nature medicine

FOCUS | LETTERS

https://doi.org/10.1038/s41591-019-0724-8

# Prediction of gestational diabetes based on nationwide electronic health records

Nitzan Shalom Artzi[1,2,8], Smadar Shilo[1,2,3,8], Eran Hadar[4,5,8], Hagai Rossman[1,2], Shiri Barbash-Hazan[4], Avi Ben-Haroush[4,5], Ran D. Balicer[6,7], Becca Feldman[6], Arnon Wiznitzer[4,5]* and Eran Segal[1,2]*

**Gestational diabetes mellitus (GDM) poses increased risk of short- and long-term complications for mother and offspring[1–4]. GDM is typically diagnosed at 24–28 weeks of gestation, but earlier detection is desirable as this may prevent or considerably reduce the risk of adverse pregnancy outcomes[5,6]. Here we used a machine-learning approach to predict GDM on retrospective data of 588,622 pregnancies in Israel for which comprehensive electronic health records were available. Our models predict GDM with high accuracy even at pregnancy initiation (area under the receiver operating curve (auROC) = 0.85), substantially outperforming a baseline risk score (auROC = 0.68). We validated our results on both a future validation set and a geographical validation set from the most populated city in Israel, Jerusalem, thereby emulating real-world performance. Interrogating our model, we uncovered previously unreported risk factors, including results of previous pregnancy glucose challenge tests. Finally, we devised a simpler model based on just nine questions that a patient could answer, with only a modest reduction in accuracy (auROC = 0.80). Overall, our models may allow early-stage intervention in high-risk women, as well as a cost-effective screening approach that could avoid the need for glucose tolerance tests by identifying low-risk women. Future prospective studies and studies on additional populations are needed to assess the real-world clinical utility of the model.**

GDM is defined as glucose intolerance that is first recognized during pregnancy. GDM is a common complication of pregnancy, occurring in 3–9% of pregnancies[7], typically diagnosed at 24–28 weeks of gestation[8]. GDM is associated with short- and long-term adverse outcomes, affecting both mothers and infants. Women with GDM are predisposed to many co-morbidities including operative delivery and type 2 diabetes mellitus (DM)[1]. Offspring of mothers with GDM are prone to adverse health outcomes including fetal macrosomia, respiratory difficulties and metabolic complications in the neonatal period, and carry a higher risk for future obesity and alteration in glucose metabolism[2–4].

The rising prevalence of GDM, reflective of the increase in type 2 DM prevalence, warrants the development of new prevention strategies[9]. While results from randomized controlled trials aimed at the prevention of GDM with nutritional and lifestyle interventions are conflicting[10], some studies have demonstrated that a major reduction in risk is possible, especially when interventions are initiated during the first or early second trimesters[5,6]. Identifying women at high risk for GDM at an early stage of pregnancy would therefore enable implementation of early intervention strategies that might prevent or reduce GDM prevalence and its associated co-morbidities.

Several studies have utilized electronic health records (EHRs) to construct prediction models for mortality[11,12] and disease onset[13–15]. However, despite progress in identifying GDM risk factors[16–18], no predictive model has thus far been established in clinical practice. Here, we constructed a model for GDM prediction based on nationwide EHR data and evaluated its performance from pregnancy initiation up to 20 weeks of gestation.

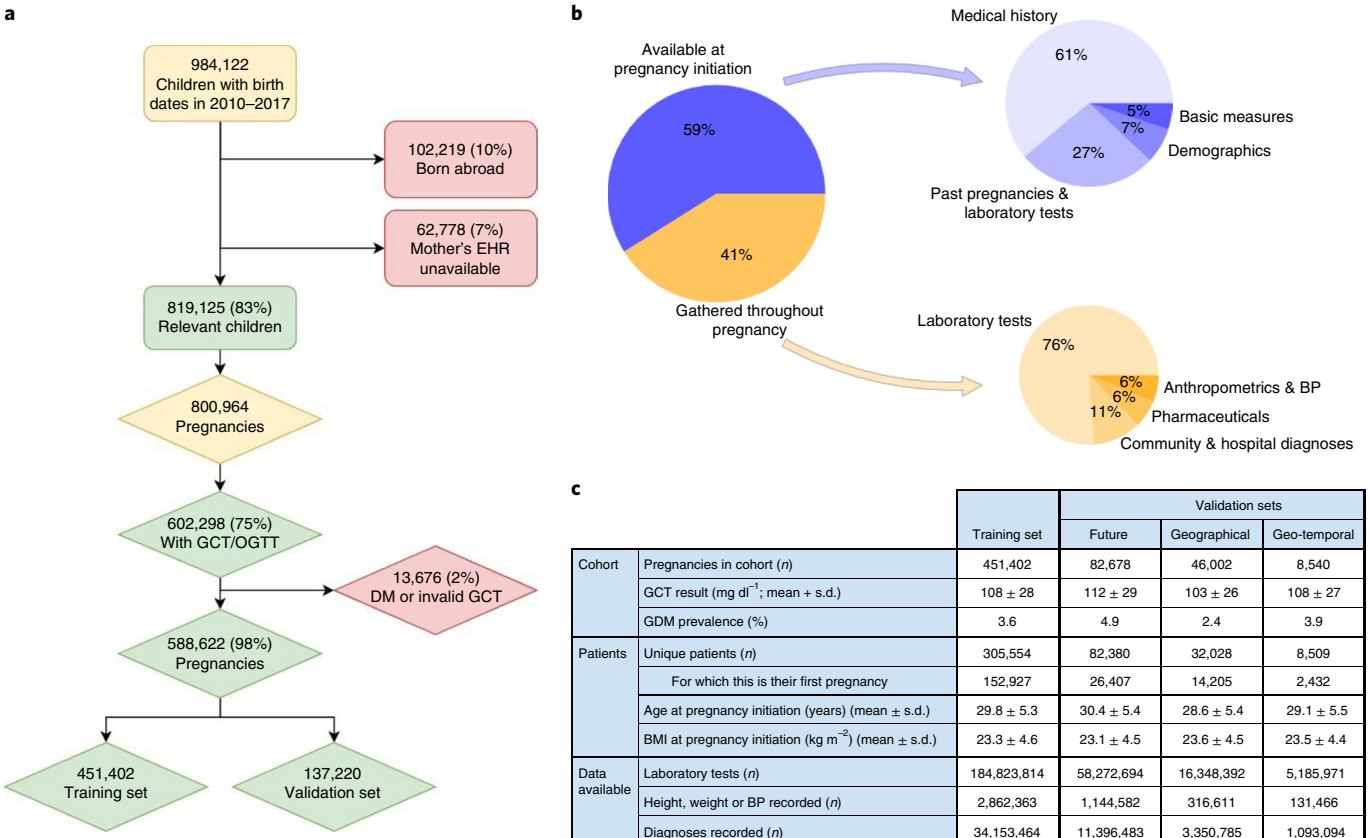
We included a total of 588,622 pregnancies from 368,351 women who gave birth between 2010 and 2017 in our cohort (Fig. 1; see Methods). The prevalence of GDM diagnosed by a two-step diagnostic test, comprising a glucose challenge test (GCT) and an oral glucose tolerance test (OGTT) at 24–28 weeks of gestation, was 3.9% (see Methods). Before any analysis, the study population was split into a training set that included 451,402 pregnancies and three validation sets: a future validation set that included 82,678 pregnancies ending in 2017 or beyond, a geographical validation set that included 46,002 pregnancies of women living in Jerusalem and a geo-temporal validation set of 8,540 pregnancies satisfying both conditions.

We first established the baseline model, termed the baseline risk score, defined as the summation of seven binary variables recommended by the National Institute of Health (NIH) as GDM risk factors[19] (see Methods). As expected, odds ratios for all parameters are >1.0 (1.28–3.92), consistent with their classification as risk factors, and the risk score is predictive of GDM status (Extended Data Fig. 1). The highest precision achieved by this score was 30%, and its auROC was 0.682.

To evaluate whether EHR-derived information might improve GDM prediction we compiled a set of 2,355 features, most of which are already available at pregnancy initiation (Fig. 1b; see Methods). We then used these to train a gradient-boosting model to predict the probability that each held-out sample (individuals not included in the training set) would develop GDM. This EHR-based model achieved an auROC of 0.854 and area under the precision-recall curve (auPR) of 0.318, compared to 0.682 and 0.097, respectively, achieved by the baseline risk score (Fig. 2a,b), all on the future validation set. The model provides 117-fold enrichment between the lowest and highest risk deciles, consistent with the predicted probabilities (Fig. 2c). The model achieved auROC of 0.875 and 0.863 for the geographical and geo-temporal validation sets, respectively (Extended Data Figs. 2–4).

We next examined whether the predictions differ in accuracy for different subsets of the population at 20 weeks of gestation by considering the following: (1) first pregnancy: women with no previous

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Pediatric Diabetes Unit, Ruth Rappaport Children's Hospital, Rambam Healthcare Campus, Haifa, Israel. [4]Helen Schneider Hospital for Women, Rabin Medical Center, Petach Tikva, Israel. [5]Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel. [6]Clalit Research Institute, Clalit Health Services, Tel Aviv, Israel. [7]Department of Public Health, Faculty of Health Sciences, Ben-Gurion University, Beer-Sheva, Israel. [8]These authors contributed equally: Nitzan Shalom Artzi, Smadar Shilo, Eran Hadar. *e-mail: ArnonW@clalit.org.il; eran.segal@weizmann.ac.il




c

| | | Training set | Validation sets | | |
|---|---|---|---|---|---|
| | | | Future | Geographical | Geo-temporal |
| Cohort | Pregnancies in cohort (*n*) | 451,402 | 82,678 | 46,002 | 8,540 |
| | GCT result (mg dl$^{-1}$; mean + s.d.) | 108 ± 28 | 112 ± 29 | 103 ± 26 | 108 ± 27 |
| | GDM prevalence (%) | 3.6 | 4.9 | 2.4 | 3.9 |
| Patients | Unique patients (*n*) | 305,554 | 82,380 | 32,028 | 8,509 |
| | For which this is their first pregnancy | 152,927 | 26,407 | 14,205 | 2,432 |
| | Age at pregnancy initiation (years) (mean ± s.d.) | 29.8 ± 5.3 | 30.4 ± 5.4 | 28.6 ± 5.4 | 29.1 ± 5.5 |
| | BMI at pregnancy initiation (kg m$^{-2}$) (mean ± s.d.) | 23.3 ± 4.6 | 23.1 ± 4.5 | 23.6 ± 4.5 | 23.5 ± 4.4 |
| Data available | Laboratory tests (*n*) | 184,823,814 | 58,272,694 | 16,348,392 | 5,185,971 |
| | Height, weight or BP recorded (*n*) | 2,862,363 | 1,144,582 | 316,611 | 131,466 |
| | Diagnoses recorded (*n*) | 34,153,464 | 11,396,483 | 3,350,785 | 1,093,094 |
| | Pharmaceuticals dispensed (*n*) | 25,098,401 | 7,996,519 | 2,317,888 | 741,957 |

**Fig. 1 | Data and cohort characteristics. a**, Cohort selection. Pregnancies were first identified by offspring birth date. Next, women with pre-existing DM, pregnancies with no record of glucose testing (50 or 100 g) and those with missing OGTT were excluded. Finally, the cohort was divided into training and validation sets (see Methods). **b**, Feature availability distribution. Pie charts are divided according to the sum of data points in each feature set. A substantial portion of the data originates from laboratory test results during current or previous pregnancies. **c**, Basic characteristics of the cohort data. Numbers of data items (laboratory tests, diagnoses and so on) before patients underwent GCT during pregnancy are presented BP, blood pressure.

record of pregnancy; (2) previous GCT: women with a record of a GCT from a previous pregnancy; and (3) high-risk: women with baseline risk score >2. Across all subgroups, our EHR-based model had higher auROC and auPR values than the baseline model (Fig. 2d).

