Predicting Gestational Diabetes Mellitus from Routinely-Collected Data in Electronic Health Records

Mark Germaine, BSc, MSc

Supervised by Dr. Graham Healy & Dr. Brendan Egan

Dr. Simon Caton, University College Dublin

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

Submitted to Dublin City University, January 2026

Declaration:

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work. I hereby certify that no Generative Artificial Intelligence (Gen AI) tools have been used in the creation of the thesis.

Signed: Mark Germaine

Date: 06-01-2026

ID No: 21268937

Acknowledgements

This PhD would not have been possible without the spirit of collaboration and support that defined so much of the work, and for that, I am deeply grateful.

First and foremost, I would like to express my sincerest gratitude to my supervisors, Dr. Graham Healy and Dr. Brendan Egan. Their guidance throughout this journey has been invaluable. It is a rare privilege to have a supervisor who is always available at the end of the phone to field any query, and I consider myself incredibly fortunate to have had that experience.

I owe a particular debt of gratitude to Dr. Amy O'Higgins of the UCD Centre for Human Reproduction. Without her partnership and expertise, the work contained within this thesis simply would not have been possible. I must also extend a professional and sincere acknowledgement to the research team at the Monash Centre for Health Research & Implementation. Collaboration has been a key theme of this project, and they were the only research team who were so open to an Open Science collaboration.

The journey of a PhD is supported by many, often behind the scenes. I am immensely grateful to the team at ML Labs, not least Angela Lally, Vicky Flanagan, Jonathan Costello, Antonella Ferrecchia, and Carla Naltchayan, who ensured that the entire process moved as smoothly as possible from an organisational point of view. Their support certainly made the process much simpler. My thanks are also extended to everyone at the Coombe Hospital who was so helpful during the crucial data collection phase, from the team at the booking clinic and the ultrasound clinic to the dedicated members of the UCD centre research team.

For their specific and expert support with the meta-analysis, a cornerstone of this work, I offer a further thanks to Dr. Mika Manninen and Dr. Ian Darragh.

Finally, on a more personal note, this work would have been impossible without the unwavering support of my family over the last four years. Anto, Debbie, I am especially grateful for their patience as I commandeered an office space that made my work so much easier. And to my partner, Diana, a massive thank you. You have been my biggest cheerleader throughout it all.

Research Outputs Arising From This Thesis

Publications

Germaine, M., Healy, G., & Egan, B. (2024). Lack of Data Sharing Despite Data Availability Statements in Studies Using Machine Learning Models for Prediction of Gestational Diabetes Mellitus. *Diabetes Care*, *47*(10), e78-e79. https://doi.org/10.2337/dc24-1483

Germaine, M., Darragh, I. A.J., Healy, G., Manninen, M., Egan, B. (2025). Machine Learning Models for Early Gestational Diabetes Prediction Using Electronic Health Records: Systematic Review and Multi-Level Meta-Analysis. (Manuscript # DMRR-25-RE-566)

Germaine, M., O'Higgins, A. C., Egan, B., & Healy, G. (2025). Label Accuracy in Electronic Health Records and Its Impact on Machine Learning Models for Early Prediction of Gestational Diabetes: 3-Step Retrospective Validation Study. *JMIR Medical Informatics*, *13*(1), e72938. doi:10.2196/72938

Germaine, M., O'Higgins, A. C., Egan, B., & Healy, G. (2025). Evaluation of Machine Learning Models for Early Prediction of Gestational Diabetes Using Retrospective Electronic Health Records from Current and Previous Pregnancies. *BMJ Digital Health & AI*, 2025-05. doi: 10.1136/bmjdhai-2025-000089

Germaine, M., Belsti, Y., O'Higgins, A. C., Egan, B., Teede, H., Healy, G., & Enticott, J. (2025). Reciprocal External Validation of GDM Risk Prediction Models Using a Machine Learning Model-Exchange Framework. (Manuscript # CIBM-D-25-09332) https://ssrn.com/abstract=5287628

Cosgrave, E. (joint first author), **Germaine, M.** (joint first author), Naughton, P., Brennan, M.M., Kearney, P.M., Healy, G., Egan, B., Turner, M., Buckley, C.M., O'Higgins, A.C. (2025). Trends in prevalence of and risk factors for obesity during pregnancy in Ireland: Longitudinal evidence from a large tertiary maternity hospital. Submitted to *BMJ Journal of Epidemiology and Community Health.*

Turner, C., McIntosh, T., Gaffney, D., **Germaine, M**., Hogan, J., & O'Higgins, A. (2025). A 10-year review of periconceptual folic acid supplementation in women with epilepsy taking antiseizure medications. *The Journal of Maternal-Fetal & Neonatal Medicine*, *38*(1), 2524094. https://doi.org/10.1080/14767058.2025.2524094

Conferences

Exploring the utility of machine learning in early classification of gestational diabetes risk: a pilot study. *6th Annual All-Ireland Postgraduate Conference in Sport Sciences, Physical Activity and Physical Education*. Dublin City University, Dublin, Ireland. 9th September, 2022.

Exploring the utility of machine learning in early classification of gestational diabetes risk: a pilot study. *Environmental Impacts on Pregnancy and Offspring Outcomes: Lessons Learned and Avenues for Intervention, The Physiological Society*. Coin Street Conference Centre, London, UK. 29th September 2022.

Trends in prevalence of obesity in pregnancy in Ireland: Longitudinal evidence from a large tertiary maternity hospital 2013-2022. *31st European Congress on Obesity* | 12-15 May 2024 - Venice, Italy. *Obes Facts* (2024) 17 (Suppl. 1): 7–515. https://doi.org/10.1159/000538577

1968-LB: Early prediction of gestational diabetes mellitus using electronic health records and machine learning. *84th Scientific Sessions American Diabetes Association*. Orlando, Florida, USA. June 2024. Diabetes 2024;73(Supplement_1):1968-LB. https://doi.org/10.2337/db24-1968-LB

Validation of gestational diabetes mellitus diagnosis in electronic health records. *Joint Irish-UK Endocrine Meeting 2024*. Belfast, Ireland. 14-15th October 2024 *Endocrine Abstracts* (2024) 104 P101 | DOI: 10.1530/endoabs.104.P101

Machine Learning for Gestational Diabetes: Challenges in Diagnosis, Early Prediction and Treatment Pathway Optimisation. *IDF Congress 2025 World Diabetes Congress*. Bankgok, Thailand. 7-10th April 2025.

Table of Contents

List of Abbreviations

American College of Obstetricians and Gynecologists	ACOG
American Diabetes Association	ADA
Area under the Receiving Operating Characteristic Curve	AUC
Artificial Intelligence	AI
Average Precision	AP
Body Mass Index	BMI
C Statistic	AUC
Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies	CHARMS
Clinical Decision Support Systems	CDSS
Clinical Team Database	CTD
Confidence Interval	CI
Continuous Glucose Monitoring	CGM
Decision Curve Analysis	DCA
Diabetes Canada	DC
Electronic Health Records	EHRs
European Union	EU
Explainable AI	XAI
Explainable Boosting Machine	EBM
Extreme Gradient Boosting, XGBoost	XGB
False Negative	FN
False Negative Rate	FNR
False Positive	FP
False Positive Rate	FPR
Fasting plasma glucose	FPG
Feature Agnostic Model	FAM
GDM label as recorded in the CTD	CTD-GDM
GDM label as recorded in the EHR	EHR-GDM
General Data Protection Regulation	GDPR
Gestational Diabetes Mellitus	GDM
Glycated Haemoglobin	HbA_{1c}
Hyperglycemia and Adverse Pregnancy Outcomes	HAPO
Identifiers	IDs
Inter Quartile Range	IQR

International Association of Diabetes in Pregnancy Study Groups	IADPSG
International Classification of Diseases	ICD
International Diabetes Federation	IDF
International Standard Classification of Occupations	ISCO
Large-for-gestational-age	LGA
Logistic Regression	LR
Machine Learning	ML
microRNAs	miRNAs
National Diabetes Data Group	NDDG
National Institute for Health and Care Excellence	NICE
Neonatal intensive care unit	NICU
Noise at Random	NAR
Nulliparous Model	NPM
Oral Glucose Tolerance Test	OGTT
Precision-Recall curve	PR-AUC
Prediction model Risk Of Bias ASsessment Tool - AI	PROBAST+AI
Preferred Reporting Items for Systematic reviews and Meta-Analyses	PRISMA
Random Forest	RF
Randomised Controlled Trial	RCT
Receiver Operating Characteristic	ROC
Reciprocal External Validation	REV
Risk of Bias	RoB
Sequential Pregnancy Model	SPM
SHapley Additive exPlanations	SHAP
Systematized Nomenclature of Medicine Clinical Terms	SNOMED
Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis	TRIPOD
Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis - Systematic Review and Meta-Analysis	TRIPOD-SRMA
True Negative	TN
True Negative Rate	TNR
True Positive	TP
True Positive Rate	TPR
Type 2 diabetes mellitus	T2DM
Validated GDM label	VAL-GDM

List of Tables

List of Figures

Predicting Gestational Diabetes Mellitus from Routinely-Collected Data in Electronic Health Records

Mark Germaine

Abstract

Machine learning (ML) techniques are increasingly applied to electronic health records (EHRs) for earlier clinical insights. Gestational diabetes mellitus (GDM), currently screened at 24–28 weeks, is ideal candidate for these models because demographic and clinical data are available before screening takes place. Therefore, this thesis examined whether first-trimester and obstetric EHR data can identify women at elevated GDM risk before standard screening.

A systematic review and meta-analysis of 38 studies (>2 million pregnancies) established the existing evidence base. Data from the Coombe Hospital EHRs were processed, including validation of GDM diagnoses against a clinical team database. ML and statistical models were developed and internally validated using first-trimester data (n=27,561) and data from previous pregnancies (n=4,005). A novel reciprocal external validation framework was implemented in collaboration with an Australian research group, to assess model transportability without direct data sharing. Finally, a developed prognostic model was prospectively validated in a clinical setting at the Coombe Hospital. This structured progression from foundational data issues to clinical application reflects a deliberate effort to address the multifaceted challenges beyond mere algorithmic performance that often hinder the translation of prognostic models into practice. Findings are reported as result followed by 95% confidence interval.

Key findings revealed moderate yet heterogeneous discrimination in the published literature to date (pooled AUC 0.75, 0.71-0.78; I^2 ~99.6%), with complex algorithms offering no advantage over logistic regression (Chapter 2). Discrepancies in the recording of GDM were found between EHRs and the CTD (14.3% FNR, 2.3% FPR), though this had minor impact on model development (Chapter 4). Incorporating data from previous pregnancies improved model performance relative to first-trimester data (AUC ~0.88 vs ~0.82; intercept ~0.040 vs ~0,035; slope ~1.032 vs 1.016), with past pregnancy alone achieving good performance (AUC ~0.86; intercept 0.050; slope ~0.984) (Chapter 5). External validation highlighted transportability challenges: declining AUC in both Irish and Australian models, with impaired calibration (Chapter 6). The prospective clinical validation showed the prognostic model achieved moderate discrimination (AUC 0.762, 0.681-0.837) and acceptable calibration (intercept 0.21, -0.15-0.57; slope 0.808, 0.53-1.08) in real-world use, resulting in 1 in 5 GDM cases identified 10-12 weeks earlier (Chapter 7). The consistent performance decline from internal to external and prospective validation is consistent with an optimism bias that is present in many prognostic modelling studies and the necessity of rigorous, multi-stage testing.

In conclusion, ML models, particularly those leveraging previous pregnancy data, show potential for early GDM risk prediction using EHRs. However, the successful clinical translation of these tools is critically dependent on data quality, multi-stage validation, and consideration of model transportability across populations. This thesis provides a comprehensive framework for developing and evaluating ML models in clinical settings incorporating EHR, highlighting the path from code to clinic.

Chapter 1

Machine Learning Prediction of Gestational Diabetes Mellitus: Introduction and Literature Review

1.1 INTRODUCTION

Gestational diabetes mellitus (GDM) is a form of diabetes presenting during pregnancy, characterised by any degree of glucose intolerance that is first recognised or initiates during the pregnancy period[1]. Nevertheless, this definition provokes debate owing to the inherent glucose intolerance associated with pregnancy, suggesting an alternative characterisation of diabetes as hyperglycaemia could be more precise[2]. The implications of developing GDM are significant, with substantial risks for both maternal and neonatal outcomes. These risks include complications during childbirth such as preterm birth, macrosomia, large-for-gestational-age infant, neonatal intensive care unit (NICU) admission, neonatal hypoglycaemia, and neonatal respiratory distress[3].

Aside from these immediate perinatal complications, GDM is associated with undesirable maternal outcomes such as gestational hypertension, pre-eclampsia, and elevated cardiovascular disease risk. Further, there are long term health implications for both the mother and the infant, including a reduced life expectancy and an increased risk of developing type 2 diabetes mellitus (T2DM) in subsequent years[4]. These implications could be potentially driven by epigenetic modifications[5] and intergenerational metabolic programming of metabolic disorders[6], thereby underpinning the importance of understanding and managing GDM effectively.

Globally, the reported incidence of GDM varies widely, spanning from 1% to over 30%. This diversity in incidence can, in part, be attributed to the absence of a uniform and globally accepted set of screening procedures and diagnostic criteria for GDM[7], a disparity that exists even within national boundaries[8]. The International Diabetes Federation (IDF) reports that one in every six pregnancies worldwide is impacted by diabetes, a trend that has been rising in recent years[9]. This rising prevalence of GDM imposes a substantial burden on public health resources. Costs associated with maternity care for pregnant women diagnosed with GDM are reported to be 34% higher than those incurred in average pregnancies[10]. Similarly, neonatal expenditure during the first year of life is typically greater for offspring born to mothers with GDM compared to those without[11]. With projections indicating that the global population affected by diabetes could surpass 250 million by 2025, this phenomenon poses a significant challenge for healthcare systems worldwide[12,13].

Despite the IDF recommendation for universal screening of all women at their initial antenatal visit to exclude the presence of pre-existing T2DM[9], the implementation of this recommendation is not universally applied. This is likely attributed to the economic cost, in

addition to logistical and resource challenges involved in conducting such comprehensive screening on a large scale. Rather, most regions predominantly employ a one-step approach as recommended by the International Association of Diabetes in Pregnancy Study Groups (IADPSG)[1]. As a result, diagnoses of GDM typically occur between the 24th and 28th weeks of gestation, a stage often too advanced for effective lifestyle interventions[14].

The Lancet highlighted the fundamental role of lifestyle factors (physical activity and nutrition) at the point of conception in influencing pregnancy outcomes and maternal-neonatal health in a series published in 2018[15–17]. Independent investigations have supported these findings, suggesting that lifestyle modifications could potentially mitigate the risk of GDM and deleterious maternal outcomes[18,19]. Further, the gestational period is frequently viewed as a 'teachable moment', a phase when women, even those habituated to sedentary lifestyles, are motivated to incorporate physical activity throughout the course of pregnancy[20]. Recognising this opportunity for lifestyle intervention and its consequential benefits regarding GDM prevention and maternal health improvement, it is important to identify those individuals most susceptible at the earliest time point. Such early identification provides a proactive avenue for intervention, thus potentially improving health outcomes for both mother and infant.

Given these challenges, the pursuit of alternative, more efficient risk prediction methodologies may be important in the area of diabetes. In this context, machine learning (ML), a subfield of artificial intelligence (AI), emerges as a potentially useful application. This technology utilises sophisticated statistical methodologies to forecast outcomes based on prior data training. The prospective integration of ML models into healthcare data analyses could provide a promising solution, potentially allowing for earlier detection and more effective management of conditions such as GDM. MLs potential for the early detection of GDM addresses a key challenge, the typical delay in diagnosis during pregnancy.

1.2 CURRENT APPROACHES TO DIAGNOSIS OF GDM

As mentioned previously, diagnostic criteria and screening methods are still not uniform between or even within nations[8] (Table 1.1). For example, countries such as Ireland and the UK implement selective screening for GDM with different diagnostic criteria and have prevalence rates of 12% and 4-5% respectively[21,22]. It is estimated ~40–50% of Irish women are never screened, with an estimated 16% of GDM diagnoses may be missed[23]. However, not discounting the potential ethnic variation in rates of diabetes, in countries such as Singapore and China where GDM is screened universally, rates of 15-20% can be observed[24–26].

GDM is most commonly diagnosed using a 75g oral glucose tolerance test (OGTT). In recent times, the IADPSG was formed and issued new guidelines for diagnosing GDM during weeks 24-28 using an OGTT and fasting plasma glucose (FPG) thresholds of 5.1 mmol/L, 1-h plasma glucose threshold of 10 mmol/L and 2-h threshold of 8.5 mmol/L[1]. However, the UK guidelines issued by the National Institute for Clinical Excellence (NICE) remain at odds with other countries in their continued support for selective risk-factor based[22] testing while the USA and Canada use the ADA[27] and ACOG[28] criteria. This can lead to inconsistent diagnosis of GDM depending on which criteria is applied to a population[29].

Aside from the variations in diagnostic criteria, discrepancies also extend to the screening methodologies (universal versus selective)[30], the timing of screening (12^{th} or 28^{th} weeks of gestation)[31], and the management of sample collection and processing[32]. These factors can significantly influence the creation and application of ML algorithms in this domain, as the target variable (GDM diagnosis) is prone to shift based on the adopted diagnostic criteria. Importantly, we must also acknowledge the fluid nature of diagnostic criteria for GDM. They are susceptible to revisions as they continually evolve in response to emerging scientific evidence and advancing research.

The diagnostic thresholds form the primary source of difference among these criteria. Hence, it is possible that pregnant women with identical plasma glucose values could receive varying diagnoses based on the geographical location of their tests. For instance, one study demonstrated considerable discrepancies in GDM diagnosis rates using three different diagnostic criteria. Using the IADPSG criteria, GDM was diagnosed in 53% of cases, while applying the ACOG and NICE criteria resulted in diagnostic rates of 35% and 18% respectively, the latter being approximately a third of the IADPSG rate[29]. Such inconsistencies carry significant implications for the development and implementation of ML algorithms in this field, considering the target (GDM diagnosis) may vary based on the selected diagnostic criteria. Furthermore, it's critical to acknowledge the evolving nature of GDM diagnostic criteria, subject to refinement as new research and data become available. An example of which is the IDF GDM Model of Care[9], which recommends screening all pregnant women for pre-existing diabetes at the first visit using a FPG, HbA_{1c}, or random glucose sampling. However, this practice is yet to achieve universal implementation.

The most commonly adopted diagnostic criteria set out by the IADPSG was developed on the basis of the findings from the Hyperglycemia and Adverse Pregnancy Outcomes (HAPO) study[33]. The multicentre study found that increased maternal glucose concentrations were associated with increased frequency of primary outcomes large birth weight, caesarean

section, neonatal hypoglycaemia and cord-blood serum c peptide, and an increase in each of the secondary outcomes. Interestingly, the study did not determine clearcut thresholds; instead, the frequency of these outcomes escalated with rising plasma glucose concentrations. This suggests a continuous risk where the likelihood of morbidity increases in line with increasing plasma glucose concentrations. Consequently, there could be considerable benefit in the early prediction of even 'mild' GDM, given that standard treatment of GDM has been demonstrated to mitigate complications associated with this disease[34].

Table 1.1. Comparison of the some of the common criteria used for the diagnosis of gestational diabetes mellitus

Criteria (mmol/L)	IADPSG	NICE	DC	ACOG*	NDDG*
Fasting Glucose	≥ 5.1	≥ 5.6	≥ 5.3	≥ 5.3	≥ 5.8
1-h Glucose	≥ 10.0	-	≥ 10.6	≥ 10.0	≥ 10.6
2-h Glucose	≥ 8.5	≥ 7.8	≥ 9.0	≥ 8.6	≥ 9.2
3-h Glucose	-	-	-	≥ 7.8	≥ 8.0
One/Two Step	One Step	One Step	Two Step	Two Step	Two Step

Abbreviations: IADPSG, International Association of the Diabetes and Pregnancy Study Groups; NICE, National Institute for Health and Care Excellence; DC, Diabetes Canada; ACOG, American College of Obstetricians and Gynecologists; NDDG, National Diabetes Data Group.
*100g oral glucose tolerance test (OGTT) used in place of 75g oral glucose tolerance test and glucose values must exceed two thresholds.
One Step approach: One 75g OGTT used to screen and diagnose.
Two step approach: 50g glucose challenge test used to screen patients and then an OGTT administered as step two.

1.3 IMPACT OF EARLY VS LATER INTERVENTION IN GDM

A series of landmark papers in The Lancet in 2018 made it clear that health and lifestyle before conception shape pregnancy outcomes and even the long-term health of children. Stephenson et al.[15] synthesised evidence from diverse settings showing that a woman who is healthy at conception had half the risk of gestational diabetes (OR 0.45, 95% CI 0.28–0.75). They noted that poor nutrition and obesity are widespread among women of reproductive age worldwide (including adolescents), and that awareness of the importance of preconception health remains low[35,36]. By contrast, interventions started after conception yield only marginal gains. An individual-patient meta-analysis of 36 antenatal diet-and-exercise trials showed just a 0.7kg reduction in gestational weight gain and a 9% decline in caesarean births (OR 0.91, 95% CI 0.83–0.99), suggesting that by the time pregnancy is underway, it may be “too little, too late”[37]. Fleming et al.[16] extended this concept by examining the biological mechanisms: exposures around the time of conception, from maternal overnutrition or undernutrition to

paternal health factors, can tigger developmental and epigenetic changes in the embryo that influence the infant's lifetime disease risk. Their review underscored that the evidence for such periconceptional programming is so strong that it "calls for new guidance" on improving parental health before pregnancy. Barker et al.[17] then addressed the next logical question, how to act on this knowledge. They reviewed intervention strategies to improve nutrition and health behaviours prior to conception, advocating for a multifaceted approach. The paper called for a "social movement" to strengthen political resolve for wide-scale intervention. Together, these papers shift the focus of prevention to the period before pregnancy, or at least at conception, identifying it as a critical window for improving maternal and offspring health.

The efficacy of lifestyle interventions *during* pregnancy remains somewhat ambiguous, with several strategies leading to limited benefits, both in terms of dietary[38,39] and physical activity interventions[40–42]. This ambiguity could partially be due to the intervention's commencement time, which is often during a more advanced stage of pregnancy. Two meta-analyses[43,44] support this, reporting that the effectiveness of lifestyle interventions is improved when they are initiated in the first trimester. A separate systematic review of 44 randomised controlled trials (RCT) found that when usual care was combined with lifestyle intervention (diet, exercise or both), gestational weight gain fell by 1.42kg and the interventions reduced pre-eclampsia risk by 26% and shoulder dystocia by 61%[45]. Further, when such lifestyle interventions are enacted, they are projected to provide a cost-saving benefit of ~$3855 per patient to the healthcare system[46].

These findings suggest that identifying and treating GDM earlier in pregnancy should improve outcomes. A recent meta-analysis[47] pooled RCTs comparing early vs routine mid-pregnancy screening, concluding early screening did not significantly reduce the risk of large-for-gestational-age (LGA) births. However, in a subgroup of trials (n=3) where universal screening at the first visit was implemented (using HbA_{1c} to identify mild hyperglycaemia), early treatment did result in significantly better outcomes. In those trials, the LGA rate was only ~2.3% in the early screening group versus 9.1% with routine later screening. Further, the 2023 TOBOGM multicentre RCT[48] (n=802) demonstrated that intervention before 20 weeks delivers a modest yet statistically significant benefit for infants. Early intervention lowered serious neonatal complications to 24.9% vs 30.5% with standard care, however, it didn't change rates of pregnancy-related hypertension or affect neonatal lean body mass.

The literature therefore presents a nuanced picture. On one hand, identifying high-risk women early and implementing lifestyle changes can prevent a subset of GDM cases and improve maternal outcomes[45]. Early onset GDM clearly signals higher risk for mother and

baby[49,50], so from a pathophysiologic standpoint, earlier intervention is justified. On the other hand, universal early treatment of GDM (using current thresholds) has shown only modest outcome improvements in RCTs[48]. This suggests that we may need better ways to target those truly in need of early intervention, and perhaps more precise or intensive interventions for that group. Therefore, if ML models can accurately identify women at an increased risk for GDM and related adverse outcomes, it may allow for earlier clinical intervention. Thee ML models could not only enhance delivery outcomes but also contribute significantly to improved maternal and infant health.

1.4 POTENTIAL ROLE OF ML IN GDM

ML constitutes a specialised subfield of AI that employs advanced statistical methods to generate predictions based on (large) data inputs[51]. The machine iteratively refines its model through a continuous process of learning from data without being explicitly programmed[52] and making subsequent corrections to enhance its predictive performance. The utility of ML has become increasingly recognised as the emergence of large datasets aligns with the computational capabilities necessary for learning from such data. This emergence has led to the integration of ML across domains, ranging from computer vision to healthcare applications, perhaps most recognisably in medical imaging techniques. Presently, ML has found relevance in several healthcare sectors, as described in Table 1.2. Collectively, the applications outlined in Table 1.2 underscore the potential of ML in healthcare domains, including GDM. For instance, clinical decision support systems (CDSS) using ML algorithms may improve clinician decision-making by prioritising patients for screening or facilitating the stratification of patients into appropriate treatment cohorts. Further, predictive modelling via ML may aid in the early identification of individuals at risk of metabolic diseases. Lastly, in light of the long term morbidity of GDM for both mother and infant, ML's capacity for population health management can play a role in detecting trends and risk profiles among those predisposed to the disease.

Table 1.2. Applications of machine learning modelling in healthcare

Use Case	Explanation	Example
Clinical decision support[53,54]	This involves the use of ML algorithms to analyse patient data and provide probabilistic predictions to healthcare professionals for prognosis, diagnosis, and treatment. To suggest a course of treatment, for example, an ML system would examine a patient's medical history, symptoms, and/or test findings. ML can assist healthcare professionals make better decisions and raise the standard of treatment or reduce the time taken to make a decision by offering recommendations that are supported by the available data.	Clinical decision support systems have been demonstrated to improve health care process measures related to performing preventive services and prescribing therapies[55].
Predictive modelling	This involves using data from sources such as electronic health records (EHRs)[56] and other sources that are analysed using algorithms to forecast outcomes, such as the probability that a patient will develop a specific condition or respond to a specific treatment. Healthcare professionals can identify individuals who are at risk for particular conditions using predictive modelling and take early action to treat or prevent the condition. By determining which treatments are most likely to be effective for a specific patient, it can also assist healthcare professionals in making better educated treatment decisions. However, notably there may be biases in EHRs that could result in socioeconomic disparities in health care[57].	An example of this in practice is demonstrated by Barack-Cohen et al.[58] who mined through 1.7 million EHRs to build a ML model of suicidal behaviour, critically making predictions 3-4 years in advance of future behaviour
Finding and developing new drugs[59]	The analysis of chemical compounds and the prediction of their potential medicinal efficacy have traditionally been long, complex, costly and depend on numerous factors, but ML algorithms can be leveraged to improve discovery. This can shorten the time and expense associated with the medication development process by assisting pharmaceutical companies in identifying good candidates for additional testing and development.	For example, this may include target identification and validation[60], identification of clinical/prognostic biomarkers[61] and analysis of digital pathology data in clinical trials[62].

Medical Imaging	ML algorithms can be used to analyse pictures from X-rays and CT scans in the field of medical imaging to help with diagnosis and therapy planning[63]. For example, a ML system might be trained to detect patterns in medical images that point to the presence of a specific ailment, enabling medical professionals to diagnose patients more precisely and suggest the best course of action.	Mayer McKinney et al.[64] developed an AI system that was capable of surpassing radiologists in breast cancer prediction based on mammograms However, caution should still be applied in this area as there can be some errors in how the models are sometimes trained, as there is not always consistency in how experts segment and label medical images thus reducing consistency of models[65].
Population health management[66]	With the rising incidence of chronic conditions, such as diabetes, identifying high risk patients as soon as possible becomes an important challenge to improve patient care and reduce costs. In order to find trends and patterns that can guide the creation of actions to promote public health, algorithms are used to analyse data on big populations. An ML system might be used, for instance, to spot trends in the prevalence of diseases or risk factors for specific disorders in a community, enabling public health organisations to develop targeted interventions to address these problems.	Chae et al.[67] used deep learning algorithms in combination with big data to predict infectious diseases in order to reduce the delay in existing surveillance systems. The deep learning algorithms were able to predict chickenpox outbreaks 1 week in advance of traditional reporting systems.
Personalised medicine	to create treatment plans that are unique to each patient's requirements and traits[68]. Healthcare professionals may achieve improved patient outcomes by using ML algorithms to analyse data on individual patients to determine the best treatment options for each patient, potentially leading to better patient outcomes.	This has been demonstrated in the case of colorectal cancer where linear regression models were developed with event-free survival analysis to predict the heterogeneity of signal transduction pathways on an individual patient level[68].

1.5 REVIEW OF ML RESEARCH IN GDM

As discussed, GDM diagnosis often occurs between the 24th to 28th weeks of gestation, too advanced for impactful lifestyle interventions[14]. Because of the varying diagnostic criteria, the first step when building an ML model for GDM is deciding whether to predict a binary outcome or a continuous measure. A classification model mirrors current clinical practice by answering a simple question: will this pregnancy meet diagnostic criteria for GDM? A regression model, in contrast, estimates plasma glucose concentrations along a continuum, letting clinicians grade risk and bypass shifting diagnostic cut-offs. Most published work still frames the task as classification, partly because the three-point OGTT produces discrete results. Yet regression remains an under-used alternative that could improve risk stratification by highlighting degrees of intolerance rather than a yes/no label. Consequently, the discussion in this thesis will primarily concentrate on this classification approach.

There is potential for data collected at the booking visit to be used an inputs to ML models that predict GDM and guide screening or treatment decisions, serving as CDSS tools[69]. Other studies predict even earlier, using preconception[26] records to identify risk before pregnancy begins, aligning with the recommendations put forth in the Lancet series. Taken together, these projects outline a 'maternal life cycle' approach[70], as demonstrated by Kumar and colleagues. ML models could be used at preconception[26], the initial antenatal appointment[71], and using perinatal visit data to predict those who may eventually develop type 2 diabetes mellitus[72], tracking metabolic risk from before conception through to the post-partum period.

Beyond diagnosis, ML also has potential utility in the management and treatment of GDM. For instance, ML algorithms may identify the need for pharmacotherapy or identify trends in plasma glucose concentrations that might signal a need for modifications in the treatment regimen[73,74]. While there are numerous potential applications of ML in GDM, the following section concentrates primarily on its utility in predicting the diagnosis of GDM during pregnancy. This includes an emphasis on studies that make predictions close to the initial antenatal visit, which could pave the way for earlier intervention. A glossary of commonly used terminology is presented in Table 1.3 to familiarise the reader before proceeding.

1.5.1 Evaluating GDM Risk Prediction Models

Across the literature, the standard measure for validating ML models for GDM prediction is the Area Under the Receiver Operating Characteristic Curve (AUC)[75]. The ROC curve plots the true-positive rate (TPR) against the false-positive rate (FPR) over every possible threshold, quantifying discrimination: the probability that a randomly chosen GDM case receives a higher predicted risk than a randomly chosen non-case. An AUC of 0.5 indicates no discriminatory power, values between 0.70-0.80 are usually deemed acceptable, 0.80-0.90 excellent, and above 0.90 outstanding[76] (Figure 1.1). Because AUC is threshold-independent and easy to interpret, it remains the primary yardstick for comparing models in this chapter.

However, it's worth noting that discrimination alone cannot assess clinical utility. Guidance on evaluating clinical risk prediction models recommend reporting calibration (how closely predicted risks match observed event rates) and decision-curve analysis (DCA), which translates model output into net clinical benefit[77–79]. Calibration, however, is arguably more critical for clinical utility as it refers to the agreement between the model's predicted probabilities and the observed outcome frequencies. A well-calibrated model that predicts a 20% risk for a group of women should find that, on average, 20% of those women actually develop GDM. Calibration is assessed visually with calibration plots and quantitatively with metrics such as the calibration slope (ideally 1.0) and intercept (calibration-in-the-large, ideally 0). Despite these recommendations, calibration plots, slope and intercept statistics, and DCA curves still appear far less often than AUC in GDM studies. The imbalance means that many published models look promising on paper but leave unanswered questions about whether their risk estimates are trustworthy or whether their adoption would improve care. Therefore, this thesis will report on both discrimination and calibration where possible, whilst acknowledging that it is an under reported statistic.

Table 1.3. Glossary of commonly used terminology in the area of machine learning risk prediction modelling

Term	Description
Artificial Intelligence	A branch of computer science dealing with the simulation of intelligent behaviour in computers.
Machine Learning	Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed[80].
Explainable AI (XAI)	Methods and techniques that make the behaviour, predictions and internal workings of an AI/ML system understandable to humans, thereby increasing transparency, trust and the ability to audit or debug the model.
Model	A program or algorithm that can find patterns or make decisions from a previously unseen dataset.
Neural Net	A machine learning algorithm comprised of layers of nodes, containing an input layer, hidden layers and an output layer.
Deep Learning	A family of machine learning methods which are based on neural networks, often containing many hidden layers, hence the term "deep".
Tree-based Models	A type of machine learning model that uses a hierarchical structure of nodes resembling a tree to make predictions based on input data.
Gradient Boosting Models	Ensemble methods that sequentially train many weak learners (almost always decision trees), each one focusing on the residual errors of the previous, and combine them via gradient-descent-style optimisation to produce a strong predictor (e.g. Gradient Boosting Machine)
Classification	A predictive modelling problem where the target of the model is a categorical class label as opposed to a continuous variable (e.g. Does the patient have diabetes? yes or no)
Regression	A predictive modelling problem where the target of the model is a continuous variable. For example, what is the expected fasting plasma glucose level measured in mmol/L?
Features	Features are variables (predictors) which will be used for the model to train and learn from. For example, anthropometric data could be used to infer diabetes risk and these data would be features in the model.
Training	Applying an algorithm to the features so that it can determine the best values for model weights and bias in order the minimise the loss function and improve the performance of the prediction.
Independent Test Set	A hold-out subset of the data (commonly 20–30%) that is not used during training or model selection, providing an unbiased estimate of performance on truly unseen data.
Internal Validation	Assessment of model performance using resampling techniques (e.g. k-fold cross-validation, bootstrapping) or a single hold-out split within the original dataset. Detects overfitting but does not test generalisability to new populations.
External Validation	Evaluation of a finalised model on an entirely independent dataset collected at a different time, place or population to judge generalisability and real-world applicability
Overfitting	The model too closely replicates the patterns in the data it is trained on and then performs poorly when it is evaluated on unseen data.
Underfitting	The model doesn't fully learn the relationship between the features and the target variables leading to a high error rate in the model when it is evaluated on both unseen data and on the data it's trained on.

1.5.2 Foundations of ML-Based GDM Prediction

ML and prognostic modelling research on GDM dates back almost two decades, with logistic regression (LR) often serving as the benchmark algorithm model[81]. However, the past decade has witnessed a surge in publications in this domain[82]. For instance, a recent meta-analysis examining studies on ML and GDM from 2004-2020 found that 40% of the included research was published in 2020 alone[83]. My own research presented in Chapter 2 demonstrates a similar pattern when examining prognostic models trained on EHRs. This increase in research underscores the scale of recent development in this area.

Early research in this field often involved the use of logistic regression models for prognostic predictions of GDM. For instance, van Leeuwen and colleagues developed a prediction model using patient history and medical characteristics with a multivariate LR model. This model was able to classify women early in pregnancy (<20 weeks) as high or low risk for GDM and reported an AUC of 0.77[84]. However, the approach used in this model differs slightly from modern ML validation practices, which involve the use of validation and test datasets. In addition, ML iterations of LR involve using the gradient descent algorithm toi minimise the cost function[85].

These practices usually involve partitioning the available data into two or three segments: a training set, a validation set, and an independent test set. The model is trained on the training set, tuned on the validation set, and finally tested on the test set to evaluate its performance on unseen data. For instance, in contrast to van Leeuwen's approach of comparing the predicted mean to the observed data, others[86] will develop the model using 70% of the data and then validate the results on unseen data (the remaining 30%). This strategy has been suggested to help to avoid overfitting[87], which can improve the model's generalisability to real-world data. Without such measures, model performance often tends to be overestimated in practice[88]. This difference in approach could be observed in prognostic models developed prior to 2016, which demonstrated acceptable levels of discrimination and calibration but varied in terms of quality and consistency[89]. However, recent evidence has suggested that having a large enough sample size and using validation or bootstrapping methods, utilising the entire dataset, yield results that generalise just as well[77]. In practice, rigorous validation (e.g. cross-validation, bootstrap sample and external validation) is crucial to mitigate optimism bias and ensure the model’s predictions hold up on new patients[90,91].

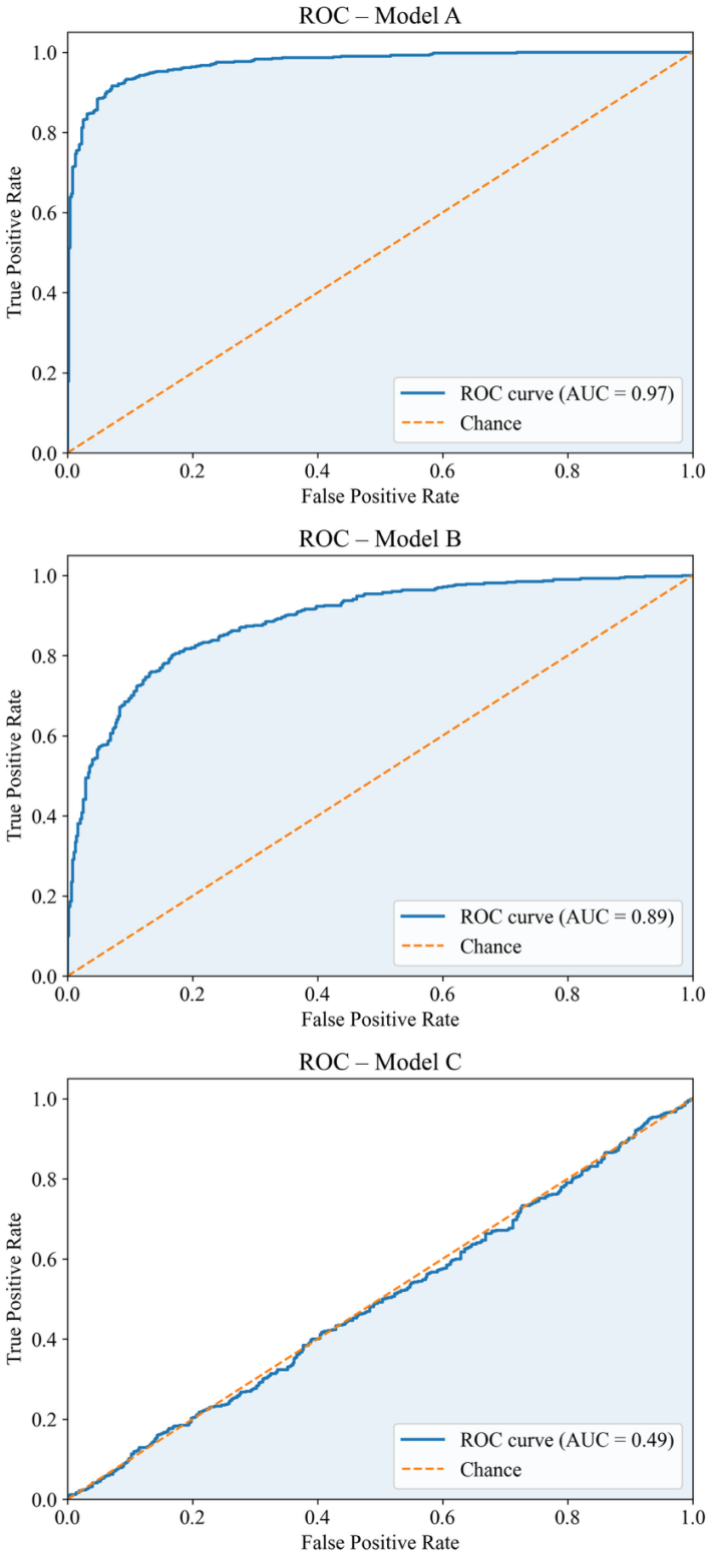


Figure 1.1. Receiver operating characteristic (ROC) curves for three dummy models. Panels A–C plot the true positive rate (sensitivity) against the false positive rate (1-specificity) across all decision thresholds for each classifier. The blue curve traces the model's performance; the light blue shaded area represents the region contributing to the area under the ROC curve (AUC), a global measure of discrimination. The red dashed diagonal indicates random classification (AUC = 0.50). Model A demonstrates outstanding discrimination with an AUC of 0.95. Model B shows excellent discrimination with an AUC of 0.83. Model C performs no better than chance (AUC = 0.48).

1.5.3 Distinguishing Diagnostic from Prognostic Modelling

The difference between diagnostic and prognostic models is significant in the field of medical research. Diagnostic models are intended to estimate the likelihood of a disease currently existing, while prognostic models aim to calculate the risk of a disease occurring in the future[92]. In the realm of ML research focused on GDM, most of the work has centred on prognostic models rather than diagnostic ones[83], risk prediction models. This focus on prognostic models means that the research has largely been oriented toward identifying women who are at a high risk of developing GDM before it manifests. However, diagnostic models could also play a significant role, particularly in terms of reducing the invasive nature, time, and biochemical cost of current diagnostic methods. Yet, there have been relatively few studies exploring the use of ML algorithms as an alternative approach to the established diagnostic criteria for GDM. As we move forward, it will be interesting to see how these models evolve and potentially transform the approach to diagnosing this condition.

In an attempt to further refine diagnostic capabilities, the combination of ML with continuous glucose monitoring (CGM) has also been explored in diagnosing GDM or identifying potentially missed diagnoses. In one pilot study, pregnant women between 12 to 35 weeks of gestation were recruited and underwent an OGTT if they had not done so already[93]. The study proposed an intriguing triangulation method for diagnosing GDM that made use of OGTT, CGM, and demographic risk factors. The researchers highlighted potential increased false positive rates with the use of OGTT alone, a conclusion drawn from the observation of low glucose variance and risk factors. However, the results of this pilot study should be interpreted with caution due to a potential confounding factor. A notable discrepancy was seen in the timing of the CGM procedure; the majority of women diagnosed with GDM underwent CGM after their OGTT (31/34), whereas most of those identified as having normal glucose tolerance had their CGM before the OGTT (24/26). This variation in timing could potentially introduce uncertainty in the accuracy of the proposed model, thereby necessitating further investigation.

The application of ML in diagnosing GDM presents promising opportunities, but the evidence remains somewhat preliminary and mixed. Novel methods leveraging ML in conjunction with biomarkers like microRNAs or CGM have shown potential to enhance the diagnostic accuracy of GDM. However, while ML holds promise for improving the diagnostic capabilities for GDM, more extensive studies are needed to substantiate these early findings and ensure the reliability and generalisability of these emerging methodologies.

1.5.4 Early Pregnancy Prediction Models

At the forefront of research within the field of GDM prognosis lies the application of ML techniques to the analysis of EHRs. EHRs typically consist of patient histories, inclusive of details on prior pregnancies, medical conditions, and potentially laboratory test results, constituting a valuable resource of potential risk factors (predictors). Making use of this resource, ML algorithms have been developed from these data, effectively identifying key risk factors of GDM and predicting the risk for patients to develop GDM[94–96]. This approach couples predictive analytics with clinician facing CDSS to enhance clinical decision-making.

A recent meta-analysis of prognostic ML models for GDM reviewed 25 studies published between 2004 and 2020[83]. The pooled AUC curve for all models reached 0.849, with the first trimester and second trimester predictions netting AUCs of 0.836 and 0.867, respectively. These results indicate improved discrimination as the pregnancy progresses but at the cost of a shorter window for intervention, possibly compromising the intended benefit for early prediction.LR accounted for 63% of the models and achieved a pooled AUC of 0.815, whereas non-LR methods reached 0.889. These findings support a separate review of pregnancy care models in which non-LR algorithms outperformed LR, leading the authors to recommend revisiting existing LR models with alternative techniques[97]. Evidence also suggests that the performance gap widens in larger cohorts: with sample sizes above 10,000, gradient-boosting machines have demonstrated greater performance than logistic regression[98]. However, given the rapid increase in the research in this area, these findings need to be revaluated with the published literature to date (Chapter 2).

Including LR in ML searches highlights a persistent validation problem. While LR is technically an ML algorithm, some researchers still deploy it mainly as a conventional statistical tool, often without the rigorous evaluation expected of predictive models. In a recent review of 109 GDM papers labelled as ML studies, 61 reported no validation at all and a further 19 tested performance only on the training data, leaving just 29 that met minimal ML validation standards[99]. Such findings suggest that poorly validated LR models should be excluded from discussions of genuine ML applications in GDM. Models that lack any form of validation (internal or external) often overestimate performance; excluding them ensures that we base conclusions on reliable, clinically sound evidence. In other words, filtering out studies that did not validate their models prevents misleading optimism in the overall assessment of predictive model accuracy.

1.5.5 Population Diversity and Model Transportability

The demographic heterogeneity of the populations on which the models are trained is another factor which should be considered. As highlighted previously, there are considerable disparities in GDM rates across various geographic locations and ethnicities that cannot be accounted for solely by testing criteria. A case in point is California, where GDM prevalence is markedly higher in Asian women (~10.5%) as opposed to Caucasian (4.5%) or African-American women (4.4%)[100], despite living in the same state with the same testing procedures. This discrepancy in Asian populations bears significant implications considering the majority of recent research has been conducted in these demographics. Indeed, a recent meta-analysis revealed that 56% of all models were trained on Asian populations, a figure which escalates to 91% for publications from 2019 onwards[83].

Tools built on such narrow data often mis-calibrate when applied elsewhere, disadvantaging under-represented groups. Responding to this concern, updated methodological frameworks now require explicit evaluation of subgroup performance, for example, the TRIPOD+AI guideline mandates reporting of discrimination and calibration across relevant strata[101], while PROBAST+AI checks that the data aligns with the target population[102]. Analyses of prediction studies confirm that models derived from homogeneous cohorts frequently show high risk of bias and limited transportability[103]. Further evidence from medical imaging research demonstrates that apparent performance can collapse when algorithms are tested in demographically distinct settings, urging multi-centre, multi-ethnic development pipelines[104]. A broader perspective likewise calls for routine auditing of error rates across sensitive attributes to ensure equitable deployment[105]. Collectively, these publications argue that future GDM models must be trained and externally validated in diverse populations, with transparent reporting and bias monitoring, to achieve safe and fair clinical adoption.

1.5.6 Invasive vs Non-Invasive Features

ML studies now test whether GDM risk can be predicted with inexpensive, non-invasive data alone, or whether adding early biochemical markers is worth the extra cost and delay. Simple anthropometric rules based on body-mass index (BMI) and waist circumference seldom exceed an AUC of 0.63 (95% CI 0.62-0.65)[106]. In contrast, a recent systematic review identified 16 models that used only history, demographic and anthropometric variables; ten were externally validated and achieved AUCs of 0.70-0.80, with maternal age, BMI and family

history of diabetes the most frequent predictors[99]. Though results are modest, they indicate an improvement over a baseline established using BMI and waist circumference.

Evidence that biochemical tests add value is growing. Du and colleagues[69] built three distinct models: a full feature set (AUC 0.79), a subset restricted to routine antenatal tests (0.74), and a lab-free version (0.66) used via a mobile application. This outcome implies that incorporating basic biochemical features may provide incremental benefit. Across studies that combine invasive and non-invasive predictors and undergo independent testing, AUCs of 0.80-0.90 are typical; higher figures almost invariably come from inadequately validated analyses[99]. Consistent with a recent biomarker review, early FPG or HbA_1c usually confers the largest single improvement, reflecting their direct link to glucose tolerance[26,107].

Early pregnancy lipid profiles may deliver a similar boost. In a pre-conception algorithm (AUC 0.89) two of the five strongest predictors were the fatty-liver index and the TG:HDL cholesterol ratio[26]. Supporting this, plasma lipids, when sampled during the first trimester, have been found to predict GDM, irrespective of BMI and ethnicity, using tree-based methodologies with an AUC of 0.880 (95% CI 0.80-0.95)[108]. This relationship aligns with findings that GDM-diagnosed women exhibit higher TGs and lower HDL cholesterol, seemingly associated with maternal obesity[109].

MicroRNAs (miRNA), an invasive measure, have demonstrated promise in producing high AUCs. Some miRNA have generated AUCs as high as 0.91. However, these studies have faced issues with inconsistent miRNA identification. This discrepancy may partly stem from measurement challenges; the miRNAs of interest might not be sufficiently abundant in the samples, resulting in their loss during analysis. For instance, while mir-23a resurfaced in a recent study, its AUC was only 0.65 (95% CI 0.4-0.8)[110], underscoring some of the persistent challenges in this otherwise promising research area. Taken together, non-invasive models outperform anthropometric baselines but appear to trail validated algorithms that include a small set of early biochemical markers, particularly glucose and lipid measures

1.5.7 Pre-Pregnancy Risk Stratification

Several studies have shown that common clinical variables, available before or at the first antenatal visit, can identify future GDM with clinically useful accuracy. Artzi et al.[94] utilized EHRs from Israel's largest healthcare provider. From the 2,355 features available in their dataset, 295 were accessible at the onset of pregnancy. Although their model's performance improved with the inclusion of data gathered during pregnancy, the researchers

achieved a promising AUC of 0.799 using just nine features obtained from questionnaires at pregnancy's outset. Using a U.S. nulliparous cohort, Donovan et al.[111] developed a five-factor score (age, BMI, ethnicity, family history of diabetes, and chronic hypertension) that achieved internal and external AUCs of 0.73 (95% CI 0.728-0.735) and 0.71 (95% CI 0.672-0.749), respectively. In Australia, Schoenaker et al.[112] combined nine demographic and lifestyle variables and reported an internally validated AUC of 0.79 (95% CI 0.76-0.83). Collectively, these studies show that risk can be quantified before conception with data already present in primary-care records. Notably some of these preconception models use clinical data that is routinely collected during the first antenatal visit. Thus, there is an argument that other clinical based risk prediction models that exclusively use clinical features could also be implemented as preconception risk prediction models.

Adding a small number of laboratory indices further improves discrimination. A Singaporean cohort[26] incorporated biochemical data into their preconception models, identifying HbA_1c as a powerful predictor of GDM (AUC 0.81) and demonstrated that its inclusion in a five-variable model increased performance to 0.93. More recently, a study in California[73] incorporated 68 features accessible one year prior to pregnancy (level 1) and 26 features available at the last menstrual period before GDM diagnosis (level 2), the majority of which were questionnaire-based. In this case, the model aimed to predict GDM treatment rather than diagnosis, and the resulting AUC from the level 1 model was 0.634 (95% CI 0.615–0.653), increasing to 0.648 (95% CI 0.630–0.667) with the addition of level 2 features. This illustrates diminishing returns when many weak predictors are added to a purely clinical model.

Women with a previous GDM pregnancy constitute a distinct high-risk group. In a German retrospective cohort, Hahn et al.[113] showed that the combination of pre-pregnancy overweight/obesity and a positive family history of diabetes identified recurrence with a positive predictive value of 96.6%. Chen et al.[114] subsequently applied gradient-boosting to thirteen early-pregnancy variables (including 1st trimester FPG) and obtained AUC values above 0.94, markedly outperforming conventional LR. Further, the conventional LR model indicated a predominance of lack of fit in the calibration plots compared to the other models. Because that algorithm incorporates biochemical data collected after conception, it cannot be classed strictly as pre-conception; nonetheless, it demonstrates that model performance can be refined as soon as the first laboratory results become available. Together, these studies suggest a pragmatic pathway: deploy inexpensive, non-invasive pre-pregnancy ML tools for early triage, and refine risk estimates when laboratory results become available.

1.5.8 Post-Diagnosis Treatment Stratification

Once GDM is diagnosed, prediction shifts from identifying risk to determining who will need pharmacotherapy. For example, a Californian study that combined diagnostic-day clinical data with CGM traces from the first post-diagnosis week improved discrimination from AUC 0.75 to 0.82, underscoring the added value of an early glycaemic profile[73]. This study suggests CGMs can assist in developing ML algorithms to analyse data, identify patterns indicative of GDM, and possibly predict glucose excursions, aiding in managing diabetes post-diagnosis[115]. An earlier study explored the prediction of required treatment post-GDM diagnosis, insulin or lifestyle modification, using FPG, 1-hour plasma glucose, and maternal characteristics in a classification tree[116]. The model showed the highest sensitivity to FPG and maternal BMI, resulting in an AUC of 0.77 (95%CI 0.73–0.81). Two UK investigations that mined CGM uploads from the GDm-Health app showed similar performance: one predicted escalation to insulin or metformin (AUC 0.80)[117], while the other predicted the number of high glucose readings in the upcoming days (mean squared error of 0.020)[115].

Complementing these primary studies, a systematic review found consistent added value for early FPG, HbA_1c and TGs, but noted that fewer than one-third of treatment-selection models had undergone any external validation[118] . Collectively, current evidence suggests that CGM-augmented or biochemistry-enhanced algorithms can stratify therapy needs with acceptable discrimination (~0.75–0.83), but independent, multi-centre validation and head-to-head comparisons with clinician judgement remain essential before these tools can be safely embedded in routine care.

1.5.9 Validation of GDM Risk Prediction Models

Rigorous external validation is now regarded as the final gateway between a promising algorithm and real-world use. Landmark papers by Ramspek et al.[90] and Steyerberg[119] emphasise that external validation, testing the unchanged model in a new, clinically relevant cohort, is the critical next step after development. The BMJ[77,91,120] emphasise this in a 3 part series. Part 1 shows how apparent (development set) discrimination can be over optimistic by ≥0.15 points on AUC, while Part 2 lays out five practical steps for an external validation study: securing an appropriate dataset, generating predictions exactly as specified, assessing discrimination, calibration and clinical utility, scrutinising key subgroups and reporting transparently. Part 3 then demonstrates that many published validations are underpowered and

argues for tailored sample size calculations that often require at least 100 events and 100 non-events, and sometimes far more, to give precise calibration estimates.

In a Dutch prospective multicentre study, Lamain-de Ruiter et al.[89] applied five first trimester logistic scores to 1,426 pregnancies and recorded a drop in AUC of 0.03-0.10 together with calibration slopes below 0.8, indicating that the original models were overfitted to their development populations. Using a similar Dutch cohort, Meertens et al.[121] evaluated eight published risk scores: only two retained an AUC above 0.70 and all required at least an intercept-plus-slope recalibration before they could be considered for use. More encouraging results emerge when development and validation are planned as one programme. Cooray et al.[122] built the PeRSonal-GDM XGBoost model using internal-external cross-validation across four Australian hospitals; when they subsequently tested the fixed model prospectively, it preserved discrimination around 0.80 and, on decision curve analysis, offered equal or greater net benefit than the routine universal OGTT pathway. Finally, an individual-participant-data meta-analysis by Ranasinha et al.[123] applied four widely cited early pregnancy scores to 16 randomised trials spanning Asia, Europe and Australia. Discrimination varied widely (AUC 0.60–0.78) and simple intercept/slope adjustment reduced misclassification by about 15%, underscoring how performance drifts as case-mix shifts.

Taken together, these studies confirm three recurrent themes. First, discrimination almost always declines when a model is transported, so researchers should expect this and plan accordingly. Second, calibration is rarely preserved; intercept or slope updating was needed in every study except Cooray's, validating Steyerberg's maxim to "recalibrate before you rebuild." Third, heterogeneity across centres and ethnic groups, seen most in the Ranasinha meta-analysis[123], means subgroup performance and fairness can no longer be optional extras but must be reported alongside population-level estimates.

In summary, without external validation, a model's performance and utility remain unproven in new settings. Only through testing a model on independent populations can we confirm that its predictive accuracy and calibration hold beyond the original training context. For this reason, Chapters 6 and 7 of this thesis are dedicated to rigorous external and prospective validations, respectively, to ensure the developed model truly generalises to real-world clinical environments.

1.5.10 Summary of Current Evidence

ML research in GDM has progressed rapidly, from a few small, non-validated LR models to high-dimensional approaches using gradient-boosting machines and, occasionally, neural networks. The number of models published is still growing. The main lessons are as follows:

- **Prognostic rather than diagnostic focus.** Most studies aim to identify women at risk before or early in pregnancy; very few attempt to replace the OGTT as a definitive diagnostic tool.
- **Early-pregnancy prediction (< 24 weeks).** Average discrimination now lies in the mid-to-high 0.8 range (AUC), rising later in gestation and when ensemble methods replace LR.
- **Population diversity.** More than half of recent models are trained and evaluated on Asian cohorts and most of the remainder in predominantly White populations; data from Latin-American, African and Middle Eastern groups are scarce, limiting generalisability.
- **Non-invasive versus invasive predictors.** Models based solely on history and anthropometry outperform simple BMI–waist rules but usually improve further when early biochemical measures, especially markers of glucose metabolism, are added.
- **Pre-conception risk stratification.** Algorithms using routine primary-care data achieve AUCs above 0.70; adding HbA_{1c} or similar tests can push performance well into the 0.8–0.9 range.
- **Treatment stratification after diagnosis.** Post-diagnosis models that combine clinical data with early CGM readings typically reach AUCs around 0.80 and could help clinicians decide when to escalate from diet to pharmacotherapy.
- **Model validation.** External validation remains the exception rather than the rule. Where it is performed, discrimination usually falls, recalibration is often required and the largest performance losses occur in demographically heterogeneous populations.

Together, these findings highlight both the promise of ML for earlier and more precise GDM care and the need for broader, rigorously validated models before routine clinical adoption.

1.6 LIMITATIONS & CONSIDERATIONS APPLYING ML IN HEALTHCARE

Integrating AI systems into routine care raises challenges that go well beyond algorithmic accuracy. A key component of this is the explainability of AI in medical contexts,

the extent to which clinicians can understand, trust and act upon model outputs[124]. Current guidance places an emphasis on ensuring these models are not only high performing but also trustworthy, transparent, and interpretable[125]. Interpretability (how readily a model's inner workings can be followed)[126], transparency (the disclosure of relevant information about the model's function)[127], and explainability (how clearly its predictions can be justified)[128], while distinct, underpin the broader concept of explainable AI (XAI) now required by many regulatory frameworks. However, historically, a tension has existed between optimising model performance and enhancing explainability[129].

Deep-learning architectures have delivered state-of-the-art results in imaging and speech tasks, yet their billions of parameters render their decision processes opaque, earning them the moniker of a "black box"[130]. Simpler models such as LR offer immediate clinical face validity (e.g. odds ratios) and make erroneous learning easier to detect, albeit sometimes at a cost in raw performance[131]. This dichotomy has resulted a common perception that highly intricate deep learning models, despite their reduced explainability, always deliver superior performance[130]. While this might hold true for complex data forms like audio and images[132], it's not necessarily the case for tabular data[131]. As evident from the literature on GDM, the input features often have a more substantial influence on model performance than the choice between a neural network or LR for tabular data structures.

The significance of explainability is perhaps best underscored by a seminal study by Caruana et al.[131]. They worked on a multicentre project aimed at predicting the risk of pneumonia by forecasting the probability of death, enabling the appropriate triaging of patients for treatment. The authors' neural network model, a layered learning model, significantly outperformed a LR model (AUC 0.86 vs 0.77). However, the LR model was deployed in preference to the neural network model as clinicians could comprehend the decision-making rationale of the former. Furthermore, a rule-based method had interpreted that patients with asthma had a lower risk of death than those without asthma. This paradox, given that asthma is actually a risk factor, was ultimately attributed to the fact that asthmatic patients typically receive more prompt treatment. Therefore, what the model had misinterpreted as 'asthma equating to lower risk' was actually 'earlier treatment equating to lower risk'. This highlights how hidden confounds can mislead uninterpretable models.

Holzinger et al.[130] delineated two categories of XAI, using Latin terminology from law. The first is post-hoc explainability, which pertains to "explaining what the model predicts in terms of what is readily interpretable". The second is ante-hoc explainability, which entails "incorporating explainability directly into the structure of an AI model". Ante-hoc

explainability is also occasionally referred to as "glass box" methods, as a counterpoint to the "black box" methods that are commonly seen in deep learning. The idea here is that the workings and decision-making process of the model can be seen and understood[126,133]. As mentioned earlier, LR is a frequently cited example of a glass box method, but it can often yield reduced performance[131]. However, there are alternative glass box methods generalized additive models that has been optimized to match the performance of black box methods, while preserving explainability in how the model makes decisions[131,133].

Post-hoc explainability involves methods applied to models that are essentially black boxes, but attempt to weigh the importance and impact of different features within the model to make its decisions more understandable. One such method involves the use of Shapley values, an approach first described by Lloyd Shapley in 1953[134] to estimate the contributions of players to a game. Today, there are software packages (like SHAP) that can be used to apply Shapley values to models. This process helps explain the relative contribution of features to the model's decisions, both mathematically and visually. Visual explanations could be especially useful in helping clinicians understand the model's results.

The debate over which method is preferable continues. However, a recent study examining the role of XAI in GDM investigated how different clinicians preferred the model's results to be explained, either by case example, feature importance, both, or neither[135]. Nearly half of all clinicians, which included obstetricians, midwives, and dietitians, wanted to see both types of explanations. Meanwhile, the majority of the remaining clinicians preferred to see just the feature importance. Other factors, such as years of experience, profession, and inclination to use an XAI support, influenced the preferred explanation method of the clinicians. Therefore, these factors should be taken into account when choosing a model for implementation in healthcare settings. In summary, high model discrimination is necessary but insufficient for clinical adoption of AI tools. Robust external validation, transparent reporting and context-appropriate explainability remain essential safeguards when models inform decisions with direct consequences for patient care.

1.7 ETHICS & ETHICAL ML IN HEALTHCARE

The demand for XAI aligns closely with the recent push for ethical AI[136]. Ethical AI and XAI intersect in the sense that a model's prediction without any supporting rationale may conceal potential biases in the data or model[137]. Bias is common in large datasets and can reinforce existing disparities[138], leading to subpar results when the models are applied in

different contexts. For instance, it was discovered that Amazon's hiring algorithm had a gender bias favouring male candidates[139]. Similar pitfalls arise in healthcare. For example, most GDM prediction models have been developed in East-Asian cohorts, populations with higher baseline GDM prevalence, while African datasets are largely absent[140,141]. Such demographic imbalance limits transportability and risks inequitable performance when models are applied to under-represented groups.

A recent systematic review in this area underscored how employing ethical and XAI methods boosts clinicians' confidence in the models' decision-making abilities and can even stimulate hypotheses about the decisions made by the models. This practice enhances the trustworthiness and acceptability of such models[53]. Several recent frameworks make bias detection and mitigation mandatory. For example, the FUTURE-AI[142] consensus guideline sets FAIR-E principles that include explicit subgroup performance reporting and bias audits before deployment. By aligning GDM-prediction research with these bias-focused guidelines, reporting subgroup metrics, conducting external validation in diverse populations and providing transparent model documentation, researcher can deliver algorithms that are both effective and transportable.

1.8 PRACTICAL BARRIERS TO CLINICAL DEPLOYMENT OF ML

AI systems in healthcare must meet a higher bar than those used in retail or entertainment, because erroneous recommendations can directly harm patients. Thus, it's critical to consider the barriers to AI deployment from the perspectives of various stakeholders, including healthcare organizations, providers, and patients[143]. Financial considerations and feasibility govern healthcare organizations' decisions about AI deployment, while patients mainly express concerns about the use of their data[144], a topic regulated in the EU by the GDPR. This concern also circles back to the points made about ethical AI. The needs of clinicians are perhaps more intricate and may depend on their specific medical practice areas.

Clinicians need more than just accurate and cost-effective ML models. They also require clinically relevant models[143] and models that can be understood from a patient's perspective[145]. Tonekaboni et al.[145] elaborate on this with the following key points:

- **Explainability**: Models with less than perfect accuracy were deemed acceptable as long as it was clear *why* the model underperforms.
- **Transparent model design**: Clinicians express the need for models that mirror the analytical process established in evidence-based medical decision-making.

- **Uncertainty**: This encompasses both model uncertainty and data uncertainty. Predictions should include confidence estimates that account for data and model variability.
- **Familiarity**: Risks should be reported and presented in units and probabilities routinely used in clinical care.
- **Feature importance**: Understanding the significance of each feature to the model was deemed "crucial."

Addressing these requirements, alongside organisational cost–benefit analyses and patient-centred privacy safeguards, is essential if AI tools for GDM, or any condition, are to achieve safe, accepted and effective implementation in routine care.

1.9 CHAPTER SUMMARY

This opening chapter has reviewed the evolution of ML methods for GDM prediction, highlighting steady gains in discrimination yet persistent shortcomings in data quality, external validity and clinical integration. Current models are constrained by heterogeneous diagnostic cut-offs, inconsistent outcome coding within EHRs, demographic imbalance in training cohorts and a paucity of prospective evaluation. These gaps frame the work that follows: the remainder of this thesis will first synthesise existing evidence (Chapter 2), then examine how label noise affects model performance (Chapter 4), before developing (Chapter 5), transporting (Chapter 6) and prospectively testing early-pregnancy (Chapter 7) prediction tools within real-world clinical pathways. By moving sequentially from evidence synthesis and data-quality appraisal to model construction, external validation and early clinical use, the thesis aims to deliver rigorous, explainable and ethically sound decision support for the timely identification and management of GDM.

Objective and Research Questions

Central Objective:

Can utilising machine learning techniques on electronic health records provide useful early predictions for gestational diabetes mellitus.

This objective will be investigated by addressing the following **research questions** which have been developed as a result of ongoing discussion with my clinical collaborators in the Coombe Hospital and UCD Centre for Human Reproduction:

RQ1. **Evidence Base.** What is the overall predictive performance, methodological quality, and heterogeneity of existing ML models for the early prediction of GDM using EHR data? (Chapter 2)

RQ2. **Label Noise**. To what extent are the data contained in the EHRs accurate, and if there is label noise, to what extent could this impact ML modelling of GDM? (Chapter 4)

RQ3. **Early Prediction**. How accurately can ML models, when applied to EHRs, predict the diagnosis of GDM? (Chapter 5)

RQ4. **Multiparous Analysis**. How accurately can ML models leverage data from a woman's previous pregnancies to predict outcomes in subsequent pregnancies, and which features from past pregnancies hold the most predictive power for future maternal outcomes? (Chapter 5)

RQ5. **External Validation**. Can an early-pregnancy GDM prediction model be externally validated on a geographically distinct population without direct sharing of patient-level data (using a model-exchange approach)? (Chapter 6)

RQ6. **Implementation**. In a clinical setting, does the model retain its predictive performance when deployed, and can it detect GDM earlier than current practice? (Chapter 7)

Chapter 2

Prediction Models for Early Gestational Diabetes Prediction Using Electronic Health Records: Systematic Review and Multi-Level Meta-Analysis

Germaine, M., Darragh, I. A.J., Healy, G., Manninen, M., Egan, B. (2026). Machine Learning Models for Early Gestational Diabetes Prediction Using Electronic Health Records: Systematic Review and Multi-Level Meta-Analysis

Chapter Overview

This chapter addresses RQ1 concerning the evidence base for using prediction models to predict GDM: *What is the overall accuracy, methodological quality, and heterogeneity of existing prediction models models for the early prediction of GDM using EHR data?*. To answer this, the chapter presents a comprehensive systematic review and multi-level meta-analysis of 38 studies, encompassing 122 individual prediction models and over two million pregnancies.

