Title: Early prediction of high risk gestational diabetes mellitus via machine learning models.

Yan-Ting Wu^{1,2,3,†}, Chen-Jie Zhang^{1,2,3,†}, Ben Willem Mol⁴, Cheng Li^{1,2,3}, Lei Chen¹, Yu Wang^{1,2,3}, Jian-Zhong Sheng⁵, Jian-Xia Fan^{1,2,3}, Yi Shi^{6*}, He-Feng Huang^{1,2,3*}

¹ International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 20030, China

² Shanghai Key Laboratory of Embryo Original Diseases, Shanghai 20030, China

³ Institute of Embryo-Fetal Original Adult Diseases Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China

⁴ Department of Obstetrics and Gynecology, Monash University, Clayton 3800, Australia

⁵ Department of Pathology and Pathophysiology, School of Medicine, Zhejiang University, Zhejiang, China

⁶ Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai 20030, China

† Yan-Ting Wu and Chen-Jie Zhang contributed equally to this work.

* Co-corresponding author: He-Feng Huang, International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, No.910, Hengshan Road, Shanghai, 200030, China

Tel: 86-21-64070434

Fax: 86-21-64474645

Email: huanghefg@sjtu.edu.cn

Yi Shi, Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China

Email: yishi@sjtu.edu.cn

Abbreviations

ADA: American Diabetes Association

AUC: Area under the curve

BMI: Body mass index

DCA: Decision curve analysis

DNN: Deep Neural Network

FPG: Fasting plasma glucose

GDM: Gestational diabetes mellitus

HbA_{1c}: Glycosylated hemoglobin A1C

HDL: High density lipoprotein

HL: Hosmer-Lemeshow test

IADPSG: International Association of Diabetes and Pregnancy Study Groups

KNN: K-Nearest Neighboring

LR: Logistic Regression

SVM: Support Vector Machine

TG: Triglyceride

Abstract

Aims: Gestational diabetes mellitus (GDM) is a pregnancy-specific disorder that can usually be diagnosed after 24 gestational weeks. So far, there is no accurate method to predict GDM in early pregnancy.

Methods: We collected data extracted from the hospital's electronic medical record system included 73 features in the first trimester. We also recorded the occurrence of GDM, diagnosed at 24-28 weeks of pregnancy. We conducted a feature selection method to select a panel of most discriminative features. We then developed advanced machine learning models, using Deep Neural Network (DNN), Support Vector Machine (SVM), K-Nearest Neighboring (KNN), and Logistic Regression (LR), based on these features.

Results: We studied 16,819 women (2,696 GDM) and 14,992 women (1,837 GDM) for the training and validation group. DNN, SVM, KNN, and LR models based on the 73-feature set demonstrated the best discriminative power with corresponding area under the curve (AUC) values of 0.92 (95%CI 0.91, 0.93), 0.82 (95%CI 0.81, 0.83), 0.63 (95%CI 0.62, 0.64), and 0.85 (95%CI 0.84, 0.85), respectively. The 7-feature (selected from the 73-feature set) DNN, SVM, KNN, and LR models had the best discriminative power with corresponding AUCs of 0.84 (95%CI 0.83, 0.84), 0.69 (95%CI 0.68, 0.70), 0.68 (95%CI 0.67, 0.69), and 0.84 (95%CI 0.83, 0.85), respectively. The 7-feature LR model had the best Hosmer-Lemeshow test outcome. Notably, the AUCs of the existing prediction models did not exceed 0.75.

Conclusions: Our feature selection and machine learning models showed superior predictive power in early GDM detection than previous methods; these improved models will better serve clinical practices in preventing GDM.

Keywords

GDM, prediction, machine learning models, early pregnancy

Research in Context section

Evidence before this study

1. A hysteretic diagnosis of GDM in the 3rd trimester is too late to prevent exposure of the embryos or fetuses to an intrauterine hyperglycemia environment during early pregnancy.
2. Prediction models for gestational diabetes are not uncommon in previous literature reports, but laboratory indicators are rarely involved in predictive indicators.
3. The penetration of AI into the medical field makes us want to introduce it into GDM predictive models.

What is the key question?

Whether the GDM prediction model established by machine learning has the ability to surpass the traditional LR model?

Added value of this study

1. Using machine learning to select features is an effective method.
2. DNN prediction model have effective discrimination power for predicting GDM in early pregnancy, but it cannot completely replace LR. KNN and SVM are even worse than LR in this study.

Implications of all the available evidence

The biggest significance of our research is not only to build a prediction model that surpasses previous ones, but also to demonstrate the advantages and disadvantages of different machine learning methods through a practical case.

Introduction

Gestational diabetes mellitus (GDM) is a common complication during pregnancy, which is very different from pure overt diabetes [1], and it affects up to 15% of pregnancies in the world [2]. Hyperglycemia is not, by itself, life-threatening for pregnant women, but it can be harmful to the fetus. Fetal symptoms include polyhydramnios, stillbirth, premature delivery, macrosomia, fetal hyperinsulinemia, and clinical neonatal hypoglycemia. The American Diabetes Association (ADA) and International Association of Diabetes and Pregnancy Study Groups (IADPSG), recommends diagnosing GDM via 2h, 75g oral glucose tolerance test (OGTT) at 24-28 weeks of pregnancy [3, 4]. However, a hysteretic diagnosis of GDM in the 3rd trimester is too late to prevent exposure of the embryos or fetuses to an intrauterine hyperglycemia environment during early pregnancy. There is accumulating evidence indicating that exposure of embryos or fetuses to a hyperglycemic environment in the uterus can lead to chronic health problems later in life [5], including obesity, diabetes, and cardiovascular diseases [6-8]. Our previous studies indicate that insulin therapy after GDM diagnosis cannot fully protect offspring from diet-associated metabolic disorders in adulthood [9]. Therefore, it is essential to establish a prediction model to forecast the possibility of GDM during early pregnancy and to provide lifestyle interventions prior to diagnosis in the 3rd trimester.

Previous research has sought to find a threshold value of fasting plasma glucose (FPG) in the first trimester through large sample studies. While elevating diagnostic criteria from FPG ≥ 5.1 mmol/l to FPG ≥ 6.1 mmol/l can obtain nearly 100% specificity, the relatively low sensitivity (1%) greatly limits the feasibility [10]. In recent years, some novel biomarkers have been reported as potential prediction parameters of GDM, including angiopoietin-like protein 8, plasma fatty acid-binding protein 4 and various adipokines [11-13]. However, these biomarkers cannot be easily obtained by hospitals undermining the prediction accuracies. Exploration of prediction models based on multiple common risk factors, such as advanced maternal age, body mass index (BMI), and family history of diabetes, provides a new perspective on solving the problem [14]. There are still only a few studies, however,

that combine demographic characteristics and laboratory testing into a prediction model. Recently, AI technology, especially supervised machine learning methods, were reported to demonstrate a powerful self-learning ability and improve accuracies in diabetic retinopathy and pediatric disease prediction/diagnosis [15]. In order to explore whether combining AI with prior clinical knowledge will boost the early prediction and diagnosis of diseases such as GDM, we developed feature selection-based machine learning models to predict the occurrence of GDM based on clinical features in the 1st trimester.

