Screening for gestational diabetes mellitus

van Leeuwen, M.

Citation for published version (APA):
van Leeuwen, M. (2012). Screening for gestational diabetes mellitus
Chapter 6

External validation of a clinical scoring system for the risk of gestational diabetes mellitus

Marsha van Leeuwen
Brent C. Opmeer
Egbert J.K. Zweers*
Evert van Ballegooie*
Henk G. ter Brugge
Harold W. de Valk
Gerard H.A. Visser
Ben Willem J. Mol

Diabetes Research and Clinical Practice. 2009; 85(1):96-101
Abstract

AIM A prediction rule for gestational diabetes mellitus (GDM) could be helpful in early detection and increased efficiency of screening. A prediction rule by means of a clinical scoring system is available, but has never been validated externally. The aim of this study was to validate the scoring system.

METHODS We used data from a prospective cohort study. Women were assigned a score based on age, Body Mass Index (BMI) and ethnicity. Performance of the scoring system was evaluated in terms of discrimination and calibration (agreement between clinical score and observed probability of GDM). We compared the efficiency of a screening strategy derived from the scoring system with conventional screening.

RESULTS We studied 1266 women. Forty-seven women had GDM (3.7%). The scoring system discriminated moderately (area under the curve = 0.64 (95% CI 0.56 - 0.72)). Calibration was limited ($\chi^2 = 8.89$, $p = 0.06$). The screening strategy derived from the scoring system reduced the number of women needed to be screened with 25% for a comparable detection rate to universal screening.

CONCLUSION Despite moderate discriminative capacity and calibration of the scoring system, the screening strategy based on the scoring system appears clinically useful. There is need for better prediction models for GDM.
Introduction

In pregnancies complicated by gestational diabetes mellitus (GDM), there is an increased rate of fetal as well as maternal complications during pregnancy and delivery. Early diagnosis and subsequent treatment of GDM could prevent these complications and therefore improve pregnancy outcome.\(^1\) GDM often is an asymptomatic condition. The optimal strategy to identify women with GDM is unknown. Some expert groups advocate the use of clinical factors to identify women at risk for GDM.\(^2\) If clinical factors can be used to estimate the probability of GDM accurately, discrimination between women at high risk and women at low risk can be made early in pregnancy or even before conception. Additional screening procedures could then be limited to women at increased risk for GDM. This would reduce the burden of screening on low risk women, whereas women at high risk could be monitored closely and treated in a timely fashion if necessary.

The probability that disease will occur can be estimated with a prognostic model or clinical scoring system. In 1997, Naylor et al. published a scoring system for the risk classification of GDM that was based on multivariable logistic regression analysis of clinical characteristics.\(^3\) Based on the variables age, Body Mass Index (BMI) and ethnicity a score was calculated for all women in the study. According to this score, women were classified as being at low, intermediate or high risk for GDM, and a screening strategy was developed based on this classification. Internal validation of the scoring system showed that the score successfully differentiated between women at high risk and women at low risk for GDM. Naylor et al. used the 100-g oral glucose tolerance test (OGTT) to establish the diagnosis of GDM. In current clinical practice however, GDM is often diagnosed using the 75-g OGTT instead of the 100-g OGTT. At present it is unknown if the scoring system by Naylor et al. is valid when using the 75 g-OGTT to diagnose GDM. The aim of this study was to validate the scoring system for GDM in an external population using the 75-g OGTT as a test for diagnosis of GDM. Moreover, we evaluated the efficiency of a screening strategy that was based on the scoring system.

