# Electronic Health Records to Predict Gestational Diabetes Risk

Bilal A. Mateen[1,2], Anna L. David[3], Spiros Denaxas[2,4,5,6,7]
1 King's College Hospital, London, United Kingdom
2 The Alan Turing Institute, London, United Kingdom
3 Elizabeth Garrett Anderson Institute of Women's Health, University College London, London, United Kingdom
4 Institute of Health Informatics, University College London, London, United Kingdom
5 Health Data Research UK, London, United Kingdom
6 The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, London, United Kingdom
7 British Heart Foundation Research Accelerator, University College London, London, United Kingdom

**Keywords**: machine learning, gestational diabetes mellitus, risk prediction, electronic health records, artificial intelligence

**Abstract** Gestational diabetes mellitus is a common pregnancy complication associated with significant adverse health outcomes for both women and infants. Effective screening and early prediction tools as part of routine clinical care are needed to reduce the impact of the disease on the baby and mother. Using large-scale electronic health records, Artzi and colleagues [1] developed and evaluated a machine learning driven tool to identify women at high and low risk of GDM. Their findings showcase how artificial intelligence approaches can potentially be embedded in clinical care to enable accurate and rapid risk stratification.

Gestational diabetes mellitus (GDM), defined as glucose intolerance during gestation, is a common complication of pregnancy. GDM is associated with an increased risk of short term and long-term complications in both mother and child, including type 2 diabetes and operative delivery. Identifying women at a higher risk of GDM, as early into the pregnancy as possible, can potentially enable the implementation of lifestyle intervention strategies thereby reducing the associated disease burden. However, GDM diagnostic criteria vary by country, illustrating the lack of consensus around the most effective way to screen in routine clinical care [i][2]. Embedding a data-driven prognostic tool within routine care can potentially address these challenges and enable clinicians to identify women at higher or lower risk earlier and with greater accuracy. In a recent work, Artzi et al. [1] present an algorithm constructed using machine learning (ML) and retrospective electronic health records (EHR) to predict GDM.

EHR are data that are generated during routine interactions with a healthcare system primarily collected for the purposes of delivering care or financial reimbursement. EHR-derived data have been shown to be an invaluable tool for research [3,4] and are widely used for observational and interventional studies. EHR datasets are however often large, complex, and heterogeneous, and thus sophisticated methods, such as ML, are often required to efficiently analyze them. ML is a subfield of AI, encapsulating computational algorithms that can identify linear or non-linear patterns from multimodal and multidimensional datasets [5,6]. Learning can generally be classified in two main types, supervised (i.e. the target label is available), and unsupervised (i.e. there is no target label). In this particular case the authors utilise a supervised learning algorithm known as a gradient-boosted tree. A simple decision tree can be thought of as a flow-chart, where at each node/branch a decision is made based on whether a feature value satisfies a certain threshold. If repeated multiple times this allows the model to correctly predict an outcome. Gradient boosting is the process of identifying the situations in which the first tree performed poorly and subsequently constructing a second tree to predict the difference (i.e. the residual error) between the correct outcome and cases in which Tree 1 was wrong. This process is then repeated a number of times, and the predicted outcome is the weighted sum of a group, or ‘ensemble’, of these trees.

Artzi and colleagues begin by identifying a suitable source of EHR data, namely that of Clalit Health Services’ database which represents more than 50% of Israel’s adult population. The sample of women, and the timeframe over which they were at risk of GDM was defined by utilizing the birth dates of registered newborns. To account for potential inaccuracies in the logging of dates of birth, the potential gestation period was defined as 32 weeks prior to the date of birth and 7 weeks after. Exploiting the fact that the Glucose Challenge Test (GCT; the first of the two-stage test for GDM) is only ever used in pregnancy, and that national policy is that all women receive this test routinely during weeks 24-28 of their pregnancy, the authors were able to identify 588,622 pregnancies between 2010-2017, corresponding to 368,351 women.

