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 *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 *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