Finally, we evaluated the ability to predict GDM at different stages of gestation up to 20 weeks, by constructing models based only on data collected before that week. The results of this analysis show that, although prediction improves by incorporating features gathered during pregnancy progression, predictions at pregnancy initiation incur only a small reduction in accuracy and still outperform the baseline model auPR by two- to threefold. This effect is more marked for women in their second or subsequent pregnancy (Fig. 2e). We ensured that the model predictions are well calibrated[20], namely that they reflect the actual expected risk of an individual and, furthermore, demonstrate the utility of the predictor by considering its decision curve[21] (Extended Data Fig. 5). Hereafter, we present results based on our full model evaluated at week 20 of pregnancy as this contains all the relevant features, but we note that models derived from earlier stages of pregnancy achieved similar results.

To gain insight into the features that contribute most to the model predictions, we used the feature attribution framework of Shapley values[22] (see Methods). Shapley analysis identified the most predictive feature for GDM diagnosis to be the GCT result from the previous pregnancy, followed by maternal age and fasting blood glucose in the first trimester (F1) (Fig. 3a). Of note, the fasting glucose test, which is performed routinely in F1, was performed for 451,685 women and was more common than glucose testing in the second trimester (F2), which was performed for 131,403 women, possibly indicating that its contribution to the model's performance was more prominent. Using the additive nature of the Shapley values, we also computed the feature importance score for feature sets by summing of Shapley values per set (Fig. 3b).

We further used Shapley values to build dependence plots that capture the nonlinear associations of every feature. Dependence plots show the Shapley value of a specific feature, representing its predicted contribution, in the form of relative risk (RR) against the feature's value (Extended Data Fig. 6; see Methods). We examined dependence plots for two well-known risk factors for GDM: prepregnancy maternal body mass index (BMI)[23] and the number of relatives diagnosed with DM[24]. For prepregnancy BMI, the RR for GDM starts to increase when the individual's BMI exceeds 21 kg m$^{-2}$, becomes a risk factor when it exceeds 24 kg m$^{-2}$ and plateaus when it is >30 kg m$^{-2}$ at RR = 1.13 (Fig. 3c). Of note, only 6.9% of the women included in our cohort had a BMI >30, possibly since higher BMI is also a risk factor for type 2 DM, and women with type 2 DM were not included in our cohort. As expected, the RR for GDM increases as the number of the first-degree family members with GDM increases, reaching RR = 1.8 in women with six relatives diagnosed with DM (Fig. 3d). Analysis of pregestational hemoglobin A1c percentage (HbA1c%) revealed an increased RR of GDM as pregestational HbA1C increased, even with HbA1C% values that are considered to be within the normal range (<5.7%). A steeper increase in RR occurs at HbA1C% > 5.9% (Fig. 3e).

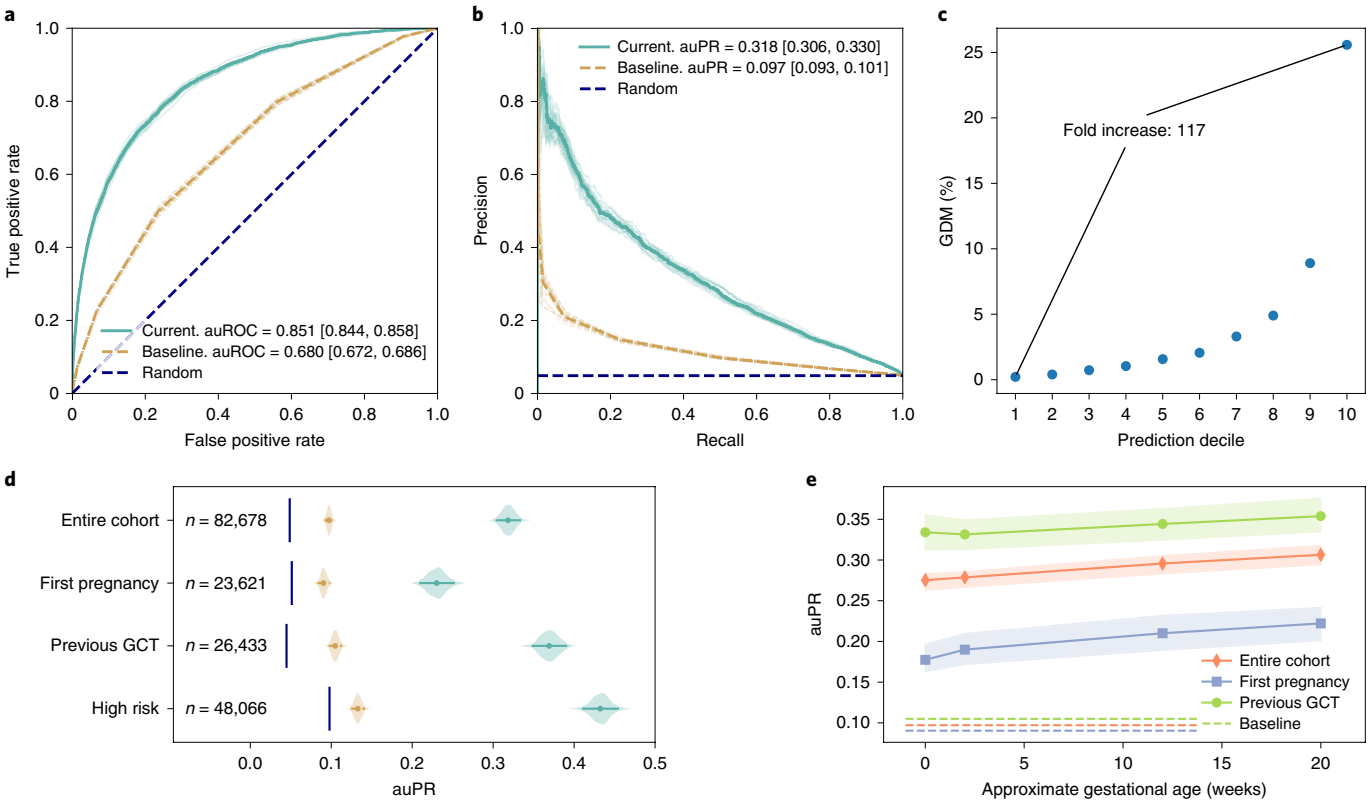



**Fig. 2 | Predictive model evaluation. a**, auROC curves comparing our model (solid) and the baseline risk score (dashed). Lighter-colored lines are auROC curves of stratified partition of the validation set (not shown in auROC); bracketed values are 95% confidence intervals calculated from a normal fit of those curves. **b**, auPR curve with the same properties as in **a**. **c**, Fraction of GDM-positive samples in every decile of predicted probability, showing 117-fold enrichment between the highest and lowest deciles. **d**, Predictions on different subsets of the cohort. auPR is shown for each subset, for our model (blue) and the baseline score (orange). Prediction for women with a previous record of GCT is more effective than for those in their first pregnancy, and prediction for women with a higher baseline risk score (>2) has the greatest deviation from the baseline risk score model. Error bars, 95% confidence intervals; dark blue lines, prevalence in each subset. Shaded area represents the distribution of the relevant score. **e**, Performance by gestational age at prediction. Each point is the evaluation score of a model built only with features available at that time point ($n$ = 82,678 for **a**–**c**; subset sample sizes are listed in **d**).

To further explore the importance of GCT in the previous pregnancy, we conducted the following analysis: for every patient we plotted the combined Shapley value for all glucose tests (GCT and OGTT, if applicable) during previous pregnancy versus the value of GCT in the previous pregnancy (Fig. 3f). This analysis revealed that the GCT result in the previous pregnancy is more predictive than GDM diagnosis in the previous pregnancy. For example, a patient with GCT of 180 mg dl$^{-1}$ will be at a higher risk of GDM in her next pregnancy irrespective of whether she is diagnosed with GDM after 100-g OGTT. On the other hand, a patient with GCT of <75 mg dl$^{-1}$ will have a GDM risk in her next pregnancy as low as 20% of the population prevalence.

Our feature contribution analysis drove us to try and establish a simpler prediction model based on a minimal number of the most influential features, as opposed to our full model based on >2,000 EHR features. To this end, we evaluated the performance of a model with only nine simple questions that a patient can answer herself. To emulate its use in practice, we trained and evaluated only on subjects with no missing values in this part of the study, resulting in a cohort of 417,601 women. This model achieves an auROC of 0.799 and auPR of 0.241, compared to 0.678 and 0.100, respectively, for the baseline model (Fig. 4a–c).

Finally, we emulated usage of our predictive model as a screening tool to identify women who are less likely to develop GDM, rather than subjecting those who fall below a certain risk threshold to the usual two-step GCT plus OGTT (GCT/OGTT) diagnostic process. To this end, we assessed the trade-off of missing diagnoses when implementing such screening across varying risk group thresholds, by analyzing the proportion of women who could avoid testing versus the predictor miss rate—that is, the percentage of GDM-positive women not accurately diagnosed by this approach (Fig. 4d). Indeed, our results show that a large proportion of the population could avoid taking the test. For example, if we permit 20% of diagnoses to be missed, which is on a par with the misdiagnosis rate of GCT[25,26], then 79% of all women with a GCT result in their previous pregnancy can avoid the test in their next pregnancy.