The analysis establishes that existing prediction models using routine EHR data achieve, on average, moderate discriminative performance, with a pooled AUC of 0.75 (95% CI 0.71-0.78). However, this performance is qualified by extreme heterogeneity across studies (I^2 ~99.6%), with a wide 95% prediction interval of 0.45 to 0.92. This variability indicates that the performance of any given model can differ markedly depending on the clinical setting, patient population, and specific features used. This finding points towards a significant credibility gap in the published literature, suggesting that many performance claims are likely optimistic and not readily generalisable. This gap provides a strong justification for the rigorous, step-by-step validation methodology adopted in the subsequent chapters of this thesis.

Importantly, the meta-analysis found no statistically significant performance advantage for complex algorithms like boosting ensembles (AUC 0.79) over simpler, more transparent linear models such as logistic regression (AUC 0.73). This suggests that for this specific clinical task using tabular EHR data, the quality and relevance of the input features are likely more important determinants of performance than the choice of algorithm. This insight has significant implications for clinical translation, where simpler, more interpretable models are often preferred. Furthermore, the review identified widespread methodological weaknesses, with 63% of studies judged to be at a high risk of bias, primarily due to inadequate validation strategies, unjustified sample sizes, and neglected calibration assessment. By systematically mapping the current state of the evidence, this chapter sets the stage for the thesis's subsequent work, which aims to directly address these identified shortcomings.

Systematic review registration: PROSPERO CRD420250651833

2.1 INTRODUCTION

Early identification of women at high-risk of GDM is a priority because timely interventions, such as lifestyle modification, has been shown to improve pregnancy outcomes[15–17]. Traditionally, GDM is diagnosed around 24-28 weeks of gestation via oral glucose tolerance testing[1], but by this stage opportunities for lifestyle intervention are limited[43,44]. An earlier prediction, using routine clinical data available at the first prenatal visit, could allow clinicians to target preventive measures weeks or months before hyperglycaemia develops, which has been shown to reduce the incidence of adverse neonatal outcomes[48].

Recent investigations have explored machine learning (ML) techniques to predict GDM risk earlier in pregnancy, leveraging the large volumes of clinical data stored in EHRs[94]. EHR data typically include demographic information, medical and obstetric history, and standard laboratory results. A variety of prediction models , from logistic regression (LR) to complex ensemble methods and neural networks, have been developed using data from pregnant women across diverse populations[99]. While individual studies often report promising ability to distinguish between patients who will develop GDM and those who will not, their results vary and models are rarely validated beyond the development setting[77,90]. Existing systematic reviews of GDM prediction models have included heterogeneous data sources, lacked meta-analysis or become outdated[83]. Given the rapid evolution of both ML methods and obstetric care practice, an up-to-date synthesis of the evidence is needed. This systematic review and meta-analysis aims to evaluate the overall accuracy of prediction models for early GDM prediction using routine EHR inputs, and to examine factors influencing model performance and validity.

The following questions are addressed: 1) What is the typical predictive performance (discrimination) of prediction models using clinical EHR data to predict GDM in the first half of pregnancy? 2) How heterogeneous are the results across studies and models, and what factors (e.g. model type, study population, risk of bias) explain this variability? 3) Are more advanced ML algorithms (e.g. ensemble or boosting models) superior to traditional approaches like LR for this task? and 4) What is the quality of these studies in terms of risk of bias, and to what extent have they adhered to best practices (such as external validation, transparency of model reporting)? 5) Does the addition of blood biomarkers to clinical risk factors improve model performance? By answering these questions, the aim is to inform researchers and clinicians about the current state of prediction models for early GDM prediction and to provide guidance on improving future model development and validation in this domain.

2.2 METHODS

2.2.1 Protocol and Registration

This review was conducted according to a pre-specified protocol registered in PROSPERO (ID CRD420250651833) and is reported in line with the Transparent Reporting of multivariable prediction model for Individual Prognosis Or Diagnosis - Systematic Review and Meta-Analysis (TRIPOD-SRMA) guidelines[146], with consideration of Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020[147] as we await the AI-specific reporting recommendations[148].

2.2.2 Study Eligibility Criteria

Inclusion Criteria:

- Studies that develop, validate, or apply prediction models to predict GDM onset based on EHR data.
- Studies that utilise structured EHR variables.
- Studies that report AUC/C-Statistic as a performance metric.
- Studies that make predictions based on data available prior to 16 weeks' gestation.

Exclusion Criteria:

- Studies that focus on biomarkers, genetic markers, or imaging techniques not routinely collected in EHR data. If biomarkers were collected as part of routine EHR, then they could be included.
- Studies that do not use prediction models for GDM prediction.
- Case reports, editorials, letters to the editor, and review articles.
- Studies that do not report discrimination performance metrics for the models evaluated.
- Studies that lack a clear definition of GDM or use inconsistent diagnostic criteria.
- Studies that use data beyond 16 weeks of gestation
- Studies that did not perform any model validation

2.2.3 Data Sources and Search Strategy

Seven databases were searched: MEDLINE, EMBASE, PubMed, Web of Science, Cochrane Library, IEEE Xplore, and Scopus. The search strategy combined terms for gestational diabetes mellitus with machine learning, AI, prediction or model, along with terms for electronic health/medical records. Searches were limited to human studies and English language. I searched records from January 2000 up to the latest available date in March, 2025

(to capture the surge in ML research over the past two decades). The full search strategies for each database are provided in Table 2.1.

Table 2.1. Search strategy used for the seven databases.

Database	Search Strategy
PubMed	((machine learning[Title]) OR (artificial intelligence[Title]) OR (AI[Title]) OR (prediction[Title]) OR (electronic health record[Title]) OR (electronic medical record[Title])) AND (gestational diabetes[Title])
MEDLINE (Ovid)	((machine learning.ti.) OR (artificial intelligence.ti.) OR (AI.ti.) OR (prediction.ti.) OR (electronic health record.ti.) OR (electronic medical record.ti.)) AND (gestational diabetes.ti.)
EMBASE (Ovid)	((machine learning:ti) OR (artificial intelligence:ti) OR (AI:ti) OR (prediction:ti) OR (electronic health record:ti) OR (electronic medical record:ti)) AND (gestational diabetes:ti)
Web of Science	TI=((machine learning) OR (artificial intelligence) OR (AI) OR (prediction) OR (electronic health record) OR (electronic medical record)) AND TI=(gestational diabetes)
Cochrane Library	TI:("machine learning" OR "artificial intelligence" OR "AI" OR "prediction" OR "electronic health record" OR "electronic medical record") AND TI:"gestational diabetes"
IEEE Xplore	(("Document Title":"machine learning") OR ("Document Title":"artificial intelligence") OR ("Document Title":"AI") OR ("Document Title":"prediction") OR ("Document Title":"electronic health record") OR ("Document Title":"electronic medical record")) AND ("Document Title":"gestational diabetes")
Scopus	TITLE(("machine learning" OR "artificial intelligence" OR "AI" OR "prediction" OR "electronic health record" OR "electronic medical record")) AND TITLE("gestational diabetes")

2.2.4 Study Selection

After removing duplicates, review articles, editorials, conference abstracts without full data, and non-English reports were excluded. Two reviewers (MG, IAJD) independently screened all titles and abstracts for potential eligibility. Full-text articles for all candidate studies were obtained and further assessed against the inclusion criteria. Covidence software was used to remove duplicates and manage the screening of eligible studies, with reviewers working independently and blinded to each other's decisions. Discrepancies were resolved through a third reviewer (BE).

2.2.5 Data Extraction

Data extraction was performed based on the CHARMS checklist (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies) and

tailored it to the specific needs of this review[149]. For each study, two reviewers (MG, IAJD) independently extracted detailed information, including: study characteristics, participant demographics, GDM prevalence and diagnostic criteria used, details of predictor variables, model development details, validation strategy, and all reported performance results. If a study reported multiple models, results were extracted for each relevant model. Information on model availability, data sharing, and code availability were also noted. The two sets of extracted data were compared for consistency, and any discrepancies were resolved by consensus or with arbitration by a third reviewer (BE) as needed. The finalised dataset is available as https://osf.io/rmbdt/

2.2.6 Risk of Bias and Quality Assessment

Two reviewers (MG, IAJD) independently appraised each study for risk of bias using the Prediction model Risk Of Bias ASsessment Tool - AI (PROBAST+AI) tool[102] . PROBAST+AI evaluates four domains; Participants, Predictors, Outcome, and Analysis, rating each as *Low*, *High*, or *Unclear* risk of bias based on signalling questions. Each study received an overall risk-of-bias judgement; per PROBAST+AI guidance, a study was marked *High* overall if any domain was rated high risk of bias, and *Low* overall only if all domains were low risk. Applicability concerns (i.e. relevance of participants, predictors, and outcome to our review question) were also assessed for each domain. Quality concern was also rated in a similar manner. Quality concern is assessed based on the model development. Reviewers compared their assessments and reached consensus on final judgements for each study. Remaining discrepancies were resolved through a third reviewer (BE).

2.2.7 Data Synthesis and Meta-Analysis

The primary measure of interest was model discrimination, measured by the C-statistic (AUC). For meta-analysis, reported AUC values were logit-transformed to stabilise variance and improve the normality of between-study effects[150–152]. For two studies, CI upper bounds were adjusted by subtracting 0.01to maintain values below 1.0 for transformation[153]. Where an included study did not report a 95% CI for the AUC, I calculated it using the Newcombe–Wilson method for a single proportion[154], as has been recommended elsewhere[150].

A multilevel random-effects meta-analysis employing restricted maximum likelihood was used, with five hierarchically nested levels to account for the non-independence of multiple performance estimates from the same study. Specifically, AUC estimates were considered to

be clustered by predictor set, within model type, within model family, within sample (or sub-cohort), and within each study. Approximate correlations that were used to construct the necessary variance-covariance matrix between multiple AUCs reported in the same study were based my own work[155,156]. In the primary model, outcome estimate correlations were assumed to be $\rho=0.95$ for models belonging to the same broad family (e.g. linear models vs ensemble models), $\rho=0.89$ for models of different types within the same family (e.g. XGBoost vs CatBoost), and $\rho=0.80$ for models sharing an identical predictor set. Models with three, four and five levels were compared using AIC, BIC and likelihood-ratio tests, and the five-level model provided the best fit. To ensure robust inference with this complex dependence structure, a cluster-robust variance estimation with small-sample correction (the clubSandwich approach) was applied to adjust standard errors and confidence intervals. This approach provides valid statistical inference even when the true within-study correlation structure may deviate from the assumed values[157].

The primary summary measure of model performance was the pooled AUC (with 95% CI), back-transformed to the probability scale from the meta-analytic logit scale. A 95% prediction interval for the AUC was also calculated, to estimate the range in which the true performance of a model would lie in a new *a priori* similar study. Heterogeneity in model performance was quantified using Cochran's Q statistic and the Higgins I^2 statistic[158]. Q value is reported with corresponding degrees of freedom and p-value, and I^2 as the percentage of total variability attributable to between-study (and between-model) heterogeneity rather than sampling error. In addition, variance components (τ^2) were estimated for each of the five levels in the multilevel model to examine the proportion of total variance contributed by between-study differences, between-sample differences, and so forth down to the predictor-set level. Model outliers and influential studies were evaluated by examining studentized residuals and Cook's distance for each effect estimate. Outlying effects were defined as those with studentized residuals exceeding approximately ±3, and highly influential effects were identified based on Cook's distance and leverage statistics[159].

Prespecified moderator analyses (meta-regressions) were carried out to explore potential sources of heterogeneity in AUC estimates. The primary moderator was model family (Linear v Bagging v Boosting). Interactions between model family and five categorical study/model characteristics (validation method, blood biomarkers, cohort type, study quality and risk of bias) were tested, as well as three continuous moderators (sample size, validation size, and number of predictors). Continuous moderators were centred and analysed similarly with interaction terms to see if, say, larger studies had higher AUCs on average. Differences

between moderator categories were tested by conducting Wald-type F-tests on the multilevel meta-regression model.

Sensitivity analyses were performed to assess the robustness of the meta-analytic results. First, the assumed within-study correlations between multiple AUCs were varied from ρ=0.7 to ρ=0.9. Next, a sensitivity analysis was conducted excluding studies and/or models identified as outliers or influential in the earlier diagnostics, to determine whether the overall results were unduly driven by those extreme findings. Finally, publication bias (small-study effects) was evaluated qualitatively with a funnel plot of model AUC vs. standard error, using one aggregated AUC per study, to avoid multiple points per study, and quantitatively with a modified Eggers regression test[160], adapted for use with the logit AUC and the robust variance meta-regression framework[152,157].

All meta-analytic estimates are presented with 95% CIs. The meta-analysis methodology followed current recommended practices for prognostic model reviews[146,149,151,161] and addresses the complex data structure by using state-of-the-art multilevel modelling and bias assessments. All meta-analyses were conducted in R (v4.2) using both the metamisc and the metafor package with the robust variance extension[157,161]. I considered $p<0.05$ as statistically significant for subgroup differences. Where data were insufficient to meta-analyse, I summarised results narratively.

2.3 RESULTS

2.3.1 Study Selection

Figure 2.1 shows the PRISMA flow diagram of study selection. An initial literature search retrieved 2,087 records from seven electronic databases (Web of Science 488; Embase 384; Scopus 372; PubMed 298; IEEE Xplore 278; MEDLINE 232; Cochrane 35) plus two additional records identified through citation searching. After removing 1,216 duplicates (27 manually and 1,189 via Covidence), 873 unique records remained for title and abstract screening. Of these, 763 were excluded, leaving 110 articles for full-text assessment. During eligibility review, 72 of the full-text articles were excluded for reasons such as the data not being EHRs (50%). Ultimately, 38 studies met all inclusion criteria and were included in the final review.

2.3.2 Study Characteristics

Thirty-eight studies (2010-2025, 82% since 2020), met the inclusion criteria, consisting of 2,095,843 pregnancies and ~143,724 GDM cases. Most were single-centre retrospective EHR analyses (26/38). Study size varied (median 4,327 women; range 97-1.1 million), with four large, multi-centre datasets (n>40,000), and ten n<1,000. Research was concentrated in Asia (26 studies, 23 from China), with smaller contributions from Europe (6), Australia (4), North America and Africa (1 each); accordingly, most cohorts were Asian, with European-origin populations representing 16% and other groups only single studies. All models relied on routine booking data, with common predictors listed in Table 2.2. An attempt was made to assess the most important predictors for models, however, the majority of studies (26/38) either did not provide this information or provided it as unstandardised coefficients. Two-thirds reported diagnostic criteria using IADPSG thresholds, with wide range across studies of prevalence of GDM (median, 16.7%; IQR, 10.6-24.4%). External validation (6 studies), calibration reporting (22 studies, rarely with slope/intercept), decision curve analysis (10) and availability of model artefacts (16), data (3), or code (1) were limited. See Tables 2.2 and 2.4 for details.

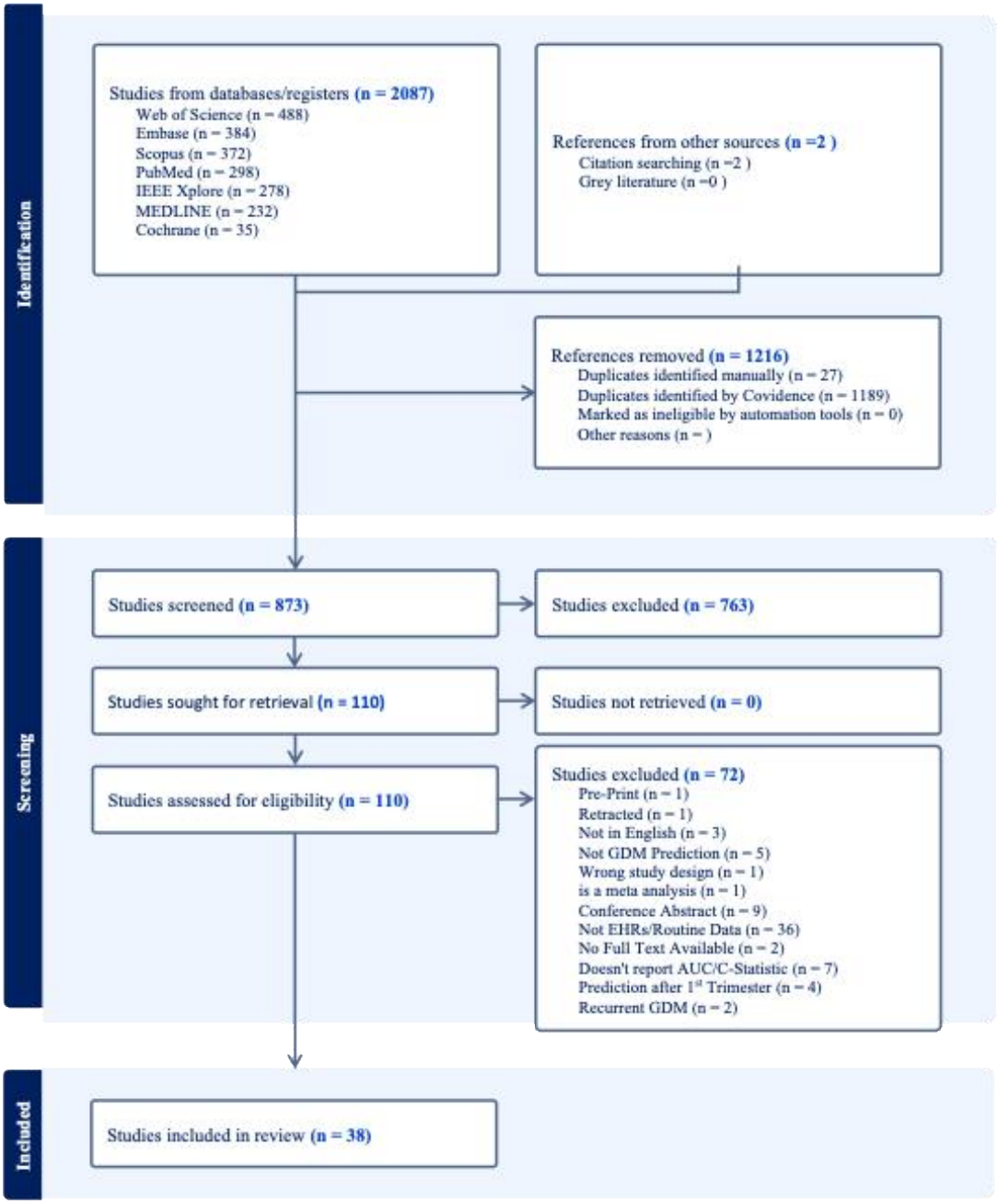


Figure 2.1. PRISMA flow diagram of study selection for meta-analysis.

Table 2.2. Key characteristics across the 38 studies included in this review.

Characteristic	n (%) or summary value
Participants	
Total pregnancies	2,095,843
Pooled GDM cases	143,724
Median sample size (IQR)	4,327 (1,067-18,749)
Median study-level GDM prevalence (range)	16.7% (2.4-71.8%)
Study design	

Table 2.2. Key characteristics across the 38 studies included in this review.

Characteristic	n (%) or summary value
Retrospective	26 (68%)
Prospective	12 (32%)
Geography	
Asia	26 (68%)
China	*23*
Middle East	*2*
Other Asia	*1*
Europe	6 (16%)
Australia	4 (11%)
North America	1 (3%)
Africa	1 (3%)
Gestational timing of prediction	
Median gestational age (range)	14 weeks (0-16 weeks)
Median gestational age	14 weeks
Studies ≤ 13 weeks	15 (39%)
Predictor prevalence (56 predictor sets)	
Maternal age	46 (82%)
BMI	44 (79%)
Family history of diabetes	30 (54%)
Previous GDM	25 (45%)
Parity / obstetric history	13 (23%)
Blood-pressure variable	11 (20%)
≥ 1 laboratory biomarker	23 (41%)
ML algorithms evaluated	
Logistic regression	34 (89%)
Random Forest	10 (26%)
XGBoost	9 (24%)
Support-vector machine	5 (13%)
Neural network / deep learning	5 (13%)
Single decision tree / CART	3 (8%)
Others (each ≤ 2 studies)	≤5%
Models per study	
Median (IQR)	1 (1–3)
Total internally validated models	122
Validation method	
Hold-out / independent split	20 (53%)
k-fold cross-validation	7 (18%)
Temporal split	5 (13%)
Bootstrapping	1 (3%)

Table 2.2. Key characteristics across the 38 studies included in this review.

Characteristic	n (%) or summary value
External validation	6 (16%)
Evaluation metrics reported	
AUC / C-statistic	38 (100%)
Calibration (plot or statistic)	22 (58%)
Decision Curve Analysis	10 (26%)
PR-curve	4 (11%)
Model / data availability	
Model artefact (equation, nomogram, web tool)	16 (42%)
Dataset publicly available	3 (8%)
Code publicly available	1 (3%)

2.3.3 Risk of Bias Assessment

Using the PROBAST+AI framework, 63% of studies were judged high overall risk of bias (24/38) and 26% (10/38) low risk. The analysis domain was the main source of concern (22/38, 61%), with issues including non-independent validation, inadequately justified complex models in small datasets, or neglected calibration. Outcome domain issues (5/38, 13%) arose GDM criteria was not defined. By contrast, participant selection (33/38, 87%) and predictor measurement (36/38, 95%) were low risk, reflecting the routine nature of the EHR data used. Development-stage overall concern for quality was rated high in 18 of the 38 studies (47%), low in 12 (32%), and unclear in 8 (21%). Inter-rater agreement on the risk of bias assessment was high, with Cohen's κ=0.803 for overall risk rating and 90% agreement. All disagreements were resolved via discussion. See Figures 2.2, 2.3 and Table 2.3 for details.

Table 2.3. Risk of bias (PROBAST+AI) summary for included studies (N=38)

Domain	Low-Risk (n,%)	High-Risk (n,%)	Unclear (n,%)	Common reasons for high risk or concerns
Participants	33 (87%)	0 (0%)	5 (13%)	Generally representative obstetric populations. Unclear in 3 due to poor reporting of enrolment.
Predictors	36 (95%)	0 (0%)	2 (5%)	Predictors were routine clinical data measured pre-outcome. A few unclear due to incomplete predictor description.
Outcome	31 (82%)	5 (13%)	2 (5%)	High risk mainly when GDM definition was non-standard or not defined.
Analysis	11 (29%)	22 (58%)	4 (11%)	High risk causes: lack of proper model validation, no calibration reporting, and small sample size without justification.

Table 2.3. Risk of bias (PROBAST+AI) summary for included studies (N=38)

Domain	Low-Risk (n,%)	High-Risk (n,%)	Unclear (n,%)	Common reasons for high risk or concerns
Overall Risk of Bias	10 (26%)	24 (63%)	4 (11%)	26 studies had ≥1 high-risk domain (most often Analysis). Only 10 had all domains low. 4 had no high domains but ≥1 unclear (classified overall unclear).
Overall Quality Concern	12 (32%)	18 (47%)	8 (21%)	High risk causes: typically small samples lacking justification that encouraged over-fitting.
Applicability: Population	35 (95%)	0 (0%)	2 (5%)	Most study populations match the review question.
Applicability: Predictors	39 (100%)	0 (0%)	0 (0%)	All used routinely collected predictors feasible in practice, so no concerns here.
Applicability: Outcome	32 (86%)	3 (8%)	2 (6%)	3 had high concern: e.g. GDM criteria (if model applied in setting with newer criteria, performance may differ).
Overall Applicability	33 (89%)	3 (8%)	1 (3%)	Most models considered applicable to broad obstetric settings.

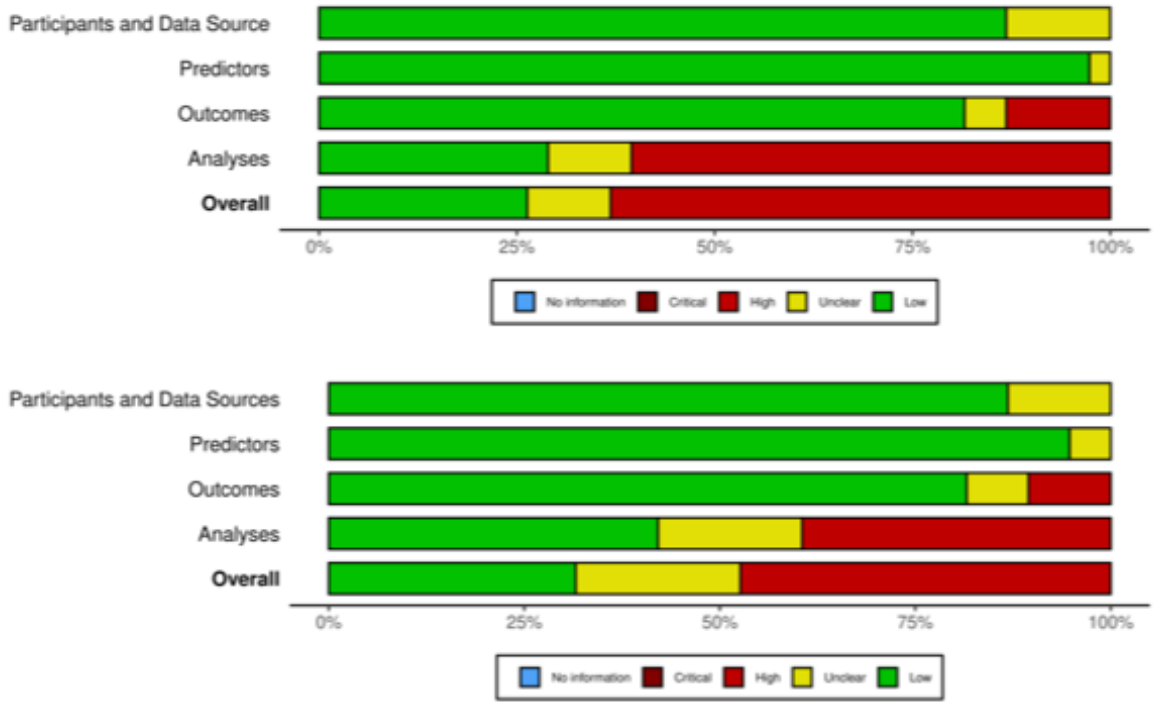


Figure 2.2. Top; Risk of bias (model evaluation) and bottom; Quality concern (model development) for expressed as the relative contributions from the four domains.

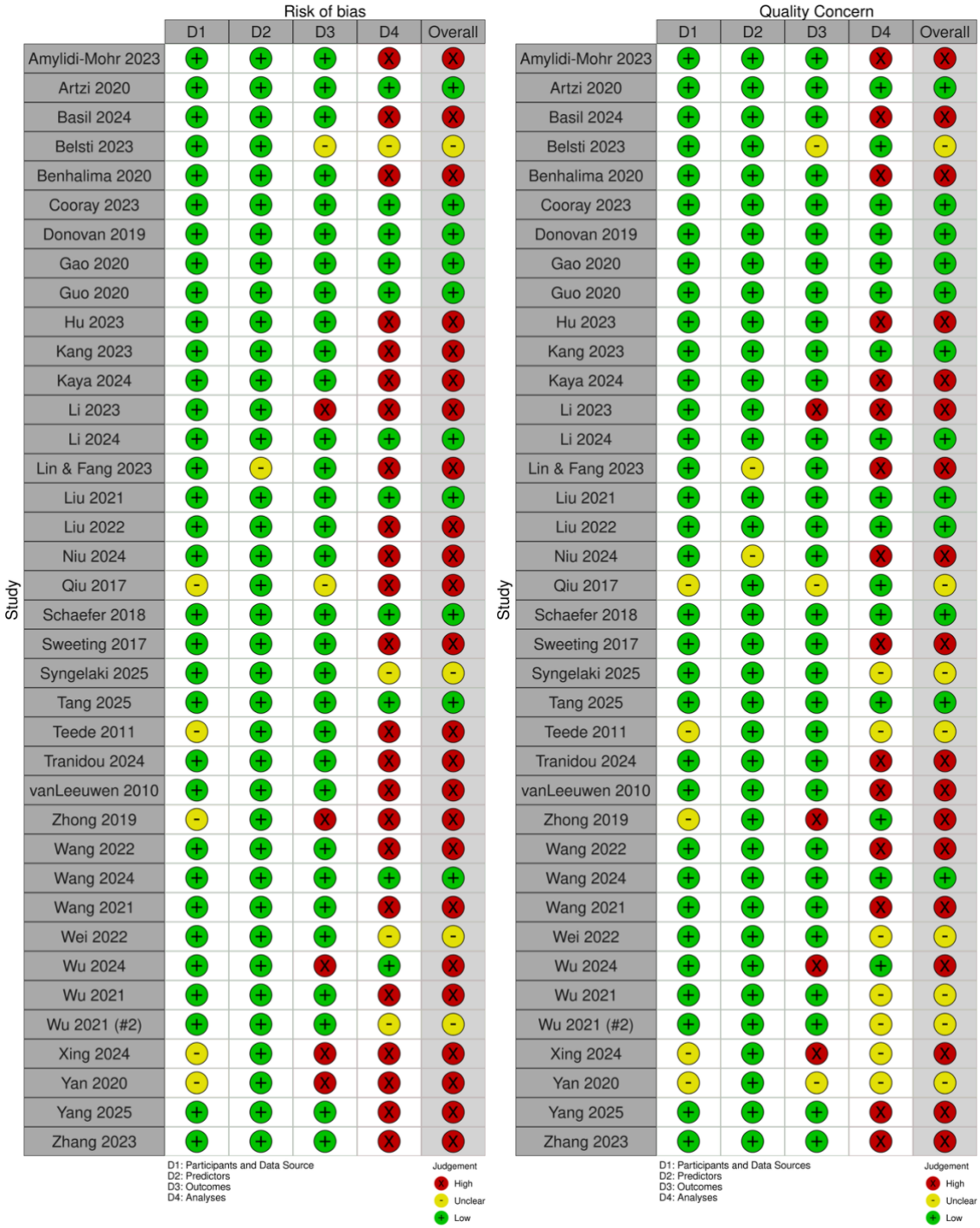


Figure 2.3. Left; Risk of bias (model evaluation) and right; Quality concern (model development) for the 38 studies, assessed using PROBAST+AI.

Table 2.4. Study-level breakdown of the characteristics, performance and clinical reporting of the 38 studies included in this review.

Author (year)	Country	Study design	Sample & Events	Gest at Prediction	Diagnostic criteria	Predictor sets	Models tested	Best AUC (95% CI)	DCA	Cal.	Ext Val
Amylidi-Mohr (2023)[162]	Switzerland	Pro	785 / 115	14 weeks	IADPSG	1	1	0.686 (0.615-0.756)	No	No	No
Artzi (2020)[94]	Israel	Retro	588,622 / 21,190	14 weeks	NIH	1[b]	1	0.799 (0.798-0.800)[a]	Yes	Plot	No
Basil (2024)[163]	Nigeria	Pro	253 / 52	12 weeks	IADPSG	1	1	0.816 (0.743-0.890)	No	HL	No
Belsti (2023)[164]	Australia	Retro	48,502 / 10,331	NS	IADPSG	4	12	0.930 (0.920-0.930)	Yes	Plot	No
Benhalima (2020)[165]	Belgium	Pro	1,843 / 230	14 weeks	IADPSG	2[b]	1	0.716 (0.686-0.754)	No	No	No
Cooray (2023)[166]	Australia	Retro	26,474 / 4,765	NS	IADPSG	1	2	0.732 (0.725-0.740)	No	Intercept, Slope, Plot	No
Donovan (2019)[111]	USA	Retro	1,160,933 / 73,139	0 weeks	ACOG	1	1	0.732 (0.728-0.735)	No	Plot	Yes
Gao (2020)[167]	China	Pro	19,331 / 1,488	15 weeks	IADPSG	1[b]	1	0.710 (0.680-0.741)	No	HL	No
Guo (2020)[168]	China	Pro	10,528 / 1,621	13 weeks	ADA	1[b]	1	0.700 (0.680-0.720)	Yes	Plot	No
Hu (2023)[169]	China	Pro	925 / 185	13 weeks	IADPSG	1[b]	2	0.745 (0.648-0.842)	Yes	Plot, HL	No
Kang (2023)[170]	South Korea	Retro	34,387 / 3,095	13 weeks	NIH	3[b]	3	0.723 (0.714-0.732)[a]	No	No	No
Kaya (2024)[171]	Turkey	Retro	97 / 19	13 weeks	IADPSG	1[b]	6	0.933 (0.864-0.990)[a]	No	No	No
Li (2023)[172]	China	Retro	673 / 182	13 weeks	NS	1[b]	1	0.753 (0.720-0.786)[a]	No	No	No
Li (2024)[173]	China	Retro	6,844 / 1,369	13 weeks	IADPSG	2[b]	4	0.996 (0.992-1.000)	No	Plot	Yes
Lin (2023)[174]	China	Retro	406 / 197	13 weeks	IADPSG	1[b]	2	0.918 (0.868-0.967)	No	No	No
Liu (2021)[175]	China	Pro	19,331 / 1,488	15 weeks	IADPSG	1[b]	2	0.742 (0.715-0.760)	No	Plot, HL	No
Liu (2022)[176]	China	Pro	6,848 / 966	13 weeks	IADPSG	3	3	0.618(0.612-0.623)	No	No	No
Niu (2024)[177]	China	Retro	6,000 / 1,740	12 weeks	IADPSG	1[b]	1	0.782 (0.759-0.806)	No	HL	No
Qiu (2017)[95]	China	Retro	4,378 / 613	13 weeks	NS	1[b]	6	0.847 (0.784-0.910)[a]	No	No	No

Schaefer (2018)[178]	China	Retro	8,381 / 1,131	14 weeks	IADPSG	1	1	0.649 (0.605-0.692)	No	Plot	No
Sweeting (2017)[179]	Australia	Pro	980 / 248	13 weeks	IADPSG	1	1	0.880 (0.850-0.920)	No	No	No
Syngelaki (2025)[180]	UK	Pro	41,587 / 4,242	14 weeks	NICE	1	1	0.757 (0.749-0.765)	No	Intercept, Slope, Plot	No
Tang (2025)[181]	China	Pro	1,904 / 352	13 weeks	IADPSG	1[b]	1	0.702 (0.645-0.758)	Yes	Plot, HL	No
Teede (2011)[182]	Australia	Retro	4,276 / 381	15 weeks	ADIPS	1	1	0.703 (0.673-0.733)[a]	No	No	No
Tranidou (2024)[183]	Greece	Pro	4,917 / 447	14 weeks	IADPSG	5[b]	1	0.678 (0.650-0.700)	No	No	No
van Leeuwen (2010)[84]	Netherlands	Pro	995 / 24	16 weeks	WHO <2013	1	1	0.770 (0.690-0.850)	No	HL	No
Zhong (2019)[184]	China	Retro	4,000 / 812	NS	NS	1[b]	6	0.91 (0.885-0.935)[a]	No	No	No
Wang (2022)[185]	China	Retro	1,285 / 391	14 weeks	IADPSG	1[b]	3	0.834 (0.785-0.882)	Yes	Plot, HL	Yes
Wang (2024)[186]	China	Retro	2,990 / 448	14 weeks	IADPSG	2	1	0.827 (0.791-0.862)	Yes	Plot	Yes
Wang (2021)[187]	China	Retro	1,640 / 328	14 weeks	IADPSG	1[b]	1	0.774 (0.733-0.814)	No	Plot, HL	Yes
Wei (2022)[188]	China	Retro	2,895 / 1,294	16 weeks	IADPSG	2	1	0.763 (0.730-0.800)	Yes	Plot, HL	No
Wu (2024)[189]	Netherlands	Retro	15,837 / 633	14 weeks	NS	1	3	0.808 (0.805-0.811)	No	Plot	Yes
Wu (2021)[190]	China	Retro	17,005 / 1,973	16 weeks	IADPSG	6[b]	2	0.711 (0.699-0.723)[a]	No	No	No
Wu (2021)[191]	China	Retro	32,190 / 4,925	14 weeks	IADPSG	2[b]	4	0.800 (0.790-0.810)	Yes	Plot, HL	No
Xing (2024)[192]	China	Retro	5,649 / 1,215	14 weeks	NS	1	1	0.617 (0.603-0.631)[a]	No	No	No
Yan (2020)[86]	China	Retro	1,600 / 805	NS	NS	2[b]	4	0.779 (0.773-0.784)[a]	No	No	No
Yang (2025)[193]	China	Retro	942 / 676	14 weeks	IADPSG	1	11	0.89 (0.85 to 0.93)[a]	No	No	No
Zhang (2023)[194]	China	Retro	924 / 235	12 weeks	IADPSG	1[b]	1	0.754 (0.718-0.790)	Yes	Plot	No

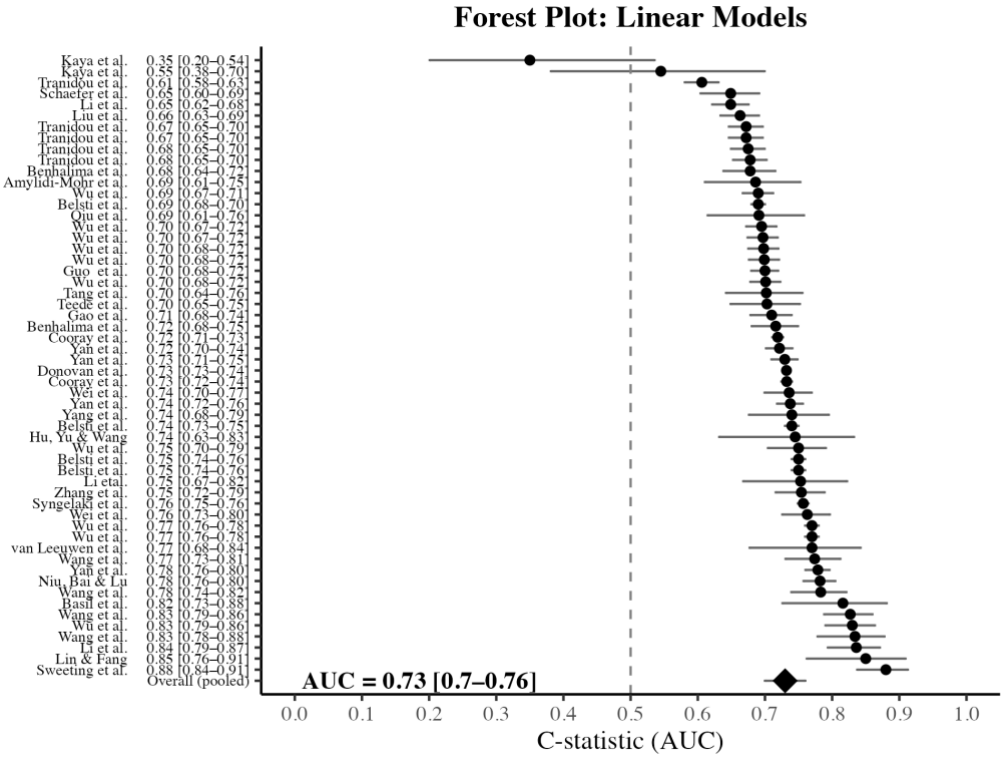
Best AUC, based on internal validation; Pro, prospective; Retro, retrospective; DCA, decision-curve analysis; Cal., Calibration (includes plots, slope, intercept, Hosmer-Lemeshow Test (HL))

[a] studies that did not report 95% CI but were computed.

[b] top performing model included blood biomarkers in predictor set.

2.3.4 Meta-Analysis of Model Performance

The AUC of 122 prediction models from the 38 studies was assessed. Figure 2.2 (forest plot) illustrates the AUC estimates for each model by study, broken down by model family and the pooled results. The pooled AUC was 0.75 (95% CI 0.71-0.78, I^2 ~99.6%), indicating moderate overall discrimination ability for early prediction of GDM with very high heterogeneity. The prediction interval was 0.45 to 0.92, meaning in some settings/models the AUC could be as low as ~0.5 or as high as ~0.9. Both extremes were seen in the dataset; the lowest reported model AUC was 0.35, and the highest was 0.996. Despite this, most models achieved AUCs in the 0.70-0.85 range, with relatively few outliers below 0.6 or above 0.9. Excluding obvious outlier models did not dramatically change the pooled estimate. The Q-statistic was 147,137 (df=121, p<0.001), indicating substantial variability among the studies. Variance component analysis suggested that the largest share of variance in AUCs was attributable to differences between models using different predictor sets (σ^2 ~0.237). The between-study variance was smaller (σ^2 ~0.097), and variance due to algorithm class was smaller still (σ^2 ~0.069). Residual variance at the model level (within a study and model family) was near zero in my model. The influential and outlier effect and study diagnostics indicated eight effects and three studies to be influential. However, only one outlier effect and study were found (See model diagnostics Figure 2.5).



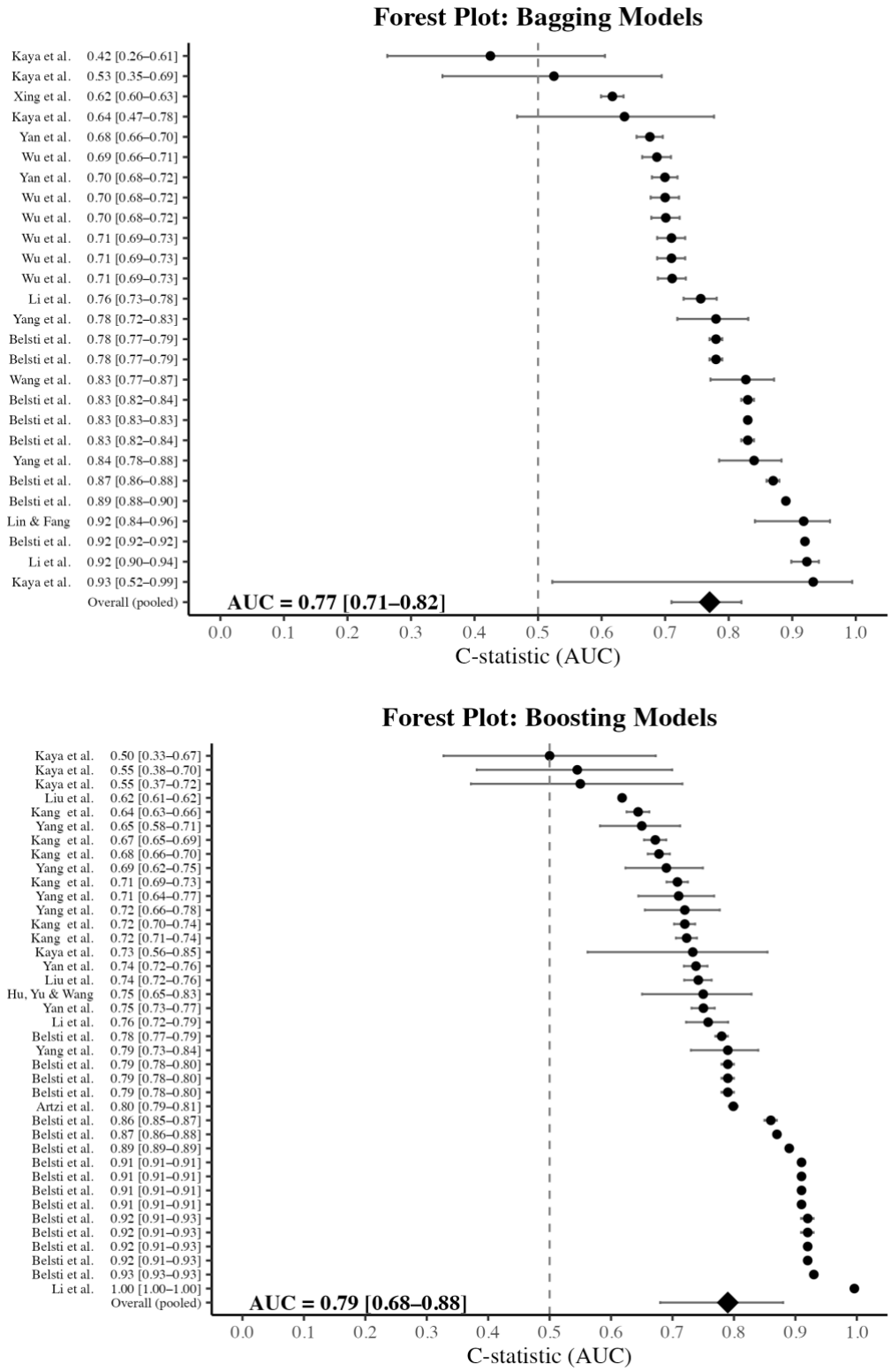


Figure 2.4. Random-effects meta-analysis of AUC of internally validated machine learning models of gestational diabetes mellitus, broken down by 56 Linear models (top), 27 Bagging models (middle) and 39 Boosting models (bottom).

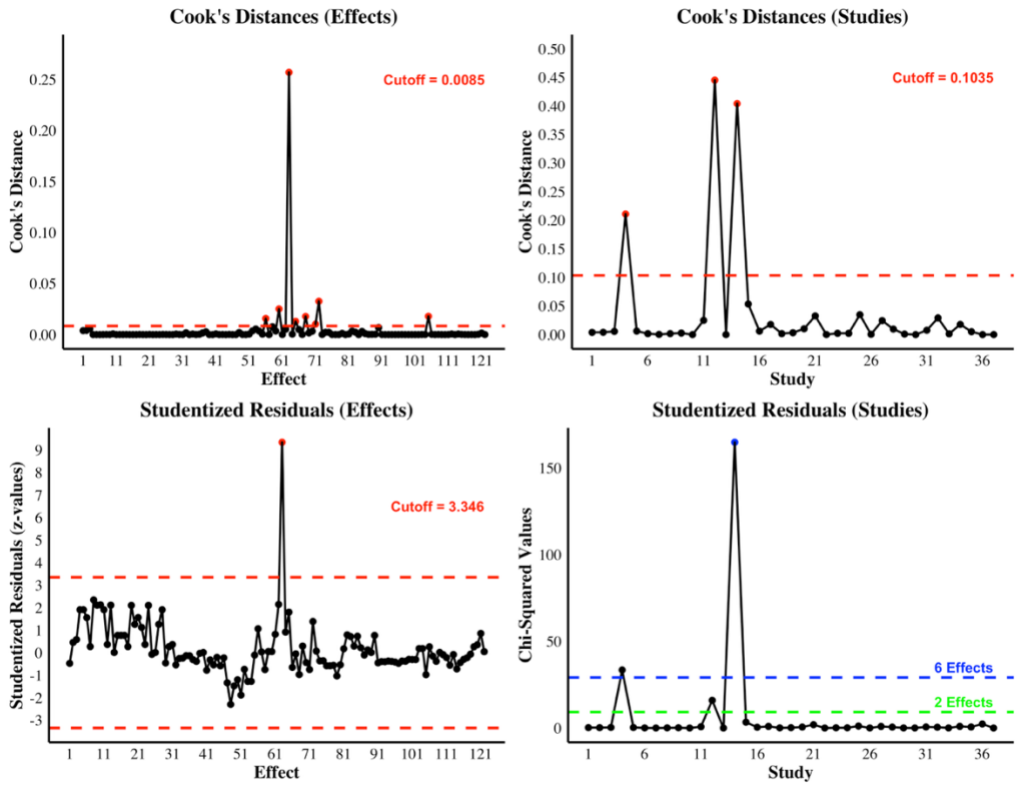


Figure 2.5. Model diagnostics. Only one outlier effect and study were found, associated with Li et al. (2024).

2.3.5 Moderator Analyses

The primary moderator of interest was model family. Pooled AUCs were 0.73 (95% CI 0.70-0.76) for linear models (k=56), 0.77 (95% CI 0.71-0.82) for bagging ensembles (k=27), and 0.79 (95% CI 0.68-0.88) for boosting ensembles (k=39). There were no statistically significant differences in model performance between model families (Ensemble boosting vs Linear models ($F(1, 9.4) = 1.83$, $p = 0.21$), Ensemble bagging vs Linear models ($F(1, 7.83) = 2.46$, $p = 0.16$), and Ensemble bagging vs Ensemble boosting ($F(1, 6.29) = 0.47$, $p = 0.52$). The secondary moderator analyses dealt with the interaction of model family and five categorical moderators (validation method, blood biomarkers, cohort type, study quality and risk of bias). The only statistically significant differences between categories within the model families were found in risk of bias moderators where the boosting and bagging model families had lower AUC in the high risk of bias models compared to the uncertain risk of bias models. Continuous moderators (total sample size, validation-sample size, and number of predictors; all centred, with sample-size variables log-transformed) showed no significant associations with AUC in any model family. The full results of the categorical and continuous moderator analyses are displayed in Table 2.6.

Table 2.5. Meta-analysis of model discrimination (AUC) by subgroup moderators

	Linear Models				Ensemble Bagging Models				Ensemble Boosting Models			
	k	**AUC**	**CI**	**sig.**	**k**	**AUC**	**CI**	**sig.**	**k**	**AUC**	**CI**	**sig.**
Model Family Overall	56	0.73	0.70, 0.76	p<0.001	27	0.77	0.71, 0.82	p<0.001	39	0.79	0.68, 0.88	p<0.001
Internal Validation												
Cross Validation	14	0.74	0.68, 0.78	p<0.001	14	0.79	0.65, 0.89	p<0.01	23	0.81	0.49, 0.95	p=0.052
Independent Test	32	0.72	0.69, 0.76	p<0.001	13	0.75	0.67, 0.81	p<0.01	16	0.79	0.58, 0.91	p<0.05
Biomarkers as Predictors												
Yes	32	0.74	0.69, 0.78	p<0.001	15	0.79	0.66, 0.89	p<0.001	14	0.83	0.57, 0.95	p<0.05
No	24	0.72	0.65, 0.77	p<0.001	12	0.73	0.47, 0.89	p=0.065	25	0.73	0.49, 0.88	p=0.057
Cohort												
Prospective	15	0.73	0.67, 0.78	p<0.001					3	0.71	0.51, 0.86	p<0.05
Retrospective	38	0.73	0.70, 0.77	p<0.001	27	0.78	0.71, 0.83	p<0.001	15	0.81	0.67, 0.90	p<0.01
Study Quality												
Low	12	0.7	0.63, 0.75	p<0.001	2	0.75	0.31, 0.95	p=0.13	11	0.82	0.47, 0.96	p=0.064
High	31	0.74	0.69, 0.79	p<0.001	19	0.78	0.68, 0.85	p<0.001	28	0.79	0.62, 0.89	p<0.01
Unclear	13	0.73	0.69, 0.77	p<0.001	6	0.72	0.65, 0.79	p<0.01				
Risk of Bias												
Low	12	0.72	0.67, 0.76	p<0.001	2	0.81	0.27, 0.98	p=0.011	4	0.9	0.20, 0.99	p=0.012
High	35	0.73	0.69, 0.77	p<0.001	17	0.74[a]	0.66, 0.80	p<0.001	19	0.7[a]	0.66, 0.74	p<0.001
Unclear	9	0.75	0.72, 0.78	p<0.001	8	0.85[b]	0.81, 0.89	p < 0.01	16	0.89[b]	0.85, 0.92	p<0.01
Continuous Variables												
Sample Size		0.73	0.70, 0.751	p<0.001		0.77	0.68, 0.84	p=0.001		0.80	0.69, 0.87	p<0.001
Validation Sample Size		0.73	0.70, 0.76	p<0.001		0.77	0.67, 0.84	p=0.001		0.81	0.69, 0.89	p=0.001
Number of Predictors		0.74	0.70, 0.76	p<0.001		0.77	0.69, 0.84	p<0.001		0.80	0.68, 0.88	p=0.001

	k	**AUC**	**CI**	**Prediction Interval**		**I²**
ALL MODELS (OVERALL)	122	0.75	0.71, 0.78	0.45	0.92	99.6%

[a] statistically significant compared to low risk of bias

[b] statistically significant compared to low risk of bias

sig. two-sided p-value from a Z-test of the null hypothesis that the subgroup's pooled AUC equals 0.50 (no better than chance).

2.3.6 Sensitivity analyses

Changing the within-study correlation between multiple AUCs (ρ = 0.70, 0.80, 0.90) and sequentially removing outliers or influential points produced minimal variation in the pooled estimate. Across all scenarios, the summary logit-AUC ranged from 1.036 to 1.091, translating to probability-scale AUCs of 0.738–0.749 with tightly overlapping 95% CI. See Table 2.5 for full details.

Table 2.6. Sensitivity Analysis.

r*	Sensitivity	Logit (AUC)	AUC	CI 95% LB	CI 95% UB
0.7	Main (all effects)	1.091	0.749	0.713	0.781
0.7	Outlier effects removed	1.047	0.740	0.711	0.767
0.7	Outlier study removed	1.036	0.738	0.709	0.765
0.7	Influential effects removed	1.073	0.745	0.719	0.770
0.7	Influential studies removed	1.037	0.738	0.717	0.758
0.7	All outliers and influentials removed	1.039	0.739	0.718	0.758
0.8	**Main (all effects)**	**1.091**	**0.749**	**0.713**	**0.781**
0.8	Outlier effects removed	1.046	0.740	0.711	0.767
0.8	Outlier study removed	1.036	0.738	0.709	0.765
0.8	Influential effects removed	1.073	0.745	0.719	0.770
0.8	Influential studies removed	1.037	0.738	0.717	0.758
0.8	All outliers and influentials removed	1.038	0.739	0.718	0.758
0.9	Main (all effects)	1.091	0.749	0.713	0.781
0.9	Outlier effects removed	1.046	0.740	0.711	0.767
0.9	Outlier study removed	1.036	0.738	0.709	0.765
0.9	Influential effects removed	1.072	0.745	0.719	0.770
0.9	Influential studies removed	1.037	0.738	0.717	0.758
0.9	All outliers and influentials removed	1.038	0.739	0.718	0.758

2.3.7 Publication bias

Neither the visual inspection of the contour enhanced funnel plot (see figure 2.6) nor the modified eggers' test (p=0.211) indicated a potential publication bias.

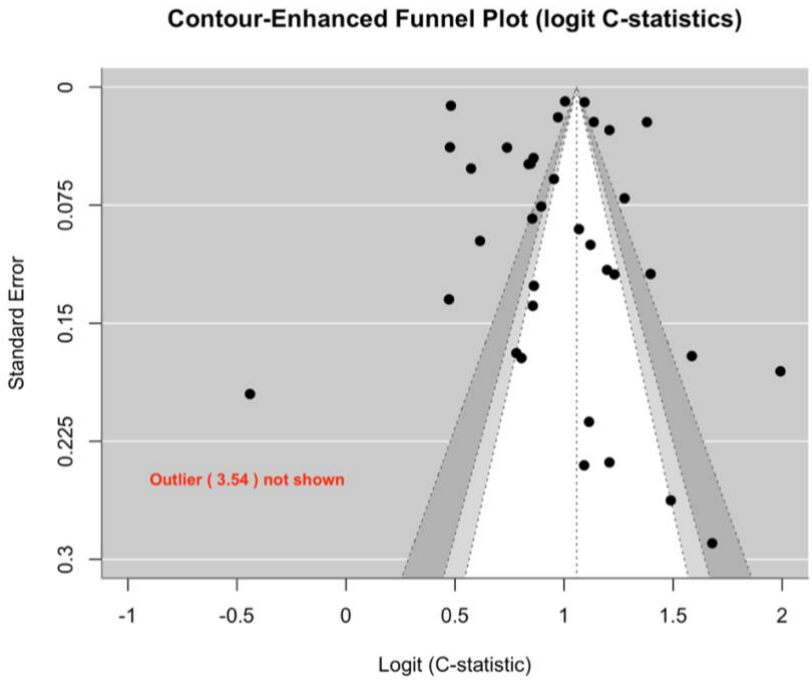


Figure 2.6. Contour enhanced funnel plot for the assessment of publication bias (outcomes estimates aggregated at the study level).

2.4 DISCUSSION

2.4.1 Main Findings

This systematic review and meta-analysis aimed to quantify the average performance of early-pregnancy GDM prediction models and assess the variability of results across studies. It synthesised evidence from 38 studies, encompassing over 2 million pregnancies and 122 prediction models, to evaluate the pooled discrimination of ML models using routinely collected EHR data for the early prediction of GDM. The main finding is that ML models, on average, achieve moderate discrimination, with a pooled AUC of 0.75 (95% CI 0.71-0.78). However, this result is characterized by extreme heterogeneity (I^2 99.6%), with a wide prediction interval (0.45-0.92), suggesting substantial variability in performance across different settings, populations, and models. The analysis indicated that a significant portion of this variance was attributable to differences in predictor sets used by the models, followed by study-level differences and model family.

In comparison with previous systematic reviews for first-trimester models using similar clinical predictors, the pooled AUC of 0.75 closely matches the median AUC of 0.71 (IQR

0.67–0.76) reported by van Eekhout et al.[195] and 0.77 (95% CI 0.69-0.84) reported by Huang et al.[196]. These are comparable to the performance of LR models (AUC 0.73) within this review. In contrast, it is lower than the 0.85 summary AUC cited by Zhang et al.[83], whose inclusion of non-routine biomarkers and data collected later in pregnancy likely inflated performance. Prior reviews such as Mennickent et al.[99] also highlighted AUCs in this range when reviewing models using clinical risk factors. AUCs in the 0.7-0.8 range are generally considered to have acceptable or moderate discriminative performance[76], however, interpretations are domain specific[77,91]. Clinically, an AUC of 0.75 means the model will assign a high-risk score to a woman who will later develop GDM 75% of the time when compared with a woman who will not. This may be useful for early risk stratification of women, however, it also means that many women will be misclassified. These models are not diagnostic but can guide early screening: false positives result in unnecessary OGTTs, while false negatives may defer or miss early testing in women who truly are at risk. Institutions need to balance this trade-off. Maternity hospitals may accept higher false-negative rates (relying on routine GTT at 28 weeks to catch these cases) to reduce excess testing, although this risks false reassurance and delayed lifestyle modifications among misclassified women.

Contrary to expectations, more complex ML algorithms such as bagging or boosting ensembles did not demonstrate a clear advantage in discriminative performance over simpler LR models for this specific task. Two earlier meta-analyses suggested non-LR techniques outperform LR for the prediction of GDM, yet one drew its conclusion from only three studies (2/3 studies supporting non-LR)[97], while the other relied on a descriptive comparison of AUCs without formal testing[83]. Indeed, looking at the descriptive AUC in the current review might also indicate boosting models (0.79) outperform LR (0.73), however, the results were not significant. These findings align the larger meta-analysis by Christodoulou et al.[197], which pooled diverse clinical prediction models and found no consistent advantage of ensemble algorithms over LR once sample size and study quality were taken into account. These findings are not insignificant, as more complex ML models typically require greater computational resources, take longer to train and have inherently less intuitive explainability[198]. Thus, there is an implementation benefit to obtaining a similar performance with simpler models.

High levels of heterogeneity are not unusual in meta-analyses of prediction models[195], given the in study populations, predictor definitions, GDM definition, and validation methods used across studies[151]. The extreme heterogeneity (I^2 99.6%) and wide prediction interval (0.45-0.92) are not too dissimilar to previous research in this area[83,195,196]. However, such heterogeneity may limit the applicability of the pooled AUC to any clinical setting. The

variance component analysis suggested the largest contributor to variance was the difference in predictors (σ^2 ~0.237), with smaller portions attributable to differences between studies (σ^2 ~0.097) and algorithm family (σ^2 ~0.069). This hierarchy suggests that the model predictors are a more significant determinant of performance variability than the specific ML algorithm chosen or general study-level factors. This aligns with findings from my own research demonstrating the impact of different predictors[156] and altering noise in data[155] on AUC, and affirm the mantra "Garbage in – garbage out"[199], that poor quality data leads to unreliable model output. Several clinical and methodological factors likely contribute significantly. Variation in the diagnostic criteria for GDM (e.g., IADPSG vs. NICE) directly alters the definition of the outcome variable, which can substantially impact model performance. Furthermore, the underlying population characteristics, such as ethnicity and baseline GDM prevalence, differ markedly across the included studies, with a heavy concentration of recent research in East Asian populations. As model performance is known to be context-dependent, this case-mix heterogeneity is a potential driver of the observed variability in reported AUCs.

2.4.2 Methodological Concerns

Using PROBAST+AI[102], there were methodological shortcomings and a lack of transparency in the majority of studies included in this review. 63% of studies were deemed to have a high risk of bias, primarily due to the 'Analysis' domain, itself primarily driven by small, unjustified sample sizes, unclear validation strategies and a failure to report appropriate clinical evaluation metrics. The high frequency of analysis concerns suggests that the performance of many published models may be overly optimistic or not reproducible in new samples. Worth noting, risk of bias was the only moderator analysis that showed any effect, where high risk ensemble models performed worse than low risk models (Table 2.6). While bias often leads to inflated performance estimates, it's possible that methodological flaws, such as inadequate sample size for model complexity or improper validation, may be detrimental to these more complex algorithms. The effects are clear in Kaya et al. These researchers built ML models based on two cohorts of 45 and 52 respectively, before applying boosting methods. With such small samples, this resulted in independent test sets of 9 and 11 respectively, far too small to provide any true evaluation[200]. Such small sample sizes render the reported AUCs (0.350-0.933) unreliable, as they are almost entirely driven by random variation in the test-set splits rather than true predictive performance. Small samples sizes have been shown to produce uncertain AUC estimates, which narrow as the sample size increases [77].

In addition to sample size and internal validation concerns, only 16% (6/38) of models underwent external validation, which will likely create barriers to implementation[90,119,201]. External validation is now recognised as an important step in the evaluation of risk prediction models and can highlight where overfitting occurs[77,91,120]. This lack of external validation could in part explain the wide predictive range. These validation concerns can be viewed in Li et al., who performed both internal and external validation of their models. On the interval validation, their boosting model achieved an AUC of 0.996 with an upper confidence bound of 1.000. However, when evaluated on an external dataset, the AUC dropped to 0.834. This is a clear indication that the model is overfitting to the training data and highlights the need for transparency in the reporting of risk prediction models. However, fewer than 10% of reviewed studies made their datasets (8%) or code (3%) publicly available, which inhibits the ability of other scientists to review for sources of data leakage or methodological errors in the development of risk prediction models. Indeed, when I recently requested data for external validation of my own model[156] from 22 authors, 14 of which had a data availability statement, only one author responded and could not share their data[202].

Finally, evaluation of the clinical impact of ML models was lacking in many studies. Calibration was reported in only 58% (22/38) of studies, often lacking detailed calibration plots or essential metrics like calibration slope and intercept, while DCA was only reported in 26% (10/38) of studies. These methodological weaknesses identified, particularly high risk of bias, insufficient external validation, and inadequate calibration reporting, are unfortunately not novel. These issues have been consistently highlighted in earlier systematic reviews of GDM prediction models[83,97,196,203] and in models for other adverse pregnancy outcomes[195]. Ruiter et al.[203], for example, concluded that most GDM prediction models were of moderate to low methodological quality and few had been externally validated. Eight years on and the same conclusions can be drawn, despite the increase research in the area. Table 2.7 summarises the key comparisons between the current review and those that have come before.

Table 2.7. Comparison with Key Prior Systematic Reviews on GDM Prediction

Feature	Current Review 2025	van Eekhout et al. 2025[195]	Zhang et al. 2022[83]	Huang et al. 2022[196]	de Ruiter et al. 2017[203]
Search Period	Jan 2000 – Mar 2025	Apr 2017 – June 2024	2000 – Oct 2020	- Mar 2022	1997 - Dec 2014
Scope: Data types	Routine EHR only	Maternal characteristics	EHR, clinical, biomarkers	No restriction	Routine clinical data
Scope: Prediction time	≤16 weeks	14 weeks	8-24 weeks	13 weeks	14 weeks

No. GDM Studies / Models	38 / 122	31 / 35	25 / 30	48 / 51	14 / 14
Pooled/Median AUC (95% CI)	0.75 (0.71-0.78)	Median 0.71	0.85 (0.8-0.9)	0.77 (0.69-0.84)	Range: 0.63-0.89
ML vs LR	No diff.	NA	ML > LR[a]	ML > LR[b]	NA
RoB/Quality	High RoB, Analysis domain	Very high RoB	High RoB	All high RoB	Most studies low quality

[a] described descriptively and based on ~5 ML studies
[b] based on 3 studies with $^2/_3$ find difference

2.4.3 Strengths and Limitations

To my knowledge, this is the largest synthesis of GDM prediction models using EHR data, incorporating 38 studies with over 2 million pregnancies, and the largest number of models meta-analysed (122). It was conducted according to a pre-registered PROSPERO protocol (CRD420250651833) and is reported in line TRIPOD-SRMA[146], CHARMS[149], and PROBAST+AI[102]. Furthermore, the meta-analysis employed advanced statistical techniques, including a multilevel random-effects model to appropriately handle the hierarchical structure of the data (multiple models per study) and robust variance estimation to ensure the validity of statistical inferences, aligning with best practices for synthesizing prediction model performance[150,151]. This approach differs to many meta-analyses in this area that tend to focus of external validations of the same model[123,195,196]. In addition to discrimination (AUC), I extracted data on calibration and decision-curve analysis, highlighting aspects of model performance that inform clinical translation.

However, the study has several limitations. The included studies had some variability in GDM diagnostic criteria and quality/transparency of EHR data, potentially introducing misclassification or information bias that could affect the pooled estimates. A meta-analysis of the calibration statistics were not possible, as the metric for calibration varied even among studies that did report it in some form, restricting the analysis to AUC. The extreme heterogeneity means the pooled AUC of 0.75 should be interpreted with caution, with a wide prediction interval (0.45-0.92). While several moderators were explored, much of the heterogeneity remains unexplained, potentially due to residual confounding from study-level factors such as differences in healthcare systems or model tuning practices. Although the formal test for publication bias was negative (p=0.211), the possibility of selective reporting of the best performing models within studies cannot be excluded. Finally, this review was

intentionally restricted to models using routinely collected EHR predictors. Therefore, the conclusions do not extend to models incorporating specialised biomarkers, genetic data or non-routine data source.

Nevertheless, this review should provide a representative summary and synthesis of the current state of the art of the evidence. The approach taken here, including the use of PROBAST+AI, TRIPOD-SRMA and multi-level meta-analytic techniques, can serve as a methodological template for future systematic reviews in the advancing field of ML in healthcare. With this in mind, the R code and data files are provided in the supplementary files so that these methods can be reproduced: https://osf.io/rmbdt/

2.4.4 Implications for Clinical Practice and Future Research

Clinicians should view early-pregnancy prediction tools for GDM prediction with cautious optimism. Across studies these models achieve only moderate discrimination and wide performance variability. External validation meta-analyses report even lower AUCs (0.69-0.78)[89,123,195,196,203], and only 6 studies in this review reported external validation, indicating there is still some progress to be made prior to implementing these models. Calibration was reported in 58% of studies and clinical usefulness in 26%, so the net benefit over current screening remains unclear. A model that discriminates yet is mis-calibrated, or fails to improve decisions, can do more harm than good[204–206]. Finally, as complex ML models gained no discrimination advantage over LR, clinical practice should prioritise refining and validating simple, transparent models based on high quality predictors. Stakeholders should therefore require robust validation, transparent reporting and demonstrated utility before approving ML prediction tools for clinical adoption.

Research progress now depends less on algorithmic novelty and more on methodological rigour and transparency. Prospective, adequately powered, multi-centre validation studies must replace small, single-centre development papers[207]. Researchers should share de-identified datasets, code and fully specified models. Where data cannot be shared, researchers should identify similar populations and attempt a model exchange approach to external validation (see Chapter 6). Widespread adoption of TRIPOD-AI[101] guidelines will ensure that published work can be scrutinised and replicated. Comparative studies should test multiple algorithms on the same diverse EHR datasets, using robust internal and external validation. Investigators should identify predictor sets that are routinely recorded and reliably

informative. Future evaluations must go beyond AUC to include, calibration decision-curve analysis and subgroup performance checks to guard against bias[77,91,120].

