Prediction of gestational diabetes based on nationwide electronic health records

Document Version:
Accepted author manuscript (peer-reviewed)

Citation for published version:
Artzi, NS, Shilo, S, Hadar, E, Rossman, H, Barbash-Hazan, S, Ben-Haroush, A, Balicer, RD, Feldman, B, Wiznitzer, A & Segal, E 2020, 'Prediction of gestational diabetes based on nationwide electronic health records', *Nature Medicine*, vol. 26, no. 1, pp. 71-76. https://doi.org/10.1038/s41591-019-0724-8

Total number of authors:
10

Digital Object Identifier (DOI):
10.1038/s41591-019-0724-8

Published In:
Nature Medicine

Prediction of Gestational Diabetes Based On Nationwide Electronic Health Records

N.S. Artzi, S. Shilo, E. Hadar, H. Rossman, S. Barbash-Hazan, A. Ben-Haroush, R.D. Balicer, B. Feldman, A. Wiznitzer, E. Segal.

ABSTRACT

BACKGROUND

Gestational diabetes mellitus (GDM) poses increased risk of short and long term complications for both mother and offspring. GDM is typically diagnosed between 24-28 weeks of gestation, yet earlier detection is desirable as it may prevent or considerably reduce the risk of adverse pregnancy outcomes.

METHODS

A computational model was constructed for predicting GDM at both pregnancy initiation and other early pregnancy stages based on electronic health records from Israel's largest health services provider. GDM was 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. 588,622 pregnant women who gave birth from 2010 to 2018 were included in the analysis.

RESULTS

The model predicted GDM with an area under the receiver operating characteristic curve (auROC) and 95% CI of 0.836 [0.828-0.843] at pregnancy initiation, significantly outperforming a baseline risk score whose auROC is 0.692 [0.669-0.696]. Predictions at week 20 achieved auROC of 0.854 [0.846-0.861]. Previous pregnancy GCT results (regardless of GDM status in previous pregnancies), age and first trimester fasting glucose contributed most to the model. A simpler model based on only 9 questions that a patient can answer achieved an auROC of 0.801 [0.787-0.815].

CONCLUSION

GDM can be accurately predicted at pregnancy initiation, even with a simple 9 questions model. The models devised can prioritize high-risk individuals for early-stage interventions aimed at preventing GDM and its adverse health outcomes, and enables a selective cost-effective screening approach by identification of low-risk patients.

INTRODUCTION

Gestational diabetes mellitus (GDM) is defined as glucose intolerance that is first recognized in pregnancy. GDM is a common complication of pregnancy, occurring in 3%-9% of pregnancies[1], typically diagnosed between 24-28 weeks of gestation[2]. GDM is associated with short and long term clinical outcomes, affecting both mothers and infants. Women with GDM are predisposed to many comorbidities including operative delivery and type 2 diabetes[3]. Offsprings of diabetic mothers are prone to adverse health outcomes such as fetal macrosomia, respiratory difficulties and metabolic complications in the neonatal period and have a higher risk for future obesity and alteration in glucose metabolism[4–6].

The rising prevalence of GDM, reflective of type 2 diabetes prevalence, warrant the development of new prevention strategies[7]. Although results from randomized controlled trials aimed at prevention of GDM with nutritional and lifestyle interventions are conflicting[8], some studies demonstrated that a major reduction of the risk is possible, especially when interventions initiate during the first or early second trimesters[9,10]. Identifying women at high GDM risk at an early stage of pregnancy will therefore enable implementation of early intervention strategies, which may prevent or reduce GDM prevalence and its associated comorbidities.

Several studies 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.

METHODS

DATA

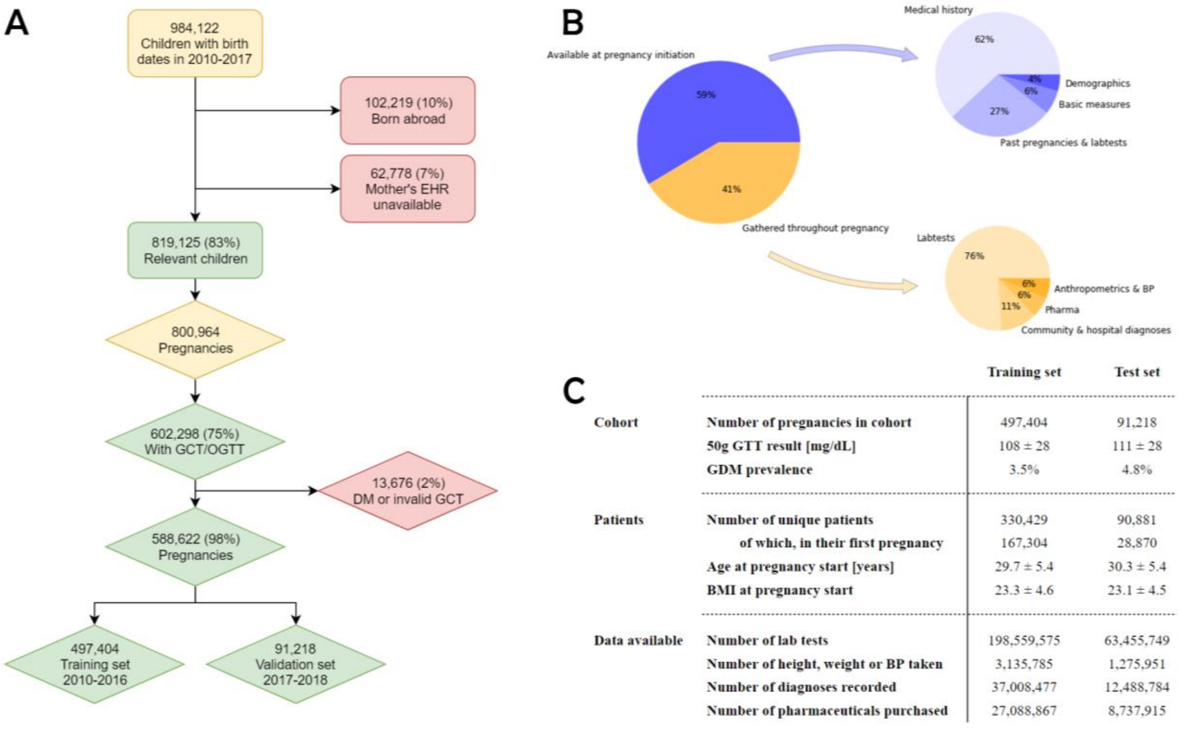
Data were extracted from the database of Clalit Health Service, the largest healthcare provider in Israel. Nearly five million individuals, representing over 50% of Israel's adult population, are currently enrolled in Clalit[19], a non-governmental, non-profit organization included in the national health insurance law in Israel. Dating back to 2002, the database contains EHRs of over 11 million patients, with over 5.4 billion numerical and categorical entries. The data analyzed included anthropometrics (height and weight), blood pressure measurements, blood and urine lab tests, diagnoses recorded by physicians, and pharmaceuticals prescribed and issued. Most of the data originates from community clinics records, but records from Clalit's 14 hospitals were also included in the analysis.