Material and Methods

We used data from a previously published prospective cohort study, which compared the performance of two screening tests for GDM.\(^4\) In this study, consecutive women with a singleton pregnancy who reported for prenatal care before 24 weeks of gestation in two hospitals in the Netherlands (the Isala Clinics in Zwolle and the University Medical Centre in Utrecht) were invited to participate. A random glucose measurement was performed in all women at intake (around 12 weeks of gestation), to detect women with undiagnosed type-1 or type-2 diabetes before pregnancy. Women with a diagnosis of pre-existing
diabetes mellitus were excluded. Medical history and characteristics of women who gave informed consent were recorded at intake. At 24 to 28 weeks of gestation all women underwent a 50-g glucose challenge as well as random glucose measurement to screen for GDM. The predefined cutoff value for an abnormal test result for the 50-g glucose challenge test was a 1-h plasma glucose value of 7.8 mmol/l. The random glucose test was considered as abnormal if the plasma glucose value ≥ 6.8 mmol/l. Women with an abnormal result on either or both screening tests underwent the 75-g oral glucose tolerance test (OGTT, reference test) within one week of the screening tests to confirm or rule out GDM. GDM was diagnosed if the fasting value of the OGTT was > 7.0 mmol/L or the two hour value of the OGTT was ≥ 7.8 mmol/L, according to the criteria of the World Health Organization (WHO). If all women with two negative screening tests would automatically be considered GDM negative, incidental cases of GDM would remain undetected, generating verification bias. To correct for verification bias the OGTT was performed in a subset of women with two negative screening test results, to estimate the fraction of diseased women that remained undetected by the screening tests (false-negative fraction). Subsequently, missing OGTT values due to the selective verification were imputed with a multiple imputation procedure. Details of this procedure have been reported in a previously published study. At the same time, other incidental missing data on continuous variables were also imputed using this multiple imputation procedure.

Table 1 shows the clinical scoring system developed by Naylor et al. The scoring system is based on converted odds ratios derived by multivariable logistic regression analysis and includes three clinical variables: age, BMI and ethnicity. Based on these variables, women are assigned a clinical risk score, with a maximum score of 10 points (Table 1). Subsequently, women are classified into three categories. Women with a clinical risk score of 0 or 1 are categorized as low risk, women with a clinical risk score of 2 or 3 are categorized as intermediate risk and women with score higher than 3 are categorized as high risk for GDM. The screening strategy based on this clinical scoring system is as follows: Low risk women are not screened. Intermediate risk women are screened with the 50-g glucose challenge test with a threshold of 7.8 mmol/L. High risk women are also screened with the 50-g glucose challenge test, however a lower threshold is set for test positivity (7.1 mmol/L).

We used the clinical scoring system to calculate individual clinical risk scores for all women in our cohort and categorized women as being low, intermediate or high risk according to the definition by Naylor et al. We compared the overall prevalence of GDM between our sample and the sample of Naylor et al. and evaluated the distribution of women over the clinical risk scores and risk categories. We assessed the validity of the
clinical scoring system by means of discrimination and calibration. Finally we evaluated the accuracy of the screening strategy based on the scoring system. Discrimination was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC). In the ROC plot the false-negative fraction was plotted against the true-positive fraction for all possible threshold values of the clinical risk score. The area under the ROC curve evaluated the ability of the clinical scoring system to distinguish women with GDM from women without GDM. The larger the AUC of the ROC curve, the better the discriminative capacity of the scoring system. An AUC of 0.5 indicates that the scoring system does no better than chance in estimating the outcome, whereas an AUC of 1.0 reflects perfect discriminative capacity. Calibration was evaluated with the \( \chi^2 \) goodness of fit test to assess the level of correspondence between predicted probabilities and the observed percentage of women with GDM per clinical risk score. If the observed percentages of GDM are close to the predicted probabilities, the scoring system is considered to be well calibrated. We used the observed percentages of women with GDM in the original patient sample of Naylor et al. as the predicted probabilities of GDM. The goodness of fit test therefore reflected whether the prevalence of GDM across risk scores in our sample was statistically different from the prevalence reported by Naylor et al.

