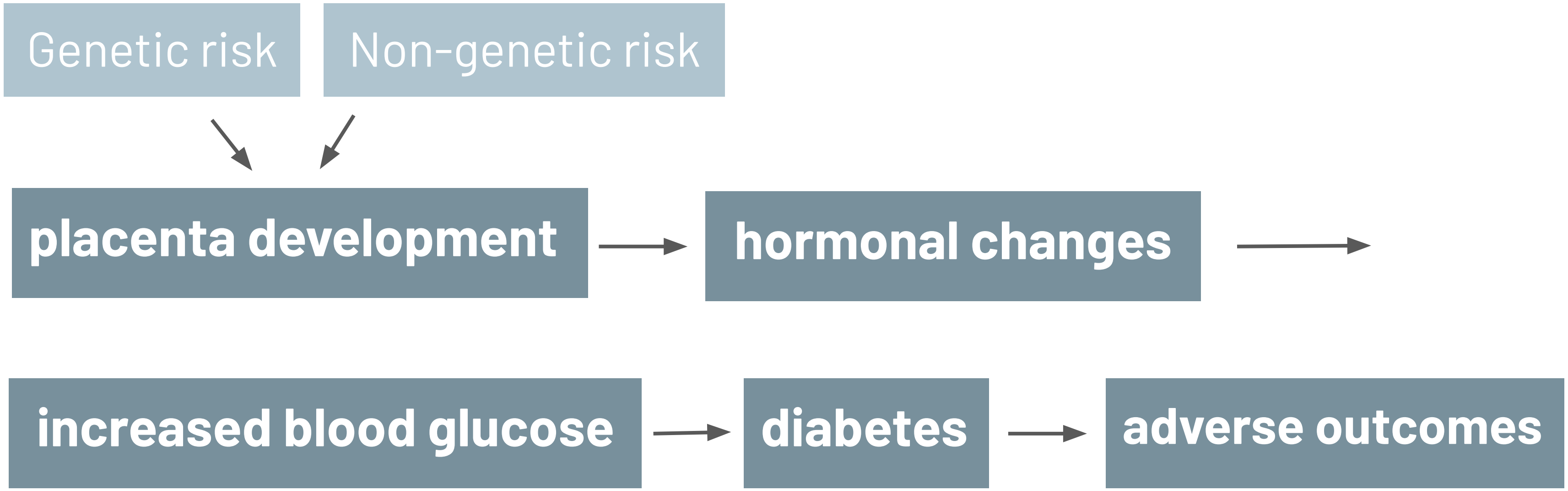
Prediction of gestational diabetes based on nationwide electronic health records

Christa Caggiano
Bioinformatics Program

Zaitlen Lab JC
Jan 27th 2019

Gestational diabetes mellitus is a common complication of pregnancy



Gestational diabetes mellitus is a common complication of pregnancy that can affect both mothers and offspring

adverse outcomes

Mothers

- Increased risk for T2D
- Risk for preeclampsia & postpartum depression
- C-section or birth injuries

Children

- Stillbirth
- Macrosomia & birth trauma
- Respiratory difficulties
- Metabolic complications

Problem: Adverse outcomes for diabetes can be prevented with early detection and monitoring

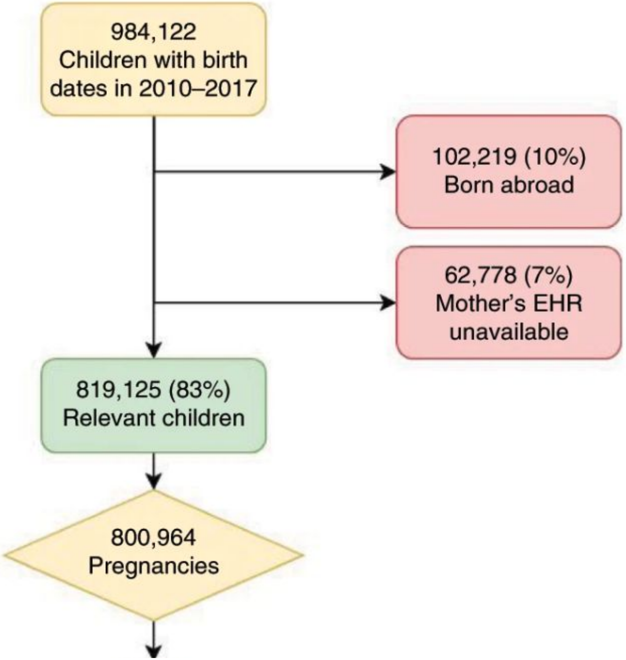
Problem: Adverse outcomes for diabetes can be prevented with early detection and monitoring