STUDY POPULATION AND TARGET DEFINITION

In Israel, GDM is diagnosed by a two-step procedure, which is performed routinely to all pregnant women during 24-28 weeks of pregnancy, in accordance with National Institutes of Health (NIH) guidelines[20]. In the first step, a 1 hour, 50g, glucose challenge test (GCT) is performed; women with glucose levels higher than 200 mg/dL receive a GDM diagnosis. Women with GCT value above 140 mg/dL are referred to the second step, in which an additional 100g, 3 hours oral glucose tolerance test (OGTT) is performed. Women with two glucose measurements above the thresholds of 95, 180, 155 and 140 mg/dL under fasting conditions, one, two and three hours after glucose intake, respectively, also receive a GDM diagnosis[2,21]. Note that although these two tests are similar in nature and sometimes denoted simply “50g GTT” and “100g GTT”, we used the “GCT” and “OGTT” notation for simplicity.

Accordingly, we defined GDM status based on GCT and OGTT results. In cases in which more than one test was available, we defined a GDM diagnosis if at least one of the tests was positive. We excluded women who were supposed to undergo a 100g OGTT due to a screen positive GCT, but had no record of the test results. Women with pre-pregnancy record of diabetes determined by a pre-pregnancy HbA1c blood test above 6.4% or a diabetes diagnosis were also excluded. In total, 588,622 pregnancies from 368,351 women were included in the cohort (**Figure 1**, Appendix 1).

In order to construct a computational model, the study population was split prior to any analysis of the data into a training set and a validation set. To emulate practical use, we defined the test set according to a temporal validation scheme[22]. Pregnancies that ended during 2017 or 2018 composed the test set, and pregnancies that ended before December 31, 2016 composed the training set. This choice thus represents a setting in which the model may be implemented in practice. Throughout this work, all results are reported on the test set, unless stated otherwise.



		Training set	Test set
Cohort	**Number of pregnancies in cohort**	497,404	91,218
	50g GTT result [mg/dL]	108 ± 28	111 ± 28
	GDM prevalence	3.5%	4.8%
Patients	**Number of unique patients**	330,429	90,881
	of which, in their first pregnancy	167,304	28,870
	Age at pregnancy start [years]	29.7 ± 5.4	30.3 ± 5.4
	BMI at pregnancy start	23.3 ± 4.6	23.1 ± 4.5
Data available	**Number of lab tests**	198,559,575	63,455,749
	Number of height, weight or BP taken	3,135,785	1,275,951
	Number of diagnoses recorded	37,008,477	12,488,784
	Number of pharmaceuticals purchased	27,088,867	8,737,915

Figure 1: Data and cohort characteristics. **A**: Cohort selection. Pregnancies were first identified by the offspring’s birth date. Second, women with pre-existing diabetes, pregnancies without a record of glucose testing (50g or 100g) and those with missing OGTT were excluded. Finally, the cohort was divided into training and validation sets (see

Methods). **B**: Feature availability distribution. Pies are divided according to the sum of datapoints in each feature set. A significant part of data originates from lab tests results during current or previous pregnancies. **C**: Basic characteristics of the cohort data. Numbers of data items (lab tests, diagnoses etc.) before the patients performed GCT during pregnancy are presented.

Abbreviations: BP - Blood Pressure, DM - Diabetes Mellitus, GDM - Gestational Diabetes Mellitus, GCT - Glucose Challenge Test, OGTT - Oral Glucose Tolerance Test

BASELINE RISK SCORE

Currently there are no validated GDM prediction tools that are employed in clinical practice. A baseline risk score was established according to a self-administered 8-points questionnaire recommended by the National Institute of Health (NIH)[23] calculating a close proxy to this score for every woman in our cohort, denoted here as their *Baseline Risk Score* (See Appendix 2). An analysis of the features and their predictive power is described in **Supp. Figure 1**.

FEATURES AND PREDICTIONS

We constructed 2355 features from the dataset, of which 295 are available at the initiation of pregnancy, and the remaining 2060 are generated from data gathered throughout the pregnancy, up to 20 weeks of gestation. The features available at the initiation of pregnancy include (1) demographics (e.g., ethnicity), (2) basic measures (e.g., age, weight, height), and medical history gathered prior to the current pregnancy, including data on (3) previous pregnancies and (4) data from non-pregnancy periods. Features gathered throughout the current pregnancy include (1) blood and urine lab tests, (2) ambulatory care clinic and hospital diagnoses, (3) anthropometrics and blood-pressure measurements, and (4) pharmaceuticals prescribed and collected. A complete list of the features, including methods for feature generation are available in Appendix 3. The percentage of feature availability per category is presented in **Figure 1B**.

Predictions were generated using a Gradient Boosting Machine (GBM) model[24], built with decision-tree base-learners. Gradient boosting is widely considered as state-of-the-art in prediction for tabular data[25], and used by many competition-winning algorithms in the field of machine learning[26,27]. We used cross-validation among the training set to set hyperparameters. Cross-validation results and exact hyperparameters values are available in Appendix 4.

MODEL INTERPRETATIONS

To understand how single features relate to the model's output we used Shapley values[28], as it is suited for complex models such as artificial neural networks and gradient boosting machines[29]. 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.

RESULTS

GDM PREDICTION MODEL

We first established a baseline, termed *Baseline Risk Score*, defined as the summation of seven binary variables that the NIH recommends to use as GDM risk factors (see Methods). Odds ratios for these seven parameters are presented in **Supp. Figure 1**. As expected, odds ratios for all the parameters are greater than one (1.28-3.92), consistent with their classification as risk factors, and the risk score is predictive of GDM status (**Supp. Figures 1B-D**). The highest precision achieved by this score is 30%, and its area under the receiver operating characteristic curve (auROC) is 0.682.

To evaluate whether EHR-derived information may improve GDM prediction, we compiled a set of 2355 features (see Methods). We then used these to train a gradient boosting model to predict the probability of each held-out individual (that is, samples not included in the training set) to develop GDM. This EHR-based model achieved an area under the receiver operating characteristic curve (auROC) of 0.854 and area under the Precision-Recall curve (auPR) of 0.318, compared to an auROC of 0.682 and auPR of 0.097 by the baseline risk score (**Figures 2A and 2B**). The model provides a 287-fold enrichment between the lowest and highest risk deciles, consistent with the predicted probabilities (**Figure 2C**).

We next examined whether the predictions differ in accuracy for different subsets of the population, consisting of (1) First pregnancy: women with no previous record of pregnancy; (2) Has prior GCT: women who have a record of a GCT from a previous pregnancy; and (3) High risk: women with Baseline Risk Score greater than 2. Across all subgroups, our EHR-based model had higher auROC and auPR values than the baseline model (**Figure 2D**).

Finally, we evaluated the ability to predict GDM at different weeks of gestation, by constructing models based only on data collected prior to that week. The results of this analysis show that although prediction improves by incorporating features gathered as pregnancy progresses, predictions at pregnancy initiation still outperform the baseline model by 2-3 folds of auPR. This effect is stronger for women in their 2nd pregnancy onwards (**Figure 2E**). Hereinafter, when we refer to “our model” we refer to the model evaluated at week 20 of pregnancy as it contains all features; of note, we achieved similar results for other models earlier in pregnancy.