We evaluated the accuracy of the proposed screening strategy in terms of number of women needed to screen to establish a diagnosis of GDM, detection rate and false-positive rate, and compared these figures to those of universal screening with the 50-g glucose challenge test in our own sample using McNemar’s test to test for agreement in screening accuracy. We also compared the accuracy measures to the figures that Naylor et al. found in their sample. In their original paper, Naylor et al. described two samples of women. The first sample was used to develop the scoring system and the subsequent screening strategies. We will refer to this group of women as Naylor’s derivation sample.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Odds Ratio (95% CI)</th>
<th>P-value</th>
<th>Score according to Naylor et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (£ 30 yr)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 – 34 yr</td>
<td>1.0 (0.7 - 1.5)</td>
<td>0.95</td>
<td>1</td>
</tr>
<tr>
<td>≥ 35 yr</td>
<td>1.6 (1.1 - 2.5)</td>
<td>0.02</td>
<td>2</td>
</tr>
<tr>
<td>BMI (£ 22.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22.1 – 25.0</td>
<td>1.8 (1.1 - 2.7)</td>
<td>0.01</td>
<td>2</td>
</tr>
<tr>
<td>≥ 25.1</td>
<td>3.2 (2.1 - 4.8)</td>
<td>&lt;0.001</td>
<td>3</td>
</tr>
<tr>
<td>Ethnicity (Caucasian)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.7 (0.3 - 1.7)</td>
<td>0.44</td>
<td>0</td>
</tr>
<tr>
<td>Asian</td>
<td>4.8 (3.0 - 7.6)</td>
<td>&lt;0.001</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>1.6 (0.7 - 3.5)</td>
<td>0.24</td>
<td>2</td>
</tr>
</tbody>
</table>
The other sample was used to internally validate the scoring system and the developed screening strategy. We will refer to this sample as Naylor’s validation sample. We used Naylor’s derivation sample for comparison of our findings to the findings of Naylor et al. and only if data on Naylor’s derivation sample were not available, we used Naylor’s validation sample for comparison of results instead. Statistical analyses were performed using SPSS version 14.0.2 (SPSS, Chicago, IL) and SAS 9.1.3. No approval of the institutional research board was required for this validation study.

Results

There were 1301 women included in the original cohort study. Information concerning ethnicity was unavailable for 35 women and therefore these women were excluded from the analysis. Of 1266 women eligible for analysis, all women underwent the random glucose test and 1246 women underwent the 50-g glucose test (98.4%). 184 women had at least one abnormal screening test of whom 146 (80%) underwent an OGTT. 38 women did not agree to undergo an OGTT despite at least one abnormal result of the screening tests. To estimate the fraction of false negative screening test results, women with negative screening test results were asked at random to undergo the OGTT. 176 women consented and underwent an OGTT. The false-negative fraction was 7.9%. In total, the OGTT was performed in 322 women (25.4%), of whom 46 women had an abnormal OGTT result. After the multiple imputation procedure to correct for verification bias the number of women diagnosed with GDM was supposed to be 47, indicating an incidence of GDM in our sample of 3.7%. Next to the missing OGTT results, 7.3% of the

<table>
<thead>
<tr>
<th>Table 2. Baseline characteristics of our sample with regard to the variables of the clinical scoring system.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDM present</td>
</tr>
<tr>
<td>Age (years)</td>
</tr>
<tr>
<td>≤ 30</td>
</tr>
<tr>
<td>31 - 34</td>
</tr>
<tr>
<td>≥ 35</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
</tr>
<tr>
<td>≤ 22.0</td>
</tr>
<tr>
<td>22.1 - 25.0</td>
</tr>
<tr>
<td>≥ 25.1</td>
</tr>
<tr>
<td>Ethnicity</td>
</tr>
<tr>
<td>Caucasian</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Asian</td>
</tr>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>

Data are n (%).
data on age and BMI was imputed using the multiple imputation procedure, as well as the missing results of the 50-g glucose challenge test (1.6%). The baseline characteristics of our sample concerning the variables of the clinical scoring system are displayed in Table 2.

