PEER REVIEW HISTORY

BMJ Medicine publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	WHO 2013 criteria for diagnosis of gestational diabetes mellitus in low-risk women in early pregnancy: an international prospective cohort study.
AUTHORS	Huhn, Evelyn; Göbl, Christian; Fischer, Thorsten; Bernasconi, Todesco; Kreft, Martina; Kunze, Mirjam; Vogt, Deborah; Dölzlmüller, Eva; Jaksch-Bogensperger, H; Heldstab, Sandra; Eppel, Wolfgang; Husslein, Peter; Ochsenbein Kölble, Nicole; Richter, Anne; Bäz, Elke; Winzeler, Bettina; Hoesli, Irene

VERSION 1 - REVIEW

REVIEWER 1	Zhang, Huijing; Peking University First Hospital, Department of Obstetrics and Gynaecology. Competing Interest: None
REVIEW RETURNED	18-Aug-2022

GENERAL COMMENTS	This is an interesting paper on the early diagnosis of GDM. I agree to accept and publish the paper.

REVIEWER 2	Sovio, Ulla; University of Cambridge, Obstetrics & Gynaecology. Competing Interest: None
REVIEW RETURNED	19-Aug-2022

GENERAL COMMENTS	This prospective multi-centre study evaluated the WHO 2013 criteria in the screening of gestational diabetes mellitus (GDM) between 12 and 15 weeks of gestation in a low-risk setting. The study addressed a clinically important topic. The manuscript was easy to follow and the authors have made suggestions for further research. I have some specific comments which may help the authors to improve their manuscript. The authors have not declared whether they investigated clustering in their multi-centre study. It looks like any within-centre clustering was not accounted for in the analysis (see for example https://doi.org/10.1093/ije/dyy117). The authors could clarify this and add any relevant information to the manuscript. Tables 1-3 should include footnotes to explain the statistics reported. These are stated in the Statistical analysis section but the tables should be possible to read independently. Also, the Abstract includes data in square brackets but it only becomes clear from the main text that these are 95% confidence intervals. Screening statistics: The diagnostic accuracy statistic is difficult to interpret when the data are imbalanced. Instead, the authors could report the diagnostic odds ratio in Table 3.

	Figure 2 plots are named violin plots in the figure title, but the figure presents box plots overlaid on dot plots of the glucose measurements (violin plots would present the probability density of the glucose measurements, see for example https://doi.org/10.1371/journal.pone.0238835). A corrected description could also be added to the Statistical analysis section. Discussion: It would be worth double-checking that the information extracted from each referenced paper is correct. I noticed that the statement "Early hyperglycaemia resulted in accelerated fetal growth < 20th weeks of gestation" (reference 22) is incorrect. The study in question reported an association between GDM diagnosed at or after 28 weeks of gestation and accelerated fetal growth from 20 to 28 weeks of gestation, but it did not include information on early hyperglycaemia or estimates of growth acceleration before 20 weeks of gestation. The manuscript contains a number of spelling mistakes and would benefit from further proofreading. For example: Under "What this study adds": "stuff" should be spelled "staff". In Introduction: "is continuously increases" should be corrected. In Methods under Participants: "chronic liver, kidneys or heart diseases" and "in previous a pregnancy" should be corrected. Discussion: the name "Jokalinen" should be spelled "Jokelainen".

REVIEWER 3	O'Shea, Paula; Mater Misericordiae University Hospital, Clinical Biochemistry & Diagnostic Endocrinology. Competing Interest: None
REVIEW RETURNED	22-Aug-2022

GENERAL COMMENTS	Title: WHO 2013 criteria for diagnosis of gestational diabetes mellitus in low-risk women in early pregnancy: an international prospective multi-centre cohort study. General comment: This prospective multi-centre study is timely and provides novel valuable data regarding the role of the one step OGTT (Standard) adjudicated by the WHO 2013 criteria in GDM prediction and diagnosis in early pregnancy in a low-risk cohort. Gestational Diabetes Mellitus (GDM) is a common pregnancy complication associated with increased morbidity and mortality for both mother and baby. The WHO 2013 criteria are usually applied between weeks '24 -28 gestation and the diagnostic/predictive utility of these decision thresholds have not been assessed prior to 24 weeks' gestation in a low-risk cohort. Currently, there is insufficient evidence to recommend early testing. However, it is suggested that early identification of women with or at risk of GDM would allow more timely intervention thus decreasing foetal exposure to hyperglycaemia. The presentation of the study and its results are of a high standard and quality. The finding of poor predictive value of the OGTT in early pregnancy using the WHO 2013 criteria is very important both for patients and those who have the privilege of caring for pregnant women. In total, 12% (74 of 636) of women had a diagnosis of GDM (late/24-28 weeks'gestation) with 35% (26 of 74) of these women, picked up

by testing in early pregnancy (12-15 weeks’ gestation) using the WHO 2013 criteria.
The potential utility of modified early-OGTT cut-offs presented here need to be validated in larger prospective & more diverse cohorts as the authors suggest.

Methodology comments:

The OGTT has poor reproducibility and standardisation of each step in the total testing process is critical to ameliorate this issue. As the glucose results are critical to the findings of the study this section needs more detail.

1. Specifically, how were the patients prepared for the OGTT / what advice was given re smoking/ diet prior to testing/ behaviour during testing?

2. Fasting prior to test:
10-14h of fasting is difficult for most but particularly challenging in pregnancy. The requirement is a minimum of 8h? Was this data on the period of fasting collected at each centre?

3. When were the OGTT performed at each centre – I assume all were performed in the morning post overnight fast?

4. How were samples collected at each of the centres (preanalytical factors):
• Was each sample for glucose measurement sent immediately on collection to the laboratory for analysis or were they sent together on completion of the OGTT?

• What steps were made to minimise the effects of glycolysis in vivo?

• What specimen tubes was whole blood collected into for glucose (e.g., Na/F or Na/F/citrate tubes/ Serum/Li heparin sample separated and processed within 30 mins of blood draw) measurement.

• Was sample collection standardised across the different centres?

5. What was the analytical performance (precision, bias, traceability/calibration) of the respective glucose methods at each laboratory where glucose was analysed.

6. Were all laboratories accredited for glucose to ISO15189;2012 standards? Suggest including a supplementary table with this data.

Statistics comments:

Did the authors recruit the appropriate number of participants to demonstrate that an early OGTT (≤15 weeks ‘gestation) might be used to predict/diagnose GDM. Moreso, how did the authors determine sample size to ensure the study is adequately powered? I see this information is provided in the study protocol, but I think a statement of sample size/ power calculation should be added to the current manuscript & referencing again the protocol paper.

Table 1 PG 12
1. Please include p-values and data on height, weight, gravidity, and blood pressure.
2. Age of women ranged 18 – 45 years. Is the data normally distributed? If not, please use median (IQR)
Please include p values in the commentary below Table 1 PG 12VLines 3 to 5.
Data on gravidity not given in Table 1 -please include.

Table 2 PG 13
1. Please include p-values

Table 3 PG 15
mg/dL – no values in table
Can the authors consider including an additional Figure detailing the ROC curves for this data? It would be helpful for comparison to other studies e.g., Corrado et al & would be in keeping with the stated primary objective in the published protocol.
Likelihood ratios (LRs) that allow one to determine how much the utilization of a particular test will alter the probability would be useful to include too.

Comment 3:
Re: Discussion PG 17 Line 7
"Setting the FPG cut-off at ≥5.1 mmol/L raise concerns for overdiagnosis, since FPG decreases until 20th weeks' of gestation."
You assessed the lowering of decision thresholds for post load glucose values. Was there a lower fasting plasma glucose concentration e.g., 4.7 mmol/L, in early pregnancy below which GDM did not develop?

