Comparison of accuracy measures of two screening tests for gestational diabetes mellitus.

Marsha van Leeuwen MD ¹, Egbert J.K. Zweers MD ²,³, Brent C. Opmeer MD, PhD ⁴, Evert van Ballegooie MD, PhD ⁵,⁶, Henk G. ter Brugge MD ⁷, Harold W. de Valk MD, PhD ², Ben W.J. Mol MD, PhD ⁸,⁹, Gerard H.A. Visser, MD, PhD, Professor ⁸

¹ Obstetrics and Gynaecology, Academic Medical Centre, Amsterdam, The Netherlands
² Internal Medicine, University Medical Centre, Utrecht, The Netherlands
³ Internal Medicine, Bronovo Hospital, The Hague, The Netherlands
⁴ Clinical Epidemiology and Biostatistics, Academic Medical Centre, Amsterdam, The Netherlands
⁵ Internal Medicine, Isala Clinics, Zwolle, The Netherlands
⁶ Internal Medicine, Bethesda Hospital, Hoogeveen, The Netherlands
⁷ Obstetrics and Gynaecology, Isala Clinics, Zwolle, The Netherlands
⁸ Perinatology and Gynaecology, University Medical Center, Utrecht, The Netherlands
⁹ Obstetrics and Gynaecology, Máxima Medical Centre, Veldhoven, The Netherlands

RUNNING TITLE: Screening for gestational diabetes mellitus.

CORRESPONDENCE:
M. van Leeuwen, MD
Academic Medical Centre
Obstetrics and Gynaecology, room H4-255
Meibergdreef 9
1105 AZ Amsterdam
The Netherlands
Marsha.vanLeeuwen@amc.uva.nl

Received for publication 22 March 2007 and accepted in revised form 31 July 2007.
ABSTRACT

Objectives To compare the accuracy measures of the random glucose test and the 50 g glucose challenge test as screening tests for gestational diabetes mellitus.

Research design and methods In this prospective cohort study pregnant women without pre-existing diabetes mellitus in two perinatal centres in The Netherlands underwent a random glucose test and a 50 g glucose challenge test between 24 and 28 weeks of gestation. If one of the screening tests exceeded predefined threshold values, the 75 g oral glucose tolerance test (OGTT) was performed within one week. Furthermore the OGTT was performed in a random sample of women in whom both screening tests were normal. Gestational diabetes was considered present when the OGTT (reference test) exceeded predefined threshold values. Receiver operating characteristic analysis was used to evaluate the performance of the two screening tests. The results were corrected for verification bias.

Results We included 1301 women. The OGTT was performed in 322 women. After correction for verification bias, the random glucose test showed an area under the ROC curve (AUC) of 0.69 (95% CI 0.61 to 0.78) whereas the glucose challenge test had an AUC of 0.88 (95% CI 0.83 to 0.93). There was a significant difference in area under the curve of the two tests of 0.19 (95% CI 0.11 to 0.27) in favour of the 50 g glucose challenge test.

Conclusions In screening for gestational diabetes mellitus, the 50 g glucose challenge test is more useful than the random glucose test.
Gestational diabetes mellitus (GDM) is estimated to occur in 2-9% of all pregnancies (1-5). It is defined as carbohydrate intolerance with onset or first recognition during pregnancy and is associated with increased rates of adverse pregnancy outcome, such as macrosomia, shoulder dystocia, birth related trauma such as fractures and nerve palsies and neonatal hypoglycaemia and jaundice. In addition, women with GDM are at substantial higher risk to develop diabetes mellitus in later life (1,6-8). Results from a randomized controlled trial show that treatment of GDM by means of dietary advice, blood glucose monitoring and insulin therapy if required, reduces the rate of serious perinatal complications without increasing the rate of caesarean delivery (1). Based on these results, identification through screening and subsequent treatment of women with GDM appears beneficial. However, consensus on the optimal policy for screening is lacking. The American Diabetes Association recommends screening based on risk factors for gestational diabetes (aged over 25, obese, close relative with diabetes mellitus, history of GDM or a previous macrosomic infant or specific ethnicity) followed by the 50 g 1 hour oral glucose challenge test as a screening test (9-11). Other methods of screening that are regularly used are (repeated) random glucose testing and fasting glucose measurement. It is indefinite which test is the most accurate in testing women for GDM.