The overall prevalence of GDM in our sample of 3.7% was not significantly different from the prevalence of GDM in Naylor’s derivation group (2.8%, \( p = 0.18 \)). Distribution of women over the risk scores and the observed percentages of GDM are shown in Figure 1. The distribution of women in our sample across the risk scores was significantly different from the distribution of women in Naylor’s validation sample \( (p < 0.001) \). In Naylor’s validation sample more women were in the low as well as in the high risk scores compared to our sample. The distribution of women across the predefined risk categories with their observed percentages of GDM are shown in Table 3. The distribution of women across the three risk categories in our population was significantly different from the distribution in the validation sample of Naylor et al. \( (p < 0.001) \).

The ROC curve of the clinical scoring system in our population had an AUC of 0.64 (95% CI 0.56 - 0.72), indicating a moderate discriminative capacity. The AUC in our sample was not significantly different from the AUC in the initial study of Naylor et al. (derivation group), which was 0.68. Since there were no results reported on the prevalence of GDM for the various clinical risk scores in Naylor’s derivation sample, we compared the prevalence of GDM in our sample with the prevalence of GDM in the internal validation sample of Naylor et al. for the various clinical risk scores to assess the calibration of the model. Figure 1 shows the correspondence between the calculated risk score and the observed probability of GDM for our sample and for Naylor’s validation sample.

![Observed rate of GDM per risk score](image)

**Figure 1.** Correspondence between the risk score and the probability of GDM for Naylor’s validation sample and our validation sample.
The prevalence of GDM increased with an increasing risk score in both samples. For most of the clinical risk scores, the prevalence of GDM in our sample was lower than the prevalence of GDM in Naylor’s validation sample. The \( \chi^2 \) goodness of fit test indicated a poor fit for the clinical scoring system in our sample (\( \chi^2 = 8.89, \) d.f. = 4, \( p = 0.06 \)).

After classifying women according to the predefined risk groups, the prevalence of GDM was 1.9%, 3.4% and 5.8% in the low, intermediate and high risk group respectively.

The performance of the screening strategy based on the clinical scoring system in our cohort is displayed in Table 4. If all women would be screened with the 50-g glucose challenge test (universal screening) the detection rate of GDM in our sample would be 68% (95% CI 55 - 80), with a corresponding false-positive rate of 10.8% (95% CI 10.4 - 11.4). If we would apply the screening strategy suggested by Naylor et al. to our sample, screening with the 50-g glucose challenge test could be omitted in 25% of the women. The detection rate in our cohort would consequently decrease to 64% (95% CI 50 - 76) with a corresponding false-positive rate of 12.6% (95% CI 12.1 - 13.1). Compared to universal screening, the decrease in detection rate was not statistically significant (\( p = 0.48 \)). The false-positive rate of the screening strategy however, was significantly higher compared to universal screening (\( p < 0.001 \)). In Naylor’s derivation sample 35% of the women could refrain from screening by using the clinical scoring system. The detection

### Table 3. Distribution of women across the predefined risk categories

<table>
<thead>
<tr>
<th>Risk category</th>
<th>Low</th>
<th>Intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of women</td>
<td>311</td>
<td>593</td>
<td>362</td>
</tr>
<tr>
<td>No. of women with GDM (%)</td>
<td>6 (1.9)</td>
<td>20 (3.4)</td>
<td>21 (5.8)</td>
</tr>
<tr>
<td><strong>Naylor’s validation sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of women</td>
<td>544</td>
<td>606</td>
<td>421</td>
</tr>
<tr>
<td>No. of women with GDM (%)</td>
<td>5 (0.9)</td>
<td>23 (3.5)</td>
<td>41 (9.7)</td>
</tr>
</tbody>
</table>