Comment 4:
HbA1c was measured in all participants between weeks'12 and 15 gestations, did you correlate with the early OGTT and late OGTT? If so, please include this data.

Comment 5:

Plasma glucose levels ≤2.5 mmol/L – how many PG results were ≤2.5 mmol/L? and what was the reason for these results? Improper sample handling/clotted sample/patient vomited post consummation of glucose load?
Were these results included in the final analysis?

Comment 6:
Even in this low-risk cohort, would clinically relevant risk factors for GDM like previous GDM history, higher BMI, ethnicity, inactivity etc at the first antenatal visit be as good as the early OGTT in predicting GDM?

Comment 7: Pg 18-19
Could you elaborate /speculate on the mechanism as to why glucose targets in women with GDM identified in early pregnancy were more difficult to achieve and so, negatively impacted the reduction of complications in early diagnosed GDM?

Comment 8:
Strengths and weaknesses

	Universal screening will include all pregnant women, those at low, medium and high risk of GDM. This study cohort reveals the utility of early OGTT testing in this study's specific population - a low-risk population. Please amend inference on PG 17 lines 23-24. Comment 9: Pg 18 lines 5-8 Re: we did not further investigate maternal characteristics and co-morbidities between early and late diagnosed GDM women and therefore cannot describe an early-onset GDM phenotype. I hope that this will be the subject of a further manuscript – if so perhaps state. Minor What this study adds Line 20: Typo – Stuff – should read staff

REVIEWER 4	Riley, Richard; University of Birmingham, Institute of Applied Health Research. Competing Interest: None
REVIEW RETURNED	12-Sep-2022

GENERAL COMMENTS	Thank you for the opportunity to review this paper for potential publication in BMJ Medicine, from a statistical perspective. The research question is clearly important and it is good to see a well-conducted multi-center cohort study used to answer the research question. I do have a few comments that arise from reading the article: 1) "The analysis set included only women with complete early and late OGTTs" – but this leads to confusion about what to do with those individuals with inconclusive results – better to consider them in sensitivity analysis, see Shinkens https://pubmed.ncbi.nlm.nih.gov/23682043/ 2) When examining the test accuracy at 12-15 weeks, the reference standard is the test result at 24-28 weeks – but surely this is also not perfect? Should we not be comparing the test accuracy to the true gold standard? Otherwise, how do we know the actual accuracy of the test at 12-15 weeks? 3) Looking at the protocol, sample size was based on precise estimation of the AUC. However, this is not the important measure – rather, we need to have precise estimates of sens, spec, PPV and NPV at the key thresholds of interest. Indeed, the authors never even report AUC in the actual paper, so I'm not sure why this changed. Sadly, this has led to a problem with wide confidence intervals at most thresholds, especially for PPV, which is a key measure for clinical decision making. E.g. at fasting glucose, a cut-off of 5.7 has a CI for PPV from 15 to 95%. So wide, that I do not see how we can make conclusions about the test performance (even ignoring the true reference standard issue) 4) Were results consistent across centres, or is there between-center heterogeneity in test accuracy? Prevalence tends to vary across settings, so this could impact upon PPV and NPV. 5) How was PPV derived? Using the observed prevalence of the outcome in the data at hand? (sometimes we take this from external data, especially if it is more accurate elsewhere) 6) "Increasing the fasting cut-off did not improve overall diagnostic accuracy" – I'm not sure I agree, as it does increase PPV, whilst barely reducing NPV. I think the authors should focus more on PPV and NPV rather than 'overall accuracy' based on sens and spec.

	7) Minor - What this study adds: stuff to be changed to staff 8) The protocol mentions many aspects that are not reported here (including ROC curves, multivariable models, random forests, lasso, etc) In summary, the wide confidence intervals for PPV, and the uncertainty about using 24-28 weeks as the truth, led to some confusion about what this study can conclusively decide in terms of the best thresholds to use and the value of tests at 12-15 weeks. There is some useful information in here, and it may be especially important if added to meta-analyses on this topic later, but regardless, I hope these comments help the authors and BMJ Medicine moving forwards.

VERSION 1 – AUTHOR RESPONSE

Comments of Reviewer 1

<u>1. Reviewer 1 wrote:</u> This is an interesting paper on the early diagnosis of GDM. I agree to accept and publish the paper.
<u>Answer:</u> *Thank you very much for your kind review!*

Comments of Reviewer 2:

<u>2. Reviewer 2 wrote:</u> The authors have not declared whether they investigated clustering in their multi-centre study. It looks like any within-centre clustering was not accounted for in the analysis (see for example https://doi.org/10.1093/ije/dyy117). The authors could clarify this and add any relevant information to the manuscript.
<u>Answer:</u> *Thank you very much for your comment. Within-centre clustering has not been accounted for. In fact, the statistical estimates that we present are all simple proportions (sensitivity, specificity, etc). To derive these, no statistical modelling was involved, hence we did not have the option to account for center-specific effects via random-effects as suggested. However, we visually examined glucose levels centre-specific. Please see our Response 30 to Chief Statistics Editor for further detailed information.*

<u>3. Reviewer 2 wrote:</u> Tables 1-3 should include footnotes to explain the statistics reported. These are stated in the Statistical analysis section but the tables should be possible to read independently. Also, the Abstract includes data in square brackets but it only becomes clear from the main text that these are 95% confidence intervals.

<u>Answer:</u> *Thank you very much for your comment. The explaination of statistical analyses were added to Table 1-3. The confidence intervals are now explained in the Abstract.*

<u>4. Reviewer 2 wrote:</u> Screening statistics: The diagnostic accuracy statistic is difficult to interpret when the data are imbalanced. Instead, the authors could report the diagnostic odds ratio in Table 3.

Answer: This is a good point. Indeed, overall diagnostic accuracy depends on disease prevalence, and can give a wrong impression on test performance. We have therefore added positive and negative likelihood ratio and the diagnostic odds ratio as prevalence-independent measures to Table 3, and also discuss these measures in the results. The sentence „Positive (LR+) and negative likelihood ratios (LR-) and diagnostic odd ratios (diag OR) were estimated as prevalence independent measures» (p12, l16-19) was added to the Statistical analysis part. Please see also our Reply 18 to Reviewer 3.

5. Reviewer 2 wrote: Figure 2 plots are named violin plots in the figure title, but the figure presents box plots overlaid on dot plots of the glucose measurements (violin plots would present the probability density of the glucose measurements, see for example https://doi.org/10.1371/journal.pone.0238835). A corrected description could also be added to the Statistical analysis section.

Answer: Thank you very much for this comment. Indeed, the box plots in Figure 2 are not violin plots. We corrected the legend of Figure 2 according to the suggested description.

6. Reviewer 2 wrote: Discussion: It would be worth double-checking that the information extracted from each referenced paper is correct. I noticed that the statement "Early hyperglycaemia resulted in accelerated fetal growth < 20th weeks of gestation" (reference 22) is incorrect. The study in question reported an association between GDM diagnosed at or after 28 weeks of gestation and accelerated fetal growth from 20 to 28 weeks of gestation, but it did not include information on early hyperglycaemia or estimates of growth acceleration before 20 weeks of gestation.

Answer: We are very sorry for the wrong citation. As the study does not fit perfectly in the context of early hyperglycemia and accelerated fetal growth we dropped the reference completely. We double-checked all citations.

7. Reviewer 2 wrote: The manuscript contains a number of spelling mistakes and would benefit from further proofreading. For example:
Under "What this study adds": "stuff" should be spelled "staff".
In Introduction: "is continuously increases" should be corrected.
In Methods under Participants: "chronic liver, kidneys or heart diseases" and "in previous a pregnancy" should be corrected.
Discussion: the name "Jokalinen" should be spelled "Jokelainen".