Paper's solution: Develop a machine learning model trained on EHR data to accurately predict risk for gestational diabetes

Outline

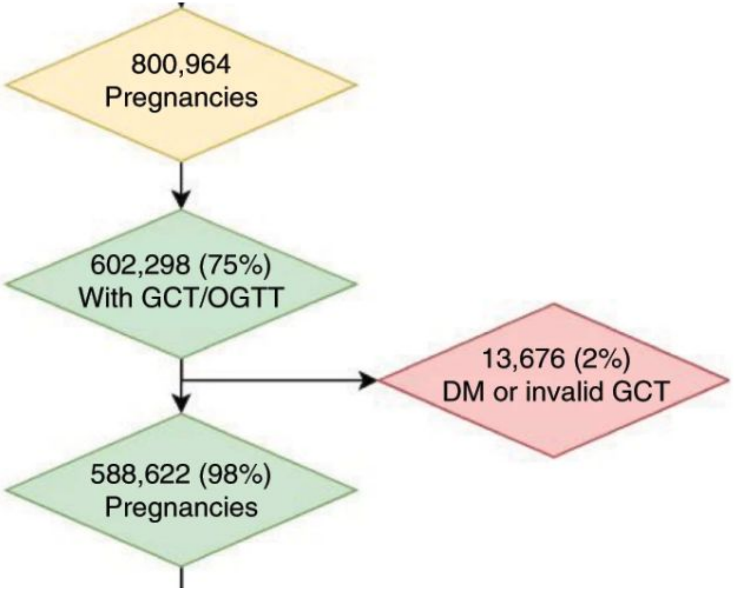
1: Data collection & design
2: Machine learning approach
3: Results
4: Conclusions and other ideas

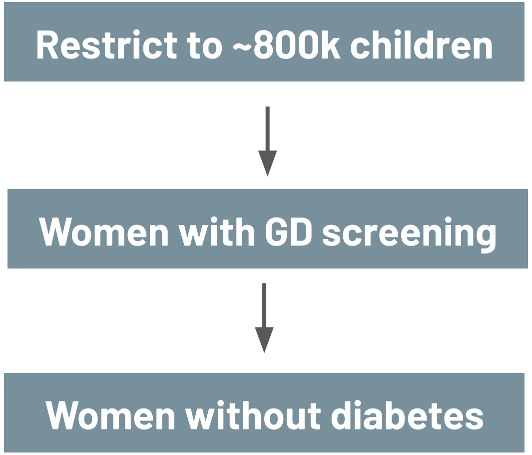
Method utilizes database from Israel's largest healthcare provider (~ 50% of Israel's adult population)



Restrict to ~800k children

Remove women with non-gestational diabetes, and those who have gestational diabetes screening tests