### Table 4. Detection rate and false positive test results with universal screening and with the selective screening strategy in our sample. The screening test used was the 50-g glucose challenge test.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>No to be screened (%)</th>
<th>Detection Rate (n)</th>
<th>P-value (^{b})</th>
<th>False-positive rate (n)</th>
<th>P-value (^{b})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual Care (^{d})</td>
<td>1266</td>
<td>68.1% (32)</td>
<td></td>
<td>10.8% (132)</td>
<td></td>
</tr>
<tr>
<td>Selective Screening (^{a})</td>
<td>955 (75.4%)</td>
<td>63.8% (30)</td>
<td>0.48</td>
<td>12.6% (153)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\(^{a}\) values are based on a total of 47 true-positive and false-negative test results. \(^{b}\) P-value for the comparison of selective screening with universal screening. \(^{c}\) values are based on a total of 1219 false-positive and true-negative test results. \(^{d}\) test 100% of the women with the 50-g glucose challenge test (threshold 7.8 mmol/L) \(^{e}\) no screening if score 0-1, screening with the 50-g glucose challenge test (threshold 7.8 mmol/L) if score 2-3, screening with the 50-g glucose challenge test (threshold 7.1 mmol/L) if score >3
rate as well as the false-positive rate of the screening strategy in Naylor’s derivation sample were significantly higher compared to our sample (detection rate 72.7%, $p = 0.84$; false positive rate 16.7%, $p = 0.012$).

Discussion

In this study we performed external validation of a clinical scoring system developed to perform selective screening for GDM. External validation is an essential step in the evaluation of a model or scoring system before it can be implemented in clinical practice,$^6$ as it assesses the performance of a developed model or scoring system in a different sample or in different circumstances than in which it was originally developed. In the present study we wanted to assess the validity of the scoring system when using the 75-g OGTT for the diagnosis of GDM instead of the 100-g OGTT. The validity of the scoring system in our sample could be considered as being unsatisfactory. The AUC of the scoring system in our sample was low, although comparable to the AUC of the scoring system in the original sample, and calibration indicated a poor fit for the individual risk score as well as for the predefined risk categories.

Naylor et al. developed their clinical scoring system by using converted odds ratios derived by multivariable logistic regression analysis. The odds ratios of all statistically significant variables were rounded to the nearest integer and added to develop the clinical risk score. No points were assigned for the reference categories of the variables. Criticism has been vented on this method of developing a clinical risk score. An odds ratio of 1.0 for all reference categories should be included in the scoring system, and when odds ratios are translated into a clinical scoring system, figures should be multiplied instead of added.$^7$ According to Naylor et al. they are aware of these flaws in their statistical framework; however they feel that the clinical scoring rule as they developed it does not lead to inferior performance of the scoring rule, whereas if official statistical procedures are followed, it would make the clinical scoring rule less understandable to practicing clinicians and unnecessarily cumbersome to use.$^8$

A possible limitation of the present study is the limited number of women in the original sample in whom a reference test (OGTT) was performed. Since not all women underwent an OGTT to determine definitive GDM status, verification bias occurred. We corrected for this verification bias by means of multiple imputation. Imputing missing data is considered an eligible method to correct for verification bias and is preferred over complete case analysis, because in complete case analysis, deleting cases with missing values leads to a loss of statistical power and biased results.$^9$ Although it would have been preferable to perform an OGTT in all women, unfortunately this was not feasible in the original study,
due to the inconvenience this would have generated for the majority of the women. The poor calibration could be the result of poor fit of the scoring system to our sample, due to for example differences in sample characteristics or patient recruitment. Naylor et al. identified age as one of the risk factors for GDM. In our cohort there was no statistically significant association between age and GDM (p = 0.14). The association between older age and an increasing risk of GDM has been described in a number of studies. A study by Coustan et al. showed an increasing incidence of GDM with increasing age. In this study 56% of the women with GDM were younger than 30 years, which is consistent with the proportion in our sample (55.3%). In the present study we did not find an association between age and the risk of GDM when using the age categories defined by Naylor et al. When using age as a continuous variable, we still did not find a significant association.