One of the main strengths of this study is that the validation framework is particularly well-defined [7]. The authors use the 8 screening questions recommended by the National Institute of Health (NIH)[ii] as the baseline with which to compare their model against. The sub-set of most recent pregnancies, (i.e. ending in 2017 or later), was used as the hold-out/external validation dataset (n = 82, 678). Moreover, in an effort to assess geographic, and geo-temporal generalizability, the authors created subsets of their (hold-out/external) validation dataset consisting of: 1) the sub-set of women living in a specific area, i.e. Jerusalem (n = 46,002); and 2) the sub-set living in the aforementioned area and born most recently, i.e. 2017-2018 (n = 8,540). The tree-based model trained on the

unrestricted set of features (2,355 in total; were 295 features are available at the start of the pregnancy, with the remaining 2,060 generated by different process including laboratory (blood) tests and ultrasound scans that occur up to week 20), produced a highly accurate model (Area Under the Curve [AUC] 0.854), which was superior to the baseline (AUC 0.682) when tested on the full hold-out sample. Furthermore, on assessment of geographic and geo-temporal generalisation using the two described sub-sets, the AUC remained high (0.875 and 0.863, respectively).

Model interpretability, i.e. the ability to provide insight into the relationship between specific prognostic features and an individual's predicted outcome, is an ongoing challenge in ML. Gradient-boosted decision tree-based models, such as the one developed in this study, can exploit specific innate properties to address this issue. Specifically, they allow for a computationally efficient method by which to derive Shapley Additive Explanations (SHAPs)[8], an approach for improving interpretability derived from coalitional game theory, which can be used to quantify the contribution of each feature value to a prediction whilst also capturing dependency between the considered features. Rather unsurprisingly, the most important feature identified in this study was the GCT in the most recent pregnancy. However, what is especially creditworthy about how SHAPs have been deployed in this study, is that the authors used it to identify the most important features, which they subsequently demonstrate can be turned into a standalone 9-question tool with only a modest reduction in AUC (0.850 -> 0.799). However, it is worth noting that the vast majority of questions in the novel tool are in fact the same as that of the baseline NIH tool used for comparison [ii] (Table 1). The difference between the two is that the former captures this information as continuous variables whereas the latter expresses them as dichotomous features based on achieving a specific threshold. The results clearly demonstrate that the additional questions about previous pregnancies significantly improve performance in situations where the data is available. Whether the improvement from the baseline NIH tool (AUC 0.678) can be fully explained by the more granular feature values in the sub-set of women experiencing their first pregnancy is more challenging to determine from the presented data.

With regards to the literature as a whole, an often-noted limitation of evaluation schemes that are based on retrospectively collected data, such as the one described by Artzi et al., is that they do not necessarily reflect the dynamic nature of complex clinical pathways. For example, a recent systematic review of diagnostic tools illustrated that despite reported improvements in accuracy, fewer than 20% of the tools having undergone robust clinical evaluation via a Randomized Controlled Trial (RCT) actually resulted in improved patient-specific outcomes [9]. As such, the importance of evaluating any potential prognostic or diagnostic tool through prospective data collection, focusing on patient-specific (or operational outcomes in the context of non-inferiority), cannot be overstated [10].

The authors rightly suggest that their tool can be used as a cost-effective screening method by identifying women at a lower risk of GDM prior to conception or at a higher risk post-conception. This is especially relevant given that over 25% (>200,000) of the relevant pregnancies that might have been considered by the authors had to be excluded as there was no recorded GCT/Oral Glucose Tolerance Test (OGTT), suggesting there is a large currently underserved population who might benefit from EHR-based predictions of their healthcare needs. Further external validation (and calibration) however would potentially be required given that the underlying EHR data used were limited to live births (GDM is a strong risk factor for stillbirth particularly with macrosomia, a baby that is large for gestational age; [11] and the fact that risk profiles vary across populations with diverse ethnic backgrounds.

In summary, the study of Artzi et al. exemplifies how EHR and ML can be leveraged to create risk stratification tools that can potentially be used as part of routine care. Further prospective studies are however required to validate and investigate the associations of the tool with patient outcomes.

**Acknowledgements**: BAM and SD are employees of The Alan Turing Institute and are supported by the EPSRC grant EP/N510129/1. ALD is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

## Resources

i) Diabetes in pregnancy: management from preconception to the postnatal period NICE guideline (NG3)
https://www.nice.org.uk/guidance/ng3

ii) Am I at risk for gestational diabetes? By NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development
https://www.nichd.nih.gov/sites/default/files/publications/pubs/Documents/gestational_diabetes_2012.pdf

## References

1. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben-Haroush A, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med. 2020;26: 71–76.