Answer: All typing errors were corrected and we did a proofreading.

Comments of Reviewer 3:

8. Reviewer 3 wrote: The OGTT has poor reproducibility and standardisation of each step in the total testing process is critical to ameliorate this issue. As the glucose results are critical to the findings of the study this section needs more detail.

Specifically, how were the patients prepared for the OGTT / what advice was given regarding smoking/ diet prior to testing/ behaviour during testing?

Answer: All women receive a short leaflet prior the OGTT 75g. The participants were asked to drink and eat normally during 2 days before testing. The woman should not eat, drink (only few sips of water) or smoke in the morning of the test. They were advised not to perform any unnecessary physical activity during the test.

9. Reviewer 3 wrote: Fasting prior to test:
10-14h of fasting is difficult for most but particularly challenging in pregnancy. The requirement is a minimum of 8h? Was this data on the period of fasting collected at each centre?

Answer: As a protocol requirement, all women were requested to stay fasting for a period of at least 10 hours. The women were asked on the day of the test as well.

10. Reviewer 3 wrote: When were the OGTT performed at each centre – I assume all were performed in the morning post overnight fast?

Answer: All OGTTs were performed in the morning between 8 and 12 am.

11. Reviewer 3 wrote: How were samples collected at each of the centres (preanalytical factors): Was each sample for glucose measurement sent immediately on collection to the laboratory for analysis or were they sent together on completion of the OGTT?

Answer: All centers were advised to send the samples directly after each blood draw to the laboratory for analysis and not to wait for the completion of the test.

12. Reviewer 3 wrote: What steps were made to minimise the effects of glycolysis in vivo? What specimen tubes was whole blood collected into for glucose (e.g., Na/F or Na/F/citrate tubes/ Serum/Li heparin sample separated and processed within 30 minutes of blood draw) measurement.

Answer: To keep the effects of glycolysis in vivo to a minimum, all centers were asked to send the samples to the laboratory directly. Only Natriumfluorid tubes (NaF) with citrat buffer were used for the study (now added to the Method section p10, l1-4). The manufacturer of the tubes were not predefined. Sarstedt (S-Monovette® GlucoEXACT) was mainly used.

13. Reviewer 3 wrote: Was sample collection standardised across the different centres?

Answer: A sampling protocol was distributed to all participating centers. The main focus was which tubes to use for OGTT and the aliquots, the times and details of centrifugation for the aliquoting of serum and plasma samples.

14. Reviewer 3 wrote: What was the analytical performance (precision, bias, traceability/calibration) of the respective glucose methods at each laboratory where glucose was analysed. Were all laboratories accredited for glucose to ISO15189;2012 standards? Suggest including a supplementary table with this data.

<u>Answer:</u> *Plasma glucose was measued by an automated colorimetric enzymatic method using the hexokinase/glucose-6-phosphate-dehydrogenase method (GLUC3 test by Roche or the Dimension Vista® Hexokinase test) and are analysed by Hitachi/Roche cobas® modular analyser, Roche Diagnostics (Rotkreuz, Switzerland)(Basel, Freiburg, Salzburg, Vienna, Zurich) or Siemens Dimension Vista® analyser, Siemens Healthcare Inc.(Aargau). Both tests have: Imprecision < 1.25%, Bias < 1.23%. The tests are ISO17025 accredited and include the obligation for external quality control (interlaboratory comparisons). All participting laboratories were ISO 17025:2017 or ISO15189 accredited. This was also added to the Methods section p10, l4-13.*

Statistics comments:

<u>15. Reviewer 3 wrote:</u> Did the authors recruit the appropriate number of participants to demonstrate that an early OGTT (≤15 weeks 'gestation) might be used to predict/diagnose GDM. Moreso, how did the authors determine sample size to ensure the study is adequately powered? I see this information is provided in the study protocol, but I think a statement of sample size/ power calculation should be added to the current manuscript & referencing again the protocol paper.

<u>Answer:</u> *We estimated a sample size of 748 women (65 women with GDM=10.9% prevalence) with a dropout rate of 15% in order to predict the development of GDM using an early OGTT 75g and/or additional biomarkers. As written in the study protocol. Referencing was added to the manuscript.*

<u>16. Reviewer 3 wrote:</u> Table 1 PG 12

1. Please include p-values and data on height, weight, gravidity, and blood pressure.

<u>Answer:</u> *P-values are not a suitable measure for assessing group differences in baseline characteristics. We would like to refer herefore to Vandenbroucke's et al. 2007 elaboration on the STROBE guidelines: „Inferential measures such as standard errors and confidence intervals should not be used to describe the variability of characteristics, and significance tests should be avoided in descriptive tables." <u>https://www.acpjournals.org/doi/10.7326/0003-4819-147-8-200710160-00010-w1</u>. Instead we have added standardized mean differences to Table 1. Data on height, weight, gravidity and systolic and diastolic pressure have been added to the Table 1. However, we will include the p-values based on the Editor`s decision.*

2. Age of women ranged 18 – 45 years. Is the data normally distributed? If not, please use median (IQR).

<u>Answer:</u> *Age was normally distributed and is hence summarized by mean and standard deviation.*

3. Please include p values in the commentary below Table 1 PG 12 Lines 3 to 5.

Answer: *See our response above regarding Table 1 PG12. Standardized mean differences (SMD) have been added to Table 1.*

17. Reviewer 3 wrote: Table 2 PG 13
Please include p-values

Answer: *Table 2 provides* ***descriptive*** *summary statistics (median and interquartile range) of glucose levels and there is* ***no inference*** *to be drawn from this table. Therefore, there are no p-values to be included in this table. Please see also our response above regarding Table 1 PG12. P-values are measures derived from statistical hypothesis tests in order to draw inference. This clearly is not our intention here. Testing GDM and non-GDM women for a difference in their glucose levels is not the objective when assessing diagnostic accuracy. The glucose levels are not the outcome that we want to test. Rather, we need to „switch" and think the other way around; the test result (GDM or not GDM) – or, the correct classification of patients as having GDM or not, based on their glucose measurements - is the outcome that we are interested in. The glucose levels are hence the predictors, and the test result is the outcome. Testing the difference in glucose levels between women with and without GDM simply would not provide any useful information to our results. However, we will include the p-values based on the Editor`s decision.*

18. Reviewer 3 wrote: Table 3 PG 15
1. mg/dL – no values in table

Answer: *We apologize for the missing conversions. The values for mg/dl were added to Table 3.*

2. Can the authors consider including an additional Figure detailing the ROC curves for this data? It would be helpful for comparison to other studies e.g., Corrado et al & would be in keeping with the stated primary objective in the published protocol.

Answer: *We now provide an additional Figure showing the ROC curves with the AUCs indicated in the Appendix (supplement Figure 4).*

3. Likelihood ratios (LRs) that allow one to determine how much the utilization of a particular test will alter the probability would be useful to include too.

Answer: *Positive and negative likelihood ratios and the diagnostic odds ratios have been added to Table 3, and are discussed. See also our Reply 4 to Reviewer 2.*

19. Reviewer 3 wrote: Discussion PG 17 Line 7
"Setting the FPG cut-off at ≥5.1 mmol/L raise concerns for overdiagnosis, since FPG decreases until 20th weeks' of gestation."
You assessed the lowering of decision thresholds for post load glucose values. Was there a lower

fasting plasma glucose concentration e.g., 4.7 mmol/L, in early pregnancy below which GDM did not develop?

<u>Answer:</u> *The lowest measured value in fasting plasma glucose concentration in early pregnancy in a woman who developed GDM was 3.1 mmol/L. This value was also measured in three women who did not develop GDM, and there was only one woman who did not develop GDM with a even lower measurement of 2.3 mmol/L.*

<u>20. Reviewer 3 wrote:</u>
HbA1c was measured in all participants between weeks'12 and 15 gestations, did you correlate with the early OGTT and late OGTT? If so, please include this data.