Methods

Data source

We collected the 2017 obstetrical electronic medical record data from the International Peace Maternal and Child Health Hospital (IPH), Shanghai Jiao Tong University School of Medicine, which served as the training group. Women with pregestational diabetes mellitus and women who terminated their pregnancy before the 75g OGTT test (24-28 weeks of pregnancy) were excluded. To improve data quality, we further filtered samples and features/biomarkers whose missing value rate was >20%. Candidate features containing sociodemographic characteristics, clinical features, and laboratory indexes in the first trimester were collected. We then collected and curated the 2018 obstetrical electronic medical record data, which served as the validation group to evaluate the prediction models. The GDM diagnostic criteria follow the IADPSG guidelines, i.e., FPG ≥ 5.1 mmol/L or 1 h ≥ 10 mmol/L or 2 h ≥ 8.5 mmol/L should be diagnosed as GDM. This study was approved by the Medical Ethical Committee of International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University (No. GKLW2019-05).

Feature selection by AI

To improve the downstream prediction power and provide a more cost-effective approach for clinical practice, the feature selection procedure was conducted to select a panel of biomarkers with the most discriminative power in distinguishing the GDM group from the control group. Specifically, we found that many features are highly correlated to each other, forming small clusters, as shown in Fig 1b, c, and Fig 2b, c; this means that a representative small cluster of features may provide sufficient discriminative power to enable early intervention. Including all redundant features from a cluster may undermine prediction effectiveness and efficiency. We followed the strategy that both our group and Xiao *et al.* previously demonstrated effective in gene selection [16, 17]. In detail, the feature selection is embedded in a cross-validation framework where both n-fold and leave-one-out cross-validation schemes are adopted. To test the performance of the n-fold cross-validation, given a feature selection method, the dataset is randomly partitioned into n equal parts where n-1 parts and the last part are treated as training dataset and testing dataset, respectively. The procedure is repeated n times so that each portion is evaluated as the testing dataset. The average classification accuracy over the whole process is considered to be the n-fold cross-validation classification accuracy of the feature selection method combined with the classifier. In this work, we chose n=10 after testing n=5, 10, 15, and 20, which all yielded similar outcomes. The random partition process was repeated 100 times. In total, the n-fold cross-validation classification accuracy is the average over a total of 1000 tests. For the leave-one-out cross-validation, each single data point (one sample) is held as a testing data point while the remaining data points are treated as the training dataset; the process is repeated for each data point and the classification accuracy is the average over the whole process. We adopted two classifiers in our study, the RBF kernel Support Vector Machine (SVM) classifier [18, 19] and the K-Nearest Neighbor (KNN) classifier [20].

All the features are initially sorted according to their absolute correlation coefficient (Pearson and Spearman) values with respect to the GDM/Control vector. The top-ranked feature is then included into the feature pool as the seed feature, and its discriminative power is determined by calculating the averaged prediction accuracy via the cross-validation process. Then, the second ranked feature is included into the feature pool and again a new prediction accuracy is calculated; if the new accuracy improves, the second ranked feature is included into the feature pool; otherwise, the third ranked feature will be tried. The process continues sequentially until all features have been considered. After this process, the features that are captured into the feature pool are treated as candidate features to be included in the final selected feature panel. Fig 1d, 1e and Fig 2d, 2e demonstrate the selected feature in each iteration. Fig 1f, 1g and Fig 2f, 2g plot the incremental trajectory of accuracies whenever including a contributing feature that stays in the selected feature pool.

Prediction methods

Based on the selected feature panel, we tested four representative prediction methods: Logistic Regression (LR) [21], KNN [20], SVM [18,19], and Deep Neural Network (DNN) [22,23]. The effectiveness of each method is again tested using cross-validation. For the DNN classifier, a sequential model with two densely connected hidden layers is devised, followed by an output layer that returns a single, continuous value. For the hidden layers, we use a rectified linear unit as the activation function and RMS_{prop} as the optimizer. Mean squared error is used as a loss function while the mean absolute error is added as an evaluation metric. All-feature vectors and selected feature vectors are imported into the model separately and then trained for 1000 epochs.

The testing set mean abs error for 7-feature sets and all-feature sets are 0.17 and 0.12.

The update rule in RMS_{prop} is detailed as follows:

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1-\beta) \frac{\delta C}{\delta w}$$

$$W_t = W_{t-1} - \frac{n}{\sqrt{E[g^2]_t}} \frac{\delta C}{\delta w}$$

Where $E(g^2)$ denotes the moving average of squared gradients and $\frac{\delta C}{\delta w}$ denotes the gradient of the cost function with respect to the weight. n is the learning rate and β is the moving average parameter.

For the LR classifier, during the training process, a sigmoid function of a given feature's linear combination is trained. For the SVM classifier, we use an RBF (Gaussian) support vector model after considering linear kernel, polynomial kernel, and RBF kernels, where the parameters were set as default as used in the LIBSVM package. For the KNN classifier, we chose $k=20$ after testing $k=1, 5, 10, 15, 20, 50, 100$, so that the k -nearest neighbors' majority voting is adopted as the prediction value.

Model evaluation

All prediction models will draw ROC curves and calculate AUC to evaluate the discrimination of models. The Hosmer-Lemeshow (HL) test is used to evaluate the calibration of each model. Finally, the decision curve analysis (DCA) is introduced to evaluate the clinical application of each model.

Results

Sample size

In our training group, a total of 2,696 GDM patients were included in the case group and 14,123 pregnant women without GDM were included in the control group. 1,837 GDM patients and 13,155 pregnant women without GDM were included in the validation group.

Features setting

The details of 73 alternative features containing sociodemographic characteristics, clinical features, and laboratory indexes in the first trimester are provided in Table 1, Table 2, and Table 3. Six features, including age, BMI, FPG, glycosylated hemoglobin A1C (HbA_{1c}), high density lipoprotein (HDL), and triglyceride (TG) were set as categorical variables apart from continuous variables.

Feature selection

In the initial step of the feature selection procedure, all the features are sorted according to their absolute Pearson correlation coefficients and Spearman correlation coefficients with respect to the GDM/Control vector, as demonstrated in Fig 1a and Fig 2b. After the model-free sequential forward feature selection procedure, 17 features were captured in the feature pool, including age, age $\dagger$, FPG, FPG $\dagger$, HbA_{1c} , $HbA_{1c}\dagger$, lipoprotein(a), apolipoprotein-A, apolipoprotein-B, TG, total triiodothyronine, total tetraiodothyronine, multiple pregnancy, multipara, smoking, family history of diabetes in a first-degree relative, previous GDM history (categorical variables denoted by $\dagger$). More details are shown in Table 4. Based on prior clinical experience and a close examination of each feature, we further narrowed the feature set into a 7-feature panel to be used in further analyses and clinical practices. Interestingly, despite obesity being a well-known risk factor for GDM, AI did not choose BMI (a clinical indicator for obesity diagnosis), instead highlighting biochemical indicators that reflect the level of lipid metabolism, such as TG. This suggests that the use of biochemical indicators of lipid metabolism to predict GDM is more accurate than anthropometric indicators.