2.5 CONCLUSION

In conclusion, prediction models using routine EHR data for early GDM prediction achieve moderate discrimination, however, performance is highly variable and often reported from studies with a high risk of bias. Complex ML algorithms appear to offer no clear, consistent advantage over simpler, more transparent model like LR. Studies consist of widespread methodological shortcomings, including lack of rigorous validation, incomplete clinical evaluation of models, and limited transparency, all of which undermine confidence in the performance and reproducibility of many models. Future research should prioritise methodological rigour, multi-centre validation, clinical evaluation and open science practices in order to progress the field. These findings provide a clear mandate for the research undertaken in this thesis. To address these gaps, the following chapters will proceed with a systematic, bottom-up approach: first, by validating the GDM outcome label within our own EHR dataset to mitigate label noise (Chapter 4); second, by developing and internally validating a suite of prognostic models on this curated data, with a focus on transparently reporting both discrimination and calibration (Chapter 5); and finally, by subjecting my primary model to both external and prospective clinical validation to assess its real-world performance and transportability (Chapters 6 and 7). This structured progression directly confronts the key limitations identified in the current body of evidence and aims to build a more robust foundation for the clinical translation of prognostic models in maternal care.

Chapter 3

Cleaning, Coding, Curating: Preparing Maternal Health Data for Machine Learning

Chapter Overview

This chapter provides the foundation for the empirical work of the thesis, creating a reliable, high-quality dataset from complex, real-world EHRs. While not tied to a single research question, this work is fundamental to answering all subsequent questions by ensuring the integrity of the data upon which all models are built and evaluated.

The chapter details the multi-stage process of sourcing, cleaning, coding, and curating several datasets, primarily from the Coombe Hospital EHR system, which initially comprised data from over 80,000 pregnancies. It details the creation of the key datasets used throughout the thesis, including the primary validated EHR dataset (n=27,561), the sequential pregnancies dataset for multiparous women (n=4,005), and the cohorts for treatment pathway prediction and prospective validation. Key processes described include the handling of high-cardinality categorical variables like ethnicity, the interpretation of missing data in consultation with clinicians, the detection and management of outliers, and the engineering of new, clinically relevant features such as inter-pregnancy weight gain and birthweight percentiles.

This chapter reframes data preparation not as a preliminary step, but as a core research contribution. The decisions made here, such as interpreting a null value for a complication as its absence or creating a high-confidence GDM label by cross-referencing with a clinical team database, are interpretive acts grounded in clinical domain knowledge. These choices have a direct impact on every subsequent result, demonstrating that data quality is an important scientific process, not merely a technical one. The deliberate creation of distinct datasets (e.g., the main cohort, the sequential cohort) also reveals the architecture of the thesis's empirical evidence. This structure was designed to enable the multi-faceted investigation that follows, with each dataset purposefully curated to address a specific research question.

3.1 INTRODUCTION

This chapter describes in detail the data sources utilised in this research, the procedures for data cleaning and preprocessing, and the composition of the final analytic datasets. A rigorous understanding of the data is essential in a machine learning (ML) study, particularly in healthcare where data quality and preparation significantly influence model performance[208]. The following sections outline each dataset used, including a pilot study dataset, the main hospital EHRs, and various supplementary data sources, and the steps taken to clean, merge, and preprocess these data. The rationale behind key preprocessing decisions (informed by clinical consultation) is discussed, and a summary of the final datasets is provided. Finally, the chapter addresses the limitations of the data and the mitigation strategies employed.

3.2 PILOT DATASET (initial study cohort)

The first dataset investigated was a pilot cohort used to explore the feasibility of early prediction of GDM. This pilot dataset stems from a prior clinical study at the Coombe Women and Infants University Hospital (described by O'Malley et al.[29]) that focused on GDM diagnosis and monitoring. The cohort consisted of 196 pregnant women, of whom ~53.5% were diagnosed with GDM under the IADPSG criteria[1]. Initially 211 variables were recorded for each participant, including demographic information, clinical measurements, and laboratory values from throughout the pregnancy. Because the aim of this research is to predict GDM early in pregnancy, the majority of these variables (available in late pregnancy) were not usable for early prediction modelling. Therefore, the feature set was subsequently reduced to 17 key predictors available by the first antenatal booking visit (~12 weeks' gestation). This included baseline maternal characteristics (age, body mass index, etc.), relevant medical history (e.g., family history of diabetes), and any early pregnancy lab tests or vitals. The pilot dataset was relatively small, and there was no missing data due to the controlled nature of data collection in the original study.

The cleaned pilot data provided an initial proof-of-concept for predictive modelling. For example, a logistic regression model trained on these data achieved an area under the ROC curve (AUC) of about 0.70 in classifying GDM, demonstrating the potential of early predictors. This was promising for two reasons. First, the relatively small size of the dataset. But second, the population had been pre-selected based on having risk factors for GDM (e.g. high BMI)[209], thus reducing the potential of this variable to predict outcomes. Further, I bootstrapped the model on 10% increments of training data, demonstrating to stakeholders how the models

improved with more data (Figure 3.1). This analysis was presented at The Physiological Society conference in London, UK, September 2022. The insights from this pilot study helped shape the approach for the larger retrospective dataset described next.

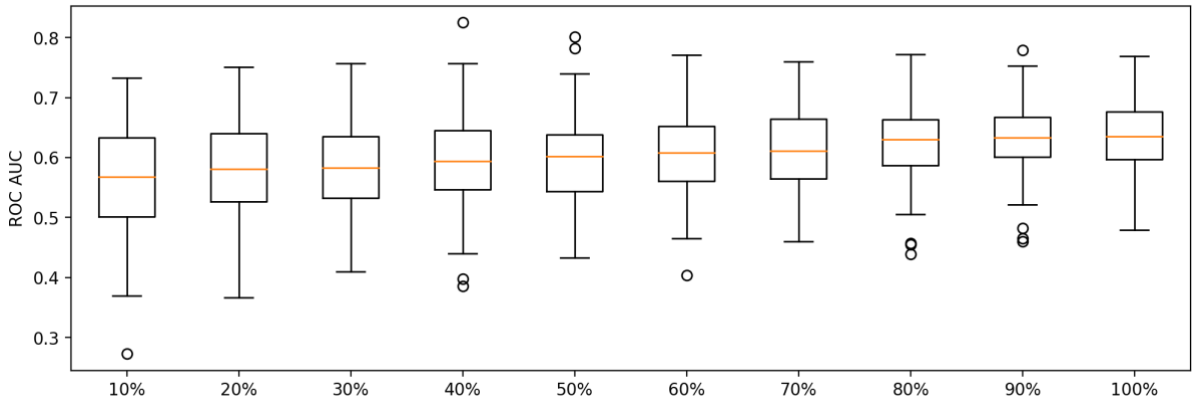


Figure 3.1. Logistic regression model trained on 10% increments to demonstrate enhanced ability to predict and reduce variance in predictions.

3.3 COOMBE HOSPITAL ELECTRONIC HEALTH RECORDS

The primary data source for this research was the Coombe Women & Infants University Hospital EHR system. The Coombe Hospital, a major maternity hospital in Dublin, uses a dedicated maternity EHR system called Euroking K2 (Euroking Maternity Software Systems, UK) to collect and manage patient data. Trained midwives and clinicians routinely enter patient information during antenatal visits using standardised electronic questionnaires, ensuring a degree of consistency in how data are recorded. This hospital delivers 7-8,000 pregnancies per year.

The data was supplied as a single flat file. Within this file, data corresponding to four distinct tables were: 1) The mother's EHR, 2) The baby's chart, 3) Previous pregnancy records, and 4) Previous birth records. For extraction, header lines were split out, assigned each data row to the matching header by column count, and four separate dataframes were built. After removing header duplicates, blank cells and rows with misplaced prefixes, I re-routed any pregnancy rows that had drifted into the Baby table, deduplicated each dataset, and finally merged the Pregnancy and Baby tables on patient ID and infant date of birth to link current pregnancies with outcomes and past obstetric history for modelling. The initial raw dataset covered the period from 2013 to 2023 and consisted of 81,798 rows and 161 columns. After an initial phase of data processing, this was reduced to 78,915 rows.

3.3.1 EHR Data Cleaning and Preprocessing

Working with EHRs poses challenges such as missing values, free-text entries, and inconsistent data entry, all of which must be addressed to ensure a reliable dataset. The goal of the preprocessing was to ensure data quality and consistency, transform variables into a suitable format for machine learning, and integrate the multiple datasets correctly, described later. All data cleaning was performed using Python (pandas library). Table 3.2 at the end of this section summarises the main cleaning steps, actions taken, and rationale.

3.3.2 Categorical Feature Processing

The Coombe EHR data presented challenges with respect to categorical variables, many of which contained a large number of unique values. For example, the 'ethnicity' column contained 1,331 unique string entries, and a column for 'Psychological problems' had 6,546 distinct values. Directly using such features in ML models, for instance through one-hot encoding, would lead to an unmanageably large and sparse feature space. Thus, the primary strategy for most categorical variables was binary encoding (e.g., coded as 1/0 or YES/NO). While this approach significantly reduced the dimensionality of the feature set, making it more amenable to modelling, it inherently involved a loss of granular detail. For instance, binarizing 'Psychological problems' meant that information about the specific type of psychological issue was not retained in that processed feature. This was true for all such features. Given this potential loss of information, features were examined post-hoc to assess their influence on ML performance. This simplification, while pragmatic for managing dimensionality, is acknowledged as a potential limitation that could impact predictive performance and the ability to uncover more nuanced relationships. Alternative, more sophisticated encoding strategies for high-cardinality categorical features (e.g., target encoding, frequency encoding, or feature hashing) were considered, but the binary encoding approach was adopted for its simplicity and effectiveness in reducing feature space in this specific context. Any features that had been binary encoded but demonstrated strong predictive performance were further investigated for key features within that category, such as Endocrine problems described below.

An important refinement to this general strategy was the pre-emptive extraction of critical information. Before applying broad binary encoding to a categorical field, specific, highly relevant pieces of information were identified and extracted into new, dedicated features. A key example is the 'Endocrine problems' column; this contained information indicating a history of GDM, this was captured as a separate 'previous GDM' feature. This

ensured that this vital predictor was preserved before the parent 'Endocrine problems' column underwent more general binarization. This demonstrates the integration of domain knowledge into the data processing pipeline, prioritising known risk factors. Other variables required tailored recoding rather than binarization.

Certain variables were handled with specific re-categorisation rules. The 'Ethnicity' column, for instance, was not binarized. Instead, its 1,331 unique values were mapped into six predefined, broader categories: Caucasian, South East Asian, Asian, Black African, Middle Eastern, and Other. This grouping was determined based on considerations of clinical relevance and ensuring sufficient numbers within each consolidated category for meaningful analysis, with less frequent ethnic groups being aggregated into the 'Other' category.

3.3.3 Addressing Missing Data

Missing data is a common issue in EHR datasets. I distinguished between two types of missing data: entire records missing key information, and specific fields missing in some records. First, I dealt with incomplete records. A small number of pregnancy records in the EHR extract were missing critical fields that are essential for my analysis (for example, a few records had no entry for maternal age, booking height/weight, or GDM outcome). Such records were not usable for modelling because they lacked fundamental predictors or the target label. Rather than attempt imputation for these crucial values, I removed any pregnancy record that was missing any critical variable. Critical variables were defined with clinical input and included: maternal age, booking weight/height, and GDM outcome label. This step led to the exclusion of a small number of cases (n=991). To ensure that there was no systematic bias in the deletion of these subset of patients, I reviewed the demographic characteristics of the deleted patients relative to the total population. The characteristics of these two cohorts are shown in Table 3.1.

Features with a very high proportion of missing values, specifically, more than 30% missing data, were generally removed from the analytical dataset. This was particularly true for newer measures that had only started to be recorded in recent years, such as COVID-19 vaccine status, as these had insufficient historical data for robust analysis. A rather large exception to this rule was made if the missingness itself could be reliably and meaningfully interpreted as a null or negative finding for that specific variable. For certain variables, missing values were not treated as unknown data points but were interpreted based on clinical context, often as indicating a negative finding (i.e., the absence of a condition). For example, in the

'shoulder dystocia' field, which was sparsely populated, a null value was assumed to indicate the absence of this complication and was accordingly coded as 'NO' or '0'. These interpretations were not made in isolation but were the result of constant conversation with clinicians at the Coombe hospital. This collaborative approach, conducted over many weeks and months, ensured that assumptions about missing data were clinically sound and contributed to a robust and reliable dataset. The majority of missing values were resolved taking this approach.

For remaining missing values in retained features after the above steps were implemented, missing values were imputed using the median value for numeric data (e.g. haemoglobin) whilst the mode was used for categorical data (e.g., 'Maternal blood taken'). While I acknowledge many recommend the imputation of missing data[210,211], arguing that listwise deletion (removing records with any missing data) can cause bias and increased variance, especially if data are not Missing Completely at Random (MCAR), the decision against universal imputation was made carefully. However, concerns also exist that imputation, particularly if misapplied or if its assumptions do not hold, might introduce its own biases or negatively impact model prediction output in EHRs[212]. Given that the evidence for data imputation typically stems from datasets much smaller than the current (~7,000 vs ~70,000), the low occurrence of absolute missing data after data cleaning, and the potential for introducing bias into the model, it was decided that the safer approach to was to remove any truly missing data instead of imputing for the critical values listed above, with the aim of resulting in a purer dataset.

3.3.4 Outlier Detection and Data Consistency

I examined all numerical variables for outliers and implausible values, as EHR data can contain data entry errors (for instance, a misplaced decimal or a typographical mistake). Maternal age boundaries were set between the age of 18 to 50, ensuring outliers above and below this range were excluded. Height and weight boundaries were similarly set, at 120cm for height and 200kg for weight. For instances where height equalled weight, these were excluded as possible duplicate entries (e.g. both entered as 120). If an obvious correction was possible (e.g., a missing decimal point), I corrected it; if not, I dropped those records to avoid skewing the data. This process was repeated for blood pressure measurements. For some laboratory values, remaining missing values were imputed as the mean value (e.g. haemoglobin), as the range of this value was very narrow regardless.

3.3.5 Feature Engineering

Beyond cleaning existing variables, several new features were derived, and existing ones were transformed to enhance their potential utility for the machine learning models.

The 'Occupation of the Mother' variable, a free-text or coded field in the EHR, was processed to serve as a proxy for socioeconomic status (SES). This involved mapping the recorded occupations to the International Standard Classification of Occupations (ISCO) framework[213]. These ISCO codes were then grouped into skill level categories, ranging from 0 (representing unemployment) to 4 (representing the highest skill level). While this provided an available measure related to SES, it was acknowledged that occupation serves as a "very weak proxy" for true socioeconomic status.

Infant birthweight, a raw measurement, was transformed into a more clinically meaningful and standardised measure, birthweight percentile. This was calculated using the baby's recorded weight and the gestational age at delivery, with reference to established international percentile ranges[214]. This conversion was also necessary for the creation of a later dataset.

The inter-pregnancy interval, which has been associated with increased GDM risk[215], was calculated as the number of days from the delivery date of one pregnancy to the delivery date of the next, given I had no conception data. Inter-pregnancy weight gain was calculated as the difference between the mother's booking weight in the later pregnancy and her booking weight in the preceding one, capturing any weight retained or gained in the interim, another recognised driver of future GDM risk[216].

3.3.6 GDM Outcome Label

Within the EHR, GDM was diagnosed and recorded according to IADPSG criteria. In practice, a lack of universal screening typically means only 50-60% are tested for GDM, resulting in an estimated under diagnosis of ~16%[23]. GDM diagnoses were extracted from the EHRs based on information recorded in a column titled "Medical problems during pregnancy." When this column contained the entry "Diabetes developed during pregnancy", the patient was coded as having GDM in a newly created column designated for this study's analysis, referred to hereafter as "EHR-GDM." Patient records not meeting this criterion were coded as not having GDM. The legacy EHR has a single structured problem field; it does not store ICD or SNOMED codes. GDM is recorded exclusively by selecting 'Diabetes developed during pregnancy' from that field's drop-down list. No alternative structured or coded location exists.

3.3.7 Final Processed EHR Dataset

Following the comprehensive multi-stage pipeline of data extraction, structuring, cleaning, validation, and feature engineering, the resultant primary analytical dataset derived from the Coombe Hospital EHRs comprised records for 73,242 pregnancies. This dataset includes records from many individual women, some of whom may have experienced more than one pregnancy during the study's observation period (2013-2023). This dataset formed the basis of two studies covered in Appendix B, 'A 10-year review of periconceptual folic acid supplementation in women with epilepsy taking antiepileptic medications', and Appendix C, 'Trends in prevalence of and risk factors for obesity during pregnancy in Ireland: Longitudinal evidence from a large tertiary maternity hospital'.

Table 3.2. Summary of data cleaning and preprocessing steps for EHRs.

Step / Issue	Action Taken
Missing critical fields in a record	Excluded entire record from dataset.
Variables with >30% missing values (sparse data fields)	Removed those variables from feature set (not used in modelling).
"Missing" indicating no condition	Imputed missing values as "No" and non-missing as "Yes" (then binary-encoded). Example: *Cardiac problems* text -> Yes/No.
Outliers or implausible values (e.g., out-of-range age, extreme or negative BMI)	Verified against other data and corrected when obvious (e.g., likely typos); otherwise removed those data points/records.
Categorical variables	Converted to binary indicators (1/0). Treated blank as "No" if applicable.
Multi-category variables with many levels or incomplete data (e.g., ethnicity, conception method)	Collapsed categories into fewer groups or binary classes. Example: ethnicity -> 6 categories.
Free-text occupational titles (socioeconomic indicator)	Mapped occupations to ISCO-08 codes & then to 0–4 skill level scale.
Free-text medical history	Consolidated into binary feature (Yes=present, No=none).
Numeric continuous variables (age, BMI, blood pressure)	Left as continuous (with normalisation later).
Residual missing values in retained features	Dropped records with any remaining missing data in features after above steps.
Calculating birth weight percentile	Computed percentile using gestational age and birth weight against reference chart (Nicolaides et al., 2018). Added as a field in outcomes dataset.
Integrating multiple datasets	Merged datasets on unique IDs (mother ID, pregnancy ID). Ensured one-to-one merges (e.g., each pregnancy links to one

Table 3.2. Summary of data cleaning and preprocessing steps for EHRs.

Step / Issue	Action Taken
	baby outcome). Excluded records where linkage was not found or inconsistent.
Sequential pregnancy feature engineering	For mothers with ≥2 pregnancies, linked prior pregnancy data to current pregnancy. Computed inter-pregnancy interval and weight change, added to current pregnancy record.
Finalising clean analytic dataset	Split data into training and validation sets (ensuring no patient overlap).

3.4 CLINICAL TEAM GDM VALIDATION DATASET (CTD)

Given how GDM was encoded within the EHR, it was important to attempt to validate this label. A separate dataset was obtained from the clinical diabetes management team at the Coombe hospital. This team maintains its own dedicated clinic records of all patients they actively diagnose and manage for GDM. The primary purpose of acquiring this dataset was to use it as an independent reference standard to validate the GDM diagnoses recorded in the main EHR system (the 'EHR-GDM' label). The dataset provided by the clinical team contained patient identifiers and the year in which each patient was managed for GDM. This CTD only covered the years from 2018 to 2022, inclusive. Records for the year 2023 were incomplete at the time of data acquisition, and maintained records of this nature did not exist prior to 2018. This temporal limitation meant that the GDM label validation process could only be applied to a subset of the main EHR data.

The data from the CTD was cross-referenced with the main EHR dataset for the overlapping period (2018-2022). The year 2020 was excluded (the detailed rationale for the exclusion of 2020 data is provided in Chapter 4 of this thesis). This validation and filtering procedure resulted in a refined dataset comprising 27,561 pregnancies for which the GDM diagnosis was considered to be reliable. This model cohort formed the basis for model development and validation in Chapters 5 and 6. The full breakdown of these features is available in Appendix D.

3.4.1 Internal Hold-out Set

In order to evaluate the developed ML models on unseen data, a validation dataset was defined as an independent hold-out subset of the Coombe EHR data. Rather than being a separate external source, this dataset was created by setting aside a portion of the hospital EHR

records that were not used in model training. Approximately 10% of the validated EHR records (~2,700 pregnancies) were reserved in this way, stratified to represent the overall cohort. The selection was done such that all records from a given patient were kept together in either the training set or the validation set (using a group-wise splitting by mother ID) to prevent any information leakage. It was treated as a stand-alone dataset during analysis: all data preprocessing steps were applied to it after being derived from the training data, without using any information from this set in the model development. This dataset was used to assess model generalisation. It should be noted that I also explored obtaining a true external dataset from another hospital for validation; however, due to data access limitations, this was not achieved[202].

3.4.2 Multiparous Pregnancies Dataset

Within the GDM-validated dataset of 27,561 pregnancies, women who had experienced more than one pregnancy recorded within this timeframe were extracted to form a distinct Multiparous Pregnancies Dataset'. The primary purpose of creating this subset was to perform analyses focusing on modelling future risk of GDM, a model that could be used during preconception or at the end of current pregnancies. Specifically, it allowed for the investigation of whether historical pregnancy information from a woman's previous pregnancy (e.g., prior GDM status, previous birth outcomes, etc) could be used as predictive features for GDM occurrence in her subsequent pregnancy. This line of research complements the broader public health move towards preconception as a time for intervention in GDM[15,17].

3.5 OGTT LABORATORY RESULTS DATASET

The Oral Glucose Tolerance Test (OGTT) laboratory dataset consists of the actual lab results for the OGTTs performed on pregnant women at the Coombe Hospital during 2018–2022. The initial research plan included using a comprehensive database of OGTT results from the hospital laboratory system, covering the years 2018-2022. The intention was to use these actual glucose values to validate GDM diagnoses by directly applying the IADPSG diagnostic thresholds and potentially as strong predictive features themselves. However, this bulk OGTT dataset unfortunately became corrupted during the extraction process from the legacy hospital system. The corruption resulted in the mixing of fasting, 1-hour, and 2-hour glucose values in

a way that lacked any recognisable systematic pattern that could be programmatically reversed to reconstruct the correct individual results.

Despite this, the value of OGTT results in predicting treatment pathways was considered significant enough to warrant a salvage operation involving manual data extraction. For a subset of 125 GDM-positive patients selected from each year of 2018-2022 period (excluding 2020), their individual OGTT results were manually looked up in the hospital's laboratory system on a case-by-case basis. The correct fasting, 1-hour, and 2-hour glucose values for the diagnostic OGTT were found and recorded, resulting in an OGTT dataset with 486 patients. This dataset was subsequently merged with the treatment pathway dataset to create a smaller dataset combining EHR data with OGTT results to predict GDM treatment pathway (Appendix I).

3.6 PROSPECTIVE CLINICAL VALIDATION DATASET

I undertook a prospective, single-centre evaluation between 1 December 2024 and 31 January 2025. Women attending either their first-trimester booking visit or routine dating ultrasound (both scheduled at 11 +0 to 13 +6 weeks' gestation) were approached in the outpatient waiting area by trained research staff. After receiving a brief verbal explanation, printed information leaflet, and opportunity to ask questions, eligible women provided written informed consent. Immediately after consent, the researcher opened the participant's routine EHR (Euroking K2) and entered the first-trimester variables required by the GDM-prediction model. A total of 298 women met eligibility criteria and provided complete follow-up data. Missingness was negligible because data were captured prospectively at the point of care; no imputation was required. This prospectively collected cohort was used exclusively for final clinical validation (Chapter 7).

Table 3.3. Summary of datasets used in the study. Each dataset is described with its time frame, sample size, and key details or purpose in the research.

Dataset	Chapter	Sample Size	Description
Pilot Dataset (Pilot Study)	Pilot	196 pregnancies	Initial ML feasibility testing for GDM prediction, methodological refinement.
Coombe EHR Dataset	Chapter 4	73,242	Primary data source for broad GDM research; supports supplementary studies on obesity and folic acid.

Table 3.3. Summary of datasets used in the study. Each dataset is described with its time frame, sample size, and key details or purpose in the research.

Dataset	Chapter	Sample Size	Description
Validated EHR Dataset	Chapter 4, 5 & 6	27,561	Core dataset for primary GDM prediction model development and evaluation, with high-confidence GDM labels.
Internal hold-out	Chapter 5	~2,700	Held-out subset of the Coombe EHR data used for model validation.
Multiparous Pregnancies Dataset	Chapter 5	4,005	Subset of EHR dataset to train a specialised model incorporating past pregnancy information.
OGTT Laboratory Results Dataset	Appendix I	486	Laboratory database of oral glucose tolerance test (OGTT). Incorporating actual glucose values in analyses, especially for treatment pathway prediction.
Clinical Validation Dataset	Chapter 7	298 births (singleton infants)	Prospective clinical validation collected in real time in the Coombe Hospital.

3.7 DATA LIMITATIONS

There are still several limitations with this dataset. First, all data were from a single tertiary maternity hospital (The Coombe Hospital). While I sought external validation sets[202] (documented in Appendix E), I was unsuccessful. The need for external validation in a different hospital or a prospective setting is recognised and is discussed in Chapter 6. Second, As with any retrospective dataset, this data were not originally collected for research, and thus prone to inconsistencies or errors. Despite data cleaning and preprocessing described here, it is still possible that a % of the data were incorrectly entered. Third, many potential risk factors were not captured in the legacy EHR system. For example, SES beyond occupation or accompanying laboratory results such as fasting plasma glucose[217] or HbA_{1c}[218] are not integrated within the system. Fourth, the exclusion of year 2020 data, while necessary for consistency, means the dataset has a gap and possibly slightly reduced sample size. The year 2020 saw different screening criteria due to Covid-19 restrictions[219]. Finally, I relied on the EHR and CTD for the GDM label. However, this label is compromised by the lack of universal screening in the hospital. Previous research in Ireland suggests that up to 16% of GDM cases are missed due to lack of screening[23], meaning the labels may be incorrect for these undiagnosed women. This "hidden GDM" means that a proportion of women labelled as non-GDM in the study (including in the validated "VAL-GDM" cohort) may have had undiagnosed GDM. This represents an

irreducible level of label noise which could attenuate the true effects of predictors and potentially bias model parameters. Finally, the mixed approach to handling missing data, which included some targeted imputations (e.g., median for haemoglobin, mode for some categorical) alongside listwise deletion for critical missing values, aimed to balance data preservation with concerns about imputation-induced bias. However, any listwise deletion still carries the risk of introducing selection bias if the data are not Missing Completely at Random (MCAR) for those instances, potentially affecting the representativeness of the final analytical sample.

3.8 CHAPTER SUMMARY

This chapter detailed the origin, extraction, cleaning and integration of all datasets underpinning the subsequent ML analyses. The Pilot Dataset served as the basis for a feasibility study, showcasing to stakeholders what is possible. The EHR database served as the primary data for this PhD and provides the foundation for most other datasets. It is directly relevant to Chapter 4, as well as appendix B & C. When merged with the clinician-validated CTD it results in the most reliable cohort, used in the modelling pipelines in Chapters 4–6; its multiparous pregnancy subset also supports model development in Chapter 5. The Diabetes Pathway dataset, with and without the OGGT Laboratory dataset, are used to predict GDM treatment, with analysis detailed in Appendix F. Finally, the Prospective Clinical dataset is used for real time clinical validation of the ML model, detailed in Chapter 7.

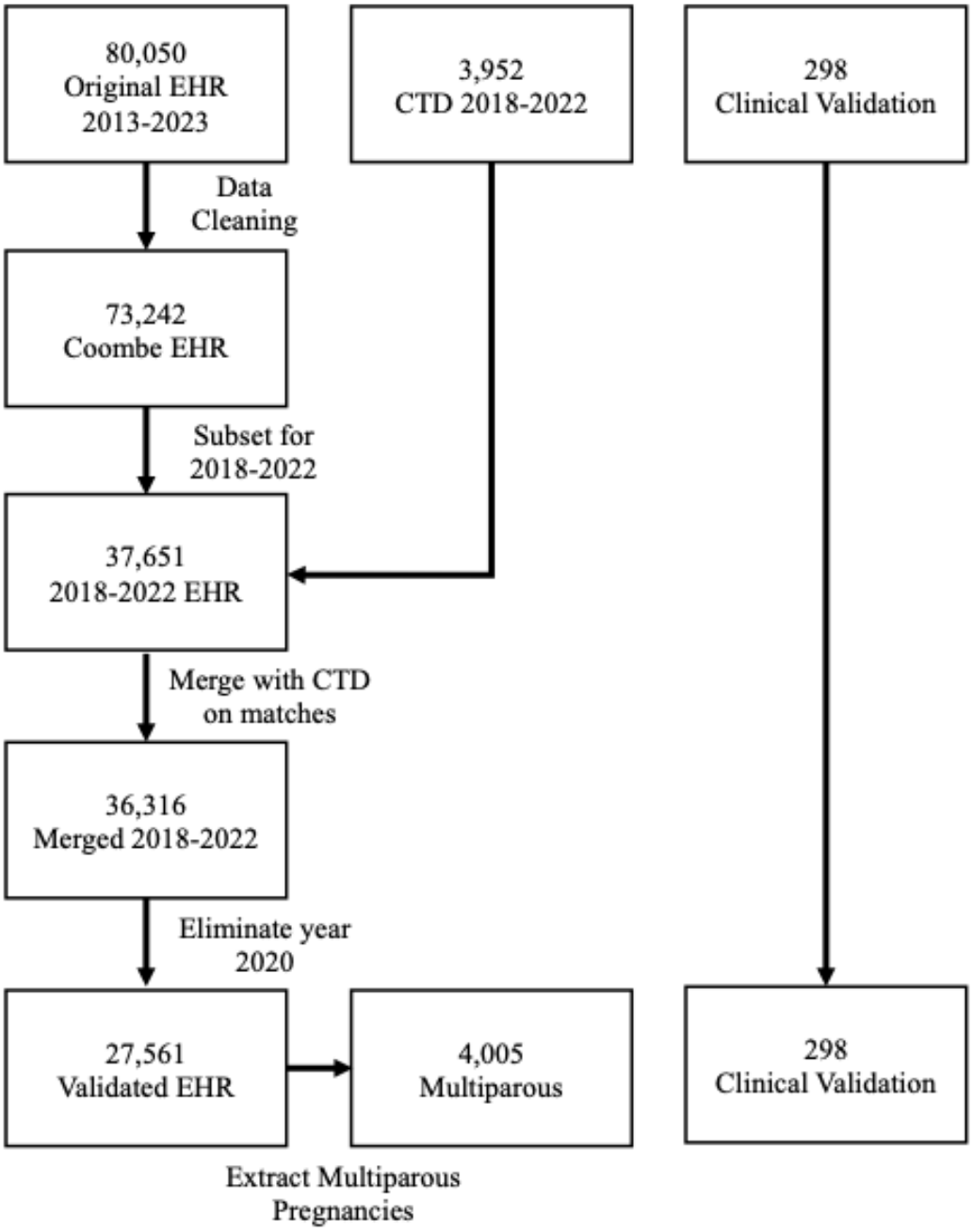


Figure 3.2. Flowchart illustrating the data pipeline from raw EHR extraction to the final analytical datasets

Chapter 4

Gestational Label Accuracy in Electronic Health Records and Its Impact on Machine Learning Models for Early Prediction of Gestational Diabetes: 3-Step Retrospective Validation Study

Germaine, M., O'Higgins, A. C., Egan, B., & Healy, G. (2025). Label Accuracy in Electronic Health Records and Its Impact on Machine Learning Models for Early Prediction of Gestational Diabetes: 3-Step Retrospective Validation Study. *JMIR Medical Informatics*, *13*(1), e72938.
DOI: 10.2196/72938

Chapter Overview

This chapter addresses the label noise research question: *To what extent are the data contained in the EHRs accurate, and if there is label noise, to what extent could this impact ML modelling of GDM?*. The study undertakes a validation of GDM diagnoses recorded in EHRs by comparing them against a ground truth database maintained by the hospital's clinical diabetes team.

The analysis identified significant discrepancies, with 564 false negative cases (1.5% of the total cohort) and 771 false positive cases (2.0%) present in the EHR data from 2018-2022. In addition to this, significant discrepancies were identified in the year 2020 due to covid-19 related disruptions to routine GDM screening.

Logistic regression models were trained on the noisy EHR labels versus one trained on the clean, validated labels. The performance during model training was negligibly affected, with both models achieving an AUC of 0.817 (95% CI: 0.803–0.832) and similar calibration slope (0.98, 0.92-1.05 vs 0.95, 0.90-1.01) and intercept (0.12, 0.05-0.20 vs 0.05, -0.02-0.13). However, the impact of label noise was more pronounced during model evaluation. This demonstrates an asymmetric impact of label noise: it is more detrimental to the evaluation of a model than to its training (when the level of noise is low). Simulations in which label noise was progressively increased confirmed this effect, showing that model performance declines as noise increases, with a particularly strong negative impact of false positives in the test set. This chapter details the act of data validation arguing that without a formal process to ensure the integrity of outcome labels, any claims about a model's performance are built on unstable foundations.

4.1 INTRODUCTION

The systematic review in Chapter 2 identified that the existing evidence base for GDM prediction is characterised by high heterogeneity and significant risk of bias. Furthermore, Chapter 3 detailed the complex process of curating the raw Coombe Hospital EHR dataset. A critical, unaddressed challenge, however, is the reliability of the GDM outcome label itself. As established in the introduction, ML models are sensitive to "label noise," and their predictions are meaningless if the "ground truth" they are trained on is flawed.

Therefore, before prognostic modelling can begin, the integrity of the EHR-derived GDM label must be established. This chapter directly addresses RQ2 (Label Noise) by performing a rigorous validation of the EHR GDM diagnosis against a 'ground truth' reference standard: the hospital's dedicated Clinical Team Database (CTD). This three-step study will first quantify the discrepancy (the label noise) between the two sources, second, investigate the impact of this noise on a baseline prediction model, and third, simulate the effects of increasing noise levels to understand the model's tolerance to data error.

EHRs are an important source of real-world data, offering detailed, longitudinal patient information historically stored in medical charts, and forming the basis of real-world evidence[220,221]. Together with advancements in artificial intelligence and machine learning (ML), EHRs are increasingly being used to develop models that improve prediction of health and disease outcomes[222]. Integration of EHRs into clinical research offers numerous opportunities for advancing healthcare delivery and patient outcomes. However, EHR data is often stored in unstructured formats like free text, requiring information extraction algorithms to enable ML applications[223]. This extraction process can introduce data quality concerns due to various issues such as data entry errors and cut and paste errors[224]. The quality and consistency of EHR data is particularly critical when the target variable, i.e. the variable being predicted, is used in ML models.

Inaccuracies in EHRs present challenges for developing and applying ML algorithms in healthcare, primarily due to the dependency on data quality and accuracy of target labels[225]. This “label noise”, which refers to inaccuracies or inconsistencies in the data labels (e.g., diagnosis codes) extracted from EHRs, can significantly impact model performance by introducing errors in the target variable, leading to potentially misleading conclusions[226]. Training ML models on unvalidated EHRs may lead to systematic errors in the model output with the potential for the model to miss, underestimate, or overestimate clinically significant relationships[227,228].

Accurate diagnosis and recording of GDM in EHRs is important not only for effective patient management, but also for informing public health strategies and economic forecasting in national healthcare planning[229,230]. EHRs are often used to train ML approaches that support clinical decision-making and care pathways that improve pregnancy outcomes[57]. However, the utility of EHRs remains a concern due to potential discrepancies in data recording practices[227]. When using ML in GDM prediction[99], the accuracy of input data is paramount because inaccuracies can lead to flawed prediction models, and ineffective or adverse clinical decisions[231].

Several studies have utilized EHRs to build ML models predicting the likelihood of developing GDM later in pregnancy[83], but none have described validation of the GDM 'label' within the EHRs. This study has three primary aims: first, to assess the accuracy of reporting of GDM diagnoses in EHRs by comparing them to a database maintained in real-time by the hospital's clinical team; second to evaluate how discrepancies in GDM reporting impact machine learning models; and third, to examine ML model performance using varying levels of simulated label noise in the dataset. By identifying discrepancies between these data sources, I aim to highlight the importance of data validation for advancing digital health and ML-driven healthcare.

4.2 METHODS

4.2.1 Study Design

A retrospective validation design was employed to assess the accuracy of GDM diagnoses recorded in the EHRs of a national maternity hospital (The Coombe Hospital, Dublin). I matched patient identifiers (IDs) between the EHR system and a reference standard established by a real-time clinical team database (CTD) of those formally diagnosed with and managed for GDM, which served as a ground truth. This approach allowed for direct comparison between the recorded GDM status in the EHRs and the validated GDM status from the CTD, enabling identification of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in the EHRs. Further, the effect of label noise on ML model performance in predicting the development of GDM (binary classification) was evaluated by firstly, examining its impact in the current EHR dataset, and then secondly, simulating progressively increasing levels of label noise to understand its effect on ML model performance both in terms of training and testing. For clarity in this chapter, I refer to two versions of the GDM outcome label: the 'EHR label', meaning the GDM diagnosis as originally recorded in

the hospital's electronic health record system; and the 'CTD-confirmed label', meaning the outcome as verified by the Clinical Team Database (considered the ground truth for GDM in this study).

4.2.2 Data Source and Validation

The EHR system serves as the repository for patient medical histories, including diagnoses, family history, and outcomes for pregnant women receiving care at the institution. The data is collected routinely from all women by trained midwives using standardized questions and is then computerized onto the electronic system of the hospital, "Euroking K2". EHRs were collected from 2018 to 2022 and consisted of over 35,000 pregnancies during this time. The dataset from the CTD spanned from 2018 to 2022, thus the timeframe for this analysis spanned from 2018 to 2022 (inclusive). Women aged 18 or above with complete information on GDM status were included in the analysis. Pregnancies with missing or incomplete data for critical variables, women without a recorded GDM status, and pregnancies with pre-existing diabetes were excluded. ML models were trained and tested on pregnancies with complete EHR data up to the 12^{th} week of gestation.

GDM diagnoses were extracted from the EHRs based on information recorded in a column titled "Medical problems during pregnancy." When this column contained the entry "Diabetes developed during pregnancy", the patient was coded as having GDM in a newly created column designated for this study's analysis, referred to hereafter as "EHR-GDM." Patient records not meeting this criterion were coded as not having GDM. The legacy EHR has a single structured problem field; it does not store ICD or SNOMED codes. GDM is recorded exclusively by selecting 'Diabetes developed during pregnancy' from that field's drop-down list. No alternative structured or coded location exists.

Patient IDs from the EHRs were then matched with a separate database maintained in real-time by the clinical team responsible for diabetes care, with patient details entered each day upon confirmation from the hospital laboratory of a diagnosis of GDM from an OGTT following the IADPSG guidelines. The matching process was an automated process, whereby the CTD dataset contained the patient ID and the year of birth. These were then merged with the EHR dataset, creating a new column containing a '1' for patients that were in the CTD dataset (and thus registered as GDM positive) and a '0' patients not present in the dataset. This matching process produced a merged dataset for validating EHR-recorded GDM diagnoses against the CTD database, leading to the creation of two comparison columns: "EHR-GDM"

for EHR-identified cases of GDM and "CTD-GDM" for cases of GDM recorded by the CTD. Finally, where both the EHR-GDM label and CTD-GDM label agreed, these data were retained for analysis. Disagreements were discarded in an attempt to create a clean dataset where we could be confident of the label integrity (Figure 4.1). The CTD was considered the definitive ground truth for GDM diagnoses, given its real-time, clinician entered, laboratory confirmed data recording process.

The validation process involved comparing the GDM diagnosis status in the EHR ("EHR-GDM") with that in the CTD ("CTD-GDM") to examine the agreement between the two datasets. This comparison allowed for the identification of TPs (positive in EHRs and present in CTD), FPs (positive in EHRs and not present in CTD), TNs (negative in EHRs and not present in CTD), and FNs (negative in EHRs and present in CTD), and thereby enabling evaluation of the accuracy of the reporting of GDM diagnosis in the EHRs. An additional column, VAL-GDM, was created indicating a positive or negative diagnoses of GDM for cases where the EHR-GDM and CTD-GDM labels matched i.e. for TPs and TNs excluding records with FPs and FNs. Thus, for the purpose of the following stage of ML modelling, only records that matched between EHR-GDM and CTD-GDM were used, reducing the risk of bias from either dataset. The true positive rate (TPR), false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) were calculated for the dataset [75].

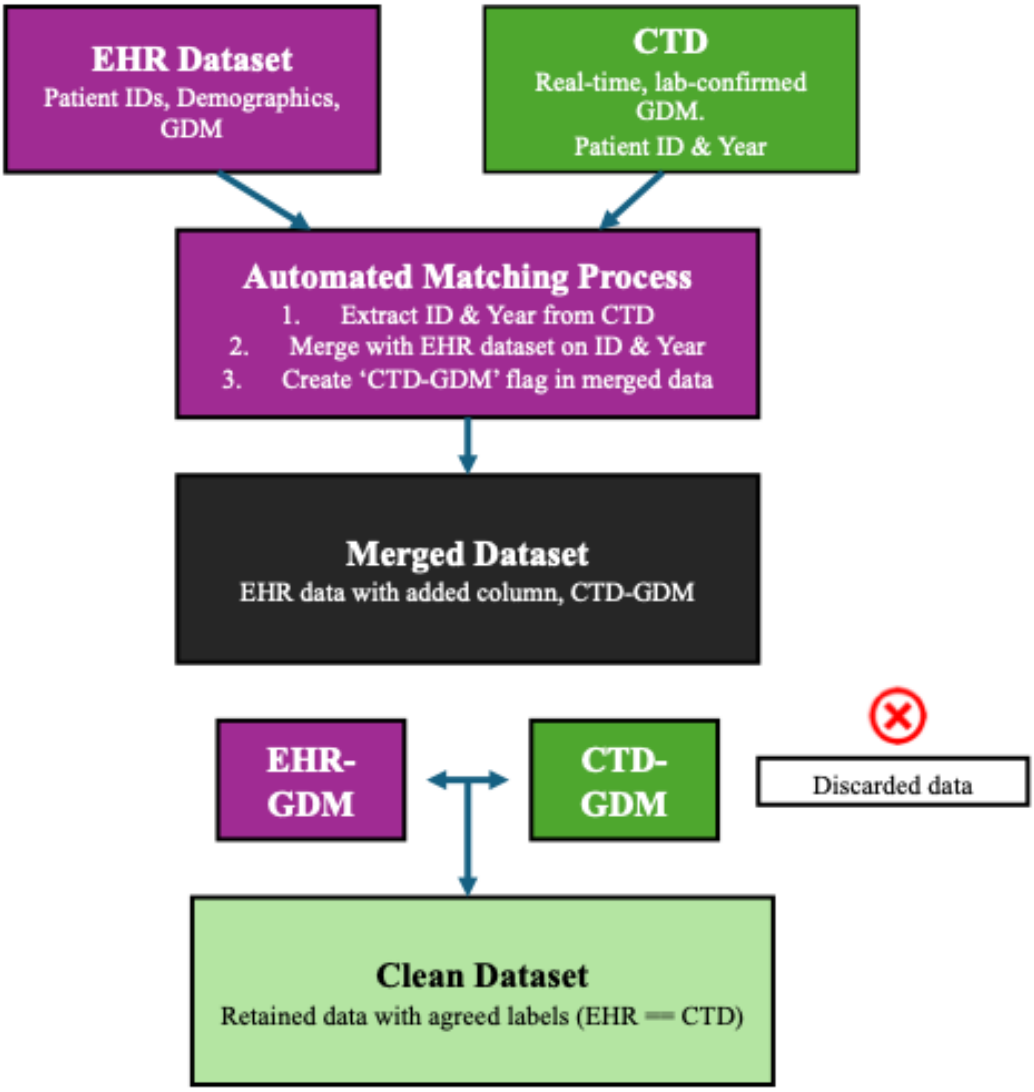


Figure 4.1. Matching process for merging the EHR dataset with the clinical team database (CTD).

4.2.3 Evaluation of Label Noise on ML Modelling

To evaluate the impact of label noise on the performance of prediction models in predicting GDM, I trained a LR model using the validated dataset, where the dataset was split into 70% training and 30% test sets for internal evaluation. Default model hyperparameters were used, as the primary objective was to compare performance across different training datasets rather than optimising hyperparameter settings. The training and testing data comprised of EHR data that was available during the first booking visit, typically the 12th week of gestation, and included 79 training features (Appendix D). The target label was GDM. The dataset contained both categorical and numerical features. Categorical features were processed using OneHotEncoder with the 'first' category dropped, and numerical features were standardized using StandardScaler. While the goal of this paper is not to produce an endpoint

AI model, a self-assessment checklist for reporting was followed to ensure adequate information about the ML model was present[232].

I trained two ML models: one with the EHR-GDM labels and the other with the VAL-GDM labels. Both models were evaluated using the same test set, which used validated VAL-GDM labels, to facilitate a direct comparison of the effects of label noise during training on a consistent test set. The year 2020 was excluded from these analyses due emerging research demonstrating reduced detection of diseases during this period [233], something that I confirm in the results below. By using both the 'raw' and 'validated' datasets, the study aimed to demonstrate the impact of label noise on model performance in the prediction of GDM, providing insights into the importance of accurate label validation in developing reliable predictive models using ML. In addition to the LR model, I replicated this process with other ML models to ensure any effects were not model specific. External validation datasets were sought but this was not successful, as documented[202].

Additionally, varying levels of label noise were introduced to determine the threshold at which label noise significantly affects model performance. This simulation was performed by progressively increasing the number of FPs and FNs in the VAL-GDM training set from 0% to 90% i.e. changing a percentage of the positive labels to negative labels (creating FNs) and changing a percentage of negative labels to positive labels (creating FPs). This approach resulted in the training of 100 different models. Next, in a separate analysis I applied this progressive noise insertion to the VAL-GDM test set to specifically assess the impact of test set label noise on model evaluation i.e. evaluating these test sets using a model trained with the 'clean' VAL-GDM labels. For reproducibility, the code used to perform the label noise simulation is made available in the Data and Resource Availability section.

4.2.4 Statistical Analysis

The validation findings were quantitatively assessed using accuracy, precision, recall, F1 and overall agreement measured by Cohen's Kappa, between the EHR-recorded GDM diagnoses (EHR-GDM) and the clinical team database (CTD-GDM). The performance of the LR ML models were evaluated using Receiver Operating Characteristic Area Under Curve (AUC) and the Average Precision score (AP). Additionally, the calibration of the model's predictions will be examined visually by calibration curves and quantitatively by the slope and intercept. The statistical and ML analysis were performed using Python version 3.8.8 with libraries including NumPy 1.23.5, pandas 1.2.4, and scikit-learn 1.2.1.

4.3 RESULTS

4.3.1 Population Characteristics

The dataset comprised 37,651 EHRs from 31,100 unique patients. The mean±SD patient age was 32±5 years, and body mass index (BMI) was 26.2±5.5 kg/m^2, with 20.7% exhibiting a BMI greater than 30.0 kg/m^2. The prevalence of GDM according to the EHRs was 11.0%, whereas the prevalence according to the CTD was 10.5%. Patient characteristics for the most important features in the machine learning models are presented in Table 1.

Table 4.1. Patient characteristics for the most important features in the machine learning models, according to the validated dataset (N=27,561). The validated dataset represents a dataset where both the EHRs and CTD agree.

Characteristics	**Mean±SD/Prevalence**
Age (years)	32±5
BMI (kg/m^2)	26.2±5.3
Systolic Blood Pressure (mmHg)	111±11
Diastolic Blood Pressure (mmHg)	67±8
Parity	0.9±1.1
Ethnic Origin of Patient	
Caucasian	87.8%
South East Asian	4.9%
Black African	2.0%
Asian	1.8%
Middle Eastern	0.6%
Latin American	0.1%
Mixed	0.1%
Other	3.0%
Occupation Skill Level (ISCO)	
Level 0 (Unemployed)	19.0%
Level 1 (Elementary occupations)	1.3%
Level 2 (Clerical and Service)	15.9%
Level 3 (Technicians & Associates)	8.6%
Level 4 (Professionals and Managers)	55.1%
Family history of diabetes mellitus	23.3%
History of GDM	3.9%
Other Endocrine Problems	21.4%
Prevalence of GDM	11.7%

4.3.2 Diagnosis Discrepancies

Of 3,952 patients with matching IDs in both databases, 3,388 were correctly identified with GDM in both EHR-GDM and CTD-GDM (9.0% TP and 85.7% TPR), while 564 lacked a corresponding GDM label in EHR-GDM (1.5% FN and 14.3% FNR) (Figure 4.2). Additionally, 771 patients were incorrectly identified with GDM in EHR-GDM without matching IDs in CTD-GDM (2.0% FP and 2.3% FPR). In EHRs there were 32,928 (87.5%) TN cases (97.7% TNR) (Figure 4.2). The accuracy, precision, F1 score, and Cohen's kappa are reported in Table 4.2.

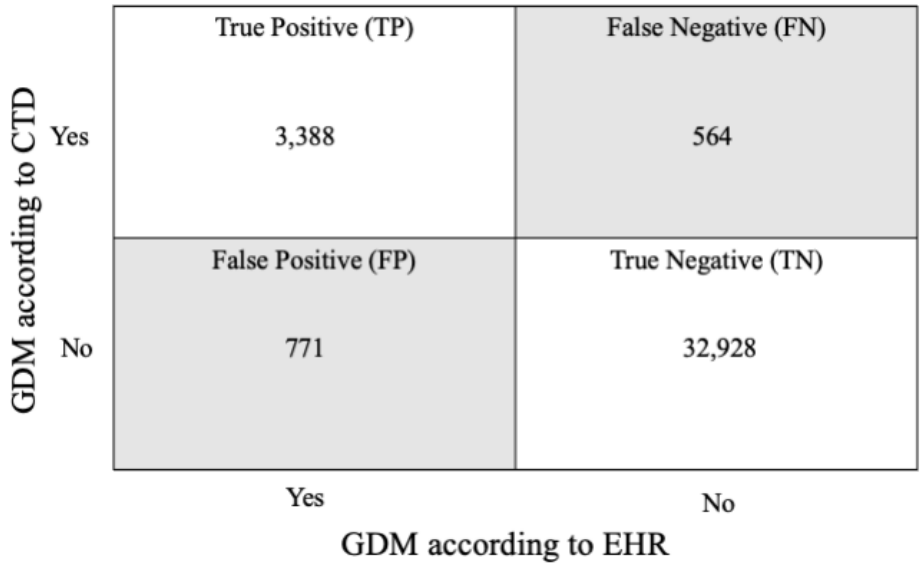


Figure 4.2. This diagram illustrates the numbers of patients classified as true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) based on the comparison of GDM status in EHRs versus the CTD (reference standard) for 2018–2022. The matrix shows 3,388 true positive cases (TP), 564 false negative cases (FN), 771 false positive cases (FP), and 32,928 true negative cases (TN) from a total sample of 37,651 records.

Table 4.2. Performance metrics for the comparison of GDM diagnoses in electronic health records (EHR) with the real-time clinical team database (CTD).

Year	Cohen's Kappa	Accuracy	Precision	Recall	F1 Score
All Years	0.82	0.96	0.81	0.86	0.84
2018	0.80	0.96	0.78	0.86	0.82
2019	0.82	0.96	0.86	0.82	0.84
2020	0.77	0.96	0.70	0.90	0.79

2021	0.86	0.97	0.89	0.87	0.88
2022	0.82	0.97	0.82	0.86	0.84
All minus 2020	0.82	0.97	0.84	0.85	0.84

4.3.3 Yearly Data Comparison

Ninety-eight patients identified in CTD lacked corresponding entries in EHRs. Sixty-seven (68%) of these discrepancies were observed in 2020 (Figure 4.3). Furthermore, GDM prevalence for both EHRs and CTD datasets revealed a notable reduction in 2020 (recorded at 10.0% in EHRs and 7.7% in CTD), indicating a deviation from the trend observed in other years (Figure 4.3). These discrepancies align with Covid related disruptions to screening practices within the hospital between March 2020 and September 2020.

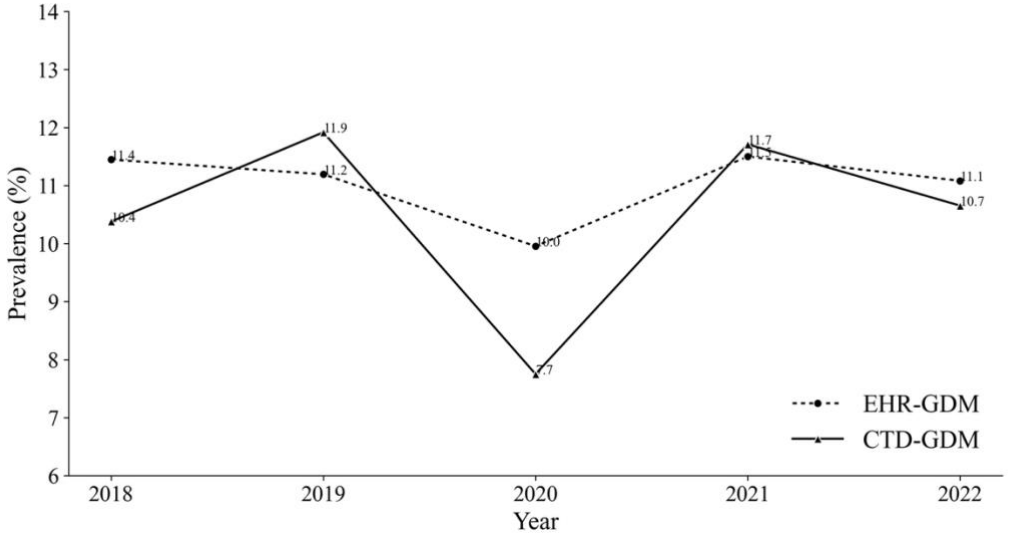


Figure 4.3. Comparison of prevalence rates of GDM diagnosis between electronic health record (EHR-GDM) data and the clinical team database (CTD-GDM) from 2018 to 2022. The solid line represents the CTD data, and the dashed line represents the EHR data.

4.3.4 Label Noise in EHRs

The performance of LR models trained using the raw (EHR-GDM) and validated (VAL-GDM) labels was evaluated using a test set with VAL-GDM labels only. The model trained using the EHR-GDM labels achieved a AUC of 0.817 (95% CI: 0.802–0.833) and an AP score of 0.451. The calibration curve is shown in Figure 4.5, with an intercept 0.12 (95% CI 0.05-0.20) and slope 0.98 (95% CI 0.92-1.05). In comparison, the model trained using the VAL-GDM labels showed a AUC of 0.817 (95% CI: 0.803–0.832) and an AP score of 0.450 (Figure 4.4), indicating a minor impact of label noise in training the model for this dataset

(intercept 0.05, -0.02-0.13; and slope 0.95, 0.90-1.01). However, when the performance of the LR ML model trained using VAL-GDM labels was evaluated on a test set with EHR-GDM labels, a AUC of 0.814 and an AP score of 0.395 was achieved, which demonstrates a greater impact of label noise when it is present in the test set.

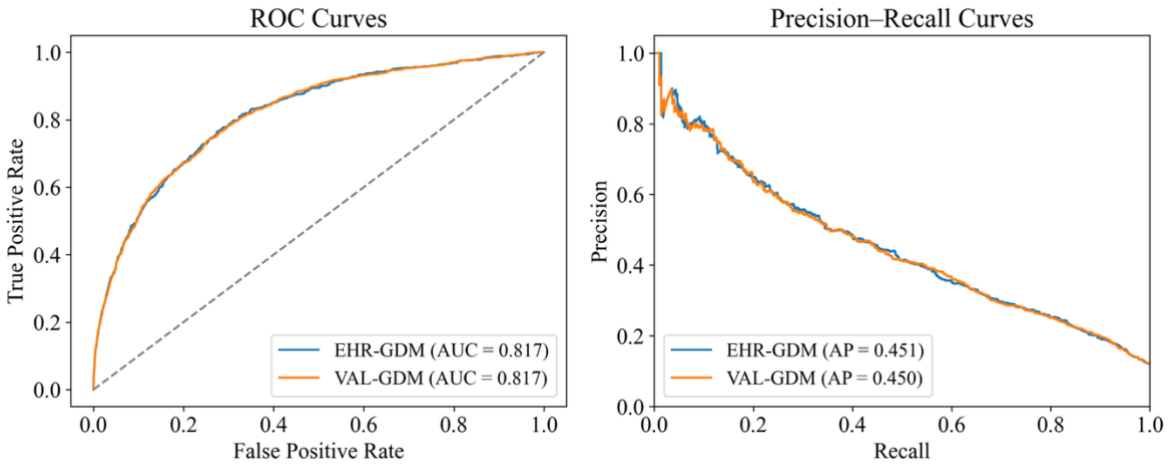


Figure 4.4. (Left) Receiver-operating-characteristic (ROC) curves and (right) precision-recall (PR) curves showing the performance of two logistic-regression models for predicting gestational diabetes mellitus (GDM). "EHR-GDM" refers to the model trained on electronic-health-record GDM labels, and "VAL-GDM" refers to the model trained on the subset of cases where the EHR and clinical team database labels agree. Both models are evaluated against the same reference labels (VAL-GDM).

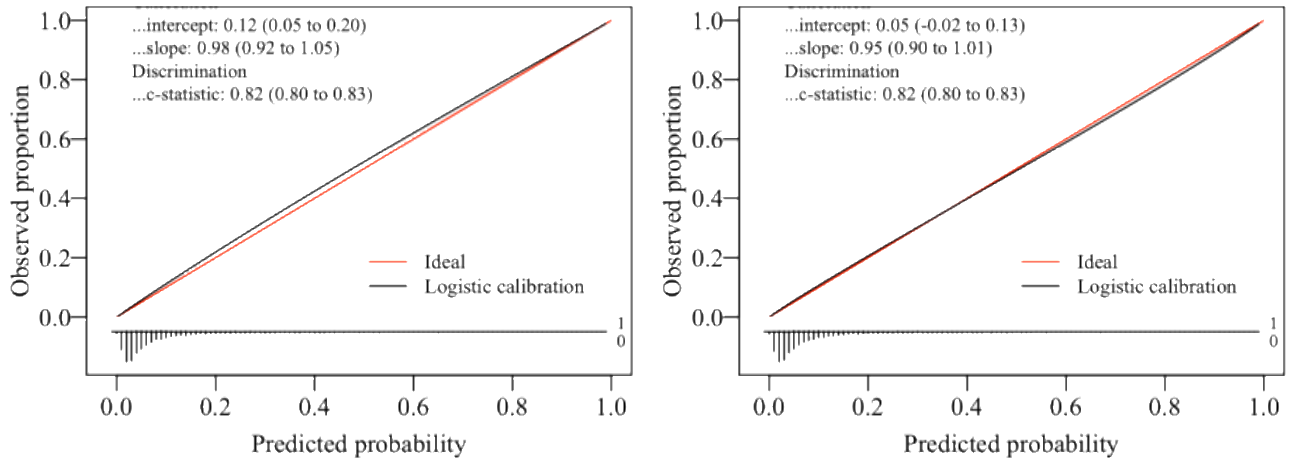


Figure 4.5. (Left) Calibration curve for the EHR-GDM model, which was trained on electronic health record GDM labels. (Right) Calibration curve for the VAL-GDM model, which was trained on the subset of cases where electronic health record and clinical team database labels agree. Both evaluated against the identical standard reference labels.

In addition to LR, Random Forest, XGBoost, and an Explainable Boosting Machine were assessed to compare their performance and robustness to label noise. As shown in Table 4.3, all three models achieved performance metrics in a similar range to the logistic regression

model, with none of the models demonstrating large changes in evaluation metrics regardless of the validation data used.

Table 4.3. Comparison of receiver operating characteristic area under the curve (AUC) and average precision for machine learning models predicting gestational diabetes mellitus (GDM) trained on the EHR data and on VAL data and validated against the VAL data.

Model	AUC		Average Precision		Intercept		Slope	
	EHR-GDM	VAL-GDM	EHR-GDM	VAL-GDM	EHR-GDM	VAL-GDM	EHR-GDM	VAL-GDM
Logistic Regression	0.817	0.817	0.451	0.450	0.093	-0.027	0.984	0.955
Random Forest	0.797	0.801	0.418	0.419	-0.747	-0.618	0.553	0.638
XGBoost	0.780	0.782	0.389	0.393	-0.427	-0.507	0.619	0.608
EBM	0.818	0.816	0.456	0.450	0.078	-0.047	0.975	0.940

4.3.5 Simulated Label Noise

The impact of simulated label noise on model performance was assessed by progressively increasing the number of FNs (False Negative Noise) and FPs (False Positive Noise) in the training set (where 0% noise equates to the original VAL-GDM labels) without modifying the testing set. The results demonstrate a decline in model performance as the level of label noise increases (Figure 4.6).

Further analysis of noise in the test set showed that model performance metrics, particularly AUC and AP scores, were sensitive to increasing levels of noise, especially FP noise. As the FP rate was increased, the AUC consistently decreased, while the AP score initially decreased before increasing. The introduction of FN into the test set had a less pronounced effect on performance compared to FP, unless both types of noise were combined, which led to a more substantial impact (Figure 4.6).

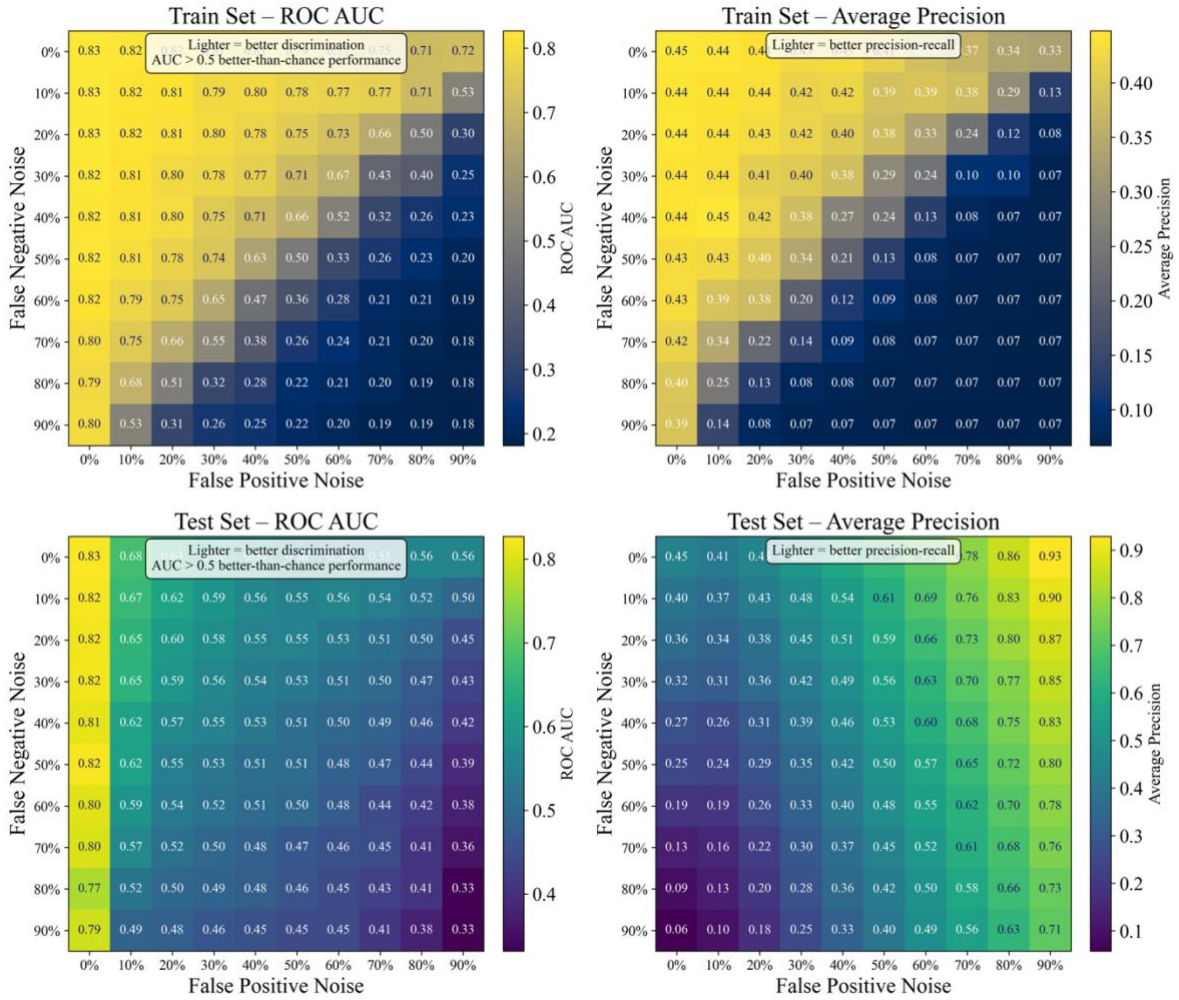


Figure 4.6. Heatmaps illustrating how gradually introducing random label noise (NAR) degrades model performance. In each panel, the x-axis denotes the percentage of true negatives flipped to false positives and the y-axis denotes the percentage of true positives flipped to false negatives. In the top-left heatmap, the ROC AUC on the training set is plotted, lighter cells signify stronger discrimination, and values above 0.5 represent performance better than random chance. The top-right heatmap presents the corresponding average precision on the same noisy training data, with lighter colours indicating a more favourable precision-recall trade-off. The bottom-left and bottom-right heatmaps repeat these experiments on the held-out test set, showing ROC AUC and average precision, respectively, under increasing label noise in the test data. The unusual behaviour of average precision is discussed in the manuscript. Each cell is annotated with the exact metric value for that combination of false-positive and false-negative noise levels.

4.4 DISCUSSION

This study highlights significant discrepancies between GDM diagnoses recorded in EHRs and those validated by the CTD. Correcting label noise in the training set had a negligible impact on the performance of an LR-based ML model developed from EHRs to predict GDM from early pregnancy data. However, correcting label noise in the test set improved the model's

average precision, underscoring the importance of accurate labelling for evaluating model performance accurately. The study also found that increasing label noise in the training set led to a gradual decline in model performance, whilst increasing FPs in the test set had a particularly strong negative impact on AUC, but counterintuitively increased AP scores. FNs had a less pronounced impact on AUC unless combined with FP, which then caused a decline in model performance.

Approximately 14% (564/3,952) of GDM cases were not recorded in the EHRs, while 18.5% (771/4,159) of positive GDM diagnoses in EHRs did not align with CTD records. Overall, there were 32,928 (87.5%) TN, 3,388 (9.0%) TP, 771 (2.0%) FP, and 564 (1.5%) FN. The FPR (2.3%) remained low in comparison to the FNR (14.3%). Similar discrepancies in accuracy of EHRs have been reported in previous studies within Irish maternal hospitals, though with higher agreement in other contexts, such as miscarriage measurements (k=0.92)[234]. More widely across Europe, wide variations exist in the accuracy of reporting in EHRs as it relates to acute cardiovascular outcomes, with sensitivity reported at <66% for heart failure diagnoses in particular[235]. A key challenge in these studies is the absence of a recommended reference standard for validating EHR data, leading to the use of various data sources[227].

The impact of COVID-19 on screening and diagnostic practices, especially in 2020, manifested in a relative reduction of 31% in GDM diagnoses i.e. 11.2% across 2018, 2019, 2021, and 2022 compared to 7.7% in 2020. This observations aligns with research indicating reduced diagnosis rates for various medical conditions during the first year of the pandemic[233], and suggests caution is warranted when utilising EHRs during this year for the purpose of healthcare modelling. The decrease in recorded GDM cases in 2020 was likely driven by changes in clinical protocols at the onset of COVID-19. The Irish Health Service Executive adopted procedures recommended by the Royal College of Obstetrics and Gynaecologists in the UK which recommended alternative testing strategies for screening pregnant women for GDM that focused on replacing the 2-h OGTT with other tests of shorter duration[219].