<u>Answer:</u> *We did measure HbA1c in the subset cohort. But the analysis of the correlation is ongoing and will be published separetely.*

<u>21. Reviewer 3 wrote:</u>
Plasma glucose levels ≤ 2.5 mmol/L – how many PG results were ≤2.5 mmol/L? and what was the reason for these results? Improper sample handling/clotted sample/patient vomited post consummation of glucose load?
Were these results included in the final analysis?

There was 1 sample for fasting glucose (n=1 from Salzburg), 2 samples for glucose after 1h (n=1 from Freiburg and n=1 from Salzburg), and 16 samples for glucose after 2h (n=2 from Aarau, n=6 from Basel, n=3 from Freiburg, n=2 from Salzburg, n=1 from Vienna and n=2 from Zurich). If reasons for the hypoglycemia were not obvious, the values were kept for the final analysis. If women vomited post consummation we did not continue to draw blood after 1 and 2 hours.

<u>22. Reviewer 3 wrote:</u>

Even in this low-risk cohort, would clinically relevant risk factors for GDM like previous GDM history, higher BMI, ethnicity, inactivity etc at the first antenatal visit be as good as the early OGTT in predicting GDM?

<u>Answer:</u> *We will thank the reviewer for this important remark. We feel that this information will overload possibly overload this manuscript. We plan to perform different prediction models in future studies.*

<u>23. Reviewer 3 wrote:</u> Pg 18-19
Could you elaborate /speculate on the mechanism as to why glucose targets in women with GDM identified in early pregnancy were more difficult to achieve and so, negatively impacted the reduction of complications in early diagnosed GDM?

<u>Answer:</u> *It seems that milder degrees of hyperglyaemia, lower than the threshold for preexisitng diabetes, but diagnosed before 24 weeks of gestation might be a phenotype of GDM which are at*

higher risk for adverse outcome and the women diagnosed after 24 weeks of gestation seems to be a milder phenotype. Therefore, treatment approach of the early diagnosed GDM phenotype might be justified but more challenging (p25, l21- p26, l2).

24. Reviewer 3 wrote:

Strengths and weaknesses
Universal screening will include all pregnant women, those at low, medium and high risk of GDM. This study cohort reveals the utility of early OGTT testing in this study's specific population - a low-risk population. Please amend inference on PG 17 lines 23-24.

Answer: The sentence was corrected accordingly.

25. Reviewer 3 wrote:

Pg 18 lines 5-8
Re: we did not further investigate maternal characteristics and co-morbidities between early and late diagnosed GDM women and therefore cannot describe an early-onset GDM phenotype.
I hope that this will be the subject of a further manuscript – if so perhaps state.

Answer: We thank you for the suggestion. Allthough we are not sure if the sample size will allow the characterisation of an early and a late phenotype in our cohort.

26. Reviewer 3 wrote: Minor
What this study adds
Line 20: Typo – Stuff – should read staff

Answer: That typing error was corrected .

Comments of Chief Statistics Editor:

27. Chief Statistics Editor wrote: "The analysis set included only women with complete early and late OGTTs" – but this leads to confusion about what to do with those individuals with inconclusive results – better to consider them in sensitivity analysis, see
Shinkens https://pubmed.ncbi.nlm.nih.gov/23682043/

Answer: Indeed this is an important point. We report the numbers and reasons for missingness/dropping out in the patient flow chart (Figure 1). In total there are 38 women with incomplete late OGTT for which no final diagnosis could be made. For these women, we provide summary statistics of the available glucose measurements in an additional Table in the Appendix (Supplementary Table 5). The most common pattern was that fasting plasma glucose was assessed but either one or both stimulated glucose measurements were not taken because of nausea and/or vomiting, forgotten 1h blood sample draw as a protocol violation or lack of time on the women´s side.

Further there are 10 women for which the early OGTT was performed not per protocol (> 15+0 weeks), 25 women with incomplete early OGTT and 16 women for which the late OGTT was performed not per protocol. Here we decided to perform a sensitivity analysis including all of the above mentioned women to the full analysis set (FAS), with a complete late OGTT to derive a final diagnosis. This approach resulted in an additional 27 women (663 compared to 636 in the FAS) and 3 additional GDM cases (77 versus 74 in the FAS). We have reestimated the diagnostic performance measures using the respective available data for each measurement time (totals differ due to varying patterns of missingness) and present these in the Appendix (Supplementary Tables 6-9). Results are consistent over the FAS and the sensitivity analysis. This information was added to the manuscript under the Results section p22, l19-p23, l8.

28. Chief Statistics Editor wrote: When examining the test accuracy at 12-15 weeks, the reference standard is the test result at 24-28 weeks – but surely this is also not perfect? Should we not be comparing the test accuracy to the true gold standard? Otherwise, how do we know the actual accuracy of the test at 12-15 weeks?

Answer: Up to date, the reference test for the diagnosis of gestational diabetes is the oral glucose tolerance test OGTT 75g in 24-28 weeks of gestation as recommended by the International Association of Diabetes and Pregnancy Study Groups Consensus Panel in »International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy.» in Diabetes Care 2010.

29. Chief Statistics Editor wrote: Looking at the protocol, sample size was based on precise estimation of the AUC. However, this is not the important measure – rather, we need to have precise estimates of sens, spec, PPV and NPV at the key thresholds of interest. Indeed, the authors never even report AUC in the actual paper, so I'm not sure why this changed. Sadly, this has led to a problem with wide confidence intervals at most thresholds, especially for PPV, which is a key measure for clinical decision making. E.g. at fasting glucose, a cut-off of 5.7 has a CI for PPV from 15 to 95%. So wide, that I do not see how we can make conclusions about the test performance (even ignoring the true reference standard issue).

Answer: Thank you for looking closely into the study protocol. There has not been a change from the study protocol. Indeed, sample size was calculated for estimation of the AUC using the early OGTT +/- glyFn and/or other biomarkers. The results reported here are part of the analysis. The objective was to assess the diagnostic performance of the WHO 2013 criteria based on the early OGTT.

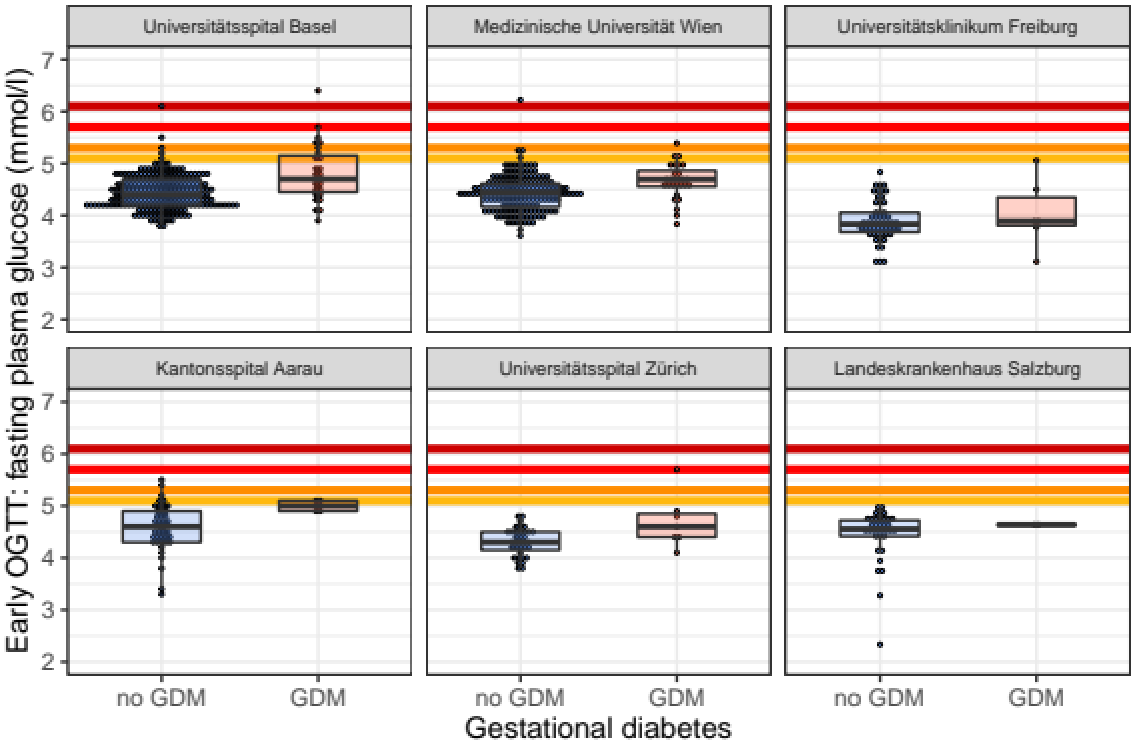
30. Chief Statistics Editor wrote: Were results consistent across centres, or is there between-center heterogeneity in test accuracy? Prevalence tends to vary across settings, so this could impact upon PPV and NPV.