The diversity in screening methods may result in unidentified cases of GDM and preventable neonatal and maternal morbidity. Establishment of an optimal, evidence based screening policy to timely detect and treat gestational diabetes could contribute to a reduction of perinatal complications. Two regularly used screening tests in the Netherlands are the random glucose test and the 50 g glucose challenge test. The objective of present study was to compare these two tests as screening tests for gestational diabetes as a first step in determining optimal screening policy in gestational diabetes mellitus.

RESEARCH DESIGN AND METHODS

Patients
In a prospective cohort study, all pregnant women attending the outpatient obstetric departments at the University Medical Centre, Utrecht and the Isala Clinics, Zwolle in the Netherlands during a two-year study period, were invited to participate. Women known with pre-existing diabetes mellitus were excluded from the study, as well as those who had not reported for prenatal care in one of the two participating hospitals before 24 weeks of gestation. Only women who delivered after 28 weeks of gestation were included in the analysis.

Data
At intake the following information was obtained: obstetric history, family history of diabetes mellitus, ethnicity (categorized as Caucasian or non-Caucasian), height, self reported weight (before pregnancy), age and smoking habits (categorized as smoking or non-smoking). The body mass index (BMI) was calculated as weight (kg)/(height(m))^2.

The following data regarding pregnancy and outcome were recorded after delivery: weight gain during pregnancy, treatment with diet or insulin, duration of pregnancy in days, birth weight of the neonate in grams, Apgar score after one and five minutes and arterial and venous pH from the umbilical cord.

In all women the random glucose test was performed at intake (±12 weeks) and between the 24th and 28th week of gestation. If the random plasma glucose measured between 24 and 28 weeks of gestation was higher than or equal to 6.8 mmol/l, the random glucose test was considered abnormal. If random plasma glucose measurement was not performed between the 24th and 28th week, a random plasma glucose at intake higher than or equal to 6.8
A 50 g oral glucose challenge test was performed in between the 24th and 28th week of gestation. The test was performed irrespective of time of the day and of the last meal. Plasma glucose was measured one hour after administration of a solution containing 50 grams of glucose. The predefined cut off value for an abnormal test result was a one hour plasma glucose value of 7.8 mmol/l.

If either the random glucose test or the 50 g oral glucose challenge test exceeded the predefined threshold value, a two hour 75 g oral glucose tolerance test (OGTT) was performed within one week to confirm or rule out the presence of gestational diabetes mellitus (reference test). The OGTT was performed in the morning after a 12 hours overnight fast and three days of minimal 150-200 g carbohydrate diet. Plasma glucose was determined before and 2 hours after administration of a 75 g glucose containing solution. Gestational diabetes was considered present if venous plasma glucose equaled or exceeded the threshold values according to the WHO criteria; less than 7.8 mmol/l after 12 hours overnight fast and ≥ 7.8 mmol/l at two hours after administration of a 75 g glucose containing solution. These criteria were also applied in the randomized controlled trial in which treatment of gestational diabetes was considered beneficial. Venous plasma glucose concentration in all tests was evaluated via glucose oxidase method (Vitros, O.C.D.) in the two perinatal centers.

**Verification bias**

When a screening test is evaluated against a reference test, ideally all participating patients should undergo both the screening and the reference test. However in practice, the reference test is seldom performed in all patients, as this test is often more invasive or expensive. If only patients with verified screening test results are used to assess the performance of the screening test, calculated accuracy measures become biased, because patients with verified disease status are often only patients with an abnormal screening test result and therefore they do not represent a random sample of the population in which the screening test is used. The bias that occurs is called (partial) verification bias (12).