GD screening in Israel is a two step process performed at 24-28 weeks

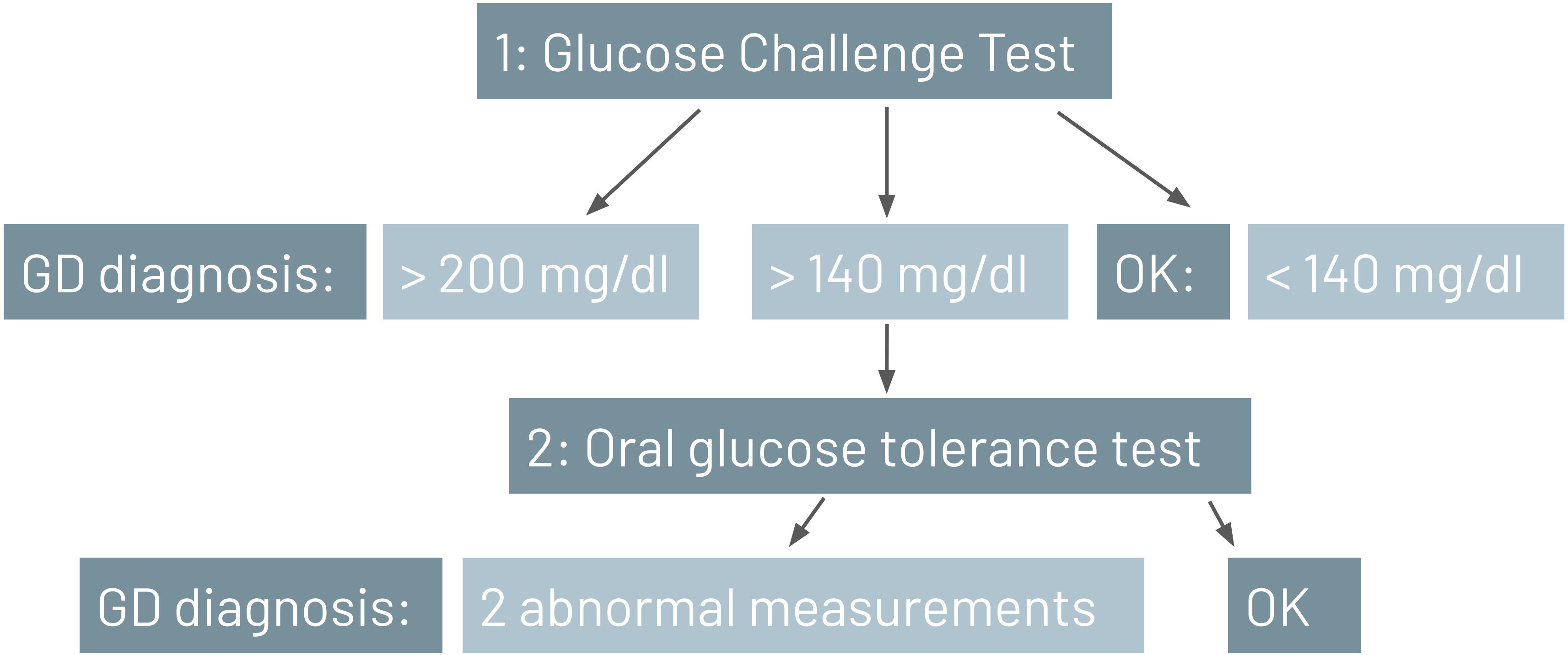
1: Glucose Challenge Test

GD diagnosis: > 200 mg/dl

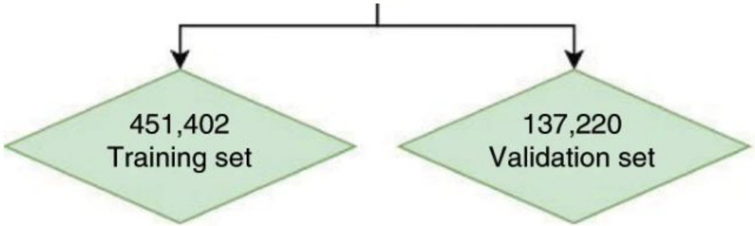
> 140 mg/dl

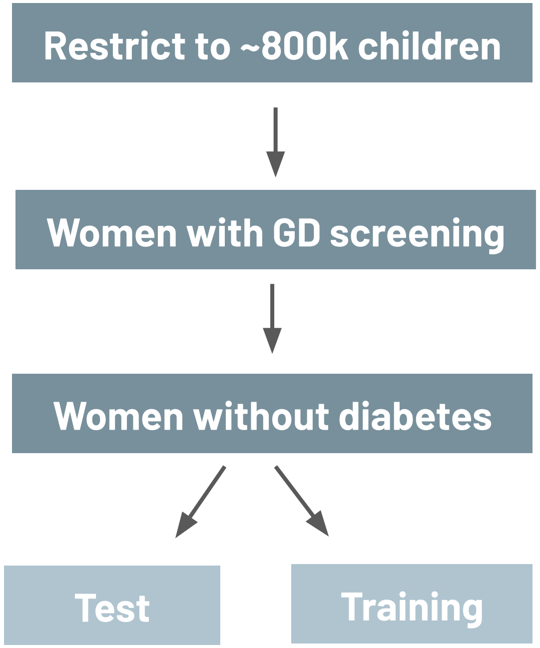
OK: < 140 mg/dl

GD screening in Israel is a two step process performed at 24-28 weeks



Split all women into test and training sets





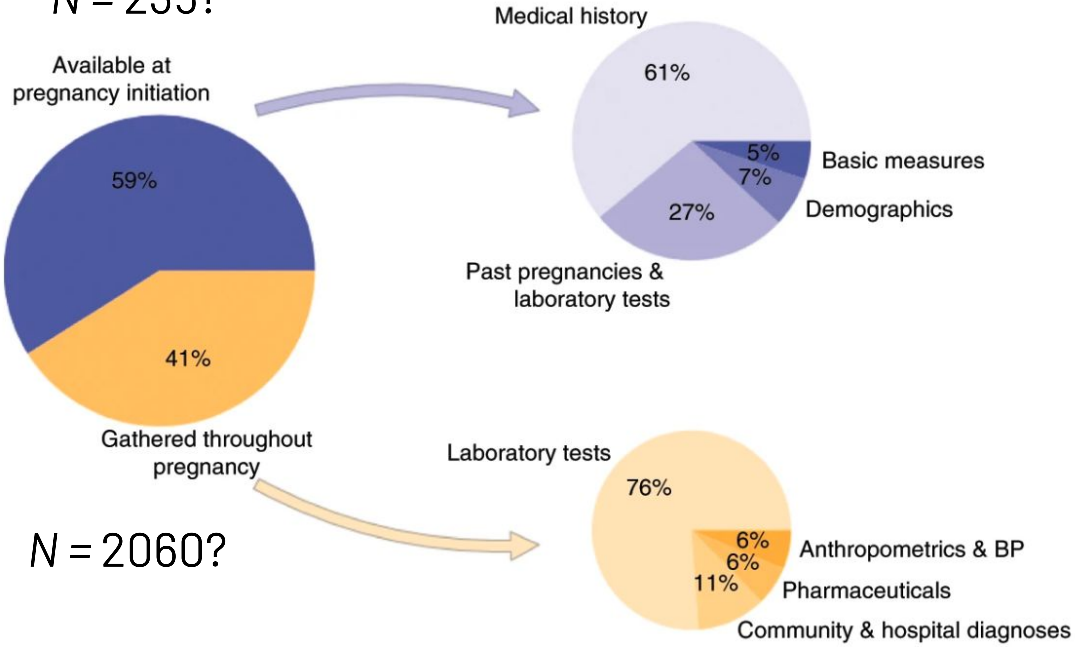
3 validation sets were used in addition to the test set