In this study, we examined the ability to utilize EHRs for prediction of GDM in the early stages of pregnancy, allowing for both early-stage interventions and effective GDM screening. Although several scoring systems for GDM risk stratification have been developed in recent years[27], they are not commonly used in routine practice and are not recommended according to current guidelines. Our results show that EHRs can be used to produce accurate predictions of GDM risk, performing substantially better than a baseline model based on commonly assessed risk factors. Our retrospective analysis demonstrates that accurate prediction of GDM is feasible even at pregnancy initiation, with an auROC of 0.836, close to the performance of a predictor constructed later on in pregnancy, which provided an auROC of 0.854. The model, when performed at pregnancy initiation, achieved >55% precision (positive predictive value) on the 0.9% highest risk, a subgroup that includes 10% of all current pregnancy GDMs.

Other than well-known risk factors for GDM such as maternal age[28] and family history of DM[29], our analysis revealed factors that were not

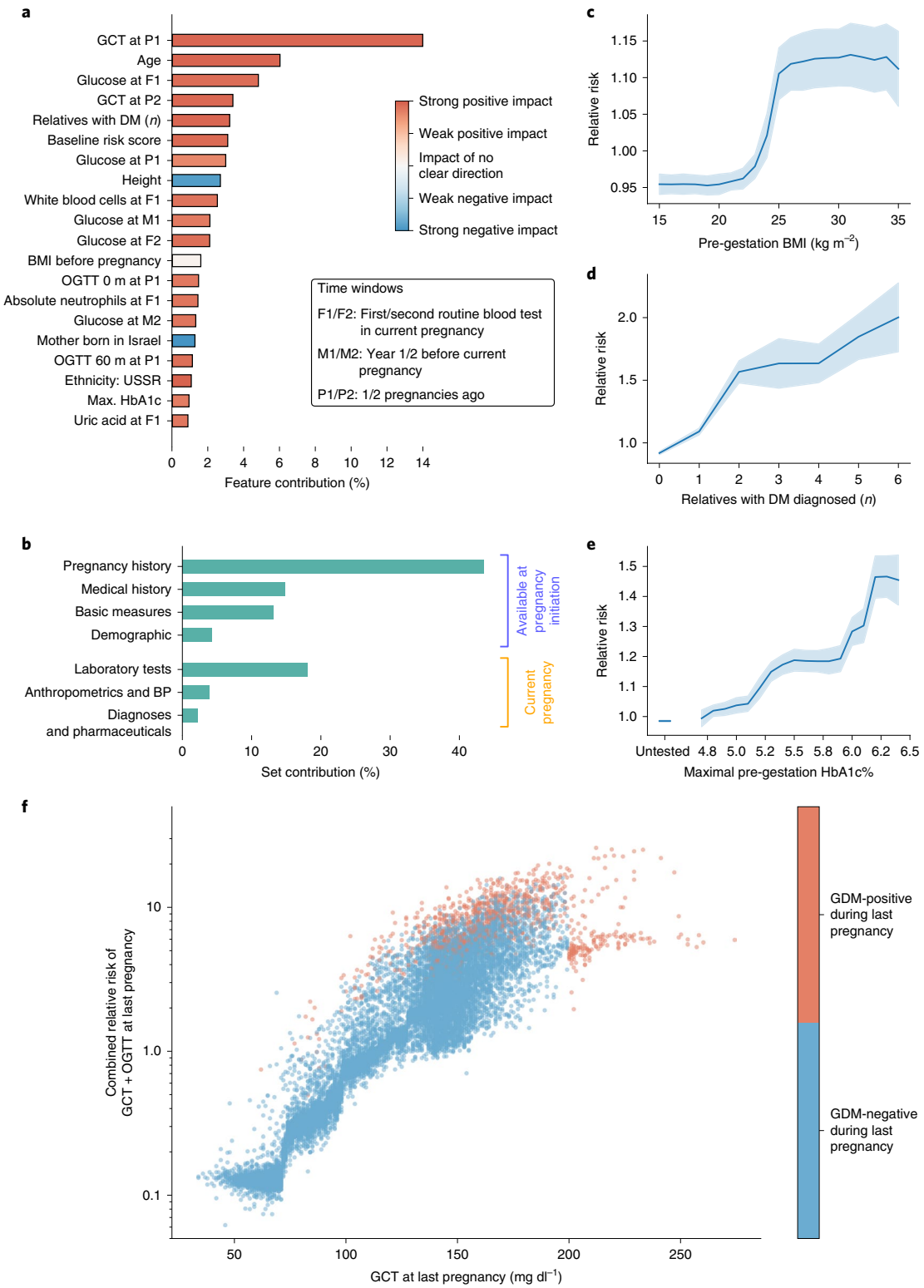



**Fig. 3 | Shapley values-based interpretation of the model. a**, Feature importance of the top 20 contributing features. Bar colors indicate direction of influence, based on the dependence plot of this feature. **b**, Analysis of contributing feature category. Shapley values were summed for each feature set, and the mean of their absolute sum was computed across all samples, producing a feature importance score for sets of features. **c**–**e**, Three examples of dependence plots, showing predicted relative risk versus feature value for BMI before pregnancy (**c**), number of first-degree relatives with diagnosed DM (**d**) and pregestation HbA1C% blood test (**e**). Bands represent s.d. of the population per bin, which is related to interactions between input features. Regions of larger vertical bands represent *x*-axis values in which other features modified the risk attributed to the *x*-axis feature ($n = 71{,}952$ for **c** and 82,678 for **d**,**e**). **f**, Combined Shapley value for all GCT/OGTT results during the previous pregnancy is plotted against GCT value during the current pregnancy. Every point is a sample and is colored according to GDM status during the previous pregnancy. While GDM in a previous pregnancy predicts that in the current pregnancy, GCT provides a continuous prediction for all women who took the test previously. Additionally, GCT in the previous pregnancy also indicates women whose GDM risk is as low as that of 20% of the population ($n = 45{,}807$).

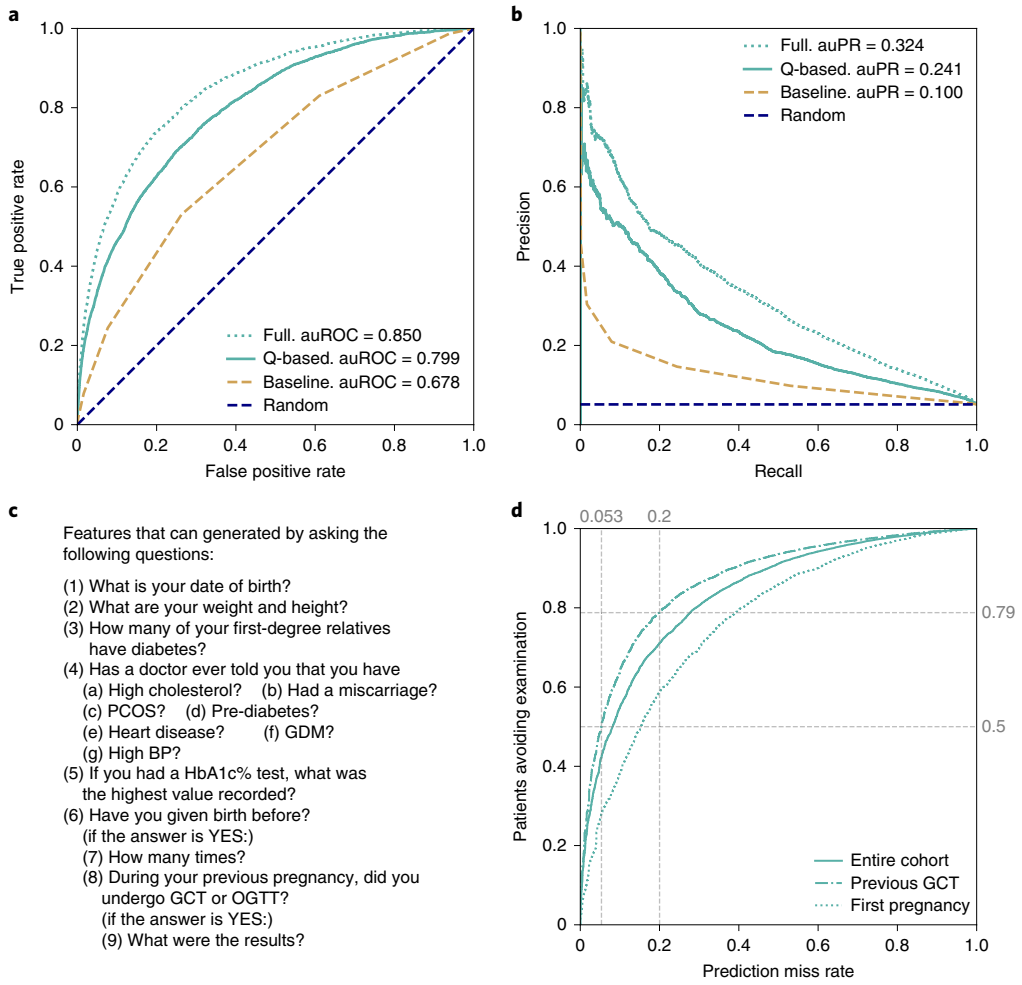



**Fig. 4 | Questionnaire-based prediction, and efficiency of the predictor as a GDM screening tool. a**,**b**, Validation results of a questionnaire-based (Q-based) predictor using nine simple questions. The auROC (**a**) and auPR (**b**) curves of the model are shown ($n = 71,952$). **c**, The list of questions that comprise the predictor. **d**, Our model as a tool for GDM screening. In this scenario, we present the trade-off of not testing low-risk patients while retaining the current system of GCT/OGTT for all others. The ratio of GCT avoidance is plotted against the predictor miss rate—that is, the percentage of GDM-positive patients that would not be diagnosed following this approach ($n = 82,678$). PCOS, polycystic ovary syndrome.

previously reported to be highly predictive of GDM. The main risk factors identified were GCT results in previous pregnancies. While women with a history of GDM in previous pregnancies are recognized as being at increased risk for GDM in the current pregnancy[30], we found that the GCT result in previous pregnancies is far more predictive (Fig. 3f). This may suggest new risk assessment guidelines based on explicit GCT values rather than on GDM diagnosis.

Although maximal prediction accuracy requires using the patient's entire EHRs, we demonstrated that nine simple questions, which can be answered by the woman herself, still enable accurate prediction (auROC = 0.799). This may allow accurate GDM risk estimation by web- or smartphone-based self-assessment tools.

Our work has several clinical applications. First, it can facilitate early-stage interventions for high-risk women. The effect of early-pregnancy interventions on GDM development is well studied but lacks a clear consensus[31]. Some studies suggest a GDM risk reduction of up to 39% with combined diet and exercise interventions during early pregnancy in high-risk pregnant women[5]. One of the challenges in attempting to analyze the efficacy of prevention strategies is the low prevalence of GDM in the population. Our prediction model may be used to identify and recruit a high-risk cohort with risk of up to 70% for GDM (Fig. 2). The current study therefore paves the way for future randomized control trials to further study both the effectiveness of the use of a model for early prediction of GDM and possible preventive interventions.