As in the present study the reference test was, according to the predefined protocol, not performed in all patients, we used the following procedure to correct for verification bias. We performed the OGTT (reference test) in an arbitrary subset of consecutive patients with two negative screening test results to determine the extent to which cases of gestational diabetes mellitus were missed by the screening tests. Subsequently we estimated OGTT measurements in women who were not subjected to an OGTT based on results of the random test and the 50 g glucose screening test as well as on patient characteristics using multiple logistic regression analysis. In other words, if the result of the OGTT was missing, OGTT values were estimated with multiple regression analysis, using the results of the two screening tests and available patient characteristics. This procedure to handle missing data is called imputation and is a commonly used, adequate technique to correct for verification bias (13,14). By using multiple imputation instead of single imputation, i.e. performing the imputation procedure multiple times instead of just once, uncertainty in the imputed values is reflected by the variation in imputed values across multiple imputed datasets, and thus by appropriately larger standard errors (15). The multiple imputation procedure was also used to impute incidental missing data on patient characteristics.

**Statistical analysis**

The distribution of continuous variables is reported as mean ± SD. We constructed two by two tables for abnormal and normal test results on the random glucose test and the 50 g glucose screening test against the OGTT. These tables reflect true positive, false positive, true negative or false negative test
Screening for gestational diabetes mellitus

results for both the random glucose test and the 50g glucose challenge test. Diagnostic accuracy (sensitivity, specificity, predictive values and likelihood ratios) and 95% confidence intervals were calculated. Receiver operating characteristic (ROC) analysis was used to evaluate the discriminatory power of the two screening tests.

RESULTS

We included 1305 women. Four women were excluded from analysis because they delivered before 28 weeks of gestation. Data from 1301 women were used for further analysis. Patient characteristics are presented in Table 1. Thereby the distribution of patient characteristics within the classification groups of the reference test (OGTT) can be compared. Figure 1 displays the flow of patients in our study based on the results of the subsequent diagnostic tests. Of all 1301 women at least one test result of the random glucose test was obtained. The random glucose test was performed at intake and between the 24th and the 28th week of gestation in 1169 (89.9%) and 1295 (99.5%) of the 1301 women, respectively. We used the results of the random glucose test obtained at intake for the 6 women (0.5%) in whom the random glucose measurement was not performed between the 24th and the 28th week of gestation. None of these six women had a random glucose test result higher than 6.8 mmol/L. The 50 g oral glucose challenge test was performed in 1281 women (98.5%).

There were 37 of 1301 women (2.8%) who had an abnormal random glucose test, whereas 167 of 1281 women (13.0%) had an abnormal 50 g glucose challenge test. There were 184 women (14.1%) with at least one abnormal test result (random glucose test or 50 g glucose challenge test or both). In 20 women (1.5%) both tests results were suspect for gestational diabetes mellitus.

The OGTT was performed in 322 women (24.8%). This included 146 of the 184 women (79.3%) with an abnormal screening test result and a subgroup of 176 women with two negative screening tests (Figure 1). Initially gestational diabetes was diagnosed in 46 women. After correction for verification bias 48 women were diagnosed with gestational diabetes mellitus (3.7%).

We used multiple imputation of the OGTT value for every patient in whom the OGTT was not performed. This would have been an adequate procedure, if the chance of verification of a screening test result depended solely on the result of the screening test. However, we calculated that the chance of verification was not completely independent of factors other than the results of the screening tests. In general, women with a history of GDM or perinatal death, increased BMI and women from the hospital in Zwolle were more likely to be verified, independent of the results of their screening tests. Due to this non random verification, there was a high prevalence of GDM in women with 2 negative screening tests who underwent an OGTT. As a result, the prevalence of GDM in the imputed dataset became unrealistically high (up to 15%). In order to obtain imputed data that are in line with the incidence of gestational diabetes mellitus in the Netherlands (estimated to be approximately 2 to 4%), we adjusted the imputation procedure by applying the following additional criterion to limit the number of cases classified as having GDM; Based on the same covariates (screening tests and patient characteristics), multiple imputation was repeated 100 times and unverified women were only classified as having GDM if they had consistently imputed OGTT values that were indicative for GDM (more than 75%). After this adjusted multiple imputation procedure, the prevalence of GDM in our sample was 3.7%; only two unverified women were classified as having gestational diabetes mellitus, whereas in all other women that were unverified, no gestational diabetes was assumed.