		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

A combination of pre-pregnancy and pregnancy data points were used as features

N = 235?

N = 2,355 total



N = 2060?

Examples of features used:

Pre-pregnancy:

- “Basic features”: Age at conception, BP, BMI

- Pregnancy history: history of GD, lab tests in previous pregnancies, number of children, miscarriages

- Other features: Baseline risk score value, prediabetes history

Pregnancy:

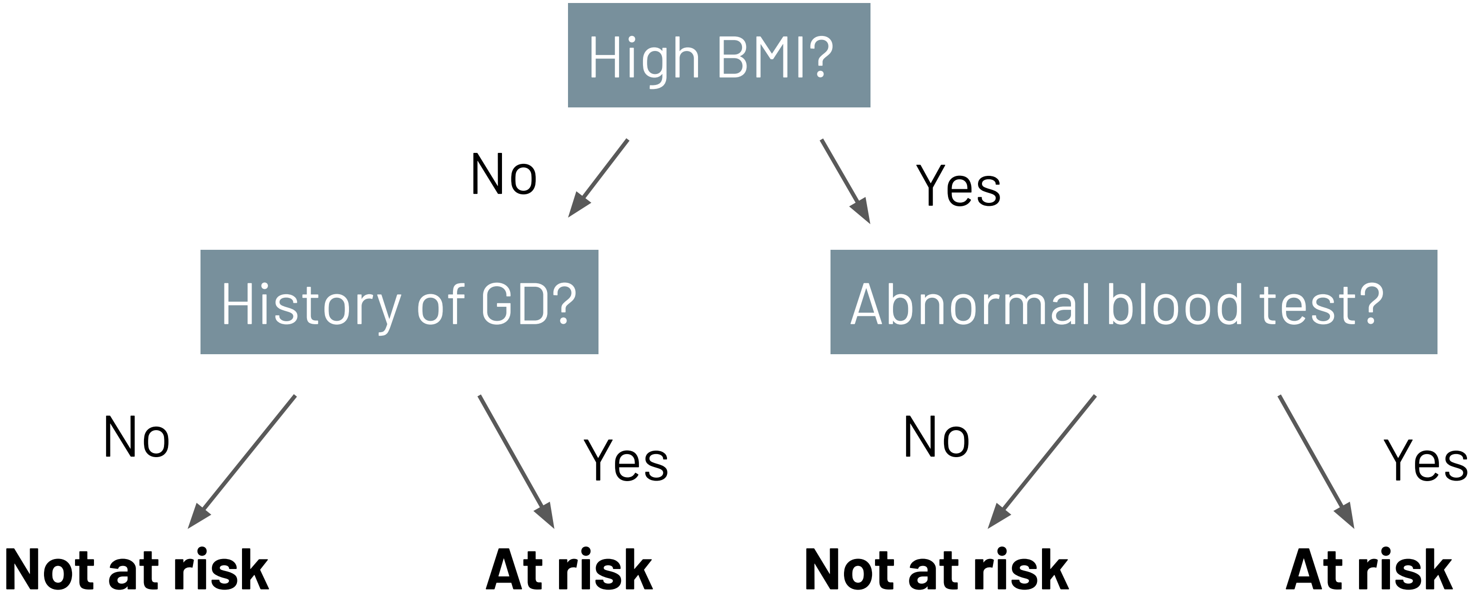
- Anthropomorphic measurements: BMI, BP, weight vs time