Correcting label noise has been shown to mitigate its adverse effects on model performance, underscoring the importance of 'clean' and accurate datasets for training and validating ML algorithms to ensure their accuracy in clinical decision support systems[236]. For example, training a model on a 'clean' dataset resulted in an accuracy of 73.6%, whereas with 30% label noise the accuracy fell to 64.1% (-9.5%)[236]. However, the current analysis demonstrated that training a LR model using EHR-GDM labels versus validated VAL-GDM yielded negligible differences in performance metrics, with AUC of 0.817 and 0.817, respectively (Figure 4.4), performance in line with previous research using EHRs to predict

GDM[83,99]. This is presumably due to the low overall representation of FN and FP in the dataset of 3.5% combined, which limited the impact of label noise on the training process.

Previous work has simulated noisy labels with artificial introduction of different levels of label noise (10%, 20%, and 40%) into the training set, and demonstrated a gradual decline in the accuracy of all models (mean AUC of all models at 10%: 81.3, 20%: 80.2, and 40%: 78.4) as label noise increased[237]. The approach taken in the present study differs in that it introduces systematic label noise using Noise at Random (NAR)[238], increasing both the FP and FN rates linearly. Introducing noise into the training set resulted in a gradual decline in model performance, with both AUC and AP scores decreasing as the level of noise increased. The model was particularly sensitive to FP, which caused a more pronounced decline in performance compared to FN. Introducing noise into the test set also impacted model performance, but the effects were more complex. The AUC consistently decreased as FP rates increased, indicating that the model's ability to distinguish between classes was compromised. However, the AP score showed a different pattern, with an initial decline followed by an increase as noise levels were increased. The introduction of FN in the test set had a less pronounced effect on performance compared to FP, unless FP and FN were combined, which led to a more marked decline in the model's overall performance.

The counterintuitive increase in the AP score as the FP rate in the test set increased can be attributed to the method of calculating AP. AP evaluates the precision-recall trade-off across different thresholds, specifically calculating the proportion of TP to the sum of (TP + FP). When the majority of the negative class in the test set is artificially converted to positive, the opportunity for FP to occur is significantly reduced. This reduction in potential FP leads to an increase in precision, which in turn increases the AP score. Additionally, this manipulation dramatically alters the (e.g. class balance from 90% negative to 90% positive), further influencing the precision-recall dynamics and contributing to the observed ostensible increase in AP. In practical terms, this finding emphasises that certain performance metrics like AP can behave unexpectedly in the presence of extensive label noise or class imbalance, underscoring the importance of using multiple evaluation metrics that are robust to changes in classes (discrimination and calibration) to fully understand model performance. These results reinforce that deploying predictive models trained on unvalidated EHR data can amplify false positive and false negative risks.

This study has several limitations which may affect the generalisability of these findings. First, the analysis was conducted using data from a single hospital, and did not perform any external validation with data from other hospitals or a formal temporal validation

using a future period. Therefore, it is uncertain whether the findings would directly generalise to different clinical settings, particularly those with different screening practices, disease encoding and EHR systems. Second, the CTD, which is treated as the ground truth, is manually maintained by the clinical team. While it is likely more accurate than the EHR, it is not immune to possible human errors or omissions. Any such errors in the CTD would affect the data validation results by erroneously labelling some EHR entries as false positives or false negatives. However, this should minimally impact ML modelling as only data that had agreement across both databases were included. Third, there is potential for model overfitting due to the lack of an external validation set and default parameters, which I attempted to mitigate with the use of k-fold cross validation, a relatively simple linear model, and a hold-out test set. Finally, NAR linearly increases the FP and FN rate in the dataset, which may not accurately reflect how errors in EHRs typically occur.

4.5 CONCLUSION

In conclusion, the identified discrepancies in EHR-recorded GDM diagnoses compared to 'true' GDM diagnoses reflect broader concerns about the accuracy of EHRs for public health and ML applications. Further, the magnitude of inaccuracies may play an important role for maximising the utility of EHRs in enhancing healthcare outcomes, particularly for conditions such as GDM. However, when these discrepancies remain a small percentage (e.g. <5%) of the dataset, like in the case of the present study, there was no notable impact on model training performance. Conversely, the risk of incorrect model evaluation increases when the test set labels are impacted by noise, as this has a more pronounced effect on performance metrics. These observations emphasise the importance of incorporating robust data cleaning, preprocessing, and validation methodologies in the development of ML models for healthcare. Future efforts should aim at developing standardised validation protocols for EHRs to ensure high data quality for training and evaluating ML algorithms. Such protocols could include harmonising how GDM diagnoses are recorded across different sites, implementing automated consistency checks (for instance, prompting for a GDM diagnosis entry in the EHR when a laboratory result confirming GDM is received), and performing regular audits comparing EHR records with reference databases or lab results. By improving the integrity of data entry and maintenance in EHR systems, these measures could reduce discrepancies and enhance the utility of EHR data for both clinical care and machine learning applications.

In summary, Chapter 4 demonstrated that cleaning the GDM labels in our dataset yields a tangible improvement in model performance. We deliberately used a straightforward logistic regression model (the same type we will carry forward into Chapter 5) to isolate the effect of label quality. Training this model on noisy labels vs. clean labels made a clear difference: the model trained on validated (clean) data achieved higher accuracy on the test set than the one trained on noisy data. This finding validates the importance of addressing label noise raised in RQ2. The next logical step is to utilise this curated dataset to develop and test the prognostic models themselves. Chapter 5 will now address this task.

Chapter 5

Evaluation of Machine Learning Models for Early Prediction of Gestational Diabetes Using Retrospective Electronic Health Records from Current and Previous Pregnancies

Germaine, M., O'Higgins, A. C., Egan, B., & Healy, G. (2025). Evaluation of Machine Learning Models for Early Prediction of Gestational Diabetes Using Retrospective Electronic Health Records from Current and Previous Pregnancies. *BMJ Digital Health & AI. 2025;1:e000089*.

DOI: 10.1136/bmjdhai-2025-000089

Chapter Overview

This chapter addresses two central research questions: early prediction (*How accurately can ML models, when applied to EHRs, predict the diagnosis of GDM?*) and multiparous prediction (*How accurately can ML models leverage data from a woman's previous pregnancy to predict outcomes in subsequent pregnancies?*). The work details the development and internal validation of several ML models using the datasets curated in Chapter 3.

For the early prediction task, models using only first-trimester data from the current pregnancy were developed and evaluated, resulting in discriminative performance of AUC 0.82 for the main cohort, and 0.81 for the nulliparous women. These models achieved good calibration, with slope ranging from 0.968 (95% CI 0.913-1.028) to 1.062 (95% CI 1.024-1.098) and intercept from -0.054 (95% CI -0.170-0.064) to 0.103 (95% CI 0.027-0.175)Incorporating data from a woman's past pregnancy led to improvements in performance, with multiparous models (that incorporated both current and past pregnancy data) achieving AUCs ~0.88 (slope 1.033; intercept 0.050). Models that used only past pregnancy data, thus enabling preconception predictions, achieved AUCs ~0.86 (slope 0.997; intercept -0.006). These insights help shift the clinical paradigm for early pregnancy prediction towards preconception risk stratification, at least in multiparous populations, aligning with public health calls for earlier intervention.

Importantly, this level of performance was achieved using only non-invasive, routinely collected EHR data. This approach lowers the barrier to clinical implementation and integration, as it does not require new or costly biochemical tests or changes to the current approach to maternal care. This chapter also highlights the need for fairness considerations, as model performance varied across ethnic subgroups, underscoring the importance of evaluating and mitigating potential biases in clinical ML tools.

5.1 INTRODUCTION

Chapter 4 established the integrity of the GDM outcome label within the Coombe EHR dataset. By quantifying the label noise and creating a validated cohort (the VAL-GDM dataset) where EHR and CTD labels agree, a reliable 'ground truth' for model development was secured. With this foundation of data quality confirmed, the logical next step is to proceed with model development itself. This chapter addresses RQ3 (Early Prediction) and RQ4 (Multiparous Analysis) by developing and internally validating a suite of prognostic models using this curated dataset. The primary objective is to determine how accurately GDM can be predicted using only routinely collected, non-invasive data from the first trimester. Furthermore, this chapter will investigate whether leveraging data from a woman's previous pregnancies—a rich source of obstetric history—can significantly improve predictive performance, thereby enabling more precise risk stratification for multiparous women and exploring the potential for pre-conception risk assessment.

A 2018 Lancet series[15–17] underscored that lifestyle modifications initiated early in pregnancy could influence maternal and neonatal outcomes. However, the efficacy of such lifestyle interventions remains unclear, with several strategies offering limited benefits in terms of dietary[38,39] and physical activity interventions[40–42], possibly due to the intervention's commencement time, which is often during late stage pregnancy. Two meta-analyses[43,44] on lifestyle interventions during pregnancy highlight this point: the timing of intervention is key. Lifestyle changes initiated during the first trimester were most effective in reducing GDM risk and improving maternal health. Furthermore, early interventions are expected to provide significant cost-saving benefits for healthcare systems by reducing complications associated with GDM[46].

Machine learning (ML) offers a promising solution for the early identification of women at risk of GDM[99], addressing a key challenge of late stage (24-28 week) diagnosis. By leveraging large datasets, such as EHRs, ML creates predictive models that can identify women at increased risk for GDM before the traditional screening window[189]. By identifying high-risk individuals at antenatal visit, or preconception [26], ML models could facilitate earlier lifestyle and clinical interventions. Such an approach could optimise the timing of interventions, potentially enhancing delivery outcomes and improving the health of both mother and child[239,240].

Despite the growing evidence on ML models for GDM prediction[83], most studies using EHRs to date have been conducted in Asian populations[202], often yielding promising but

population-specific results. Furthermore, most have primarily utilised data from the current pregnancy, with limited incorporation of past pregnancy information. One study incorporated data from a previous pregnancy to predict GDM, but it involved only East Asian women with prior GDM (n=553) and also used a current first trimester glucose measurement[114]. These constraints underscore the need for broader examinations of whether including previous pregnancy data can enhance early GDM prediction across diverse populations, including those in Europe, and whether such models can generalise beyond women already known to be at high risk. One rationale for analysing a past pregnancy model is to explore a potential preconception risk tool. If a woman has had a prior pregnancy, her data from that pregnancy could stratify her risk even before conceiving again. This aligns with a preventative strategy: identifying high-risk individuals in interpregnancy intervals for targeted interventions.

Therefore, the aim of this study was to develop and evaluate the performance of ML models in predicting GDM using EHR data collected in the first trimester and to determine whether incorporating data from previous pregnancies could improve predictive performance. By including data from past pregnancies, the aim was to evaluate the potential for predicting GDM risk even before conception, thereby allowing for preconception risk assessment. Additionally, by developing models tailored to both nulliparous and multiparous populations, the study explored potential differences in performance while keeping in mind future applications in clinical decision support systems (CDSS)[53]. Finally, this study also investigated using a reduced set of clinically relevant features, recognising that limiting the feature set (variables) to be collected could enhance the practicality of model deployment by clinicians in a healthcare facility.

5.2 METHODS AND ANALYSIS

5.2.1 Study Design and Population

This retrospective cohort study employed ML techniques on EHRs from The Coombe Hospital, Dublin, spanning a five-year period (2018–2022). Ethical approval was granted by The Coombe Hospital Research Ethics Committee (Study No. 06–2023). The primary aim was to develop ML models that could predict the diagnosis of GDM by using EHR data collected during the first antenatal visit (~12th week of gestation). Three distinct cohorts were used to build the ML models (see Study Populations). The study was designed following the TRIPOD+AI guidelines for reporting clinical prediction models [101]. Patients or members of the

public were not involved in the design, conduct, reporting, or dissemination plans of this study due to its retrospective nature.

5.2.2 Inclusion and Exclusion Criteria

Eligible records were from women were 18 years or older, attended an antenatal visit at or before 16 weeks, with complete EHR data up to 16 weeks, and validated GDM status. Exclusions included women with pre-existing diabetes (type 1 or 2), first antenatal visit after 16 weeks, those with missing or incomplete data for critical variables, invalidated GDM status, and pregnancies from 2020 (see below).

5.2.3 Study Populations

Three distinct study populations were defined. The First-Trimester population included all eligible pregnancies with data up to the 16th week of gestation. The Nulliparous population, a subset of the general population, focused on women experiencing their first pregnancy. Finally, the Multiparous population identified women with at least one previous pregnancy in the dataset, allowing for the inclusion of historical data to predict GDM in subsequent pregnancies and modelling based on past pregnancy alone.

5.2.4 Data Source and Validation

Data were routinely collected by trained midwives using standardized questionnaires and entered into the hospital's electronic health system, "Euroking K2" (Euroking Maternity Software Systems, UK)[241]. GDM was diagnosed according the IADPSG criteria[1]. EHR entries indicating “Diabetes developed during pregnancy” were labelled as GDM, while all others were labelled non-GDM. GDM status was cross-referenced with a real-time clinical database (2018–2022)[155]. Records were included only if both databases provided consistent labels. Data from 2020 were excluded due to COVID-19 related fluctuations in GDM diagnosis[155,242]. The final dataset included 27,561 pregnancies and 108 retained variables. This sample size was deemed sufficient based on the BMJ guidelines for calculating the required sample size for developing a clinical prediction model[200]. Individual prediction uncertainty was further assessed by calculating effective sample size[243].

5.2.5 Data Cleaning and Preprocessing

Data quality was ensured through extensive cleaning and preprocessing. Critical variables included maternal age, body mass index (BMI) at booking, key medical history items (e.g., family history of diabetes, endocrine problems), and obstetric history (e.g., previous GDM, parity). Pregnancy records missing any critical variables were excluded, and features with more than 20% missing data were removed unless the missing value could be interpreted as a null finding. Inconsistencies (e.g., negative BMI) were either corrected if plausible or excluded if not. Categorical variables were simplified; for instance, the "Cardiac Problems" feature, initially containing over 2,600 string values, was simplified into a binary "NO"/"YES," with nulls interpreted as "NO" after clinician consultation. Any remaining missing data were then removed.

Two deviations from this general approach were the "Occupation of the Mother" mapped to the International Standard Classification of Occupations (ISCO)[213] and grouped into skill level categories, with 4 representing the highest skill level and 0 representing unemployment. Second, birthweight percentiles were calculated using the weight of the baby and the gestational age at delivery, based on the percentile ranges presented by Nicolaides et al[214].

For multiparous pregnancies, two additional features, Inter-Pregnancy Interval (days from the delivery date of one pregnancy to the conception date of the next) and Inter-Pregnancy Weight Gain (difference in maternal booking weight between successive pregnancies), were calculated. Clinically relevant features were selected by modelling the full EHR, retaining the 15 highest-importance variables, and applying backward elimination until cross-validated performance fell; we also dropped ISCO skill level given its limited availability to maximise external validity.

5.2.6 Model Development

Four ML models were developed to address clinically relevant different aspects of the data: (1) First-trimester models with data up to week 16th of gestation (routinely collected 11-13 weeks' gestation), (2) Nulliparous pregnancy models using first trimester data, (3) Multiparous pregnancy models that incorporated previous pregnancy variables to predict GDM in subsequent pregnancies, (4) Past pregnancy models that use data only available from the

previous pregnancy to predict future GDM. Model cohort information is presented in Table 5.1 and a list of all variables in the EHR are reported in the Appendix D. The chosen ML algorithms were Random Forest Classifier (RF), Logistic Regression (LR), XGBoost Classifier (XGB), and Explainable Boosting Machine (EBM). LR has historically been popular in GDM modelling [83], but recent evidence suggests that advanced models like RF and XGB yield better results[97,98]. EBMs have been shown to match the performance of these advanced models while retaining interpretability[244]. A Dummy Classifier from scikit-learn was included as a baseline to establish a minimal threshold for performance.

Table 5.1. Model development sets for GDM prediction models.

Model Development Set	No. Pregnancies	Predictors	GDM %
First-Trimester Population	27,561	N=9 Ethnicity, Family History of Diabetes (binary), Previous History of GDM (binary), Endocrine Problems (PCOS/thyroid conditions, binary), Parity (integer), Age (integer), BMI (float), Blood Pressure (float).	11.6
Nulliparous Population	11,623	N=7 Ethnicity, Family History of Diabetes (binary), Endocrine Problems (PCOS/thyroid conditions, binary), Age (integer), BMI (float), Blood Pressure (float).	9.9
Multiparous Population	4,005	N=8 Ethnicity, Age (integer), Interpregnancy weight change (float), History of GDM (binary), BMI (float), Time between pregnancies (integer), Birthweight Percentile (float), Family history of diabetes (binary)	12.2
Past Pregnancy	4,005	N=6 Ethnicity, Age (integer), History of GDM (binary), BMI (integer), Birthweight Percentile (float), Family History of Diabetes (binary)	12.2

5.2.7 Data Preprocessing

A pipeline was developed for data preprocessing and modelling. Categorical variables underwent one-hot encoding with the first category dropped, and numerical variables were standardized using StandardScaler. Scikit-learn's ColumnTransformer ensured separate treatment of categorical and numerical data, and the processed DataFrame retained patient identifiers to facilitate group-based splitting. To avoid data leakage, the dataset was split into training, validation, and test sets using GroupShuffleSplit, ensuring all records from the same

patient were allocated to the same split. An 80–10-10 initial split created a training set (80%) and a validation (10%) and hold-out (10%) set. This approach ensured patient-specific splits, providing separate data for training, hyperparameter tuning, and model evaluation.

5.2.8 Model Training

Hyperparameter tuning using RandomizedSearchCV with a stratified k-fold cross-validation strategy was used for each model. This approach involved a randomized search over specified parameter ranges (Table 5.2) with 10 iterations, evaluating model performance on multiple splits of the training set. Internal validation occurred at three stages: (1) 5-fold cross-validation (CV) during RandomizedSearchCV, (2) additional hyperparamter tuning based on the 10% validation set, and (3) final evaluation on an out-of-fold CV on the entire dataset. External validation was attempted but I was unable to source an external dataset[202]. The final hyperparameter space for each model is reported in Table 5.2.

Table 5.2. Hyperparameter space searched for the GDM models during training

Hyperparameter Search Space (n_iter = 10)	
Random Forest Classifier	• n_estimators: [100, 200, 300] • max_depth: [None, 10, 20, 30] • min_samples_split: [2, 5, 10] • min_samples_leaf: [1, 2, 4]
Logistic Regression	• C: np.logspace(-4, 4, 20) • solver: ['liblinear', 'lbfgs']
XGB Classifier	• n_estimators: [100, 200, 300] • learning_rate: [0.01, 0.1, 0.2] • max_depth: [3, 4, 5, 10]
Explainable Boosting Classifier	• learning_rate: [0.01, 0.1, 1] • max_leaves: [3, 10, 20]

5.2.9 Model Performance Evaluation

The performance of the ML models were primarily evaluated, on the test set, using the Area Under the Receiver Operating Characteristic Curve (AUC) to measure discrimination with bootstrapping (1,000 iterations) employed to compute 95% confidence intervals for the

AUC, each created by sampling, with replacement, the same number of observations as the original test set[245]. Model calibration was assessed both qualitatively with the use of calibration plots, and quantitatively, with the use of slope, intercept and observed to expected ratio (O:E ratio), as has been recommended for clinical prediction models [77]. AUC measures the model's ability to distinguish between classes without being sensitive to class imbalances[75], while calibration assesses the agreement between the observed and predicted outcomes. Models were further evaluated for net clinical benefit with decision-curve analyses[205,206]. Model performance was also assessed using the average precision score (AP), confusion matrices, model sensitivity, specificity and F-1 score with a default threshold of 0.5 for determining the binary classification from the classifier output, while Brier score was calculated to measure agreement between predicted probabilities and observed outcomes across all models[78]. Final performance was obtained with a 5-fold stratified outer CV on the entire data set, producing out-of-fold probabilities for every pregnancy.

Further, to determine whether the models' predictive capabilities generalised across diverse patient backgrounds, performance metrics were also stratified by ethnicity. The AUCs were analysed with a linear mixed-effects model in which Ethnicity was a fixed factor and each dataset × classifier row was entered as a random intercept, thereby treating the six ethnicity-specific AUCs produced by the same model as repeated measures. Degrees of freedom and p-values were obtained with the Kenward–Roger (Satterthwaite) approximation, and post-hoc Sidak-adjusted contrasts tested every ethnicity against the Caucasian reference group.

5.2.10 Addressing Class Imbalance

The dataset exhibited class imbalance, with a GDM rate of 11.6%. Pilot experiments with the Synthetic Minority Over-sampling Technique did not yield improvements, and adjusting class weights resulted in a bias toward higher false positives compared to false negatives. This trade-off was considered undesirable for clinical application. Therefore, no further techniques were applied.

5.2.11 Feature Importance and Interpretability

Feature importance was determined after each model had been fully trained on the training set, ensuring that no information from the test set influenced model parameters. Relative importance scores from RF (feature_importances_) and coefficient-derived odds

ratios from LR provided basic importance measures. In XGB, SHapley Additive exPlanations (SHAP) summarised each feature's contribution to predictions, while EBM's built-in interpretability tools offered global and per-feature visualisations (interpret package). Although SHAP values were computed on the test data for final interpretability assessments, all feature importance metrics reflected the model parameters learned from the training set only.

5.2.12 Software and Computational Resources

Model development were conducted in Python 3.8.8, with pandas (1.2.4) for data manipulation, numpy (1.23.5), scikit-learn (1.2.1) for ML workflows, seaborn (0.11.1) and matplotlib (3.3.4) for visualisation, xgboost (1.7.6), shap (0.41.0), and interpretML (0.3.0). Computations were carried out on an Apple M1 workstation with 16 GB RAM. Training times ranged from several minutes models (e.g., LR) to longer durations for EBM. Linear mixed-effects model were performed in R version 4.5.0 (R Core Team, 2025) running inside RStudio 2025.05.0 (build 496) (installer file RStudio-2025.05.0-496.dmg, Posit Software, Boston, MA).

5.3 RESULTS

5.3.1 Population Characteristics

Initially 37,651 pregnancy records were identified, of which 10,090 were excluded (2,696 for missing/incomplete/invalid data and 7,394 from 2020), resulting in N = 27,561 for analysis. Table 5.3 presents the baseline characteristics of the study participants across these populations, including key demographics, medical history, and clinical measurements. Further details of baseline characteristics by ethnicity are provided in the Table 5.3.

Table 5.3: Patient characteristics of the validated dataset for GDM prediction.

Characteristics	N=27,561 (%)	Rate of GDM (%)
Ethnicity		
Caucasian	24,180 (87.8)	10.0
Black African	554 (2.0)	14.3
Southeast Asian	1,360 (4.9)	33.9
Other	824 (3.0)	13.7
Asian	489 (1.8)	21.7
Middle Eastern	154 (0.6)	13.6
Family history of Diabetes		

No	21,154 (76.7)	8.4
Yes	6,407 (23.3)	21.9
Previous History of GDM		
No	26,483 (96.1)	9.5
Yes	1,078 (3.9)	62.5
Other Endocrine Problems		
No	21,707 (78.8)	8.5
Yes	5,854 (21.2)	22.8
Current GDM		
No	24,373 (88.4)	0
Yes	3,188 (11.6)	100
Parity		
0	11,377 (41.3)	9.9
1	9,875 (35.8)	11.4
>=2	6,309 (22.9)	15.0
Age		
Age (Mean ± SD)	32 ± 5	
>=40	2,178 (7.9)	17.2
<40	25,383 (92.1)	11.1
BMI		
BMI (Mean ± SD)	26.2 ± 5.3	
<25	13,695 (49.7)	4.9
25 to <30	8,423 (30.6)	11.6
30 to <35	3,537 (12.8)	24.2
35 to <40	1,280 (4.6)	33.0
>=40	626 (2.3)	41.9
Blood Pressure		
Systolic BP (Mean ± SD)	111 ± 11	
Diastolic BP (Mean ± SD)	67 ± 8	

5.3.2 Discrimination and Calibration of GDM Prediction Models

With nine routinely recorded predictors, the first-trimester logistic-regression model attained an AUC of 0.819 (95% CI 0.811-0.827), calibration slope 1.010, intercept 0.013, and average precision (AP) 0.441. Random forest, XGBoost and EBM showed virtually identical discrimination (AUC 0.817-0.818) and good calibration (slopes 0.988-1.062, intercepts -0.017 to 0.103).

Among the nulliparous pregnancies, EBM achieved the highest AUC (0.814, 0.799-0.827), with slope 1.004, intercept 0.017 and AP 0.342; LR, XGB and RF ranged from 0.805

to 0.813. Brier scores were 0.076-0.077 and specificity exceeded 0.98 for all models, although sensitivity remained modest (0.05-0.11).

In multiparous women, adding previous-pregnancy variables increased discrimination: EBM reached an AUC of 0.885 (0.867-0.900), slope 0.994 and intercept 0.001, with an AP of 0.620, sensitivity 0.464 and Brier score 0.068. RF, LR and XGB followed closely (AUC 0.874-0.878). Calibration intercepts stayed within ±0.10 and O:E ratios approximated 1.0.
Using past pregnancy features alone, EBM maintained good performance (AUC 0.860, 0.839-0.879; slope 1.028; intercept 0.051; AP 0.556) and RF, LR and XGB yielded AUC values between 0.854 and 0.858. Probability calibration remained acceptable (slopes 0.935-1.028, intercepts -0.094 to 0.051).

The full performance metrics are reported in Table 5.5. Across all settings, Brier scores clustered between 0.068 and 0.083, indicating good calibration, with predicted probabilities closely aligning with observed outcomes. AUC curves for the logistic-regression and XGBoost models are shown in Figure 5.1A and 5.1B, the corresponding precision–recall curves in Figure 5.1C and 5.1D, and the calibration plots in Figure 5.1E and 5.1F.

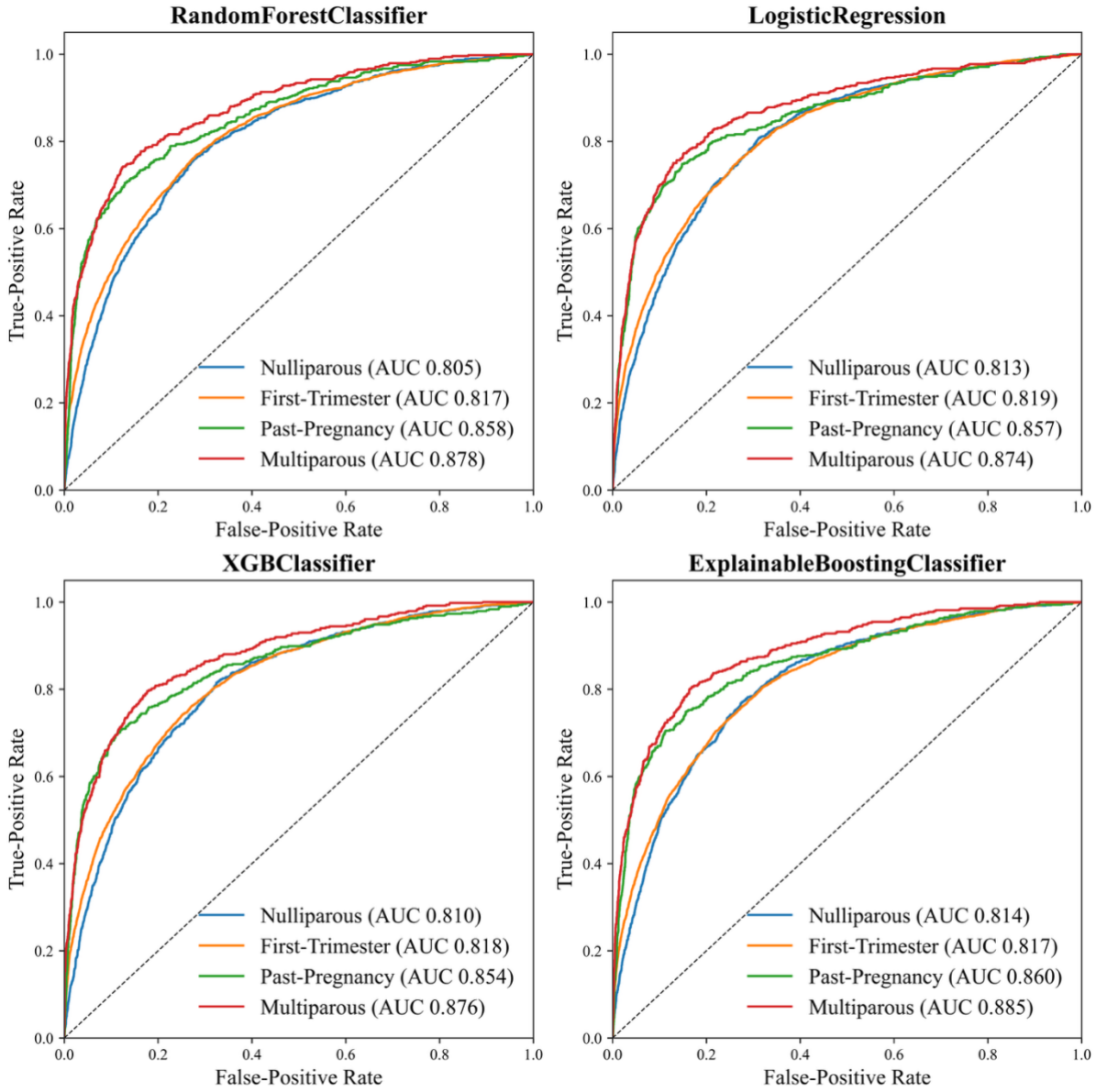


Figure 5.1. Model discrimination (AUC) for RandomForest (top left), LogisticReression (top right), XGBoost (bottom left) and ExplainableBoosting (bottom right). AUCs are reported for each model population: Nulliparous, First-Trimester, Past-Pregnancy and Multiparous.

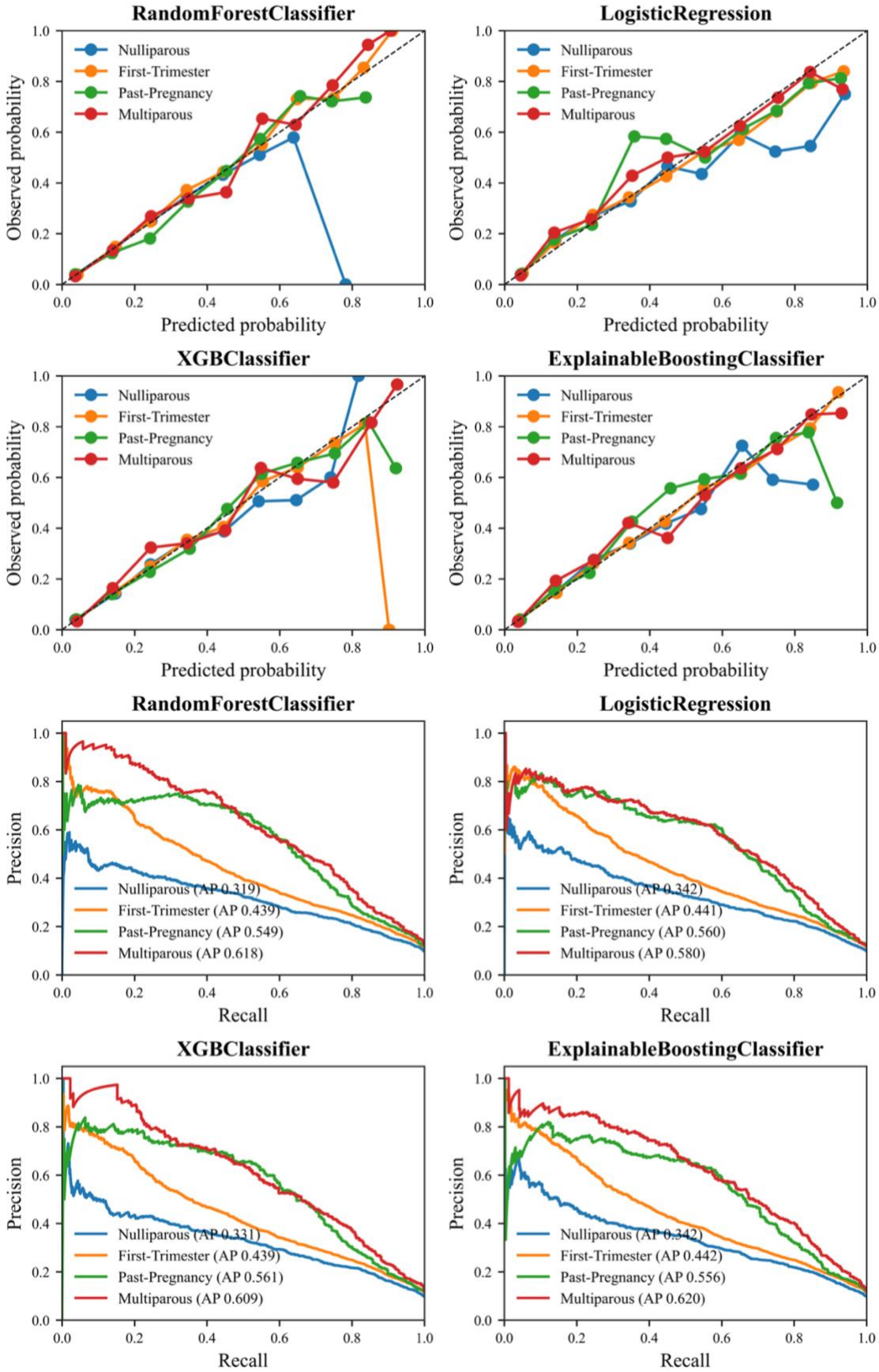


Figure 5.2. Model performance and calibration for Logistic Regression (left panel) and XGBoost (right panel). Subplots A, C, and E show the AUC, average precision, and calibration plots, respectively, for

the Logistic Regression models, while subplots B, D, and F show the corresponding plots for the XGBoost models.

5.3.3 Net Clinical Benefit

Decision-curve analysis (Figure 5.3) was used to translate statistical performance into clinical utility following recommendations[77,205,206]. Net benefit peaks around the lower thresholds (~0.10 at 5%) and tapers as the threshold rises, reflecting the usual trade-off between missed cases and false positives. Curves for RF, LR, XGB and EBM almost overlap, indicating negligible differences in clinical utility once discrimination and calibration are comparable. The multiparous models, boosted by previous-pregnancy information, achieve the highest net benefit, whereas the nulliparous curves approach the treat-none line sooner, mirroring their lower sensitivity. Crucially, none of the model curves fall below the treat-none baseline at thresholds <50%, suggesting that use of these models would not introduce net clinical harm. Sensitivity and specificity (with PPV/NPV, FPR, counts per-100 and net benefit) for each model at decision thresholds 0.10–0.50 across all cohorts, are reported in Table 5.4.

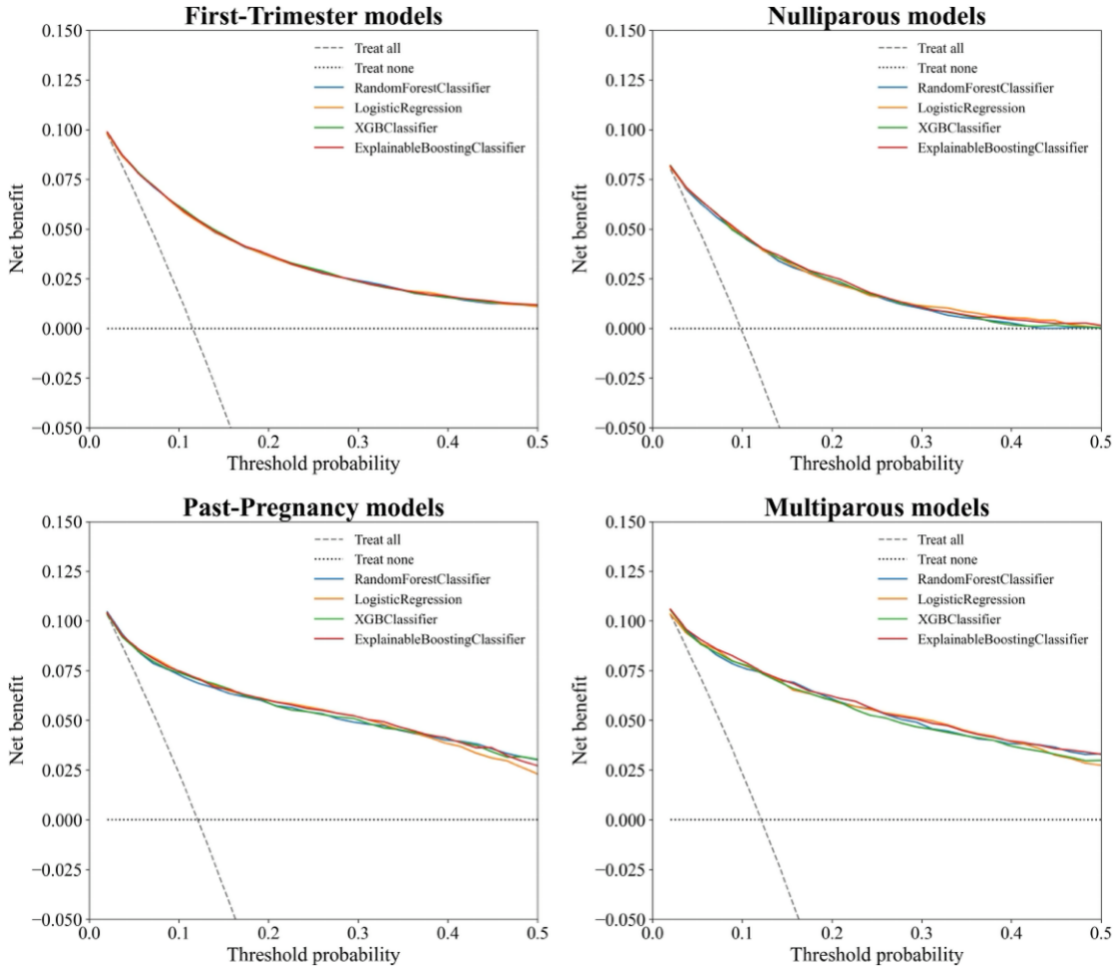


Figure 5.3. Decision-curve analysis for the models applied across the four cohorts.

Table 5.4. Threshold-specific performance and clinical trade-offs for candidate models across cohorts

Cohort & Model	Threshold	Sensitivity	Specificity	FPR	PPV	NPV	TP/100	FP/100
First-Trimester								
Explainable Boosting Machine	0.1	0.77	0.72	0.28	0.26	0.96	9	25
	0.2	0.55	0.88	0.12	0.38	0.94	6	10
	0.3	0.37	0.95	0.05	0.49	0.92	4	5
	0.4	0.26	0.98	0.02	0.59	0.91	3	2
	0.5	0.20	0.99	0.01	0.67	0.90	2	1
Logistic Regression	0.1	0.74	0.74	0.26	0.27	0.96	9	23
	0.2	0.50	0.90	0.10	0.41	0.93	6	8
	0.3	0.36	0.95	0.05	0.50	0.92	4	4
	0.4	0.27	0.97	0.03	0.58	0.91	3	2
	0.5	0.22	0.98	0.02	0.64	0.91	3	1
Random Forest Classifier	0.1	0.78	0.70	0.30	0.26	0.96	9	26
	0.2	0.55	0.88	0.12	0.37	0.94	6	11
	0.3	0.36	0.95	0.05	0.50	0.92	4	4
	0.4	0.24	0.98	0.02	0.61	0.91	3	2
	0.5	0.18	0.99	0.01	0.69	0.90	2	1
XGBoost Classifier	0.1	0.77	0.71	0.29	0.26	0.96	9	25
	0.2	0.55	0.88	0.12	0.37	0.94	6	11
	0.3	0.38	0.95	0.05	0.48	0.92	4	5
	0.4	0.25	0.98	0.02	0.59	0.91	3	2
	0.5	0.19	0.99	0.01	0.68	0.90	2	1
Nulliparous								
Explainable Boosting Machine	0.1	0.74	0.75	0.25	0.24	0.96	7	22
	0.2	0.51	0.89	0.11	0.34	0.94	5	10
	0.3	0.31	0.95	0.05	0.40	0.93	3	5
	0.4	0.16	0.98	0.02	0.48	0.91	2	2
	0.5	0.08	0.99	0.01	0.55	0.91	1	1
Logistic Regression	0.1	0.71	0.77	0.23	0.25	0.96	7	21
	0.2	0.44	0.91	0.09	0.35	0.94	4	8
	0.3	0.29	0.96	0.04	0.42	0.92	3	4
	0.4	0.18	0.98	0.02	0.49	0.92	2	2
	0.5	0.11	0.99	0.01	0.51	0.91	1	1
Random Forest Classifier	0.1	0.74	0.74	0.26	0.23	0.96	7	24
	0.2	0.51	0.88	0.12	0.32	0.94	5	11
	0.3	0.31	0.95	0.05	0.39	0.93	3	5
	0.4	0.14	0.98	0.02	0.46	0.91	1	2
	0.5	0.05	0.99	0.01	0.51	0.91	1	0
XGBoost Classifier	0.1	0.72	0.75	0.25	0.24	0.96	7	23
	0.2	0.51	0.89	0.11	0.33	0.94	5	10
	0.3	0.33	0.94	0.06	0.39	0.93	3	5
	0.4	0.16	0.98	0.02	0.43	0.91	2	2

	0.5	0.07	0.99	0.01	0.52	0.91	1	1
Multiparous								
Explainable Boosting Machine	0.1	0.80	0.83	0.17	0.40	0.97	10	15
	0.2	0.64	0.93	0.07	0.54	0.95	8	7
	0.3	0.57	0.95	0.05	0.62	0.94	7	4
	0.4	0.51	0.96	0.04	0.65	0.93	6	3
	0.5	0.46	0.97	0.03	0.71	0.93	6	2
Logistic Regression	0.1	0.77	0.85	0.15	0.41	0.96	9	13
	0.2	0.61	0.94	0.06	0.57	0.95	7	6
	0.3	0.57	0.95	0.05	0.61	0.94	7	4
	0.4	0.53	0.96	0.04	0.64	0.94	6	4
	0.5	0.44	0.97	0.03	0.67	0.93	5	3
Random Forest Classifier	0.1	0.80	0.79	0.21	0.35	0.97	10	18
	0.2	0.68	0.90	0.10	0.49	0.95	8	9
	0.3	0.58	0.94	0.06	0.57	0.94	7	5
	0.4	0.49	0.96	0.04	0.65	0.93	6	3
	0.5	0.43	0.98	0.02	0.73	0.93	5	2
XGBoost Classifier	0.1	0.80	0.82	0.18	0.37	0.97	10	16
	0.2	0.66	0.91	0.09	0.51	0.95	8	8
	0.3	0.54	0.95	0.05	0.58	0.94	7	5
	0.4	0.48	0.96	0.04	0.65	0.93	6	3
	0.5	0.42	0.98	0.02	0.71	0.92	5	2
Past-Pregnancy								
Explainable Boosting Machine	0.1	0.75	0.83	0.17	0.38	0.96	9	15
	0.2	0.62	0.93	0.07	0.56	0.95	7	6
	0.3	0.58	0.95	0.05	0.62	0.94	7	4
	0.4	0.53	0.96	0.04	0.65	0.94	6	4
	0.5	0.44	0.97	0.03	0.67	0.93	5	3
Logistic Regression	0.1	0.73	0.86	0.14	0.42	0.96	9	12
	0.2	0.61	0.94	0.06	0.57	0.95	7	6
	0.3	0.58	0.95	0.05	0.62	0.94	7	4
	0.4	0.52	0.96	0.04	0.63	0.94	6	4
	0.5	0.44	0.97	0.03	0.64	0.93	5	3
Random Forest Classifier	0.1	0.76	0.80	0.20	0.34	0.96	9	18
	0.2	0.65	0.91	0.09	0.49	0.95	8	8
	0.3	0.59	0.94	0.06	0.58	0.94	7	5
	0.4	0.53	0.96	0.04	0.64	0.94	6	4
	0.5	0.45	0.97	0.03	0.69	0.93	5	2
XGBoost Classifier	0.1	0.74	0.83	0.17	0.37	0.96	9	15
	0.2	0.63	0.92	0.08	0.53	0.95	8	7
	0.3	0.58	0.95	0.05	0.60	0.94	7	5
	0.4	0.53	0.96	0.04	0.65	0.94	6	4
	0.5	0.45	0.97	0.03	0.69	0.93	5	2

Table 5.5. Comparative Performance Metrics of GDM Machine Learning Models Evaluated on the Test Set Across Various Datasets. AUC (Area Under the Receiver Operating Characteristic Curve), AP (Average Precision Score), O:E Ratio (Observed to Expected Ratio).

Dataset and Model	AUC (95% CI)	Calibration Slope	Calibration Intercept	O:E Ratio	AP	Sensitivity	Specificity	F1 Score	Brier Score
First-Trimester Models									
Random Forest Classifier	0.817 (0.808-0.825)	1.062 (1.024-1.098)	0.103 (0.027-0.175)	1.001	0.439	0.184	0.989	0.291	0.082
Logistic Regression	0.819 (0.811-0.827)	1.010 (0.974-1.046)	0.013 (-0.067-0.090)	0.997	0.441	0.217	0.984	0.324	0.082
XGBoost Classifier	0.818 (0.810–0.826)	1.004 (0.968-1.039)	0.007 (-0.067-0.077)	1.000	0.439	0.193	0.988	0.301	0.082
Explainable Boosting Machine	0.817 (0.809–0.826)	0.988 (0.952-1.022)	-0.017 (-0.093-0.054)	1.002	0.442	0.199	0.987	0.307	0.082
Dummy Classifier	0.497 (0.484–0.511)	-0.001	-1.993		0.122	0.107	0.887	0.111	0.209
Nulliparous Models									
Random Forest Classifier	0.805 (0.791-0.819)	0.974 (0.917-1.038)	-0.042 (-0.160-0.080)	1.002	0.319	0.053	0.995	0.095	0.077
Logistic Regression	0.813 (0.799-0.827)	0.995 (0.937-1.055)	-0.008 (-0.138-0.131)	1.001	0.342	0.112	0.988	0.184	0.076
XGBoost Classifier	0.810 (0.795-0.823)	0.968 (0.913-1.028)	-0.054 (-0.170-0.064)	1.001	0.331	0.066	0.993	0.117	0.076
Explainable Boosting Machine	0.814 (0.799-0.827)	1.004 (0.946-1.063)	0.017 (-0.104-0.142)	1.008	0.342	0.085	0.992	0.147	0.076
Dummy Classifier	0.503	0.001	-2.196		0.098	0.108	0.899	0.106	0.178
Multiparous Models									
Random Forest Classifier	0.878 (0.859–0.894)	1.073 (0.999-1.160)	0.099 (-0.054-0.267)	0.997	0.618	0.431	0.978	0.543	0.069
Logistic Regression	0.874 (0.855–0.891)	1.030 (0.958-1.111)	0.048 (-0.127-0.237)	1.001	0.580	0.441	0.970	0.533	0.071
XGBoost Classifier	0.876 (0.857–0.893)	1.031 (0.960-1.114)	0.012 (-0.154-0.194)	0.980	0.609	0.423	0.976	0.529	0.070
Explainable Boosting Machine	0.885	0.994	0.000	1.005	0.620	0.464	0.974	0.620	0.068

	(0.867–0.900)	(0.927-1.074)	(-0.164-0.184)						
Dummy Classifier	0.459 (0.429–0.492)	-0.013	-2.290		0.123	0.065	0.852	0.062	0.248
Past Pregnancy Models									
Random Forest Classifier	0.858 (0.837–0.875)	0.976 (0.904-1.058)	-0.037 (-0.186-0.121)	0.998	0.549	0.447	0.973	0.544	0.072
Logistic Regression	0.857 (0.836–0.876)	0.995 (0.924-1.076)	-0.008 (-0.184-0.182)	0.999	0.560	0.441	0.965	0.521	0.072
XGBoost Classifier	0.854 (0.832–0.873)	0.935 (0.866-1.016)	-0.094 (-0.252-0.072)	1.002	0.561	0.454	0.972	0.548	0.072
Explainable Boosting Machine	0.860 (0.839–0.879)	1.028 (0.957-1.112)	0.051 (-0.116-0.229)	1.004	0.556	0.443	0.970	0.533	0.072
Dummy Classifier	0.459 (0.429–0.492)	-0.013	-2.290		0.123	0.065	0.852	0.062	0.248

5.3.4 Model Performance Across Ethnicities

Performance metrics stratified by ethnicity are shown in Table 5.6. A linear mixed-effects model with a random intercept for each dataset × classifier row found a significant effect of ethnicity on model AUC ($F_{(5, 99.5)} = 88.36$, $p<0.001$). Compared with Caucasian patients (mean AUC 0.84), discrimination was significantly lower in Black African patients (AUC 0.772; $p<0.001$), Asian patients (AUC 0.757; $p<0.001$) and Southeast Asian patients (AUC 0.755; $p<0.001$), and significantly higher in the "Other" category (AUC 0.903; $p<0.001$). Discrimination in Middle Eastern patients (AUC 0.828) did not differ significantly ($p=0.78$). Generally, the models maintained performance for Caucasian patients (AUC >0.8), however, reduced performance in some smaller or more diverse subgroups. Conversely, the models continued to perform well among Other category, although small sample sizes may have skewed these results. For the First-trimester & Nulliparous models, no individuals in any ethnic group had $n_{eff}<30$. For the Past-Pregnancy models a notable proportion of individuals in ethnic groups had $n_{eff}<30$: 100% for Middle Eastern, 30% for Black African , and 25% for Asian.

5.3.5 Feature Importance

The feature importance analysis identified the top predictors of GDM, including a history of GDM, maternal BMI at booking, maternal age, family history of diabetes, ethnicity, inter-pregnancy weight gain, time between pregnancies, and ethnicity. The SHAP plots for the XGB are shown in Appendix F. The feature importance for all of the models can be found in Appendix F.

Table 5.6. Performance of machine learning models predicting across the different ethnicity sub-groups. Performance measured by AUC evaluated against the validation set combined with the test set. Minimum of 15 samples required.

Dataset and Model	Caucasian	Southeast Asian	Black African	Asian	Middle Eastern	Other
First-Trimester Models	n=24,180	n=1,360	n=554	n=489	n=154	n=824
Random Forest Classifier	0.808	0.747	0.748	0.704	0.761	0.877
Logistic Regression	0.809	0.759	0.769	0.724	0.758	0.860
XGBoost Classifier	0.810	0.751	0.756	0.717	0.786	0.866
Explainable Boosting Machine	0.809	0.753	0.768	0.711	0.718	0.866
Nulliparous Models	n=10,022	n=511	n=167	n=222	n=49	n=406
Random Forest Classifier	0.804	0.727	0.700	0.747	0.803	0.865
Logistic Regression	0.811	0.727	0.722	0.731	0.854	0.870
XGBoost Classifier	0.810	0.732	0.697	0.719	0.752	0.855

Explainable Boosting Machine	0.805	0.729	0.668	0.741	0.786	0.864
Past Pregnancy Models	n=3,648	n=138	n=70	n=51	n=19	n=77
Random Forest Classifier	0.845	0.752	0.790	0.750	0.853	0.911
Logistic Regression	0.837	0.764	0.792	0.793	0.941	0.926
XGBoost Classifier	0.858	0.734	0.810	0.778	0.853	0.932
Explainable Boosting Machine	0.844	0.774	0.783	0.827	0.882	0.913
Multiparous Models						
Random Forest Classifier	0.874	0.758	0.794	0.840	0.824	0.955
Logistic Regression	0.867	0.770	0.823	0.788	0.912	0.953
XGBoost Classifier	0.874	0.699	0.806	0.806	0.824	0.967
Explainable Boosting Machine	0.880	0.776	0.845	0.799	0.824	0.963

5.4 DISCUSSION

This study evaluated the performance of several ML models for early prediction of GDM using data available at the first antenatal visit, and examined whether incorporating information from previous pregnancies enhances predictive performance. Overall, the findings confirmed that incorporating prior pregnancy history can improve early risk prediction for GDM. For example, the Multiparous models that included both first-trimester and previous pregnancy features achieved an AUC of 0.885, higher than models using first-visit data alone, suggesting that a woman's obstetric history is highly informative for forecasting GDM in a subsequent pregnancy. Notably, even a simplified model using only past pregnancy features achieved discrimination of AUC 0.860, highlighting that a focused set of clinical predictors from previous pregnancies can achieve most of the predictive signal.

In the First-trimester models, the LR model performed similarly to more complex models such as XGB and EBM (AUC ~0.82). This contrasts with much of the existing literature, highlighted in a recent meta-analysis that found LR models achieved a pooled AUC of 0.815, compared to 0.889 for non-linear models[83]. This discrepancy suggests that more sophisticated ML algorithms tend to perform better, potentially capturing complex, non-linear relationships in the data. However, the comparisons in the meta-analysis were not always direct comparisons of model performance, as the models were often tested on different datasets, each with varying characteristics, sample sizes, and feature availability. Nevertheless, even when looking at studies that implemented LR alongside more advanced models on the same dataset the pattern of increased performance with non-linear models like XGB or RF remains evident[164,173,174,246,247]. The lack of a gap in LR performance may indicate that the relationships

between early-pregnancy predictors and GDM are largely linear or additive, meaning a well-specified linear model can capture them adequately. It is also possible that the complex models were limited by the data quality or volume, or by the fact that only routine, non-invasive features were intentionally used. This underscores an important point for clinical machine learning, more complex is not always better, especially if interpretability and ease of use are priorities. From a clinician's perspective, a simpler model that offers similar performance might be preferable for integration into practice, due to its transparency and reliability[145].

Following recommended guidance[205,206], clinical usefulness was assessed by plotting net benefit against the decision threshold with "treat-all" and "treat-none" comparators (Fig. 5.3). In DCA the trade-off is encoded by the threshold: the relative harm of a false positive to the benefit of a true positive; equivalently, one additional true positive justifies up to extra false positives. Thus, at thresholds 0.10, 0.20, 0.30, 0.40, 0.50 the implied tolerances are approximately 9, 4, 2.3, 1.5 and 1 false positives per true positive, respectively. Interpretation focuses on clinically plausible thresholds and on ranges where a model's curve exceeds the default strategies[205,206]. Across cohorts, curves for the candidate models were closely aligned and showed positive net benefit within the lower–moderate threshold range, reinforcing the observation that simpler models (e.g. LR) can offer comparable clinical utility to more complex approaches when discrimination is similar. To anchor threshold selection in practice, Table 5.4 reports sensitivity and specificity (with PPV/NPV, FPR and counts per-100) at thresholds 0.10, 0.20, 0.30, 0.40, 0.50; these operating points make the trade-off explicit and enable alignment with service capacity and the relative consequences of false positives versus missed cases.

A strength of this study is the exclusive use of non-invasive, routinely collected data available at booking, which enhances the practicality and scalability of implementing the model in clinical settings. With the exception of height, body mass and blood pressure, all predictors in the models were questionnaire-based or existing records. Among previous studies that used non-invasive features, the majority demonstrate moderate predictive power (AUC 0.7-0.8)[99], with two exceptions to date that demonstrate much better performance[164,179]. In comparison, my models that used non-invasive features achieved AUCs ranging from 0.819 in the First-trimester models up to 0.885 in the multiparous models that included obstetric history, suggesting that leveraging a patient's obstetric history and optimising the ML methodology can substantially boost performance even without biochemical markers. Notably, the Multiparous EBM model demonstrated the highest performance by achieving AUCs of 0.885, which is in the range reported by Belsti et al.[164] (0.921) and Sweeting et al.[179] (0.880), and in the range of models that use biochemical predictors in addition to non-invasive features[174].

Using only previous pregnancy variables in multiparous women improved model performance compared to the First-trimester and Nulliparous models, with EBM achieving an AUC of 0.860. These results align favourably with findings from other studies that emphasise the importance of early pregnancy and preconception data for GDM prediction, particularly when considering non-invasive data collection. For instance, Artzi et al.[94] utilised features accessible 'at the beginning of pregnancy' and achieved an AUC of 0.799 using just nine features collected from questionnaires. It could be argued that this is a similar approach to the First-trimester models without blood pressure measures, which have minimal impact on the model. Additional work has demonstrated the benefit (AUC 0.930) of incorporating biochemical markers like glycosylated haemoglobin (HbA_{1c}) with other features collected during the preconception stage of pregnancy[26]. Thus, these data demonstrate the potential for predicting diagnosis of GDM with data available prior to conception. The inclusion of inter-pregnancy factors, such as weight gain and time interval between pregnancies, provided an additional perspective in this study, enhancing the models' predictive power to an AUC 0.904. This finding aligns with work demonstrating the value of including more comprehensive data in the prediction of recurrent GDM, which achieved an AUC of 0.942 with Light Gradient Boosting and 0.924 with XGB, by incorporating biomarkers such as the OGTT results from the index pregnancy, and fasting plasma glucose (FPG) and triglycerides in the first trimester of the ongoing pregnancy[114].

The reported AUC values in this study demonstrate consistency across some subgroups, which is important for ensuring generalisability of the models, while performing poorly across others. The Nulliparous models achieved an AUC of ~0.81, indicating promising predictive performance in identifying diagnosis of GDM among nulliparous pregnancies. These figures appear higher than those reported in other studies that have focused on nulliparous pregnancies, such as Kang et al.[170], who reported variability in AUC between multiparous (0.720) and nulliparous (0.672) populations, and Cooray et al.[166] (AUC 0.732) and Donovan et al.[111] (AUC 0.710), both of which focused on nulliparous populations using LR. However, direct comparisons should be interpreted cautiously. The underlying cohorts differed in age, BMI, ethnicity, GDM prevalence and data-collection protocols, and our models have not yet been externally validated. Nonetheless, the consistency of performance within our own cohort suggests that the selected predictors capture risk factors for nulliparous women, warranting further validation in independent populations.

Beyond parity-based analysis, I also explored model performance across ethnic subgroups (Table 5.5). While results remained consistent among Caucasian participants,

performance declined in certain minority groups, particularly Asian, Southeast Asian, and Black African populations. This finding is consistent with previous research in Ireland demonstrating a marked decrease in model performance in non-Caucasian populations[69]. In contrast, models continued to perform favourably for Other and Middle Eastern individuals, though the small sample sizes for these subgroups limit the generalisability of these findings. However, models trained on diverse populations in California, USA, noted the opposite trend, whereby models performed well in Hispanic populations but tended to underperform in Caucasian populations[111]. This highlights the need for including more diverse data in model training.

The feature importance analysis confirmed that a history of GDM, maternal BMI, and ethnicity were strong predictors in the early pregnancy GDM models, and aligns with meta-analytical analysis in this field[83]. Family history of diabetes and maternal age were also consistently ranked as important features. Other frequent predictors that were not available in the current feature set were FPG, triglycerides and HbA_{1c}. For the Multiparous and Past Pregnancy models, interpregnancy weight gain and the birthweight percentile of the previous child emerged as highly informative predictors, underscoring the value of a detailed obstetric history.[114]. These findings are confirmatory, as they reinforce the known clinical relevance of these features, and their effectiveness as strong predictors in different studies and populations. This consistency across different model architectures increases confidence that the models are capturing true biological and demographic signals rather than dataset artifacts.

In this study, I developed subset versions of each model by selecting the most clinically relevant features, aiming to enhance the potential for CDSS[53] by relying on a limited set of predictors rather than an entire EHR. The models presented here represent similar model performance to training on the full EHRs (Appendix H), with the reduced risk of overfitting to the data. This streamlined approach could facilitate earlier GDM risk prediction, potentially as early as 12 weeks' gestation or even preconception, thereby enabling more proactive interventions such as earlier diagnostic testing long before standard screening. However, to integrate these models seamlessly into antenatal workflows, clinicians need clear guidelines for classifying high-risk and moderate-risk patients, and the chosen thresholds must carefully balance the risks of false positives (unnecessary interventions) and false negatives (missed high-risk cases). It should be noted that the Past-Pregnancy model is applicable even before a new pregnancy begins (a woman's risk can be assessed interpregnancy), whereas the First-Trimester and Nulliparous models apply at the booking visit for any pregnancy. The Multiparous model, requiring prior gestational data, applies only to women with at least one

previous pregnancy. In our cohort, 40% of pregnancies were to nulliparous women and 60% to multiparous women, so a two-model strategy would be needed in practice to cover all patients.

This study has several limitations. First, the dataset used for model development was derived from a single institution, which may limit the generalisability of the findings. External validation using data from other populations is necessary to confirm the models' robustness, as emphasised by others[111,173,187], who validated their model performance across independent populations. The use of historical data from previous pregnancies also means that model performance may vary based on the quality and availability of such data across healthcare systems. The method by which features were pre-processed could have resulted in a loss of potentially important information. One-hot encoding broad categorical features, such as "Endocrine problems," into binary variables may have discarded valuable context that could have improved predictions. Moreover, I did not have access to blood-based biomarkers in this study, such as FPG or HbA_{1c}, which have been shown to be strong predictors of GDM[26]. While the goal of this study was to develop a model that could be used at the booking visit in the first trimester without the need for extensive laboratory tests, incorporating point-of-care biomarkers could further enhance model performance. Socioeconomic and educational status were also not directly available in this dataset. Previous research has shown that educational attainment, particularly in women, is correlated with health outcomes, including the risk of GDM and inter-pregnancy weight gain[248]. The hospital in this study does not universally screen women for GDM, suggesting that some of the dataset contains undiagnosed GDM[23], despite best efforts to validate the labels. Finally, the models did not include OGTT results, limiting predictions to the IADPSG criteria. Modelling GDM using OGTT results could provide a more versatile tool applicable to multiple diagnostic standards.

5.5 CONCLUSION

In conclusion, these ML models, particularly those incorporating data from previous pregnancies, have the potential early in pregnancy to identify women at greater risk of later diagnosis of GDM. Early identification may allow for timely interventions, which could mitigate the adverse maternal and foetal outcomes associated with GDM. Future research should focus on validating the developed models in external datasets to assess their generalisability. Incorporating additional features, such as lifestyle factors and biomarkers, may further improve model performance and sensitivity. Additionally, prospective studies

evaluating the integration of these models into clinical workflows would be valuable to determine their impact on clinical decision-making and patient outcomes. The next critical step is to verify these models in an independent population (Chapter 6), to ensure that this performance isn't just an artifact of our development sample. These findings warrant external validation to ensure these predictors have similar effects in different healthcare settings and populations.

Chapter 6

Reciprocal External Validation of GDM Risk Prediction Models Using a Machine Learning Model-Exchange Framework

This article has been submitted to the International Journal of Medical Informatics (Manuscript # IJMEDI-D-25-02039) as:

Germaine, M., Belsti, Y., O'Higgins, A. C., Egan, B., Teede, H., Healy, G., & Enticott, J. (2025). Reciprocal External Validation of GDM Risk Prediction Models Using a Machine Learning Model-Exchange Framework

DOI: https://ssrn.com/abstract=5287628

Chapter Overview

This chapter addresses the external validation research question: *Can we overcome traditional data sharing challenges when working with sensitive data, such as patient EHRs, and perform external validation of base models?* The chapter introduces and implements a novel reciprocal external validation framework, a methodological approach designed to address the validation gap that exists in clinical risk prediction research. In this framework, two independent research groups from Ireland and Australia exchanged pre-trained GDM prediction models and the associated data preprocessing pipelines for external validation, a process that avoid the need to share sensitive patient-level data.

The validation resulted in a decline in performance for both models when applied to the respective external cohort. The DCU model's (Chapter 5) AUC fell from 0.82 in its development set to 0.69 on the Australian data, while the Monash model's AUC dropped from 0.93 to 0.77 on the Irish data. Furthermore, both models exhibited miscalibration in the new settings, confirming that their risk estimates were not reliable without local adjustment. This performance decline was attributed to key differences between the cohorts, including but not limited to GDM screening policies, which resulted in different prevalence rates (11.7% in Ireland vs. 21.1% in Australia), and differing population demographics.

The goal of external validation is not necessarily to confirm high performance but to realistically assess it. The REV framework revealed the models' lack of transportability and their context-dependency, important information for ensuring safe and responsible clinical deployment. The REV framework itself stands as a generalisable methodological contribution to the field, offering a practical, privacy-preserving solution that could be adopted by other researchers to ensure more rigorous and collaborative validation, thereby accelerating the translation of trustworthy ML into clinical practice.

6.1 INTRODUCTION

Chapter 5 successfully developed and internally validated a first-trimester prognostic model, achieving good discrimination (AUC ~ 0.82) on the Coombe hospital cohort. However, performance on a development dataset is often optimistic and does not guarantee the model will generalise to other populations. Following thia, the next essential step, as mandated by clinical prediction modelling guidelines[77,90,91,120], is rigorous external validation. This process is critical to test a model's generalisability and to assess for "optimism bias" before any clinical implementation. However, performing such validation is often hindered by practical and legal barriers to sharing sensitive patient-level data[202]. To address this, this chapter introduces and utilizes a Reciprocal External Validation (REV) framework. The primary research question is therefore not whether data sharing can be overcome, but rather: how does the Irish first-trimester LR model (developed in Chapter 5) perform when tested against the Australian (Monash) cohort, and vice versa? This reciprocal test will directly assess model transportability, a key component of our reframed RQ5.

In Australia, GDM affected 17.9% of births in 2021–22, more than doubling from 9.3% in 2012–13[249]. In Ireland, the most recent evidence similarly demonstrates rising GDM since adopting the IADPSG guidelines, from 3.1% in 2008, 12.4% in 2012 and up to 14.8% in 2017[3,250]. GDM poses both short- and long-term health risks for pregnant women and their newborns, including complications such as macrosomia, pre-eclampsia, and an increased likelihood of developing type 2 diabetes[251]. Beyond health consequences, the economic burden associated with GDM-related healthcare costs is substantial. In Australia, the annual healthcare costs attributed to GDM have been estimated at $71.6 million[249] and, in Ireland, costs associated with maternity care for pregnant women diagnosed with GDM are reported to be 34% higher than those incurred in average pregnancies[10]. These costs further emphasise the need for early detection and management of GDM.

There are several well-established predictors of GDM. These include advanced maternal age, family history of diabetes, previous history of GDM, high BMI, history of macrosomia, polycystic ovarian syndrome, and use of medications such as corticosteroids and antipsychotics[195,249]. Furthermore, blood glucose concentrations assessed in early gestation (before 20 weeks) are a continuous measure of risk[33]. Combining risk factors to provide a personalised assessment of the risk of developing GDM via risk prediction models are emerging approaches for early identification of GDM. These are often developed with the aim

to enable timely interventions such as lifestyle modifications, including physical exercise, dietary adjustments[252,253].

Despite the availability of various GDM risk prediction models globally, their clinical implementation remains low[90,119,201]. This is for a number of reasons[201,253] and a key reason is due to a lack of external validations that test model performance using new data (i.e. data not used in model development). Without external validations, there is limited confidence in the model applicability across diverse populations and different geographical settings. Clinical risk prediction experts globally, regardless of clinical discipline, promote that comprehensive external validations are now critical to advance the field[77,91,120]. Without this, the next steps in implementing developed risk prediction models are hindered, and this restricts the progress in advancing digital health and AI-driven healthcare[90,201].