Table 2 displays results of the comparison of the two screening tests in terms of accuracy
measures calculated after correction for verification bias.

Comparison of accuracy measures after correction for verification bias resulted in an almost 5 times higher sensitivity in favour of the 50 g glucose challenge test compared to the random glucose test (70.2% (95% CI 57.1-83.3) versus 14.6% (95% CI 4.6-24.6)). The random glucose test had less false positive test results and was therefore more specific 97.6% (95% CI 96.6-98.5) versus 89.1% (95% CI 87.4-90.9). Positive predictive values for both tests were comparable, as were the negative predictive values. The likelihood ratio of an abnormal test result was larger for the 50 g glucose challenge test than for the random glucose test. The likelihood ratio of a normal test was smaller for the 50 g glucose challenge test. The area under the ROC curve was larger for the 50 g glucose challenge test (0.88 (95% CI 0.83 to 0.93) than for the random glucose test (0.69 (95% CI 0.61 to 0.78). There was a significant difference in the areas under the curve of the two tests of 0.19 (95% CI 0.11 to 0.27).

CONCLUSIONS

Evidence for screening for gestational diabetes mellitus is often inconsistent and difficult to interpret due to various screening methods and thresholds applied internationally. An evidence based policy could increase the number of identified women with gestational diabetes mellitus and therefore reduce the number of neonatal and maternal complications by providing adequate monitoring and treatment for these women. For this purpose, the present study compared the random glucose test and the 50 g glucose challenge test as screening tests for gestational diabetes. The area under the curve was larger for the 50 g glucose challenge test, indicating that the 50 g glucose challenge test was a better predictor for gestational diabetes mellitus than the random glucose test.

A potential weakness in present study is the number of missing reference tests, due to which verification bias occurred. Because verification was apparently not performed at random, other characteristics than only the screening test results influenced the chance of verification. An intuitive and straightforward procedure to correct for verification bias would be to calculate the ratio diseased / non diseased from the results of the verified patients stratified by screening test results, and to extrapolate this ratio to the unverified patients (12,16). However, this mathematical correction can only be applied if verification of patients is performed completely at random, in other words if the chance of verification is truly independent of other factors as for example patient characteristics. In addition, this results in an adjustment at the sample level, as for individual unverified patients, the disease status according to the reference test remains unknown. In order to correct for verification at the individual level, accounting for factors that influence the chance of verification, imputation techniques can be used to estimate disease status accounting for these factors (17).

There are several strategies to deal with incomplete data, also within the context of partial verification (17). As in our study various imputation strategies consistently lead to a considerable higher number of cases, this would consequently imply unrealistically high prevalence rates. We therefore had to apply an additional criterion to limit the number of cases classified as having gestational diabetes, by means of repeating the multiple imputation procedure for the OGTT 100 times, only classifying women as having GDM if they had consistently imputed values for the OGTT that were indicative for GDM (more than 75 out of 100 times). Further research is required to evaluate which approach is preferred, thereby also accounting for the epidemiological context of the study.

The overall prevalence of gestational diabetes in the literature varies from 2-9% (1). In western countries such as the Netherlands, the incidence is more often towards 2% than 9%. Hypothetically the incidence of gestational diabetes could be
systematically underestimated in the literature (if these estimates have been based solely on selectively verified patients). In that case, we also underestimated the incidence of gestational diabetes and consequently our approach would have been suboptimal. However, it is not very plausible that for years the incidence of gestational diabetes has been underestimated, so application of the described method should have corrected properly for this verification bias (18,19).