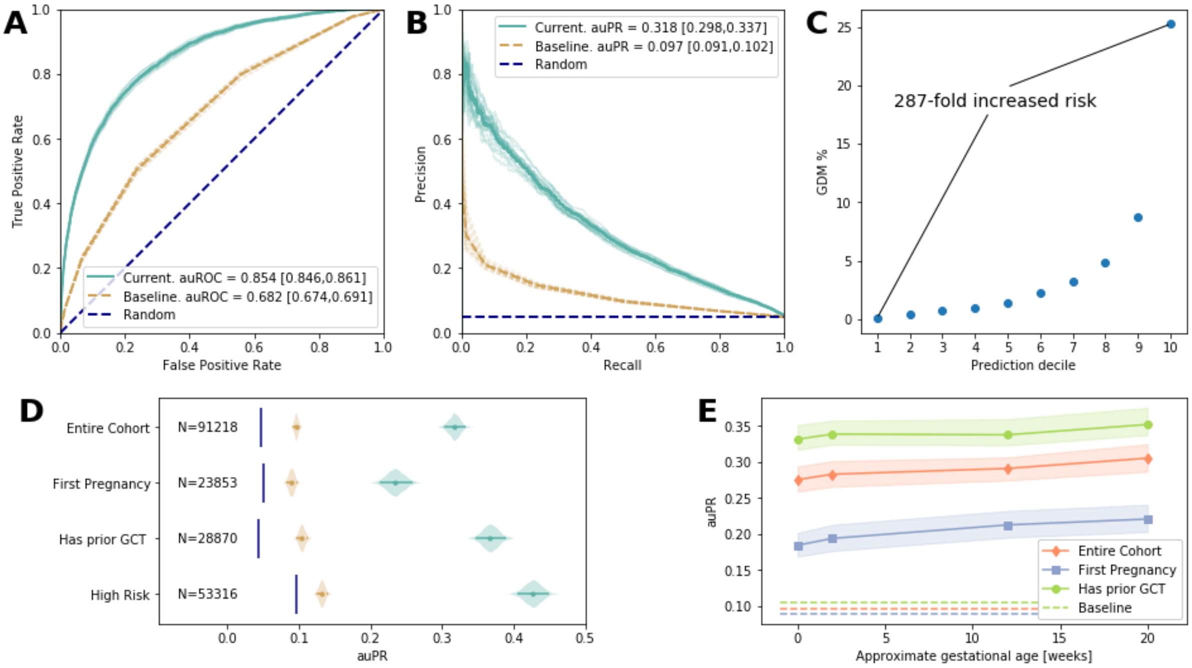


Figure 2: Predictive model evaluation. **A**: Receiver Operating Characteristic (ROC) curve, comparing our model (solid) and the Baseline Risk Score (dashed). Lighter colored lines are ROC curves of stratified partition of the test set (not shown in ROC); bracketed values are 95% confidence intervals 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, showing a 287-fold enrichment between the highest and lowest deciles . The mean predicted probability of each decile is also shown, demonstrating the calibration of the model. **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 prior record of GCT is more effective than those in their first pregnancy, and predictions for women with a higher baseline risk score (>2) have the greatest deviation from the baseline risk score model. Error bars show 95% confidence intervals, and dark blue lines show the prevalence in each subset. **E**: Time-dependent analysis. Every point is the evaluation score of a model built only with features available at this time point.

Abbreviations: auPR/auROC - Area under the PR/ROC curve, GCT - Glucose Challenge Test, GDM - Gestational Diabetes Mellitus, PR - Precision-Recall, ROC - Receiver-Operator-Characteristic,,

MODEL INTERPRETATION

To gain insight into the features that contribute most to the model predictions, we used the feature attribution framework of Shapley values[28] (see Methods). Shapley analysis (**Figure 3A**) identified the most predictive feature for GDM diagnosis to be the GCT result in the previous pregnancy, followed by maternal age and fasting blood glucose in the first trimester.

To further explore the importance of a 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 the GCT in the previous pregnancy (**Figure 1F**). This analysis revealed that the GCT result in the previous pregnancy is more predictive than GDM diagnosis in previous pregnancy. For example, a patient with a GCT of 180 mg/dL will be in higher risk of GDM in her next pregnancy irrespective of whether she will be diagnosed with GDM after the 100g OGTT. On the other hand, a patient with a glucose level below 75 mg/dl after a GCT will have a GDM risk in her next pregnancy as low as 1/5 of the population prevalence.

The additive nature of the Shapley values allows construction of a feature importance score for feature sets, by summing of Shapley values per set. The results of this analysis (for the sets defined in Methods) are presented in Figure 3B.

We further used Shapley values[28] to build *Dependence Plots* that capture the non-linear 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 (Appendix 5). We examined dependence plots for two well known risk factors for GDM: pre-pregnancy maternal BMI[30], and the number of relatives diagnosed with diabetes mellitus (DM)[31]. For pre-pregnancy BMI, the RR for GDM starts to increase above 21, reaches above 1 in BMI values above 24, and plateaus at 1.2 in BMI above 30 (**Figure 3C**). As expected, the RR for GDM increases as the number of the first degree family members with GDM increases, reaching a RR of 1.8 in women with 6 relatives diagnosed with DM (**Figure 3D**). Analysis of pre-gestation HbA1C revealed an increase of the RR for GDM with an increase in the pregestational HbA1C, even in values that are considered to be within the normal range (less than 5.7%). A steeper increase in the RR occurs at HbA1C> 5.9% (**Figure 3E**).