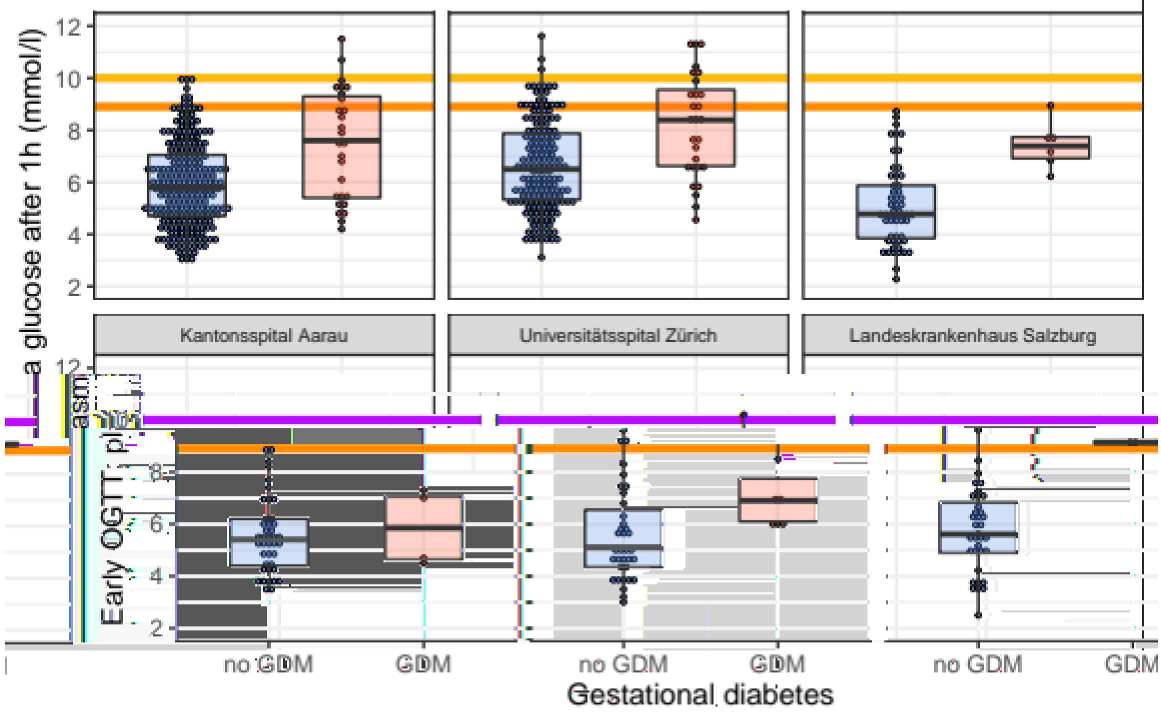
Answer: *This is a good point. The number of patients provided to the analysis set varied considerably between the six centres (37 to 254), with a „point prevalance" of GDM ranging from 5 to 17 percent (however , these proportions have to be taken with a grain of salt and taking into account the respective sample size of the center). We have summarised this information in Supplementary Table 4 (provided in Appendix).*

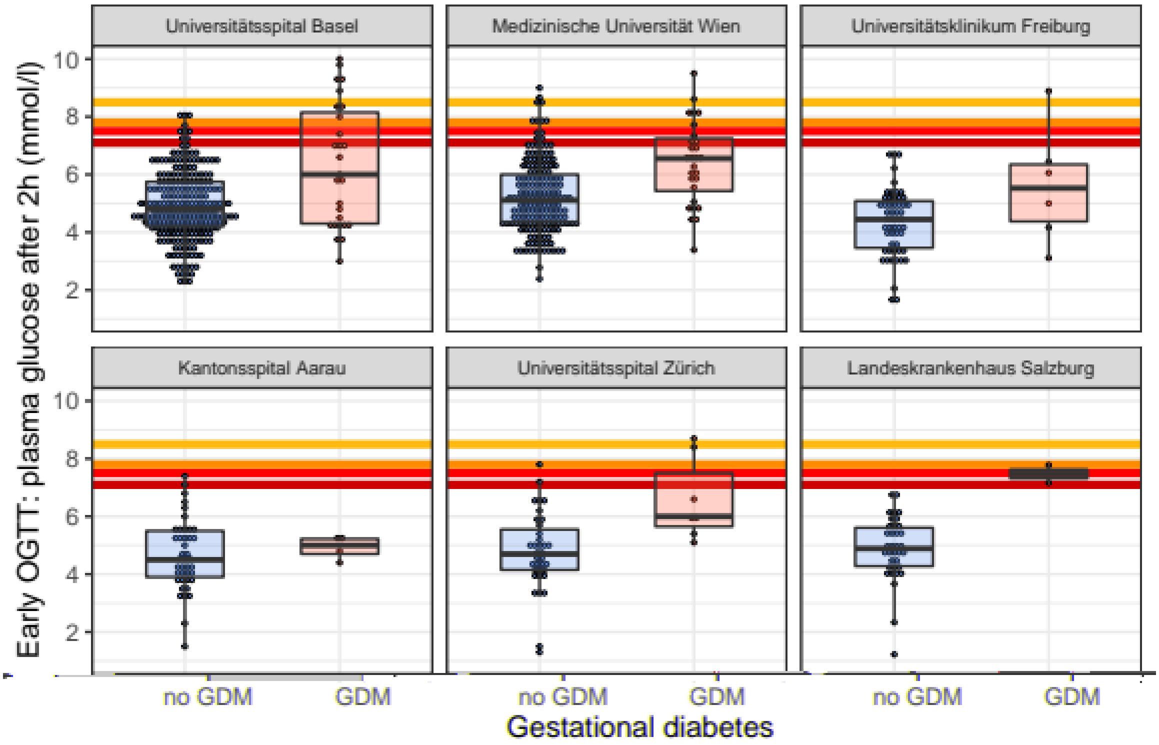
Centre	**N**	**GDM**
Universitätsspital Basel (GDM)	254	27 (11%)
Medizinische Universität Wien (GDM)	198	28 (14%)
Universitätsklinikum Freiburg (GDM)	60	6 (10%)
Kantonsspital Aarau (GDM)	45	4 (9%)
Universitätsspital Zürich (GDM)	42	7 (17%)
Landeskrankenhaus Salzburg (GDM)	37	2 (5%)
Total	636	74 (12 %)

Since the number of women with GDM is low (<10) in 4 centres, we decided to visually present the individual glucose measurements by each centre separately, with the respective cut-offs indicated (i.e. original Figure 2 of the manuscript splitted into subplots by centre). These Figures are provided in the Appendix as Supplementary Figures 5a-c. Considering the figures and the respective resulting numbers (such as true positives etc), deriving centre-specific diagnostic measures did not seem sensible to us.