Results from the present study show that the 50 g glucose challenge test has an almost fivefold higher sensitivity compared to random glucose testing. To our knowledge, these two screening tests have only been equated in the same sample two times before. Mc Elduff et al. found their results in favour of the 50 challenge test, whereas Mathai et al. found similar sensitivity for both tests and a higher specificity for the random test if both tests were performed in the 26th to 30th week of gestation (20,21). A number of studies compared the 50-g glucose challenge test with measurement of fasting glucose. Perucchini et al. found the results in favour of the fasting glucose measurement (22), whereas Rey et al. showed the 50 g glucose challenge test to be superior (23). Other studies investigating the test characteristics of the glucose challenge test, reported sensitivities ranging from 58-80% (24,25) for a specificity of around 65% (25). In these studies, thresholds for an abnormal result of the challenge test ranged from 7.2 to 7.8 mmol/l. In the present study, a pre-defined cut off value for an abnormal test result was set at 7.8 mmol/l. If thresholds were set lower than 7.8 mmol/l, sensitivity of the 50 g glucose challenge test would increase, at the expense of a decreased specificity.

The random glucose test is a fast, simple and relatively inexpensive test. Accuracy of random glucose measurement is less frequently studied than the glucose challenge test. Nasrat and colleagues evaluated random glucose measurement, which revealed a sensitivity of 16% and a specificity of 96% using a threshold value of 7.0 mmol/l or 6.4 mmol/l if evaluated within or more than 2 hours postprandial (26). Jowett et al. also concluded that random glucose measurement is not sufficiently sensitive for screening on gestational diabetes mellitus (27). Results from the present study are in accordance with results from those two groups, using a threshold value for an abnormal test result of 6.8 mmol/l. As high sensitivity is key to any screening test, random glucose testing is not an accurate method to screen women for gestational diabetes, as still five out of six women with gestational diabetes would be missed.

In conclusion we recommend that despite easy implementation, low costs and relative high specificity, random glucose measurement should not be used as a screening test for gestational diabetes mellitus. Until superior screening alternatives become available, the 50 g glucose challenge test should be preferred as screening test for gestational diabetes mellitus.

ACKNOWLEDGEMENTS

This research was supported by a grant from Novo Nordisc, Alphen aan den Rijn, The Netherlands and by grant 917. 46.346 in the VIDI-program of ZonMW, The Hague, The Netherlands. The funding sources had no involvement in the design, analysis or reporting of this study.

The abstract of this paper was accepted for oral presentation on the 4th International Symposium of Diabetes and Pregnancy, which was held in Istanbul, Turkey in March 2007.
Reference List