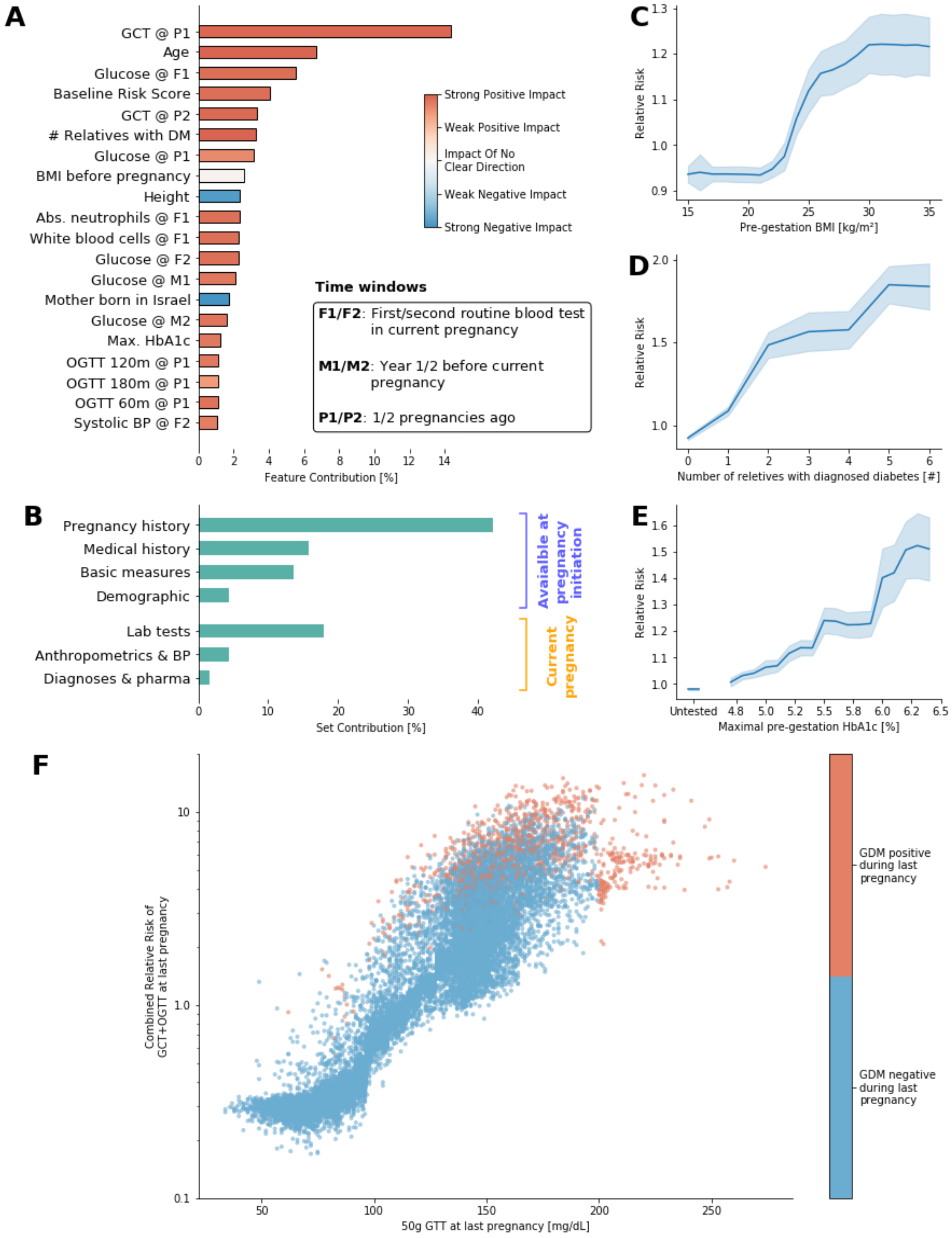


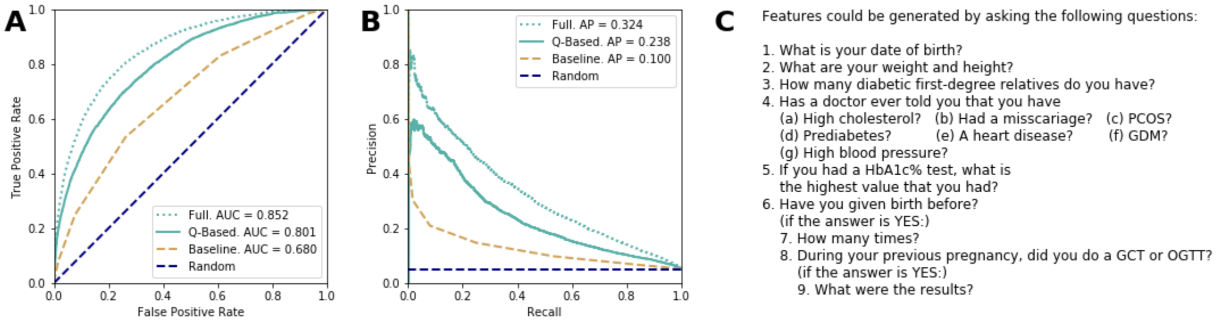
Figure 3: Shapley values based interpretation of the model. **A**: Feature importance of the top 20 contributing features. Bar colors indicate direction of influence, based on the Dependence Plot of this feature. **B**: Contributing feature category analysis. Shapley values were summed for each feature 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, number of first-degree relatives with diagnosed diabetes mellitus (DM) and pre-gestation HbA1C% blood test. Bands represent SD of the population per bin, which is connected 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. **F**: The combined Shapley Value for all GCT/OGTT results during last pregnancy is plotted against GCT value during last pregnancy. Every point is a sample, and it is colored according to GDM status during the previous pregnancy. While GDM in a previous pregnancy predicts GDM in the current one, GCT provides a continuous predictions for all women who took the test before. Additionally, GCT in previous pregnancy also points to women who are in a GDM risk as low as 1/5 of the population risk.

Abbreviations: BP - Blood Pressure, DM - Diabetes Mellitus, GCT - Glucose Challenge Test, GDM - Gestational Diabetes Mellitus, OGTT - Oral Glucose Tolerance Test, SD - Standard Deviation

SIMPLE PREDICTION MODEL

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 over 2000 EHR features. To this end, we evaluated the performance of a model based on 9 simple questions that a patient can answer herself. This predictor achieves an auROC of 0.801 and auPR of 0.238, compared to 0.680 and 0.100, respectively, for the baseline model (**Figure 4**).



Features could be generated by asking the following questions:

1. What is your date of birth?
2. What are your weight and height?
3. How many diabetic first-degree relatives do you have?
4. Has a doctor ever told you that you have
   (a) High cholesterol? (b) Had a misscariage? (c) PCOS?
   (d) Prediabetes? (e) A heart disease? (f) GDM?
   (g) High blood pressure?
5. If you had a HbA1c% test, what is
   the highest value that you had?
6. Have you given birth before?
   (if the answer is YES:)
   7. How many times?
   8. During your previous pregnancy, did you do a GCT or OGTT?
      (if the answer is YES:)
      9. What were the results?

Figure 4: **Questionnaire-based prediction**. **A-B**: Validation results of a questionnaire-based predictor, using 9 simple questions. Both ROC (A) and PR (B) curves of the model are shown. **C**: The list of questions that assemble the predictor.

Abbreviations: GCT - Glucose Challenge Test, GDM - Gestational Diabetes Mellitus, OGTT - Oral Glucose Tolerance Test, PCOS - Polycystic Ovary Syndrome.

Efficiency of our predictor as a GDM screening tool

We next analyzed if we can emulate usage of our predictor as a screening tool for identifying women who are less likely to develop GDM. In this approach, for a given threshold, only part of the population will undergo the usual two-steps GCT/OGTT diagnosis process. We therefore analyzed the ratio of women who avoid testing versus the predictor miss rate, i.e., the percentage of GDM positive patients that would not be diagnosed following this procedure (**Figure 5B**). The steep increase in the resulting curve indicates that a large fraction of the population can avoid taking the test. For example, if we permit 20% of diagnoses to be missed - which is in the reported scale of missed diagnoses from GCT[32,33] - then 79% of all women that have a GCT result in their previous pregnancy can avoid the test in their next pregnancy.

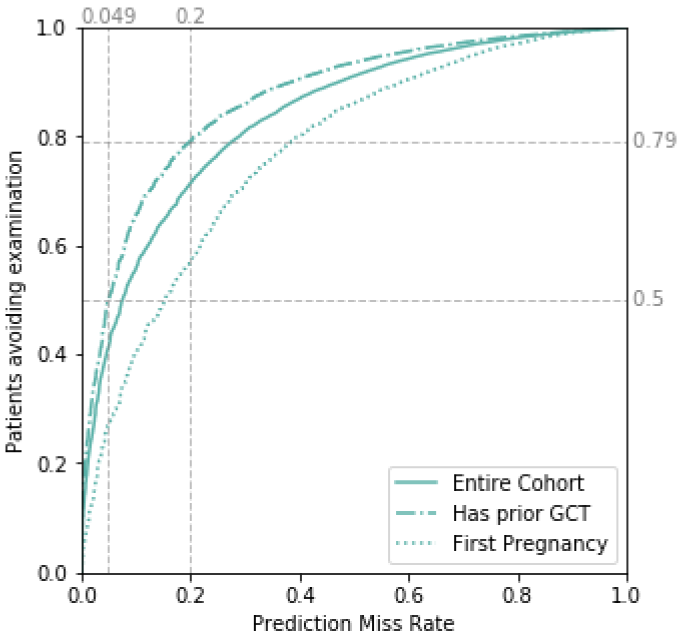


Figure 5: **Efficiency of our predictor as a GDM screening tool**. We consider avoiding the examination of low-risk patients, while retaining current system of GCT/OGTT for all others. The ratio of avoided GCTs is plotted versus the predictor miss rate, that is the percentage of GDM positive patients that would not be diagnosed following this procedure.