Comparing the two largest centres (Basel and Wien), there are no striking differences in data distribution visible. We would therefore conclude that these two populations are comparable. Comparing the other four smaller centres, we see some variability in the data dsitributions and median tendencies but taking into account the respective sample sizes we would conclude that these centres are comparable as well.


Universitätsspital Basel
Medizinische Universität Wien
Universitätsklinikum Freiburg
Kantonsspital Aarau
Universitätsspital Zürich
Landeskrankenhaus Salzburg
Early OGTT: fasting plasma glucose (mmol/l)
7
6
5
4
3
2
no GDM
GDM
Gestational diabetes


12
10
8
6
4
2
a glucose after 1h (mmol/l)
Kantonsspital Aarau
Universitätsspital Zürich
Landeskrankenhaus Salzburg
Early OGTT: pl
no GDM
GDM
Gestational diabetes



<u>*31. Chief Statistics Editor wrote:*</u> How was PPV derived? Using the observed prevalence of the outcome in the data at hand? (sometimes we take this from external data, especially if it is more accurate elsewhere).

<u>Answer:</u> PPV was indeed derived based on the prevalenc of our study population. For transparency we have added this information to the manuscript (p14, l3-5): „Positive and negative predictive value were derived based on the observed prevalence in the analysis set.“

<u>*32. Chief Statistics Editor wrote:*</u> “Increasing the fasting cut-off did not improve overall diagnostic accuracy” – I’m not sure I agree, as it does increase PPV, whilst barely reducing NPV. I think the authors should focus more on PPV and NPV rather than ‘overall accuracy’ based on sens and spec.

<u>Answer:</u> Thank you for pointing this out, this objection is correct; also the diagnostic odds ratio increases. We have hence rephrased the sentence to „Increasing the fasting cut-off slightly improved test performance (increasing PPV whilst barely reducing NPV) and effectiveness (diagnostic odds ratio).“

<u>*33. Chief Statistics Editor wrote:*</u> Minor - What this study adds: stuff to be changed to staff.

<u>Answer:</u> This typo was corrected.

<u>*34. Chief Statistics Editor wrote:*</u> The protocol mentions many aspects that are not reported here (including ROC curves, multivariable models, random forests, lasso, etc).

<u>*Answer:*</u> *This is correct and no discrepancy, but we appologize for not being clear enough. As a first paper oft he study, we aimed to adress the diagnostic performance of the 2013 WHO criteria and other suggested cut-offs for development of GDM. A further objective of the study is the development of a novel screening/prediction algortithm for GDM based on OGTT and/or novel biomarkers and maternal characteristics. The last mentioned analyses will include the mentioned aspects and require the more complex methods. These analyses are still ongoing and will be reported as soon as possible. See also our response to point 29.*

<u>*35. Chief Statistics Editor wrote:*</u> In summary, the wide confidence intervals for PPV, and the uncertainty about using 24-28 weeks as the truth, led to some confusion about what this study can conclusively decide in terms of the best thresholds to use and the value of tests at 12-15 weeks. There is some useful information in here, and it may be especially important if added to meta-analyses on this topic later, but regardless, I hope these comments help the authors and BMJ Medicine moving forwards.

<u>*Answer:*</u> *Thank you. We hope that the additional information provided and the transparent reporting is helpful for understanding and interpreting our results.*

VERSION 2 – REVIEW

REVIEWER 1	Zhang, Huijing; Peking University First Hospital, Department of Obstetrics and Gynaecology. Competing Interest: None
REVIEW RETURNED	04-Jan-2023

GENERAL COMMENTS	Nil

REVIEWER 3	O'Shea, Paula; Mater Misericordiae University Hospital, Clinical Biochemistry & Diagnostic Endocrinology. Competing Interest: None
REVIEW RETURNED	24-Dec-2022

GENERAL COMMENTS	I am satisfied with the responses to my queries/comments. I believe that the input/suggestions from all reviewers has (I hope you agree), much improved the quality of the manuscript. The study findings are valuable, add to the current knowledge, and I am confident will greatly appeal to the readership of the bmjmedicine.

REVIEWER 4	Riley, Richard; University of Birmingham, Institute of Applied Health Research. Competing Interest: None
REVIEW RETURNED	20-Jan-2023

GENERAL COMMENTS	I thank the authors for a very comprehensive and clear response and revision. I am generally happier and we are getting there, but do have some important comments remaining: 1) The abstract states the aim is to assess ‘the “gold standard” for the diagnosis of gestational diabetes mellitus (GDM) by 75g oral glucose tolerance test (OGTT) between 24 and 28 weeks of gestation’ – But actually the aim is to assess the test at between 10 and 14 weeks, not 24 and 28 weeks, as the latter is the reference standard. 2) The paper is suggesting to lower cut-off values, yet the abstract does not give results for lower cut-off levels. It says “Lowering the post-load glucose values ameliorate the detection rate with 53% at a false positive rate of 10%.” – but what cut-off level is this, and what are the sens, spec, PPV and NPV at these cut-offs. It’s very vague. 3) I asked the authors to provide results by center. They do this for prevalence, but not for other key measures such as PPV and NPV, sens and spec. They refer to Figures 5a to 5c, but these present distributions, and not test accuracy results. 4) To address point 3, I do think a forest plot of the results by center would be welcome, with a meta-analysis result and estimate of heterogeneity provided, for example using the Cochrane DTA meta-analysis methods approach. 5) Page 18/19: “WHO 2013 criteria showed a low sensitivity …” – make it clear that this is the 2013 criteria applied at an earlier timepoint than the 24-28 recommended by WHO. 6) I mentioned that the gold-standard is also measured with error, and the authors state that the reference test for the diagnosis of gestational diabetes is the oral glucose tolerance test OGTT 75g in 24-28 weeks of gestation as recommended by the International Association of Diabetes and Pregnancy Study Groups Consensus Panel. This may be the case, but this does not negate my point that could this reference test also be measured with misclassification? I do think this point needs clear discussion. Does the 24-28 test every misclassify? Surely it does? 7) The abstract suggests to lower the cut-off to improve the number who will be treated correctly but surely this could also lead to potential harm (treated unnecessarily)? There is a trade-off which needs to be discussed. Also, the authors themselves state that “There is a paucity of studies showing a sufficient effect of early diagnosis and intervention in these women” – so I think the abstract statement that lowering the cut-off could potentially improve neonatal and maternal outcome. 8) As mentioned in previous review, the sample size section should be explicit that this study was powered on the AUROC, and not on sens and spec and specific cut-points. This is very important for transparency. 9) Table 3 – please also provide the sample size(s) used in these analyses – also in supp material tables 8 to 9.