- Clinical and hospital diagnoses: top 300 most common community clinic diagnoses

- Laboratory tests

Gradient Boosting Machines are built off decision trees

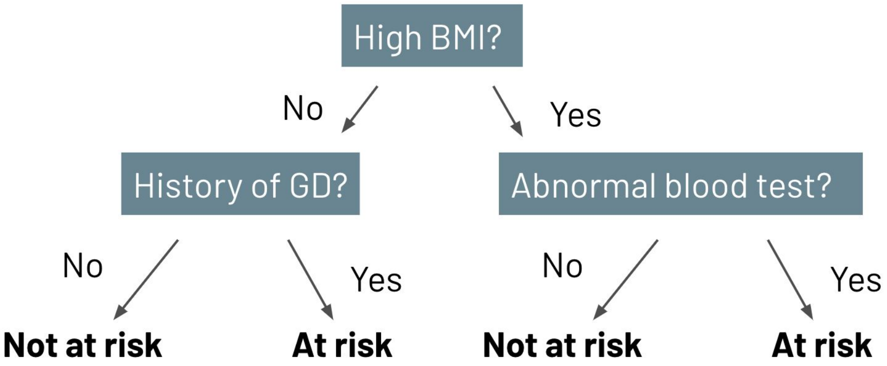
Ex: *Is a patient at risk for gestational diabetes?*