Abbreviations: GCT - Glucose Challenge Test, GDM - Gestational Diabetes Mellitus

DISCUSSION

In this study, we examined the ability to utilize EHRs for predicting GDM in early stages of pregnancy. Although several risk score systems for GDM diagnosis have been developed in recent years[34], they are not commonly used in routine practice and are not recommended by current guidelines. Our results show that EHRs can be used to produce accurate predictions of GDM risk, performing significantly better than a baseline model based on commonly assessed risk factors. Our 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 reaches an auROC of 0.854.

The effect of early-pregnancy intervention on GDM development is well studied but without a clear consensus[35], as some studies suggest a GDM risk reduction of up to 39% with combined diet and exercise interventions during pregnancy in high-risk pregnant women [9]. One of the challenges in attempting to analyze the efficacy of prevention strategies is the low prevalence of GDM in the population. Our predictor may be used to identify and recruit a high risk cohort with risk of up to 70% for GDM. Our study therefore paves the way to future randomized control trials to further study both the use of a model for early prediction of GDM and possible preventive interventions.

Other than well known risk factors for GDM such as maternal age[36] and family history of diabetes[37], our analysis reveals factors that were not previously described to be highly predictive of GDM. The main risk factors identified were GCT results in previous pregnancies. While the medical system already addresses women with a history of GDM in previous pregnancies as being at increased risk for GDM in the current pregnancy[38], here we have shown that the GCT result is far more predictive (**Figure 1F**). This may suggest new risk assessment guidelines which will be based on explicit GCT value, and not GDM diagnosis.

Although maximal prediction accuracy requires using the patient's entire EHR, we demonstrated that nine simple questions that can be answered by the patient herself still enable accurate prediction (auROC of 0.801). This may allow patients to get accurate GDM risk estimation by web- or smartphone-based self-assessment tools.

One of the major issues regarding GDM diagnosis is whether universal or selective screening should be used[38,39]. Currently, a 50g GCT or a similar universal screening method is commonly used, followed by the 100g OGTT if needed[40]. However, 20% of the GDM cases are estimated to be missed using this screening approach[32,33]. Our results suggest that a more efficient approach may be established by using the prediction model for the identification of low risk women who can avoid the GCT and the OGTT altogether, therefore creating a selective, cost-effective screening method. As most of the population is predicted to have a low chance of GDM, this could be highly effective, as demonstrated in **Figure 5B**. Avoiding 50% of the GCTs of patients who previously did a GCT would result in only 5% miss rate when diagnosing GDM according to the two steps approach guidelines. Additionally, women with high risk for GDM development may be referred directly to the diagnostic 100g OGTT and avoid the screening test. Accurate selective screening is highly desirable, as it can both reduces costs and physical inconvenience for women at low or high risk for GDM. The utility of this approach should be tested in prospective designated clinical trials.

This study has several limitations. First, our predictor is based on retrospective EHR which have inherent biases and are influenced by the interaction of the patient with the health system[41]. However, these biases are reduced here since the data contains information originating from a non-governmental, non-profit organization which includes the majority of the Israeli population, and since the outcome of the model is based on routine pregnancy tests. Another limitation is that our data does not contain information regarding lifestyle habits, previously shown to be associated with GDM development[42]. Finally, the predictor was trained and validated on Israeli population data. Although the applicability to other populations needs to be shown, the size of our data, our validation process, and the fact that our analysis validated the utility of established risk factors for GDM development supports its ability to generalize to other populations.

In conclusion, our work demonstrates that accurate and calibrated predictions of GDM before and early on in pregnancy can be achieved. These results can have many implications for the health of pregnant women and their offspring. Our predictor may be the basis for a selective screening process for GDM diagnosis, and for identification and implementation of early-stage pregnancy interventions in order to prevent or reduce the development of GDM and its associated adverse health outcomes.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Donovan, P. J. & McIntyre, H. D. Drugs for gestational diabetes. *Aust. Prescr.* **33,** 141–144 (2010).
2. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2018. *Diabetes Care* **41,** S13–S27 (2018).
3. 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).
4. Lowe, W. L. *et al.* Hyperglycemia and Adverse Pregnancy Outcome Follow-up Study (HAPO FUS): Maternal Gestational Diabetes Mellitus and Childhood Glucose Metabolism. *Diabetes Care* **42,** 372–380 (2019).
5. Lowe, W. L. *et al.* Association of gestational diabetes with maternal disorders of glucose metabolism and childhood adiposity. *JAMA* **320,** 1005–1016 (2018).
6. Zhao, P. *et al.* Maternal gestational diabetes and childhood obesity at age 9-11: results of a multinational study. *Diabetologia* **59,** 2339–2348 (2016).
7. Hunt, K. J. & Schuller, K. L. The increasing prevalence of diabetes in pregnancy. *Obstet. Gynecol. Clin. North Am.* **34,** 173–99, vii (2007).
8. Bain, E. *et al.* Diet and exercise interventions for preventing gestational diabetes mellitus. *Cochrane Database Syst. Rev.* CD010443 (2015). doi:10.1002/14651858.CD010443.pub2
9. 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).
10. 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).

11. Avati, A. *et al.* Improving palliative care with deep learning. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 311–316 (IEEE, 2017). doi:10.1109/BIBM.2017.8217669
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. *arXiv* (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. Data – Clalit Research Institute. at <http://clalitresearch.org/about-us/our-data/>
20. Vandorsten, J. P. *et al.* NIH consensus development conference: diagnosing gestational diabetes mellitus. *NIH Consens. State Sci. Statements* **29,** 1–31 (2013).
21. Monitoring of Pregnancy and Medical Examinations During Pregnancy, Ministry of Health. at <https://www.health.gov.il/English/Topics/Pregnancy/during/examination/Pages/permanent.aspx>
22. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **4,**

40–79 (2010).

23. U.S. 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? at <https://www.nichd.nih.gov/sites/default/files/publications/pubs/Documents/gestational_diabetes_2012.pdf>

24. Mining, D. *et al.* The Elements of Statistical Learning.

25. Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* (2014).

26. Omar, K. XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison Semester Project. (2018).

27. biendata competitions. KDD Cup 2018 - Winners List. at <https://biendata.com/competition/kdd_2018/winners/>

28. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* (2017).

29. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2,** 749–760 (2018).

30. Chu, S. Y. *et al.* Maternal obesity and risk of gestational diabetes mellitus. *Diabetes Care* **30,** 2070–2076 (2007).

31. 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).

32. van Leeuwen, M. *et al.* Glucose challenge test for detecting gestational diabetes mellitus: a systematic review. *BJOG* **119,** 393–401 (2012).

33. Donovan, L. *et al.* Screening tests for gestational diabetes: a systematic review for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* **159,** 115–122 (2013).

34. 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).

35. Shepherd, E. *et al.* Combined diet and exercise interventions for preventing gestational diabetes mellitus. *Cochrane Database Syst. Rev.* **11,** CD010443 (2017).

36. 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).

37. Di Cianni, G. *et al.* Prevalence and risk factors for gestational diabetes assessed by universal screening. *Diabetes Res. Clin. Pract.* **62,** 131–137 (2003).

38. 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).

39. Davey, R. X. Selective versus universal screening for gestational diabetes mellitus: an evaluation of predictive risk factors | The Medical Journal of Australia. *The Medical Journal of Australia* (2001).

40. 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).

41. 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).

SUPPLEMENTARY MATERIAL

APPENDIX 1: TARGET AND COHORT DEFINITIONS

The version of Clalit Health database that we used does not include exact delivery dates for all women, however it has 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, and looking for GCTs and OGTTs in the relevant period prior to the delivery, namely 32 weeks before the logged date of birth to 7 weeks after the logged date of birth. The fact that GCTs are only used in pregnancy in Israel, and the fact that we looked for pregnancy period to begin with means the tests we see are all pregnancy-related.