Another impactful application is in aiding the construction of effective GDM screening approaches. One of the major issues regarding GDM diagnosis is whether universal or selective screening should be used[30,32]. Currently, a 50-g GCT or a similar universal screening method is commonly used, followed by 100-g OGTT if needed[33]. However, 20% of GDM cases are estimated to be missed using this screening approach[25,26]. Our results suggest that a more efficient approach could be established by using the prediction model for the identification of low-risk women who can then avoid the GCT and OGTT altogether, or for high-risk women who may be referred directly to a single-step 100-g OGTT, thereby creating a selective, cost-effective screening method. Because most of the population is predicted to have a low likelihood of GDM this could be highly effective, as demonstrated in Fig. 4d. Avoiding 50% of GCTs in women who previously underwent a GCT would result in a miss rate of only 5% when diagnosing GDM according to the two-step approach guidelines. Additionally, women at high risk for GDM development could be referred directly to the diagnostic 100-g OGTT and thus avoid the screening test, potentially increasing overall adherence. Accurate selective screening is highly desirable, as it reduces both costs and physical inconvenience for women at low or high risk for

GDM. The utility of this approach should be tested in designated prospective clinical trials.

Our study has several limitations. First, our prediction model is based on retrospective EHR data that have inherent biases and are influenced by the interaction of the patient with the health system[34]. However, these biases are reduced here since the data contain information originating from a nongovernmental, nonprofit organization that includes the majority of the Israeli population, and since the outcome of the model is based on routine pregnancy tests that are comprehensively documented in the EHRs. Another limitation is that the data do not contain information regarding lifestyle or dietary habits, previously shown to be associated with GDM development[35]. Furthermore, we assessed GDM only in pregnancies that resulted in a recorded birth but, given the very low incidence of stillbirths in pregnancies with and without GDM[36], this is not expected to bias our results heavily. Additionally, the performance of our simple prediction model was calculated using data ascertained from the EHRs and was not based on an actual self-reported survey. It is possible that, as a self-assessed tool, its performance will be different due to questionnaire response biases. Finally, the predictor was trained and validated on EHRs of the Israeli population. Although applicability to other populations needs to be shown, the size of the data, the validation process, and the fact that the analysis validated the utility of established risk factors for GDM development, all support its ability to generalize to other populations.

In conclusion, our work demonstrates that accurate and calibrated predictions of GDM before and early in pregnancy can be achieved. These results could have many implications for the health of pregnant women and their offspring. Our predictive model could become the basis for a selective screening process for GDM diagnosis, and for identification and implementation of early-stage pregnancy interventions to prevent or reduce the development of GDM and its associated adverse health outcomes. Future prospective studies, as well as those on other populations, are needed to evaluate the clinical impact of the model.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-019-0724-8.

Received: 23 July 2019; Accepted: 26 November 2019;
Published online: 13 January 2020

## References

1. Lowe, L. P. et al. Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study: associations of maternal A1C and glucose with pregnancy outcomes. *Diabetes Care* **35**, 574–580 (2012).
2. Lowe, W. L. et al. Association of gestational diabetes with maternal disorders of glucose metabolism and childhood adiposity. *JAMA* **320**, 1005–1016 (2018).
3. Scholtens, D. M. et al. Hyperglycemia and Adverse Pregnancy Outcome Follow-up Study (HAPO FUS): maternal glycemia and childhood glucose metabolism. *Diabetes Care* **42**, 381–392 (2019).
4. Zhao, P. et al. Maternal gestational diabetes and childhood obesity at age 9–11: results of a multinational study. *Diabetologia* **59**, 2339–2348 (2016).
5. Koivusalo, S. B. et al. Gestational diabetes mellitus can be prevented by lifestyle intervention: the Finnish gestational diabetes prevention study (RADIEL): a randomized controlled trial. *Diabetes Care* **39**, 24–30 (2016).
6. Wang, C. et al. A randomized clinical trial of exercise during pregnancy to prevent gestational diabetes mellitus and improve pregnancy outcome in overweight and obese pregnant women. *Am. J. Obstet. Gynecol.* **216**, 340–351 (2017).
7. Donovan, P. J. & McIntyre, H. D. Drugs for gestational diabetes. *Aust. Prescr.* **33**, 141–144 (2010).
8. American Diabetes Association. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes Care* **41**, S13–S27 (2018).
9. Hunt, K. J. & Schuller, K. L. The increasing prevalence of diabetes in pregnancy. *Obstet. Gynecol. Clin. N. Am.* **34**, 173–199 (2007).
10. Bain, E. et al. Diet and exercise interventions for preventing gestational diabetes mellitus. *Cochrane Database Syst. Rev.* CD010443 https://doi.org/10.1002/14651858.CD010443.pub2 (2015).
11. Avati, A. et al. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* **18**(Suppl 4), 122 (2018).
12. Silva, I., Moody, G., Scott, D. J., Celi, L. A. & Mark, R. G. Predicting in-hospital mortality of ICU patients: the PhysioNet/Computing in Cardiology Challenge 2012. *Comput. Cardiol. (2010)* **39**, 245–248 (2012).
13. Razavian, N., Marcus, J. & Sontag, D. Multi-task prediction of disease onsets from longitudinal lab tests. Preprint at arXiv https://arxiv.org/abs/1608.00647 (2016).
14. Oh, J. et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect. Control Hosp. Epidemiol.* **39**, 425–433 (2018).
15. Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).
16. Danilenko-Dixon, D. R., Van Winter, J. T., Nelson, R. L. & Ogburn, P. L. Universal versus selective gestational diabetes screening: application of 1997 American Diabetes Association recommendations. *Am. J. Obstet. Gynecol.* **181**, 798–802 (1999).
17. Qiu, H. et al. Electronic health record-driven prediction for gestational diabetes mellitus in early pregnancy. *Sci. Rep.* **7**, 16417 (2017).
18. Syngelaki, A. et al. First-trimester screening for gestational diabetes mellitus based on maternal characteristics and history. *Fetal Diagn. Ther.* **38**, 14–21 (2015).
19. US Department of Health and Human Services, National Institutes of Health & Eunice Kennedy Shriver National Institute of Child Health and Human Development. *Am I at Risk for Gestational Diabetes?* https://www.nichd.nih.gov/sites/default/files/publications/pubs/Documents/gestational_diabetes_2012.pdf (2012).
20. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
21. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
22. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Proc. Syst.* **30**, 4765–4774 (2017).
23. Chu, S. Y. et al. Maternal obesity and risk of gestational diabetes mellitus. *Diabetes Care* **30**, 2070–2076 (2007).
24. Williams, M. A., Qiu, C., Dempsey, J. C. & Luthy, D. A. Familial aggregation of type 2 diabetes and chronic hypertension in women with gestational diabetes mellitus. *J. Reprod. Med.* **48**, 955–962 (2003).
25. van Leeuwen, M. et al. Glucose challenge test for detecting gestational diabetes mellitus: a systematic review. *BJOG* **119**, 393–401 (2012).
26. Donovan, L. et al. Screening tests for gestational diabetes: a systematic review for the US Preventive Services Task Force. *Ann. Intern. Med.* **159**, 115–122 (2013).
27. Lamain-de Ruiter, M. et al. External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. *BMJ* **354**, i4338 (2016).
28. Lao, T. T., Ho, L.-F., Chan, B. C. P. & Leung, W.-C. Maternal age and prevalence of gestational diabetes mellitus. *Diabetes Care* **29**, 948–949 (2006).
29. Di Cianni, G. et al. Prevalence and risk factors for gestational diabetes assessed by universal screening. *Diabetes Res. Clin. Pract.* **62**, 131–137 (2003).
30. Teh, W. T. et al. Risk factors for gestational diabetes mellitus: implications for the application of screening guidelines. *Aust. N. Z. J. Obstet. Gynaecol.* **51**, 26–30 (2011).
31. Shepherd, E. et al. Combined diet and exercise interventions for preventing gestational diabetes mellitus. *Cochrane Database Syst. Rev.* **11**, CD010443 (2017).
32. Davey, R. X. Selective versus universal screening for gestational diabetes mellitus: an evaluation of predictive risk factors. *Medical J. Aust.* **174**, 118–121 (2001).
33. Kalter-Leibovici, O. et al. Screening and diagnosis of gestational diabetes mellitus: critical appraisal of the new International Association of Diabetes in Pregnancy Study Group recommendations on a national level. *Diabetes Care* **35**, 1894–1896 (2012).
34. Phelan, M., Bhavsar, N. A. & Goldstein, B. A. Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *EGEMS (Wash DC)* **5**, 22 (2017).
35. Zhang, C. & Ning, Y. Effect of dietary and lifestyle factors on the risk of gestational diabetes: review of epidemiologic evidence. *Am. J. Clin. Nutr.* **94**, 1975S–1979S (2011).
36. Dudley, D. J. Diabetic-associated stillbirth: incidence, pathophysiology, and prevention. *Clin. Perinatol.* **34**, 611–626 (2007). vii.

## Methods

**Data.** Data were extracted from the database of Clalit Health Services (Clalit), the largest healthcare provider in Israel. Almost 5 million individuals, representing over 50% of Israel's adult population, are currently enrolled in Clalit[37], a nongovernmental, nonprofit organization included in the national health insurance law in that country. Dating back to 2002, the database contains EHRs of over 11 million patients and with >5.4 billion numerical and categorical entries. The data analyzed included anthropometrics (height and weight), blood pressure (BP) measurements, blood and urine laboratory tests, diagnoses recorded by physicians, and pharmaceuticals prescribed and dispensed. Most of the data originate from community clinics records, but records from Clalit's 14 hospitals were also included in the analysis.

**Study population and outcome definition.** In Israel, GDM is diagnosed by a two-step procedure performed routinely for all women at 24–28 weeks of pregnancy, in accordance with NIH guidelines[38]. In the first step, a 1-h, 50-g, GCT is performed; women with glucose levels >200 mg dl$^{-1}$ receive a GDM diagnosis. Women with a GCT value >140 mg dl$^{-1}$ are referred to a second step, in which an additional 100-g, 3-h OGTT is performed. Women with two glucose measurements above the thresholds of 95, 180, 155 and 140 mg dl$^{-1}$ under fasting conditions (zero), 1, 2 and 3 h after glucose intake, respectively, also receive a GDM diagnosis[8,39]. Note that although these two tests are similar in nature and sometimes denoted simply as 50-g GTT and 100-g GTT, we use the GCT and OGTT notation for simplicity.