Comprehensive external validations can consist of model performance evaluations conducted using existing secondary data from other settings[91]. Despite the rise in electronic routine health data[254], accessing this data for external validation efforts may be hindered by challenges such as restricted access to appropriate datasets and privacy concerns related to medical health records. For example, requests to authors who have conducted similar work in different geographic areas to use their data for external validation often lead to non-responses or refusals[202]. This inability to access data for external validation underscores the need for alternative solutions.

This study proposes a Reciprocal External Validation approach to tackle the data access challenge. This is a collaborative process where two independent groups exchange their risk prediction models and mutually validate each other's work ensuring both models are evaluated with external data. This approach fosters communication, transparency, and knowledge-sharing, while enhancing the reliability and generalizability of the models, all which are key elements within the expert guidance of Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis (TRIPOD-AI)[101]. It may also provide a pathway forward to enhance the field by potentially enabling future external validations to occur globally, thereby overcoming the current dearth which is limiting the field. This study aims to conduct a case study on the reciprocal external validation of GDM prediction models.

6.2 METHODS

6.2.1 Study Design

Two independently developed GDM risk prediction models were externally validated using a model-exchange approach rather than direct data sharing (no patient-level data were exchanged). Each research team (from Ireland and Australia) has an existing GDM risk prediction model developed using its local cohort, and then exchanged the saved models for validation on the other site's data. The Irish group had identified the Australian group from the peer-reviewed literature in a preliminary study[202]. This collaboration was established to provide a strenuous test of model transportability between two high-income countries with different population demographics (e.g., predominantly Caucasian in Ireland vs. highly diverse in Australia) and different GDM screening policies (risk-based vs. universal), representing a challenging external validation scenario.

6.2.2 Base Models for External Validation

The Monash model is a gradient-boosted decision tree classifier (CatBoost) developed in 2023 from data across three large maternity hospitals within Monash Health. It included eight predictors (mix of both categorical and continuous variables): prior GDM history, ethnicity (six categories), family history of diabetes, past poor obstetric history (defined as a history of preeclampsia or eclampsia; delivering a macrosomic baby or shoulder dystocia), maternal age, height, weight, and parity. This model was developed using 48,502 singleton pregnancies collected between January 2016 and June 2021, with a GDM prevalence of 21.3%. It demonstrated strong performance during internal validation, and subsequently was temporally validated using the latest dataset in the same setting[164].

The DCU model is a logistic regression model developed in 2024 from data contained in the EHRs of a single large tertiary hospital in Dublin, Ireland. The model chosen for this external validation was the First-Trimester LR model from Chapter 5. This model was selected for external validation over the other machine learning models (e.g., RF, EBM) also developed in Chapter 5 for two primary reasons. First, as shown in Table 5.5, its discrimination and calibration were indistinguishable from the more complex ensemble models. Second, its inherent interpretability, simple structure, and ease of implementation make it a more pragmatic and transparent candidate for potential clinical translation, rendering its external validation a high priority. Nine predictors were selected from the EHRs: parity (continuous), maternal age (continuous), BMI (continuous), ethnicity (6 categories), Other Endocrine

condition (PCOS or thyroid disorders; binary), family history of diabetes (binary), previous history of GDM (binary), and systolic and diastolic blood pressure at booking (continuous). This model was trained on 27,561 pregnancies from the Coombe Hospital in Dublin (2018–2022, excluding 2020) with a validated GDM prevalence of 11.7%[155], and it also performed well in its development cohort[156].

6.2.3 Sample size calculation for external validation

The existing sample size determination software[255] was used to calculate the minimum sample size required for external validation for both models. The sample size for external validation of the Irish GDM risk prediction model was calculated using the 'pmvalsampsize' R package, with an AUC of 0.81 and a GDM prevalence of 21.1% from the Monash dataset. Targeting a precision of 0.1 for the AUC, 0.2 for the calibration slope, and 1.0 for the O/E ratio, the minimum required sample size was determined to be 4,501 participants with at least 950 GDM events. Similarly, with the same assumption and targets with an AUC of 0.93 and a GDM prevalence of 11.7% from the Irish dataset, the minimum sample size required to validate the Monash GDM model externally becomes 4501 participants with 527 events.

6.2.4 Validation Data and Study Populations

For external validation, the full Irish cohort served as the validation set for the Monash model, while the Australian cohort was used to validate the DCU model. Both datasets were derived from routine clinical practice in tertiary hospitals. In both settings, GDM was diagnosed according to the IADPSG criteria[1], although the approach to screening differed. Both Ireland and Australia provide maternity care within tax-funded, universal health systems. Reflecting this universal access, 74% of Australian hospital births in 2022 were in public hospitals[256], and Irish studies estimate that roughly 75.2% of women give birth as public patients[257]. Inclusion criteria for both cohorts included routinely collected first trimester data in singleton pregnancies over the age of 18. Each dataset represents a large, real-world hospital population for its region, making them suitable for assessing model transportability. Handling of missing data were described elsewhere[156,164].

6.2.5 Data Preparation

Before external validation, both datasets underwent a process of variable harmonisation. Each dataset recorded ethnicity in raw form but used different classification schemas for model development. The DCU model categorised participants into six broad groups (Caucasian, Black African, Southeast Asian, Asian, Middle Eastern, Other), whereas the Monash model employed a more detailed structure, including categories such as Southern and Central Asian, South-East and North-East Asian, and Middle Eastern, North African, or Sub-Saharan African. Each group applied the other's mapping to its raw data, ensuring that comparable ethnicity labels were produced in both datasets to enable external validation. Additionally, a new binary variable, history of poor obstetric outcomes (shoulder dystocia, pre-eclampsia, macrosomia), was created in the Irish cohort to match the respective measure in the Australian cohort. If women did not have a previous pregnancy, this variable was set to 0. Finally, the Monash model was primarily trained on numerical data, requiring a scaler for feature standardisation, whereas the DCU model incorporated a preprocessor handling both numerical and categorical variables. These scalers and preprocessors were exchanged alongside the models to ensure that each external dataset underwent the same transformations used during the original model development.

6.2.6 External Validation Protocol

The research teams exchanged their prediction models and associated preprocessing pipelines (saved in .pkl format). Upon receipt of each external model and pre-processor, the receiving team loaded these components into its local data analysis environment. The full respective cohorts were then processed using the transferred pipeline, after which the prediction model generated estimated probabilities of GDM for the external cohort. No model recalibration was performed. Crucially, no model recalibration (e.g., intercept or slope adjustment) was performed. This was a deliberate methodological choice to assess the model's 'out-of-the-box' transportability and raw performance in a new population, as recommended for an initial external validation[90,258].

6.2.7 Statistical Analysis and Performance Metrics

Model performance was evaluated by comparing predicted probabilities with actual GDM outcomes using discrimination and calibration metrics. Model discrimination was

assessed by the area under the receiver operating characteristic curve (AUC)[75], with bootstrapping (1,000 iterations) employed to compute 95% confidence intervals for the AUC. Model calibration was assessed both visually, using calibration plots and quantitatively, using slope and intercept, to evaluate the agreement between predicted probabilities and observed outcomes, as recommended[77,91]. For completeness, the Brier score and the observed-to-expected (O:E) ratio in each validation are reported. Model performance on external datasets was compared with that on the development dataset to illustrate the degree of performance decline observed when the models were applied externally. Further analyses examined generalisability and fairness by assessing predictive performance by ethnic group, parity, and previous GDM history. All analyses followed the TRIPOD-AI reporting recommendations for prediction model validation and evaluation[101].

6.3 RESULTS

6.3.1 Demographic and Clinical Characteristics

As summarised in Table 1, the Australian and Irish validation cohorts differed notably in ethnic composition, GDM prevalence, and family history of diabetes. While the Australian cohort represents a highly diverse population within Australia's universal healthcare system, the Irish cohort is predominantly Caucasian .The Australian cohort (N=35,064) featured fewer participants identifying as Caucasian (49.5%) and higher proportions of Southeast Asian (17.5%) and Black African (27.2%) ethnicities, while the Irish cohort (N=27,651) was predominantly Caucasian (87.8%). A greater percentage of women in the Australian cohort also reported a family history of diabetes and a previous history of GDM. Despite similar mean BMI values between the two cohorts, the Australian cohort had a slightly lower mean maternal age. The prevalence of GDM was 21.1% (7,389) in the Australian cohort compared to 11.7% (3,188) in the Irish cohort.

Table 6.1. Comparison of baseline sociodemographic characteristics of validation datasets.

Variable Name	Australian Cohort (N=35,064)	GDM (%) Prevalence	Irish Cohort (N=27,561)	GDM (%) Prevalence
Ethnicity				
Caucasian	17374 (49.5)	3586 (20.6)	24,180 (87.8)	10.0

Black African	9536 (27.2)	1884 (19.8)	554 (2.0)	14.3
Southeast Asian	6130 (17.5)	1512 (24.7)	1,360 (4.9)	33.9
Other	901 (2.6)	181 (20.1)	824 (3.0)	13.7
Asian	629 (1.8)	128 (20.3)	489 (1.8)	21.7
Middle eastern	494 (1.4)	98 (19.8)	154 (0.6)	13.6
Family history of Diabetes				
No	21001 (59.9)	3623 (17.3)	21,154 (76.7)	8.4
Yes	14063 (40.1)	3766 (26.8)	6,407 (23.3)	21.9
Previous History of GDM				
No	32389 (92.4)	5826 (18.0)	26,483 (96.1)	9.5
Yes	2675 (7.6)	1563 (58.4)	1,078 (3.9)	62.5
Current GDM				
No	27675 (78.9)		24,373 (88.3)	
Yes	7389 (21.1)		3,188 (11.7)	
Parity				
0	13983 (39.9)	2750 (19.7)	11,377 (41.3)	9.9
1	12900 (36.8)	2755 (21.4)	9,875 (35.8)	11.4
>=2	8181 (23.3)	1884 (23.0)	6,309 (22.9)	15.0
Age				
Age (Mean ± SD)	30.6 ± 5.1		32 ± 5	
>= 40	1422 (4.0)	485 (34.0)	2,178 (7.9)	17.2
<40	33638 (96.0)	6904 (20.5)	25,383 (92.1)	11.1
BMI				
BMI (Mean ± SD)	26.0 ± 6.1		26.2 ± 5.3	
<25	17143 (48.9)	2595 (15.1)	13,695 (49.7)	4.9
25 to <30	10091 (28.8)	2381 (23.6)	8,423 (30.6)	11.6
30 to <35	4538 (12.9)	1267 (27.9)	3,537 (12.8)	24.2
35 to <40	1912 (5.5)	618 (32.3	1,280 (4.6)	33.0
>=40	1380 (3.9)	528 (38.3)	626 (2.3)	41.9

6.3.2 Model Performance: Discrimination and Calibration

When validated on the Irish cohort, the Monash model's AUC declined from 0.93 (in its development set) to 0.77 (0.762-0.778) (Figure 1A). The calibration plot (Figure 2A) visually indicates risk is overestimated across all predicted probabilities, with a slope of 1.278

(1.229-1.329), an intercept of -0.573 (-0.634—0.511), an O:E Ratio: 0.536 ad a Brier score of 0.11.

The AUC of the DCU model also decreased from 0.819 in the development setting to 0.694 (0.688-0.701) in external validation (Figure 1B). The DCU model had a calibration slope of 0.55, an intercept of 0.17, and a Brier score of 0.16, visually indicating underprediction at lower probabilities, good calibration at intermediate probabilities, and overprediction at higher probabilities (Figure 2B). The timeframe from first contact model exchange was 4 months, after which the models were externally validated within 7 days.

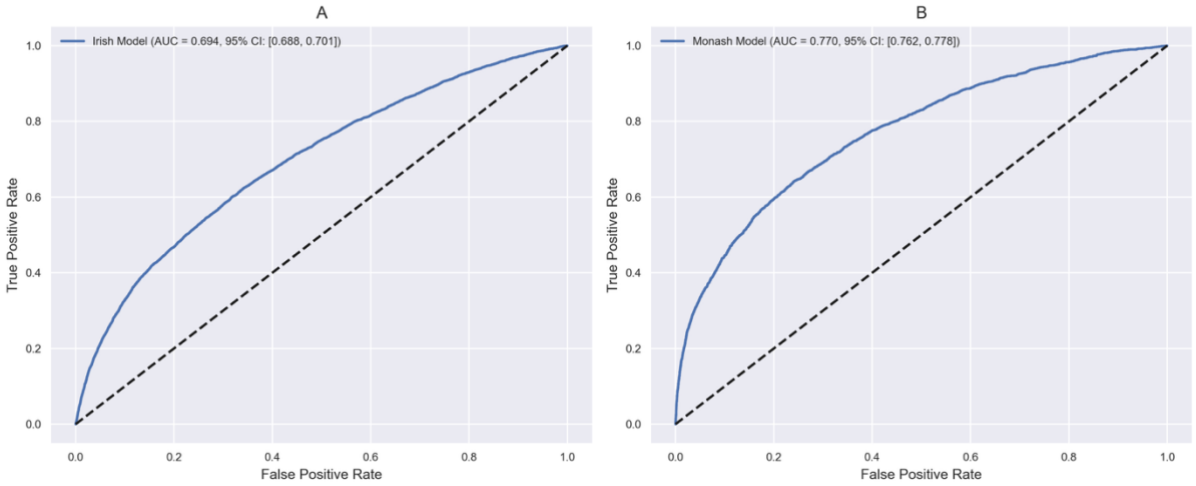


Figure 6.1: Discrimination: The Receiver Operating Characteristic Curve (ROC) for (A) the DCU model validated on the Monash dataset, and (B) the Monash model validated on the Irish dataset.

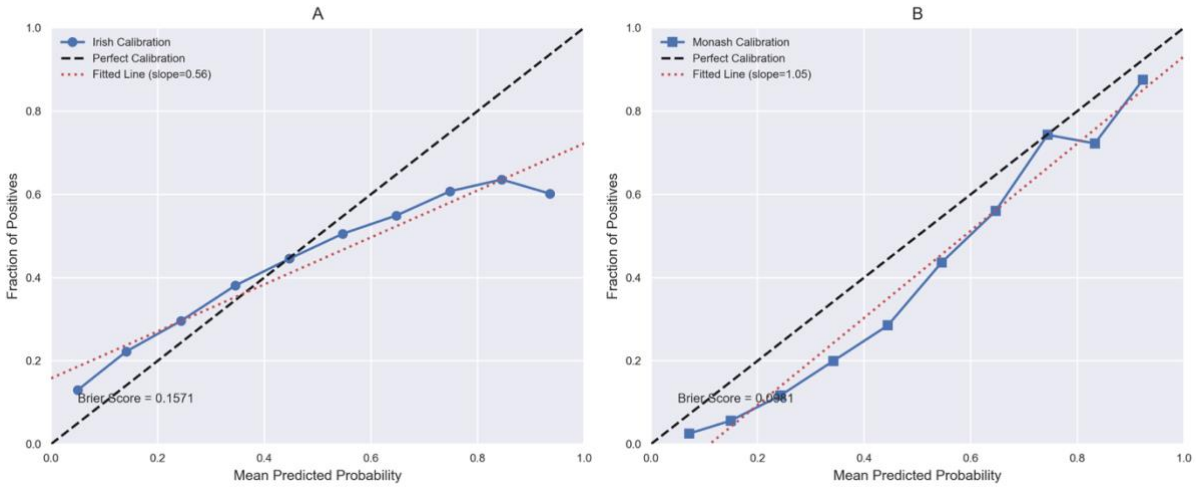


Figure 6.2: Calibration: Calibration plot for (A) the DCU model validated on the Monash dataset, and (B) the Monash model validated on the Irish dataset.

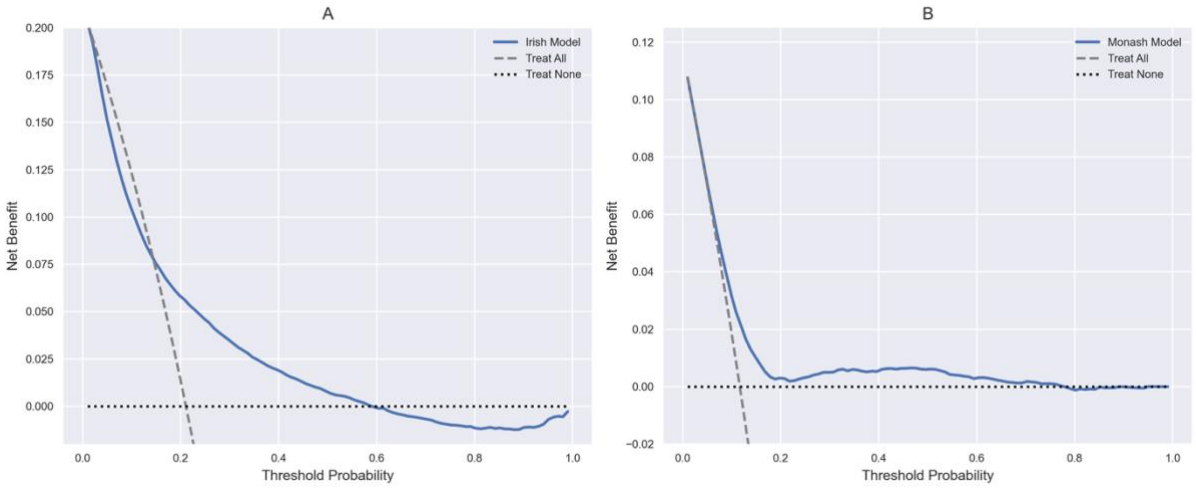


Figure 6.3. Decision curve analysis for (A) the DCU model validated on the Monash dataset, and (B) the Monash model validated on the Irish dataset.

6.3.3 Model Fairness

When the Monash model was validated, discrimination varied substantially among ethnic subgroups (Appendix H). The highest AUC was observed for Caucasian participants (0.755, 0.745-0.765), whereas individuals of Southeast or Northeast Asian origin exhibited the lowest AUC (0.648, 0.555-0.734). Middle Eastern, North African, or Sub-Saharan African participants had an AUC of 0.720 (0.643-0.755), Southern and Central Asian individuals 0.700 (0.693-0.750), and those classified as Other 0.743 (0.701-0.785). Conversely, when the DCU model was validated (Appendix H.2), the AUC ranged from 0.68 to 0.71 across ethnic groups. Performance was highest for individuals classified as Asian (0.71) and somewhat lower for Middle Eastern (0.68) and Southeast Asian (0.68). Black African and Other groups achieved intermediate values (0.69 and 0.70, respectively), while Caucasians demonstrated an AUC of 0.70.

Both models exhibited variations in predictive performance according to parity (Appendix H.3 & H.4). When the DCU model was validated, the AUC was lowest among nulliparous women (parity 0, n=13,983) at 0.65, rising to 0.72) for parity 1 (n=12,900) and 0.71 for parity ≥2 (n=8,181). In contrast, the Monash model validated achieved higher overall values, with AUCs of 0.761 (0.747-0.775) for parity 0 (n=11,377), 0.773 (0.761-0.784 for parity 1 (n=9,875), and 0.791 (0.779-0.802) for parity ≥2 (n=6,309).

Validation of the Monash model showed stronger performance among women without a previous GDM diagnosis (AUC 0.725, 0.715-0.734; n=26,483) compared to those who had a documented history of GDM (AUC 0.637, 0.604-0.672; n=1,078). Similarly, DCU model

showed stronger performance among women without a previous GDM diagnosis (AUC 0.65, n=32,389) compared to those who had a documented history of GDM (AUC 0.58, n=2,675), as illustrated in Appendix H.5 & H.6. All results are reported in Table 2.

Table 6.2. Discrimination and calibration results from the external validation of both the Irish and Monash models, including fairness metrics.

Metric	Monash model on Irish cohort	DCU model on Australian cohort
Discrimination (AUC)	0.77 (0.761–0.779)	0.694 (0.688-0.701)
Calibration slope	1.278 (1.229-1.329)	0.55
Calibration intercept	–0.573 (-0.634—0.511)	0.17
O:E ratio	0.536	–
Brier score	0.11	0.16
Ethnic subgroups AUC (range)	0.648 – 0.755	0.68 – 0.71
Caucasian	0.755 (0.745-0.765)	0.70
Middle Eastern/North African/Sub-Saharan African	0.720 (0.643-0.755)	0.68
Southern/Central Asian	0.700 (0.693-0.750)	–
Southeast/Northeast Asian	0.648 (0.555-0.734)	0.68
Other	0.743 (0.701-0.785)	0.70
Parity AUC		
Parity 0	0.761 (0.747-0.775)	0.65
Parity 1	0.773 (0.761-0.784)	0.72
Parity ≥2	0.791 (0.779-0.802)	0.71
Prior GDM history AUC		
No prior GDM	0.725 (0.715-0.734)	0.65
Prior GDM	0.637 (0.604-0.672_	0.58

6.4 DISCUSSION

This study proposed and implemented a reciprocal model-exchange validation process for risk prediction models. In this approach, two independent research teams from different parts of the globe (Ireland and Australia) exchanged their models (instead of sharing any patient data) for external validation. This novel strategy circumvents data-access barriers that often hinder multi-centre validation[259], while still adhering to best-practice recommendations for external validation studies[77,91,120]. I demonstrated the practical implementation of this approach by exchanging GDM prediction models and externally validating them.

Many predictive models never undergo rigorous external validation due to barriers in data sharing particularly when using EHRs[260]. This reciprocal model-exchange validation process exemplifies a novel solution to a well-recognised challenge in risk prediction model research, directly addressing calls in the literature for innovative strategies to facilitate external

validation, which is often a neglected yet crucial step[261]. The model-exchange method allowed us to assess the performance of two GDM models without breaching data privacy. Typical barriers extend further to heterogeneous ethics-and-governance requirements that demand site-specific approvals, sometimes across countries[77]; time consuming data-sharing contracts that can leave "open data" unavailable in practice[262]; limited interoperability between EHR platforms that hampers technical integration of models[263]; and institutional worries about liability or reputational damage in the wake of any breach[264]. This external validation method thus not only demonstrates the generalisability (or limits thereof) of each GDM risk model but is a practical example of how to implement multi-site external validation in situations where data cannot be freely shared, like health data. Such strategies are increasingly important to accelerate the advancement and responsible translation of prediction models into clinical practice, ensuring they are rigorously validated and, when necessary, recalibrated or updated before deployment[91,261].

In this practical demonstration of reciprocal external validation of GDM models, each model's performance decreased notably on the external data. The Monash model's AUC dropped from 0.93 to 0.77 on the Irish cohort, and the DCU model's AUC fell from 0.81 to 0.69 on the Australian cohort. Calibration estimates confirmed systematic risk misestimation in these external validations , each model tending to over or under-predict GDM probabilities outside its training domain, with calibration-in-the-large of -0.573 for the Monash model and 0.17 for the DCU model; slopes were 1.278 and 0.55 respectively. This indicates that recalibration would be necessary for clinical use in this new population. This poor calibration is a direct and expected consequence of the methodological decision (Section 6.2.6) to apply the model without any form of recalibration. These findings are consistent with prior research showing that GDM prediction models usually exhibit reduced performance after external validation[89,121].

The magnitude of the model performance declines observed in our cross-country validations appear to be larger than typically seen in validations within a single country or similar healthcare setting. For example, models validated on external cohorts from the same country have shown more modest performance drops[111,173,187]. Whereas validations spanning very different populations (e.g. a model developed in Europe applied in the US) can result in AUC declines comparable to the present study[189]. These results thus reinforce that differences in healthcare context and population characteristics may substantially impact a model's generalisability.

A likely contributor to the performance gap is the difference in GDM screening practice between the two cohorts. In Australia universal 75 g OGTT screening yields a GDM prevalence of ~21%, whereas Ireland applies a risk factor approach, with only ~60% of women tested and a lower observed prevalence of ~12%. Findings from Irish research suggest that as many as 16% of true GDM cases may go undetected[23]. Undiagnosed women are therefore misclassified as non GDM, skewing predictor distributions, blunting model coefficients and inflating apparent calibration error when models are exchanged. This "hidden case" bias also hampers identification of history based predictors such as prior GDM or macrosomia (Appendix I.5-6).

Population composition, particularly ethnicity, provides a second explanation. The Australian cohort is ethnically diverse, whereas the Irish cohort is more than 85% Caucasian. Southeast Asian women illustrate the problem: they carried a 33% GDM rate in Ireland but 24.7% in the Monash cohort (Table 1), so the DCU model may have overweighted this subgroup. Across large series, non-European groups (South Asian, East Asian, Middle Eastern) show two to three-fold higher baseline risk than European women[7,189,265,266]. Genomic studies confirm that the excess is not purely environmental: a type 2 diabetes polygenic risk score increased GDM odds by 45% in more than 5,000 South Asian pregnancies, independent of BMI or family history[267], and half of the loci influencing gestational glycaemia in a 116,000 pregnancy Chinese genome-wide study were East Asian specific[268]. Models developed in largely European samples therefore tend to under-predict risk in high prevalence ethnic groups and over-predict in low prevalence groups, precisely the bow shaped calibration pattern observed after model exchange. Re-estimating a version of the early pregnancy Monash model with a six-level ethnicity variable largely restored discrimination on external data[166], showing that simple recalibration or ethnicity specific models may be required for equitable performance.

Taken together, systematic differences in (i) case ascertainment and (ii) ethnic and genetic risk profiles may explain part of the cross-country drop in performance and underscore a broader clinical message. Before deployment, prediction tools should undergo local external validation and, if necessary, recalibration or updating, especially when screening practices or ancestral composition differ from those in the derivation cohort[77,91]. Where ethnic heterogeneity is large, parallel ethnic specific models or adapted risk thresholds may be needed; conversely, in homogeneous settings universal models may overestimate risk and lead to unnecessary intervention. Finally, the evidence that polygenic scores capture ethnic specific liability suggests that future GDM tools could combine clinical and genomic predictors, but

only after transparent, multi-site validation that guards against precisely the calibration failures documented here.

Despite its strengths, this study still experienced limitations in external validation. The differences in GDM screening approaches and demographics between the two cohorts raise important considerations regarding the accuracy and completeness of training data, which could affect the reliability of GDM prediction models when implemented in new populations. Inaccurate predictions could lead to missed diagnoses or delayed interventions, ultimately affecting maternal–foetal outcomes[43,44]. These considerations affecting variable availability and potential misclassification remain obstacles to achieving consistent performance across heterogeneous cohorts. Recent work underscores these challenges, emphasising the need for continuous model refinement and validation in diverse populations[166]. While REV offers a solution to data sharing barriers, it's important to acknowledge that it inherently limits the depth of diagnostic analysis compared to scenarios where pooled individual patient data is available. For instance, detailed examination of specific predictor distributions in misclassified cases across datasets, or fine-tuning recalibration strategies, is more challenging without direct access to the external site's raw data.

6.5 CONCLUSION

In conclusion, reciprocal model exchange offers a concise, privacy preserving route to the rigorous external validation that prediction models must clear before clinical use. Demonstrated here with GDM models from Ireland and Australia, the approach demonstrated how screening policy and population mix can erode performance. By enabling two sites to test each other's model without sharing raw data, the framework directly addresses the "validation gap" identified in recent BMJ guidance on machine learning evaluation[77,91,120] and aligns with the transparency advocated by TRIPOD+AI[101]. Future research should apply this strategy across multiple sites and focus on enhancing model adaptability (e.g. through recalibration) to account for differences in screening and population characteristics. Further advances will require globally harmonised, interoperable datasets with common variable definitions and outcome criteria, supporting multi-site collaborations of sufficient scale to obtain precise estimates across diverse populations. The final, and most critical, test is to move beyond retrospective data entirely and assess the model's performance and feasibility in a live, prospective clinical workflow. Chapter 7 details this final validation stage.

Chapter 7

Prospective Clinical Validation of a First Trimester Machine Learning Model for Gestational Diabetes Prediction in Routine Care

Chapter Overview

This chapter addresses the implementation research question: *Can we assess the deployment validity of the model in a clinical setting, does the model maintain predictive performance when deployed, and can it detect GDM earlier than current diagnosis?* The chapter presents a prospective, single-centre clinical evaluation where the first-trimester LR model from Chapter 5 was integrated into a routine antenatal workflow at the Coombe Hospital. This study represents the final stage of the thesis's "code to clinic" journey, moving beyond the retrospective analyses in the preceding chapters to real-world assessment.

In this clinical implementation. The model demonstrated moderate discriminative ability with an AUC of 0.762 (95% CI 0.681-0.837) and acceptable calibration. At its pre-specified risk threshold, the model functioned as a useful "rule-in" tool, with high specificity (95%), but modest sensitivity (37%). The application of the model also resulted in 1 in 5 women being diagnosed with GDM 10-12 weeks earlier than with standard screening protocols, enabling them to enter the diabetes care pathway earlier.

This chapter also adds to the validation attrition narrative that is central to this thesis. The model's performance declined from an interval validation AUC of ~0.82 (Chapter 5) to a prospective AUC of 0.76. This decline demonstrates the validation gap between a model's performance on a curated, retrospective dataset and its performance in a live, prospective setting, providing a realistic benchmark and an argument against the optimism bias seen in less rigorously validated models. By integrating the ML tool into a busy clinical workflow and demonstrating its ability to function as intended, this chapter provides a rare example of a code to clinic pipeline, bridging the gap between algorithm development and clinical impact.

7.1 INTRODUCTION

The preceding chapters have documented a 'validation attrition' narrative. The first-trimester model developed in Chapter 5 showed promising internal validation (AUC ~0.82), but Chapter 6 revealed a significant performance drop when the model was externally validated on a retrospective Australian cohort (AUC ~0.69). This 'validation gap' highlights that retrospective performance is not a reliable proxy for real-world utility. The final, and most critical, test is to move beyond retrospective data entirely and assess the model's performance and feasibility in a live, prospective clinical workflow. This chapter addresses RQ6 (Implementation) by prospectively validating the first-trimester model within the routine antenatal care setting at the Coombe Hospital. This study evaluates the model's performance when applied in real-time to new patients, providing a true benchmark of its deployment validity. The primary aim is to assess the model's real-world discriminative and calibration performance. A secondary aim is to quantify the clinical trade-offs of its implementation, specifically its ability to detect GDM cases earlier than standard care and the real-world impact of its false negative rate at a pre-specified clinical threshold.

GDM is a common obstetric complication that poses serious risks to both mother and baby[269]. Women with GDM face increased chances of hypertensive disorders (e.g. preeclampsia) during pregnancy[270] and a higher lifetime risk of type 2 diabetes[271], while their offspring are prone to macrosomia, neonatal hypoglycaemia, and future metabolic disease[272–274]. Globally, 14–15% of pregnancies are affected under current diagnostic criteria and rising alongside obesity and older maternal age[7,269]. Early identification of GDM (in the first trimester) is desirable, as timely nutritional or pharmacological interventions could mitigate hyperglycaemia exposure and improve outcomes[48]. However, standard practice typically relies on universal screening at 24–28 weeks' gestation via OGTT, meaning that preventive measures often begin only in mid-pregnancy. Earlier risk stratification (at the booking visit ~12 weeks) using clinical risk factors is variably implemented and has only modest predictive value[84]. There is no consensus on optimal early screening, especially in high-risk populations such as those with obesity or certain ethnic backgrounds[269].

In recent years, ML models have shown promise for predicting GDM from first-trimester data, potentially improving upon conventional risk factor-based tools[83]. Numerous retrospective studies have built early-pregnancy prediction models incorporating demographics, clinical history, or novel biomarkers, often reporting high discriminative performance in internal testing[99]. For example, ML models using first-trimester clinical and

biochemical features have achieved area under the ROC curve (AUC) values around 0.80–0.90 in development cohorts [83,99]. Despite these advances, translation to clinical practice remains limited. Most published models have not been externally validated in different cohorts (see Chapter 2), nor prospectively in real-world settings, and concerns remain about generalisability, usability, and impact on care[275]. Indeed, a growing number of AI-based decision support tools demonstrate excellent *in silico* performance, but few have yet shown benefit or been rigorously assessed in live clinical workflows. Early-stage clinical evaluation, at small scale and under real-use conditions, is now recognised as a crucial step to ensure an AI system's actual performance, safety, and integration with human users before any large trials or deployment[275].

I have previously developed a first-trimester ML prediction model for GDM using EHR data from a large retrospective cohort The Coombe Hospital[156] (Chapter 5). The final model (n=27,561, 11.6% GDM prevalence) achieved good discrimination (AUC 0.819) and good calibration in internal validation, suggesting that a limited set of routinely collected features could feasibly stratify GDM risk at the booking visit. However, as noted in my prior work and by others, external validation and prospective trials are needed to confirm the model's utility and impact in practice. An independent external validation of the model was undertaken in an Australian maternity cohort (to evaluate generalisability across populations), which resulted in a substantial decrease in performance (AUC 0.694) (see Chapter 6). Following this, the present study represents the next developmental phase: a prospective clinical (implementation) validation of the model within routine antenatal care at the original institution. Thus, the aim of this study is to evaluate the ML tool's performance and usability when integrated into live clinical workflows, meaning the model generated risk predictions in real time during clinic visits. The primary aim of this study is to evaluate the model's real-world predictive performance (discrimination and calibration) and clinical feasibility when integrated into a live workflow (RQ6). A key secondary aim is to quantify the clinical trade-offs of this implementation at a pre-specified risk threshold.

7.2 METHODS

7.2.1 Study Design

I conducted a prospective, single-centre, early-stage clinical evaluation (DECIDE-AI Stage 2a-2b) of a previously developed ML prediction model[156] (Chapter 5). Ethical approval was granted by the Coombe Hospital Research Ethics Committee (REC ref Study No. 16 –

2024). All participants provided written informed consent using the patient information leaflet approved by the REC. Women were approached at their first-trimester booking visit. For those who consented, routinely collected EHR data were entered into the locked ML model, which classified each participant as either high- or low-risk for developing GDM. Women in the high-risk group entered the intervention arm, while those deemed low-risk continued with standard care and underwent the routine 26-to-28-week 75 g OGTT. The study workflow is summarised schematically in Figure 1.

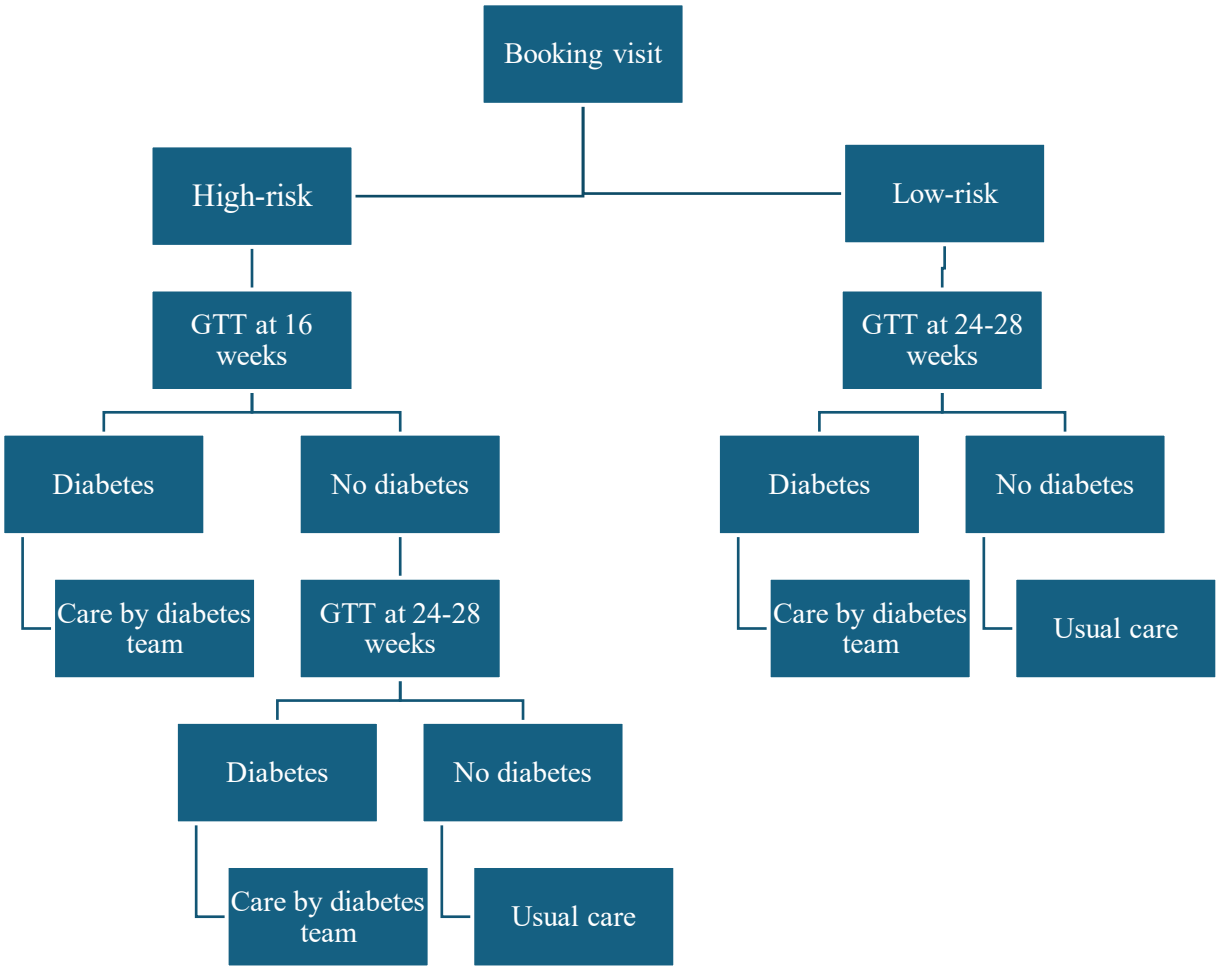


Figure 7.1. Schematic overview of the study design.

7.2.2 Setting and participants

The study took place at the Coombe Hospital (tertiary maternity centre, Dublin) and recruitment process was embedded within the routine clinical workflow. Pregnant women attending either their first antenatal booking visit or dating ultrasound (both scheduled at 11+0 to 13+6 weeks' gestation) between 1 December 2024 and 31 January 2025 were identified by

clinic staff and approached by a single, dedicated researcher (MG) while in the outpatient waiting area, prior to their clinical consultation. I approached consecutive eligible patients in the antenatal booking and dating scan waiting areas, provided a brief study explanation and patient information leaflet, and obtained informed consent for participation. Inclusion criteria were: age ≥18 years, singleton pregnancy, attending first trimester booking, and no type 1 or type 2 diabetes previously diagnosed. Exclusion criteria were: late booking (>16 weeks' gestation), inability to provide informed consent (e.g. language barrier or cognitive impairment), multiple pregnancy (twins or higher order), or known pre-existing diabetes.

7.2.3 Intervention: ML Prediction Tool and Integration

The ML prediction tool evaluated is the model developed in Chapter 5 that uses first visit clinical data to predict the likelihood of GDM later in pregnancy. In brief, the model is First-Trimester LR algorithm developed in Chapter 5 and externally validated in Chapter 6. The model takes routine clinical data from the booking visit and outputs the probability of GDM diagnosis later in pregnancy. Key predictive features include maternal demographics (age, ethnicity), anthropometrics (BMI), obstetric history (parity, history of GDM), family history of diabetes, other endocrine condition (PCOS/Hypothyroidism) and blood pressure. The final model uses a set of 9 top features identified in development to balance accuracy with interpretability (see Chapter 5). The choice of a 'simple' LR model over more complex algorithms (e.g., XGBoost, Random Forest) was a deliberate decision. The systematic review in Chapter 2 found no consistent evidence of superior performance from complex models for this task. This was confirmed in our own internal validation (Chapter 5, Table 5.5), where the LR model's discrimination (AUC 0.819) and calibration were indistinguishable from the ensemble methods. Given this performance parity, LR was selected for prospective validation as it offers significant advantages for clinical translation, namely high interpretability, computational simplicity, and ease of integration into a real-time clinical workflow[198].

For this study, the model was deployed within the hospital and run in real time for each consented participant after the booking visit. Immediately after consent, the researcher opened the participant's routine EHR (Euroking K2) and entered the first-trimester variables required by the GDM-prediction model. A total of 298 women met eligibility criteria and provided complete follow-up data. Missingness was negligible because data were captured prospectively at the point of care; no imputation was required. The 9 required predictor variables (Chapter 5, Table 5.3) were manually transcribed from the EHR into the standalone, pre-validated ML

model interface (a secure Python-based application). The model executed locally, generating a risk score (probability) and binary classification (High/Low) within seconds. This result was recorded by the researcher in the secure study database, but was not written back to the EHR and was not visible to the clinical team. The output was categorised as "high-risk" if the predicted probability met or exceeded a pre-specified threshold (set during model development to target ~10% identification rate), and "low-risk" otherwise. This threshold was determined based on the goal of the prediction tool: to identify high-risk patients while reducing the number of false negatives, which would send additional women for testing unnecessarily. As all women would be screened at the 26-28 week period, it was important not to conduct too many tests early in pregnancy due to hospital resource constraints. All model outputs were logged for audit and traceability purposes, aligning with FUTURE-AI principles of traceability and reliability in deployment[142].

To simulate a potential future use-case (where high-risk women might be offered earlier testing), the study protocol did include an optional early OGTT at ~16 weeks' gestation for participants whom the model classified as high-risk. Women in the high-risk category were informed by the research team about the option of an earlier OGTT (in addition to the routine 24–28 week test, free of charge) for research purposes, but this was not mandated by clinical staff. Clinical providers were aware that some patients might have an extra OGTT due to research, but the model's risk designation was not used to direct any therapeutic interventions at that stage. Those who underwent an early OGTT and were diagnosed with GDM at 16 weeks were referred for standard GDM care at that point (thus ethical obligations were upheld despite the tool being officially inactive in clinical decisions). Participants who had a negative early OGTT continued routine antenatal care and were referred for the standard 24–28 week OGTT (Figure 7.1).

7.2.4 Outcome Definition (Reference Standard)

The primary outcome was the development of GDM, as diagnosed by the standard 2-hour 75 g OGTT at 24–28 weeks of gestation, or a positive OGTT at week 16 in high-risk patients (as predicted by the ML tool). I used the IADPSG criteria to define a positive OGTT[1]. Participants with an early OGTT were assessed by the same criteria. OGTT results were extracted from hospital lab records by the study team, and GDM status (positive/negative) was assigned accordingly. In addition, I recorded the gestational age at GDM diagnosis for each case (to distinguish early-diagnosed GDM from conventional timing). Women who did not

complete an OGTT by pregnancy's end (due to missed appointment or other reasons) were considered to have unknown GDM status and were excluded from the primary analysis of predictive performance. Only participants with a completed reference standard (OGTT) result were included in the final outcome analysis, in accordance with TRIPOD guidelines for handling missing outcomes[101].

7.2.5 Statistical Analysis

Following CONSORT-AI extensions, screening, enrolment, and follow-up were recorded and reported a CONSORT adapted table 7.1 in the Results[276]. Baseline maternal characteristics are presented overall and stratified by GDM outcome; continuous variables are summarised as mean±SD and categorical variables as n (%). Model performance was evaluated prospectively: discrimination was quantified by the AUC with a bootstrap 95% CI[75]; at the pre-specified "high-risk" cut-off I calculated sensitivity, specificity, positive and negative predictive values, and overall accuracy, with CIs derived by bootstrapping[277]. Calibration was assessed with a calibration plot, intercept and slope statistics, Brier score, and the Hosmer–Lemeshow test[204]. Decision-curve analysis quantified the net benefit of using the model to trigger early OGTT across threshold probabilities from 0% to 30%, compared with strategies of testing all or none[121,205]. Analyses were conducted in Python 3.11, using scikit-learn for performance metrics, and statsmodels plus SciPy for calibration and decision-curve calculations. Group differences were analysed in R 4.3.0 using Welch's two-sample t-test for continuous variables and χ^2 tests (switching to Fisher's exact when expected counts < 5) for categorical variables, calculated from the summary statistics.

Using G*Power 3.1.9.7, it was calculated that 36–52 GDM-positive cases were needed to detect the published birthweight difference between early- and late-diagnosed GDM (α=0.05, 1−β=0.80)[239]. Given the hospital's ~12% GDM prevalence[155,156] and the 16% under-diagnosis associated with Ireland's non-universal screening[23], recruitment was set at a target of 300 women to secure at least 36 GDM cases. However, the sample-size calculations using Riley et al.[255], following their 2025 guidance to include at least 200 outcome events for fair and precise risk estimates[278], indicated that 1,350 participants would be needed to prospectively validate the model with a 95% CI no wider than ±0.05. It is explicitly acknowledged that this sample size (target n=300, 52 events observed) is modest for a primary validation of a prediction model and may result in wide confidence intervals for performance metrics, limiting the precision of the estimates[255,278]. This study was therefore designed not as a definitive

validation, but as an early-stage clinical evaluation (DECIDE-AI Stage 2a-2b)[275] focused on assessing feasibility, workflow integration, and preliminary real-world performance.

7.3 RESULTS

7.3.1 Recruitment and Participant Flow

Over the two-month enrolment period, 398 women were approached; 299 (75%) consented and 298 completed baseline assessment (Table 7.1). By study close, 235 participants (79% of those enrolled; 59% of those approached) had a definitive 24–28 week OGTT result. The 63 without an outcome comprised 17 not yet due an OGTT, 35 who missed the appointment, nine who declined testing, and three early pregnancy losses. Thus, the analysis cohort numbered 235, of whom 52 (22%) were diagnosed with GDM, 10 at week 16 and 42 at routine screening. The 10 early diagnoses are relative to a total of 27 early OGTTs triggered by the ML tool predictions. Participant flow is documented in Table 7.1.

Table 7.1. Recruitment and participant flow.

Stage	n	% of approached (398)
Approached	398	100%
Consented	299	75.1%
Booking not completed	1	—
Recruited	298	74.9%
No OGTT performed*	35	11.7%
Declined OGTT	9	3.0%
Miscarriage before OGTT	3	1.0%
Awaiting OGTT (as of 22-05-2025)	16	5.7%
Completed OGTT (analysis set)	**235**	**58.8%**

*unknown reasons (administrative/logistical).

7.3.2 Baseline characteristics

Among the 235 analysable participants, mean age was 33±5 years, BMI 27.9±5.7 kg/m^{-2}, and gestational age at recruitment 12.8±2.1 weeks. The cohort was 73.6% Caucasian; the main minority groups were Indian (9.4%), Brazilian (3.8%) and Pakistani (3.8%). Current smoking was reported by 5.1% and vaping by 6.8%; 31.1% had a first-degree family history of diabetes. Prior GDM was documented in 9.8% of women. BMI, previous GDM and family history of diabetes were all higher in the GDM positive patients (p<0.05 for all). These baseline characteristics are summarised in Table 7.2.

Table 7.2. Baseline characteristics of the patients enrolled in the prospective clinical validation.

Variable (booking visit)	All recruited (n=298)	Analysis set (n=235)	GDM Positive (n=52)	GDM Negative (n=183)
Age, years - mean±SD	33±5	33±5	34±5	33±5
Gravidity - mean±SD	2.4±1.6	2.5±1.7	2.7±1.6	2.4±1.7
Parity - mean±SD	0.9±1.1	0.9±1.2	1.1±1.1	0.8±1.2
BMI, kg/m^{-2} - mean±SD	27.9±5.7	28.4±5.9	31.6±6.1	27.4±5.5*
Gestational age at recruitment, weeks - mean±SD	12.8±2.1	12.7±1.9	12.6±2	12.7±1.9
Systolic BP - mean±SD	111±11	111±11	114±11	110±11*
Diastolic BP - mean±SD	69±8	69±8	70±8	68±8
Current smokers - n (%)	16 (5.4%)	12 (5.1%)	4 (7.7%)	8 (4.4%)
Current Vaping - n (%)	23 (7.7%)	16 (6.8%)	3 (5.8%)	13 (7.1%)
Previous GDM - n (%)	27 (9.1%)	23 (9.8%)	15 (28.9%)	8 (4.4%)*
Family history diabetes - n (%)	88 (29.5%)	73 (31.1%)	23 (44.2%)	50 (27.3%)*
Known endocrine disorder (thyroid/PCOS) - n (%)	56 (18.8%)	50 (21.3%)	11 (21.2%)	39 (21.3%)
Ethnicity - n (%)				
Caucasian	221 (74.2%)	173 (73.6%)	33 (63.5%)	140 (76.5%)
Indian	25 (8.4%)	22 (9.4%)	8 (15.4%)	14 (7.7%)
Brazilian	12 (4.0%)	9 (3.8%)	3 (5.8%)	6 (3.3%)
Pakistani	11 (3.7%)	9 (3.8%)	2 (3.9%)	7 (3.83%)
Other (≤ 3.5% each)	39 (9.7%)	22 (9.4%)	6 (11.5%)	16 (8.7%)

*indicates statistically different than GDM Positive *($p<0.05$)*

7.3.3 Model Predictive Performance

The ML model showed moderate discrimination with an AUC of 0.762 (95% CI 0.681-0.837, Figure 7.3). Calibration was acceptable (intercept -0.045; slope 0.808; Brier score 0.14; Hosmer–Lemeshow p=0.112, Figure 7.2). Using the pre-specified high-risk threshold (~10% of participants), sensitivity was 37% and specificity 95%; positive and negative predictive values were 66% and 84%, respectively. Decision-curve analysis indicated net benefit over "test-all" or "test-none" strategies for threshold probabilities between 10% and 30%.

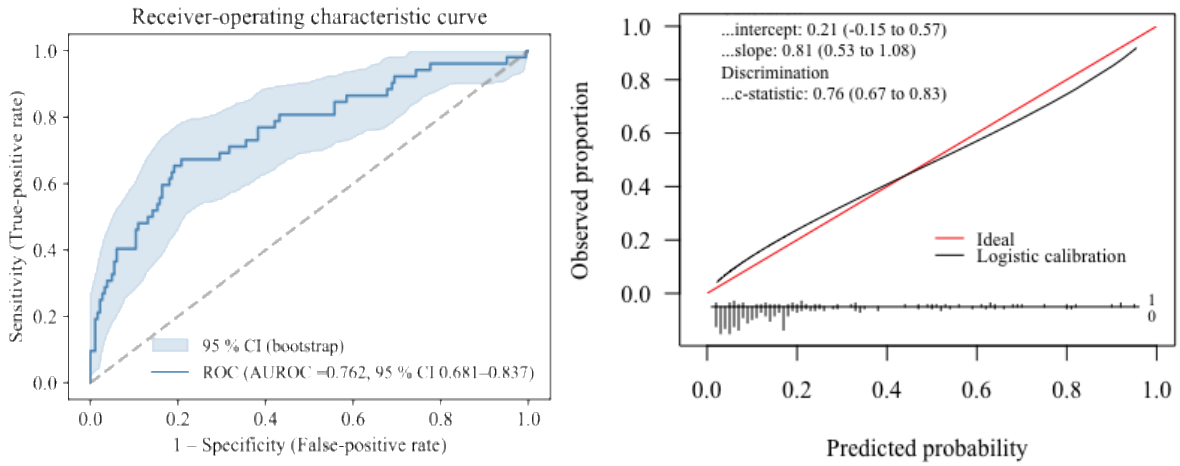


Figure 7.2. The ROC curve (left) and calibration plot (right) based on the ML tool predictions of GDM probability.

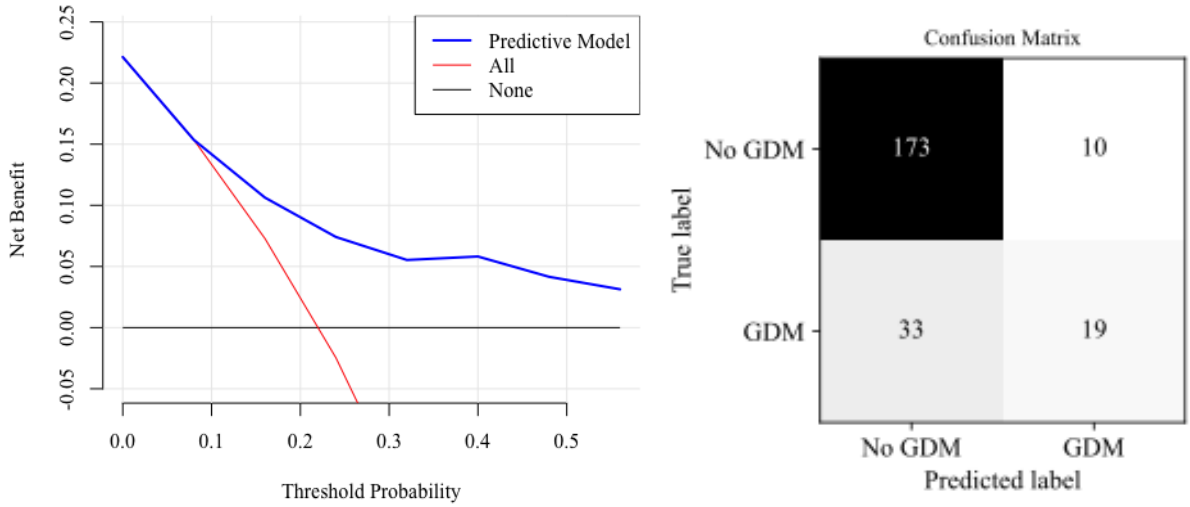


Figure 7.3. The decision-curve analysis and confusion matrix of the output of the ML tool relative to the actual GDM outcomes. The confusion matrix represents the sensitivity and specificity at the threshold used in the clinical deployment.

7.3.4 Clinical Classification Outcomes

At the pre-specified high-risk threshold, designed to identify ~10% of participants, the model produced the following confusion matrix against the final OGTT diagnosis: 19 true positives, 10 false positives, 173 true negatives, 33 false negatives. Thus, the tool generated nearly twice as many true positives as false positives (19 TP vs 10 FP). The model identified 32 women as high-risk in total; 27 attended an early OGTT at 16 weeks. Among the 27 women identified by the tool, 10 were diagnosed with GDM at the 16-week OGTT. Of the 17 who tested negative, seven were GDM positive at 24–28 weeks, nine remained normoglycaemic, and one is still awaiting her 28-week test. Two women who did not attend the early OGTT

were positive at 28 weeks, one did not attend and was subsequently negative, one delivered abroad and was untested, and one miscarried.

7.3.5 Sensitivity Analysis

When the 63 women without an OGTT were classified exactly as predicted (best-case), AUC remained similar at 0.791 (95% CI 0.721–0.860) and calibration changed only modestly (intercept = –0.074, slope = 0.917), while sensitivity and specificity improved slightly to 40% and 96%, respectively. Coding all missing outcomes as non-GDM produced virtually the same operating characteristics (AUC 0.775, intercept –0.267), whereas the extreme worst-case assumption that every missing woman had GDM halved the calibration slope (0.287) and reduced AUC to 0.570 but left specificity unchanged. Results are reported in Table 7.3.

Table 7.3. Sensitivity analysis for the best case, complete case, all negative and all positive missing values.

Scenario	n	AUC (95% CI)	CITL	Slope	Brier	TP	FP	TN	FN	Sens %	Spec %	PPV %	NPV %
Complete-case	235	0.762 (0.681-0.837)	-0.045	0.808	0.14	19	10	173	33	0.365	94.5	65.5	84.0
Best-case	298	0.791 (0.721–0.860)	–0.074	0.917	0.115	22	10	233	33	0.400	95.9	68.8	87.6
All NEG	298	0.775 (0.707–0.846)	–0.267	0.836	0.116	19	13	233	33	0.365	94.7	59.4	87.6
All POS	298	0.570 (0.505–0.644)	0.073	0.287	0.275	22	10	173	93	0.191	94.5	68.8	65.0

CITL, calibration-in-the-large; **TP**, true positives; **FP**, false positives; **TN**, true negatives; **FN**, false negatives; **Sens**, sensitivity; **Spec**, specificity; **PPV**, positive predictive value; **NPV**, negative predictive value

7.4 DISCUSSION

To my knowledge, this is among the first prospective clinical validation that integrated a first-trimester ML risk prediction model for GDM into routine antenatal practice, evaluating its real-time performance and usability. The model demonstrated moderate discriminative ability with an AUC of 0.76 (95% CI 0.68-0.84). This performance is lower than its retrospective internal validation (AUC 0.819, Chapter 5)[156], but shows an improvement compared to an independent external validation in an Australian maternity cohort (AUC 0.694, Chapter 6). Calibration of predictions remained good, suggesting that the model's risk

estimates were generally well-aligned with observed incidence after minor shifts. The model demonstrated a high specificity (~95%) but low sensitivity (~37%) at the pre-specified risk threshold. In effect, the tool identified about one in three future GDM cases by 12 weeks while mistakenly identifying only ~5% of non-GDM women. This trade-off resulted in 1 in 5 GDM cases being diagnosed and managed roughly 10-12 weeks earlier than usual. These findings provide valuable real-world evidence that such models can work in practice, though with below ideal performance and with important considerations for clinical utility.

Previous retrospective studies have reported higher discrimination (AUCs of 0.80–0.90) for similar ML models when temporally validated, but prospective validations are rare[94,166,168,178]. A recent external validation of several published GDM prediction models found that most exhibited moderate performance (AUCs ~0.70) in new populations[121], which is in line with the AUC ~0.76 I observed.

A notable observation was the prevalence of GDM in this study's cohort (22.2%) was nearly double that of the model's training dataset (11.6%). Several factors might explain this. Firstly, this study in effect employed universal OGTT screening for all participants who reached 24-28 weeks, which is not currently standard practice in Ireland. This may have identified cases that might otherwise go undetected under risk-factor-based screening protocols[23]. Secondly, volunteer bias could have played a role; women with known risk factors for GDM (thus already being tested for GDM)[279], or a heightened awareness of their health[280], might have been more inclined to participate, potentially skewing the cohort towards a higher baseline GDM risk. The higher baseline risk can affect a model's apparent performance; for instance, positive predictive value naturally increased (because GDM was more common), but sensitivity at a fixed threshold decreased (because many more women just above the risk threshold turned out to have GDM). These factors likely contributed to the decline from AUC 0.81 to 0.76 in this validation. It reinforces that a model may need recalibration or threshold adjustment when applied to a population with different risk profiles[119].

This study's primary contribution is not the evaluation of the model's impact on clinical outcomes, which would require a large-scale randomised controlled trial. Rather, the impact lies in its demonstration of clinical feasibility and its quantification of a shift in the diagnostic timeline. First, this study confirmed the feasibility of integrating a real-time, risk prediction tool into a busy antenatal clinic workflow. Data for the 9-variable model was successfully extracted from the EHR at the point of care, and a risk score was generated without disrupting clinical practice. Second, the study quantifies the potential for earlier diagnosis. At the chosen threshold, the model identified 19 true positive GDM cases. Of these, 10 women were

diagnosed via the early OGTT at 16 weeks (Table 7.1). This represents 19% (10/52) of all GDM cases in this cohort being diagnosed and referred to the specialist diabetes care team 10–12 weeks earlier than standard practice. This "impact" is therefore the creation of a window for earlier intervention, the benefits of which are supported by trials such as TOBOGM[281].

The model's performance, with a high specificity (95%) but modest sensitivity (37%), suggests its utility as a "rule-in" rather than a "rule-out" tool for early GDM detection[84,282]. While it missed a substantial proportion of GDM cases (33 FNs vs. 19 TPs, 63.5% FNR), its low false-positive rate (10 FPs, 5.5% FPR) is an improvement over many traditional risk-factor scoring systems, which often have sensitivities in the 20-40% range but lower specificity[83,282]. The FNR (63.5%) observed is a direct consequence of the high risk threshold chosen for this feasibility study. This threshold was deliberately set to identify only a small, high-priority group (~10% of participants) for early testing, thereby minimising the burden of false positives on clinic resources. This FNR is not a fixed property of the model, but rather a point on a trade-off spectrum. As the Decision Curve Analysis (Figure 7.3) illustrates, this threshold can be adjusted. If clinicians wished to reduce the FNR (i.e., miss fewer cases), they could lower the risk threshold (e.g., from 10% to 5%). This would correctly identify more true positives (increasing sensitivity) but would come at the cost of also identifying more false positives (decreasing specificity), thereby sending more women for unnecessary early testing. The optimal threshold is ultimately a clinical and economic decision, balancing the benefit of early detection against the cost and burden of over-testing. The ML tool was efficient in identifying a smaller group of high-risk women with a relatively high yield (PPV of 66% where 19 of 29 identified as high-risk and tested were true positives). This has practical value; resources for early testing or preventive measures could be concentrated on a small high-risk group with relatively high yield (PPV ~66% in this case). Decision curve analysis also supported a net benefit for using the model within a reasonable threshold range, indicating clinical value in terms of “treated” (early tested) vs “untreated” trade-offs.

Implementing this early prediction tool in practice would mean that roughly 1 in 8 pregnant women undergo an OGTT 10–12 weeks earlier than usual. In this study, 10 women (4% of the cohort) were found to have GDM by 16 weeks and could start management in the late first or early second trimester. Prior evidence suggests that earlier treatment of GDM might improve some outcomes. For instance, the recent TOBOGM trial showed a reduction in macrosomia and neonatal respiratory distress with early vs standard GDM treatment (-5.6%, with a 95% CI of -10.1% to -1.2%)[48]. However, it's important to consider potential behavioural changes. Women informed of a high-risk status who then test negative on an early OGTT, or

those classified as low-risk by the model, might alter their behaviour or become complacent, potentially affecting the results of the standard 24-28 week OGTT[283]. While all participants were counselled that an early negative test did not preclude later GDM, this psychological impact warrants further investigation.

Resource implications are also a factor. Offering early OGTTs to 1 in 8 women represents a manageable increase in testing burden for the hospital. However, scaling this would depend on local laboratory capacity. While this study did not assess cost-effectiveness, earlier diagnosis and treatment might yield downstream cost savings by preventing complications, as suggested by other economic evaluations of early GDM treatment which indicated a cost saving of $1,373 per patient (95% CI -$3,749 to $642)[281]. In fact, diagnosing women before 14 weeks of gestation was more effective and tended to be less costly (~$5,500 savings). Still, these benefits must be balanced against the inconvenience and discomfort of an extra OGTT for those identified but ultimately not GDM[284].

This study demonstrated the feasibility of deploying an AI tool within a clinical workflow. The tool functioned as a decision support for scheduling an investigation, rather than a directive for therapy, which likely aided acceptability among clinicians. This aligns with a human-in-the-loop approach advocated in early AI implementations[275]. As actual risk scores were not shown to providers, I did not evaluate the impact of interpretability. Future deployments could explore providing explanations for high-risk identifications (e.g., key contributing factors from the logistic regression model), which may enhance trust and adoption, consistent with FUTURE-AI principles of transparency.[142].

The study has several limitations. First, as stated in the methods, the sample size was relatively small (n=235 analysable), with only 52 GDM cases. This yields imprecise estimates for some performance measures (e.g. the 95% CI on sensitivity was ±12%). The study was intended as an initial pilot implementation; a larger sample or multi-site trial would be needed to confirm findings and better estimate effectiveness. Second, the single-centre nature may limit generalisability. Although the hospital serves a diverse urban population, the results may not directly translate to other settings with different demographics or clinical practices. External validation in other centres (beyond the Australian cohort in Chapter 6) would strengthen evidence of generalisability. Third, there is potential selection bias: the women who volunteered for the study might differ from those who declined. For example, the slightly elevated GDM prevalence suggests volunteers may have had more risk factors or interest in their pregnancy health. This could exaggerate the model's PPV, since high-risk identifications were more likely to truly be GDM in a higher-risk group. A comparison of basic data between

participants and non-participants wasn't possible, but this volunteer effect should be acknowledged. Fourth, about 21% of enrolled women did not complete the OGTT (Table 7.1); while I excluded them from primary analysis, if their outcomes systematically differed, that could bias results. Encouragingly, including them as negatives in a sensitivity analysis did not markedly change the AUC or calibration (Table 7.3). Because the 'best-case' and 'neutral' scenarios show performance very similar to our main finding (AUC ~0.76), we can be reasonably confident that the missing data did not artificially inflate the results. The thresholds set at the start of the study is another limitation. These were based on the prevalence of the original dataset and with a view to decreasing false positives. However, the rate of GDM was higher than anticipated and thus the thresholds may need to be revised for future studies. Finally, this study was not designed to evaluate improvements in clinical outcomes or cost-effectiveness directly. While we can infer potential benefits from earlier detection, robust evidence for improved maternal or neonatal outcomes would require a specifically designed randomised controlled trial.

These findings paves the way for further refinement and testing of GDM prediction models. Next steps include updating the model with the new prospective data. For instance, using the combined retrospective and prospective data to re-calibrate or even re-train the algorithm. According to PROBAST-AI considerations, model updating can reduce bias when applying a model in a shifted setting[102]. Further, the exploration of whether incorporating simple recalibration (adjusting intercept/slope) or adding new predictors (biochemical or imaging) can improve sensitivity without sacrificing specificity is warranted. Previous research has typically demonstrated that biochemical features, such as FPG or HbA_{1c} taken early in pregnancy can be predictive of GDM[162,285,286]. Additionally, expanding to a multi-centre study would be valuable to ensure the model's fairness and universality. From a clinical trial perspective, a randomised controlled trial (RCT) is also warranted. Such an RCT could, for example, randomise women identified as high-risk by the ML model to an immediate early intervention pathway versus standard care and compare maternal and neonatal outcomes. These results suggest that about a third of GDM cases might be identified early using this tool; an RCT would determine if acting on this early information leads to tangible clinical benefits.

7.5 CONCLUSION

In summary, real-time use of a first-trimester ML model for GDM prediction was achieved in routine care, yielding moderate predictive performance and enabling a subset of

GDM cases to be identified and managed earlier in pregnancy. The tool was successfully integrated with clinical workflow, and provided decision support (early testing) that appears beneficial without adverse effects. The model's performance declined compared to development, highlighting the importance of prospective validation and possibly model recalibration. While not a standalone replacement for standard screening, the ML tool can augment current practice by stratifying risk at booking, which could personalise care (e.g. early interventions for high-risk patients). Further research, including larger trials and external validations, is warranted to investigate whether clinical benefits exist and to refine the model for improved sensitivity. This study underscores that trustworthy and deployable AI in obstetrics is feasible, provided it is evaluated transparently and rigorously. It also paves the way for integrating predictive analytics into prenatal care to ultimately improve maternal and offspring health outcomes.

Chapter 8

Summary, General Discussion and Future Recommendations

8.1 OVERVIEW OF THESIS AND KEY CONTRIBUTIONS

This thesis aimed to comprehensively develop, validate, and assess the clinical feasibility of ML models for early GDM prediction, leveraging both current first-trimester and historical pregnancy data. The research process took a structured approach, starting with a qualitative and quantitative assessment of the current state-of-the-art relating to ML and GDM risk prediction models (Chapter 2). This review of the quality concerns, bias concerns and methodological weaknesses across studies could be taken into consideration when designing ML based risk prediction tools. This understanding laid the grounds for a critical examination of the quality of the data itself, with detailed description of data sources, data processing methodologies and assumptions detailed in Chapter 3. Subsequently, I progressed to ensuring the validity of the recorded GDM status in the EHRs (Chapter 4). Research then focused on the development of ML models using the processed data sources, using routine clinical EHR data in different cohorts within the EHRs. In particular, I examined the impact of including data from previous pregnancies and building preconception models (Chapter 5). The empirical work concluded in multi-faceted validation of these models, employing an innovative Reciprocal External Validation (REV) framework (Chapter 6) and a real-world prospective clinical validation study (Chapter 7). The overarching aim of this thesis remained at its core throughout, to guide the journey from code to clinic, ensuring models move beyond computation to real-world care.