GCTs and OGTTs appear in the labtests data under five distinct tests: one for 1 hour 50g GCT result, and four for fasting, 1h, 2h and 3h 100g OGTT results. We defined GDM in accordance to practice, regardless of the order of the tests, and without consideration of whether a relevant diagnosis was recorded. In case more than one test was conducted, we considered a positive result in a single test to be positive. We chose to exclude a small number of pregnancies (n=8228 in the training set and n=1,525 in the test set, 1.6%) for women that had a GCT result of 140 mg/dL or higher, but did not have a record of a OGTT.

We defined our cohort according to the relevant date of delivery. The main cohort included pregnancies that ended between January 1st, 2010 to December 31st, 2016, and the validation cohort included pregnancies that ended between January 1st, 2017 to December 31st, 2017.

Our sole exclusion criteria was pre-pregnancy notion of (non-gestational) diabetes. Normally women with DM do not take a GCT during pregnancy, but it appears that some (<0.2%) do. To address that, we excluded patients who had one of the following markers prior to pregnancy start: (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 diabetes, this metric is extremely inaccurate in our data as some non-fasting patients still take the test, and therefore we decided not to use it.

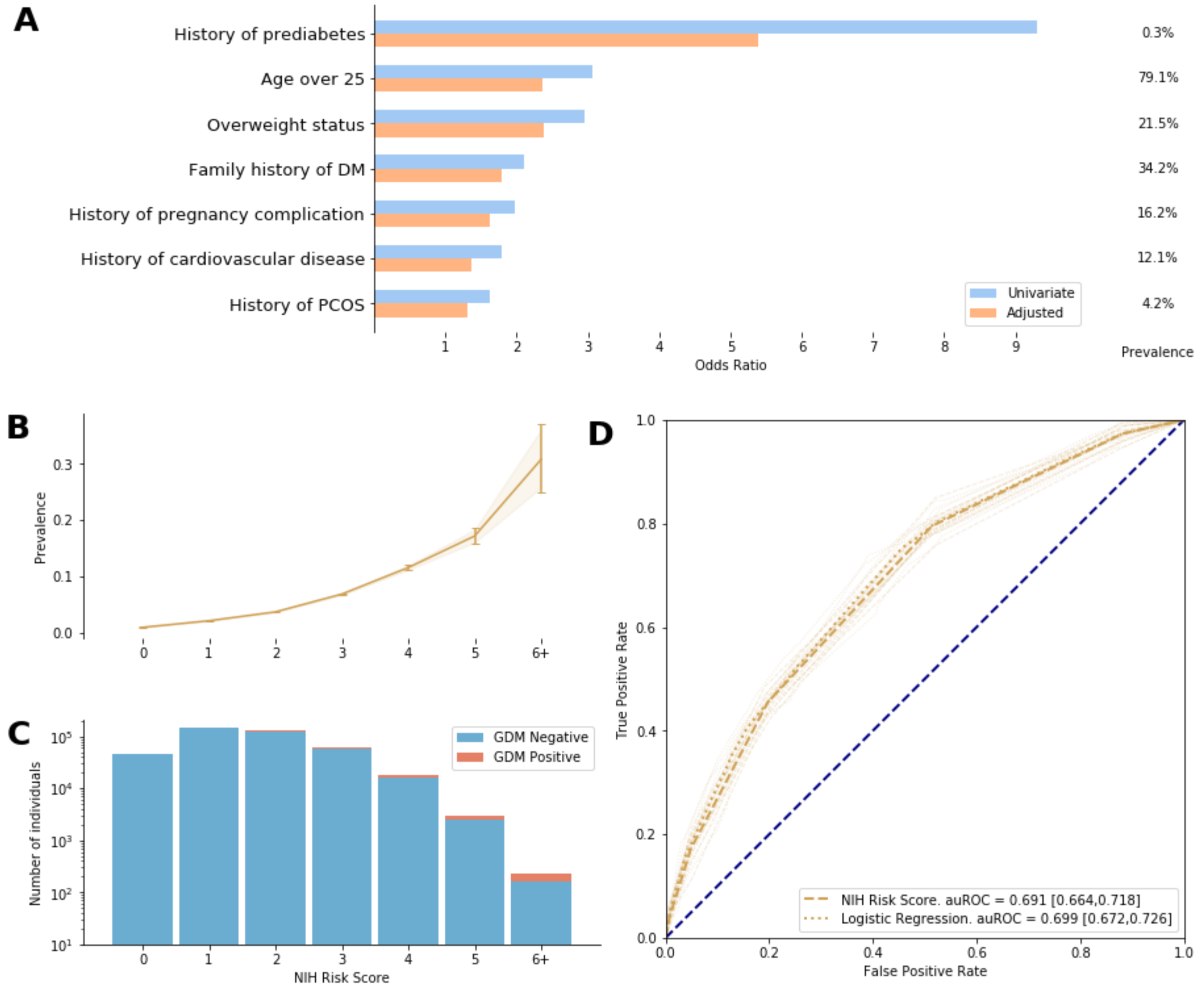
APPENDIX 2: BASELINE RISK SCORE DEFINITION

The original Risk Score suggested by the NIH[23] includes eight parameters, of which all except for ethnicity are relevant for the Israeli population. We therefore included seven parameters in our score, defined according to the following binary variables:

1. Overweight status: true if non pregnancy BMI is higher than 25 kg/m^2. If there is no record of BMI prior to the pregnancy, we consider that as false.
2. Family history of diabetes: 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.
3. Age: true if the patient was 25 or more years of age at pregnancy start.

4. History of pregnancy complication: the logic OR operation of the following markers:
    a. History of GDM according to GCTs and OGTTs, defined similar to the target
    b. History of miscarriage or stillbirth, seen in a form of a diagnosis with ICD9 632, 634.x, 635.x or 637.x
    c. History of a liveborn baby with birth weight higher than 4 kg. Note that birth weight is only logged for deliveries done in Clalit owned hospitals (about 30% of the deliveries)
5. History of PCOS: true if the patient has at least one diagnosis of PCOS, ICD9 code 256.4. Only diagnoses available at pregnancy start are considered.
6. Problems with insulin or blood sugar: true if the patient has at least one diagnosis of prediabetes, either according to ICD9 codes 790.2x or by a HbA1c test in the range 5.7% to 6.4%. Only diagnoses and tests available at pregnancy initiation are considered.
7. High blood pressure, high cholesterol, and/or heart disease: the logic OR operation of the following markers:
    a. History of high BP, defined as two or more BP tests with systolic BP over 140 or diastolic BP over 90. Blood pressure measurements taken during pregnancies are not included in this analysis.
    b. Recorded relevant ICD9 of 401.x, 272.x 390.x-449.x

The final Baseline Risk Score is, then, the number of “true” entries in the above list, and therefore ranges between 0-7. An analysis of the odds ratio of the constructing variables, as well as a comparison to a logistic regression model from the above binary variables is presented in **Supp. Figure 1**. Of note, the logistic regression model does not significantly improve performance.



Supp. Figure 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 test set. Logistic regression model does not suppress the Naive summation in the risk score.

APPENDIX 3: FEATURE GENERATION MECHANISM

We constructed 2355 features from the dataset. The following list describes the generation mechanism for each. As suggested by previous works[43], missing values were inherently handled by the gradient boosting predictor[44].