Another reason for an unsatisfactory calibration could be differences in test protocol or analysis. Because we wanted to evaluate if the clinical scoring system was valid when using the 75-g OGTT as a diagnostic test, we used the 2 hour 75-g OGTT to diagnose GDM instead of the 3 hour 100-g OGTT. Since the original model was developed using the 3 hour 100-g OGTT as a diagnostic test, validation with the 2 hour 75-g alternative might have lead to inadequate estimation of the performance of the clinical scoring system in our sample. Application of a different reference test to validate the clinical scoring system might conceal the true origin of the poor calibration. In the present study we could not differentiate between poor fit due to application of a different reference test, or poor fit due to differences in sample characteristics. We have found however, that the scoring system has an unsatisfactory fit to our population when using the 75-g OGTT as a reference test. Santos-Ayarzagoitia described higher accuracy of the 100-g OGTT. The results of the HAPO study however show that there is an association between the results of the 75-g OGTT and a number of important perinatal complications. Since the 75-g OGTT nowadays is a frequently applied diagnostic test for GDM, in clinical practice as well as in a number of important studies, it is important that a clinical scoring system to perform selective screening is also applicable when using the 75-g OGTT.

Another explanation for the relatively poor performance of the scoring system is the slightly different gestational age at which the 50-g glucose challenge test was performed. In our sample the 50-g glucose challenge test was performed between 24 and 28 weeks of gestation, whereas in the study by Naylor et al. the 50-g glucose challenge test was performed at 25 to 27 weeks. In GDM insulin sensitivity decreases progressively with gestational age, leading to rising glucose values as pregnancy progresses. In line with this change in glucose tolerance, it is possible that women in our sample would have had different results on the 50-g glucose test, if the test was performed from 25 weeks to 27 weeks instead of 24 weeks to 28 weeks, leading to inaccurate estimates of the
performance of the scoring system. This would especially be the case if many women in our dataset would have a 50-g glucose challenge test result that is close to the predefined cutoff value of the 50-g glucose challenge test of 7.8 mmol/L. Since there were 73 women who had a 50 g glucose challenge test result ranging from 7.3 mmol/L to 7.8 mmol/L, of which 6 women had GDM. This false-negative rate might have been lower if women would have undergone the 50-g glucose challenge test later in pregnancy.

Although the performance of the scoring system was relatively poor in our sample, the performance of the screening strategy that was developed based on the scoring system was satisfactory. The detection rate of the selective screening strategy was comparable to the detection rate of universal screening in our cohort. By using the proposed screening strategies, screening with the 50-g glucose challenge test can be omitted in nearly 25% of the women in our sample, though with a higher false-positive rate and therefore at the cost of an increased rate of unnecessary performed diagnostic OGTTs of 1.8% (false-positive rate).

Some international expert groups recommend reduction of the upper limit of normal of the fasting venous plasma glucose of the OGTT from 7.0 mmol/L to 6.0 mmol/L, as this is considered to be more representative of the physiological changes in pregnancy. Results from a large cohort study (HAPO study) show that there is a relation between the fasting value of the OGTT and the risk of a number of perinatal and maternal outcomes and complications. Next to being more representative of the physiological changes in pregnancy, lowering the upper limit of normal fasting glucose values could also result in detecting more women at risk for complications. If these women are treated in a timely fashion, pregnancy outcome might be improved.

In conclusion, this study demonstrates that the discriminative capacity of the clinical scoring system developed by Naylor et al. in our sample was relatively poor and that the clinical scoring system estimated the risk of GDM only moderately using the 75-g OGTT as a diagnostic test for GDM. Performance of the screening strategy based on the scoring system however was still adequate, resulting in a reduction of rate of women needed to be screened of 25%, with a detection rate comparable to universal screening. Possibilities of another prediction model or a clinical scoring system for our population are worthwhile to explore since risk estimation of GDM was not optimal in our sample. A new prediction model or scoring system with additional or different prognostic factors or covariates could estimate the risk of GDM in our population more accurately, possibly improving the process of selective screening for GDM even further, leading to better patient care as well as to cost-effective management.
References