Development of prediction models

Eight prediction models were developed: KNN, SVM, LR and DNN models were all developed for both 7-feature and all-feature sets. The LR model with 7 features is shown in Table 5.

Discrimination of different models

The AUC of different models is provided in Fig 3a and Table 6. All-feature DNN, SVM, KNN, and LR models obtained their best discriminative power with corresponding AUC of 0.92 (95%CI 0.91, 0.93), 0.82 (95%CI 0.81, 0.83), 0.63(95%CI 0.62, 0.64), and 0.85(95%CI 0.84, 0.85), respectively. The 7-feature DNN, SVM, KNN and LR models obtained their best discriminative power with corresponding AUC of 0.84(95%CI 0.83, 0.84), 0.69(95%CI 0.68, 0.70), 0.68 (95%CI 0.67, 0.69), 0.84(95% CI 0.83, 0.85). The discrimination effect of each model is shown visually through violin plots (Figs 3b and 3c). The all-feature DNN demonstrated the best discrimination ability, and the traditional LR models resulted in higher AUC than KNN and SVM models. The optimal sensitivity and specificity of each model in certain Pt value ranges are given in Table 6. We also summarize previous prediction models or diagnostic criteria (Supplemental Table S1). Existing prediction models did not obtain the level of discrimination our LR and DNN models achieved (Supplemental Fig S1). In fact, the AUC of existing prediction models did not exceed 0.75.

Calibration of different models

The HL test was used to test the calibration of the LR, SVM, and DNN models (Fig 4). KNN models could not provide risk probability for every case, so the HL test was not suited for KNN models. The P value of six different models was <0.0001 in HL test. The 7-feature models (Fig 4a, 4b, 4c) have better HL test performance than the all-feature models (Fig 4d, 4e, 4f). The 7-feature LR model revealed the most accurate calibration among all prediction models.

Clinical use

The DCA of different models is presented in Fig 5. DCA is a useful method for evaluating the benefits of a prediction model across a range of patient preferences for accepting risks of overtreatment and undertreatment. This helps facilitate decisions about prediction, test selection, and use of these models. DCA showed that the range between the "treat-all" and "treat-none" line had a net benefit. For example, DCA showed that if the Pt value was set $>4\%$, and $<96\%$, using the 7-feature LR model to predict high risk status of GDM added more benefit than either the treat-all-patients scheme or the treat-none scheme.

Discussion

With the help of our model-free feature selection procedure, 17 features were screened out from the 73-feature set. Based on clinical experience and cost concerns, 7 features were finally chosen as the final predictive panel to construct AI models. Unlike

many GDM prediction models previously used, this study added demographic data, blood biochemical indicators, and other data that can more easily be detected and collected during routine examinations. We validated previous predictive models in our validation group; the results further confirmed that our predictive model is more effective in terms of outcomes, such as AUC.

The 7-feature set includes age, personal history of GDM, family history of diabetes in a first-degree relative, fetal number, FPG, glycosylated hemoglobin A_{1c}, and triglyceride. Although the 7 features as a prediction panel has never been reported before, some of these features were individually reported as risk factors of GDM in previous researches [24, 25]. In addition to well-known risk factors, such as advanced age, previous GDM history, family history of diabetes, and blood glucose, our prediction model also included multiple pregnancy, HbA_{1c}, and blood lipid levels. Twin pregnancies have an increased risk of GDM and higher rates of adverse pregnancy outcomes occur in GDM twin pregnancies [26]. HbA_{1c} reflects blood glucose levels in the last 1-2 months, which circumvents the contingency brought by the single fasting blood glucose measurement [27, 28]. Previous studies have determined that insulin resistance is associated with obesity, the mechanisms involve dysregulated adipocytokines secreted by adipocytes that impair glucose tolerance and reduce insulin sensitivity [29, 30]. Adopting blood TG levels instead of BMI can better reflect the true obesity of pregnant women in early pregnancy, because BMI does not represent the body fat rate very well [31]. A meta-analysis shows that TG levels are significantly elevated among women with GDM compared to women without insulin resistance. This difference persists across all three trimesters of pregnancy [32], although the causal relationship between TG levels and GDM is still not clear [33]. The 7 features chosen in this study are not only significant from a computational point of view but are also supported by previous clinical studies. Interestingly, all 17 indicators selected by machine learning can be supported by relevant clinical literature, which also demonstrates that using machine learning to select features is an effective method.

Our study is the first GDM prediction model that combines feature selection, advanced machine learning technology, and traditional LR methods. Because our final prediction models were built upon a relatively small feature panel, the efficiency differences between prediction methods can be ignored. In terms of prediction efficacy, as shown in Fig 3 and Fig S1, DNN is better than LR, followed by SVM and KNN. The merit of DNN is that it can capture subtle nonlinear relationships between features and the outcome. However, the DNN approach may suffer from overfitting, and DNN is a black box to end-users, making it difficult to explain the contribution of each input features [34]. The LR model is more straightforward, where each feature's contribution to the prediction is salient, especially in real time clinical practices where efficiency needs to be considered.

Our prediction model was designed to predict GDM based on the electronic health record of the first trimester, which is three months earlier than standard GDM diagnosis in 24-28th week of pregnancy. The uncertain alteration of hormonal milieu during subsequent months increases the difficulty of predicting GDM via our LR model. The increased release of human placental growth hormone (hPGH), human chorionic somatomammotropin, progesterone, and placental growth hormone from placenta reprogrammed by maternal physiology to become insulin resistant in the second and third trimesters of pregnancy [35], are especially disruptive. Due to these limitations, it is impossible to predict all GDM patients in early pregnancy. However, our LR, SVM, and DNN models have a high success rate (AUC more than 0.80), and outperform existing early detection techniques.

The HL test was adopted to evaluate the calibration of prediction models [36]. The KNN model does not provide Pt value (it only provides "Yes" or "No"), so the HL test and the DCA curve did not contain KNN model result. The P values of all models for HL test were <0.0001, which implied that the model calibrations were not optimal. Excellent discrimination and dissatisfied calibration mean that the LR and DNN models are able to distinguish high risk status of GDM in early pregnancy, while the specific risk probability provided by these models can be improved further [37]. Although none of the models pass the HL test, the 7-feature LR model revealed better calibration than the DNN model. Notably, the deficiency of an appropriate assessment method for advanced machine learning models may lead to failure in calibration. Because the Pt values calculated from DNN and SVM models don't represent risk probability, unlike the LR model, the HL test may be an inadequate method to evaluate the calibration of DNN and SVM models.

DCA was applied in this study for the assessment of clinical utility. This novel method offers insight for clinical consequences based on threshold probability (Pt) [38]. According to the results of the calculated DCA curves, the LR model and 7-feature DNN model were able to detect about 5 positive GDM cases per 100 pregnant women in early pregnancy without increasing false positives when the Pt value was set at 0.17 (Youden index). Generally, pregnant women with a high risk of GDM in the first trimester would receive physical activity and face-to-face nutrition intervention with nutritionists, rather than insulin injections [39,40]. The side effect of misdiagnosis brought by our models was slight, because lifestyle intervention is not an aggressive therapy. Therefore, relative low Pt values (0.17) are acceptable for doctors and patients. The all-feature DNN model revealed better DCA results when the Pt value is set to 0.2. Additionally, while the optimal Pt value (Youden index) was set at 0.03 for the 7-feature SVM model, 0.05 for the all-feature SVM model, and 0.02 for the all-feature LR model, these models added no additional benefit than the intervene-all-patients scheme.