We identified pregnancies by birth records and then looked for a GCT or an OGTT in the relevant time period. The version of the Clalit database that we used does not include exact delivery dates for all women, but it has the approximate (±1 month) birth date of every child. As such, we defined our cohort by collecting all birth dates of children of Clalit-insured mothers, looking for GCTs and OGTTs in the relevant period before the delivery, namely 32 weeks before the logged date of birth to 7 weeks following. The facts that in Israel, GCTs are used in pregnancy only and that we initially looked for the pregnancy period, means that the tests included are all pregnancy related.

We defined GDM status based on GCT and OGTT results. GCTs and OGTTs appear in the laboratory test data under five distinct tests: one for 1-h, 50-g GCT and four for fasting, 1-, 2- and 3-h, 100-g OGTT results. We defined GDM in accordance with normal practice, regardless of the order of the tests and without consideration of whether a relevant diagnosis was recorded. In cases where more than one test was conducted, we considered a positive result of a single test as positive. Women who were supposed to undergo a 100-g OGTT due to a screen-positive GCT, but for whom we had no record of the test results, were excluded ($n$ = 9,753, 1.6%). Women with a prepregnancy record of DM were also excluded. Normally women with DM do not take a GCT during pregnancy, but it appears that some (<0.2%) do. To address this issue, we excluded patients who had one of the following markers before the start of pregnancy: (1) a recorded diagnosis of DM, defined as any of the ICD9 codes in 250.x or 357.2, or (2) a recorded non-pregnancy HbA1c% blood test of 6.5 or higher. Note that, although fasting glucose could also be used to diagnose DM, this metric is extremely inaccurate in our data because some non-fasting patients still take the test, and therefore we decided not to use it.

To emulate its use in practice, we considered two validation cross-sections: (1) a future validation set that included pregnancies ending in 2017 or 2018, and (2) a geographical validation set that included pregnancies of patients for which the main clinic locality was Jerusalem. We used the intersection of both conditions—pregnancies of women mostly visiting Jerusalem clinics and who gave birth from 2017 onwards—as a geo-temporal validation set, posing the highest generalization challenge to the model. The training comprised pregnancies of women mostly visiting any other locality and which ended before 31 December 2016. This choice thus represents a setting in which the model may be implemented in practice.

**Baseline risk score.** Currently there are no validated GDM prediction tools employed in clinical practice that represent current standards of care. The NIH recommends a self-administered, eight-item questionnaire to clinicians to determine GDM risk at pregnancy initiation[19]. To establish how a simple set of questions performs as a baseline indicator of current clinical practice, we adapted this questionnaire and calculated a score for every woman in our cohort to establish what we term a baseline risk score. Since the question regarding Hispanic/African American race was irrelevant to our cohort, this score was defined as the summation of seven binary variables. We therefore included seven parameters in our score, defined according to the following binary variables.

(1) Overweight status: true if non-pregnancy BMI is >25 kg m$^{-2}$. Prepregnancy BMI measurements were available for 78% of the women included in the cohort. If there was no record of BMI before the pregnancy, we considered that as false.
(2) Family history of DM: true if a first-degree relative (parent or sibling) has at least one diagnosis of DM, defined as any of the ICD9 codes in 250.x or 357.2. Only diagnoses available at pregnancy initiation are considered; 34.1% of the women included in the cohort had at least one first-degree relative with DM.
(3) Age: true if the patient was at least 25 years of age at pregnancy initiation.
(4) History of pregnancy complication: the logic odds ratio operation of the following markers:
    (A) History of GDM according to GCT and OGTT, defined similarly to the target.
    (B) History of miscarriage or stillbirth, seen in the form of a diagnosis with ICD9 632, 634.x, 635.x or 637.x.
    (C) History of a liveborn baby of birth weight >4 kg (note that birth weight is logged only for deliveries in Clalit-owned hospitals (about 30% of deliveries)).
(5) History of polycystic ovary syndrome (PCOS): true if the patient has at least one diagnosis of PCOS (ICD9 code 256.4). Only diagnoses available at pregnancy initiation were considered.
(6) Problems with insulin or blood sugar: true if the patient has at least one diagnosis of prediabetes, either according to ICD9 code 790.2x or following a HbA1c blood test in the range 5.7–6.4%. That test is not performed routinely in this population and therefore was available for only 13% of the cohort. Only diagnoses and tests available at pregnancy initiation were considered.
(7) High BP, high cholesterol and/or heart disease: the logic odds ratio operation of the following markers:
    (A) History of high BP, defined as two or more BP tests with systolic BP >140 or diastolic BP >90. Prepregnancy BP measurements were available for 82.9% of the cohort. Measurements taken during pregnancy were not included in this analysis.
    (B) Recorded relevant ICD9 of 401.x, 272.x or 390.x–449.x.

The final baseline risk score is, then, the number of 'true' entries in the above list, and therefore ranges between 0 and 7. An analysis of the odds ratio of the constructing variables, as well as comparison to a logistic regression model from the above binary variables, is presented in Extended Data Fig. 1. Of note, the logistic regression model does not substantially improve performance.

**Features and predictions.** We constructed 2,355 features from the dataset, of which 295 are available at the initiation of pregnancy with the remaining 2,060 generated from data gathered throughout the pregnancy, up to 20 weeks of gestation. The features available at the initiation of pregnancy include (1) demographics (for example, ethnicity), (2) basic measures (for example, age, weight and height) and medical history gathered before the current pregnancy, including (3) data from previous pregnancies and (4) data from non-pregnancy periods. Features gathered throughout the current pregnancy include (1) blood and urine laboratory tests, (2) ambulatory care clinic and hospital diagnoses, (3) anthropometrics and BP measurements and (4) pharmaceuticals prescribed and collected. A complete list of features, including methods for feature generation, is shown below. The percentage of feature availability per category is presented in Fig. 1b.

Predictions were generated using a gradient-boosting machine model[40] built with decision-tree base-learners. Gradient boosting is widely considered as state of the art in prediction for tabular data[41], and is used by many competition-winning algorithms in the field of machine learning[42,43]. As suggested by previous works[44], missing values were inherently handled by the gradient-boosting predictor[45]. We used the gradient-boosting predictor trained with the LightGBM[46] Python package. Hyperparameters were selected following a cross-validated grid search, with the following settings selected:

- num_boost_round = 603
- num_leaves = 20
- learning_rate = 0.05
- feature_fraction = 0.2
- bagging_fraction = 0.8
- bagging_freq = 5
- min_data_in_leaf = 4

For each of the 2,355 features, the following list describes the generation mechanism:

(1) Features available at pregnancy initiation ($n$ = 295):
    (A) Demographics (41 features):
        (i) Was the patient born in Israel (true/false)?
        (ii) Features describing ethnicity: 15 features breaking down the origin of the patient's ancestors, as logged in their country of origin. Countries were clustered into 14 categories, corresponding to Israel's major ethnic groups: North Africa, Iraq, Iran, Yemen, East Europe, West Europe, ex-USSR, North America, Latin America, Arab, Mediterranean, Ethiopia, Asia and Africa. Another feature logs the percentage of unknown origin.
        (iii) Socioeconomic data of the locality where the patient attended most clinic visits. Although personalized socioeconomic data were not available, we generated some estimates using the data made available by Israel's Central Bureau of Statistics[47]. Features include locality type (length 20, 1-hot vector) and locality religion breakdown (length 5, summing to one vector).

(B) Basic measures (seven features):
(i) Age at pregnancy initiation.
(ii) Weight, height and BMI: only samples available before the current pregnancy and outside past pregnancies were considered; the median for all samples from those aged 18 and above was calculated.
(iii) Systolic and diastolic BP: only samples available before the current pregnancy and outside past pregnancies were considered; the median for all samples from those aged 18 and above was calculated.
(iv) Number of children born in current pregnancy: one for single child, two for twins, and so on.

(C) Pregnancy history (103 features):
(i) History of GDM:
(a) Any history of GDM according to past pregnancy GCT and OGTT (true/false).
(b) GDM status in each of the last three pregnancies.
(ii) History of miscarriage: seen in the form of a diagnosis with ICD9 632, 634.x, 635.x or 637.x.
(iii) Largest baby weight: maximal birth weight recorded (note that birth weight was available for only 25% of the cohort).
(iv) Number of previous births: number of children born before the current pregnancy.
(v) Laboratory tests during last three pregnancies:
(a) Median values during each pregnancy of the following laboratory tests (the 25 most common—75 features).
(b) Median values during each pregnancy of fasting glucose and HbA1c% (six features).
(c) GCT and OGTT results, if available (15 features).

(D) Medical history outside of pregnancy (144 features):
(i) Number of first-degree relatives (parent or sibling) with at least one diagnosis of DM, defined as any of the ICD9 codes in 250.x or 357.2. Only diagnoses available at pregnancy initiation are considered.
(ii) History of PCOS, according to ICD9 code 256.4. Only diagnoses available at pregnancy initiation are considered.
(iii) History of prediabetes:
(a) Diagnoses: true if the patient has at least one diagnosis of prediabetes according to ICD9 code 790.2.x.
(b) Maximal HbA1c% logged.
(c) Joint prediabetes definition: either according to diagnosis or by a HbA1c test in the range 5.7–6.4%. Only diagnoses and tests available at pregnancy initiation are considered.
(iv) Features related to high BP, high cholesterol and/or heart disease:
(a) Number of high BP tests with systolic BP >140 or diastolic BP >90. BP measurements taken during pregnancy are not included in this analysis.
(b) Recorded relevant ICD9 of 401.x (hypertension), 272.x (high cholesterol) and 390.x–449.x (heart diseases) (three true/false features):
(v) Baseline risk score value.
(vi) Laboratory tests during the past 5 years (132 features): logging the median value in every window M1–M5 (see Time windows). We considered the 25 most commonly used tests, plus glucose and HbA1c%. We considered only data gathered outside of pregnancy periods for these features.
(vii) Coefficients ($n = 2$) of a linear regression for fasting glucose versus time (only if three or more measurements are available).

(2) Features gathered throughout current pregnancy ($n = 2{,}060$):

(A) Laboratory tests (524 features): median values of the 250 most commonly used laboratory tests during F0–F2 (see Time windows).

(B) Clinic and hospital diagnoses (906 features): counts of the 300 most common diagnoses in community clinics and the 10 most common diagnoses in hospitals, plus 'other' counts for all non-top diagnoses, for each window in F0–F2 (see Time windows).

(C) Anthropometrics and BP measurements (27 features):
(i) Medians of weight, height, BMI, systolic and diastolic BP and time interval between measurements of GCT, for each window in F0–F2 (see Time windows).
(ii) Coefficients (2) of a linear regression for weight versus time, for 10–20 weeks of gestation (only if three or more measurements are available).
(iii) Coefficients (2 × 2) of a linear regression for systolic/diastolic BP versus time, for 0–20 weeks of gestation (only if three or more measurements are available).