The structured progression of this thesis, moving from assessment of the state-of-the-art, to developing ML models in maternal care and finally to validation in settings that attempt to mirror real-world clinical use, offers a narrative of the clinical ML development lifecycle. Whilst these models still require robust external validation, the research has made several primary contributions to the field:

Comprehensive Evaluation of ML for GDM: This work provides a contemporary and thorough evaluation of ML models for early GDM prediction using routine EHR data. The systematic review and meta-analysis (Chapter 2) identified important performance benchmarks, highlighted the extreme heterogeneity in reported outcomes and highlighted persistent methodological flaws within the existing literature.

Emphasis on Dat Quality and Label Integrity: An important consideration of this thesis is the emphasis placed on the impact of data quality and label validity within EHRs. Data quality issues are rarely discussed in detail when examining the existing literature, however, my experience has shown how challenging real-world datasets can be. In addition to details in

Chapter 3, the study presented in Chapter 4 aimed to demonstrate the potential impact that unreliable labels can have on ML performance, particularly when present in evaluation data.

Development & Validation of GDM Prediction Models: The research developed and internally validated Irish based GDM risk prediction models demonstrating moderate to strong performance (AUC ~0.81-0.88, Chapter 5). An important finding was the improvement in model performance achieved by incorporating data from previous pregnancies, opening the possibility for risk prediction models to operate during the preconception period or very early risk assessment in multiparous women.

Innovation in External Validation Methodology: The introduction, implementation, and appraisal of a REV framework (Chapter 6) represents an important methodological innovation. This framework offers a pragmatic and privacy-preserving solution to the pervasive challenge of data sharing that often hinders robust external validation of ML models in healthcare AI.

Bridging the Translational Gap with Prospective Clinical Validation: Finally, this thesis presents a prospective clinical validation study (Chapter 7). Such studies, which evaluate ML models in live clinical workflows, are relatively rare yet important for assessing real-world performance, safety, challenges and integration capabilities. This component of the research provided insights into the practicalities of deploying AI tools at the point of care and demonstrated the potential for earlier GDM diagnosis.

Taken together, this thesis contributes to the growing literature on the application of ML for the prediction of GDM and offers some practical lessons for researchers, clinicians, and policymakers. This body of work suggests that algorithmic advances alone are not enough for clinical impact; meaningful progress also requires careful attention to data quality, step-by-step validation, and a clear view of potential value to patients.

8.2 SYNTHESIS AND DISCUSSION OF MAIN FINDINGS

The research presented in this thesis unfolds through a sequence of studies, each building upon the previous, to provide a progressively deeper understanding of the potential and challenges of using ML for early GDM prediction.

8.2.1 Evidence Base: Systematic Review and Meta-Analysis (Chapter 2)

The systematic review and meta-analysis presented in Chapter 2, encompassing 38 studies, 122 models and over 2 million pregnancies, established a baseline for early pregnancy

ML models using routine EHR data. The primary finding was that these models achieve, on average, moderate discriminative performance, with an AUC of 0.75 (95% CI 0.71-0.78). An AUC in this range is generally considered to offer acceptable discriminative ability, meaning the model has a 75% chance of correctly assigning a higher risk score to a woman who will develop GDM compared to one who will not. This level of performance aligns with findings from other systematic reviews, such as those by van Eekhout et al.[195] and Huang et al.[196], who reported median or pooled AUCs of 0.71 and 0.77, respectively, for first-trimester models using similar clinical predictors. However, this overall figure was significantly qualified by the observation of extreme heterogeneity across studies (I^2 99.6%), with a very wide 95% prediction interval (0.45-0.92) for the AUC. This variability suggests that the performance of any given ML model for GDM prediction can differ markedly depending on the specific clinical setting, patient population characteristics, precise feature sets utilized, and the chosen ML algorithm and its implementation.

A noteworthy finding from this meta-analysis was the lack of a clear statistically significant advantage of more complex ML algorithms (e.g. bagging or boosting ensembles) over simpler, traditional LR models when applied to routine EHR data for GDM prediction. While descriptive AUCs were slightly higher for boosting models (0.79) compared to linear models including LR (0.73), these differences were not statistically significant. This finding is contrary to previous recent meta-analyses in early GDM prediction[83,196] which have generally found non-LR models to outperform LR. However, one of the studies was assessed descriptively based on 5 models[83], while the other drew that conclusion based on 2/3 models favouring non-LR[196]. In fact, my findings align with a larger meta-analysis[197] which pooled diverse clinical prediction models and found no consistent advantage of ensemble algorithms over LR once sample size and study quality were taken into account. These findings are not insignificant, as more complex ML models typically require greater computational resources, take longer to train and have inherently less intuitive explainability[198]. The implication is that simpler, more interpretable models like LR, if well-specified and trained on high-quality data, can be just as effective for this specific predictive task.

This review also highlighted issues concerning the methodological rigour of existing literature. A high risk of bias was identified in 63% of studies, primarily arising within the 'Analysis' domain of the PROBAST+AI[102] tool. Upon reflection, even this may underrepresent the problem. van Eekhout et al.[195] and Huang et al.[196] reported 86% and 100% high risk of bias respectively, almost exclusively resulting from the Analysis domain. Common issues included the use of non-independent validation sets, lack of detail concerning missing data, inadequately

justified small datasets (which increases the risk of overfitting), and a frequent neglect of calibration assessment. These concerns are not restricted to models in GDM prediction. A recent systematic review assessed the methodological quality of prediction models using ML techniques and found 87% at high risk of bias, with the Analysis domain again identified as a key area of concern[103]. The insights gained from this review were instrumental in informing the methodological approach of the current thesis, emphasising the need for rigorous validation, careful attention to data quality, and transparent reporting to address these identified gaps in the field.

8.2.2 Data Preparation, Processing and Label Accuracy (Chapters 3 & 4)

Chapter 3 detailed the comprehensive process of sourcing, cleaning, coding, and curating a large retrospective EHR dataset from The Coombe Hospital for this thesis. This process underscored the inherent complexities and challenges of working with real-world clinical data, which are primarily collected for patient care rather than research. Issues such as missing data for key variables, inconsistencies in data entry, the need to transform categorical features, and the requirement for careful feature engineering were all encountered and addressed. This work formed the foundation on which the subsequent chapters could be developed.

Building upon this, Chapter 4 investigated the accuracy of GDM diagnoses as recorded in the EHRs, comparing them against a maintained database from the hospital's specialist clinical diabetes team (CTD), which served as the reference standard. This validation revealed notable discrepancies: while 3,388 GDM cases were identified in both databases, 564 cases present in the CTD lacked a GDM label in the EHRs, and 771 cases labelled as GDM in the EHRs were not found in the CTD records. Overall, the dataset comprised 87.5% true negatives, 9.0% true positives, 2.0% false positives, and 1.5% false negatives when comparing EHR to CTD for the 2018-2022 period (excluding 2020). The year 2020 was excluded as I found Covid-19 protocols[219] had disrupted routine GDM screening practices from March 2020 until at least September 2020, leading to fluctuations in recorded GDM prevalence and further complicating data consistency and interpretation.

When an LR model was trained using the "raw" EHR-GDM labels (containing the identified noise), its discriminative performance (AUC 0.817) was virtually identical to a model trained on "validated" VAL-GDM labels (AUC 0.817), where only matching EHR and CTD labels were used. This suggested that, at the levels of noise present in this dataset, the

training process itself was relatively resilient. However, the picture changed when evaluating the model. If the model trained on validated labels was then tested against a set with noisy EHR-GDM labels, its apparent performance, particularly the AP score, was notably degraded (AP 0.395 with noisy test labels vs. 0.450 with validated test labels). This implies that while models might learn adequately from slightly imperfect training data if the underlying signal is strong, their true performance can be significantly misrepresented if the benchmark against which they are evaluated (the test set labels) is itself flawed. Simulations involving the introduction of progressively higher levels of random label noise confirmed that model performance (both AUC and average precision) degrades as noise increases, with a particularly detrimental effect of false positives in the test set on AUC. While the impact of the existing low-level noise on training was minimal here, the study underscores the importance of quantifying label noise and understanding its potential impact, especially for robust model evaluation.

8.2.3 Model Development & Evaluation: Early & Preconception Prediction (Chapter 5)

Chapter 5 focused on the development and internal validation of ML models for early GDM prediction using the validated EHR dataset. Models developed using only first-trimester data from the current pregnancy achieved good discriminative performance, with an AUC of approximately ~0.82. Consistent with the findings of the meta-analysis (Chapter 2), LR performed comparably to more complex algorithms like XGBoost and Explainable Boosting Machines (EBM) for this task (e.g., LR AUC 0.819, XGBoost AUC 0.818). Furthermore, models trained specifically on nulliparous women maintained good performance (AUC ~0.81).

A key finding from this chapter was the significant improvement in predictive performance when incorporating data from women's previous pregnancies. These multiparous models, which included data from past pregnancies in addition to first-trimester data, achieved the highest AUCs (~0.88), despite having a smaller population to model (4,005 as opposed to 27,561). Notably, models using data from past pregnancies alone (i.e., for preconception risk assessment or very early assessment in a new pregnancy) achieved good discrimination (AUC ~0.86). This highlights the information contained within a woman's obstetric history could offer a promising avenue for identifying high-risk individuals even before conception or at the very earliest stages of a subsequent pregnancy. This aligns with evidence calling for earlier interventions during the preconception period to help prevent GDM[15–17].

However, the study also observed variations in model performance across different ethnic subgroups, with lower AUCs for Black African, Asian and Southeast Asian patients, for instance. This finding underscores the need to evaluate and address model fairness and to ensure that predictive tools do not exacerbate existing health disparities[287], a key consideration in guidelines like TRIPOD+AI[101]. The similar performance of LR with more complex models for first-trimester data, suggest that for this specific predictive task using routine EHR variables, simpler, more interpretable models may be sufficient and potentially preferable for clinical translation.

8.2.4 From Code to Clinic: External and Prospective Validation (Chapters 6 & 7)

The process of taking an ML model from development to potential clinical use requires rigorous and multifaceted validation. Recent guideline papers have described how to undertake internal validation[77], external validation[91,120] and early-stage clinical evaluation[275] in models driven by AI. I recognised the importance of seeking external data early on, which led to my letter to the editor detailing the struggles attaining data that is supposedly 'available on request'. I soon figured out that those requests are soundly ignored (Appendix E). However, a research team in Australia did respond, opening the avenue for the explorations in Chapter 6.

Chapter 6 introduced and implemented a novel REV framework. This approach was conceived to address the pervasive challenge of data sharing that often hinder external validation of clinical prediction models[261,288,289]. By exchanging pre-trained models and their processing pipelines, rather than patient-level data, my first-trimester LR model was evaluated on an Australian cohort and their model was evaluated on my cohort. The REV resulted in a decrease in AUC for both models, my model from 0.82 to 0.69 and the Australian model from 0.83 to 0.77. Calibration was also notably impaired for both models in the external setting, highlighting how models ideally need to be recalibrated in new settings[78,290]. These findings illustrate the "validation gap" or optimism bias frequently discussed in prediction model literature[123,291], where performance in new, unseen populations is often considerably lower than in the development cohort. The REV framework itself, however, proved to be a viable and valuable method for conducting external validation, offering a potential pathway to more widespread and realistic benchmarking of clinical AI models.

Chapter 7 presents the findings of a prospective clinical evaluation of the first-trimester LR model, conducted at the Coombe Hospital. This study, aligning with early-stage clinical evaluation principles like those in the DECIDE-AI framework[275], aimed to assess the model's

performance and usability when integrated into a live clinical workflow. The model demonstrated moderate discriminative ability with an AUC of 0.762 and acceptable calibration (intercept -0.045, slope 0.808). A key practical outcome was the model's ability to facilitate earlier GDM diagnosis. By identifying approximately 10% of participants as high-risk for an optional early OGTT (around 16 weeks), 19 out of 52 GDM cases, representing approximately 19% of all GDM cases, were diagnosed and could be managed 10-12 weeks sooner than via standard 24–28 week screening. The consistent decline in first-trimester model performance from internal validation (Chapter 5) to external validation (Chapter 6) and then to prospective clinical validation (Chapter 7) remains a key finding of this thesis. The findings are summarised in Table 8.1.

Table 8.1. Comparison of first-trimester GDM prediction model performance across validation stages.

Performance Stage	Dataset	AUC (95% CI)	Calibration Slope	Calibration Intercept
Internal Validation (Chapter 5)	Irish – Coombe EHRs	0.819 (0.811-0.827)	1.010	0.013
External Validation (Chapter 6)	Australian – Monash Cohort	0.694 (0.688-0.701)	0.550	0.170
Prospective Validation (Chapter 7)	Irish - Coombe Prospective	0.762 (0.681-0.837)	0.808	-0.045

The implications of these findings reinforce four key messages. First, they reinforce the need for rigorous, multi-stage validation and evaluation of GDM risk prediction tools. Second, frameworks like REV offer a valuable, timely and efficient mechanism to facilitate external validation of models where data sharing is often challenging. Third, prospective studies, despite their resource requirements, play an important role not only to evaluate the performance of models in real-time, but also to explore aspects like clinical utility, workflow integration, user acceptance and impact on patient pathways and outcomes. Finally, the observed miscalibration in external settings suggests that strategies for model recalibration or local adaptation will likely be necessary for the widespread and equitable deployment of GDM prediction models across diverse clinical environments[119].

8.3 ADDRESSING THE RESEARCH QUESTIONS

This thesis systematically addressed its predefined research questions, resulting in an evaluation of the central hypothesis.

RQ1. **Evidence Base.** What is the overall accuracy, methodological quality, and heterogeneity of existing ML models for the early prediction of GDM using EHR data?

The systematic review (Chapter 2) indicated that existing models achieve moderate pooled AUC (0.75) but with high heterogeneity (I^2 99.6%) and high risk of bias in 63% of studies. No clear superiority of complex algorithms over LR was found for routine EHR data.

RQ2. **Label Noise**. To what extent are the data contained in the electronic health records accurate, and if there is label noise, to what extent could this impact machine learning modelling of gestational diabetes mellitus? (Chapter 4)

The findings revealed that GDM diagnoses in the studied EHRs exhibited inaccuracies when compared to a clinical team database. In this specific dataset, the level of noise (affecting <5% of the cohort used for modelling) had a negligible impact on the *training* of the logistic regression model. However, the presence of noise in the *evaluation* dataset was shown to more significantly affect performance metrics, particularly average precision. Furthermore, simulated increases in label noise demonstrated a clear degradation in model performance, underscoring the importance of data quality.

RQ3. **Early Prediction**. How accurately can machine learning models, when applied to electronic health records, predict the diagnosis of gestational diabetes mellitus? (Chapter 2, 5 & 7)

The models developed in this thesis (Chapter 5) demonstrated good performance in internal validation, with AUCs around 0.82 for first-trimester data and up to 0.88 when incorporating data from previous pregnancies. The prospective clinical validation (Chapter 7) provided a real-world performance estimate of AUC 0.762 for the first-trimester model. Thus, ML models show potential to predict GDM early, albeit with performance that attenuates in more stringent validation settings.

RQ4. **Multiparous Analysis**. How accurately can machine learning models leverage data from a woman's previous pregnancies to predict outcomes in subsequent pregnancies? (Chapter 5)

The results from Chapter 5 demonstrated that incorporating data from previous pregnancies improves GDM prediction performance, with models achieving AUCs up to 0.88. Key features from the past pregnancy, such as birthweight percentile, and calculating interpregnancy weight gain offer enhanced predictive performance over first-trimester data alone. Further, using past pregnancy data alone opens up the possibility of preconception risk prediction.

RQ5. **External Validation**. Can we overcome traditional data sharing challenges when working with sensitive data, such as patient electronic health records, and perform external validation of base models? (Chapter 6)

The REV framework presented in Chapter 6 successfully enabled the external validation of GDM prediction models between Irish and Australian research groups without the need for direct patient-level data sharing. While the validation highlighted significant transportability issues due to population and screening differences, the methodology itself proved a viable solution to data access barriers.

RQ6. **Implementation**. Can we assess the deployment validity of the model in a clinical setting, does the model maintain predictive performance when deployed, and can it detect GDM earlier than current diagnosis? (Chapter 7)

The study in Chapter 7 confirmed that the ML model could be successfully deployed in a real-time clinical setting. Its predictive performance (AUC 0.762) was lower than in internal validation but still demonstrated potential utility. Importantly, the model enabled the diagnosis of 1 in 5 GDM cases approximately 10-12 weeks earlier than standard screening pathways.

Central Research Question:

Can utilising machine learning techniques on electronic health records provide useful early predictions for gestational diabetes mellitus.

The findings of this thesis offer qualified support for this objective. Useful early predictions were demonstrably achieved, with models identifying women at risk of GDM well before standard screening. The earlier diagnosis of a subset of GDM cases in the prospective study provides a direct pathway to potentially enhancing early interventions. The potential to

influence diagnostic/screening practices is evident, as such tools could support more targeted or risk-stratified approaches. However, direct evidence that these models improve maternal outcomes is not provided, as this would require dedicated intervention trials (e.g., RCTs). The literature suggests early GDM intervention can be beneficial, and this thesis provides tools that could facilitate such earlier intervention. The journey through the chapters reveals that while ML can provide these useful predictions, the path to reliable and impactful clinical tools is complex, heavily dependent on data quality, rigorous multi-stage validation, and careful consideration of the implementation context. The research questions, addressed sequentially, effectively map out a lifecycle for the development and evaluation of clinical prediction models, from understanding the existing evidence and data landscape through to real-world testing, aligning with the principles espoused in guidelines such as TRIPOD+AI and evaluation frameworks like DECIDE-AI.

8.4 STRENGTHS AND NOVELTY OF THE THESIS

This thesis has several notable strengths and elements of novelty in the context of predicting GDM from routine electronic health records EHRs. A primary strength of this thesis lies in its adoption of a comprehensive multi-stage research lifecycle. The thesis did not confine itself to a single facet of ML model development but instead navigated the entire clinical ML development lifecycle: from a broad, critical assessment of the existing literature (Chapter 2), through data acquisition, cleaning, and an investigation into outcome label integrity (Chapters 3 and 4), into model development leveraging both first-trimester and prior pregnancy history (Chapter 5), and concluding in rigorous validation stages. These validations included an innovative approach to external validation (Chapter 6) and, importantly, a prospective clinical validation (Chapter 7). Such an end-to-end methodological arc, encompassing systematic review, data quality analysis, original model development, and robust multi-modal validation within a single cohesive body of work, is a distinguishing feature of this research. The aim pursued here was to provide a more holistic and grounded perspective on the subject, whilst aligning with best practice recommendations at each stage[77,91,101,102,146,275,292].

The focus on data quality and the practical realities of model validation is another significant strength. Real-world EHR data are notoriously complex, and issues such as poor data quality and inaccurate outcome labels represent major impediments to reliable model development[231]. This thesis dedicated substantial effort to these foundational aspects. Chapter 3 detailed the processes of data sourcing, cleaning, coding, and curation. Chapter 4 provided

an examination of GDM label accuracy and the impact of label noise. This proactive approach to data quality aligns with contemporary guidance underscoring the critical importance of meticulous data assessment in clinical prediction modelling[88,231].

The systematic review and meta-analysis (Chapter 2) was the largest and most contemporary synthesis of evidence on ML for early GDM prediction using routine EHR data at its undertaking. The findings highlighted performance benchmarks but also extreme heterogeneity and methodological weaknesses in existing literature, informing both this thesis and the broader research community. The Chapter was also developed in accordance with the latest guidelines for SRMA[146], risk of bias assessment[102] and meta-analytic modelling[150,151] for clinical risk prediction models using AI.

The development and successful implementation of the REV framework (Chapter 6) stands as a key methodological innovation. In a field where data sharing for external validation is a major impediment due to privacy and governance concerns[90,119], REV offers a pragmatic and privacy-preserving solution. This framework has the potential to be generalised to other clinical prediction tasks using structured data, thereby facilitating more widespread and realistic benchmarking of AI models in healthcare[91,101].

Furthermore, the prospective clinical validation study (Chapter 7) is a notable strength and novelty. Such studies, which evaluate ML models within live clinical workflows, are relatively rare yet important for understanding actual performance, safety, and integration challenges. This component of the research provided invaluable insights into the practicalities of deploying AI tools at the point of care and demonstrated the tangible potential for earlier GDM diagnosis.

Finally, this thesis also contributes by focusing on the Irish healthcare context, developing and validating GDM prediction models using EHR data from a large Irish maternity hospital. Given that much of the existing research in this area originates from Asian populations (Chapter 2), this work helps to address a geographical gap and provides insights relevant to European healthcare systems with similar demographic profiles or screening practices. Chapter 6 further highlighted this point, demonstrating how model performance declines when transported across continents.

8.5 LIMITATIONS OF THE THESIS

While this thesis offers significant contributions, it is important to acknowledge its limitations. Transparency regarding these constraints is crucial not only for academic integrity but also for contextualizing the findings and guiding future research endeavours[198,293].

A primary limitation is the single-centre focus for the primary dataset used in model development[294]. The main GDM prediction models (Chapter 5) were developed using EHR data exclusively from The Coombe Hospital. As highlighted in prognostic research literature, models developed and validated within a single institution may not generalise effectively to other patient populations or healthcare settings due to differences in demographics, GDM prevalence, clinical practices, and data systems[119,207]. The observed performance decline during external (Chapter 6) and prospective validation (Chapter 7) illustrates this "single-centre bottleneck" and how models can become overfit to their development environment, requiring local validation and potential recalibration before wider deployment[119,290].

The reliance on retrospective data collection for the main EHR dataset (Chapters 3, 4, and 5) constitutes another limitation[295]. Retrospective data, not originally gathered for research, can be prone to biases like selection bias and unmeasured confounding[296]. While Chapter 7 involved prospective validation, the initial model architecture was derived from these retrospective data. As PROBAST+AI[102] guidance suggests, and as experienced in this thesis (Chapter 3 vs. Chapter 7 data collection confidence), prospective data collection is generally preferred for developing robust prediction models due to higher data quality and control over variable measurement. Despite extensive cleaning (Chapter 3), residual data quality issues from the retrospective data could impact model robustness and generalizability.

The GDM screening protocol in Ireland, which is predominantly risk-factor based rather than universal, presents another limitation. With estimates suggesting that 40-50% of women may not be screened and up to 16% of GDM cases might be missed[23], the GDM outcome label, even when validated against the CTD, may not capture all true GDM cases. This "hidden GDM" could reduce true predictor effects, impact model calibration, and affect the generalisability of the models to settings with universal screening. This “hidden GDM” was somewhat apparent in Chapter 7, which saw almost double the rate of GDM when what was effectively universal screening was applied in the Coombe Hospital. However, recruitment bias also likely played a role. This issue highlights how ML models must deal with systemic healthcare practices, where the ground truth itself is influenced by the diagnostic process.

The prospective clinical validation study (Chapter 7), while a strength, had a relatively small sample size (n=235, 52 GDM cases), which can lead to imprecise performance estimates of the model performance, particularly among minority groups[243,255,278]. Missing OGTT outcomes for ~21% of enrolled women, despite sensitivity analyses, could introduce bias if their true outcomes systematically differed. Potential selection bias among volunteers might also have influenced observed GDM prevalence and model PPV. Further, not yet having the data from the delivery of these infants restricts any analysis of whether the early identification of high-risk GDM cases results in clinical benefit.

Furthermore, the thesis acknowledges the absence of formal Patient and Public Involvement (PPI) in the design and conduct of the retrospective components (Chapters 2-6). While the prospective study (Chapter 7) involved patient consent, earlier PPI is increasingly emphasised in healthcare AI research to ensure developed tools are relevant, acceptable, and address genuine patient needs[198,297]. The lack of such involvement in initial stages is a limitation. Further, this thesis did not quantify the cost associated with the tools or the potential cost-benefit of early diagnosis, an important consideration for clinical implementation.

Finally, the handling of missing data in the main EHR dataset (Chapter 3) warrants discussion, particularly considering literature highlighting that missing data are often poorly handled and reported in prediction model studies[210,231]. This thesis employed a mixed approach of some targeted imputations alongside listwise deletion for records with missing critical predictor values. While methods like multiple imputation offer potential solutions, they also come with their own pitfalls[211,298]. The listwise deletion used for critical variables, although aiming to avoid imputation-induced bias for these key variables, is known to risk selection bias if the data are not Missing Completely at Random (MCAR)[299,300], potentially affecting the representativeness of the sample and the generalizability of model parameters.

8.6 CLINICAL IMPLICATIONS

The findings of this thesis, particularly from prospective and external validation, carry important clinical implications. While the vast majority of published prediction models never see clinical use[301], often due to a lack of a specific clinical decision-making process they could meaningfully inform or optimize[302], this research demonstrates a potential pathway towards enhancing maternal care through earlier GDM risk detection. A key clinical implication is the potential for earlier GDM risk identification and, consequently, earlier intervention. The prospective clinical validation (Chapter 7) showed the first-trimester ML model could identify

1 in 5 of women later diagnosed with GDM approximately 10-12 weeks earlier than standard screening. Current protocols typically schedule OGTTs at 24-28 weeks. The ML model enables risk assessment at the first booking visit (~12 weeks), facilitating a proactive stance. Early diagnosis allows timely interventions known to improve outcomes[48]. This earlier diagnosis could lead to more efficient resource allocation, concentrating early preventive efforts on high-risk individuals.

The performance characteristics observed in the prospective validation, moderate discrimination (AUC 0.762) with high specificity (95%) but modest sensitivity (37%) at the chosen risk threshold, suggest that the current ML tool is better suited as a "rule-in" instrument rather than a "rule-out" one. The model can potentially identify a smaller group of high-risk women who would benefit most from earlier attention or testing, while minimising the number of false positives. This targeted strategy could be particularly valuable in resource-constrained healthcare systems, allowing for more efficient allocation of resources for early GDM management. Decision-curve analysis supported the net benefit of using the model within a reasonable range of risk thresholds (Chapter 5 & 7).

However, the translational pathway is not without its challenges. The consistent finding of performance decline when models are validated externally (Chapter 6) or prospectively (Chapter 7) underscores an important reality: models developed in one specific context are unlikely to perform identically in another without adaptation. This consistent decline necessitates that healthcare institutions planning to adopt such ML tools must consider strategies for local model validation and recalibration to ensure their accuracy and reliability within the specific patient population and clinical environment[101,119,207]. The validation attrition observed is not a failure of the models per se, but rather a reflection of the inherent complexities and heterogeneities of real-world healthcare. Healthcare institutions should not simply adopt “off-the-shelf” models, because rigorous local validation and adaptation are essential for safety, effectiveness, and equity[207,290,292].

Furthermore, the significant concern with regard to the validity of the GDM label should not be understated. If the hospital screening practices align with findings from prior research, this would mean there is a systematic under diagnosis of GDM cases, resulting in further label noise. Thus, any model that would ever be deployed in a clinical setting would need to ensure that it was developed on a dataset where the outcomes are certain to prevent further bias in the model.

Finally, clinician and patient engagement is important for successful translation. Clinicians need to trust the tools, understand their capabilities and limitations, and see their

value in improving patient care. Patients need to be informed about how these tools are being used and be comfortable with the recommendations derived from them. Efforts to enhance model explainability (XAI), as touched upon in Chapter 1, will be vital in building this trust and facilitating shared decision-making. The prospective study in Chapter 7, where clinicians were aware of the research tool, represents an initial step in this engagement process and further efforts are being made to evaluate their experience.

8.7 RECOMMENDATIONS FOR FUTURE RESEARCH

Building upon the findings, insights, and limitations identified, this section outlines recommendations for future research to advance GDM prediction, enhance ML model robustness and clinical utility, and facilitate responsible translation into maternal care. Table 8.2 summarises these recommendations.

Future research must prioritise enhancing methodological rigor and ensuring transparent reporting. As highlighted by the systematic review in Chapter 2 and the broader literature[195,196,203], many existing GDM prediction model studies suffer from methodological weaknesses and incomplete reporting, undermining confidence and hindering reproducibility. Further emphasised in Chapter 2 is just how rapid the number of publications has been increasing in this area, often with high risk of bias due to poor methodological rigour. Fortunately, there has also been an increase in comprehensive reporting guidelines. Guidelines such as TRIPOD+AI[101], DECIDE-AI[275], FUTURE-AI[142], CONSORT-AI[276], SPIRIT-AI[303], TRIPOD-SRMA[146] and PROBAST+AI[102] include detailed descriptions of study population, data, preprocessing, model development, validation, and full model specification to allow independent scrutiny.

Alongside this, there must be consistent reporting of key performance metrics. As demonstrated by the systematic review (Chapter 2) and emphasised by established guidance[77,299,304], all studies should, at a minimum, report measures of discrimination (e.g., AUC with confidence intervals) to quantify the model's ability to separate patients with different outcomes[75], and calibration (e.g., calibration plots accompanied by calibration slope and intercept) to assess the agreement between predicted risks and observed event rates[77,299,304,305]. Furthermore, to evaluate clinical utility, decision curve analysis should be reported where appropriate, as it translates model output into net clinical benefit across relevant risk thresholds[77,205,206]. This thesis found these metrics frequently underreported, yet they are essential for a broader understanding of a model's potential value.

Furthermore, robust sample size justification is important to ensure the results of the model are generalisable. Many studies reviewed in Chapter 2 were at high risk of bias due to inadequately justified, often small, sample sizes, which increases the risk of overfitting and the development of unstable models[255,258,278]. Further, even when the total sample size appears adequate, there may be an underrepresentation of minority classes. As such, it also recommended that studies assess the effective sample size for sub-populations to ensure the results are not just generalisable but generalisable to all populations within the dataset[243]. A failure to take these into account could lead to further bias in models and exacerbate health inequalities[57,306].

Efforts should continue to improve model validation and generalisability. This thesis underscored the decrease in performance as models were externally and prospectively validated (Table 8.1). Thus, as has been strongly advocated in the literature[90,119,201,207], future work should aim to conduct large-scale, multi-centre external validation studies. Evaluating GDM models across diverse populations, healthcare systems, EHR platforms, and screening policies is important to assess generalisability and identify factors influencing transportability. Given observed miscalibration, it is important to investigate and apply model updating and recalibration techniques[101,119,290]. Research into efficiently adapting models to local settings will be essential. To facilitate these external validations, future work should build upon the REV approach described in Chapter 6. This approach also requires that more researchers must make both their models and their code publicly available at the time of publication, aligning with RECORD guidelines[307]. Reporting of code, models and data preprocessing pipelines is essential for the widespread external validation and recalibration of GDM prediction models.

Once robust external validation has been established, researchers should progress to early stage clinical evaluation of such models[275]. The ultimate goal of developing clinical prediction models is to improve patient outcomes and healthcare efficiency. Therefore, the next step is to move beyond predictive assessments to rigorous evaluations of clinical impact. RCTs are the gold standard for this, comparing care guided by the ML prediction tool against standard care, and measuring effects on processes of care (e.g., timing of diagnostic OGTT, uptake of interventions), maternal and neonatal health outcomes, and overall cost-effectiveness. Alongside RCTs, dedicated usability and workflow integration studies are needed. Even the most accurate model will fail to deliver benefits if it is cumbersome for clinicians to use or disrupts established clinical pathways. Finally, comprehensive health economic analyses will be essential to determine the value proposition of ML-driven GDM screening and management strategies from a healthcare system perspective.

Table 8.2. Summary of Key Recommendations for Future Research in ML for GDM Prediction.

Recommendations	Rationale/Link to Thesis Findings & Literature
Methodological Rigor & Reporting	
1. Adherence to Comprehensive Reporting Guidelines.	Rigoroulsy adhere to guidelines like TRIPOD+AI, CONSORT-AI, and DECIDE-AI. The meta-analysis (Ch 2) and literature highlight widespread reporting deficiencies, hindering appraisal and replication. This thesis attempted adherence to these guidelines whenever possible.
2. Consistent Reporting of Key Performance Metrics**.**	Report discrimination (AUC with CIs), calibration (plots, slope, intercept), and DCA. This thesis (Ch 2) and other reviews show these metrics are often underreported, yet essential for judging true model value.
3. Robust Sample Size Justification.	Employ established methods for sample size calculation for model development and validation. Many existing models suffer from high risk of bias due to inadequate sample sizes (Ch 2). Effective sample size estimations should also be considered in calculations to ensure results are generalisable to minority populations.
Model Validation & Generalizability	
4. Conduct Large-Scale, Multi-Centre External Validation Studies.	Validate GDM models across diverse populations, systems, and EHR platforms. This thesis demonstrated significant performance declines in external settings (Ch 6, Table 8.1), emphasising that single-centre validations are insufficient.
5. Investigate and Apply Model Updating and Recalibration Techniques.	Explore methods for efficiently updating/recalibrating models for local settings. Observed miscalibration in external settings (Ch 6) highlights the need for local adaptation for safe deployment.
6. Further Develop and Utilize Privacy-Preserving Validation Frameworks.	Expand on frameworks like REV or explore federated learning. Overcoming data-sharing barriers (addressed by REV in Ch 6) is key to broader external validation.
Enhancing Model Performance & Scope	
7. Investigate the Incremental Value of Additional Predictors.	Explore incorporating FPG, HbA_{1c}, detailed socioeconomic data, lifestyle factors and data from ultrasound scans. These predictors, unavailable in this thesis, are known have some predictive value for GDM and could improve model performance.

8. Further Develop and Validate Models for GDM Treatment Pathways.	Expand on preliminary work (Appendix F) to predict need for diet, metformin, or insulin. Stratifying treatment needs post-diagnosis is another area for personalized medicine using ML.
9. Extend ML Methodologies to Other Adverse Pregnancy Outcomes.	Apply learnings to preeclampsia, preterm birth, foetal growth abnormalities. The infrastructure and expertise developed (data handling, validation strategies) can be leveraged for broader impact in maternal care.
Clinical Impact & Implementation	
10. Conduct Randomized Controlled Trials (RCTs).	Design RCTs to assess if ML-driven earlier risk stratification and subsequent lifestyle intervention improve clinical outcomes and are cost-effective. While models predict risk, RCTs are needed to prove tangible health benefits from acting on these predictions (Ch 7.4, 8.3).
11. Improve Data Infrastructure and Interoperability.	Better data infrastructure is foundational for robust and generalizable AI tools.
12. Investigate Human Factors, Implementation Science, and Ethical Considerations.	Investigate clinician/patient acceptability, usability, workflow integration, algorithmic bias, fairness, and PPI. Successful AI translation depends heavily on these non-algorithmic factors. This thesis noted limited PPI in retrospective stages.

While the models developed using routine EHR data demonstrated good performance, there is always scope for improvement. Future studies should investigate the incremental value of incorporating novel predictors not extensively available for this thesis. This could include early pregnancy biomarkers like fasting plasma glucose or HbA_{1c}, and also data from dating or anatomy scans. Typically, women attend clinics for a dating scan around the same time as their booking visit, in addition to basic laboratory tests. A joined-up data system could make use of EHRs, laboratory results and data extracted from dating scans. Further, data taken from anatomy scans, usually ~20 weeks of gestation, may further refine these models[308,309]. The advantage with these predictors is that they are still routinely collected data, which means there is no additional data collecting burden on healthcare providers. While this thesis found simpler models often suffice, the availability of richer, multi-modal datasets might warrant further exploration of advanced ML techniques, provided that a strong emphasis is maintained on explainability to ensure clinical acceptance.

Beyond predicting GDM itself, the ML framework developed here could be extended to predict other GDM-related outcomes. One direction, initiated in Appendix I of this thesis, is the development of models to predict GDM treatment pathways, identifying which women are likely to manage GDM with lifestyle changes alone versus those who will require pharmacological interventions like metformin or insulin. Another avenue for further exploration is the prediction of the metabolic burden reflected in the OGTT itself. By reframing the OGTT as a regression task rather than a binary threshold, we can estimate each woman's full glucose curve. The benefit of this approach would be two-fold: first, it may help identify those at most need of intervention early in pregnancy, and second, the models would be transportable across different GDM diagnostic criteria as it would be predicting the OGTT result in terms of $mmol.L^{-1}$. A third strand of work already piloted in Appendix J uses the same feature set to anticipate downstream obstetric outcomes tightly linked to hyperglycaemia, including small-for-gestational-age (SGA) and large-for-gestational-age (LGA) infants, macrosomia, low birth-weight and pre-term birth. Collectively, these extensions would transform the current single-endpoint model into a suite of algorithms that 1) guide intensity of antenatal surveillance, 2) tailor pharmacological escalation, and 3) identify pregnancies needing neonatal support, thereby offering a comprehensive decision-support ecosystem around hyperglycaemia in pregnancy.

8.8 CONCLUDING REMARKS

The research journey, from synthesising existing evidence and preparing local data to developing models and subjecting them to a rigorous multi-stage validation process including a novel external validation method and a prospective clinical study, has yielded valuable insights. The findings indicate that prognostic models, particularly those for first-trimester clinical data, demonstrate promise for identifying women at an elevated risk of GDM earlier than current standard screening practices. This early identification capability holds the potential to facilitate timely interventions, which may ultimately improve maternal and neonatal outcomes.

However, the pathway from a promising algorithm to a clinically impactful and responsibly deployed tool is complex and fraught with challenges. This research underscores that the clinical translation of these predictive tools is dependent on data quality, robust and transparent multi-stage validation (internal, external, and prospective), careful consideration of model transportability across diverse populations and healthcare settings, and a nuanced understanding of the clinical context in which these tools will be used. The consistent observation of performance decline as models were tested in increasingly realistic settings serves as an important reminder of the optimism bias inherent in much of the initial prognostic modelling literature and the need for rigorous evaluation.

Ultimately, this thesis provides not only a set of GDM prediction models for a specific context but also a broader framework and a narrative of the research lifecycle for developing and evaluating clinical prediction tools in maternal healthcare. This thesis highlights both the potential of data-driven approaches and the responsibilities that accompany their development and deployment. The findings and methodologies presented herein will inform future efforts to refine, validate, and ethically integrate these technologies into clinical practice, with the overarching goal of enhancing the health and wellbeing of mothers and their infants. The journey from code to clinic requires ongoing methodological rigour, interdisciplinary collaboration, and a commitment to patient-centred care.

On a personal reflection, I started this journey with the mindset that ML techniques had moved beyond traditional statistical approaches for building prediction models, and that statistical approaches needed to adapt to enable robust model development and evaluation. However, I have ended up at the opposite conclusion. Much of the current ML research rests on small, often artificial datasets, lacks formal sample size estimation, and shows only a thin

grasp of the clinical context required for sound risk model assessment. In the end, I now see ML in healthcare as a promising yet still-emerging science with many lessons left to learn.

REFERENCES

1. International Association of Diabetes and Pregnancy Study Groups Consensus Panel. International Association of Diabetes and Pregnancy Study Groups Recommendations on the Diagnosis and Classification of Hyperglycemia in Pregnancy. *Diabetes Care* **33**, 676–682 (2010).
2. Huhn, E. A., Rossi, S. W., Hoesli, I. & Göbl, C. S. Controversies in Screening and Diagnostic Criteria for Gestational Diabetes in Early and Late Pregnancy. *Frontiers in Endocrinology* **9**, (2018).
3. O'Sullivan, E. P. *et al.* Atlantic Diabetes in Pregnancy (DIP): the prevalence and outcomes of gestational diabetes mellitus using new diagnostic criteria. *Diabetologia* **54**, 1670–1675 (2011).
4. Reece, E. A. The fetal and maternal consequences of gestational diabetes mellitus. *The Journal of Maternal-Fetal & Neonatal Medicine* **23**, 199–203 (2010).
5. Elliott, H. R., Sharp, G. C., Relton, C. L. & Lawlor, D. A. Epigenetics and gestational diabetes: a review of epigenetic epidemiology studies and their use to explore epigenetic mediation and improve prediction. *Diabetologia* **62**, 2171–2178 (2019).
6. Patti, M.-E. Intergenerational Programming of Metabolic Disease: Evidence from Human Populations and Experimental Animal Models. *Cell Mol Life Sci* **70**, 1597–1608 (2013).
7. McIntyre, H. D. *et al.* Gestational diabetes mellitus. *Nat Rev Dis Primers* **5**, 1–19 (2019).
8. O'Higgins, A., Dunne, F., Lee, B., Smith, D. & Turner, M. J. A national survey of implementation of guidelines for gestational diabetes mellitus. *Ir Med J* **107**, 231–233 (2014).
9. International Diabetes Federation. IDF GDM Model of Care. *International Diabetes Federation* (2015).
10. Gillespie, P., Cullinan, J., O'Neill, C. & Dunne, F. Modeling the Independent Effects of Gestational Diabetes Mellitus on Maternity Care and Costs. *Diabetes Care* **36**, 1111–1116 (2013).
11. Chen, Y. *et al.* Cost of gestational diabetes mellitus in the United States in 2007. *Popul Health Manag* **12**, 165–174 (2009).
12. Moran, P. S. *et al.* Economic burden of maternal morbidity – A systematic review of cost-of-illness studies. *PLOS ONE* **15**, e0227377 (2020).
13. King, H., Aubert, R. E. & Herman, W. H. Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. *Diabetes Care* **21**, 1414–1431 (1998).

14. Moholdt, T. & Hawley, J. A. Maternal Lifestyle Interventions: Targeting Preconception Health. *Trends Endocrinol Metab* **31**, 561–569 (2020).
15. Stephenson, J. *et al.* Before the beginning: nutrition and lifestyle in the preconception period and its importance for future health. *The Lancet* **391**, 1830–1841 (2018).
16. Fleming, T. P. *et al.* Origins of lifetime health around the time of conception: causes and consequences. *The Lancet* **391**, 1842–1852 (2018).
17. Barker, M. *et al.* Intervention strategies to improve nutrition and health behaviours before conception. *The Lancet* **391**, 1853–1864 (2018).
18. Lin, X., Yang, T., Zhang, X. & Wei, W. Lifestyle intervention to prevent gestational diabetes mellitus and adverse maternal outcomes among pregnant women at high risk for gestational diabetes mellitus. *J Int Med Res* **48**, 0300060520979130 (2020).
19. Brown, J. *et al.* Lifestyle interventions for the treatment of women with gestational diabetes. *Cochrane Database of Systematic Reviews* https://doi.org/10.1002/14651858.CD011970.pub2 (2017) doi:10.1002/14651858.CD011970.pub2.
20. Mottola, M. F. *et al.* 2019 Canadian guideline for physical activity throughout pregnancy. *Br J Sports Med* **52**, 1339–1346 (2018).
21. O'sullivan, J. B. & Mahan, C. M. CRITERIA FOR THE ORAL GLUCOSE TOLERANCE TEST IN PREGNANCY. *Diabetes* **13**, 278–285 (1964).
22. National Institute for Clinical Excellence. Diabetes in pregnancy: management from preconception to the postnatal period. *NICE Guidelines* **[NG3]**, (2015).
23. Avalos, G. E., Owens, L. A., Dunne, F., & for the ATLANTIC DIP Collaborators. Applying Current Screening Tools for Gestational Diabetes Mellitus to a European Population: Is It Time for Change? *Diabetes Care* **36**, 3040–3044 (2013).
24. Gao, C., Sun, X., Lu, L., Liu, F. & Yuan, J. Prevalence of gestational diabetes mellitus in mainland China: A systematic review and meta-analysis. *Journal of Diabetes Investigation* **10**, 154–162 (2019).
25. Lee, K. W. *et al.* Prevalence and risk factors of gestational diabetes mellitus in Asia: a systematic review and meta-analysis. *BMC Pregnancy and Childbirth* **18**, 494 (2018).
26. Kumar, M. *et al.* Automated Machine Learning (AutoML)-Derived Preconception Predictive Risk Model to Guide Early Intervention for Gestational Diabetes Mellitus. *International Journal of Environmental Research and Public Health* **19**, 6792 (2022).
27. American Diabetes Association. Gestational Diabetes Mellitus. *Diabetes Care* **27**, s88–s90 (2004).

28. ACOG Practice Bulletin No. 190: Gestational Diabetes Mellitus. *Obstetrics & Gynecology* **131**, e49 (2018).
29. O'Malley, E. & Turner, M. J. Diagnostic criteria for gestational diabetes mellitus. *Australian and New Zealand Journal of Obstetrics and Gynaecology* **60**, E16–E17 (2020).
30. Hod, M. *et al.* The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: A pragmatic guide for diagnosis, management, and care. *International Journal of Gynecology and Obstetrics* **131**, S173–S211 (2015).
31. Poon, L. C. *et al.* The International Federation of Gynecology and Obstetrics (FIGO) Initiative on Preeclampsia (PE): A Pragmatic Guide for First Trimester Screening and Prevention. *Int J Gynaecol Obstet* **145**, 1–33 (2019).
32. Daly, N. *et al.* Impact of Implementing Preanalytical Laboratory Standards on the Diagnosis of Gestational Diabetes Mellitus: A Prospective Observational Study. *Clinical Chemistry* **62**, 387–391 (2016).
33. Hyperglycemia and Adverse Pregnancy Outcomes. *New England Journal of Medicine* **358**, 1991–2002 (2008).
34. Crowther, C. A. *et al.* Effect of Treatment of Gestational Diabetes Mellitus on Pregnancy Outcomes. *New England Journal of Medicine* **352**, 2477–2486 (2005).
35. Mitchell, E. W., Levis, D. M. & Prue, C. E. Preconception Health: Awareness, Planning, and Communication Among a Sample of US Men and Women. *Matern Child Health J* **16**, 31–39 (2012).
36. Stephenson, J. *et al.* How Do Women Prepare for Pregnancy? Preconception Experiences of Women Attending Antenatal Services and Views of Health Professionals. *PLOS ONE* **9**, e103085 (2014).
37. Group, T. I. W. M. in P. (i-W. C. Effect of diet and physical activity based interventions in pregnancy on gestational weight gain and pregnancy outcomes: meta-analysis of individual participant data from randomised trials. *BMJ* **358**, j3119 (2017).
38. Tieu, J., Shepherd, E., Middleton, P. & Crowther, C. A. Dietary advice interventions in pregnancy for preventing gestational diabetes mellitus. *Cochrane Database of Systematic Reviews* https://doi.org/10.1002/14651858.CD006674.pub3 (2017) doi:10.1002/14651858.CD006674.pub3.
39. Martis, R. *et al.* Treatments for women with gestational diabetes mellitus: an overview of Cochrane systematic reviews. *Cochrane Database of Systematic Reviews* https://doi.org/10.1002/14651858.CD012327.pub2 (2018) doi:10.1002/14651858.CD012327.pub2.

40. Rönö, K. *et al.* Effect of a lifestyle intervention during pregnancy—findings from the Finnish gestational diabetes prevention trial (RADIEL). *J Perinatol* **38**, 1157–1164 (2018).
41. Peaceman, A. M. *et al.* Lifestyle Interventions Limit Gestational Weight Gain in Women with Overweight or Obesity: LIFE-Moms Prospective Meta-Analysis. *Obesity* **26**, 1396–1404 (2018).
42. Poston, L. *et al.* Effect of a behavioural intervention in obese pregnant women (the UPBEAT study): a multicentre, randomised controlled trial. *The Lancet Diabetes & Endocrinology* **3**, 767–777 (2015).
43. Guo, X.-Y. *et al.* Improving the effectiveness of lifestyle interventions for gestational diabetes prevention: a meta-analysis and meta-regression. *BJOG: An International Journal of Obstetrics & Gynaecology* **126**, 311–320 (2019).
44. Song, C., Li, J., Leng, J., Ma, R. C. & Yang, X. Lifestyle intervention can reduce the risk of gestational diabetes: a meta-analysis of randomized controlled trials. *Obes Rev* **17**, 960–969 (2016).
45. Thangaratinam, S. *et al.* Effects of interventions in pregnancy on maternal weight and obstetric outcomes: meta-analysis of randomised evidence. *BMJ* **344**, e2088 (2012).
46. Bailey, C. *et al.* A Comparison of the Cost-Effectiveness of Lifestyle Interventions in Pregnancy. *Value in Health* **25**, 194–202 (2022).
47. McLaren, R. A. *et al.* Early screening for gestational diabetes mellitus: a meta-analysis of randomized controlled trials. *American Journal of Obstetrics & Gynecology MFM* **4**, (2022).
48. Simmons, D. *et al.* Treatment of Gestational Diabetes Mellitus Diagnosed Early in Pregnancy. *New England Journal of Medicine* **388**, 2132–2144 (2023).
49. Immanuel, J. & Simmons, D. Screening and Treatment for Early-Onset Gestational Diabetes Mellitus: a Systematic Review and Meta-analysis. *Curr Diab Rep* **17**, 115 (2017).
50. Yefet, E. *et al.* Risk for fetal malformations and unfavorable neonatal outcomes in early-onset gestational diabetes mellitus. *J Endocrinol Invest* **47**, 1181–1190 (2024).
51. Helm, J. M. *et al.* Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Curr Rev Musculoskelet Med* **13**, 69–76 (2020).
52. Simon, H. A. 2 - WHY SHOULD MACHINES LEARN? in *Machine Learning* (eds Michalski, R. S., Carbonell, J. G. & Mitchell, T. M.) 25–37 (Morgan Kaufmann, San Francisco (CA), 1983). doi:10.1016/B978-0-08-051054-5.50006-6.

53. Antoniadi, A. M. *et al.* Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences* **11**, 5088 (2021).
54. Mould, D., D'Haens, G. & Upton, R. Clinical Decision Support Tools: The Evolution of a Revolution. *Clinical Pharmacology & Therapeutics* **99**, 405–418 (2016).
55. Bright, T. J. *et al.* Effect of Clinical Decision-Support Systems. *Ann Intern Med* **157**, 29–43 (2012).
56. Adkins, D. E. Machine Learning and Electronic Health Records: A Paradigm Shift. *AJP* **174**, 93–94 (2017).
57. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine* **178**, 1544–1547 (2018).
58. Barak-Corren, Y. *et al.* Predicting Suicidal Behavior From Longitudinal Electronic Health Records. *AJP* **174**, 154–162 (2017).
59. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* **18**, 463–477 (2019).
60. Jeon, J. *et al.* A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med* **6**, 57 (2014).
61. Zhan, F. *et al.* The molecular classification of multiple myeloma. *Blood* **108**, 2020–2028 (2006).
62. Lee, G. *et al.* Nuclear Shape and Architecture in Benign Fields Predict Biochemical Recurrence in Prostate Cancer Patients Following Radical Prostatectomy: Preliminary Findings. *Eur Urol Focus* **3**, 457–466 (2017).
63. Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine Learning for Medical Imaging. *Radiographics* **37**, 505 (2017).
64. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
65. Willemink, M. J. *et al.* Preparing Medical Imaging Data for Machine Learning. *Radiology* **295**, 4–15 (2020).
66. Panicacci, S. *et al.* Population Health Management Exploiting Machine Learning Algorithms to Identify High-Risk Patients. in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)* 298–303 (2018). doi:10.1109/CBMS.2018.00059.

67. Chae, S., Kwon, S. & Lee, D. Predicting Infectious Disease Using Deep Learning and Big Data. *International Journal of Environmental Research and Public Health* **15**, 1596 (2018).
68. Nwaokorie, A. & Fey, D. Personalised Medicine for Colorectal Cancer Using Mechanism-Based Machine Learning Models. *International Journal of Molecular Sciences* **22**, 9970 (2021).
69. Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L. & Mooney, C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Sci Rep* **12**, 1170 (2022).
70. Filippi, V., Chou, D., Barreix, M., Say, L. & Group (MMWG), the W. M. M. W. A new conceptual framework for maternal morbidity. *International Journal of Gynecology & Obstetrics* **141**, 4–9 (2018).
71. Kumar, M. *et al.* Population-centric risk prediction modeling for gestational diabetes mellitus: A machine learning approach. *Diabetes Research and Clinical Practice* **185**, 109237 (2022).
72. Kumar, M. *et al.* Machine Learning–Derived Prenatal Predictive Risk Model to Guide Intervention and Prevent the Progression of Gestational Diabetes Mellitus to Type 2 Diabetes: Prediction Model Development Study. *JMIR Diabetes* **7**, e32366 (2022).
73. Liao, L. D. *et al.* Development and validation of prediction models for gestational diabetes treatment modality using supervised machine learning: a population-based cohort study. *BMC Med* **20**, 307 (2022).
74. Pustozerov, E. A. *et al.* Machine Learning Approach for Postprandial Blood Glucose Prediction in Gestational Diabetes Mellitus. *IEEE Access* **8**, 219308–219321 (2020).
75. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).
76. Mandrekar, J. N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology* **5**, 1315–1316 (2010).
77. Collins, G. S. *et al.* Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819 (2024).
78. Efthimiou, O. *et al.* Developing clinical prediction models: a step-by-step guide. *BMJ* **386**, e078276 (2024).
79. Riley, R. D. *et al.* Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *BMJ* **388**, e080749 (2025).

80. Tang, A. *et al.* Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can Assoc Radiol J* **69**, 120–135 (2018).
81. Caliskan, E., Kayikcioglu, F., Öztürk, N., Koc, S. & Haberal, A. A population-based risk factor scoring will decrease unnecessary testing for the diagnosis of gestational diabetes mellitus. *Acta Obstetricia et Gynecologica Scandinavica* **83**, 524–530 (2004).
82. Kleinrouweler, C. E. *et al.* Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* **214**, 79-90.e36 (2016).
83. Zhang, Z. *et al.* Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. *Journal of Medical Internet Research* **24**, e26634 (2022).
84. van Leeuwen, M. *et al.* Estimating the risk of gestational diabetes mellitus: a clinical prediction model based on patient characteristics and medical history. *BJOG: An International Journal of Obstetrics & Gynaecology* **117**, 69–75 (2010).
85. An Introduction to Statistical Learning: with Applications in R | SpringerLink. https://link.springer.com/book/10.1007/978-1-4614-7138-7?view=modern&utm_source=tomegenius.
86. Yan, J. *et al.* A Prediction Model of Gestational Diabetes Mellitus Based on First Pregnancy Test Index. in *Health Information Science* (eds Huang, Z., Siuly, S., Wang, H., Zhou, R. & Zhang, Y.) 121–132 (Springer International Publishing, Cham, 2020). doi:10.1007/978-3-030-61951-0_12.
87. Kheng, T. Y. *Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things*. (BoD – Books on Demand, 2021).
88. Neeman, T. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating by Ewout W. Steyerberg. *International Statistical Review* **77**, 320–321 (2009).
89. Ruiter, M. L. *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).
90. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* **14**, 49–58 (2020).
91. Riley, R. D. *et al.* Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ* **384**, e074820 (2024).

92. van Smeden, M., Reitsma, J. B., Riley, R. D., Collins, G. S. & Moons, K. G. Clinical prediction models: diagnosis versus prognosis. *Journal of Clinical Epidemiology* **132**, 142–145 (2021).
93. Di Filippo, D. *et al.* Continuous Glucose Monitoring for the Diagnosis of Gestational Diabetes Mellitus: A Pilot Study. *J Diabetes Res* **2022**, 5142918 (2022).
94. Artzi, N. S. *et al.* Prediction of gestational diabetes based on nationwide electronic health records. *Nat Med* **26**, 71–76 (2020).
95. Qiu, H. *et al.* Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy. *Sci Rep* **7**, 16417 (2017).
96. Mateen, B. A., David, A. L. & Denaxas, S. Electronic Health Records to Predict Gestational Diabetes Risk. *Trends in Pharmacological Sciences* **41**, 301–304 (2020).
97. Sufriyana, H. *et al.* Comparison of Multivariable Logistic Regression and Other Machine Learning Algorithms for Prognostic Prediction Studies in Pregnancy Care: Systematic Review and Meta-Analysis. *JMIR Medical Informatics* **8**, e16503 (2020).
98. Seto, H. *et al.* Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Sci Rep* **12**, 15889 (2022).
99. Mennickent, D., Rodríguez, A., Farías-Jofré, M., Araya, J. & Guzmán-Gutiérrez, E. Machine learning-based models for gestational diabetes mellitus prediction before 24–28 weeks of pregnancy: A review. *Artificial Intelligence in Medicine* **132**, 102378 (2022).
100. Hedderson, M. *et al.* Racial/Ethnic Disparities in the Prevalence of Gestational Diabetes Mellitus by BMI. *Diabetes Care* **35**, 1492–1498 (2012).
101. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).
102. Moons, K. G. M. *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* **388**, e082505 (2025).
103. Navarro, C. L. A. *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* **375**, n2281 (2021).
104. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat Med* **30**, 2838–2848 (2024).
105. Liu, M. *et al.* A translational perspective towards clinical AI fairness. *npj Digit. Med.* **6**, 1–6 (2023).

106. Song, Z. *et al.* Prediction of gestational diabetes mellitus by different obesity indices. *BMC Pregnancy and Childbirth* **22**, 288 (2022).
107. Rathnayake, H. *et al.* Advancement in predictive biomarkers for gestational diabetes mellitus diagnosis and related outcomes: a scoping review. https://doi.org/10.1136/bmjopen-2024-089937 (2024) doi:10.1136/bmjopen-2024-089937.
108. Hou, G. *et al.* Maternal plasma diacylglycerols and triacylglycerols in the prediction of gestational diabetes mellitus. *BJOG: An International Journal of Obstetrics & Gynaecology* **130**, 247–256 (2023).
109. O'Malley, E. G. *et al.* Maternal obesity and dyslipidemia associated with gestational diabetes mellitus (GDM). *European Journal of Obstetrics and Gynecology and Reproductive Biology* **246**, 67–71 (2020).
110. Pour, S. J. *et al.* Analysis of serum circulating MicroRNAs level in Malaysian patients with gestational diabetes mellitus. *Sci Rep* **12**, 20295 (2022).
111. Donovan, B. M. *et al.* Development and validation of a clinical model for preconception and early pregnancy risk prediction of gestational diabetes mellitus in nulliparous women. *PLoS One* **14**, e0215173 (2019).
112. Schoenaker, D. A. J. M., Vergouwe, Y., Soedamah-Muthu, S. S., Callaway, L. K. & Mishra, G. D. Preconception risk of gestational diabetes: Development of a prediction model in nulliparous Australian women. *Diabetes Res Clin Pract* **146**, 48–57 (2018).
113. Hahn, S., Körber, S., Gerber, B. & Stubert, J. Prediction of recurrent gestational diabetes mellitus: a retrospective cohort study. *Arch Gynecol Obstet* **307**, 689–697 (2023).
114. Chen, M., Xu, W., Guo, Y. & Yan, J. Predicting recurrent gestational diabetes mellitus using artificial intelligence models: a retrospective cohort study. *Arch Gynecol Obstet* **310**, 1621–1630 (2024).
115. Yang, J. *et al.* Machine Learning-Based Risk Stratification for Gestational Diabetes Management. *Sensors* **22**, 4805 (2022).
116. Eleftheriades, M. *et al.* Prediction of insulin treatment in women with gestational diabetes mellitus. *Nutr. Diabetes* **11**, 1–5 (2021).
117. Velardo, C. *et al.* Toward a Multivariate Prediction Model of Pharmacological Treatment for Women With Gestational Diabetes Mellitus: Algorithm Development and Validation. *Journal of Medical Internet Research* **23**, e21435 (2021).
118. Benham, J. L. *et al.* Precision gestational diabetes treatment: a systematic review and meta-analyses. *Commun Med* **3**, 1–13 (2023).

119. Steyerberg, E. W. Updating for a New Setting. in *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (ed. Steyerberg, E. W.) 399–429 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-030-16399-0_20.
120. Riley, R. D. *et al.* Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. *BMJ* **384**, e074821 (2024).
121. Meertens, L. J. E. *et al.* External validation and clinical utility of prognostic prediction models for gestational diabetes mellitus: A prospective cohort study. *Acta Obstet Gynecol Scand* **99**, 891–900 (2020).
122. Cooray, S. D. *et al.* Development, validation and clinical utility of a risk prediction model for adverse pregnancy outcomes in women with gestational diabetes: The PeRSonal GDM model. *EClinicalMedicine* **52**, 101637 (2022).
123. Ranasinha, S. *et al.* External validation of risk prediction model for gestational diabetes: Individual participant data meta-analysis of randomized trials. *Int J Med Inform* **190**, 105533 (2024).
124. Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? Preprint at https://doi.org/10.48550/arXiv.1712.09923 (2017).
125. Hudec, M., Bednárová, E. & Holzinger, A. Augmenting Statistical Data Dissemination by Short Quantified Sentences of Natural Language. *Journal of Official Statistics* **34**, 981–1010 (2018).
126. Gilpin, L. H. *et al.* Explaining Explanations: An Overview of Interpretability of Machine Learning. in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 80–89 (2018). doi:10.1109/DSAA.2018.00018.
127. Bhatt, U. *et al.* Explainable machine learning in deployment. in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 648–657 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3351095.3375624.
128. Bhatt, U., Andrus, M., Weller, A. & Xiang, A. Machine Learning Explainability for External Stakeholders. Preprint at https://doi.org/10.48550/arXiv.2007.05408 (2020).
129. Bologna, G. & Hayashi, Y. Characterization of Symbolic Rules Embedded in Deep DIMLP Networks: A Challenge to Transparency of Deep Learning. *Journal of Artificial Intelligence and Soft Computing Research* **7**, 265–286 (2017).
130. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery* **9**, e1312 (2019).

131. Caruana, R. *et al.* Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1721–1730 (Association for Computing Machinery, New York, NY, USA, 2015). doi:10.1145/2783258.2788613.
132. Esteva, A. *et al.* Deep learning-enabled medical computer vision. *npj Digit. Med.* **4**, 1–9 (2021).
133. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. Preprint at https://doi.org/10.48550/arXiv.1909.09223 (2019).
134. Shapley, L. S. 17. A Value for n-Person Games. in *17. A Value for n-Person Games* 307–318 (Princeton University Press, 2016). doi:10.1515/9781400881970-018.
135. Du, Y., Antoniadi, A. M., McNestry, C., McAuliffe, F. M. & Mooney, C. The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations. *Applied Sciences* **12**, 10323 (2022).
136. Eitel-Porter, R. Beyond the promise: implementing ethical AI. *AI Ethics* **1**, 73–80 (2021).
137. Guidotti, R. *et al.* A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51**, 93:1-93:42 (2018).
138. Birhane, A. Algorithmic injustice: a relational ethics approach. *Patterns* **2**, 100205 (2021).
139. *Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women *. Ethics of Data and Analytics* 296–299 (Auerbach Publications, 2022). doi:10.1201/9781003278290-44.
140. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L. & Bonham, V. L. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. *Health Affairs* **37**, 780–785 (2018).
141. Hense, H.-W., Schulte, H., Löwel, H., Assmann, G. & Keil, U. Framingham risk function overestimates risk of coronary heart disease in men and women from Germany—results from the MONICA Augsburg and the PROCAM cohorts. *European Heart Journal* **24**, 937–945 (2003).
142. Lekadir, K. *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).

143. Singh, R. P. *et al.* Current Challenges and Barriers to Real-World Artificial Intelligence Adoption for the Healthcare System, Provider, and the Patient. *Translational Vision Science & Technology* **9**, 45 (2020).
144. Zhang, B. & Dafoe, A. U.S. Public Opinion on the Governance of Artificial Intelligence. in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 187–193 (Association for Computing Machinery, New York, NY, USA, 2020). doi:10.1145/3375627.3375827.
145. Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. in *Proceedings of the 4th Machine Learning for Healthcare Conference* 359–380 (PMLR, 2019).
146. Snell, K. I. E. *et al.* Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ* **381**, e073538 (2023).
147. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
148. Cacciamani, G. E. *et al.* PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nat Med* **29**, 14–15 (2023).
149. Moons, K. G. M. *et al.* Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine* **11**, e1001744 (2014).
150. Debray, T. P. *et al.* A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res* **28**, 2768–2786 (2019).
151. Debray, T. P. A. *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, i6460 (2017).
152. Snell, K. I., Ensor, J., Debray, T. P., Moons, K. G. & Riley, R. D. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* **27**, 3505–3522 (2018).
153. Hedges, L. V. & Olkin, I. CHAPTER 9 - Random Effects Models for Effect Sizes. in *Statistical Methods for Meta-Analysis* (eds Hedges, L. V. & Olkin, I.) 189–203 (Academic Press, San Diego, 1985). doi:10.1016/B978-0-08-057065-5.50014-2.
154. Newcombe, R. G. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**, 857–872 (1998).

155. Germaine, M., O'Higgins, A. C., Egan, B. & Healy, G. Gestational Diabetes Diagnoses in Electronic Health Records: A Three-Step Study of Label Accuracy and Its Impact on Machine Learning Models for Early Prediction. Preprint at https://doi.org/10.2196/preprints.72938 (2025).
156. Germaine, M., O'Higgins, A. C., Egan, B. & Healy, G. Evaluation of Machine Learning Models for Early Prediction of Gestational Diabetes Using Retrospective Electronic Health Records from Current and Previous Pregnancies. 2025.05.12.25327431 Preprint at https://doi.org/10.1101/2025.05.12.25327431 (2025).
157. Pustejovsky, J. E. & Tipton, E. Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prev Sci* **23**, 425–438 (2022).
158. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557–560 (2003).
159. Viechtbauer, W. & Cheung, M. W.-L. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* **1**, 112–125 (2010).
160. Egger, M., Smith, G. D., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634 (1997).
161. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**, 1–48 (2010).
162. Amylidi-Mohr, S. *et al.* First-trimester glycosylated hemoglobin (HbA1c) and maternal characteristics in the prediction of gestational diabetes: An observational cohort study. *Acta Obstet Gynecol Scand* **102**, 294–300 (2023).
163. Basil, B., Mba, I. N., Myke-Mbata, B. K., Adebisi, S. A. & Oghagbon, E. K. A first trimester prediction model and nomogram for gestational diabetes mellitus based on maternal clinical risk factors in a resource-poor setting. *BMC Pregnancy Childbirth* **24**, 346 (2024).
164. Belsti, Y. *et al.* Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *Int J Med Inform* **179**, 105228 (2023).
165. Benhalima, K. *et al.* Estimating the risk of gestational diabetes mellitus based on the 2013 WHO criteria: a prediction model based on clinical and biochemical variables in early pregnancy. *Acta Diabetol* **57**, 661–671 (2020).
166. Cooray, S. D. *et al.* Temporal validation and updating of a prediction model for the diagnosis of gestational diabetes mellitus. *J Clin Epidemiol* **164**, 54–64 (2023).

167. Gao, S. *et al.* Development and validation of an early pregnancy risk score for the prediction of gestational diabetes mellitus in Chinese pregnant women. *BMJ Open Diabetes Research and Care* **8**, e000909 (2020).
168. Guo, F. *et al.* Nomogram for prediction of gestational diabetes mellitus in urban, Chinese, pregnant women. *BMC Pregnancy Childbirth* **20**, 43 (2020).
169. Hu, X., Hu, X., Yu, Y. & Wang, J. Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm. *Front Endocrinol (Lausanne)* **14**, 1105062 (2023).
170. Kang, B. S. *et al.* Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. *Sci Rep* **13**, 13356 (2023).
171. Kaya, Y. *et al.* The early prediction of gestational diabetes mellitus by machine learning models. *BMC Pregnancy Childbirth* **24**, 574 (2024).
172. Li, R. *et al.* Construction and validation of risk prediction model for gestational diabetes based on a nomogram. *Am J Transl Res* **15**, 1223–1230 (2023).
173. Li, Y.-X., Liu, Y.-C., Wang, M. & Huang, Y.-L. Prediction of gestational diabetes mellitus at the first trimester: machine-learning algorithms. *Arch Gynecol Obstet* https://doi.org/10.1007/s00404-023-07131-4 (2023) doi:10.1007/s00404-023-07131-4.
174. Lin, Q. & Fang, Z.-J. Establishment and evaluation of a risk prediction model for gestational diabetes mellitus. *World J Diabetes* **14**, 1541–1550 (2023).
175. Liu, H. *et al.* Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China. *Diabetes/Metabolism Research and Reviews* **37**, e3397 (2021).
176. Liu, R. *et al.* Stacking Ensemble Method for Gestational Diabetes Mellitus Prediction in Chinese Pregnant Women: A Prospective Cohort Study. *J Healthc Eng* **2022**, 8948082 (2022).
177. Niu, Z.-R., Bai, L.-W. & Lu, Q. Establishment of gestational diabetes risk prediction model and clinical verification. *J Endocrinol Invest* https://doi.org/10.1007/s40618-023-02249-3 (2023) doi:10.1007/s40618-023-02249-3.
178. Schaefer, K. K. *et al.* Prediction of gestational diabetes mellitus in the Born in Guangzhou Cohort Study, China. *Int J Gynaecol Obstet* **143**, 164–171 (2018).
179. Sweeting, A. N. *et al.* First trimester prediction of gestational diabetes mellitus: A clinical model based on maternal demographic parameters. *Diabetes Res Clin Pract* **127**, 44–50 (2017).