Limitations of this study include a limited sample size and the fact that all the data used are collected from a single center. The diversity of laboratory testing between different hospitals caused by different laboratory instruments may also influence the effect of prediction and extrapolation. In future work, we will collect multi-center clinical data to verify the extrapolation of these prediction models. However, these shortcomings do not affect the demonstration that our feature selection and machine learning-based methodology is effective in early GDM prediction.

In conclusion, this study established prediction models in early pregnancy for the early intervention of GDM. Combining machine learning with clinical experience in feature selection has enormous utilization potential in establishing GDM prediction models based on demographic characteristics and laboratory testing. Our 7-feature LR and all-feature DNN prediction model have effective discrimination power for predicting GDM in early pregnancy. However, different modeling methods have distinct advantages and disadvantages, and the traditional LR model cannot be completely replaced yet.

Acknowledgments

We thank all the people who helped to collect the data and the graduate students who took part in the statistical analysis.

Author contributions

Yan-Ting Wu contributed to the research idea and was the project coordinator. Chen-Jie Zhang contributed to the manuscript drafting. Ben Willem Mol contributed to manuscript modification. Cheng Li contributed to figures production and modification. Lei Chen contributed to the data export from the hospital electronic medical record system. Yu Wang, Jian-Zhong Sheng, and Jian-Xia Fan contributed to providing useful project suggestions. Yi Shi contributed to the machine learning algorithm and statistical analysis. He-Feng Huang contributed project guidance and financial support.

Funding

This study was supported by the National Key Research and Development Program of China (2018YFC1002804, 2016YFC1000203), the National Natural Science Foundation of China (81671412, 81661128010), Foundation of Shanghai Municipal Commission of Health and Family Planning (20144Y0110), and Clinical Skills Improvement Foundation of Shanghai Jiaotong University School of Medicine (JQ201717).

Ethical approval

This study was approved by the Medical Ethical Committee of International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University (No. GKLW2019-05).

Competing interests

No potential conflicts of interest relevant to this article were reported.

Data availability

Please contact the corresponding author for the availability of the original data.

References

1. Gabbe SG. Gestational diabetes mellitus. *N Engl J Med* 1986; 315(16):1025-1026.
2. Chiefari E, Arcidiacono B, Foti D, Brunetti A. Gestational diabetes mellitus: an updated overview. *J Endocrinol Invest* 2017;40(9):899-909.
3. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2018. *Diabetes Care* 2018;41(Suppl 1):S13-S27.
4. Weinert LS. International Association of Diabetes and Pregnancy Study Groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy: comment to the International Association of Diabetes and Pregnancy Study Groups Consensus Panel. *Diabetes Care* 2010;33(7):e97, e98.
5. Ding GL, Wang FF, Shu J et al. Transgenerational glucose intolerance with Igf2/H19 epigenetic alterations in mouse islet induced by intrauterine hyperglycemia. *Diabetes* 2012;61(5):1133-42.
6. Dabelea D, Hanson RL, Lindsay RS et al. Intrauterine exposure to diabetes conveys risks for type 2 diabetes and obesity: a study of discordant sibships. *Diabetes* 2000;49(12):2208-2211.
7. Dabelea D, Mayer-Davis EJ, Lamichhane AP et al. Association of intrauterine exposure to maternal diabetes and obesity with type 2 diabetes in youth: the SEARCH Case-Control Study. *Diabetes Care* 2008;31(7):1422-1426.
8. Tam WH, Ma R, Ozaki R et al. In utero exposure to maternal hyperglycemia increases childhood cardiometabolic risk in offspring. *Diabetes Care* 2017;40(5):679-686.
9. Zhu H, Chen B, Cheng Y et al. Insulin therapy for gestational diabetes mellitus does not fully protect offspring from diet-induced metabolic disorders. *Diabetes* 2019;68(4):696-708.
10. Zhu WW, Yang HX, Wei YM et al. Evaluation of the value of fasting plasma glucose in the first prenatal visit to diagnose gestational diabetes mellitus in China. *Diabetes Care* 2013;36(3):586-590.
11. Leong I. Diabetes: ANGPTL8 as an early predictor of gestational diabetes mellitus. *Nat Rev Endocrinol* 2018;14(2):64.
12. Ning H, Tao H, Weng Z, Zhao X. Plasma fatty acid-binding protein 4 (FABP4) as a novel biomarker to predict gestational diabetes mellitus. *Acta Diabetol* 2016;53(6):891-898.
13. Bao W, Baecker A, Song Y, Kiely M, Liu S, Zhang C. Adipokine levels during the first or early second trimester of pregnancy and subsequent risk of gestational diabetes mellitus: A systematic review. *Metabolism* 2015;64(6):756-764.
14. Naylor CD, Sermer M, Chen E, Farine D. Selective screening for gestational diabetes mellitus. Toronto Trihospital Gestational Diabetes Project Investigators. *N Engl J Med* 1997;337(22):1591-1596.
15. Ting D, Cheung CY, Lim G et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic Populations With Diabetes. *JAMA* 2017;318(22):2211-2223.

16. Shi Y, Cai Z, Xu L, Ren W, Goebel R, Lin G. A model-free greedy gene selection for microarray sample class prediction. 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (IEEE CIBCB) 2006;406-417.
17. Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. GENOME RES 2011(11):1878-1887.
18. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2011;2(27):1-27.
19. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002;46:389-422.
20. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77-87.
21. Cramer JS. The origins of logistic regression (Technical report). 119. Tinbergen Institute. 2002; 167-178.
22. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504-7.
23. Deng L, Yu D. Deep learning: methods and applications. Foundations Trends Signal Processing 2014;7(3-4):197-387.
24. Coustan DR. Gestational diabetes mellitus. Clin Chem 2013;59(9):1310-1321.
25. Buschur E, Stetson B, Barbour LA. Diabetes in pregnancy. Endotext (Internet). 2018;South Dartmouth (MA): MDText.com, Inc.; 2000-.
26. Ooi S, Wong VW. Twin pregnancy with gestational diabetes mellitus: A double whammy? Diabetes Care 2018;41(2):e15-e16.
27. Rahbar S, Blumenfeld O, Ranney HM. Studies of an unusual hemoglobin in patients with diabetes mellitus. Biochem Biophys Res Commun 1969;36(5):838-843.
28. Koenig RJ, Peterson CM, Jones RL, Saudek C, Lehrman M, Cerami A. Correlation of glucose regulation and hemoglobin A1c in diabetes mellitus. N Engl J Med 1976;295(8):417-420.
29. Yamauchi T, Kamon J, Waki H et al. The fat-derived hormone adiponectin reverses insulin resistance associated with both lipodystrophy and obesity. Nat Med 2001;7(8):941-946.
30. Steppan CM, Bailey ST, Bhat S et al. The hormone resistin links obesity to diabetes. Nature 2001;409(6818):307-312.
31. Provencher MT, Chahla J, Sanchez G et al. Body mass index versus body fat percentage in prospective national football league athletes: Overestimation of obesity rate in athletes at the national football league scouting combine. J Strength Cond Res 2018; 32(4):1013-1019.
32. Ryckman KK, Spracklen CN, Smith CJ, Robinson JG, Saftlas AF. Maternal lipid levels during pregnancy and gestational diabetes: a systematic review and meta-analysis. BJOG 2015;122(5):643-651.
33. Savage DB, Watson L, Carr K et al. Accumulation of saturated intramyocellular lipid is associated with insulin resistance. J Lipid Res 2019;60(7):1323-1332.
34. Collins GS, Moons K. Reporting of artificial intelligence prediction models. LANCET 2019;393(10181):1577-1579.
35. Barbour LA, McCurdy CE, Hernandez TL, Kirwan JP, Catalano PM, Friedman JE. Cellular mechanisms for insulin resistance in normal pregnancy and gestational diabetes. DIABETES CARE 2007;30 Suppl2: S112-S119.
36. Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results. the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat 2000;5(4):251-253.
37. Alba AC, Agoritsas T, Walsh M et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA 2017;318(14):1377-1384.
38. Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. JAMA 2015;313(4):409-410.
39. Tobias DK, Zhang C, van Dam RM, Bowers K, Hu FB. Physical activity before and during pregnancy and risk of gestational diabetes mellitus: a meta-analysis. DIABETES CARE 2011;34(1):223-229.
40. Simmons D, Devlieger R, van Assche A et al. Effect of physical activity and/or healthy eating on GDM risk: The DALI lifestyle study. J Clin Endocrinol Metab 2017;102(3):903-913.