A. Features that are available at pregnancy initiation (295 features):

1. 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. World's 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. Socio-economic data of the locality the patient attended most clinic visits. Although personalized socio-economic data were not available, we generated some estimates using the data available by Israel's Central Bureau of Statistics[45]. Features include locality type (length 20, 1-hot vector) and locality religion breakdown (length 5, summing to 1 vector).

2. Basic measures (7 features):
    i. Age at pregnancy initiation
    ii. Weight, height and BMI. Only samples available before the current pregnancy and outside past pregnancies were considered, median for all samples with age 18 and up was calculated.
    iii. Systolic and Diastolic blood pressure. Only samples available before the current pregnancy and outside past pregnancies were considered, median for all samples with age 18 and up was calculated.
    iv. Number of children born in current pregnancy - 1 for single born, 2 for twins etc.
3. Pregnancy history (103 features):
    i. History of GDM:
        a. Any history of GDM according to past pregnancies' GCTs and OGTTs (True/False).
        b. GDM status in each of the last 3 pregnancies.
    ii. History of miscarriage: seen in a 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 is only available for 25% of the cohort.
    iv. Number of previous births: number of children born before the current pregnancy.
    v. Lab tests during last 3 pregnancies:
        a. Median values during each pregnancy of the following tests, the 25 most common lab tests: HB, HCT, RBC, MCV, MCH, MCHC, WBC, PLT, LYM%, NEUT%, MONO%, EOS%, BASO%, LYMP (abs), NEUT (abs), MONO (abs), EOS (abs), BASO (abs), MPV, RDW, Urine culture, MICRO%, MACRO%, HYPO%, HYPER% (75 features).
        b. Median values during each pregnancy of fasting glucose and HbA1c% (6 features).

c. GCT and OGTT results, if available (15 features).

4. Medical history outside of pregnancy (144 features):

   i. Number of first degree relative (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 start are considered.

   iii. History of prediabetes:

      a. Diagnoses: true if the patient has at least one diagnosis of prediabetes according to ICD9 codes 790.2x.

      b. Maximal HbA1c% logged.

      c. Joint prediabetes definition: either according to diagnosis or by a HbA1c test in the range 5.7% to 6.4%. Only diagnoses and tests available at pregnancy initiation are considered.

   iv. Features related to high blood pressure, high cholesterol, and/or heart disease:

      a. Number of high BP tests with systolic BP over 140 or diastolic BP over 90. Blood pressure measurements taken during pregnancies 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) (3 True/False features).

   v. Baseline Risk Score value (Appendix 1)

   vi. Lab tests during the last 5 years (132 features): logging the median value in every window M1-M5 (see Time Windows ahead). We considered the 25 most common tests: HB, HCT, RBC, MCV, MCH, MCHC, WBC, PLT, LYM%, NEUT%, MONO%, EOS%, BASO%, LYMP (abs), NEUT (abs), MONO (abs), EOS (abs), BASO (abs), MPV, RDW, Urine culture, MICRO%, MACRO%, HYPO% and HYPER%, plus Glucose and HbA1c%. We only considered data gathered outside of pregnancy periods for these features.

   vii. Coefficients (2) of a linear regression for fasting glucose vs. time (only if 3 or more measurements were available).

B. Features gathered throughout current pregnancy (2060 features):

1. Lab tests (524 features): Median values of 250 most common lab tests during F0-F2 (see Time Windows ahead).

2. Clinic and hospital diagnoses (906 features): Counts of 300 most common diagnoses in community clinics and 100 most common diagnoses in hospitals, plus “other” count for all the non-top diagnoses, for each window in F0-F2 (see Time Windows ahead).

3. Anthropometrics and blood-pressure measurements (27 features):

    i. Medians of weight, height, BMI, systolic and diastolic BP and the time between the measurement to the GCT, for each window in F0-F2 (see Time Windows ahead).
    ii. Coefficients (2) of a linear regression for weight vs. time, for 10-20 weeks of gestation (only if 3 or more measurements were available)..
    iii. Coefficients (2x2) of a linear regression for systolic/diastolic blood pressure vs. time, for 0-20 weeks of gestation (only if 3 or more measurements were available)..

4. Pharmaceuticals (603 features): Counts of 300 most common medications in, plus "other" count for all the non-top medications, for each window in F0-F2 (see Time Windows ahead).

Time Windows

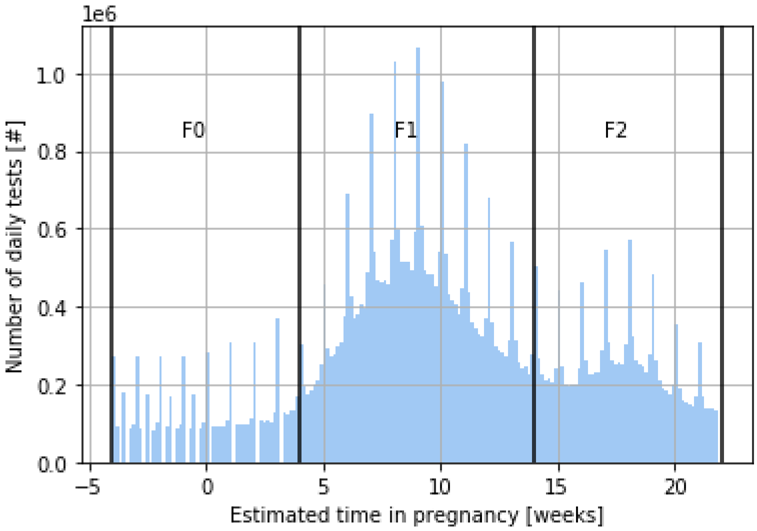
The following time windows were defined for feature calculation:

A. Windows during pregnancy were defined according to the usual medical examination pregnancy schedule for Israeli women[21], we defined the following relative-time windows:
    - **F0**: from 30 to 22 weeks before the GCT, representing -4 to 4 weeks of gestation.
    - **F1**: from 22 to 12 weeks before the GCT, representing 4 to 14 weeks of gestation. This window includes the period in which women attend the first blood test during pregnancy, which is recommended during 6-12 weeks of gestation.
    - **F2**: from 12 to 4 weeks before the GCT, representing 14 to 22 weeks of gestation. This window includes the period in which women attend the second blood test during pregnancy (triple test), which is recommended during 16-18 weeks of gestation.

   This choice is backed by the test population in the data, as seen in **Supp. Figure 2**.

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

C. Past pregnancies period are denoted by **P1** (last pregnancy), **P2** (two pregnancies ago) and **P3** (three pregnancies ago). Pregnancies were located according to the birth date of a child, and pregnancy period was defined as 40 weeks before that date plus 2.5 months in each direction to cover randomization of birth dates.