(D) Pharmaceuticals (603 features): counts of the 300 most common medications, plus 'other' count for all non-top medications, for each window in F0–F2 (see Time windows).

**Time windows.** The following time windows were defined for feature calculation:

(1) Windows during pregnancy were defined according to the usual medical examination pregnancy schedule for Israeli women[39]; this choice is supported by the test population in the data, as seen in Extended Data Fig. 7. We defined the following relative-time windows:

(A) F0: from 30 to 22 weeks before GCT, representing −4 to 4 weeks of gestation.
(B) F1: from 22 to 12 weeks before GCT, representing 4–14 weeks of gestation. This window includes the period in which women attend the first blood test during pregnancy, which is recommended at 6–12 weeks of gestation.
(C) F2: from 12 to 4 weeks before GCT, representing 14–22 weeks of gestation. This window includes the period in which women attend the second blood test during pregnancy (triple test), which is recommended at 16–18 weeks of gestation.

(2) For medical history outside pregnancy periods, we defined five 1-year windows covering the 5 years preceding the date of approximate gestation, named M1 (last year before pregnancy) to M5 (5 to 4 years before pregnancy).

(3) Past pregnancy periods are denoted by P1 (most recent pregnancy), P2 (pregnancy preceding P1) and P3 (pregnancy preceding P2). Pregnancies were situated according to the birth date of the child, and pregnancy period was defined as 40 weeks before that date plus 2.5 months in either direction, to cover randomization of birth dates.

**Model interpretations.** To understand how single features relate to the model output we used Shapley values[22], which are suited for complex models such as artificial neural networks and gradient-boosting machines[48]. Originating in game theory, Shapley values partition the prediction result of every sample into the contribution of each constituent feature value by estimating the difference between models with subsets of the feature space. By averaging over all samples, Shapley values estimate the contribution of each feature to the overall model predictions.

To draw dependence plots, we converted the resulting Shapley value to RR. In Shapley analysis, the log-odds (LO) of the predicted probability is calculated according to

$$\mathrm{LO} = \phi_0 + \phi_1 + \cdots + \phi_d$$

where $\phi_0$ is the 'base' Shapley value (the logit of the population prevalence, $P_0$), and $\phi_i$ for $i \in \{1,\ldots,d\}$ are the Shapley values related to features $1,\ldots,d$. The predicted probability based on a single feature is then

$$P_i = S(\phi_0 + \phi_1)$$

where

$$S(x) = \frac{1}{1 + e^{-x}}$$

is the sigmoid function, the inverse of the logit function. We therefore defined the relative risk related to a single feature and sample as

$$\mathrm{RR}_i = \frac{P_i}{P_0} = \frac{S(\phi_0 + \phi_i)}{S(\phi_0)}$$

For a set $D = \{i,j,\ldots\}$ of features, this definition extends to

$$\mathrm{RR}_i = \frac{P_i}{P_0} = \frac{S(\phi_0 + \sum_{i \in D} \phi_i)}{S(\phi_0)}$$

To plot the dependence plot, we calculated mean and standard deviations of the RR for each bin of feature value, and presented those versus the mean feature value. This resembles a standard dependence plot, only with RR rather than the Shapley values presented.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The data that support the findings of this study originate from Clalit Health Services. Restrictions apply to the availability of these data and they are therefore not publicly available. Due to restrictions, these data can be accessed only by request to the authors and/or Clalit Health Services.

## Code availability

The code that supports the findings of this study is tailored to the data and the fields of the Clalit Health Services database, and is thus not provided since it is of no use as a standalone without access to the data per se. The algorithmic models used the standard Python code package scikit-learn, which is publicly available.

## References


37. *Data*. Clalit Research Institute; http://clalitresearch.org/about-us/our-data/ (accessed 23 July, 2019).
38. Vandorsten, J. P. et al. NIH consensus development conference: diagnosing gestational diabetes mellitus. *NIH Consens. State Sci. Statements* **29**, 1–31 (2013).
39. State of Isreal Ministry of Health. *Monitoring of Pregnancy and Medical Examinations During Pregnancy* https://www.health.gov.il/English/Topics/Pregnancy/during/examination/Pages/permanent.aspx (accessed 23 July, 2019).
40. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).
41. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
42. Omar, K. *XGBoost and LGBM for Porto Seguro's Kaggle Challenge: A Comparison Semester Project* (ETH, 2018).
43. Biendata Competitions. *KDD Cup of Fresh Air* https://biendata.com/competition/kdd_2018/winners/ (accessed 23 July 2019).
44. Josse, J., Prost, N., Scornet, E. & Varoquaux, G. On the consistency of supervised learning with missing values. Preprint at arXiv https://arxiv.org/abs/1902.06931 (2019).
45. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* - KDD 2016 (eds Krishnapuram, B. et al.) 785–794 (ACM Press, 2016).
46. Ke, G. et al. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree* https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf (2017).
47. CBS. *Regional Statistics Section* https://www.cbs.gov.il/EN/settlements/Pages/default.aspx?mode=Yeshuv (accessed 10 July 2018).
48. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).


## Acknowledgements

We thank G. Barabash, E. Barkan, I. Kalka and members of the Segal group for discussions. E.S. is supported by the Crown Human Genome Center, by D. L. Schwarz, J. N. Halpern and L. Steinberg, and by grants funded by the European Research Council and the Israel Science Foundation.

## Author contributions

N.S.A., S.S. and E.H. conceived the project, designed and conducted the analyses, interpreted the results and wrote the manuscript, and are listed in random order. H.R. conducted the analyses and wrote the manuscript. S.B.-H., A.B.-H., R.D.B. and B.F. interpreted the results. A.W. and E.S. conceived and directed the project and analyses, designed the analyses, interpreted the results, wrote the manuscript and supervised the project.

## Competing interests

The authors declare no competing interests.

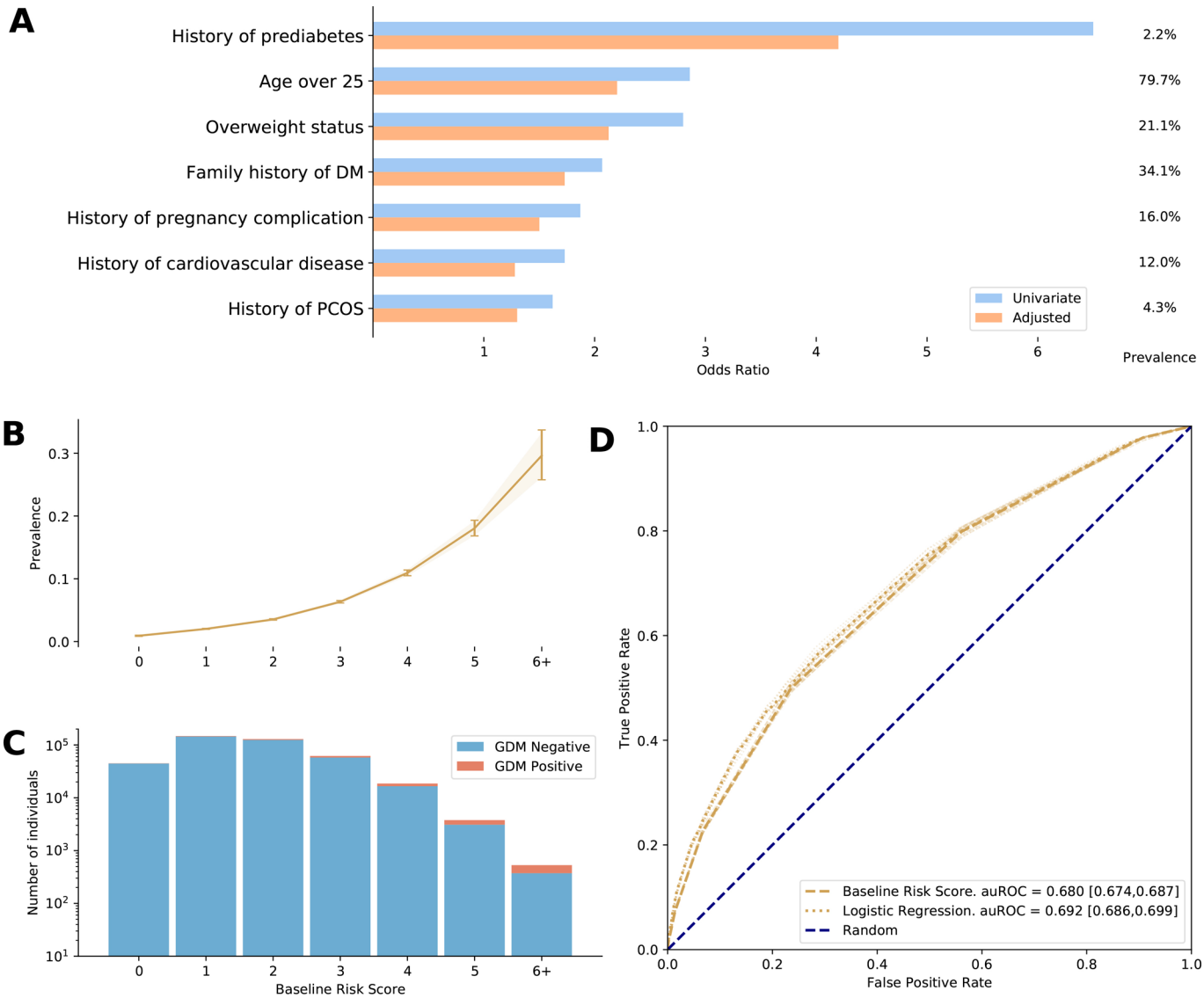
## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41591-019-0724-8.