Figure legend

Fig 1. Feature selection results. (a) The Pearson correlation coefficients between each feature and the GDM/non-GDM label vector, overall the samples. Bar plots from left to right are absolute values from high to low. (b) The Pearson correlation coefficients between all the features over vectors of all the samples. (c) The feature-way hierarchical clustering results using distance metric based on Pearson correlation coefficients. (d and e). The 10-fold cross-validation based detailed prediction outcomes of each feature selection iteration. Yellow elements are predicted GDM cases and blue elements are predicted non-GDM cases (d) is seeking best accuracy. (e) is seeking AUC. (f) and (g) Feature selection trajectory guided by classification accuracy and AUC respectively, under 10-fold cross-validation framework.

Fig 2. Feature selection results. (a) The Spearman correlation coefficients between each feature and the GDM/non-GDM label vector, over all the samples. Bar plots from left to right are absolute values from high to low. (b) The Spearman correlation coefficients between all the features over vectors of all the samples. (c) The feature-way hierarchical clustering results using distance metric based on Spearman correlation coefficients. (d) and (e) The leave-one-out cross-validation based detailed prediction outcomes of each feature selection iteration. Yellow elements are predicted GDM cases and blue elements are predicted non-GDM cases, (d) is seeking the best accuracy, (e) is seeking the best AUC. (f) and (g) Feature selection trajectory guided by classification accuracy and AUC respectively, under leave-one-out cross-validation framework.

Fig 3. Discriminative power comparison between different prediction models. (a) The ROC curves of different prediction models based on the 7-feature panel (*) and all-feature panel (**). (b) and (c) are the violin plot comparisons of predicted score distribution using different prediction models, with 7-feature panel and all-feature panel, respectively.

Fig 4: Calibration of different models. P value of all prediction models in HL tests is <0.0001. The 7-feature models (a, b, c) have better HL test performance than the versus all-feature models (d, e, f). Because if the model incorporates all features without selection, the model will inevitably over-fit, which will greatly reduce the calibration of models.

Figure 5

Fig 5: Decision curve analysis of different models. When prediction models have best discrimination (Pt sets at Youden index), 7-feature LR model and 7-feature DNN model can detect five positive GDM cases per 100 pregnant women in early pregnancy without increasing false positives (a, c), all-feature DNN model can detect 7.5 positive GDM cases per 100 pregnant women in early pregnancy without increasing false positive (c). While the net benefit brought by all-feature LR model and SVM models is no more than intervene-all-patients scheme under versus Pt value (a, b).

Table 1 Sociodemographic characteristics of GDM and non-GDM cases.

Characteristic	2017 Training Group			2018 Validation Group		
	GDM cases	Controls	P value	GDM cases	Controls	P value
	N=2696	N=14123		N=1837	N=13155	
	No. (%)	No. (%)		No. (%)	No. (%)	
Age (year), median(IQR)	32(29 to 36)	30(28 to 34)	<0.0001	33(30 to 36)	30(28 to 33)	<0.0001
<38	2340(86.8)	13240(93.7)		1577(85.8)	12323(93.7)	
≥38	356(13.2)	883(6.3)	<0.0001	260(14.2)	832(6.3)	<0.0001
Weight before pregnancy, median(IQR)	56.0 (52.0 to 62.0)	54.5 (50.0 to 59.0)	<0.0001	60.0 (53.5 to 65.0)	55.0 (50.0 to 59.0)	<0.0001
Height, median(IQR)	161.0(158.0 to 165.0)	162.0(159.0 to 165.0)	<0.0001	161.3(158.0 to 165.0)	162.0(160.0 to 165.0)	<0.0001
BMI(kg/m ²), median(IQR)	21.6(20.1 to 23.6)	20.7(19.3 to 22.3)	<0.0001	22.6(20.7 to 24.8)	20.8(19.5 to 22.1)	<0.0001
≤23	1620 (60.1)	9926 (70.2)	<0.0001	1050 (57.2)	11064 (84.1)	<0.0001

>23	1076 (39.9)	4197 (29.7)		787 (42.8)	2091 (15.9)	
Drinking	7 (0.3)	39(0.3)	1-0000	32 (1.7)	218 (1.7)	0.7705
Smoking	14 (0.5)	81 (0.6)	0.8883	13 (0.7)	61 (0.5)	0.1565
Educational Background						
Primary school degree	5 (0.2)	10 (0.1)		3 (0.2)	6 (0.05)	
Junior high school degree	102 (3.8)	286(2)		55 (3.0)	295 (2.2)	
High school degree	162 (6)	727 (5.1)	<0.0001	117 (6.4)	647 (4.9)	<0.0001
University degree and above	2427 (90.0)	13100 (92.8)		1662 (90.5)	12207 (92.8)	
Family history of diabetes in a first-degree relative	439 (16.3)	763 (5.4)	<0.0001	228 (12.4)	705 (5.4)	<0.0001

Table 2 Clinical features of GDM and non-GDM cases in the first trimester.