REVIEWER 5	Pressman, Eva; University of Rochester, OB/Gyn. Competing Interest: None
REVIEW RETURNED	24-Feb-2023

GENERAL COMMENTS	This is a well done prospective multicenter study evaluating the use of a 75g oral glucose tolerance test at 12-15 weeks in low risk patients who also underwent testing at 24-28 weeks. While this

	study provides new information on potential cutoffs for early screening, its overall impact on the field is limited. This topic has been studied several times before using slightly different methods, most often HgbA1C at the first prenatal visit but 2 studies have used the 2 hour 57 g load used in this study (D Simmons, J Nema, C Parton, et al. The treatment of booking gestational diabetes mellitus (TOBOGM) pilot randomised controlled trial BMC Pregnancy Childbirth, 18 (2018) and CA Vinter, MH Tanvig, MH Christensen, et al. Lifestyle intervention in Danish obese pregnant women with early gestational diabetes mellitus according to WHO 2013 criteria does not change pregnancy outcomes: results from the LiP (lifestyle in pregnancy) study. Diabetes Care, 41 (2018)) but were not referenced in this paper. A recent meta-analysis on this topic (McLaren RA Jr, Ruymann KR, Ramos GA, Osmundson SS, Jauk V, Berghella V. Early screening for gestational diabetes mellitus: a meta-analysis of randomized controlled trials. Am J Obstet Gynecol MFM. 2022 Aug 27;4(6):100737) would also be a good reference. Given that literature, a few other points to address in the introduction and discussion include: 1. GDM has been related to placental hormonal effect on insulin resistance which likely increases with increasing gestation. Would diagnosis at 20 weeks make more sense than diagnosis at 12 weeks? 2. Though earlier diagnosis and treatment may improve the incidence of fetal macrosomia, other studies have not shown many other benefits 3. The methods indicate Hgb A1C was only done if there was a concern for pre-existing diabetes. Other studies have used Hgb A1C as a screening method. Please discuss why you did not.

VERSION 2 – AUTHOR RESPONSE

Comments of Reviewer 3:

1. Reviewer 3 wrote: I am satisfied with the responses to my queries/comments. I believe that the input/suggestions from all reviewers has (I hope you agree), much improved the quality of the manuscript. The study findings are valuable, add to the current knowledge, and I am confident will greatly appeal to the readership of the BMJ Medicine.

Answer: We thank you very much for your kind review.

Comments of Chief Statstics Editor/Reviewer 4:

Comments to the Author

I thank the authors for a very comprehensive and clear response and revision. I am generally happier and we are getting there, but do have some important comments remaining:

2. Reviewer 4 wrote: The abstract states the aim is to assess ‘the “gold standard” for the diagnosis of gestational diabetes mellitus (GDM) by 75g oral glucose tolerance test (OGTT) between 24 and 28

weeks of gestation' – But actually the aim is to assess the test between 10 and 14 weeks, not 24 and 28 weeks, as the latter is the reference standard.

Answer: We thank you very much for the comment. The abstract now states the following: „To evaluate the 75g oral glucose tolerance test (OGTT) in early pregnancy using the WHO 2013 criteria – normally used between 24 and 28 weeks of gestation to diagnose gestational diabetes mellitus (GDM) – and additionally, to test newly proposed cut-off values. " (p4, l2-7)

3. Reviewer 4 wrote: The paper is suggesting to lower cut-off values, yet the abstract does not give results for lower cut-off levels. It says "Lowering the post-load glucose values ameliorate the detection rate with 53% at a false positive rate of 10%." – but what cut-off level is this, and what are the sens, spec, PPV and NPV at these cut-offs. It's very vague.

Answer: We changed the abstract accordingly: „Lowering the post-load glucose values (OGTT 75g cut-offs of 5.1/8.9/7.8mmol/L) ameliorate the detection rate with 53% [95% CI 0.41; 0.64] and NPV (0.94 [95% CI 0.91; 0.95]), specificity (0.91 [95% CI 0.88; 0.93]) and PPV (0.42 [95% CI 0.32; 0.53])) at a false positive rate of 9% (LR+ of 5.59 [95% CI 4.0-7.81], LR- of 0.64 [95% CI 0.52-0.77], diag OR of 10.07 [6.26-18.31])." Because of the restricted word count, we left out the confidence interval. We hope, that this still works for the Editorial office. (p5, l4-9)

4. Reviewer 4 wrote (point 3 and 4): I asked the authors to provide results by center. They do this for prevalence, but not for other key measures such as PPV and NPV, sens and spec. They refer to Figures 5a to 5c, but these present distributions, and not test accuracy results. To address point 3, I do think a forest plot of the results by center would be welcome, with a meta-analysis result and estimate of heterogeneity provided, for example using the Cochrane DTA meta-analysis methods approach.

Answer: We thank for the idea and conducted a meta-analysis for the WHO 2013 criteria. To the section ´Statistical analysis´ the following description was added: Between-center heterogeneity was evaluated using a meta-analysis for the WHO 2013 criteria. Forest plots for the diagnostic measures were derived. Models were fit with center as random effect (random intercept); logistic regression was used for proportions (sensitivity, specificity, accuracy, PPV and NPV), and the Mantel-Haenszel method was used for the likelihood ratios (LR+ and LR-). For the Generalized Linear Mixed Models (GLMMs), no weights for centres are provided by this approach." (p14, l2-9)

We report in the section ´Results´: „In order to address potential between-center heterogeneity we performed a meta-analyses and show forest plots for the diagnostic measures derived from the WHO 2013 (see supplementary Figure 6)." (p21, l20-22)

6. Reviewer 4 wrote: Page 18/19: "WHO 2013 criteria showed a low sensitivity …" – make it clear that this is the 2013 criteria applied at an earlier timepoint than the 24-28 recommended by WHO.

Answer: Thank you very much for your advice. We now emphasise the test being evaluated in „early pregnancy" by stating: „The WHO 2013 criteria – the standard for GDM screening between 24 and 28 weeks of gestation – showed a low sensitivity of 0.35 [95% CI 0.24; 0.47] and a high specificity of 0.96 [95% CI 0.95; 0.98] in early pregnancy." (p19, l5-8)

7. Reviewer 4 wrote: I mentioned that the gold-standard is also measured with error, and the authors state that the reference test for the diagnosis of gestational diabetes is the oral glucose tolerance test OGTT 75g in 24-28 weeks of gestation as recommended by the International Association of Diabetes and Pregnancy Study Groups Consensus Panel. This may be the case, but this does not negate my point that could this reference test also be measured with misclassification? I do think this point needs clear discussion. Does the 24-28 test every misclassify? Surely it does?

Answer: We thank you very much for the comment. We added to the ´Discussion`section: „As already mentioned in the introduction section, the 2013 WHO cut-off values are derived from the HAPO study published in 2008.(3) The HAPO study reported a linear relationship between maternal fasting and post laod and perinatal adverse outcomes like birth weight > 90th percentile, cord c-peptid > 90th percentile as well as neonatal body fat > 90th percentile. No glucose value was soley strongly correlated with the outccomes and no single value was superior to the others to predict GDM diagnosis. The defined cut-off values represent the average glucose values at which the odds were 1.75 times increased to develop the mentioned outcomes, based on fully adjusted logistic regression models. Finally, only one cut-off value were needed to screen-positive and diagnose GDM. Though, the „no threshold"-effect of the OGTT 75g lower reproducability which could lead to missclassification by the reference test. Additionally, there are a quite few factors which can influence test reproducability and test accuracy preanalytically (time of fasting maternal diet, smoking, excercise or stress, type oft he sample, collection tube, storage and transportation etc.), analytically and postanalytically.(21,22) A Chinese study showed an overall reproducability of only 65.6% in men and non-pregnant women between two OGTTs performed in a 6 weeks interval.(23) The reproducability could not be improved even in the high risk group with high HbA1c, BMI or waist-to-hip ratio. There is only one small study reporting low overall reproducability of 74.2% using an OGTT 75g in pregnancy in a Sub-Saharan African population.(24) In the end, these aspects are important for the interpretation of our results. In the background of the reported low reproducability of the OGTT 75g, the glucose results might lead to misclassification of the GDM diagnosis in the predefined period of screening between 24 and 28 weeks of gestation and it surely have an impact on the resuts of the OGTT 75g in early pregnancy as well. As the OGTT 75g is the reference standard in pregnancy, these difficulties cannot be overcome in the moment, but biomarkers of glucose metabolism or continuous glycaemic monitoring (CGM) might improve diagnostic reproducability and accuracy and are currently under evaluation.(11,25,26)" (p24, l11-p25, l19)

8. Reviewer 4 wrote: The abstract suggests to lower the cut-off to improve the number who will be treated correctly but surely this could also lead to potential harm (treated unnecessarily)? There is a trade-off which needs to be discussed. Also, the authors themselves state that "There is a paucity of studies showing a sufficient effect of early diagnosis and intervention in these women" – so I think the abstract statement that lowering the cut-off could potentially improve neonatal and maternal outcome.