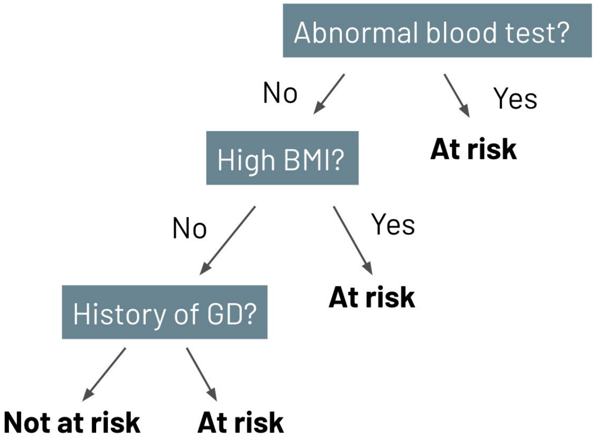


Some decision trees perform better than others

Patient 1 w/ GD: not high BMI, abnormal blood test, no history of GD

Patient 2 w/ GD: high BMI, normal blood test, history of GD



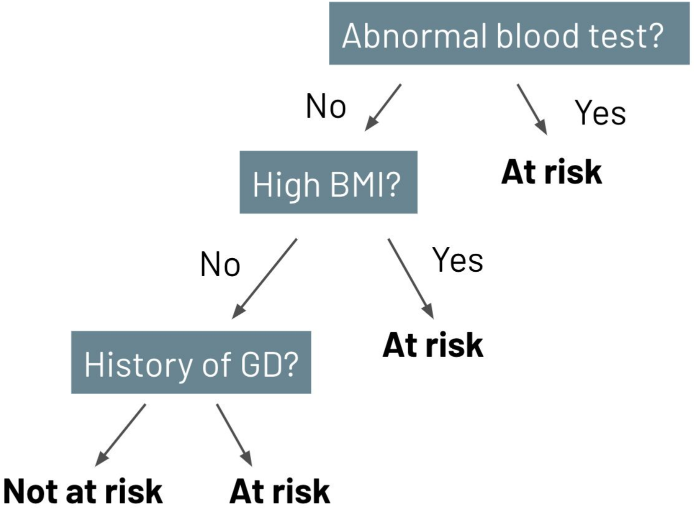


CODE

https://github.com/christacaggiano/gradient-boosting

Decision trees overfit to the training data

Patient 3 w/o GD: high BMI, normal blood test, no history of GD



CODE

Gradient boosting creates many trees, with each new tree "learning" from its predecessor's mistakes

Tree 0: "average" (log odds) of all samples in training set

Tree 1: Best tree to fit residuals between true data + prediction of tree 0

Tree 1: Best tree to fit residuals between true data + prediction of tree 1

Gradient boosting creates many trees, with each new tree "learning" from its predecessor's mistakes

Learning rate to prevent overfitting to training data

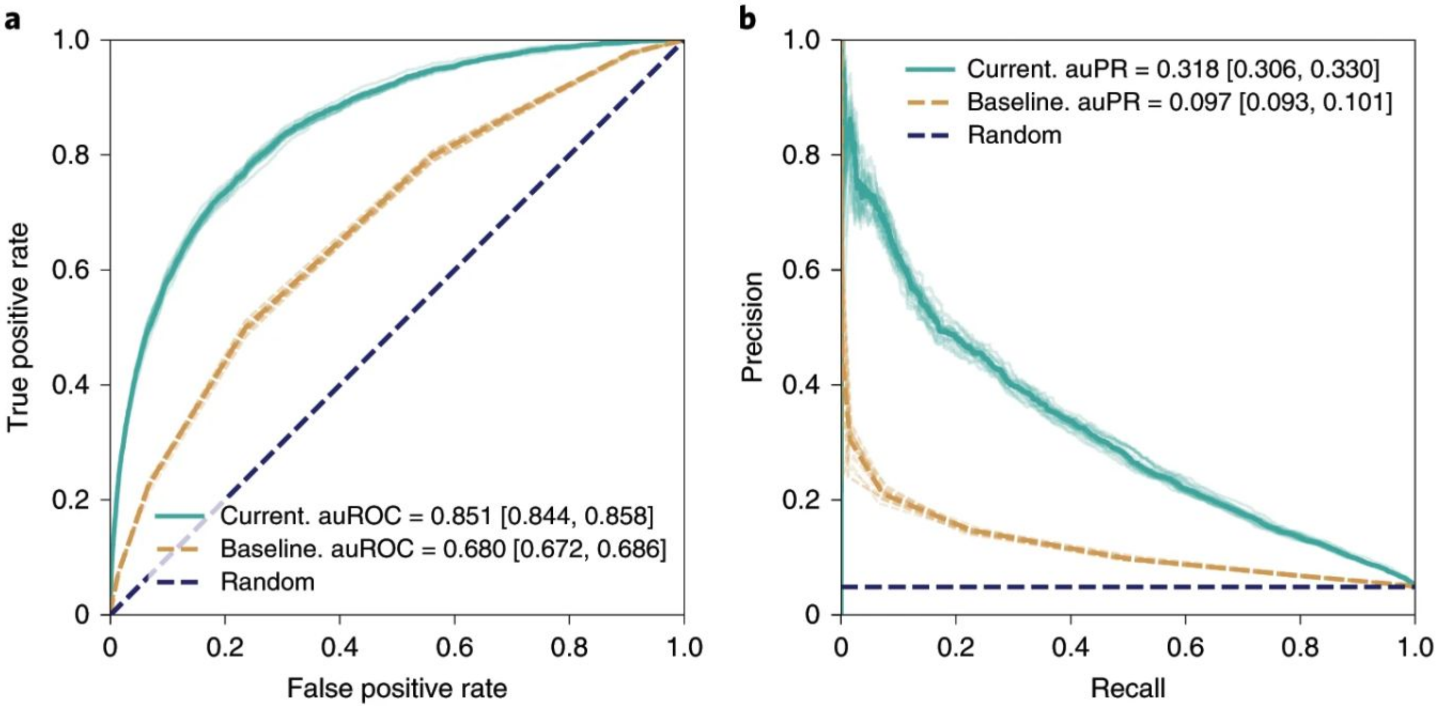
Tree 0: "average" (log odds) of all samples in training set = prediction 0