Characteristic	2017 Training Group		P value	2018 Validation Group		P value
	GDM cases	Controls		GDM cases	Controls	
	N=2696 No. (%)	N=14123 No. (%)		N=1837 No. (%)	N=13155 No. (%)	
SBP (mmHg), median(IQR)	114(107 to 122)	110(102 to 117)	<0.0001	116(107 to 124)	110(102 to 117)	<0.0001
DBP (mmHg), median(IQR)	71(65 to 77)	68(62 to 73)	<0.0001	72(66 to 78)	68(62 to 74)	<0.0001
PCOS	13 (0.5)	30 (0.2)	0.0195	29 (1.6)	65 (0.5)	<0.0001
Personal history of GDM	132 (4.9)	44 (0.3)	<0.0001	94 (5.1)	44 (0.3)	<0.0001
Natural pregnancy	2180 (80.9)	12324(87.3)	<0.0001	1432 (80.0)	11462 (87.1)	<0.0001
Multiple pregnancy	110 (4.1)	379(2.7)	<0.0001	77 (4.2)	386 (2.9)	<0.0001
Multipara	1053 (39.1)	4486 (31.8)	<0.0001	713 (38.8)	4008 (30.5)	<0.0001

SBP, systolic blood pressure; DBP, diastolic blood pressure; PCOS, polycystic ovary syndrome

Table 3. Laboratory indexes of pregnant women in early pregnancy.

Characteristics		2017 Training Group			2018 Validation Group		
		GDM cases	Controls	P value	GDM cases	Controls	P value
		N=2696	N=14123		N=1837	N=13155	
Abbreviation	Full name	Median (interquartile range)	Median (interquartile range)		Median (interquartile range)	Median (interquartile range)	
Hepatorenal function							
AST(U/L)	Aspartate transaminase	22.0(17.0 to 24.0)	20.0(17.0 to 23.0)	0.0166	19.0(17.0 to 23.0)	19.0(17.0 to 23.0)	0.4344
ALT(U/L)	Alanine transaminase	14.0(9.0 to 22.0)	12.0(9.0 to 19.0)	<0.0001	13.0(9.0 to 22.0)	9.0(8.0 to 18.0)	<0.0001
UA(μmol/L)	Uric acid	243.0(213.0 to 280.0)	236.0(207.0 to 268.0)	<0.0001	252.0(220.0 to 288.0)	240.6(210.0 to 274.0)	<0.0001
Tbil(μmol/L)	Total bilirubin	6.60(5.39 to 8.14)	6.69(5.38 to 8.05)	0.9846	6.90(5.64 to 8.30)	6.93(5.70 to 8.29)	0.5664
Dbil(μmol/L)	Direct bilirubin	2.98(2.43 to 3.67)	3.07(2.49 to 3.82)	<0.0001	2.83(2.35 to 3.48)	2.93(2.38 to 3.66)	<0.0001
TBA(μmol/L)	Total bile acid	2.40(1.70 to 3.60)	2.90(2.10 to 4.10)	<0.0001	3.30(2.60 to 4.40)	3.70(2.90 to 4.90)	<0.0001
HCY(μmol/L),	Homocysteine	3.1(2.1 to 4.9)	3.4(2.3 to 5.1)	<0.0001	3.0(3.0 to 4.0)	4.0(3.0 to 4.0)	<0.0001
CK (U/L)	Creatine kinase	44.0(36.0 to 53.0)	44.0(36.0 to 53.0)	0.7145	44.0(36.0 to 53.0)	44.0(36.0 to 53.0)	0.8630
LDH (U/L)	Lactic dehydrogenase	151.7(138.0 to 166.0)	151.0(138.0 to 164.0)	0.1673	162.6(147.0 to 179.0)	162.0(148.0 to 177.0)	0.8415
GGT (U/L)	γ-glutamyltranspeptidase	13.0(10.0 to 19.0)	12.0(9.0 to 16.0)	<0.0001	14.0(9.0 to 20.0)	11.0(9.0 to 15.0)	<0.0001
UREA(mmol/L)	Urea nitrogen	3.3(2.9 to 3.8)	3.2(2.8 to 3.7)	<0.0001	3.4(3.0 to 3.9)	3.3(2.9 to 3.8)	<0.0001
CREA(μmol/L)	Creatinine	48.0(44.0 to 53.0)	49.0(45.0 to 53.0)	0.0159	49.0(45.0 to 54.0)	50.0(45.0 to 54.0)	0.0603
ALP(U/L)	Alkaline phosphatase	44.0(38.0 to 51.0)	43.0(37.0 to 49.0)	<0.0001	44.0(39.0 to 51.0)	42.3(37.0 to 49.0)	<0.0001
TP(g/L)	Serum total protein	70.7(68.4 to 73.0)	70.6(68.3 to 73.0)	0.3272	71.2(68.9 to 73.5)	71.2(68.8 to 73.6)	0.9357
ALB(g/L)	Serum albumin	42.0(40.5 to 43.4)	42.0(40.5 to 43.5)	0.3354	41.4(39.9 to 42.9)	41.5(40.1 to 43.0)	0.0040
ALBC	Serum albumin concentration	3.51(3.33 to 3.70)	3.51(3.34 to 3.70)	0.6668	3.41(3.26 to 3.54)	3.40(3.25 to 3.53)	0.0291
GLO(g/L)	Seroglobulin	28.7(26.9 to 30.9)	28.6(26.8 to 30.6)	0.0459	29.8(28.0 to 31.7)	29.6(27.7 to 31.5)	0.0094
AG	Albumin-globulin ratios	1.5(1.4 to 1.6)	1.5(1.4 to 1.6)	0.0125	1.4(1.3 to 1.5)	1.4(1.3 to 1.5)	0.0002
Lipid metabolism							
HDL(mmol/L)	High-density lipoprotein	1.83(1.58 to 2.07)	1.90(1.67 to 2.15)	<0.0001	1.81(1.53 to 2.09)	1.91(1.67 to 2.17)	<0.0001
<0.9mmol/l, No. (%)		5(0.2)	18(0.1)		8(0.4)	22(0.2)	
≥0.9mmol/l, No. (%)		2691(99.8)	14105(99.9)	0.4009	1829(99.6)	13133(99.8)	0.0246
LDL(mmol/L)	Low-density lipoprotein	2.63(2.23 to 3.08)	2.50(2.12 to 2.89)	<0.0001	2.75(2.32 to 3.22)	2.57(2.16 to 3.02)	<0.0001
APO-A(g/L)	Apolipoprotein-A	2.01(1.93 to 2.08)	1.98(1.94 to 2.02)	<0.0001	2.14(2.00 to 2.29)	2.14(2.01 to 2.27)	0.4658
APO-B(g/L)	Apolipoprotein-B	0.89(0.84 to 0.94)	0.85(0.84 to 0.88)	<0.0001	0.80(0.71 to 0.91)	0.74(0.65 to 0.85)	<0.0001