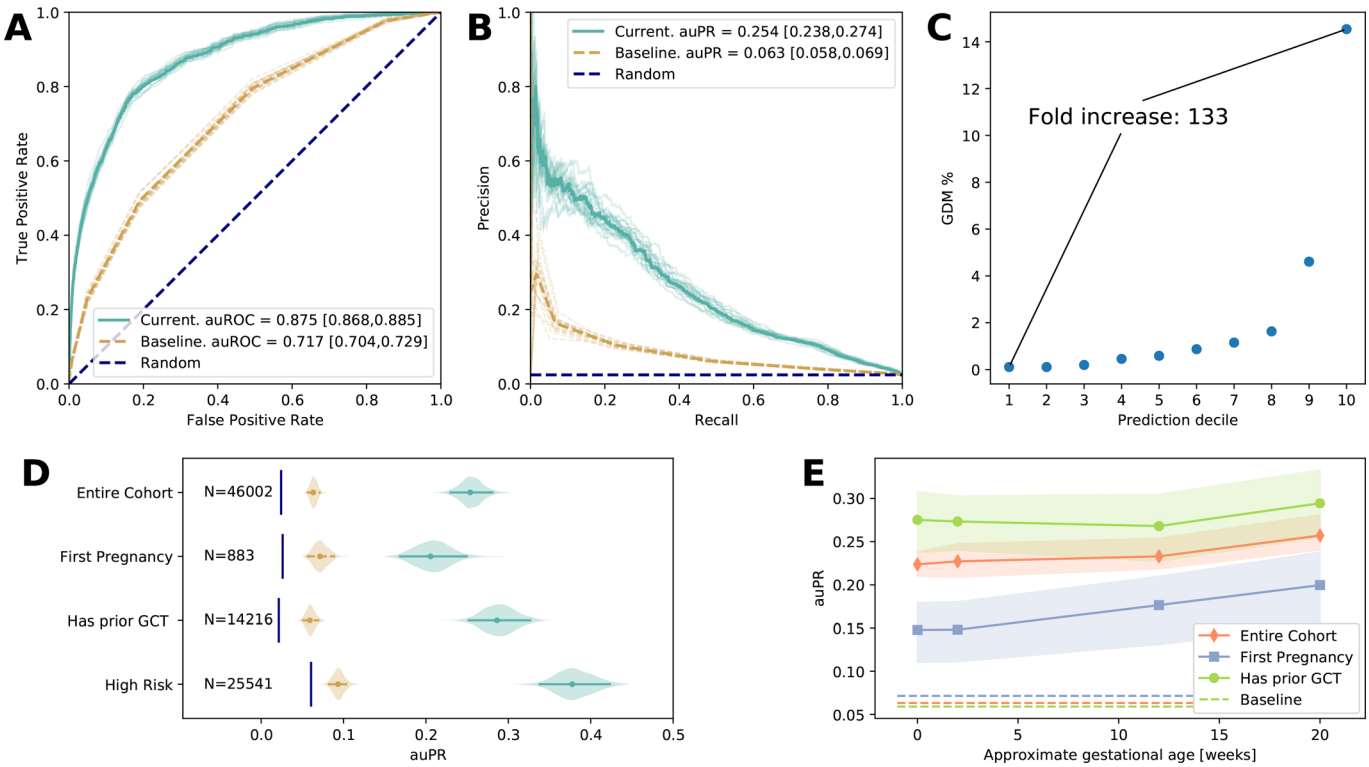
**Supplementary information** is available for this paper at https://doi.org/10.1038/s41591-019-0724-8.

**Correspondence and requests for materials** should be addressed to A.W. or E.S.

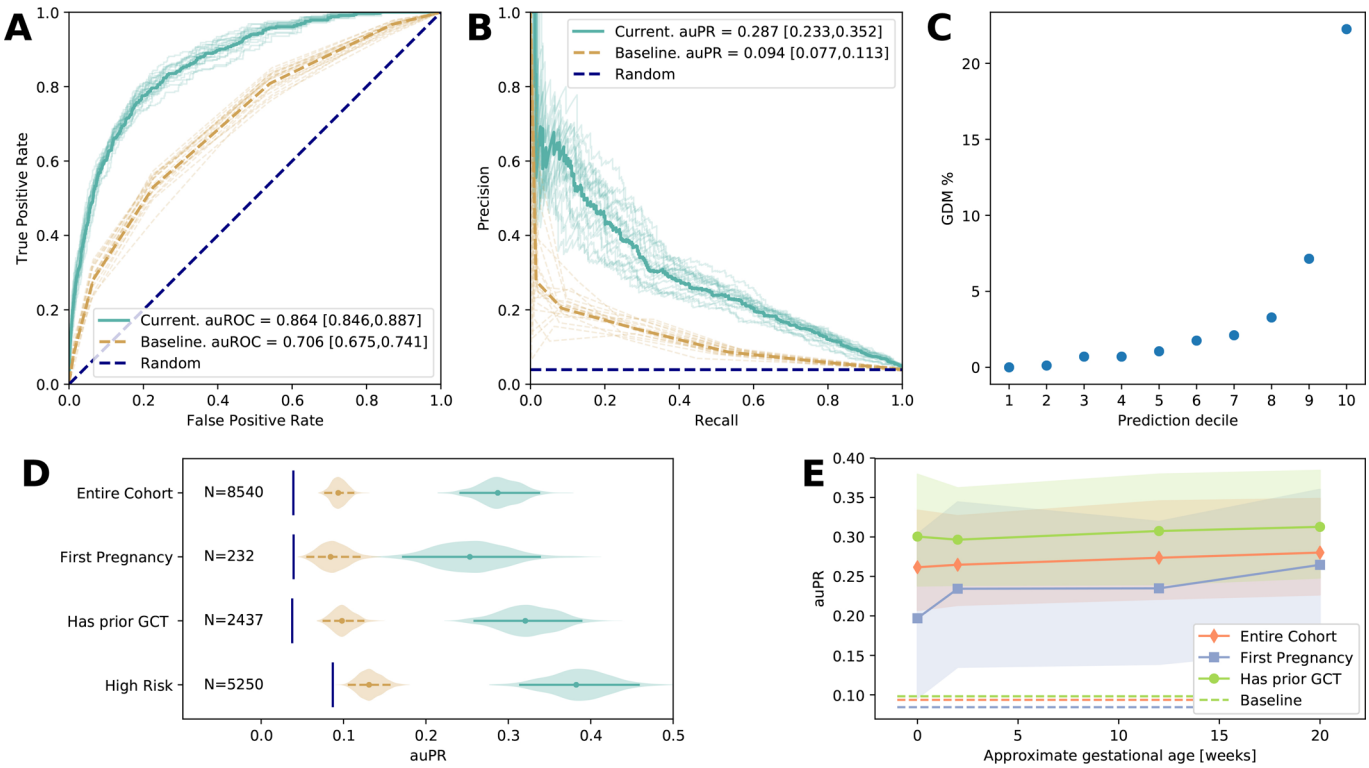
**Peer review information** Joao Monteiro was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.




**Extended Data Fig. 1 | Baseline prediction, based on *Baseline Risk Score*. a**: Odds ratio for the risk score composing parameters. Adjusted odds ratios were derived from a logistic regression model, both values are presented on the training set. **b**: Prevalence among women grouped by risk score. Error bars represent 90% confidence intervals on the train set. **c**: Histogram of risk scores in the training set. **d**: ROC curve for NIH Risk Score and for a logistic regression model trained on its constructing parameters. Results are reported on the future validation set. Logistic regression model does not suppress the Naive summation in the risk score. (n = 82,678 for all panels).



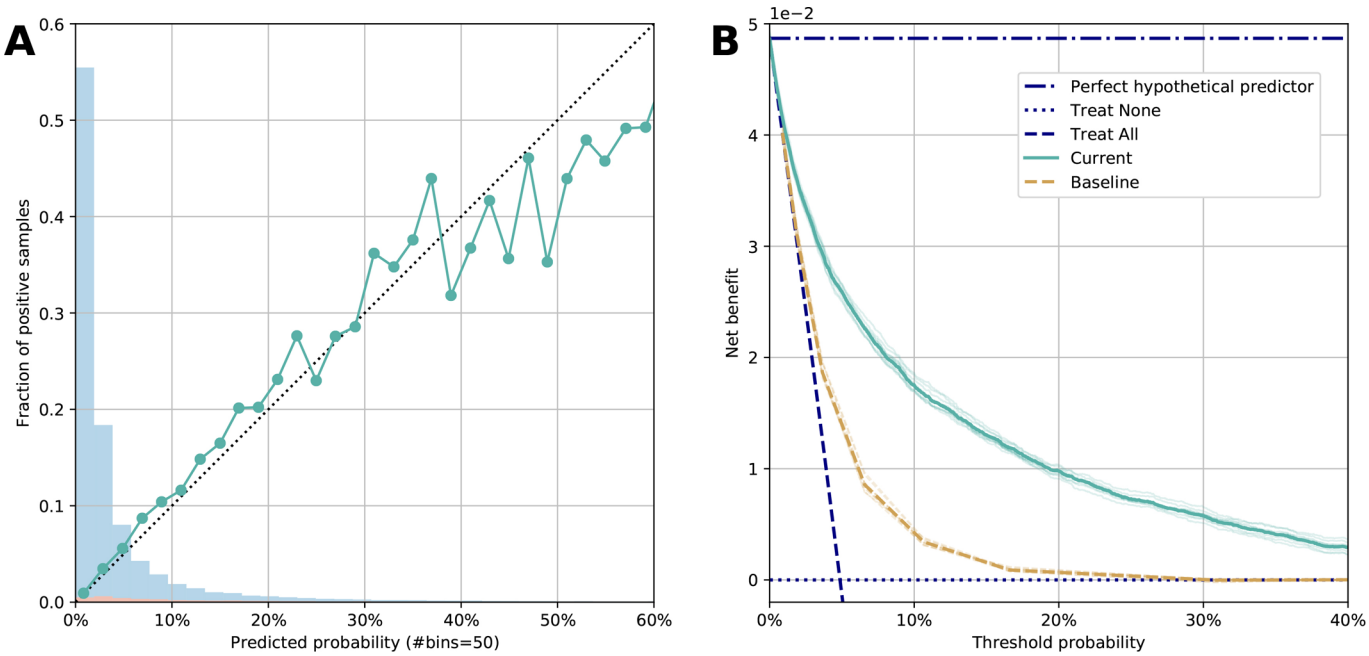

**Extended Data Fig. 2 | Evaluation of the model on the geographical validation set. a**: Receiver Operating Characteristic (ROC) curve, comparing our model (solid) and the Baseline Risk Score (dashed). Lighter colored lines are ROC curves of stratified partition of the validation set (not shown in ROC); bracketed values are 95% confidence intervals calculated through a normal fit of those curves. **b**: Precision-Recall (PR) curve, with the same properties as in A. **c**: The fraction of GDM-positive samples in every decile of the predicted probability. **d**: Predictions on different subsets of the cohort. auPR is shown for each subset, for our model (blue) and the baseline score (orange). Error bars show 95% confidence intervals, and dark blue lines show the prevalence in each subset. Shaded area is the distribution of the relevant score. **e**: Performance by gestational age at prediction. Every point is the evaluation score of a model built only with features available at this time point. (n = 46,002 for panels A-C. Subset sample sizes are listed in panel D).



