6) which might have increased costs at that level of analysis. Some words to cover this would be helpful.

Noted: although the differences in Tables 5 and 6 between screen positive women were fairly minor and (possibly) indicative of a lower risk group, we take the point that such outcome improvements would represent lower inpatient costs if data were assessed prospectively as part of a randomised trial (or possibly an observational study). We have added a comment to the discussion to this effect as it seemed more appropriate here than in the methods.