LPA(mg/L)	Lipoprotein(a)	157.8(101.5 to 185.9)	191.2(173.3 to 210.9)	<0.0001	103.0(46.0 to 215)	123.0(57.0 to 232.0)	<0.0001
TG(mmol/L)	Triglyceride	1.47(1.15 to 1.89)	1.22(0.97 to 1.52)	<0.0001	1.53(1.20 to 1.96)	1.24(0.98 to 1.57)	<0.0001
≤2.8, No. (%)		2545(94.4)	13892(98.4)		1705(92.8)	12882(97.9)	
>2.8, No. (%)		151(5.6)	231(1.6)	<0.0001	132(7.2)	273(2.1)	<0.0001
TC(mmol/L)	Total cholesterol	4.59(4.14 to 5.10)	4.45(4.01 to 4.92)	<0.0001	4.63(4.16 to 5.11)	4.44(3.99 to 4.92)	<0.0001
Glycometabolism							
FPG(mmol/L)	Fasting plasma glucose	4.77(4.49 to 5.13)	4.50(4.30 to 4.70)	<0.0001	4.86(4.56 to 5.19)	4.54(4.33 to 4.73)	<0.0001
<5.1mmol/l, No. (%)		1930(71.6)	13779(97.6)		1223(66.6)	12756 (97.0)	
≥5.1mmol/l, No. (%)		766(28.4)	344(2.4)	<0.0001	614(33.4)	399 (2.8)	<0.0001
HbA1c(%)	Glycosylated hemoglobin A1c	5.3(5.1 to 5.5)	5.1(5.0 to 5.3)	<0.0001	5.4(5.3 to 5.6)	5.2(5.1 to 5.4)	<0.0001
≤5.7, No. (%)		2517(93.4)	14052(99.5)		1596(86.9)	13024(99.0)	
>5.7 , No. (%)		179(6.6)	71(0.5)	<0.0001	241(13.1)	131(1.0)	<0.0001
GAL(%)	Glycated albumin	13.2(12.4 to 13.9)	12.9(12.2 to 13.6)	<0.0001	13.2(12.5 to 14.1)	12.9(12.2 to 13.6)	<0.0001
GALC	Glycated albumin concentration	0.41(0.38 to 0.45)	0.40(0.37 to 0.44)	<0.0001	0.40(0.37 to 0.43)	0.39(0.36 to 0.41)	<0.0001
Hematological parameters							
LYMPH(10 ⁹ /L)	Lymphocytes	2.02(1.72 to 2.37)	1.98(1.69 to 2.30)	<0.0001	2.01(1.72 to 2.36)	1.92(1.65 to 2.25)	<0.0001
LYMPH PCT	Lymphocyte percentage	22.7(19.5 to 26.7)	23.0(19.7 to 26.7)	0.1477	22.9(19.6 to 26.6)	22.9(19.5 to 26.7)	0.8177
MONO(10 ⁹ /L)	Monocytes	0.67(0.56 to 0.81)	0.69(0.57 to 0.83)	<0.0001	0.68(0.56 to 0.81)	0.69(0.57 to 0.83)	0.0003
MONO PCT	Monocyte percentage	7.2(6.3 to 8.2)	7.5(6.5 to 8.5)	<0.0001	7.3(6.3 to 8.3)	7.6(6.6 to 8.6)	<0.0001
NEUT(10 ⁹ /L)	Neutrophil count	7.32(6.29 to 8.48)	7.28(6.25 to 8.39)	0.0723	7.28(6.25 to 8.35)	7.19(6.14 to 8.29)	0.0373
NEUT PCT	Neutrophil percentage	75.5(72.4 to 78.6)	75.5(72.5 to 78.4)	0.8071	75.5(72.4 to 78.4)	75.5(72.5 to 78.5)	0.3566
EOS(10 ⁹ /L)	Eosnophil count	0.11(0.07 to 0.17)	0.11(0.07 to 0.17)	0.0974	0.11(0.07 to 0.17)	0.11(0.07 to 0.17)	0.3489
BASO(10 ⁹ /L)	Basophil count	0.02(0.02 to 0.03)	0.02(0.02 to 0.03)	0.8758	0.02(0.01 to 0.03)	0.03(0.02 to 0.04)	0.5586
WBC(10 ⁹ /L)	White blood cell count	9.1(8.3 to 9.7)	9.1(8.3 to 9.7)	0.0595	9.1(8.3 to 9.7)	9.0(8.2 to 9.6)	0.0013
PLT(10 ⁹ /L)	Platelets count	250.4±59.1	243.7±54.8	<0.0001	251.0(217.0 to 292.0)	240.0(207.0 to 276.0)	<0.0001
RBC(10 ¹² /L)	Red blood cell count	4.26(4.04 to 4.50)	4.18(3.97 to 4.41)	<0.0001	4.27(4.05 to 4.51)	4.18(3.96 to 4.40)	<0.0001
HGB(g/L)	Hemoglobin	128.0(122.0 to 134.0)	126.0(120.0 to 132.0)	<0.0001	128.0(122.0 to 134.0)	126.0(119.0 to 132.0)	<0.0001
Nutrition element							
CA(mmol/L)	Calcium	2.35(2.29 to 2.40)	2.34(2.28 to 2.40)	<0.0001	2.33(2.28 to 2.38)	2.32(2.27 to 2.36)	<0.0001
VB12(pmol/L)	Vitamin B12	333.0(267.0 to 410.0)	335.0(266.0 to 412.0)	0.5422	338.0(270.0 to 420.0)	345.0(274.0 to 423.0)	0.0299

25-OH-VD(nmol/L)	25-OH-Vitamin D	44.3(34.0 to 54.7)	43.4(32.5 to 55.3)	0.0134	55.2(44.2 to 68.2)	54.7(42.8 to 67.4)	0.0269
FA	Folic acid	30.9(26.9 to 33.7)	30.3(26.3 to 33.5)	<0.0001	32.1(27.4 to 35.4)	31.4(26.8 to 34.9)	<0.0001
Thyroid Function Test							
TGAb(IU/ml)	Antithyroglobulin antibody	1.20(0.80 to 3.00)	1.20(0.80 to 3.10)	0.0351	1.50(1.00 to 3.50)	1.50(1.00 to 3.60)	0.6335
TSH(mIU/L)	Thyroid stimulating hormone	1.56(1.09 to 2.20)	1.63(1.14 to 2.29)	<0.0001	1.61(1.10 to 2.25)	1.68(1.19 to 2.33)	<0.0001
TT3(nmol/L)	Total triiodothyronine	2.10(2.00 to 2.23)	2.02(1.97 to 2.08)	<0.0001	2.10(1.90 to 2.40)	2.00(1.80 to 2.30)	<0.0001
FT3(pmol/l)	Free triiodothyronine	4.80(4.40 to 5.10)	4.60(4.30 to 5.00)	<0.0001	4.50(4.20 to 4.90)	4.40(4.10 to 4.70)	<0.0001
TT4(pmol/l)	Total tetraiodothyronine	114.2(106.6 to 119.0)	116.0(112.6 to 120.1)	<0.0001	115.5(99.3 to 131.8)	118.9(102.8 to 134.2)	<0.0001
FT4(nmol/L)	Free tetraiodothyronine	13.6(11.9 to 14.9)	14.0(12.5 to 15.2)	<0.0001	13.2(11.1 to 14.4)	13.6(11.9 to 14.8)	<0.0001

Continuous table

Table 4 Selecting features by KNN.