2. Committee on Practice Bulletins—Obstetrics. ACOG Practice Bulletin No. 190: Gestational Diabetes Mellitus. Obstet Gynecol. 2018;131: e49–e64.

3. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health. 2019;1: e63–e77.

4. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc. 2019;26: 1545–1559.

5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25: 44–56.

6. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and AI research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness. arXiv [cs.CY]. 2018. Available: http://arxiv.org/abs/1812.10404

7. Mateen B, Sonabend R. All I want for Christmas is…Rigorous validation of predictive models to prevent hasty generalisations. Significance. 2019;16: 20–24.

8. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2: 749–760.

9. Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, Ioannidis JPA. Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol. 2014;67: 612–621.

10. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA. 2020. doi:10.1001/jama.2019.20866

11. Stacey T, Tennant P, McCowan L, Mitchell EA, Budd J, Li M, et al. Gestational diabetes and the risk of late stillbirth: a case-control study from England, UK. BJOG. 2019;126: 973–982.

**TABLE**

**Table 1: Comparison of the NIH Screening Tool[ii] and Novel 9-Q Tool developed by Artzi et al. [1].**

| | NIH Baseline | Novel 9-Q Tool |
|---|---|---|
| 1 | Overweight status: true if non-pregnancy BMI is >25 kg m–2. If there was no record of BMI before the pregnancy, it was considered as false. | What are your height and weight? |
| 2 | Family history of diabetes mellitus (DM): true if a first-degree relative (parent or sibling) has at least one diagnosis of DM, defined as any of the International Classification of Diseases (ICD-9) diagnosis codes in 250.x or 357.2 in the patient EHR. Only diagnoses available at pregnancy initiation were considered. | How many of your first-degree relatives have DM? |
| 3 | Age: true if the patient was at least 25 years of age at pregnancy initiation. | What is your date of birth? |
| 4 | History of pregnancy complication: the logic odds ratio operation of the following markers:<br>(A) History of GDM according to GCT and OGTT<br>(B) History of miscarriage or stillbirth, seen in the form of a diagnosis with ICD-9 632, 634.x, 635.x or 637.x.<br>(C) History of a liveborn baby of birth weight >4 kg. | Has a doctor ever told you that you have had?<br>1) GDM<br>2) A miscarriage |
| 5 | High blood pressure (BP), high cholesterol and/or heart disease: the logic odds ratio operation of the following markers:<br>(A) History of high BP, defined as two or more BP tests with systolic BP >140 or diastolic BP >90. Measurements taken during pregnancy were not included in this analysis. Recorded relevant ICD-9 of 401.x, 272.x or 390.x–449. x. | Has a doctor ever told you that you have had?<br>1) High BP<br>2) High cholesterol<br>3) Heart disease |
| 6 | History of polycystic ovary syndrome (PCOS): true if the patient has at least one diagnosis of PCOS (ICD-9 code 256.4). Only diagnoses available at pregnancy initiation were considered. | Has a doctor ever told you that you have had PCOS? |
| 7 | Problems with insulin or blood sugar: true if the patient has at least one diagnosis of prediabetes, either according to ICD-9 code 790.2x or following a HbA1c blood test in the range 5.7–6.4%. Only diagnoses and tests available at pregnancy initiation were considered. | If you have had a HbA1c blood test, what was the highest recorded value? |
| 8 | Are you Hispanic/Latina, African American, American Indian, Alaska Native, Asian American, or Pacific Islander? | ***No equivalent question*** |

| | | |
|---|---|---|
| 9 | ***No equivalent question*** | Have you ever given birth, if so, how many times? |
| 10 | ***No equivalent question*** | Did you undergo GCT or OGTT in your last pregnancy, if so, what are the results? |

*Note: Questions that are identical between the NIH screening questions and the Artiz et al. (2020) tool are highlighted using green colored rows. The questions presented in red colored rows are those that are principally similar, but the NIH tool dichotomizes them into true/false based on a pre-specified threshold whereas the novel 9-Q tool captures them as continuous variables. Question 8 was excluded at the outset as it was not relevant in the population of interest. It is to be noted that one of the challenges in assessing the clinical applicability of the 9-Q screening tool, is that the performance on the subset of women in their first pregnancy is not reported; results from this specific sub-sample would in effect nullify the importance/contribution of the last two questions (Q9-10), at which point it would be possible to ascertain to what degree the improved accuracy metrics can be solely explained to be a result of dichotomization of the input features.*