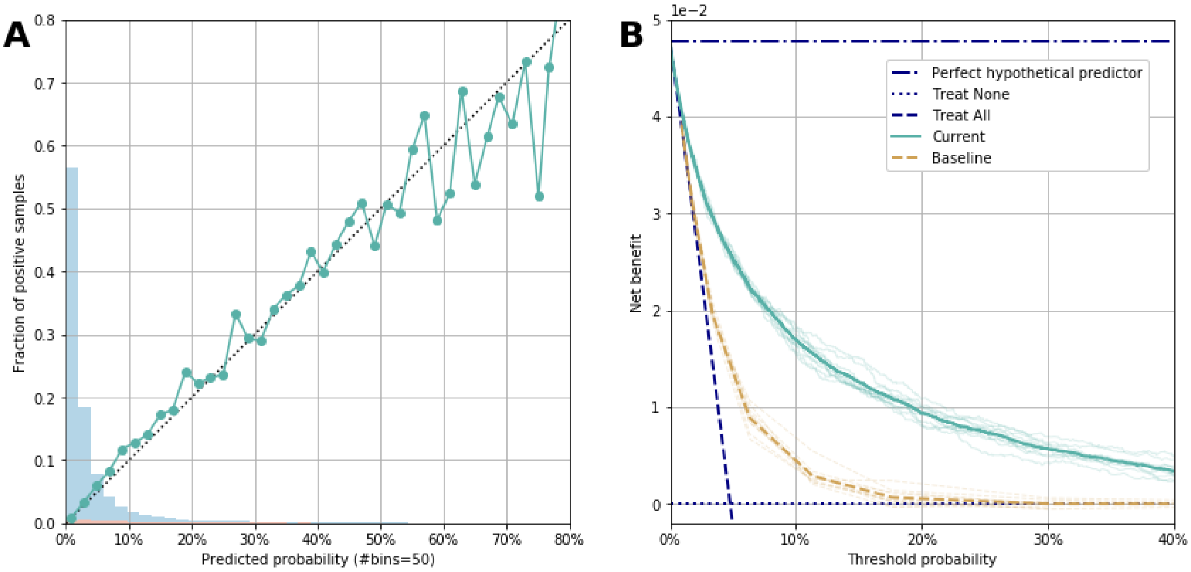
Supp. Figure 2: histogram of lab tests during pregnancy, showing the window definition of F0, F1 and F2. The peaks showing are weekly, and represents the fact that patients tend to see a doctor in the same day of the week.

APPENDIX 4: PREDICTION DETAILS

We used 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
- leraning_rate = 0.05
- feature_fraction = 0.2
- bagging_fraction = 0.8
- bagging_freq = 5
- min_data_in_leaf = 4

We ensured that the model predictions are well-calibrated[47], namely that its predictions reflect the actual expected risk of an individual (**Supp. Figure 3A**). We furthermore demonstrate the utility of the predictor by considering its Decision Curve[48] (**Supp. Figure 3B**).



Supp. Figure 3: Basic utility of the predictor. **A**: Calibration curve, showing the fraction of positive samples per bin versus the mean predicted probability of the bin. **B**: Decision curve, showing the net benefit versus the threshold probability, for both predictor and baseline. The predictor outperforms the baseline at all thresholds.

APPENDIX 5: DEPENDENCE PLOTS

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

$$LO = \phi_0 + \phi_1 + \ldots + \phi_d$$

where ϕ_0 is the "base" Shapley value (the logit of the population prevalence P_0), and ϕ_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

$$P_i = S\left(\phi_0 + \phi_i\right)$$

where

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

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

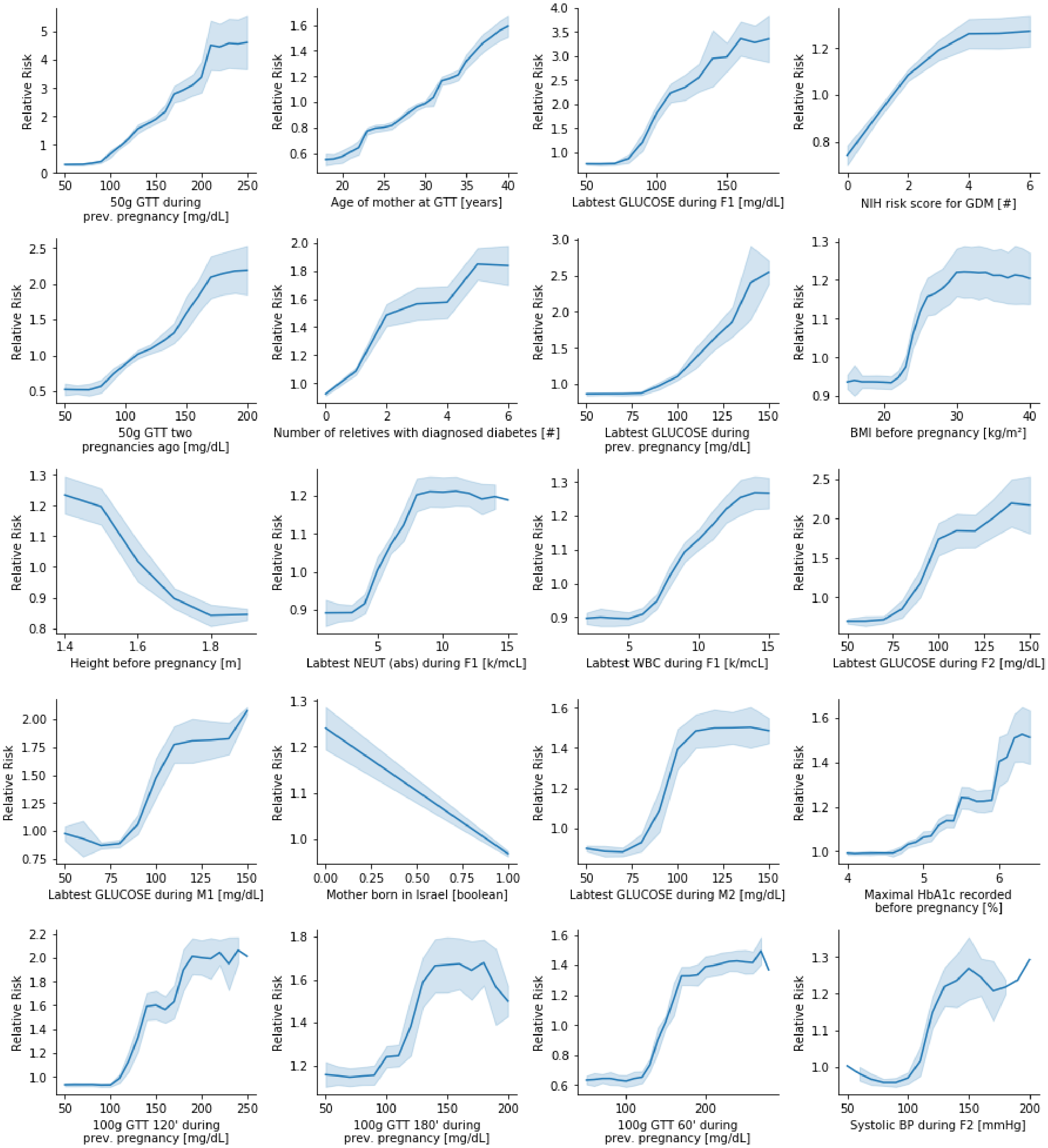
$$RR_i = \frac{P_i}{P_0} = \frac{S\left(\phi_0 + \phi_i\right)}{S\left(\phi_0\right)}$$

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

$$RR_i = \frac{P_i}{P_0} = \frac{S\left(\phi_0 + \sum_{i \in D} \phi_i\right)}{S\left(\phi_0\right)}$$

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

Supp. Figure 4 contains the dependence plots of 20 most significant features, ranked according to the mean absolute Shapley value.



Supp. Figure 4: Additional dependence plots. Top 20 features are shown (ordered left to right, top to bottom). In each 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.

REFERENCES

42. 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).

43. Josse, J., Prost, N., Scornet, E. & Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv* (2019).

44. 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 '16* 785–794 (ACM Press, 2016). doi:10.1145/2939672.2939785

45. CBS Site. at <https://www.cbs.gov.il/en/pages/default.aspx>

46. Ke, G. *et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *undefined* (2017).

47. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21,** 128–138 (2010).

48. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26,** 565–574 (2006).