Tree 1: Best tree to fit residuals between true data + prediction 0 = prediction 1

Tree 2: Best tree to fit residuals between true data + prediction 1 = prediction 2

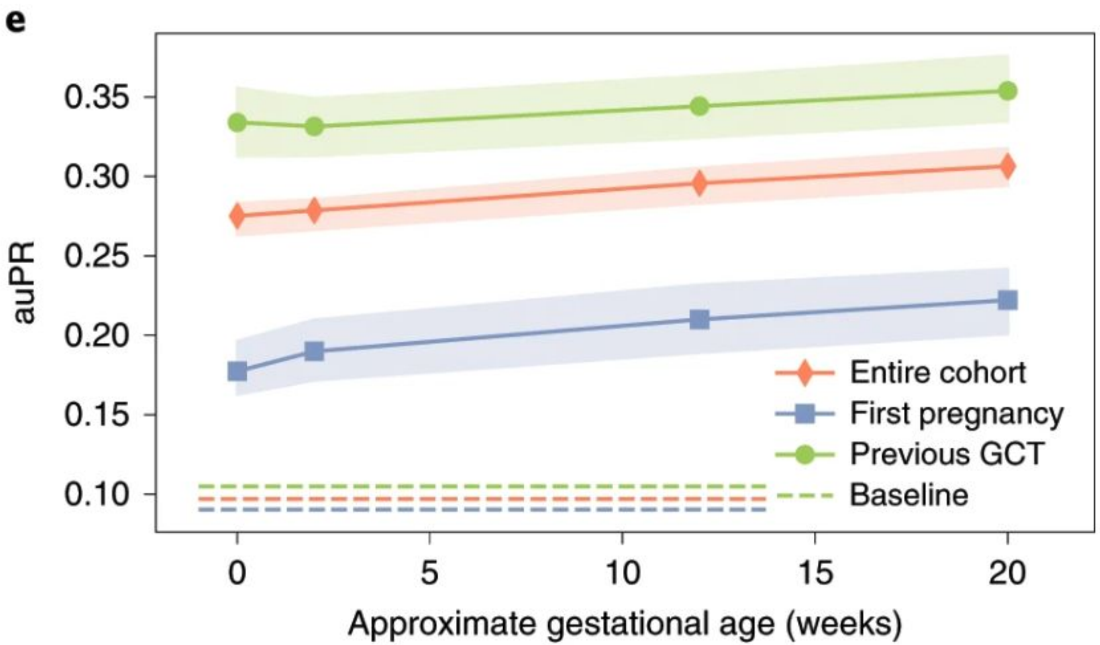
CODE

Gradient boosting performs better than baseline model

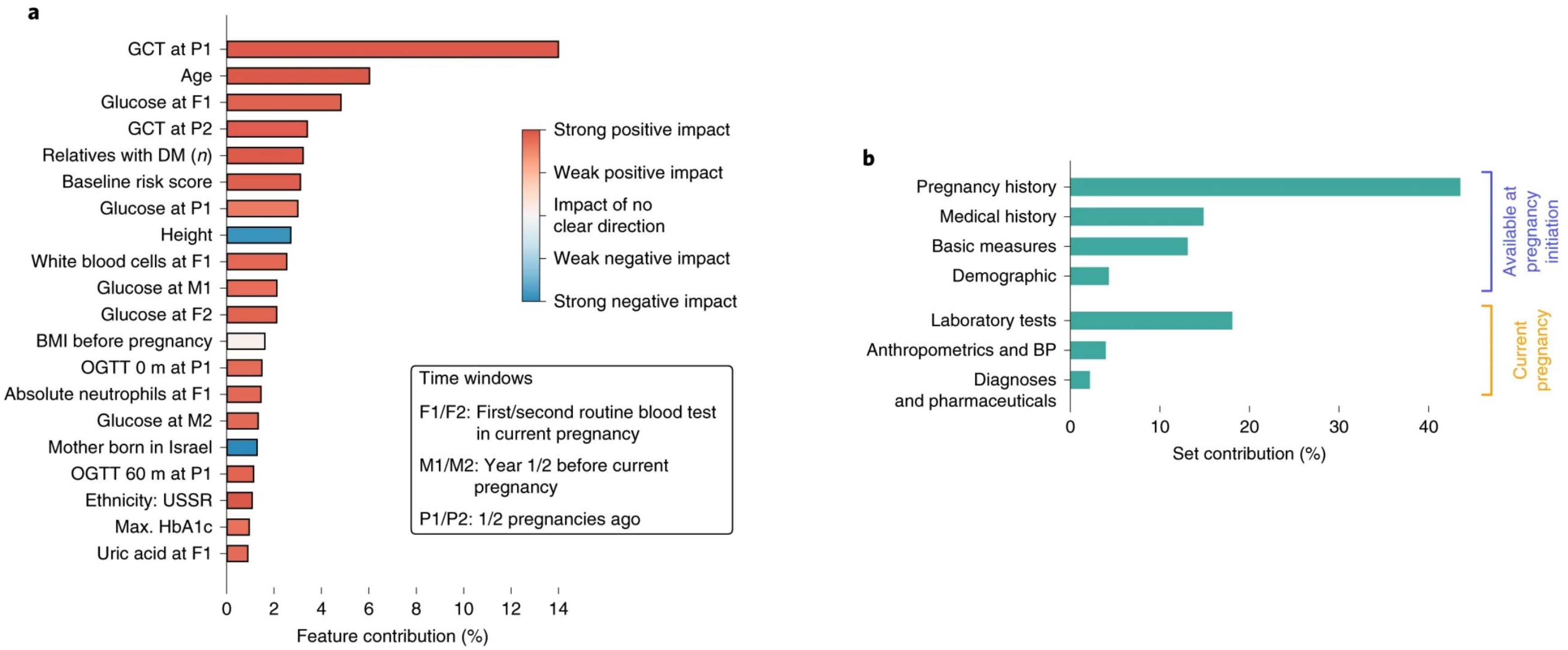


Precision = TP/TP+FP, Recall = TP/TP+FN

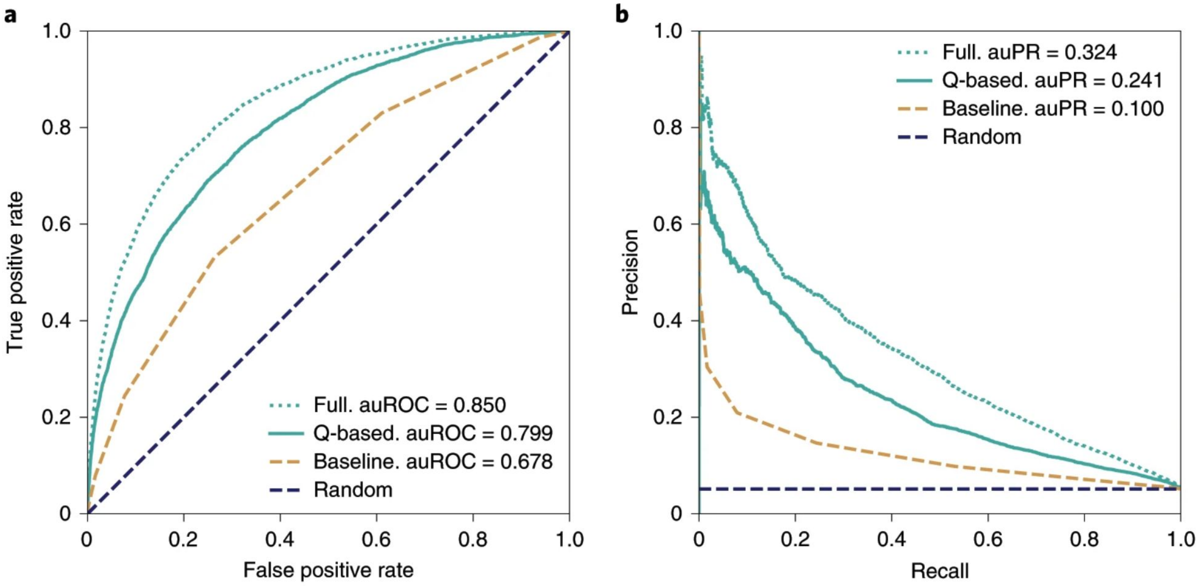
Their model can predict GD fairly accurately at the beginning of pregnancy



Using Shapley values, the relative contribution of features are estimated



They developed a simple 8 question survey that can predict GD with small drop of performance



c

Features that can generated by asking the following questions:

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