180. Syngelaki, A., Wright, A., Gomez Fernandez, C., Mitsigiorgi, R. & Nicolaides, K. H. First-Trimester Prediction of Gestational Diabetes Mellitus Based on Maternal Risk Factors. *BJOG: An International Journal of Obstetrics & Gynaecology* **132**, 972–982 (2025).
181. Tang, Y. *et al.* Development and validation of a risk prediction model for gestational diabetes mellitus in women of advanced maternal age during the first trimester. *The FASEB Journal* **39**, e70334 (2025).
182. Teede, H. J., Harrison, C. L., Teh, W. T., Paul, E. & Allan, C. A. Gestational diabetes: development of an early risk prediction tool to facilitate opportunities for prevention. *Aust N Z J Obstet Gynaecol* **51**, 499–504 (2011).
183. Tranidou, A. *et al.* Prediction of Gestational Diabetes Mellitus in the First Trimester of Pregnancy Based on Maternal Variables and Pregnancy Biomarkers. *Nutrients* **16**, 120 (2024).
184. Zhong, W. *et al.* Gestational Diabetes Mellitus Prediction Based on Two Classification Algorithms. in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* 1–7 (2019). doi:10.1109/CISP-BMEI48845.2019.8965819.
185. Wang, N. *et al.* Development and Validation of Risk Prediction Models for Gestational Diabetes Mellitus Using Four Different Methods. *Metabolites* **12**, 1040 (2022).
186. Wang, X. *et al.* Establishment and validation of a prediction model for gestational diabetes. *Diabetes Obes Metab* https://doi.org/10.1111/dom.15356 (2023) doi:10.1111/dom.15356.
187. Wang, Y. *et al.* Risk Prediction Model of Gestational Diabetes Mellitus in a Chinese Population Based on a Risk Scoring System. *Diabetes Ther* **12**, 1721–1734 (2021).
188. Wei, Y. *et al.* Risk prediction models of gestational diabetes mellitus before 16 gestational weeks. *BMC Pregnancy Childbirth* **22**, 889 (2022).
189. Wu, Y. *et al.* Early prediction of gestational diabetes mellitus using maternal demographic and clinical risk factors. *BMC Res Notes* **17**, 105 (2024).
190. Wu, Y. *et al.* A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women. *Diabetes Research and Clinical Practice* **179**, 109001 (2021).
191. Wu, Y.-T. *et al.* Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning. *The Journal of Clinical Endocrinology & Metabolism* **106**, e1191–e1205 (2021).

192. Xing, J. *et al.* Enhancing gestational diabetes mellitus risk assessment and treatment through GDMPredictor: a machine learning approach. *J Endocrinol Invest* **47**, 2351–2360 (2024).
193. Yang, Z. *et al.* An early prediction model for gestational diabetes mellitus created using machine learning algorithms. *International Journal of Gynecology & Obstetrics* **n/a**,.
194. Zhang, H. *et al.* Integration of clinical demographics and routine laboratory analysis parameters for early prediction of gestational diabetes mellitus in the Chinese population. *Front Endocrinol (Lausanne)* **14**, 1216832 (2023).
195. van Eekhout, J. C. A. *et al.* First-Trimester Prediction Models Based on Maternal Characteristics for Adverse Pregnancy Outcomes: A Systematic Review and Meta-Analysis. *BJOG: An International Journal of Obstetrics & Gynaecology* **132**, 243–265 (2025).
196. Huang, Q.-F. *et al.* Clinical First-Trimester Prediction Models for Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis. *Biol Res Nurs* **25**, 185–197 (2023).
197. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* **110**, 12–22 (2019).
198. Vollmer, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **368**, l6927 (2020).
199. Kilkenny, M. F. & Robinson, K. M. Data quality: "Garbage in – garbage out". *HIM J* **47**, 103–105 (2018).
200. Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441 (2020).
201. Royen, F. S. van, Moons, K. G. M., Geersing, G.-J. & Smeden, M. van. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *European Respiratory Journal* **60**, (2022).
202. Germaine, M., Healy, G. & Egan, B. Lack of Data Sharing Despite Data Availability Statements in Studies Using Machine Learning Models for Prediction of Gestational Diabetes Mellitus. *Diabetes Care* **47**, e78–e79 (2024).
203. Lamain-de Ruiter, M. *et al.* Prediction models for the risk of gestational diabetes: a systematic review. *Diagn Progn Res* **1**, 3 (2017).

204. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* **27**, 621–633 (2020).
205. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research* **3**, 18 (2019).
206. Calster, B. V. *et al.* Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *European Urology* **74**, 796–804 (2018).
207. Van Calster, B., Steyerberg, E. W., Wynants, L. & van Smeden, M. There is no such thing as a validated prediction model. *BMC Medicine* **21**, 70 (2023).
208. Li, Y., Sperrin, M., Martin, G. P., Ashcroft, D. M. & van Staa, T. P. Examining the impact of data quality and completeness of electronic health records on predictions of patients' risks of cardiovascular disease. *International Journal of Medical Informatics* **133**, 104033 (2020).
209. Rahnemaei, F. A. *et al.* Association between body mass index in the first half of pregnancy and gestational diabetes: A systematic review. *SAGE Open Med* **10**, 20503121221109911 (2022).
210. Nijman, S. *et al.* Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *Journal of Clinical Epidemiology* **142**, 218–229 (2022).
211. Held, U. *et al.* Methods for Handling Missing Variables in Risk Prediction Models. *American Journal of Epidemiology* **184**, 545–551 (2016).
212. Stiglic, G., Kocbek, P., Fijacko, N., Sheikh, A. & Pajnkihar, M. Challenges associated with missing data in electronic health records: A case study of a risk prediction model for diabetes using data from Slovenian primary care. *Health Informatics J* **25**, 951–959 (2019).
213. Ganzeboom, H. B. G. & Treiman, D. J. Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research* **25**, 201–239 (1996).
214. Nicolaides, K. H., Wright, D., Syngelaki, A., Wright, A. & Akolekar, R. Fetal Medicine Foundation fetal and neonatal population weight charts. *Ultrasound in Obstetrics & Gynecology* **52**, 44–51 (2018).
215. Chou, J. S., Packer ,Claire H., Mittleman ,Murray A. & and Valent, A. M. Association of interpregnancy interval and gestational diabetes mellitus. *The Journal of Maternal-Fetal & Neonatal Medicine* **35**, 10545–10550 (2022).

216. Martínez-Hortelano, J. A. *et al.* Interpregnancy Weight Change and Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis. *Obesity* **29**, 454–464 (2021).
217. Riskin-Mashiah, S., Damti, A., Younes, G. & Auslender, R. First trimester fasting hyperglycemia as a predictor for the development of gestational diabetes mellitus. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **152**, 163–167 (2010).
218. Bender, W., McCarthy ,Clare, Elovitz ,Michal, Parry ,Samuel & and Durnwald, C. Universal HbA1c screening and gestational diabetes: a comparison with clinical risk factors. *The Journal of Maternal-Fetal & Neonatal Medicine* **35**, 6430–6436 (2022).
219. Neville, J. *et al.* Impact of changes in gestational diabetes mellitus diagnostic criteria during the COVID-19 pandemic. *Ir J Med Sci* https://doi.org/10.1007/s11845-025-03926-3 (2025) doi:10.1007/s11845-025-03926-3.
220. Garrison Jr., L. P., Neumann, P. J., Erickson, P., Marshall, D. & Mullins, C. D. Using Real-World Data for Coverage and Payment Decisions: The ISPOR Real-World Data Task Force Report. *Value in Health* **10**, 326–335 (2007).
221. Berger, M. L. *et al.* Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiology and Drug Safety* **26**, 1033–1039 (2017).
222. Wong, J., Murray Horwitz, M., Zhou, L. & Toh, S. Using Machine Learning to Identify Health Outcomes from Electronic Health Record Data. *Curr Epidemiol Rep* **5**, 331–342 (2018).
223. Ford, E., Carroll, J. A., Smith, H. E., Scott, D. & Cassell, J. A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association* **23**, 1007–1015 (2016).
224. Bowman, S. Impact of Electronic Health Record Systems on Information Integrity: Quality and Safety Implications. *Perspect Health Inf Manag* **10**, 1c (2013).
225. Frénay, B. & Kaban, A. A Comprehensive Introduction to Label Noise. in (2014).
226. Yang, J., Triendl, H., Soltan, A. A. S., Prakash, M. & Clifton, D. A. Addressing label noise for electronic health records: insights from computer vision for tabular data. *BMC Medical Informatics and Decision Making* **24**, 183 (2024).
227. Nissen, F., Quint, J. K., Morales, D. R. & Douglas, I. J. How to validate a diagnosis recorded in electronic health records. *Breathe* **15**, 64–68 (2019).

228. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107–115 (2021).
229. Cebul, R. D., Love, T. E., Jain, A. K. & Hebert, C. J. Electronic Health Records and Quality of Diabetes Care. *New England Journal of Medicine* **365**, 825–833 (2011).
230. Veinot, T. C., Zheng, K., Lowery, J. C., Souden, M. & Keith, R. Using electronic health record systems in diabetes care: emerging practices. in *Proceedings of the 1st ACM International Health Informatics Symposium* 240–249 (Association for Computing Machinery, New York, NY, USA, 2010). doi:10.1145/1882992.1883026.
231. de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digit. Med.* **5**, 1–13 (2022).
232. Cabitza, F. & Campagner, A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *International Journal of Medical Informatics* **153**, 104510 (2021).
233. Burus, T. *et al.* Undiagnosed Cancer Cases in the US During the First 10 Months of the COVID-19 Pandemic. *JAMA Oncology* https://doi.org/10.1001/jamaoncol.2023.6969 (2024) doi:10.1001/jamaoncol.2023.6969.
234. San Lazaro Campillo, I. *et al.* Assessing the concordance and accuracy between hospital discharge data, electronic health records, and register books for diagnosis of inpatient admissions of miscarriage: A retrospective linked data study. *Journal of Obstetrics and Gynaecology Research* **47**, 1987–1996 (2021).
235. Davidson, J., Banerjee, A., Muzambi, R., Smeeth, L. & Warren-Gash, C. Validity of Acute Cardiovascular Outcome Diagnoses Recorded in European Electronic Health Records: A Systematic Review. *Clinical Epidemiology* **12**, 1095–1111 (2020).
236. Bernhardt, M. *et al.* Active label cleaning for improved dataset quality under resource constraints. *Nat Commun* **13**, 1161 (2022).
237. Ju, L. *et al.* Improving Medical Images Classification With Label Noise Using Dual-Uncertainty Estimation. *IEEE Transactions on Medical Imaging* **41**, 1533–1546 (2022).
238. Sáez, J. A. Noise Models in Classification: Unified Nomenclature, Extended Taxonomy and Pragmatic Categorization. *Mathematics* **10**, 3736 (2022).
239. Yasuda, S. *et al.* Differences in the birthweight of infants born to patients with early- or mid-to-late-detected gestational diabetes mellitus who underwent guideline-based glycemic control. *Journal of Diabetes and its Complications* **35**, 107850 (2021).

240. Hosseini, E., Janghorbani, M. & Shahshahan, Z. Comparison of risk factors and pregnancy outcomes of gestational diabetes mellitus diagnosed during early and late pregnancy. *Midwifery* **66**, 64–69 (2018).
241. Reynolds, C. M. E., O'Malley, E. G., Egan, B., Sheehan, S. R. & Turner, M. J. Maternal Weight Trajectories in Successive Pregnancies and Their Association With Gestational Diabetes Mellitus. *Diabetes Care* **43**, e33–e34 (2020).
242. Panaitescu, A. M. RESUMING ADEQUATE SCREENING FOR GESTATIONAL DIABETES MELLITUS DURING THE ONGOING COVID-19 PANDEMIC. *Acta Endocrinol (Buchar)* **17**, 278–279 (2021).
243. Thomassen, D., le Cessie, S., van Houwelingen, H. C. & Steyerberg, E. W. Effective sample size: A measure of individual uncertainty in predictions. *Statistics in Medicine* **43**, 1384–1396 (2024).
244. Bosschieter, T. M. *et al.* Interpretable Predictive Models to Understand Risk Factors for Maternal and Fetal Outcomes. *J Healthc Inform Res* https://doi.org/10.1007/s41666-023-00151-4 (2023) doi:10.1007/s41666-023-00151-4.
245. DiCiccio, T. J. & Efron, B. Bootstrap confidence intervals. *Statistical Science* **11**, 189–228 (1996).
246. Cubillos, G. *et al.* Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. *BMC Pregnancy Childbirth* **23**, 469 (2023).
247. Ye, Y. *et al.* Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study. *Journal of Diabetes Research* **2020**, e4168340 (2020).
248. Maslin, K. *et al.* Interpregnancy maternal weight change is not associated with offspring weight and obesity at age 2 years. *Int J Obes* **48**, 1402–1413 (2024).
249. Health AIo, Welfare. Diabetes: Australian facts. *Canberra: AIHW* (2024).
250. McMahon, L. E., O'Malley, E. G., Reynolds, C. M. E. & Turner, M. J. The impact of revised diagnostic criteria on hospital trends in gestational diabetes mellitus rates in a high income country. *BMC Health Services Research* **20**, 795 (2020).
251. Ye, W. *et al.* Gestational diabetes mellitus and adverse pregnancy outcomes: systematic review and meta-analysis. *BMJ* **377**, e067946 (2022).
252. Cooray, S. D., Thangaratinam, S. & Teede, H. J. Prediction modelling to personalise care for gestational diabetes. *BJOG* **128**, 655–656 (2021).
253. Tiruneh, S. A., Rolnik, D. L., Teede, H. & Enticott, J. Temporal validation of machine learning models for pre-eclampsia prediction using routinely collected maternal

characteristics: A validation study. *Computers in Biology and Medicine* **191**, 110183 (2025).

254. Graydon, C., Teede, H., Sullivan, C., De Silva, K. & Enticott, J. Chapter 2 - Driving impact through big data utilization and analytics in the context of a Learning Health System. in *Big Data Analytics for Healthcare* (ed. Keikhosrokiani, P.) 13–22 (Academic Press, 2022). doi:10.1016/B978-0-323-91907-4.00019-4.
255. Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine* **40**, 4230–4251 (2021).
256. Australia's mothers and babies, Place of birth. *Australian Institute of Health and Welfare* https://www.aihw.gov.au/reports/mothers-babies/australias-mothers-babies/contents/labour-and-birth/place-of-birth (2025).
257. Reynolds, C. M. E. *et al.* Trends in private maternity care in Ireland's capital during and after the Great Economic Recession 2009–2017. *Ir J Med Sci* **190**, 933–940 (2021).
258. Riley, R. D. *et al.* Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine* **38**, 1276–1296 (2019).
259. Li, X. & Cong, Y. Exploring barriers and ethical challenges to medical data sharing: perspectives from Chinese researchers. *BMC Med Ethics* **25**, 132 (2024).
260. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* **24**, 198–208 (2017).
261. Riley, R. D. *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* **353**, i3140 (2016).
262. Watson, H. *et al.* Delivering on NIH data sharing requirements: avoiding Open Data in Appearance Only. *BMJ Health Care Inform* **30**, e100771 (2023).
263. Sharma, V. *et al.* Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform* **28**, (2021).
264. Carter, P., Laurie, G. T. & Dixon-Woods, M. The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics* **41**, 404–409 (2015).
265. Hedderson, M. M., Darbinian, J. A. & Ferrara, A. Disparities in the risk of gestational diabetes by race-ethnicity and country of birth. *Paediatr Perinat Epidemiol* **24**, 441–448 (2010).

266. Jenum, A. K. *et al.* Impact of ethnicity on gestational diabetes identified with the WHO and the modified International Association of Diabetes and Pregnancy Study Groups criteria: a population-based cohort study. *European Journal of Endocrinology* **166**, 317–324 (2012).
267. Lamri, A. *et al.* The genetic risk of gestational diabetes in South Asian women. *eLife* **11**, e81498 (2022).
268. Gu, Y. *et al.* Genetic architecture and risk prediction of gestational diabetes mellitus in Chinese pregnancies. *Nat Commun* **16**, 4178 (2025).
269. Sweeting, A. *et al.* Epidemiology and management of gestational diabetes. *The Lancet* **404**, 175–192 (2024).
270. Östlund, I., Haglund, B. & Hanson, U. Gestational diabetes and preeclampsia. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **113**, 12–16 (2004).
271. Bellamy, L., Casas, J.-P., Hingorani, A. D. & Williams, D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. *The Lancet* **373**, 1773–1779 (2009).
272. Linder, T. *et al.* Impact Of Prepregnancy Overweight And Obesity On Treatment Modality And Pregnancy Outcome In Women With Gestational Diabetes Mellitus. *Front. Endocrinol.* **13**, (2022).
273. Beltrand, J. *et al.* Neonatal Diabetes Mellitus. *Front Pediatr* **8**, 540718 (2020).
274. Bukhari, I., Iqbal, F. & Thorne, R. F. Research advances in gestational, neonatal diabetes mellitus and metabolic disorders. *Front. Endocrinol.* **13**, (2022).
275. Vasey, B. *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* **377**, e070904 (2022).
276. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* **26**, 1364–1374 (2020).
277. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1–26 (1979).
278. Riley, R. D. *et al.* Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *The Lancet Digital Health* 100857 (2025) doi:10.1016/j.landig.2025.01.013.
279. King, T. L. *et al.* Survey of willingness to participate in clinical trials and influencing factors among cancer and non-cancer patients. *Sci Rep* **15**, 1626 (2025).

280. Shrank, W. H., Patrick, A. R. & Alan Brookhart, M. Healthy User and Related Biases in Observational Studies of Preventive Interventions: A Primer for Physicians. *J GEN INTERN MED* **26**, 546–550 (2011).
281. Haque, M. M. *et al.* Cost-effectiveness of diagnosis and treatment of early gestational diabetes mellitus: economic evaluation of the TOBOGM study, an international multicenter randomized controlled trial. *eClinicalMedicine* **71**, (2024).
282. Farrar, D. *et al.* Risk factor screening to identify women requiring oral glucose tolerance testing to diagnose gestational diabetes: A systematic review and meta-analysis and analysis of two pregnancy cohorts. *PLOS ONE* **12**, e0175288 (2017).
283. Tymstra, T. & Bieleman, B. The psychosocial impact of mass screening for cardiovascular risk factors. *Fam Pract* **4**, 287–290 (1987).
284. Lachmann, E. H. *et al.* Barriers to completing oral glucose tolerance testing in women at risk of gestational diabetes. *Diabetic Medicine* **37**, 1482–1489 (2020).
285. Falcone, V. *et al.* Early Assessment of the Risk for Gestational Diabetes Mellitus: Can Fasting Parameters of Glucose Metabolism Contribute to Risk Prediction? *Diabetes Metab J* **43**, 785–793 (2019).
286. Gunasekaran, U. *et al.* First prenatal visit HbA1c improves gestational diabetes (GDM) prediction. *Diabetes* **63**, A340 (2014).
287. Panteli, D. *et al.* Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health* **10**, e428–e432 (2025).
288. van Panhuis, W. G. *et al.* A systematic review of barriers to data sharing in public health. *BMC Public Health* **14**, 1144 (2014).
289. Simpson, C. L. *et al.* Practical Barriers and Ethical Challenges in Genetic Data Sharing. *International Journal of Environmental Research and Public Health* **11**, 8383–8398 (2014).
290. Bullock, G. S. *et al.* Improving Clinical Utility of Real-World Prediction Models: Updating Through Recalibration. *The Journal of Strength & Conditioning Research* **37**, 1057 (2023).
291. Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. A. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology* **68**, 25–34 (2015).

292. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* **35**, 1925–1931 (2014).
293. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W. & Collins, G. S. Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association* **26**, 1651–1654 (2019).
294. Samaga, D. *et al.* Single-center versus multi-center data sets for molecular prognostic modeling: a simulation study. *Radiation Oncology* **15**, 109 (2020).
295. Wang, S. *et al.* Risk-prediction models for intravenous immunoglobulin resistance in Kawasaki disease: Risk-of-Bias Assessment using PROBAST. *Pediatr Res* **94**, 1125–1135 (2023).
296. Nagurney, J. T. *et al.* The Accuracy and Completeness of Data Collected by Prospective and Retrospective Methods. *Academic Emergency Medicine* **12**, 884–895 (2005).
297. Bryant, E. A., Scott, A. M., Greenwood, H. & Thomas, R. Patient and public involvement in the development of clinical practice guidelines: a scoping review. https://doi.org/10.1136/bmjopen-2021-055428 (2022) doi:10.1136/bmjopen-2021-055428.
298. Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393 (2009).
299. HARRELL Jr., F. E., Lee, K. L. & Mark, D. B. Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine* **15**, 361–387 (1996).
300. Sisk, R., Sperrin, M., Peek, N., van Smeden, M. & Martin, G. P. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* **32**, 1461–1477 (2023).
301. Steyerberg, E. W. *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLOS Medicine* **10**, e1001381 (2013).
302. Snooks, H. *et al.* Effects and costs of implementing predictive risk stratification in primary care: a randomised stepped wedge trial. *BMJ Qual Saf* **28**, 697–705 (2019).
303. Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K. & Calvert, M. J. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* **26**, 1351–1363 (2020).
304. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).

305. Van Calster, B. *et al.* A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology* **74**, 167–176 (2016).
306. Gianfrancesco, M. A. & Goldstein, N. D. A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Med Res Methodol* **21**, 234 (2021).
307. Benchimol, E. I. *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine* **12**, e1001885 (2015).
308. Jin, D. *et al.* Gestational Diabetes Mellitus: Predictive Value of Fetal Growth Measurements by Ultrasonography at 22–24 Weeks: A Retrospective Cohort Study of Medical Records. *Nutrients* **12**, 3645 (2020).
309. Rekawek, P. *et al.* Large-for-gestational age diagnosed during second-trimester anatomy ultrasound and association with gestational diabetes and large-for-gestational age at birth. *Ultrasound in Obstetrics & Gynecology* **56**, 901–905 (2020).

APPENDICES

Appendix A. DCU Graduate Training Elements Completed

CA684	Machine Learning *(7.5 ECTS)*
COMP50060 (UCD)	ML CRT Bootcamp *(10 ECTS)*
CA660	Statistical Data Analysis *(7.5 ECTS)*
CA675	Cloud Technologies *(7.5 ECTS)*
EE611	Enterprise Experience for Graduate Research Students *(15 ECTs)*

Appendix B. Folic Acid Supplementation in Pregnancy

Title: A 10-year review of periconceptual folic acid supplementation in women with epilepsy taking antiepileptic medications

Authors: Turner C[1], McIntosh T[2], Gaffney D[3], Germaine M[4], Hogan J[1], O'Higgins A[1].

Source: UCD Centre for Human Reproduction, The Coombe Hospital, Dublin, Ireland.

Affiliations: 1. UCD Centre for Human Reproduction, The Coombe Hospital, Dublin, Ireland. 2. School of Pharmacy, Applied Sciences and Public Health, Robert Gordon University, Aberdeen, Scotland. 3. Vaccination Service, Health Services Executive, Dublin, Ireland. 4. School of Computing and School of Health and Human Performance, Dublin City University, Dublin, Ireland.

Corresponding author: Professor Amy O'Higgins, Director, UCD Centre for Human Reproduction.

Email: amy.ohiggins@ucd.ie

Keywords: Women with epilepsy, periconceptual folic acid supplementation, prevention of neural tube defects, antiepileptic medications, national folic acid guidelines.

Abstract

Background

Epidemiological studies have reported that women with epilepsy who are taking anti-epileptic medications have an increased risk of Neural Tube Defects. Periconceptual folic acid supplementation potentially prevents two-thirds of cases. International guidelines recommend that women at increased risk of a pregnancy complicated by a Neural Tube Defect who could become pregnant should start high-dose (5mg daily) oral folic acid at least three months before conceiving. The purpose of the study was to examine supplementation in women taking antiepileptic medications who delivered a baby weighing >499g during the ten years 2013-2022 in a large maternity hospital.

Methods

The hospital's computerised database contains standardised maternal clinical and sociodemographic details which were entered at the first antenatal visit in the obstetric records. The data on all women with epilepsy taking antiepileptic medications was anonymised before coding.

Results

In the ten years, 75,869 women delivered a baby weighing >499g. Of the deliveries, 632 (0.83%) were to women with epilepsy. Of these, 250 (0.33%) were taking anti-epileptic medications when they presented for antenatal care. The most frequently prescribed medications were lamotrigine n=98 (33.8%) and levetiracetam n=89 (30.7%). Monotherapy was prescribed in 211 (84.4%) women and polytherapy in 39 (15.6%). Three (1.2%) women took no folic acid before or after conception and 59 (23.6%) only took it after conception. Of the 188 (75.2%) who took folic acid before conception, 164 (65.6%) took high-dose 5mg and 24 (9.6%) took low-dose 0.4mg. No maternal characteristics were associated with taking high-dose folic acid before conception. Compliance with the national guidelines in the 16 women taking valproate was 30.8% compared with 76.0% for the other 234 women taking medications ($p<0.03$). Compliance in the 211 women receiving monotherapy was 72.5% compared with 25.6% in the 39 women receiving polytherapy ($p<0.03$).

Conclusions

Two-thirds of women taking antiepileptic medications complied with national guidelines on high-dose periconceptual folic acid supplementation. A particular concern is the suboptimal

compliance in women prescribed valproate and polytherapy, which are two cohorts at higher risk of a neural tube defect. These findings need to be communicated to women with epilepsy in their reproductive years and their doctors.

349 words

Background

Neural Tube Defects (NTDs), which include anencephaly, spina bifida and encephalocele, are serious congenital malformations due to failure of the neural tube to close during early embryonic life, 21-28 days post-conception [1]. Anencephaly affects the brain and is associated with fetal death before or shortly after birth. Spina bifida is the failure of the spinal portion of the neural tube to close. It varies widely in severity and is associated with increased mortality and morbidity, which may be lifelong. An encephalocele is when a sac containing cerebrospinal fluid +/- meninges +/- brain herniates outside the skull due to a defect in normal skull formation. The prevalence of NTDs varies widely geographically from 0.5-6.0/1000 births. Globally, it is estimated that 300,000 babies with NTDs are born annually, resulting in 88,000 deaths and 8.6m disability-adjusted life-years [2].

There is strong evidence that the majority of NTDs may be prevented by periconceptual folic acid oral supplementation [3]. The best evidence comes from the Medical Research Council (MRC) international Randomised Controlled Trial (RCT) [4]. The folic acid dose used was 4mg daily and the trial reported a 71% decrease in NTDs in the treatment arm compared with controls. The decrease was so marked it was decided on ethical grounds to stop the trial early.

This reduction in NTDs was confirmed in a Hungarian RCT of 4753 pregnancies, using a multivitamin containing folic acid 0.8mg daily [5]. There were six cases of neural tubes defects in controls and none in the treatment arm. The findings were consistent with the MRC RCT. A recent systematic review in the United States found that observational studies provide further evidence of the benefit of folic acid supplementation and no evidence of harm, with no increase in multiple gestation, autism, or maternal cancer [6].

Epilepsy is a chronic neurological disorder characterised by recurrent, unprovoked seizures which may require anti-epileptic medications (AEMs) [7]. It affects an estimated 70 million people worldwide [8]. There is strong evidence that women with epilepsy (WWE) taking AEMs are at increased risk of major congenital malformations, including NTDs, cardiac and

craniofacial anomalies, compared with women without epilepsy [3,8-13]. The risk of congenital malformations in WWE depends on the type of anti-epileptic medications, for example, the risk is highest with valproate. The incidence is also increased with exposure to AEMs in the first trimester compared with later in pregnancy, with higher doses of AEMs and with polytherapy [8].

Following the RCTs, public health policies in Ireland recommended that all women who could become pregnant should take over-the-counter (OTC) oral folic acid 0.4mg daily for at least three months before they become pregnant and should continue for the first trimester [4,5,13]. National and international guidelines also recommended that women with a higher risk of NTDs should be prescribed high-dose folic acid supplement, although the optimal dosage remains uncertain [3,13-17].

In 2019, the Irish Department of Health published an updated report recommending that "All women who may possibly become pregnant within the next three months, whether intentionally or not, are advised to take oral folic acid 0.4mg daily to prevent neural tube defects (NTDs)" [13] "Women who are at increased risk of a pregnancy complicated by a NTD should arrange to see their doctor, because they may need a prescription-only folic acid 5mg daily" [13]. These recommendations are important because recent audits have shown that improvements in the national incidence of NTDs have stalled in Ireland and voluntary food fortification with folate has fallen [13,18]. The recommendations are supported by The Irish College of General Practitioners, The Institute of Obstetricians and Gynaecologists and The Royal College of Physicians of Ireland [19,20].

Previous studies on folic acid supplementation in pregnancy found that, in general, compliance rates were suboptimal and in women with unplanned pregnancies were at highest risk of suboptimal supplementation [3,21]. However, our review of the literature did not identify studies which specifically examined folic acid supplementation in WWE taking AEMs in early pregnancy, despite this being a group at increased risk of having a baby with a NTD.

Furthermore, there is scant information on what medications WWE in Ireland are taking for seizure control in early pregnancy. The only published report on the prevalence of epilepsy in Ireland drew on data from different data sources, including prescription drug data, but did not

present data on the different AEMs prescribed for the general population, including pregnant women [22].

In women with epilepsy taking AEMs who delivered a baby weighing >499g grams in a large maternity hospital between 2013-2022 inclusively, the aims of the study were to determine:

1. What were the overall folic acid supplementation rates, and did they follow the 2019 national recommendations in Ireland?
2. What was the timing and dosage of folic acid supplementation when women attended for their first hospital antenatal visit?
3. Was compliance with national folic acid recommendations associated with maternal characteristics?
4. What were the different antiepileptic medications prescribed and were they prescribed as monotherapy or polytherapy?

Methods

The hospital is a large, university maternity hospital which accepts patients from all socioeconomic groups across the rural-urban divide. The hospital accepts patients, with or without private health insurance, from Dublin and the surrounding counties. Currently about two-thirds of mothers reside in Dublin and one-third outside [23]. About two-thirds of mothers who attend are born in the Republic of Ireland and one third are international immigrants.

In this study, women who attended the hospital for pregnancy care had a consultation at their first visit with a trained midwife, who took a confidential history that was computerised in real-time using a standardised questionnaire and barcode system. The history included the current pregnancy, previous obstetric and gynaecological events, previous medical and surgical problems, social circumstances, lifestyle details, and the woman's current medications. Body Mass Index (BMI) was calculated after the accurate measurement of weight and height. A written record of the computerised history formed part of the clinical records.

The database included women who deliver a baby weighing >499g which is the standard birth cut-off weight recommended by the World Health Organization. The data for the ten years

2013-2022 was previously validated by researchers for other studies in the UCD Centre for Human Reproduction.

An anonymised dataset was created from the hospital's database with details on all the WWE who delivered during the years 2013-2022 inclusively. The diagnosis of epilepsy was based on the International League Against Epilepsy 2017 Classification [7]. No WWE taking AEMs were excluded from the study. The dataset was cleaned and coded for analysis using IBM SPSS Statistics (Version 27.0)**.**

Folic acid supplementation was defined as a categorical variable as follows: "none", "preconception use", "postconception use", "pre- and postconception use". The AEMs drug names were standardised before coding. The maternal sociodemographic and clinical characteristics at the first antenatal visit were analysed. Maternal employment was categorised based on Ireland's Central Statistics Office classification.

The study was approved by The Coombe Hospital's Research Ethics Committee and by the Research Ethics Committee, School of Pharmacy and Life Sciences, Robert Gordon University (No. S351). Data were irreversibly anonymised prior to analysis. Differences in maternal characteristics and medication regimes between those who were and were not compliant with the national folic acid supplementation guidelines were compared using Student's t-test. A p value <0.05 was considered statistically significant.

Results

Based on the hospital's records, 75,869 women delivered a baby weighing >499g during the ten years. Of the deliveries, 632 (0.83%) were to WWE. Of the 632, 250 (0.33%) were taking AEMs when they first presented for antenatal care. The characteristics of the study population are shown in Table 1.

Table 1: Characteristics of Study Population (n=250).

	Percentage (%)	Number (n)
Nulliparas	40.0%	100
Multiparas	60.0%	150
Age >34 years	38.8%	97

Unplanned pregnancy	24.4%	61
Infertility treatment	6.8%	17
Obesity (BMI >29.9kg/m^2)	21.2%	53

Table 2: Characteristics of general Hospital Population 2013-2022 (n=75,869).

	Percentage (%)	Number (n)
Nulliparas	40.2%	30,499
Multiparas	59.8%	45,370
Age >34 years	35.4%	26,896
Unplanned pregnancy	25.3%	19,228
Infertility treatment	6.0%	4,571
Obesity (BMI >29.9 kg/m^2)	17.6%	13,354

The study population was similar to that of the general obstetric population of the hospital (Table 2). Of the 250, 28 (11.2%) were current smokers and 8 (3.2%) women reported a family history of a NTD. Twenty-five (10.0%) reported they were unemployed at presentation compared with an unemployment rate in the hospital population of 21.5% in 2016 which fell to 16.9% in 2022 [23]. Of the 250, 84 (33.6%) had been trying to conceive for more than one year and 84 (33.6%) gave a history of a previous miscarriage.

Three (1.2%) women took no folic acid before or after conception and 59 (23.6%) took it only after conception. Of the 188 who took folic acid before conception, 164 took the 5mg high-dose and 24 took the 0.4mg low-dose supplement: overall 65.6% of the 250 WWE taking AEMs followed the national guideline of preconception high-dose supplementation. Of the 188 who took preconceptual folic acid, 49 stopped during the pregnancy.

Of the 100 nulliparas, 34 (34.0%) did not follow the guidelines compared with 57 (38.0%) in 150 multiparas. There was no relationship between compliance rates and maternal age, BMI category or occupation. Of the 250 women, 189 (75.6%) had planned their pregnancy. Of the 61 (24.4%) with unplanned pregnancies, 17 reported failed contraception and 7 of these had been taking oral contraception. Of these 7 women, five were prescribed levetiracetam, one lamotrigine and one lacosamide. Of those who planned their pregnancy, 17 required assisted reproduction. The compliance rate with the national guidelines was 123 (65.1%) in the 189

women who had planned their pregnancy compared with 46 (75.4%) in the 61 that had not (NS).

Table 3: List of antiepileptic medications taken in early pregnancy by women with epilepsy (n=250).

Name of Medication	Percentage (%)	Number (n)
Lamotrigine	33.8%	98
Levetiracetam	30.7%	89
Carbamazepine	12.4%	36
Sodium Valproate	5.5%	16
Pregabalin	2.4%	7
Topiramate	2.4%	7
Lacosamide	2.4%	7
Brivaracetam	2.1%	6
Other (<5 women)	8.3%	24
Total	100%	290

*38 women were prescribed two medications, and one woman was prescribed three for seizure control.

Monotherapy was prescribed in 211 (84.4%) women and polytherapy in 39 (15.6%) (Table 3). In the first 5 years of the study, 12 out of 121 (9.9%) were prescribed polytherapy compared with 27 out of 129 (20.9%) in the second five years (NS). In the first 5 years, 9 out of 121 (7.4%) were prescribed valproate compared with 7 out of 129 (5.4%) in the second five years (NS). There were no changes in the type of AEMs prescribed over the decade.

Of the 16 women prescribed valproate, 12 (75.0%) had planned their pregnancy but only five (31.3%) followed the folic acid guidelines; 13 were prescribed valproate as monotherapy and three as polytherapy. The compliance rate for valproate as monotherapy was four out of 13 (30.8%) compared with 149 out of 198 (76.0%) for other WWE taking monotherapy ($p<0.03$). Four (25%) of the women taking valproate were immigrants from developing countries. The compliance rate in the 211 women receiving monotherapy was 153 (72.5%) compared with 10 (25.6%) in 39 women receiving polytherapy ($p<0.03$).

Discussion

Over the 10 years two-thirds (65.6%) of the 250 WWE taking AEMs in early pregnancy followed the national guidelines on preconceptual high-dose folic acid supplementation for the prevention of NTDs. A particular concern is the suboptimal compliance in women prescribed valproate and in women prescribed polytherapy because these two cohorts are at even higher risk of a pregnancy complicated by NTD than other WWE on AEMs [24]. No maternal characteristics were found that could predict which women followed the guidelines.

The prevalence of WWE of 0.83% in our study is consistent with the 0.86% estimated for the female population in Ireland in 2005 [22]. For women aged 25-34 years the estimated prevalence was 0.78-0.85% and aged 35-44 years it was 0.96-1.05%. This 2005 Irish study was the first national study published in Europe on the prevalence of epilepsy in adults.

Over 90% of WWE will deliver a healthy baby [11]. The risk of major congenital malformations in WWE depends on the type of drug prescribed, the number of drugs and the dosage. The risk reaches 2.6%-5.5% in women taking carbamazepine. It increases to 6.7%-10.3% for valproate and 15.0% with polytherapy [14]. It is notable that the risks cited in the North American register are lower than in the European registers. This may be explained by mandatory folic acid food fortification which has been linked with a 19% decrease in NTDs in the United States [25].

A recent systematic meta-analysis of monotherapy for epilepsy in pregnancy identified 49 high-quality studies which had examined all congenital malformations in the child [26]. The increased risk with valproate was highest consistently across comparisons with other monotherapy AEMs with the absolute risk ranging from 5%-9%. The risk of malformation was also increased for carbamazepine, phenobarbital, phenytoin, lamotrigine and topiramate. For certain AEMs, the risk was dose-dependent, particularly valproate. For other AEMs, particularly newer medications, data was limited. The review did not confine itself to NTDs and did not review folic acid supplementation.

In 2022, the Medicines and Healthcare products Regulatory Agency (MHRA) published a 2018-2021 analysis of AEMs in females aged 0-52 years in England [27]. There was a decrease in valproate prescribing annually but 20,192 were still prescribed valproate in September 2021. Of these 8,107 were aged 16-44 years and 208 had started that month. A total of 832 women

studied had 938 pregnancies and 247 of them were prescribed valproate in the month they became pregnant. Of these, 130 received polytherapy.

In 2023, the MHRA published a report recommending that no patient (male or female) under 55 years should be prescribed valproate unless recommended by two specialists because of the risks of congenital malformations in pregnancy and of male infertility [24]. It recommended that valproate should only be taken during pregnancy or in women of childbearing age if other AEMs were ineffective or not tolerated. This recommendation, however, has proved to be controversial [28].

The report focused on overall major congenital malformations and found that the risks of polytherapy with AEMs, including valproate, during pregnancy carried a higher risk of malformations than polytherapy of AEMs without valproate. The report also found that valproate increases the risk of neurodevelopmental problems which persists through all three trimesters.

The MHRA report accepted that periconconceptual folic acid reduced the background risk of NTDs but could identify no evidence that folic acid prevented valproate-induced NTDs. To answer this research question would require a large, multi-centre study where WWE prescribed valproate before pregnancy and for the first six weeks of pregnancy would be randomised to different doses of folic acid. Such a study would be expensive to conduct and challenging to power statistically. Prenatal exposure to dual therapies increased the risk of neurodevelopmental disorders. There is also evidence that folic acid supplementation in the first trimester may reduce the risk in women taking AEMs of neurodevelopmental disorders [14].

In general, compliance with folic acid supplementation guidelines internationally is suboptimal [3,13]. An Irish study of 42,362 women presenting for antenatal care in 2009 to 2013 found that 18,473 (43.9%) had started folic acid before pregnancy [21]. In the same hospital 3,715 (50.7%) of the women started folic acid before pregnancy in 2022 [23].

In a Polish study of 1455 women at high risk of fetal anomalies, 46.8% reported taking folic acid before pregnancy [29]. In a recent 10-year Israeli study in 282 non-pregnant WWE of childbearing age taking AEMs, 22% were taking folic acid supplements [30]. Previous

hospital-based studies have not examined the dosage and timing of folic acid supplementation in pregnant WWE prescribed AEMs.

In an observational study using the UK and Ireland Register, 1935 (44.3%) of the 4365 WWE who became pregnant took preconception folic acid [31]. There were eight cases of a NTD in the 1935 compared with eight cases in the 2,430 who did not take preconceptual folic acid (NS). Details on folic acid dosing and timing was limited. In a case-control report from the Slone Birth Defects Study (1988-2015), women prescribed AEMs were more likely to have a baby with a NTD [32].

The epidemiology of NTDs in babies born to WWE is challenging [33]. National registers for either NTDs or epilepsy are limited, especially in developing countries. There are variations in the findings from different international registries due to differences in methodology. In general, registry data is inferior to that collected in a clinical trial and may be subject to bias and variability. Large registries, however, do serve an important function because they may identify adverse outcomes early for newer AEMs.

The optimal dose of preconceptual folic acid in WWE at high risk of a NTD remains unknown [12]. The 5mg dose is recommended in Europe, including Ireland [10]. The American College of Obstetricians and Gynecologists recommends 4mg based on the MRC RCT [25]. The optimal dose has not been determined in either low-risk or high-risk women.

The spectrum of opinion is wide with the principal investigator in the MRC RCT advocating high-dose for all women and others arguing against high-dose even for women at increased risk [3,15]. However, there is a strong theoretical rationale for prescribing high-dose folic acid for women taking AEMs that have antifolate properties [15]. A study of 68 adults taking various AEMS found that serum concentrations of both folate and vitamin B12 were lower ($p<0.05$) in adults than they were before starting medications [34]. Potentially the dose of folic acid could be titrated based on maternal plasma concentrations.

There is a lack of RCTs in women at increased risk of a NTD due to maternal medical disorders, including WWE, and future studies may not be large enough to find the optimum dose statistically. Furthermore, there may be a need for higher doses in Europe because, unlike North America, it has not implemented mandatory folic acid food fortification [35]. There are plans

to fortify non-wholemeal flour with folic acid in the United Kingdom (UK), but not in the European Union.

The optimal dose may also vary according to the type of AEM. Enzyme-inducing AEMs increase the risk of folate deficiency during pregnancy, in particular, phenytoin and carbamazepine. Valproate, although not enzyme-inducing, disrupts folate absorption and folate-dependent coenzymes, which may explain why it is associated with the highest risk of NTDs in WWE [37]. An Irish 1995 to 2016 review identified 29 cases of fetal valproate syndrome, including three with NTDs [38]. Of the 29, 52% experienced developmental delay and 40% speech delay.

Apart from preventing NTDs, folic acid supplementation may have other benefits in WWE. There is evidence from an International Register on AEMs and pregnancy that spontaneous miscarriage was lower in women prescribed high-dose folic acid compared with those taking low-dose folic acid [10]. This also supports the recommendation that women should continue folic acid in the first trimester.

The suboptimal folic acid compliance in the cohorts of women prescribed valproate and polytherapy is a particular concern. It is unexplained because the hospital does not have access to the clinical records of the woman's neurologist or general practitioner. One possible explanation is that these cohorts of WWE did not have a recent review of their medications by a neurologist and thus, were not advised to change their AEMs or about the importance of high-dose folic acid. The four immigrants taking valproate, for example, may never have been reviewed by a neurologist in Ireland.

A UK survey of 144 healthcare professionals from 94 hospitals highlighted the fragmentation of care between the maternity and neurology services for WWE with less than a third of hospitals providing joint obstetric/neurology clinics [39]. Another advantage of joint clinics is that it helps dose adjustments. While carbamazepine and lamotrigine appear to be the safer options during pregnancy, lamotrigine is associated with pharmacokinetic changes in drug levels which can lead to breakthrough seizures.

The sociodemographic characteristics were similar in the study population compared with the hospital population. In particular, the 24.4% incidence of unplanned pregnancy was similar. In

the previous study from the hospital between 2009 and 2013, the women most likely to comply with recommendations for preconceptual low-dose folic acid were those who planned their pregnancy, who were more than 30 years old, non-obese, Irish-born and employed professionally [21].

The 24.4% incidence of unplanned pregnancies in our study was lower than the 55% incidence reported in a Scottish report on WWE which was similar to other studies [17,40]. If oral contraceptives are taken correctly, the failure rate in healthy women is 1% and in WWE is 3%-6% [40]. It is not known how many WWE use oral contraception in Ireland but seven (11.5%) of the unplanned pregnancies in this study were relying on oral contraception.

Interactions between AEMs and oral hormonal contraceptives are important because there is a risk of both failed contraception and reduced seizure control. The risk of failed contraception is more likely with AEMs that induce liver enzymes. The Clinical Programme for Epilepsy in Ireland recommends additional barrier contraception for women taking a combination of oral contraception and AEMs [20]. Based on the results of our study, there is a strong case for WWE taking oral contraceptives also being prescribed high-dose folic acid.

Only three (1.2%) women in our study took no folic acid compared to 6.6% of the obstetric population in the 2009-2013 study. No associations were found between maternal sociodemographic characteristics and folic acid consumption. However, the number of WWE who did not follow national recommendations was only 81 women in our study compared with 23,569 in the general obstetric population in the previous study [21]. The study, therefore, was not powered statistically to identify maternal characteristics that might predict compliance with the guidelines.

In this study, no annual trends were observed in AEMs prescribed. In a Spanish study from the EURAP register of AEMs prescribed as monotherapy in pregnancy, 2008-2015 was compared with 2001-2007 [41]. There was an increase in levetiracetam and lamotrigine prescribing and a decrease in carbamazepine, phenytoin and phenobarbital. There is no single source of Irish data about valproate prescribing during pregnancy and there are confidentiality barriers which inhibit the linking of datasets. This is the first hospital-based study to provide data on AEMs in pregnant women in Ireland and can be used as a baseline for future research studies.

In July 2020, a summary of a three-month Irish survey of WWE taking valproate was published [42]. There have been several public health initiatives nationally about prescribing valproate in women of childbearing age since 2014. Of the 152 respondents, 83% were aware that valproate caused serious birth defects and 66% were aware it caused neurodevelopmental problems in the offspring. However, 29% had never discussed the risks with a healthcare professional, only 27% were aware of the Pregnancy Prevention Programme and only 30% reported that they received the Patient Alert Card when renewing their prescription. Communication gaps between WWE taking valproate and their general practitioners, neurologists and community pharmacists were shown.

A strength of our study is that approximately one in eight deliveries nationally were to women who gave birth in the hospital and over a decade there was a relatively large cohort of pregnant women taking AEMs. The study also included women from diverse backgrounds socioeconomically. Our findings, therefore, may be applicable in other settings nationally and internationally.

A potential weakness of the study is that the information on the women's medications was self-reported, and compliance was not verifiable. A limitation is that we do not have the details of the paediatric follow-up and thus, the risk of congenital anomalies or autism in the 250 babies was not determined. A further limitation is that the data included women who had more than one baby during the decade. The exact number cannot be determined because the data was anonymised and, furthermore, it cannot be assumed that their AEMs or supplementation compliance was the same for each pregnancy.

All the women studied required a prescription for their AEMs in the months before they became pregnant, usually from a neurologist or general practitioner. Thus, in a third of the women studied an opportunity was missed to add folic acid 5mg to the prescription of WWE who could become pregnant in the near future. As a result, the risk of a preventable NTD remained high in women who already face healthcare challenges with seizure management. A recent UK drug utilisation study found that the prescribing of AEMs in pregnancy had increased to 1.4% in 2018 compared with 0.6% in 1995 driven largely by indications other than epilepsy [43].

Conclusions

The key findings in this study should be communicated to WWE in their reproductive years and to their doctors, but also to advocacy groups and public health policymakers. The suboptimal folic acid supplementation compliance in women prescribed valproate or polytherapy highlights the need for a preconceptual review of AEMs by a neurologist for WWE who may become pregnant, whether intentionally or not, in the near future. At the neurology consultation, the opportunity should be taken to advice the woman about future contraception [44]. This ideal will be challenging to implement in Ireland where there is a shortage of neurologists and a growing waiting list for outpatient consultations, even for patients presenting with seizures [45].

There is an opportunity to achieve seizure control by changing valproate to a safer option or changing to monotherapy as recommended in Irish and Scottish reports [17,20]. WWE may also be identified who are prescribed AEMs that induce liver enzymes which interact with oral contraceptives. They too should be prescribed high-dose folic acid. Ideally WWE prescribed AEMs should be reviewed during pregnancy in a joint obstetric/neurology antenatal clinic, supported by a pharmacist, particularly if the AEMs require adjustment.

Word count 4155.

List of abbreviations

AEMs Antiepileptic medications
BMI Body Mass Index
MHRA Medical and Healthcare products Regulatory Authority
MRC Medical Research Council
NTDs Neural Tube Defects
RCTs Randomised Control Trials
UK United Kingdom
WWE Women with epilepsy

Authors' contributions

CT cleaned, coded and analysed the data and wrote the first drafts. TMcI contributed to study design and to editing the texts. JH contributed to the clinical analysis. DG and MG contributed to statistical analysis and editing the text. AO'H contributed to data collection and analysis, to editing and finalising the text. All authors read and approved the last version.

Funding
No funding was received for this study.

Ethics approval
Research Ethics Committee, The Coombe Hospital, Dublin.
Research Ethics Committee, School of Pharmacy and Life Sciences, Robert Gordon University (no. S351).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

References

1.Avagliano L, Massa V, George TM et al.
Overview on Neural Tube Defects: from development to physical characteristics.
Birth Defects Res 2019; 111:1455-67.

2. Zaganjor L, Sekkarie A, Tsang BL et al
Describing the Prevalence of Neural Tube Defects Worldwide: A Systematic Literature Review
Plos One 2016; 11(4):e0151586.

3. Wald NJ 2022.
Folic acid and neural tube defects: Discovery, debate and the need for policy change.
J Med Screening 2022; 29: 138-146.

4. Medical Research Council Vitamin Study Research Group.
Prevention of neural tube defects: results of the Medical Research Council Vitamin Study.
Lancet 1991; 338:131-137.

5.Czeizel AE, Dudas I.
Prevention of the first occurrence of neural tube defects by periconceptual vitamin supplementation.
N Eng J Med *1992;* 327:1832-1835.

6.Viswanathan M, Urrutia RP, Hudson KN et al.
Folic Acid Supplementation to Prevent Neural Tube Defects: A Limited Systematic Review Update for the U.S. Preventive Services Task Force. Evidence Synthesis No. 230.
JAMA 2023; 330:460-466.

7. Perucca E, French JA, Aljandeel G et al.
Which terms should we use to describe medications used in the treatment of seizure disorders? An ILAEV position paper.
Epilepsia 2024; 65:533-541.

8. Li Y, Meador KJ.
Epilepsy and Pregnancy.

Continuum (Minneap Minn) 2022; 28:34-54.

9. Weston J, Bromley R, Jackson CF et al.
Monotherapy treatment of epilepsy in pregnancy: congenital malformation outcomes in the child.
Cochrane Database Syst Review 2016; 11: CD010224.

10. Nucera B, Brigo F, Trina E et al.
Treatment and care of women with epilepsy before, during, and after pregnancy: a practical guide.
Ther Adv Neurol Disord 2022; 15:1-31.

11. Blaszczyk B, Miziak B, Pluta R et al.
Epilepsy in Pregnancy – Management Principles and Focus on Valproate.
Int J Mol Sci 2022; 23:1369.

12. Wilson RD, O'Connor DL.
Maternal folic acid and multivitamin supplementation: International clinical evidence with considerations for the prevention of folate-sensitive birth defects.
Preventitive Medicine Reports 24 2021; 101617.

13. McAvoy H.
Folic Acid Supplementation.
Department of Health Folic Acid Policy Committee. Ireland. 2019.

14. Tomson T, Battino D, Bromley R et al.
Management of epilepsy in pregnancy: a report from the International League against Epilepsy Task Force on Women and Pregnancy.
Epileptic Disord 2019 6: 497-517.

15. American College of Obstetricians and Gynecologists.
Prepregnancy Counseling. Committee Opinion no.762.
Fertil Steril 2019; 111:32-42.

16. Dwyer ER. Filion KB, MacFarlane AJ et al.
Who should consume high-dose folic acid supplements before and during early pregnancy for the prevention of neural tube defects?
BMJ 2022; *337e: 0*67728.

17. Stephen L, Frier E, Leavy Y et al.
Standards of Care for Women with Epilepsy of Childbearing Age.
Scottish Government Obstetric Neurology Working Group.
https://perinatalnetwork.scot/wp-content/uploads/2023/02/2023-02-21-Pregnancy-Standards-of-Care-for-Women-with-Epilepsy.pdf [accessed online September 24, 2024].

18. McDonnell R, Delany V, O'Mahony MM et al.
An Audit of Neural Tube Defects in the Republic of Ireland 2012-2015.
IMJ 2018; 111;712. PMID 30376230.

19. O'Connor R, Moriarty T, Moran M et al.
Good Practice Points. Epilepsy in Adults.
Irish College of General Practitioner Quality and Safety in Practice Committee. July 2020
https://www.icgp.ie/speck/properties/asset/asset.cfm?type=LibraryAsset&id=69767A51%2D2F43%2D4CE3%2D9F9AA55F76779755&property=asset&revision=tip&disposition=inline&app=icgp&filename=Good%5FPractice%5FPoints%5F%2D%5FEpilepsy%5Fin%5FOLICACIDdults%5FSummary%2Epdf [accessed online September 24, 2024].

20. Royal College of Physicians in Ireland.
Practice Guide for the Management of Women with Epilepsy. July 2018
https://www.hse.ie/eng/about/who/acute-hospitals-division/woman-infolicacidnts/clinical-guidelines/practice-guide-for-mgt-of-women-with-epilepsy.pdf [accessed online September 24, 2024].

21. McKeating A, Farren M, Cawley S et al.
Maternal folic acid supplementation trends 2009-2013.
Acta Obstet Gynecol Scan 2015*;* 94:727-733.

22. Brainwave: The prevalence of epilepsy in Ireland. Summary Report. May 2009.

The Irish Epilepsy Association
https://www.epilepsy.ie/sites/www.epilepsy.ie/files/imported/other_articles/072574A1-DC3F-979A-39B33BD6D4954E1D.pdf [accessed online September 25, 2024].

23. Annual Clinical Reports 2013-2022.
The Coombe Hospital, Dublin, Ireland.
https://static1.squarespace.com/static/5df2275ab592b63d2d87fd51/t/655de5dbbab60e7abee6c9folic_acid/1700652545560/The+Coombe_Annual+Report+2022_Final+Web.pdf [accessed online September 25, 2024].

24. Medicines and Healthcare Products Regulatory Agency
Valproate: review of safety data and expert advice on management of risks.
Public Assessment Report November 2023
https://www.gov.uk/government/publications/valproate-review-of-safety-data-and-expert-advice-on-management-of-risks [accessed online September 25, 2024].

25. American College of Obstetricians and Gynecologists.
Neural Tube Defects. Practice Bulletin. 2017.
Obstet Gynecol 2017; 130:e279-285.

26. Bromley R, Adab A, Bluett-Duncan M et al.
Monotherapy treatment of epilepsy in pregnancy: congenital malformation outcomes in the child (Review).
Cochrane Database of Systematic Reviews 2003 Issue 8. CD010224.
https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD010224.pub3/full
[accessed online September 25, 2024].

27. Bullard I.
Antiepileptic use in females aged 0 to 54 in England (April 2018 to September 2021).
Medicines and Pregnancy Register Published March 31, 2022.
https://digital.nhs.uk/data-and-information/publications/statistical/mi-medicines-and-pregnancy-registry/antiepileptic-use-in-females-aged-0-to-54-in-england-april-2018-to-september-2021 [accessed online September 25, 2024].

28. Editorial.
Reducing the uses of valproate: a controversial decision.
Lancet Neurology 2024; 23:123.

29. Wojtowicz A, Babczyk D, Galas A et al/
Evaluation of the prevalence of folic acid supplementation before conception and through the first 12 weeks of pregnancy in Polish women at high risk of fetal anomalies.
Ginekol Pol 2022; 93:489-495.

30. Gandelman-Martin R, Theitler J.
Folic acid supplementation in women of childbearing with epilepsy: No association with type or number of antiepileptic drugs.
Birth Defects Res 2024; 116: e2283.

31. Morrow JI, Hunt SJ, Russell AJ et al.
Folic acid use and major congenital malformations in offspring of women with epilepsy: a prospective study from the UK Epilepsy and Pregnancy Register.
J Neurol Neurosurg Psychiatry 2009; 80:506-511.

32. Petersen JM, Parker SE, Benedum CM et al.
Periconceptual folic acid and risk for neural tube defects among higher risk pregnancy
Birth Defects Res 111:1501-1512.

33. Tornes L, Harden CL.
Antiepileptic drug pregnancy registries: do the latest concur?
Therapy 2010; 7:517-526.

34. Huang H, Zhou H, Wang N et al.
Effects of antiepileptic drugs on the serum folate and vitamin B12 in various epileptic patients.
Biological Reports 2016; 5:413-416.

35. Sadat-Hossieny Z, Robalino CP, Pennell RB et al.
Folate Fortification of Food: Insufficient for women with epilepsy.
Epilepsy Behav 2021; 117:107688.

36. Broughan JM, Martin D, Higgins T et al.
Prevalence of neural tube defects in England prior to the mandatory fortification of non-wholemeal flour with folic acid: a population-based cohort study.
Arch Dis Child 2023*; 325856.*

37. Ragueneau-Majlessi I, Levy RH, Janik F.
Levetiracetam does not alter the pharmacokinetics of an oral contraceptive in healthy women.
Epilepsia 2002*;* 43:697-702.

38. Yunos HM, Green A.
Fetal Valproate Syndrome: the Irish experience.
IMJ *2018;* 187:965-968.

39. Taylor E, Junaid F, Khattak H et al.
Care of pregnant women with epilepsy in the United Kingdom: A national survey of healthcare professionals.
Eur J Obstet Gynecol Reprod Biol 2022*;* 276:47-55.

40. Reimers A, Brodtkorb E, Sabers A.
Interaction between hormonal contraception and antiepileptic drugs: Clinical and mechanistic considerations.
Seizure 2015; 28:66-70.

41. Ferri MM, Mayor PP, Lopez-Fraile P.
Comparative study of antiepileptic drug use during pregnancy over a period of 12 years in Spain. Efficacy of the newer antiepileptic drugs lamotrigine, levetiracetam and oxcarbazepine.
Neurolgia 2018*; 33:78-84.*

42. Epilepsy Ireland Valproate Survey 2020.
Summary Findings July 2020 of Survey December 2019-February 2020.
https://www.epilepsy.ie/sites/www.epilepsy.ie/files/2020%20Epilepsy%20Ireland%20Valproate%20Survey%20results%20FINAL.pdf [accessed online September 25, 2024].

43. Madley-Dowd P, Rast J, Ahlqvist VH et al.
Trends and patterns of antiseizure medication prescribing during pregnancy between 1995 and 2018 in the United Kingdom: A cohort study.
Br J Obstet Gynaecol 2024; 131:15-25.

44. Stephen L, Tomson T, Harden C et al.
Management of epilepsy in women.
Lancet Neurology 2019 18:481-491.

45. Dwyer B, Flynn C, Murphy S.
A Quality Improvement Project on Seizure Referral Triage: Preliminary Outcomes and Implications for Further Research.
IMJ 2022; 115:600-604.

TABLES

Table 1: Characteristics of Study Population (n=250).

	Percentage (%)	Number (n)
Nulliparas	40.0%	100
Multiparas	60.0%	150
Age >34 years	38.8%	97
Unplanned pregnancy	24.4%	61
Infertility treatment	6.8%	17
Obesity (BMI >29.9kg/m^2)	21.2%	53

Table 2: Characteristics of Hospital Population 2013-2022 (n=75,869).

	Percentage (%)	Number (n)
Nulliparas	40.2%	30,499
Multiparas	59.8%	45,370
Age >34 years	35.4%	26,896
Unplanned pregnancy	25.3%	19,228
Infertility treatment	6.0%	4,571

Obesity (BMI >29.9 kg/m^2)	17.6%	13,354

Table 3: List of antiepileptic medications taken in early pregnancy by women with epilepsy (n=250).

Name of Medication	Percentage (%)	Number (n)
Lamotrigine	33.8%	98
Levetiracetam	30.7%	89
Carbamazepine	12.4%	36
Sodium Valproate	5.5%	16
Pregabalin	2.4%	7
Topiramate	2.4%	7
Lacosamide	2.4%	7
Brivaracetam	2.1%	6
Other (<5 women)	8.3%	24
Total	100%	290

*38 women were prescribed two medications, and one woman was prescribed three for seizure control.

Appendix C. Obesity Trends

Target Journal: BMJ Journal of Epidemiology and Community Health

Title: Trends in prevalence of and risk factors for obesity during pregnancy in Ireland: Longitudinal evidence from a large tertiary maternity hospital.

AUTHOR LIST

Ellen Cosgrave[1] (joint first author), Mark Germaine[2 3,4](joint first author), Peter Naughton[5] Margaret M. Brennan[6], Patricia M. Kearney[7], Graham Healy[2], Brendan Egan[3], Michael Turner[8], Claire M. Buckley[7], Amy C O'Higgins[8]

AFFILIATIONS

[1] Child Health Programme, Health Service Executive, Dublin, Ireland

[2] School of Computing, Dublin City University, Dublin 9, Ireland

[3] School of Health and Human Performance, Dublin City University, Dublin 9, Ireland

[4] Research Ireland Centre for Research Training in Machine Learning, Dublin City University, Dublin 9, Ireland

[5] Global Health Programme, Health Service Executive, Dr Steeven's Hospital, Dublin 8

[6] TCD Institute of Population Health, Russell Centre, Tallaght Cross West, Tallaght, Dublin 24

[7] School of Public Health, University College Cork

[8] UCD Centre for Human Reproduction, The Coombe Hospital, Dublin 8, Ireland

ORCIDs

Ellen Cosgrave 0000-0003-1802-5161

Mark Germaine 0000-0002-7862-7714

Amy O'Higgins 0000-0002-2020-1585

Brendan Egan 0000-0001-8327-9016

Graham Healy 0000-0001-6429-6339

CORRESPONDING AUTHOR

Dr Ellen Cosgrave

ABSTRACT WORD COUNT: 250

WORD COUNT: 2996

REFERENCES: 27

TABLES: 3

FIGURES: 2

SUPPLEMENTARY TABLES: 7

SUPPLEMENTARY FIGURES: 0

What is already known on this topic - Obesity during pregnancy is linked to short and long term adverse maternal and neonatal outcomes including gestational diabetes mellitus, pre-eclampsia, increased delivery interventions and greater healthcare costs and service burden. Despite this knowledge, there is a lack of recent, objectively measured data on the prevalence of obesity during pregnancy in the Republic of Ireland, necessitating this study.

What this study adds - This study provides up to date data indicating a significant increase in the prevalence of obesity during pregnancy from 2013 to 2022. It offers insight into obesity trends among young adults, allowing for broader societal inferences. It also identifies key factors associated with obesity during pregnancy, including maternal age, parity and ethnicity.

How this study might affect research, practice or policy - The findings highlight the need for both universal and targeted strategies to address the rising trend of obesity during pregnancy. Policymakers and healthcare providers should use findings to inform population health policies, service planning and resource allocation both within maternity services and across the broader young adult population. This up-to-date data can also support benchmarking with other European countries and long-term planning to manage obesity-related complications and healthcare demands beyond maternity services.

Abstract: 250

Background

This study investigates the prevalence and trends of obesity during pregnancy (BMI >29.9kg/m²) among women attending a large tertiary maternity hospital in Dublin, Ireland, in which 1 in 8 deliveries occur nationally.

Methods

The study involved secondary analysis of electronic health records from The Coombe Hospital from 2013-2022. Body Mass Index (BMI) was calculated using objective measurement of height and weight. Trends in prevalence of obesity during pregnancy and associated 95% Confidence Intervals were estimated by year. Changes in prevalence were assessed statistically. Linear and logistic regression were used to examine risk factors associated with obesity during pregnancy.

Results

There were 74,233 pregnancies. There was a significant increase in mean BMI from 25.5kg/m² in 2013 to 26.5kg/m² in 2022. Prevalence of obesity during pregnancy rose from 16.3% (95%CI 15.2%-17.5%) in 2013 to 22.3% (95%CI 22.1%-23.5%) in 2022, a 36.8% relative increase ($p<0.001$). Prevalence of BMI ≥39.9 kg/m² increased from 1.7% (95%CI 1.4%-2.0%) to 2.5% (95%CI 2.1%-2.9%) ($p<0.001$), a 47.1% relative increase while the prevalence of optimal range BMI (18.5-24.9 kg/m2) decreased. Obesity prevalence increased across all age groups, with the highest rise among those aged 20-24 years. Elevated BMI was positively associated with maternal age, parity, ethnicity, lower skills level, unplanned pregnancy, psychological problems and smoking.

84 **Conclusion**

85 This study offers current estimates of obesity during pregnancy in Ireland, allowing
86 comparisons with European trends. The rising obesity rates across all age groups have
87 significant public health implications. Findings will guide healthcare planning and inform
88 universal and targeted public health strategies.

89

Introduction 575

Obesity (BMI>29.9kg/m²) is an escalating public health concern, especially in high-income countries.[1] Rates during pregnancy are increasing and associated with increased healthcare interventions, costs and adverse short and long-term maternal and child outcomes.[2,3] Risks include gestational diabetes, pre-eclampsia and increased delivery interventions for mothers and increased neonatal complications for babies.[4,5,6,7] Notably, 30-50% of women with obesity develop gestational diabetes, conferring a 60% lifetime risk of type 2 diabetes.[8,9] The intrauterine environment also influences longer-term child health, with metabolic, inflammatory and epigenetic changes induced by obesity during pregnancy associated with higher risk of later-life obesity, type 2 diabetes and cardiovascular disease.[10,11]

Pregnancy represents a "teachable moment" to influence health behaviours and improve family health.[12] Given high healthcare utilisation during this period, pregnancy represents an opportune data capture point to examine broader population health trends; large scale, objective BMI measurement can serve as a proxy indicator of BMI trends in the wider population.

There is limited up-to-date, objectively measured data on prevalence and risk factors for obesity during pregnancy in Ireland.[13] Such evidence is essential to identify high-risk subpopulations, support effective population-level interventions, and enable international comparisons. This study aims to address this gap by analysing trends and risk factors for obesity during pregnancy at a large tertiary maternity hospital in Dublin, Ireland, from 2013 to 2022.

Methods 539

Study design and setting

This retrospective cohort study involved a secondary analysis of electronic health records of women attending the Coombe Hospital in Dublin, Ireland from 2013-2022.[14] Longitudinal analysis of prevalence of obesity during pregnancy was conducted. Factors associated with obesity during pregnancy were assessed cross-sectionally by year. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines were followed.