	10-Fold (Accuracy)	LOO (Accuracy)	10-Fold (AUC)	LOO (AUC)
Selected features	FPG†	Fasting plasma glucose	Fasting plasma glucose	FPG
	Lipoprotein(a)	FPG†	FPG†	HbA1c
	Total triiodothyronine	Lipoprotein(a)	Lipoprotein(a)	Family history of diabetes in a first-degree relative
	Age†	Total triiodothyronine	Total triiodothyronine	Triglyceride
	Multiple pregnancy	Age	Triglyceride	Age
		Total tetraiodothyronine	Age	Total triiodothyronine
		APO-A	HbA1c†	Lipoprotein(a)
		Multipara	Total tetraiodothyronine	Age†
		Multiple pregnancy	Age†	Total tetraiodothyronine
			APO-B	Multipara
			Multipara	APO-A
			Previous GDM	Multiple pregnancy
			Multiple pregnancy	Previous GDM
			Smoking	

- .12 †Categorical variable Age <38: 0, age≥38: 1; fasting plasma glucose<5.1 mmol/l: 0, FPG≥ 5.1 mmol/l: 1;
.13 HbA1c<5.7: 0, HbA1c ≥5.7: 1
.14 Using the 10-Fold method, 5 features are selected to obtain the best accuracy; using the LOO method, 9 features are selected to obtain the best accuracy; using 10-Fold method, 14 features are
.15 selected to obtain best ROC area; using LOO method, 13 features are selected to obtain best ROC area.
.16

Table 5 Multivariate analysis for the 7-feature LR model.

	β	Adjusted Odds Ratio (95% CI)	P value
Intercept	-12.3055	-	<0.0001
Age [†]	0.517	1.677(1.436 to 1.958)	<0.0001
Previous GDM	2.4097	11.131(7.436 to 16.662)	<0.0001
Family history of diabetes in a first-degree relative	1.1413	3.131(2.707 to 3.621)	<0.0001
Multiple pregnancy	0.4077	1.503(1.170 to 1.932)	0.0014
FPG [†]	2.7893	16.268(13.748 to 19.253)	<0.0001
Glycosylated hemoglobin A1c	1.803	6.068(5.006 to 7.354)	<0.0001
Triglyceride	0.5527	1.738(1.612 to 1.874)	<0.0001

[†] Categorical variable: age <38: 0, age ≥38: 1; fasting plasma glucose <5.1 mmol/l: 0, FPG ≥ 5.1 mmol/l: 1.

Table 6 Sensitivity and specificity of different models.

Prediction model	AUC[95%CI]	Best Pt threshold	Sensitivity	Specificity	Youden index
LR*	0.84[0.83-0.85]	0.17	71%	80%	0.51
LR**	0.85[0.84-0.86]	0.02	67%	86%	0.53
KNN*	0.68[0.67-0.69]	-	38%	98%	0.36
KNN**	0.63[0.62-0.64]	-	28%	98%	0.26
SVM*	0.69[0.68-0.70]	0.04	40%	96%	0.36
SVM**	0.82[0.81-0.83]	0.05	61%	89%	0.5
DNN*	0.84[0.83-0.85]	0.17	75%	76%	0.51
DNN**	0.92[0.91-0.93]	0.2	94%	81%	0.75

*: 7-feature model; **: All-feature model

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Supplementary data

Supplement table S1. Risk factors, scoring system for the prediction of GDM and controversial diagnostic criteria reported in the literature.

Study	Risk factors	Risk calculation/diagnostic criteria
Naylor et al.(1), 1997	Age	<30: 0; 31 ~ 34: 1; ≥35: 2
	BMI	<22: 0; 22.1 ~ 24.9: 1; ≥25: 2
	Race	White or black: 0; East Asian: 1; South Asian: 2
Caliskan et al.(2), 2004	Age	<25: 0; ≥25: 1
	BMI	<25: 0; ≥25: 1
	1 st-degree relative with DM	No: 0; Yes:1
	Previous BW >4000g	No: 0; Yes:1
	History of adverse outcome	No: 0; Yes:1
IADPSG(3), 2010	Fasting plasma glucose in the early pregnancy	5.1≤FPG<7.0mmol/L: GDM
van Leeuwen et al.(4), 2010	BMI, race, previous GDM, 1 st- or 2 nd-degree relative with DM,	Probability of GDM = 1/[1+exp(-b)]; b = (-6.1 + (0.13*BMI) + (0.5*multipara with history of GDM) - (0.67*multipara without history of GDM) + (0.83*non-Caucasian ethnicity))
Teede et al.(5), 2011	Age	<25: 0; 25 ~ 34: 1; ≥35: 2
	BMI	<20: 0; 20 ~ 34.9: 1; ≥35: 2
	Race	White: 0; African, East Asian, and South Asian: 1
	1 st-degree relative with DM	No: 0; Yes:1
	Previous GDM	No: 0; Yes:1
Nanda et al.(6), 2011	Age, BMI, race, previous GDM, previous BW >90th centile	Probability of GDM = 1/[1+exp(-b)]; b=(-8.68947 + (0.10852*BMI) + (0.05365*age) + (1.00312*South Asian) + (0.88785*East Asian) + (3.72259*previous GDM) + (0.67673*previous BW >90th centile))
Sweeting et al.(7), 2017	Age, BMI, race, previous GDM, parity, family history of DM	Probability of GDM = 1/[1+exp(-b)]; b = (-5.9515 + (0.08462*BMI) + (0.0354*age) + (1.7033*South Asian) + (1.5500*East Asian) + (2.8624*previous GDM) + (3.0678*family history of DM) - (0.5836*parity))

- 1.Naylor CD, Sermer M, Chen E, Farine D. Selective screening for gestational diabetes mellitus. Toronto Trihospital Gestational Diabetes Project Investigators. N Engl J Med 1997;337:1591–6.
2. Caliskan E, Kayikcioglu F, Ozturk N, Koc S, Haberal A. A population-based risk factor scoring will decrease unnecessary testing for the diagnosis of gestational diabetes mellitus. ActaObstetGynecol Scand. 2004;83:524–30.
3. Metzger BE, Gabbe SG, Persson B et al. International Association of Diabetes and Pregnancy Study Groups Consensus Panel. International Association of Diabetes and Pregnancy Study Groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. Diabetes Care. 2010;33:676–682.
4. van Leeuwen M, Opmeer BC, Zweers EJ, et al: Estimating the risk of gestational diabetes mellitus: a clinical prediction model based on patient characteristics and medical history. BJOG. 2010;117(1):69–75.
- 5.Teede HJ, Harrison CL, Teh WT, Paul E, Allan CA. Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention. Aust N Z J ObstetGynaecol. 2011;51(6):499-504.
- 6.Nanda S, Savvidou M, Syngelaki A, et al. Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. PrenatDiagn.

2011;31(2):135–41.

7.Sweeting AN, Appelblom H, Ross GP, et al. First trimester prediction of gestational diabetes mellitus: A clinical model based on maternal demographic parameters. Diabetes Res ClinPract. 2017;127:44-50.

.71
.72
.73
.74
.75
.76
.77
.78
.79
.80
.81
.82
.83
.84
.85
.86
.87
.88
.89
.90
.91
.92
.93

Supplementary Figure S1

Supplement Figure S1: Discriminative power of previous models. The AUC of all previous prediction model or diagnostic criteria is less than 0.75. Nanda et al prediction model has the best AUC validated by our data, followed by Naylor et al model, van Leeuwen et al model, Sweeting et al model, IADPSG diagnostic criteria in the early pregnancy, Teede et al model, and Caliskan et al model.