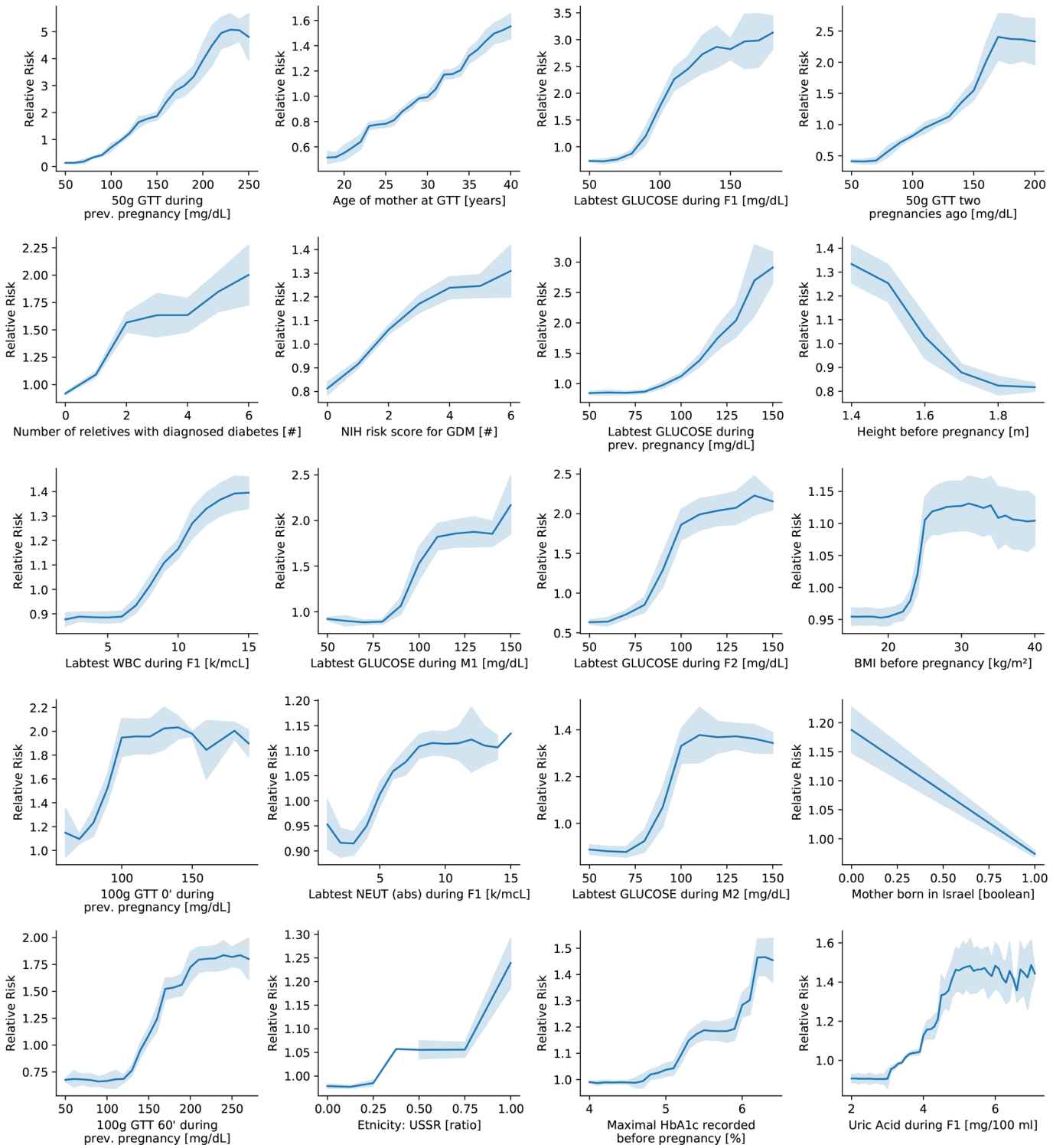

**Extended Data Fig. 3 | Evaluation of the model on the geo-temporal validation set. a**: Receiver Operating Characteristic (ROC) curve, comparing our model (solid) and the Baseline Risk Score (dashed). Lighter colored lines are ROC curves of stratified partition of the validation set; bracketed values are 95% confidence intervals calculated through a normal fit of those curves. **b**: Precision-Recall (PR) curve, with the same properties as in A. **c**: The fraction of GDM-positive samples in every decile of the predicted probability. **d**: Predictions on different subsets of the cohort. auPR is shown for each subset, for our model (blue) and the baseline score (orange). Error bars show 95% confidence intervals, and dark blue lines show the prevalence in each subset. Shaded area is the distribution of the relevant score. **e**: Performance by gestational age at prediction. Every point is the evaluation score of a model built only with features available at this time point. (n = 8,540 for panels A-C. Subset sample sizes are listed in panel D).

| Evaluation results in different validation sets | | | | | | |
|---|---|---|---|---|---|---|
| | Future validation set (n = 82,678) | | Geographical validation set (n = 46,002) | | Geo-temporal validation set (n = 8,540) | |
| | Full Model | Baseline RS | Full Model | Baseline RS | Full Model | Baseline RS |
| auROC | 0.851 (0.847-0.855) | 0.680 (0.675-0.686) | 0.875 (0.866-0.885) | 0.717 (0.708-0.728) | 0.864 (0.856-0.880) | 0.706 (0.683-0.732) |
| auPR | 0.318 (0.307-0.329) | 0.097 (0.095-0.100) | 0.254 (0.241-0.270) | 0.063 (0.061-0.068) | 0.287 (0.248-0.341) | 0.094 (0.081-0.111) |
| %GDM is lowest decile | 0.22% | | 0.11% | | 0% | |
| %GDM in highest decile | 25.6% | | 14.6% | | 22.2% | |
| PPV@TPR= 10% | 62.5% | | 53.3% | | 60.7% | |

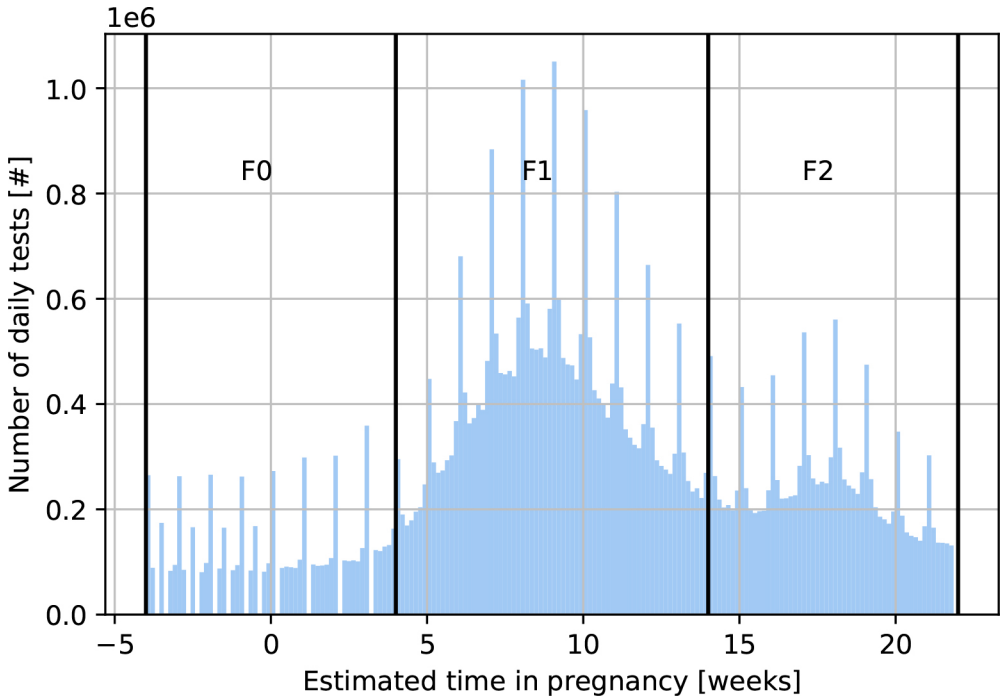
**Extended Data Fig. 4 |** Evaluation results in different validation sets.




**Extended Data Fig. 5 | Basic utility of the predictor. a**: Calibration curve, showing the fraction of positive samples per bin versus the mean predicted probability of the bin. Blue and red bars represent the ratio of negative/positive samples in the bin, respectively. **b**: Decision curve, showing the net benefit versus the threshold probability, for both predictor and baseline. The predictor outperforms the baseline at all thresholds. (n = 82,678 for all panels).




**Extended Data Fig. 6 | Additional dependence plots.** Top 20 features are shown (ordered left to right, top to bottom). In each the mean predicted relative risk is plotted versus feature value. Bands represent SD area of the population per bin, which is connected to interactions between input features. (n = 82,678).




**Extended Data Fig. 7 | Histogram of lab tests during pregnancy, showing the window definition of F0, F1 and F2.** The peaks showing are weekly, and represents the fact that patients tend to see a doctor in the same day of the week.

nature research

Corresponding author(s): Prof Eran Segal

Last updated by author(s): 19/11/2019

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Retrospective Electronic health records data originates from Clalit healthcare |
| Data analysis | Data were analyzed in Python 3.5, using packages Scikit-learn 0.12, LightGBM 2.2 and SHAP 0.29. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study originates from Clalit healthcare but restrictions apply to the availability of these data and so are not publicly available. Due to restrictions it can only be accessed through requests from the authors and/or the Clalit healthcare organisation

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | No sample size calculation was preformed since sample size was determined by the number of pregnancies identified in the electronic health records. Train and tested on this sample size, our models predict GDM with high accuracy (AUC=0.84) |
| Data exclusions | Exclusion criteria were pre-established and included Pre-gestational diabetes and missing values of oral glucose tolerance test required for GDM diagnosis |
| Replication | We used cross-validation on the training set, and resampling from the validation sets to demonstrate a replication of our results. All attempts were successful |
| Randomization | Randomization was not applicable in this study since we analysed existing retrospective Electronic health records |
| Blinding | Blinding was not applicable in this study since we analyzed existing anonymized retrospective Electronic health records |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |
| ☐ | ☒ Clinical data |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | We included a total of 588,622 pregnancies from 368,351 women who gave birth between 2010-2017 in our cohort. The prevalence of GDM diagnosed by a two-step diagnostic test composed of a glucose challenge test (GCT) and an oral glucose tolerance test (OGTT) during 24-28 weeks of gestation, was 3.9%. Mean age at pregnancy initiation was 29.1, Mean BMI was 23.5 |
| Recruitment | The study is based on retrospective data originating from a non-governmental, non-profit organization which includes the majority of the Israeli population, and the outcome of the model is based on routine pregnancy tests that are comprehensively documented in the EHR, so the risk of selection bias is low |
| Ethics oversight | The study protocol was approved by Rabin Medical Center Institutional Review Board (IRB) |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

| | |
|---|---|
| Clinical trial registration | n/a |
| Study protocol | n/a |
| Data collection | Retrospective clinical data from Electronic health records |
| Outcomes | GDM development |