Answer: We thank you very much for the comment. To the ´Conclusion´ section in the abstract, the sentence: „However, randomized controlled trials showing a beneficial effect of early intervention are ambiguous." were added.. (p5, l13-16)

There are very interesting studies which were published recently about the topic. Though we discuss the potential results of these studies in a separate paragraph in the section `Discussion` after „It seems that milder degrees of hyperglyaemia, lower than the threshold for preexisitng diabetes, but diagnosed before 24 to 28 weeks of gestation might be a phenotype of GDM which are at higher risk for adverse outcome and the women diagnosed after 24 weeks of gestation seems to be a milder phenotype. Therefore, treatment approach of the early diagnosed GDM phenotype might be justified but more challenging. and benefits of an early intervention were lacking.(7).": „A RCT published in 2020 (the Early Gestational Diabetes Screening in the Gravid Obese Woman (EGGO trial)(30) investigated early screening between 14 and 20 weeks of gestation by a two step screening approach (first 50g GCT followed by a 100g OGTT using Carpenter and Coustan criteria) in an obese population. The study reported no improvement in the primary outcome (macrosomia defined as a birth weight > 4000g) in the early diagnosed GDM group (n=29 < 20 weeks of gestation). The intervention group size was low and underpowered for the early intervention group. Particularly obesity is known to be an independent and most prevalent risk factor for neonatal macrosomia which might not be improved easily by glycemic control alone. Another Danish study about lifestyle intervention in very obese women also reported no improvement in primary obstetric and metabolic outcomes.(31) Simmons et al. recently published results of the TOBOGM trial (The treatment of booking gestational diabetes mellitus).(32) Women with at least one risk factor for hyperglycemia were tested in 4 to 20 weeks of gestation (mean 15.6 weeks of gestation) with a 2h-OGTT 75g. 802 women were randomized to immediate (n=400) or deferred treatment (n=393) (dietary advice or pharmacotherapy). The immediate intervention led to a modestly improved adverse neonatal outcome (birth at < 37 weeks`gestation, birth trauma, birth weight of > 4500g, respiratory distress, phototherapy, stillbirth, neonatal death or shoulder dystocia) (24.9% versus 30.5%, adjusted relative risk of 0,82, 95% CI, 0.68-0.98)) and no material didferences in pregnancy-related hypertension and neonatal lean body fat.A flaw of the study might be the not strongly hyperglycaemia related adverse outcomes of preterm birth, respiratory distress syndrome or phototherapy. Secondary outcomes like LGA infants (16.8% versus 19.6%) and neonatal hypoglycemia < 40mg/dl within 72 hours (18.9 versus 22.7%) could be improved by a slight increase in small for gestational age (SGA) infants (12% versus 9.2%), especially in women with lower glycemic ranges. The higher rate of SGA might be a possitble harm, but the neonatal pH/Apgar status or the possible increased neonatal intensive care unit admission seen in the pilot study in this group of neonates was not further evaluated in the main study.(33)" (p27, l9-p28, l19)

9. Reviewer 4 wrote: As mentioned in previous review, the sample size section should be explicit that this study was powered on the AUROC, and not on sens and spec and specific cut-points. This is very important for transparency.

Answer: We added a new paragraph „Sample size considerations" to the´Methods´section: „Of note, the sample size was calculated based on an area under the curve (AUC) of a newly proposed screening method combining OGTT 75g and new biomarkers like glyFn. The power calculation was performed using a proposed true AUC of 0.9 with a lower boundary of 0.8 (95% CI>0.8) which led with a power of 90% and an α-level of 5% and offsetting a dropout of 15% a total sample size of 748.(11) It is important to know that the study was not powered for the diagnostic power of specific cut-off values.» (p12, l23-p13, l8)

10. Reviewer 4 wrote: Table 3 – please also provide the sample size(s) used in these analyses – also in supp material tables 8 to 9.

Answer: We apologize for the missing cross tables which were meant to be included in table 3 and in two newly uploaded supplementary table 6 and 7. We have added the cross tables now to table 3 and to supplementary tables 8 and 9 relabeling both to supplementary tables 6 and 7.

Comments of Reviewer 5:

11. Reviewer 5 wrote: This topic has been studied several times before using slightly different methods, most often HgbA1C at the first prenatal visit but 2 studies have used the 2 hour 75 g load used in this study (D Simmons, J Nema, C Parton, et al. The treatment of booking gestational diabetes mellitus (TOBOGM) pilot randomised controlled trial

BMC Pregnancy Childbirth, 18 (2018) and CA Vinter, MH Tanvig, MH Christensen, et al. Lifestyle intervention in Danish obese pregnant women with early gestational diabetes mellitus according to WHO 2013 criteria does not change pregnancy outcomes: results from the LiP (lifestyle in pregnancy) study. Diabetes Care, 41 (2018)) but were not referenced in this paper. A recent meta-analysis on this topic (McLaren RA Jr, Ruymann KR, Ramos GA, Osmundson SS, Jauk V, Berghella V. Early screening for gestational diabetes mellitus: a meta-analysis of randomized controlled trials. Am J Obstet Gynecol MFM. 2022 Aug 27;4(6):100737) would also be a good reference.

Answer:We thank the reviewer for the comment. The TOBOGM sutdy has published their results recently in the NEJM. The TOBOGM study, the LiP and the EGGO study were added to the text under section in the section `Discussion`. See also our comment to Point 8 (Reviewer 4).

The recently published meta-analysis by Jauk et al. is mentioned in the ´Introduction´ section: Nevertheless, the WHO 2013 criteria have not been assessed for use in pregnancy before 24 weeks of gestation, especially in a low-risk population which might benefit by an early GDM screening by lowering the LGA rates at birth.(7)" (p8, l3-4) and mentioned in the `Discussion` section (see also comments to Point 8 (Reviewer 4).

12. Reviewer 5 wrote: GDM has been related to placental hormonal effect on insulin resistance which likely increases with increasing gestation. Would diagnosis at 20 weeks make more sense than diagnosis at 12 weeks?

Answer: This is a very good question. We don`t know a study which is comparing screening at 12 to 20 weeks of gestation. But the anti-insulin, hyperglycaemic effect of the placental hormones are tremendously important and screening at 20 weeks might improve detection rates of the OGTT 75g. We are discussing that between the lines in the ´Discussion´ section about the possible different phenotypes of early and late diagnosed women. The window of possible interventions are getting smaller the nearer we diagnose to the normal initiation of treatment at 24-28 weeks. Finally, that topic need to be clarified further in future studies.

13. Reviewer 5 wrote: Though earlier diagnosis and treatment may improve the incidence of fetal macrosomia, other studies have not shown many other benefits.

Answer: This point is also now disussed in the section `Discussion` (see answer to point 8 Reviewer 4.

14. Reviewer 5 wrote: The methods indicate Hgb A1C was only done if there was a concern for pre-existing diabetes. Other studies have used Hgb A1C as a screening method. Please discuss why you did not.

Answer: We did measure HbA1c in a subset cohort. A manuscript about the correlation of the early OGTT 75g, HbA1c and other markers has been prepared and will be published separetely.