<table>
<thead>
<tr>
<th></th>
<th>GDM present</th>
<th>GDM not present</th>
<th>GDM not verified</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>46</td>
<td>276</td>
<td>979</td>
<td>1301</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td>30.8 ± 4.6</td>
<td>30.6 ± 4.9</td>
<td>30.8 ± 5.0</td>
<td>30.8 ± 4.9</td>
</tr>
<tr>
<td><strong>BMI before pregnancy (kg/m²)</strong></td>
<td>25.6 ± 4.4</td>
<td>25.7 ± 5.2</td>
<td>23.8 ± 4.4</td>
<td>24.2 ± 4.6</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>37 (82.2)</td>
<td>247 (90.5)</td>
<td>848 (89.4)</td>
<td>1132 (89.4)</td>
</tr>
<tr>
<td>Non Caucasian</td>
<td>8 (17.8)</td>
<td>26 (9.5)</td>
<td>101 (10.6)</td>
<td>135 (10.6)</td>
</tr>
<tr>
<td><strong>Family history of diabetes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>13 (28.9)</td>
<td>55 (20.1)</td>
<td>185 (19.5)</td>
<td>253 (19.7)</td>
</tr>
<tr>
<td>No</td>
<td>32 (71.1)</td>
<td>218 (79.9)</td>
<td>783 (80.5)</td>
<td>1033 (80.3)</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8 (17.4)</td>
<td>46 (17.0)</td>
<td>170 (17.5)</td>
<td>224 (17.4)</td>
</tr>
<tr>
<td>No</td>
<td>38 (82.6)</td>
<td>225 (83.0)</td>
<td>799 (82.5)</td>
<td>1062 (82.6)</td>
</tr>
<tr>
<td><strong>Hospital</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utrecht</td>
<td>22 (47.8)</td>
<td>99 (35.9)</td>
<td>874 (89.3)</td>
<td>995 (76.5)</td>
</tr>
<tr>
<td>Zwolle</td>
<td>24 (52.2)</td>
<td>177 (64.1)</td>
<td>105 (10.7)</td>
<td>306 (23.5)</td>
</tr>
<tr>
<td><strong>Obstetric history 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Previous spontaneous abortion</td>
<td>15 (32.6)</td>
<td>84 (30.4)</td>
<td>317 (32.4)</td>
<td>416 (32.0)</td>
</tr>
<tr>
<td>No previous spontaneous abortion</td>
<td>31 (67.4)</td>
<td>192 (69.6)</td>
<td>662 (67.6)</td>
<td>885 (68.0)</td>
</tr>
<tr>
<td><strong>Obstetric history 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nullipara</td>
<td>19 (43.2)</td>
<td>112 (40.9)</td>
<td>422 (44.3)</td>
<td>553 (43.5)</td>
</tr>
<tr>
<td>Multipara with history of GDM</td>
<td>2 (4.5)</td>
<td>8 (2.9)</td>
<td>6 (0.6)</td>
<td>16 (1.3)</td>
</tr>
<tr>
<td>Multipara without history of GDM</td>
<td>23 (52.3)</td>
<td>154 (56.2)</td>
<td>525 (55.1)</td>
<td>702 (55.2)</td>
</tr>
<tr>
<td><strong>Obstetric history 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nullipara</td>
<td>19 (43.2)</td>
<td>112 (40.7)</td>
<td>422 (44.3)</td>
<td>553 (43.5)</td>
</tr>
<tr>
<td>Multipara with perinatal mortality</td>
<td>4 (9.1)</td>
<td>17 (6.2)</td>
<td>47 (4.9)</td>
<td>68 (5.3)</td>
</tr>
<tr>
<td>Multipara without perinatal mortality</td>
<td>21 (47.7)</td>
<td>146 (53.1)</td>
<td>484 (50.8)</td>
<td>651 (51.2)</td>
</tr>
</tbody>
</table>

Data are means ± SD or n (%).
Table 2: Results of the 2x2 table and accuracy measures calculated after correction for verification bias.

<table>
<thead>
<tr>
<th></th>
<th>Random Glucose test</th>
<th>1 hour 50 g Glucose Challenge test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OGTT abnormal</td>
<td>OGTT normal</td>
</tr>
<tr>
<td>Screening test abnormal</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>屏考</td>
<td>33</td>
</tr>
<tr>
<td>Screening test normal</td>
<td>41</td>
<td>1223</td>
</tr>
<tr>
<td></td>
<td>屏考</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>48</td>
<td>1253</td>
</tr>
<tr>
<td></td>
<td>屏考</td>
<td>47</td>
</tr>
</tbody>
</table>

Sensitivity (%) | 14.6 (4.6-24.6) | 70.2 (57.1-83.3) |
Specificity (%) | 97.6 (96.6-98.5) | 89.1 (87.4-90.9) |
Positive predictive value (%) | 18.9 (6.3-31.5) | 19.8 (3.7-25.8) |
Negative predictive value (%) | 96.8 (91.0-100.0) | 98.7 (97.1-100.0) |
LR for an abnormal test result | 6.1 (2.8-13.2) | 6.5 (5.1-8.3) |
LR for a normal test result | 0.88 (0.78-0.98) | 0.33 (0.22-0.52) |
Diagnostic OR | 7.0 (2.9-16.8) | 19.4 (6.8-31.9) |
AUC | 0.69 (0.61-0.78) | 0.88 (0.83-0.93) |

All accuracy measures are displayed with 95% CIs. LR = Likelihood Ratio, OR = Odds Ratio, AUC = Area under the Curve.
Figure 1. Screening and diagnostic test results before and after correction for verification bias. The figures in the diagram represent the number of women with the specific combination of test results before (after) correction of verification bias. Figures between parentheses represent the number of women after correction for verification bias.