Study population

All those who attended for their first antenatal visit at the Coombe Hospital from 2013-2022 and subsequently delivered a baby weighing ≥500g, with complete objectively measured height and weight data were included. The Coombe is one of the largest maternity hospitals in Europe, serving women from all socioeconomic groups, from urban and rural areas, with and without private health insurance.[13] Approximately 1 in 8 deliveries in Ireland occur in this hospital.[13] Despite being a single-centre study, demographics of the study population align closely with the national population as per Irish National Perinatal Reports.[15]

Variables

Data were routinely computerised by trained midwives using a standardised data collection system as part of the medical records at the first antenatal visit. BMI was calculated based on objective measurement of weight and height and categorised into World Health Organizations (WHO) classifications (see appendix).[16]

Other variables included age, parity, pregnancy type (singleton/multiple), ethnicity, maternal occupation (a proxy for socioeconomic status categorised into five ordinal skill levels as per the International Standard Classification of Occupations (ISCO) classification system[17]: Level 0=unemployed, Level 1 =Elementary Occupations, Level 2 = [Clerical Support Workers, Service and Sales Workers], Level 3=Technicians & Assoc Technicians, Level 4 = Professional & Managers) pregnancy intention, presence of psychological problems (anxiety and/or depression), and cigarette use (see data dictionary, supplementary table 1).

Ethics and data management

The Coombe Hospital Research Ethics Committee (Study No. 06–2023) granted ethical approval. An anonymised data file was provided securely to the research team who analysed the data.

Statistical analysis

Anonymised data were analysed using SPSS version 29.0 and R. Descriptive analyses summarised the study population by year. Obesity prevalence during pregnancy was estimated for the total sample and stratified by key characteristics. Yearly trends were assessed, and differences between 2013 and 2022 were tested using the Chi-squared test. Participant characteristics by BMI category were compared for 2013 and 2022. To account for repeat pregnancies within individuals over the 10-year period, a multivariable linear mixed effects model was used with a woman-level dummy variable as a random effect. This identified factors independently associated with obesity during pregnancy reported as βeta co-efficients, 95% CI, and p-values. Multivariable binary logistic regression identified changes in risk factors between 2013 and 2022, with results reported as adjusted odds ratios (aOR), 95% CI, and p-values. For

the adjusted models, a stepwise backward elimination approach was used, sequentially removing predictors with p-values >0.05 until the model with best fit was achieved.

Results 771

Demographic information

BMI data were available for 74,233 pregnancies among 51,530 women. Those with no data on height or weight were excluded from the final analyses (n=991, 1.3%). Characteristics of missing observations are included in supplementary table 2.

The characteristics of the study population analysed by year are shown in Table 1. Between 2013 and 2022, mean age increased from 31.3 to 32.8 years. Nulliparous pregnancies rose from 38.6% (95%CI 37.5%-39.7%) to 41.6% (95%CI 40.4%-42.8%), a 7.8% relative increase. Multiple pregnancies decreased 14.3% from 4.2% (95%CI 3.8%-4.7%) to 3.6% (95%CI 3.2%-4.1%). The proportion in the highest skill category increased 35.3% from 42.8% (95%CI 41.7%-43.9%) to 57.9% (95%CI 56.7%-59.1%). White European ethnicity declined 3.2%, Asian ethnicity increased 59.3% and Afro-Caribbean ethnicity declined 17.9%. Unplanned pregnancies decreased 19.2%, from 29.2% (95%CI 28.2%-30.3%) to 23.6% (95%CI 22.8%-24.7%). Pregnancies following infertility treatment increased 85.1% from 4.7% (95%CI 4.2%-5.2%) to 8.7% (95%CI 8.0%-9.4%).

Table 1: Participant characteristics by year of first antenatal visit, 2013-2022

	2013 N=7543 %	2014 N=7802 %	2015 N=7792 %	2016 N=7941 %	2017 N=7679 %	2018 N=7790 %	2019 N=7510 %	2020 N=6874 %	2021 N=6830 %	2022 N=6472 %	Total N=74233 %	Change 2013 vs 2022 %
Age (years)												
mean (years)	31.3	31.7	32.0	32.1	32.3	32.4	32.6	32.6	32.9	32.8	32.3	4.8
SD (years)	5.4	5.4	5.3	5.4	5.4	5.4	5.4	5.4	5.3	5.4	5.4	0.0
<20	1.9	1.8	1.8	1.9	1.6	1.3	1.4	1.4	1.1	1.2	1.5	-36.8
20-24	10.1	9.5	8.1	8.2	8.0	8.0	7.4	7.3	6.7	7.5	8.1	-25.7
25-29	22.7	20.6	19.4	18.5	18.3	17.4	17.0	16.6	15.6	16.0	18.3	-29.5
30-34	36.0	36.3	37.3	36.9	34.5	35.2	34.9	35.3	36.0	34.4	35.7	-4.4
35-39	23.8	25.7	27.6	28.0	30.9	31.1	31.3	31.4	31.9	32.1	29.3	34.9
40·	5.5	6.2	5.9	6.6	6.7	7.0	7.9	8.0	8.6	8.8	7.1	60.0
Parity												
Nulliparous	38.6	39.3	38.6	40.0	40.8	42.2	41.6	40.6	39.1	41.6	40.2	7.8
Multiparous	61.4	60.7	61.4	60.0	59.2	57.8	58.4	59.4	60.9	58.4	59.8	-4.9
Pregnancy type												
Singleton	95.8	96.4	95.	95.6	95.4	96.0	95.5	95.9	96.9	96.4	95.9	0.6
Multiple	4.2	3.6	4.3	4.4	4.6	4.0	4.5	4.1	3.1	3.6	4.1	-14.3
Skills level												
0	29.4	27.0	25.4	23.7	22.6	21.0	19.6	18.5	18.5	17.3	22.5	-41.2
1	1.4	1.4	1.2	1.3	1.3	1.3	1.4	1.5	1.4	1.2	1.3	-14.3
2	17.6	18.5	18.1	18.4	16.9	17.8	16.8	15.7	14.1	14.4	16.9	-18.2
3	8.8	8.2	8.5	8.0	8.6	8.4	8.7	9.2	8.3	9.3	8.6	5.7
4	42.8	44.9	46.9	48.6	50.7	51.5	53.4	55.1	57.7	57.9	50.7	35.3
Ethnicity												
White European %	88.3	90.0	90.0	89.9	89.5	89.1	87.9	87.8	87.8	85.5	88.7	-3.2
Asian %	5.4	4.8	4.9	4.4	5.2	5.8	6.2	6.3	6.8	8.6	5.8	59.3
Afro-Caribbean %	2.8	2.8	2.3	2.3	2.1	2.0	2.0	1.9	1.9	2.3	2.2	-17.9
Middle Eastern %	0.5	0.4	0.5	0.6	0.6	0.4	0.6	0.5	0.7	0.5	0.5	0.0
Other %	2.9	2.0	2.2	2.8	2.5	2.7	3.4	3.6	2.8	3.1	2.8	6.9
Pregnancy intention												
Planned %	66.1	66.1	68.2	67.4	68.2	68.8	67.4	69.1	70.9	67.7	67.9	2.4
Unplanned %	29.2	28.8	26.5	26.9	25.5	24.8	24.9	24.7	23.3	23.6	25.9	-19.2
Infertility treatment %	4.7	5.1	5.3	5.7	6.4	6.5	7.7	6.2	5.8	8.7	6.2	85.1
BMI (kg/m2)												
Mean (kg/m2)	25.5	25.4	25.3	25.6	25.8	25.9	26.1	26.2	26.5	26.5	25.9	
SD (kg/m2)	5.1	5.2	4.9	5.1	5.2	5.2	5.3	5.4	5.4	4.4	5.2	
<18.5	2.0	2.4	2.3	2.0	1.8	1.8	1.7	1.6	1.5%	1.6	1.9	-20
18.5 – 24.9	52.9	54.8	55.1	52.1	51.3	50.4	48.6	48.6	45.6	45.3	50.7	-14.3
25.0 – 29.99	28.8	26.8	27.8	29.3	28.8	29.7	30.2	29.6	31.6	30.8	29.3	6.9
30 - 34.99	10.8	10.0	9.9	10.8	11.4	11.9	12.7	12.7	13.9	14.4	11.8	33.3
35.0 – 39.99	3.9	4.3	3.6	4.1	4.7	4.1	4.7	5.2	4.7	5.6	4.4	12.8
40.0 – 44.9	1.1	1.3	0.9	1.2	1.4	1.6	1.6	1.6	2.0	1.7	1.4	54.5
45.0 49.9	0.4	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.5	0.5	0.4	25.0
>/= 50	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.1	50.0

FOOTNOTE

N = count, ISCO = International Standard Classification of Occupations

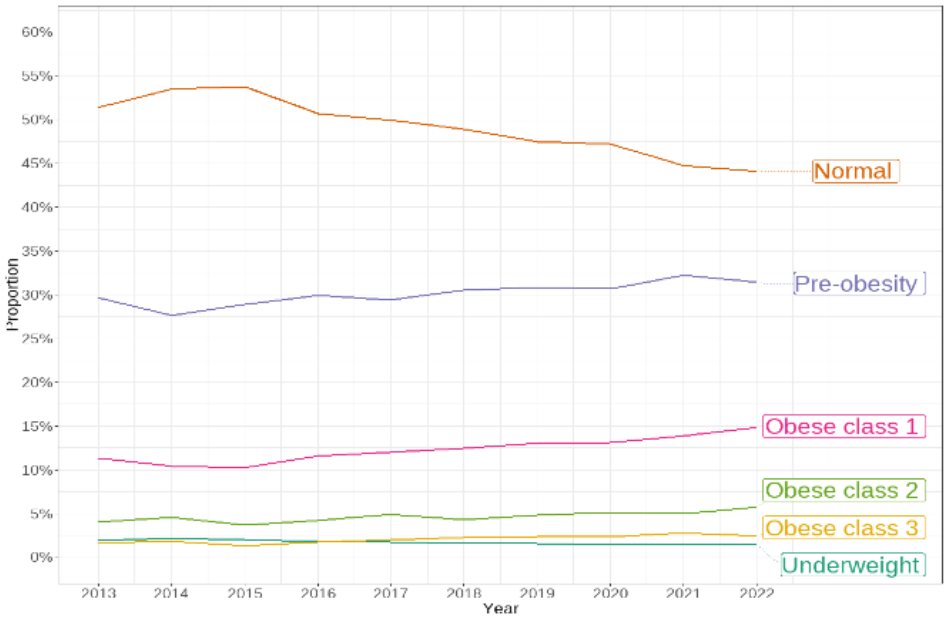


Figure 1: BMI trends over time by BMI category 2013-22

Obesity prevalence

Obesity during pregnancy increased 36.8% from 16.3% (95%CI 15.2%-17.5%) in 2013 to 22.3% (95%CI 22.1%-23.5%) in 2022 ($p<0.001$). Mean BMI increased from $25.5 \pm 5.1 kg/m^2$ to $26.5 \pm 5.4 kg/m^2$. Class III obesity increased 47.1% from 1.7% (95%CI 1.4%-2.0%) to 2.5% (95%CI 2.1%-2.9%), while normal BMI decreased from 51.4% to 44.0% (Figure 1).

Obesity rates increased across all age groups, with largest increases among those aged 20–24 (+65.0%), 30–34 (+39.4%), and 35–39 (+34.3%) (Table 2). Individuals aged 40+ saw a smaller increase (+12.3%). Obesity rose more among nulliparous (+51.6%) than multiparous (+32.6%), and among singleton (+39.0%) compared to multiple pregnancies (+11.9%). Rates rose across all skill levels, with the largest increases in skill levels 1 (+63.0%), 2 (+52.8%), and 3 (+51.9%). Increases were observed in Asian (+6.0%), White European (+46.5%), and Other (+13.7%) groups. Obesity increased among both planned (+39.2%) and unplanned pregnancies (+38.7%), and pregnancies following infertility treatment (23.2%).

Figure 2 outlines how although obesity was initially higher among those aged 40+, rising rates among younger women narrowed this gap by 2022. Obesity remained consistently higher among multiparous. All socio-economic groups saw increases, but the steepest occurred in lower and middle groups, with the middle group diverging from the most affluent over time.

Table 2: Obesity prevalence analysed by participant characteristics, 2013-2022

	2013 N=7543 (%)	2014 N=7802 (%)	2015 N=7792 (%)	2016 N=7941 (%)	2017 N=7679 (%)	2018 N=7790 (%)	2019 N=7510 (%)	2020 N=6874 (%)	2021 N=6830 (%)	2022 N=6472 (%)	Total Sample N=74233 (%)	Change in % 2013-2022	P value
BMI ≥ 29.9 kg/m2	16.3	16.0	14.7	16.6	18.0	18.2	19.5	20.2	21.3	22.3	18.2	36.8	<0.001
Age													
<20	10.6	7.9	8.6	12.8	8.3	10.9	12.1	12.4	9.0	16.9	10.7	59.4	.187
20-24	14.0	15.9	13.0	16.2	17.7	20.3	21.6	23.6	23.9	23.1	18.4	65.0	<0.001
25-29	17.7	17.1	15.3	18.7	21.0	19.4	22.6	22.0	23.6	23.5	19.7	32.8	<0.001
30-34	15.5	14.2	14.8	15.6	16.8	18.6	18.6	20.0	21.1	21.6	17.5	39.4	<0.001
35-39	16.6	17.1	14.5	15.0	18.2	16.1	18.4	19.0	20.2	22.3	17.7	34.3	<0.001
40+	20.3	21.5	18.2	24.6	17.8	21.4	20.2	20.0	21.0	22.8	20.8	12.3	0.337
Parity													
Nulliparous	12.6	13.6	12.0	13.7	15.3	15.4	16.2	18.3	18.0	19.1	15.3	51.6	<0.001
Multiparous	18.6	17.6	16.5	18.6	19.9	20.2	21.9	21.4	23.4	24.6	20.1	32.3	<0.001
Pregnancy type													
Singleton	15.9	16.2	14.6	16.6	18.0	18.0	19.4	20.0	21.3	22.1	18.1	39.0	<0.001
Multiple	24.3	12.0	18.0	15.5	18.8	23.9	22.3	23.2	21.2	27.2	20.4	11.9	0.435
Skills level													
0	20.2	21.4	18.7	20.6	23.1	24.2	23.9	25.6	25.1	27.5	22.5	36.1	<0.001
1	16.5	15.3	15.6	18.1	16.7	15.0	23.4	19.6	22.7	26.9	18.8	63.0	0.089
2	16.3	15.4	16.3	16.4	20.6	19.4	23.6	24.1	25.6	24.9	19.7	52.8	<0.001
3	15.8	14.7	14.8	15.1	18.9	16.3	19.2	18.6	22.5	24.0	17.9	51.9	<0.001
4	13.7	13.4	12.0	14.9	14.8	15.7	16.5	17.5	18.8	19.7	15.8	43.8	<0.001
Ethnicity													
Asian	13.4	11.0	12.6	11.7	14.8	12.5	19.0	14.0	16.9	14.2	14.2	6.0	0.001
Afro-Caribbean	39.4	41.7	33.9	31.3	36.8	35.5	36.9	29.5	37.7	36.1	36.2	-8.4	<0.001
White European	15.7	15.5	14.4	16.6	17.6	18.2	19.2	20.5	21.4	23.0	18.0	46.5	<0.001
Middle Eastern	30.6	27.3	15.8	19.6	30.4	14.7	17.8	22.2	6.5	20.6	20.3	-32.7	0.342
Other	14.6	14.1	12.6	12.8	21.1	19.2	17.7	18.0	20.0	16.6	16.8	13.7	0.575
Pregnancy intention													
Planned	15.3	14.6	13.8	15.5	16.9	16.2	18.4	18.6	20.1	21.3	17.0	39.2	<0.001
Unplanned	18.1	19.6	17.4	19.3	21.3	23.8	22.6	24.8	25.3	25.1	21.4	38.7	<0.001
Infertility treatment	18.1	14.8	12.9	16.2	16.8	17.5	19.1	19.4	18.9	22.3	17.8	23.2	0.124

FOOTNOTE

N = count, ISCO = International Standard Classification of Occupations

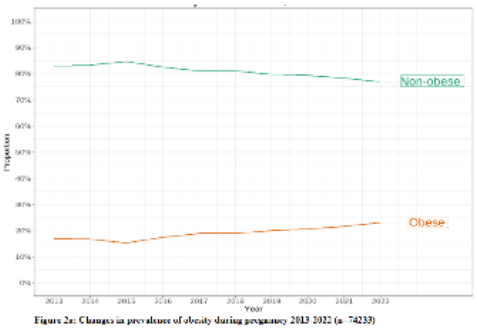


Figure 2a: Changes in prevalence of obesity during pregnancy 2013-2022 (n=74233)

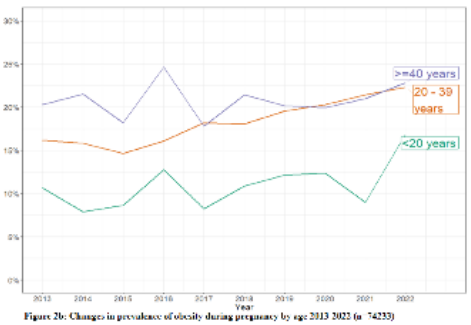


Figure 2b: Changes in prevalence of obesity during pregnancy by age 2013-2022 (n=74233)

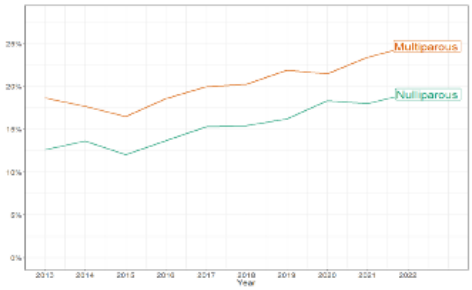


Figure 2c: Changes in prevalence of obesity during pregnancy by parity 2013-2022 (n=74233)

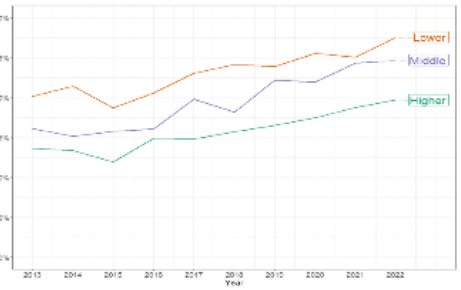


Figure 2d: Changes in prevalence of obesity during pregnancy by socioeconomic group 2013-2022 (n=74233)

As outlined in supplementary table 3, obesity class III increased from 2013 to 2022 across all age groups except 40+, among both nulliparous and multiparous, all skill levels, planned and unplanned pregnancies, those following infertility treatment, and among individuals of White European and Afro-Caribbean ethnicity.

Adjusted analysis

Table 3 presents associations between predictors and BMI from 2013 to 2022. Compared to individuals aged 25–29, those under 20 and 20–24 had lower BMI (β=-2.270 and -0.782, respectively; $p<0.001$), while those aged 30–34, 35–39, and 40+ had higher BMI (βs=0.237-0.352; $p<0.001$), peaking in 40+. Multiparous individuals had higher BMI (β=0.399, $p<0.001$). Compared to skill level 4, level 0 had the highest BMI (β=0.454), followed by levels 1–3 (βs=0.276-0.312; all $p\leq0.020$. Compared to White Europeans, Afro-Caribbean ethnicity was associated with higher BMI (β=2.230), while Asian (β=-0.558) and "Other" ethnicities (β=-0.270) had lower BMI. Unplanned pregnancies (β=0.083, p=0.002), and infertility treatment (β=0.101, p=0.089), were associated with higher BMI, though the latter was not significant. Psychological problems were associated with higher BMI (β=0.471, $p<0.001$). Current and past smoking were associated with higher BMI (β=0.203 and 0.369, respectively; $p<0.001$). BMI increased by an average of 0.214 units per year ($p<0.001$), rising 1.723 units between 2013 and 2022. The intercept (24.720) represents the predicted BMI in 2013 for a woman with the following characteristics: nulliparous, aged 25–29, highest skill level, White European, planned pregnancy, no psychological issues, non-smoker. This increased to 26.443 by 2022, shifting from normal to pre-obesity.

Logistic regression (supplementary table 4) identified how, in 2013, those under 20 (aOR 0.69, 95%CI 0.54–0.88, p=0.003) and aged 20–24 (aOR 0.54, 95%CI 0.30–0.95,

p=0.032) had lower odds of obesity versus ages 25–29. This was not seen in 2022 (aOR 0.91, p=0.503; aOR0.58, p=0.087, respectively). Multiparty was associated with obesity in 2013 (aOR 1.42, 95%CI 1.23-1.64, p<0.001) and 2022 (aOR 1.28, 95%CI 1.12-1.46, p<0.001). Having psychological problems was associated with obesity in 2022 (aOR 1.27, 95%CI 1.11-1.45, p<0.001) but not in 2013 (aOR 1.12, 95%CI 0.95-1.31, p=0.182). Compared to being a non-smoker, being an ex-smoker was associated with obesity in 2022 (aOR 1.39, 95%CI 1.22-1.59, p<0.001) but not in 2013 (aOR 0.95, 95%CI 0.83-1.10, p=0.499). Compared to White European ethnicity, Afro Caribbean ethnicity was associated with obesity in both 2013 (aOR 2.80, 95%CI 2.07-3.79, p<0.001) and 2022 (aOR 2.04, 95%CI 1.43-2.80, p<0.001). Asian ethnicity was inversely associated with obesity in 2022 (aOR 0.62, 95%CI 0.48-0.80, p<0.001) but not 2013 (aOR 0.78, 95%CI 0.58-1.05, p=0.100).

Table 3: Linear mixed effects model output: Participant characteristics, socioeconomic and lifestyle risk factors associated with BMI during pregnancy, 2013-2022

Variable	2013-2022 (N=74,233)		
Age (years)			
<20	-2.270	-2.460—2.081	<0.001
20-24	-0.782	-0.873—0.691	<0.001
25-29	Ref.		<0.001
30-34	0.237	0.173-0.301	<0.001
35-39	0.284	0.200-0.367	<0.001
40+	0.352	0.227-0.477	<0.001
Parity			
Nulliparous	Ref.		
Multiparous	0.399	0.356-0.443	<0.001
Skill level			
0	0.454	0.373-0.536	<0.001
1	0.276	0.043-0.509	0.020
2	0.312	0.228-0.397	<0.001
3	0.266	0.158-0.374	<0.001
4	Ref.		
Ethnicity			
White European 5	Ref.		
Asian 2	-0.558	-0.728—0.389	<0.001
Afro-Caribbean 3	2.230	1.967-2.493	<0.001
Middle Eastern 4	0.083	-0.436-0.603	0.754
Other 1	-0.270	-0.482—0.059	0.012
Pregnancy intention			
Planned	Ref.		
Unplanned	0.083	0.030-0.136	0.002
Infertility treatment	0.101	-0.012-0.217	0.089
Psychological problems			
No	Ref.		
Yes	0.471	0.398-0.544	<0.001
Cigarette use			
Never	Ref.		
Current	0.203	0.090-0.316	<0.001
Past	0.369	0.292-0.446	<0.001
Year			
2013	Ref.		
2014	0.041		0.355
2015	0.110		0.008
2016	0.402		<0.001
2017	0.650		<0.001
2018	0.915		<0.001
2019	1.074		<0.001
2020	1.336		<0.001
2021	1.670		<0.001
2022	1.723		<0.001

Intercept = 24.720; Model adjusted for maternal age, parity, ethnicity, skills level, psychological problems, and smoking status.

Discussion 1111

Key findings

This study assessed trends in obesity during pregnancy among women attending a large Irish maternity hospital from 2013-2022. It provides up-to-date, objective data demonstrating a 36.8% relative increase in obesity prevalence and a decline of 51.4% in normal BMI over the decade, with only 44.0% of pregnancies in the normal range by 2022.

Findings align with international trends,[18] and highlight the urgent need for strategies to address rising obesity during pregnancy. Obesity increases the risk of maternal and neonatal complications, earlier onset of metabolic and cardiovascular disease,[5] and gestational diabetes, [19] with significant healthcare costs.[2021] Intergenerational effects may also raise offspring's risk of obesity and cardiovascular disease,[22] with wider implications for population health. Class III obesity rose by 47.1%, a clinically and economically significant trend, given its association with the highest risk of pregnancy complications, interventional deliveries, and other peripartum interventions, contributing to increased maternal and neonatal morbidity and costs.[23,3]

Obesity prevalence rose across all age groups, with most rapid increases among younger (aged 20–29) and nulliparous women. By 2022, younger age (<25), previously protective, was no longer associated with lower obesity risk—indicating that obesity now affects the full reproductive-age population. This shift has significant implications for health service demand, particularly as many affected women are early in their reproductive life.

Multiparity remained a persistent risk factor for obesity, underscoring the need for postpartum interventions to support healthy weight maintenance and inform health promotion among

women of reproductive age. Obesity was strongly socially patterned, with lower skill levels consistently linked to higher BMI. Although obesity increased across all groups, steepest rises occurred among lower and middle skill levels, widening health disparities. These findings reflect patterns seen in other high-income countries and reinforce the need for equitable, targeted policy responses.[24] Ethnic disparities were evident: Afro-Caribbean women had significantly higher BMI than White Europeans, while Asian women had lower BMI. Asian ethnicity was protective against obesity in 2022 but not in 2013, suggesting a changing risk profile. These trends may reflect the influence of cultural, dietary, structural, and healthcare access factors and highlight the need for continued monitoring and culturally tailored public health strategies.

Unplanned pregnancy was associated with higher BMI, suggesting fewer opportunities to optimise health pre-conception, underscoring the need for health promotion regardless of pregnancy intention. BMI was linked with smoking and psychological issues, consistent with previous research,[25,26] and reinforcing the value of integrated mental and behavioural health support in reproductive care. Psychological problems became a significant obesity risk only in 2022, indicating a rising trend.

Among low-risk women (nulliparous, White European, aged 25–29, high SES, planned pregnancy, non-smoker, no mental health issues), average BMI rose from 24.73 in 2013 to 26.44 in 2022, shifting from normal to pre-obese. If current trends continue, projected BMI for 2025 is 27.29 kg/m², suggesting that even low-risk women are increasingly entering pregnancy overweight—reflecting a broader cultural shift in weight norms.

Findings highlight the urgent need for population-level interventions targeting women of reproductive age—particularly before pregnancy—to support healthy weight and improve long-term maternal and child health. Encouragingly, past initiatives have shown positive outcomes, reinforcing the effectiveness of proactive, preconception-focused strategies.[27] A combined approach involving universal and tailored interventions, including preconception and inter-pregnancy care, can leverage pregnancy as a critical window to promote lifelong healthy behaviours.

Strengths and limitations

Study strengths include the large sample size and routinely collected, objectively measured data with minimal missingness. Standardized height and weight measurements by trained midwives minimised errors, enhancing reliability. Limitations include the use of hospital-based rather than population-based data, which may affect national representativeness—though previous analyses support the hospital's representativeness.[11] Excluding participants with missing data—likely late antenatal attendees or transfers—and those not using hospital antenatal services may reduce internal validity. Small numbers in some ethnic minority groups may limit generalisability. Using ISCO skills level as the sole proxy for socioeconomic status, without markers like education or deprivation, may limit identification of high-risk groups. Despite these limitations, this study provides robust, up-to-date estimates and a valuable basis European comparisons and policy development.

Conclusion

This study presents updated estimates of obesity during pregnancy in Ireland, using a longitudinal approach to allow international comparisons. Obesity rose significantly from 2013 to 2022, with widening socioeconomic disparities. Increasing prevalence across all age groups is likely to increase maternal and infant morbidity, strain healthcare services and impact long-

term population health. The unequal distribution of highlights the need for equitable, needs-based service planning.

These findings call for urgent action from policymakers and healthcare providers. Innovative policies and both universal and targeted preventative and therapeutic interventions are needed to reduce obesity prevalence and guide resource allocation based on population needs. Future research should explore cross-European comparisons to deepen understanding of regional patterns and inform international best practices.

Acknowledgments

The authors are grateful to The Coombe Hospital for their collaboration without whom this analysis would not be possible. During preparation of this manuscript OpenAI's ChatGPT version 4o was used to improve readability and reduce word count in certain paragraphs where necessary. The authors reviewed all content and take full scientific responsibility for its accuracy and integrity.

Data Availability

Due to patient confidentiality and data use agreements, individual-level data cannot be shared publicly.

Funding and Assistance

Statistical guidance on the linear mixed effects model was provided by the Centre for Support and Training in Analysis and Research (CSTAR) at University College Dublin.

Conflicts of Interest Disclosure

The authors have no conflicts of interest to declare.

Author Contributions

All authors contributed to the conception and design of the study. AOH served as data owner and facilitated access to conduct the analysis. MG processed the raw electronic health records into a clean dataset, which EC subsequently analysed. ED led the data analysis and interpretation and drafted the manuscript. PN developed the figures. EC, MG, PN and MB inputted into the data interpretation. AOH, PK, CB, and MT provided oversight of the study's conception and design, analysis, and interpretation. All authors reviewed and provided edits in response to iterative manuscript drafts, and approved the final manuscript.

Supplementary Material

Appendix 1:

World Health Organization Classification of Body Mass Index	
<18.5 kg/m^2	Underweight
18.5-24.9 kg/m^2	Normal weight
25.0-29.9 kg/m^2	Pre-obesity
>30.0 kg/m^2	Obese
30.0-34.9 kg/m^2	Obesity Class I
35.0-39.9 kg/m^2	Obesity Class II
≥40.0 kg/m^2	Obesity Class III

Source: World Health Organization, 2000; kg/m^2: kilograms per metre squared

Supplementary Table 1: Variables included in the analysis

Variable name	Role of variable	Type of variable	
Year	Exposure	Numeric	As entered
Age	Exposure	Numeric Categorical	As entered <20 years, 20-24 years, 25-29 years, 30-34 years, 30-34 years 35-39 years, >/= 40 years
Singleton/multiple	Exposure	Categorical	Singleton/multiple
Ethnicity	Exposure	Categorical	Asian, Afro-Caribbean, White European, Middle Eastern, Other
Employment status	Exposure	Categorical	ISCO level 0-4
Socioeconomic status	Exposure	Categorical	ISCO level 0 = lower, ISCO level 1-3 = middle, ISCO level 4 = higher
Parity Pregnancy intention	Exposure Exposure	Categorical Categorical	Nulliparous/multiparous Planned/unplanned
Assisted reproduction	Exposure	Categorical	Yes/No
Psychological problems	Exposure	Categorical	Yes/No
Cigarette use	Exposure	Categorical:	Current smoker/Ex-smoker/Non-smoker
BMI in the first trimester/at first antenatal visit	Outcome	Numeric Categorical	As entered Underweight, normal, pre-obesity, Obese Class I,II,III Obese vs Non-obese

Supplementary Table 2: Comparison of included and excluded observations

		Included observations N=74233	Excluded observations N=991
Age (years)	**mean**	32.3	32.9
	SD	5.4	5.8
	Age <20 n	1149	17
	Age <20 %	1.5%	1.7%
	Age 20-24 n	6022	79
	Age 20-24 %	8.1%	8.0%
	Age 25-29 n	13569	136
	Age 25-29 %	18.3%	13.7%
	Age 30-34 n	26508	347
	Age 30-34 %	35.7%	35.0%
	Age 35-39 n	21734	323
	Age 35-39 %	29.3%	32.6%
	Age 40+ n	5251	89
	Age 40+ %	7.1%	9.0%
Parity	**Nulliparous n**	29860	399
	Nulliparous %	40.2%	40.3%
	Multiparous n	44373	591
	Multiparous %	59.8%	59.7%
Ethnicity	**White European n**	65821	884
	White European %	88.7%	89.2%
	Asian n	4282	43
	Asian %	5.8%	4.3%
	Afro-Caribbean n	1666	25
	Afro-Caribbean %	2.2%	2.5%
	Middle Eastern n	399	14
	Middle Eastern %	0.5%	1.4%
	Other n	2065	25
	Other %	2.8%	2.5%

Supplementary Table 3: Participant characteristics by BMI among total sample and in 2013 and 2022

		Total N= 74233					**2013 N=7543**					**2022 N=6472**				
		Underweight / normal weight N= 38998 N (%)	**Pre-obesity N= 21741 N (%)**	**Obese Class I N= 8728 N (%)**	**Obese Class II N= 3303 N (%)**	**Obese Class III N= 1463 N (%)**	**Underweight / normal weight N= 4143 N (%)**	**Pre-obesity N= 2172 N (%)**	**Obese Class I N= 814 N (%)**	**Obese Class II N= 291 N (%)**	**Obese Class III N= 123 N (%)**	**Underweight / normal weight N=3035 N (%)**	**Pre-obesity N=1993 N (%)**	**Obese Class I N=930 N (%)**	**Obese Class II N=361 N (%)**	**Obese Class III N=153 N (%)**
Total	**n**	38998	21741	8728	3303	1463	4143	2172	814	291	123	3035	1993	930	361	153
	%	52.5%	29.3 %	11.8 %	4.4%	2.0%	54.9%	28.8 %	10.8 %	3.9%	1.%	46.9%	30.8%	14.4 %	5.6%	2.4%
Age (years)	**Age <20**	789	237	79	40	4	99	27	9	6	0	46	18	8	4	1
	Age <20	68.7%	20.6 %	6.9%	3.5%	0.3%	70.2%	19.1 %	6.4%	4.3%	0.0%	59.7%	23.4%	10.4 %	5.2%	1.3%
	Age 20-24	3367	1548	719	265	123	472	183	68	25	14	244	129	72	26	14
	Age 20-24	55.9%	25.7 %	11.9 %	4.4%	2.0%	61.9%	24.0 %	8.9%	3.3%	1.8%	50.3%	26.6%	14.8 %	5.4%	2.9%
	Age 25-29	7137	3756	1668	688	320	923	485	188	78	36	459	334	149	68	27
	Age 25-29	52.6%	27.7 %	12.3 %	5.1%	2.4%	54.0%	28.4 %	11.0 %	4.6%	2.1%	44.3%	32.2%	14.4 %	6.6%	2.6%
	Age 30-34	14148	7722	3010	1122	506	1520	777	281	98	42	1079	666	327	109	46
	Age 30-34	53.4%	29.1 %	11.4 %	4.2%	1.9%	55.9%	28.6 %	10.3 %	3.6%	1.5%	48.5%	29.9%	14.7 %	4.9%	2.1%
	Age 35-39	11132	6746	2560	887	409	934	565	211	65	23	966	648	291	118	54
	Age 35-39	51.2%	31.0 %	11.8 %	4.1%	1.9%	51.9%	31.4 %	11.7 %	3.6%	1.3%	46.5%	31.2%	14.0 %	5.7%	2.6%
	Age 40+	2425	1732	692	301	101	195	135	57	19	8	241	198	83	36	11

	Age 40+	46.2%	33.0 %	13.2 %	5.7%	1.9%	47.1%	32.6 %	13.8 %	4.6%	1.9%	42.4%	34.8%	14.6 %	6.3%	1.9%
Parity	**Nulliparous**	17147	8140	2951	1109	513	1776	769	244	84	38	1348	831	334	123	58
	Nulliparous	57.4%	27.3 %	9.9%	3.7%	1.7%	61.0%	26.4 %	8.4%	2.9%	1.3%	50.0%	30.8%	12.4 %	4.6%	2.2%
	Multiparous	21851	13601	5777	2194	950	2367	1403	570	207	85	1687	1162	596	238	95
	Multiparous	49.2%	30.7 %	13.0 %	4.9%	2.1%	51.1%	30.3 %	12.3 %	4.5%	1.8%	44.7%	30.8%	15.8 %	6.3%	2.5%
Plurality	**Singleton**	17147	8140	2951	1109	513	3997	2078	763	273	115	2939	1918	889	345	146
	Singleton	57.4%	27.3 %	9.9%	3.7%	1.7%	55.3%	28.8 %	10.6 %	3.8%	1.6%	47.1%	30.8%	14.3 %	5.5%	2.3%
	Multiple	21851	13601	5777	2194	950	146	94	51	18	8	96	75	41	16	7
	Multiple	49.2%	30.7 %	13.0 %	4.9%	2.1%	46.1%	29.7 %	16.1 %	5.7%	2.5%	40.9%	31.9%	17.4 %	6.8%	3.0%
ISCO Skill level	**Skill level 0**	8108	4815	2299	1002	453	1155	617	290	104	54	452	359	193	75	40
	Skill level 0	48.6%	28.9 %	13.8 %	6.0%	2.7%	52.0%	27.8 %	13.1 %	4.7%	2.4%	40.4%	32.1%	17.2 %	6.7%	3.6%
	Skill level 1	527	276	124	41	21	53	33	12	5	0	44	13	10	7	4
	Skill level 1	53.3%	27.9 %	12.5 %	4.1%	2.1%	51.5%	32.0 %	11.7 %	4.9%	0.0%	56.4%	16.7%	12.8 %	9.0%	5.1%
	Skill level 2	6343	3736	1602	602	272	718	394	142	53	21	409	289	155	51	25
	Skill level 2	50.5%	29.8 %	12.8 %	4.8%	2.2%	54.1%	29.7 %	10.7 %	4.0%	1.6%	44.0%	31.1%	16.7 %	5.5%	2.7%
	Skill level 3	3288	1939	734	283	122	367	191	62	32	11	271	184	93	33	18
	Skill level 3	51.6%	30.5 %	11.5 %	4.4%	1.9%	55.4%	28.8 %	9.4%	4.8%	1.7%	45.2%	30.7%	15.5 %	5.5%	3.0%
	Skill level 4	20732	10975	3969	1375	595	1850	937	308	97	37	1859	1148	479	195	66
	Skill level 4	55.1%	29.2 %	10.5 %	3.7%	1.6%	57.3%	29.0 %	9.5%	3.0%	1.1%	49.6%	30.6%	12.8 %	5.2%	1.8%
Ethnicity	**Asian**	34887	19076	7641	2903	1314	3706	1912	683	253	110	2599	1663	825	312	135
	Asian	53.0%	29.0 %	11.6 %	4.4%	2.0%	55.6%	28.7 %	10.2 %	3.8%	1.7%	47.0%	30.1%	14.9 %	5.6%	2.4%
	Afro-Caribbean	2272	1404	453	118	35	237	119	42	8	5	275	204	54	20	5
	Afro-Caribbean	53.1%	32.8 %	10.6 %	2.8%	0.8%	57.7%	29.0 %	10.2 %	1.9%	1.2%	49.3%	36.6%	9.7%	3.6%	0.9%
	White European	526	537	348	173	82	65	64	58	20	6	46	48	26	17	10
	White European	31.6%	32.2 %	20.9 %	10.4 %	4.9%	30.5%	30.0 %	27.2 %	9.4%	2.8%	31.3%	32.7%	17.7 %	11.6 %	6.8%
	Middle Eastern	179	139	55	16	10	16	9	7	2	2	19	8	5	1	1
	Middle Eastern	44.9%	34.8 %	13.8 %	4.0%	2.5%	44.4%	25.0 %	19.4 %	5.6%	5.6%	55.9%	23.5%	14.7 %	2.9%	2.9%
	Other	1134	585	231	93	22	119	68	24	8	0	96	70	20	1	2
	Other	54.9%	28.3 %	11.2 %	4.5%	1.1%	54.3%	31.1 %	11.0 %	3.7%	0.0%	48.2%	35.2%	10.1 %	5.5%	1.0%
Pregnancy intention	**Planned**	26963	14909	5647	2063	852	2799	1421	523	172	70	2078	1371	611	237	87
	Planned	53.5%	29.6 %	11.2 %	4.1%	1.7%	56.1%	28.5 %	10.5 %	3.5%	1.4%	47.4%	31.3%	13.9 %	5.4%	2.0%
	Unplanned	9643	5466	2541	1048	530	1150	655	252	98	49	685	459	242	95	47

	Unplanned	50.2%	28.4%	13.2%	5.5%	2.8%	52.2%	29.7%	11.4%	4.4%	2.2%	44.8%	30.0%	15.8%	6.2%	3.1%
	Infertility treatment	2392	1366	540	192	81	194	96	39	21	4	272	163	77	29	19
	Infertility treatment	52.3%	29.9%	11.8%	4.2%	1.8%	54.8%	27.1%	11.0%	5.9%	1.1%	48.6%	29.1%	13.8%	5.2%	3.4%

Supplementary Table 4: Participant characteristics, socioeconomic and lifestyle risk factors associated with obesity during pregnancy in 2013 and 2022

	2013 n=7543			**2022 n=6472**		
Age (years)	**N (%) obesity**	**aOR**	**p-value**	**N (%) obesity**	**aOR**	**p-value**
<20	15 10.6%	**0.69 (0.54-0.88)**	**0.003**	13 16.9%	0.91 (0.70-1.19)	0.503
20-24	107 14.0%	**0.54 (0.30-0.95)**	**0.032**	112 23.1%	0.58 (0.31-1.08)	0.087
25-29	302 17.7%	**1**		244 23.7%	1	
30-34	421 15.5%	0.92 (0.78-1.09)	0.319	482 21.7%	1.01 (0.85-1.22)	0.866
35-39	299 16.6%	0.94 (0.78-1.13)	0.487	463 22.4%	1.01 (0.83-1.22)	0.935
40+	84 20.3%	1.10 (0.83-1.47)	0.511	130 23.1%	0.99 (0.77-1.29)	0.992
Parity						
Nulliparous	366 12.6%	1		515 19.2%	1	
Multiparous	862 18.6%	**1.42 (1.23-1.64)**	**<0.001**	929 24.8%	**1.28 (1.12-1.46)**	**<0.001**
Skill level						
0	448 20.2%	**1.45 (1.22-1.72)**	**<0.001**	308 27.7%	**1.47 (1.23-1.75)**	**<0.001**
1	17 16.5%	1.19 (0.69-2.03)	0.535	21 26.9%	1.41 (0.84-2.37)	0.194
2	216 16.3%	**1.22 (1.01-1.47)**	**0.036**	231 25.0%	**1.29 (1.08-1.54)**	**0.005**
3	105 15.8%	1.18 (0.94-1.49)	0.162	144 24.1%	**1.31 (1.07-1.61)**	**0.010**
4	442	1		740	1	

	13.7%			19.8%		
Ethnicity						
White European	1046 15.7%	1		1272 23.1%	1	
Asian	55 13.4%	0.78 (0.58-1.05)	0.100	79 14.2%	**0.62 (0.48-0.80)**	**<0.001**
Afro-Caribbean	84 39.4%	**2.80 (2.07-3.79)**	**<0.001**	53 36.8%	**2.04 (1.43-2.80)**	**<0.001**
Middle Eastern	11 30.6%	**2.09 (1.01-4.32)**	**0.047**	7 20.6%	0.84 (0.36-1.95)	0.683
Other	32 14.9%	0.87 (0.59-1.28)	0.470	33 16.6%	**0.68 (0.46-1.00)**	**0.050**
Pregnancy intention						
Planned	765 15.3%	1		935 21.4%	1	
Unplanned	399 18.1%	1.14 (0.99-1.32)	0.071	384 25.2%	1.10 (0.95-1.27)	0.214
Infertility treatment	64 18.1%	1.18 (0.86-1.61)	0.305	125 22.4%	1.21 (0.96-1.52)	0.103
Psychological problems						
No	978 15.8%	1		975 20.6%	1	
Yes	250 18.3%	1.12 (0.95-1.31)	0.182	469 27.0%	**1.27 (1.11-1.45)**	**<0.001**
Cigarette use						
Never	659 16.6%	1		817 19.7%	1	
Current	406 15.3%	0.99 (0.81-1.22)	0.939	501 26.8%	1.24 (0.98-1.57)	0.078
Past	163 17.5%	0.95 (0.83-1.10)	0.499	126 28.0%	**1.39 (1.22-1.59)**	**<0.001**

Model adjusted for maternal age, parity, ethnicity, skills level, psychological problems, and smoking status; aOR = adjusted odds ratio; kg/m^2: kilograms per metre squared.

Supplementary Table 5: Obesity prevalence stratified by age and parity: total sample 2013-2022

	Nulliparous		**Multiparous**	
	Non-obese N=25287	Obese N=4573	Non-obese N=35452	Obese N=8921
Age <20	936 90.2%	102 9.8%	90 81.1%	21 18.9%
Age 20-24	3169 83.3%	637 16.7%	1746 78.8%	470 21.2%
Age 25-29	5417 83.8%	1051 16.2%	5476 77.1%	1625 22.9%
Age 30-34	9467 85.7%	1581 14.3%	12403 80.2%	3057 19.8%
Age 35-39	5086 84.4%	938 15.6%	12792 81.4%	2918 18.6%
Age 40+	1212 82.1%	264 17.9%	2945 78.0%	830 22.0%

Supplementary Table 6: Obesity prevalence stratified by age and parity: 2013 versus 2022 only

Parity	**Nulliparous**				**Multiparous**			
Year	**2013**		**2022**		**2013**		**2022**	
Obesity Status	Non-obese	Obese	Non-obese	Obese	Non-obese	Obese	Non-obese	Obese
	N=2545	N=366	N=2179	N=515	N=3770	N=862	N=2849	N=929
Age <20	117	13	57	11	9	2	7	2
	90.0%	10.0%	83.8%	16.2%	81.8%	18.2%	77.8%	22.2%
Age 20-24	397	53	254	71	258	54	119	41
	88.2%	11.8%	78.2%	21.8%	82.7%	17.3%	74.4%	25.6%
Age 25-29	654	105	427	98	754	197	366	146
	86.2%	13.8%	81.3%	18.7%	79.3%	20.7%	71.5%	28.5%
Age 30-34	873	126	837	179	1424	295	908	303
	87.4%	12.6%	82.4%	17.6%	82.8%	17.2%	75.0%	25.0%
Age 35-39	399	50	458	121	1100	249	1156	342
	88.9%	11.1%	79.1%	20.9%	81.5%	18.5%	77.2%	22.8%
Age 40+	105	19	146	35	225	65	293	95
	84.7%	15.3%	80.7%	19.3%	77.6%	22.4%	75.5%	24.5%

Supplementary Table 7: Obesity prevalence by ethnicity stratified by parity

	Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
Ethnicity: Nulliparous	**Asian n**	13	11	8	7	24	10	28	14	19	28	162
	Asian %	10.5%	9.4%	6.8%	7.1%	16.6%	6.1%	15.5%	8.0%	11.4%	11.1%	10.5%
	Afro-Caribbean n	6	9	4	6	5	10	10	6	8	14	78
	Afro-Caribbean %	17.6%	22.0%	12.1%	15.0%	13.9%	23.3%	22.7%	14.3%	21.6%	28.6%	19.5%
	White European n	338	390	344	415	434	468	456	470	433	461	4209
	White European %	12.8%	13.8%	12.4%	14.2%	15.2%	15.8%	16.4%	19.2%	18.4%	20.3%	15.7%
	Middle Eastern n	2	1	1	0	3	1	0	3	1	0	12
	Middle Eastern %	15.4%	10.0%	7.1%	0.0%	20.0%	7.7%	0.0%	23.1%	5.3%	0.0%	9.3%
	Other n	7	6	4	6	12	17	12	17	19	12	7
	Other %	7.3%	9.8%	6.5%	5.9%	14.1%	17.9%	10.3%	14.7%	20.9%	10.8%	7.3%
Ethnicity: Multiparous	**Asian n**	42	30	40	34	35	46	60	46	60	51	444
	Asian %	14.6%	11.8%	15.1%	13.6%	13.7%	16.1%	21.2%	18.0%	20.0%	16.7%	16.2%
	Afro-Caribbean n	78	82	58	50	55	45	45	32	41	39	525
	Afro-Caribbean %	43.6%	46.3%	38.7%	36.0%	43.3%	40.2%	42.9%	36.8%	44.1%	39.8%	41.4%
	White European n	708	699	667	767	776	793	812	766	850	811	708
	White European %	17.6%	16.7%	15.8%	18.2%	19.3%	20.0%	21.2%	21.3%	23.3%	24.9%	17.6%
	Middle Eastern n	9	8	5	10	11	4	8	5	2	7	69
	Middle Eastern %	39.1%	34.8%	20.8%	27.0%	35.5%	19.0%	23.5%	21.7%	7.4%	25.9%	25.6%
	Other n	25	16	18	23	29	23	33	27	19	21	234
	Other %	20.3%	16.8%	16.1%	18.4%	26.6%	20.4%	24.1%	20.9%	19.2%	23.9%	20.7%

References

[1] The Lancet Gastroenterology Hepatology. Obesity: another ongoing pandemic. Lancet Gastroenterol Hepatol. 2021;6(6):411. doi:10.1016/S2468-1253(21)00143-6

[2] Poston L, Caleyachetty R, Cnattingius S, Corvalán C, Uauy R, Herring S, et al. Preconceptional and maternal obesity: epidemiology and health consequences. Lancet Diabetes Endocrinol. 2016;4(12):1025–36.

[3] Denison FC, Norwood P, Bhattacharya S, Duffy A, Mahmood T, Morris C, et al. Association between maternal body mass index during pregnancy, short-term morbidity, and increased health service costs: a population-based study. BJOG. 2014;121(1):72-81.

[4] Langley-Evans SC, Pearce J, Ellis S. Overweight, obesity and excessive weight gain in pregnancy as risk factors for adverse pregnancy outcomes: A narrative review. J Hum Nutr Diet. 2022;35(2):250-264. doi:10.1111/jhn.12999

[5] Marchi J, Berg M, Dencker A, Olander EK, Begley C. Risks associated with obesity in pregnancy, for the mother and baby: a systematic review of reviews. *Obes Rev*. 2015;16(8):621-638. doi:10.1111/obr.12288

[6] Ijas H, Morin-Papunen L, Keranen AK, Bloigu R, Ruokonen A, Puukka K, et al. Pre-pregnancy overweight overtakes gestational diabetes as a risk factor for subsequent metabolic syndrome. Eur J Endocrinol. 2013; 169(5):605–611. 10.1530/EJE-13-0412

[7] Leddy MA, Power ML, Schulkin J. The impact of maternal obesity on maternal and fetal health. Rev Obstet Gynecol. 2008;1(4):170-178.

[8] Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. Lancet. 2009;373(9677):1773–1779.

9 Catalano PM. Management of obesity in pregnancy. Obstet Gynecol. 2007;109(2 Pt 1):419–433.

[10] Şanlı E, Kabaran S. Maternal Obesity, Maternal Overnutrition and Fetal Programming: Effects of Epigenetic Mechanisms on the Development of Metabolic Disorders. Curr Genomics. 2019 Sep;20(6):419-427.

[11] Scheidl TB, Brightwell AL, Easson SH, Thompson JA. Maternal obesity and programming of metabolic syndrome in the offspring: searching for mechanisms in the adipocyte progenitor pool. *BMC Med*. 2023;21(1):50. Published 2023 Feb 13. doi:10.1186/s12916-023-02730-z

[12] Rockliffe L, Peters S, Heazell AEP, Smith DM. Understanding pregnancy as a teachable moment for behaviour change: a comparison of the COM-B and teachable moments models. Health Psychol Behav Med. 2021 Dec 30;10(1):41-59. doi: 10.1080/21642850.2021.2014851. PMID: 34993005; PMCID: PMC8725882.

[13] Reynolds CME, Egan B, McMahon L, O'Malley EG, Sheehan SR, Turner MJ. Maternal obesity trends in a large Irish university hospital. Eur J Obstet Gynecol Reprod Biol. 2019 Jul;238:95-99. doi: 10.1016/j.ejogrb.2019.05.003.

[14] Reynolds CME, O'Malley EG, Egan B, Sheehan SR, Turner MJ. Maternal Weight Trajectories in Successive Pregnancies and Their Association With Gestational Diabetes Mellitus. Diabetes Care. 2020 Jan 16;43(3):e33–4.

[15] Healthcare Pricing Office. Perinatal Statistics Report 2021. Dublin: Health Service Executive; 2024.

[16] World Health Organization (WHO). Obesity: preventing and managing the global epidemic. Report of a WHO consultation. WHO technical report series 894. Geneva: World Health Organization; 2000.

[17] International Labour Organization. (2012). *International Standard Classification of Occupations: ISCO-08*. Geneva: International Labour Office.

[18] World Health Organization. WHO European Regional Obesity Report 2022. Copenhagen: WHO Regional Office for Europe; 2022.

[19] G.E. Avalos, L.A. Owens, F. Dunne, for the ATLANTIC DIP Collaborators, Applying Current Screening Tools for Gestational Diabetes Mellitus to a European Population: Is It Time for Change?, Diabetes Care 36 (2013) 3040–3044. https://doi.org/10.2337/dc12-2669.

[20] Gillespie P, O'Neill C, Avalos G, O'Reilly M, Dunne F; ATLANTIC DIP Collaborators. The cost of universal screening for gestational diabetes mellitus in Ireland. Diabet Med. 2011 Aug;28(8):912-8.

doi: 10.1111/j.1464-5491.2011.03293.x. Erratum in: Diabet Med. 2016 May;33(5):701. doi: 10.1111/dme.13101. PMID: 21418093.

[21] P. Gillespie, J. Cullinan, C. O'Neill, F. Dunne, Modeling the Independent Effects of Gestational Diabetes Mellitus on Maternity Care and Costs, Diabetes Care 36 (2013) 1111–1116. https://doi.org/10.2337/dc12-0461.

[22] Fleming TP, Watkins AJ, Velazquez MA, Mathers JC, Prentice AM, Stephenson J, et al. Origins of lifetime health around the time of conception: causes and consequences. *Lancet*. 2018 May 5;391(10132):1842–52. doi:10.1016/S0140-6736(18)30312-X.

[23] Chu SY, Kim SY, Schmid CH, Dietz PM, Callaghan WM, Lau J, et al. Maternal obesity and risk of cesarean delivery: a meta-analysis. *Obes Rev*. 2007 Sep;8(5):385–94.

[24] Heslehurst N, Rankin J, Wilkinson JR, Summerbell CD. A nationally representative study of maternal obesity in England, UK: trends in incidence and inequalities in 619,323 births, 1989–2007. *Int J Obes (Lond)*. 2010 Feb;34(3):420–8. doi:10.1038/ijo.2009.235

[25] Tuthill EH, Turner MJ. Association of self-reported maternal depression and obesity at the first antenatal visit. *Ir J Med Sci*. 2021 Jun;190(2):555–561. doi:10.1007/s11845-021-02665-5

[26] Griffiths A, Shannon OM, Brown T, Davison M, Swann C, Jones A, et al. Associations between anxiety, depression, and weight status during and after pregnancy: A systematic review and meta-analysis. Obes Rev. 2024 Mar;25(3):e13668. doi: 10.1111/obr.13668.

[27] van Dammen L, Wekker V, de Rooij SR, Painter RC, Maas AHEM, Roseboom TJ, et al. A life course perspective on women's reproductive health and the need for a preconception care approach. *Am J Obstet Gynecol*. 2021 Jul;225(1):B2–B10. doi:10.1016/j.ajog.2020.11.1245

Appendix D. Features in the clean EHR dataset.

Variable	Value Count/Range
Ethnic Origin of Patient	{'CAUCASIAN': 24180, 'SOUTH EAST ASIAN': 1360, 'OTHER': 824, 'BLACK': 554, 'ASIAN': 489, 'MIDDLE EASTERN': 154}
Cardiac Problems	{'NO': 26027, 'YES': 1534}
Raised BP outside preg.	{'NO': 26959, 'IN PAST': 334, 'CURRENT': 268}
VV's/clotting probs.	{'NO': 25380, 'YES': 2181}
Urinary/kidney probs.	{'NO': 20874, 'YES': 6687}
Fits/epilepsy	{'NO': 23708, 'YES': 3853}
Jaundice/liver disease	{'TATTOO/PIERCING': 14834, 'NO': 12211, 'YES': 516}
Resp. system disease	{'NO': 22588, 'YES': 4973}
Metabolic Disorders	{'NO': 27537, 'YES': 24}
Digestive tract disease	{'NO': 23209, 'YES': 4352}
Serious infections	{'NO': 20883, 'YES': 6678}
Connective Tissue	{'NO': 27167, 'YES': 394}
Orthopaedic Probs	{'NO': 20212, 'YES': 7349}
Operations in past	{'NO': 18217, 'YES': 9344}
Gynae. probs./operations	{'NO': 16207, 'YES': 11354}
STD - ever	{'NO': 20295, 'YES': 7266}
Anaemia/blood disorder	{'NO': 20464, 'YES': 7097}
Anaemia Treatment	{'NONE': 21447, 'YES': 6114}
Blood transfusion - ever	{'NO': 26354, 'YES': 1207}
Blood/Blood products	{'YES': 27504, 'NO': 57}
Cigarette Alternatives	{'NO': 26611, 'YES': 950}
Folic acid	{'POSTCONCEPTION': 12503, 'BOTH': 10791, 'PRECONCEPTION': 3439, 'NO': 828}
Folic acid length	{'0-3 months': 14520, '3-6 months': 6852, '6-12 months': 3165, 'longer than 12 months': 3024}
Psychological prob	{'NO': 20623, 'YES': 6938}
Illicit drugs ever	{'NO': 25487, 'YES': 2074}
Illicit Drugs - Partner	{'NO': 26458, 'YES': 1103}
Allergies	{'NO': 21558, 'YES': 6003}
FH Hypertension	{'NO': 14316, 'YES': 13245}
FH of Cancer	{'NO': 19806, 'YES': 7755}
FH TB	{'NO': 27241, 'YES': 320}
FH Blood disorder	{'NO': 24516, 'YES': 3045}
FH - Heart Problems	{'NO': 16342, 'YES': 11219}
FH Thyroidism	{'NO': 21533, 'YES': 6028}
FH Diabetes	{'NO': 21154, 'YES': 6407}
FH Multiple preg.	{'NO': 20717, 'YES': 6844}

FH Mental Illness	{'NO': 23890, 'YES': 3671}
FH-Congenital Abnormality	{'NO': 21730, 'YES': 5831}
Pregnant Before	{'YES': 18867, 'NO': 8694}
Misc/TOP/Ect Before	{'NO': 18333, 'YES': 9228}
Pregnant no of Times	{'min': 0, 'max': 16, 'range': 16}
Proteinuria at booking	{'NO': 27433, 'YES': 128}
Systolic BP at booking	{'min': 58, 'max': 195, 'range': 137}
Diastolic BP at booking	{'min': 40.0, 'max': 120.0, 'range': 80.0}
Planned pregnancy	{'PLANNED PREGNANCY': 19027, 'UNPLANNED PREGNANCY': 6719, 'INFERTILITY TREATMENT': 1815}
Infertility Treatment	{'NO': 25791, 'YES': 1770}
Pregnant on contraception	{'NO': 26344, 'YES': 1217}
Cycle Regular	{'YES': 21789, 'NO': 5772}
Tests/investigations	{'NO': 15312, 'YES - OTHER': 9593, 'EPAU': 2656}
Weight at Booking	{'min': 32.4, 'max': 175.2, 'range': 142.79999999999998}
Gravida	{'min': 0.0, 'max': 16.0, 'range': 16.0}
Parity (not inc.multiple)	{'min': 0, 'max': 11, 'range': 11}
Livebirths - No	{'min': 0.0, 'max': 11.0, 'range': 11.0}
Stillbirths - No	{'min': 0, 'max': 2, 'range': 2}
Neonatal deaths - No	{'min': 0, 'max': 2, 'range': 2}
Deaths after 28 days - No	{'min': 0, 'max': 2, 'range': 2}
Caesarean Sections - No	{'min': 0, 'max': 6, 'range': 6}
Terminations - No	{'min': 0, 'max': 4, 'range': 4}
Miscarriages - No	{'min': 0, 'max': 11, 'range': 11}
Ectopics - No	{'min': 0, 'max': 3, 'range': 3}
Previous moles - No.	{'min': 0, 'max': 1, 'range': 1}
Drug Abuse	{'NO': 27465, 'YES': 96}
Fertility treatment	{'NO': 25815, 'YES': 1746}
Hb Electrophoresis	{'NOT PERFORMED': 25535, 'NORMAL': 1578, 'ABNORMAL': 448}
Serology Result	{'NEGATIVE': 27319, 'NOT PERFORMED': 152, 'POSITIVE': 90}
H.I.V. Status	{'NEGATIVE': 27463, 'UNKNOWN': 54, 'POSITIVE': 44}
Last Haemoglobin	{'min': 6.1, 'max': 18.7, 'range': 12.6}
Bleeding during Pregnancy	{'NO': 24292, 'YES': 3269}
Hypertension during preg	{'NO': 26176, 'YES': 1385}
Hypertension Treatment	{'NONE': 26603, 'YES': 958}
Fetal problems	{'NO': 22080, 'YES OTHER': 4651, 'MACROSOMIA': 830}
Scan Abnormalities	{'NO': 26936, 'YES OTHER': 414, 'LGA': 127, 'SGA': 84}
Age at booking	{'min': 15, 'max': 53, 'range': 38}

GDM	{'min': 0, 'max': 1, 'range': 1}
Height of Mother (m)	{'min': 1.26, 'max': 1.95, 'range': 0.69}
BMI	{'min': 14.5, 'max': 61.1, 'range': 46.6}
Smoke Now	{'NO': 16465, 'STOPPED': 8636, 'YES': 2460}
Ever Smoked	{'NO': 16459, 'YES': 11102}
Skill Level	{'min': 0, 'max': 4, 'range': 4}
Hx_GDM	{'min': 0, 'max': 1, 'range': 1}
Other Endocrine probs	{'min': 0, 'max': 1, 'range': 1}

Appendix E. Letter to the Editor

Lack of Data Sharing Despite Data Availability Statements in Studies Using Machine Learning Models for Prediction of Gestational Diabetes Mellitus

Mark Germaine,[1,2,3] Graham Healy,[1] and Brendan Egan[2,4]

Diabetes Care 2024;47:e78–e79 | https://doi.org/10.2337/dc24-1483

Recent advancements in artificial intelligence and machine learning (ML) research allow for the mining of electronic health records (EHRs) for predicting health outcomes. One application is that ML models can be developed to predict likelihood of gestational diabetes mellitus (GDM) by using data taken from EHRs obtained early in pregnancy. We have completed preliminary work developing such models using EHR data collected in the first trimester (1). An important feature of ML modeling is the use of an independent data set for external validation to ensure the model's generalizability across different data sets. However, obtaining such a data set has proven challenging and illustrates broader issues regarding data sharing and the implementation of open science principles.

In an attempt to acquire data for external validation, we contacted authors from 22 published articles describing studies that aimed to predict GDM using EHRs (Table 1), and we sought access to a sample subset of their data sets for the purpose of external validation of our model. These studies were identified from a systematic literature search, which we performed up to March 2024, for ML models developed to predict GDM using data from EHRs.

We contacted the respective authors on three separate occasions between 18 April and 17 June 2024. All listed e-mail addresses were contacted simultaneously. The first e-mail detailed the purpose of our request, the importance of external validation for improving model reliability, and assurances regarding data confidentiality. Follow-up emails served as concise reminders of the significance of their contribution to advancing research in ML models for GDM prediction.

Of the 22 articles, 14 had data availability statements indicating that data were available upon request, 3 stated data were not available, and 5 did not have any data availability statement. Despite our efforts, the response rate was unequivocally low. Only one author group (corresponding to two articles) responded positively, expressing a willingness to validate our model independently using their data set, but were unable to share their data directly. Out of the remaining 20 articles, one e-mail address was no longer valid, and another e-mail address elicited an automatic reply but without further response with follow-up emails. Authors from the remaining 18 articles did not provide any response (Table 1).

While it is recognized that the availability of research data declines rapidly with article age (2), the median publication date of the identified articles was 2021. Only 4 out of 22 articles were published prior to 2020. We expected that this recent publication timeframe would result in greater likelihood of data availability.

Difficulties in acquiring data despite the presence of data availability statements is not uncommon. Nonresponses and refusals when attempting to conduct an individual patient data meta-analysis have been previously reported (3). Analysis of data sharing practices in *The BMJ* found that despite a strong data sharing policy, actual sharing rates were low (4). Only 4.5% of the articles shared their data sets, although a higher rate of 24% was observed for articles describing clinical trials. Ambiguous policy wording and a lack of incentives for researchers were identified as being among several barriers to data sharing (4).

Our experience, along with findings above and from others (5), suggests that these data availability statements often do not translate into actual data sharing and highlights two major issues. First, it underscores poor practices around data availability statements. Despite these statements, the sharing of data described in published articles remains inconsistent and unreliable. Second, the lack of data sharing poses a substantial barrier to the external validation of predictive models in ML. Without access to external data sets, it is challenging to ensure the generalizability and robustness of ML models, which will ultimately affect the utility of ML for advancing digital health and artificial intelligence–driven health care.

[1]School of Computing, Dublin City University, Dublin, Ireland
[2]School of Health and Human Performance, Dublin City University, Dublin, Ireland
[3]SFI Centre for Research Training in Machine Learning, Dublin City University, Dublin, Ireland
[4]Florida Institute for Human and Machine Cognition, Pensacola, FL

Corresponding author: Brendan Egan, brendan.egan@dcu.ie

Received 18 July 2024 and accepted 25 July 2024

Table 1—Details of studies included in data access inquiries

Article no.	Year	Sample size range	Country	Data availability statement	Author response	Outcome of request
1	2010	<1,000	The Netherlands	N	N	E-mail address no longer valid; NDS
2	2013	1,000–5,000	Vietnam	N	N	NDS
3	2017	1,000–5,000	China	N	N	NDS
4	2017	<1,000	Australia	N	N	NDS
5	2019	>500,000	U.S.	Y	N	Contains "Accessible Data" link, but no data available in repository; NDS
6	2020	>500,000	Israel	Y	N	NDS
7	2020	1,000–5,000	China	Y	N	NDS
8	2020	5,001–50,000	China	Y	N	NDS
9	2021	5,001–50,000	China	Y	N	NDS
10	2021	5,001–50,000	China	Y	N	NDS
11	2021	5,001–50,000	China	Y	N	Automatic reply acknowledging emails, but no further response; NDS
12	2021	1,000–5,000	China	Y	N	NDS
13	2022	1,000–5,000	China	Y	N	NDS
14	2023	50,001–500,000	Japan	Y	N	NDS
15	2023	5,001–50,000	South Korea	DNA	N	NDS
16	2023	<1,000	China	DNA	N	NDS
17	2023	5,001–50,000	Australia	Y	Y	Cannot share data but willing to validate model in own data set; NDS
18	2023	5,001–50,000	Australia	DNA	Y	Cannot share data but willing to validate model in own data set; NDS
19	2023	1,000–5,000	Chile	Y	N	NDS
20	2023	<1,000	China	Y	N	NDS
21	2024	5,001–50,000	China	Y	N	NDS
22	2024	5,001–50,000	China	N	N	NDS

Shown are details of 22 studies, identified by a systematic literature search, which aimed to develop ML models to predict GDM from data in EHRs and to whose authors we sent data access inquiries. DNA, data not available; N, no; NDS, no data shared; Y, yes.

These challenges highlight the need for clearer policies and better incentives to promote data sharing and support the open science movement. This gap between the ideal of open science and the reality of data accessibility emphasizes the need for more robust mechanisms to ensure data availability and to support the reproducibility of scientific findings. To advance the field, it is important to establish more dependable mechanisms for data sharing. This includes reinforcing the commitment of authors and journals to uphold data availability statements in practice as well as developing clearer policies and incentives to promote data sharing. The move toward open science has encouraged the inclusion of data availability statements to promote transparency and reproducibility in research, but the cultural shift toward open data is still evolving, and there remains significant room for improvement.

Funding. This work has emanated from research supported in part by a grant from Science Foundation Ireland under grant number 18/CRT/6183.

Duality of Interest. No potential conflicts of interest relevant to this article were reported.

Author Contributions. M.G. contacted the authors, compiled the data, and wrote the manuscript. G.H. and B.E. contributed to the discussion and reviewed and edited the manuscript. B.E. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

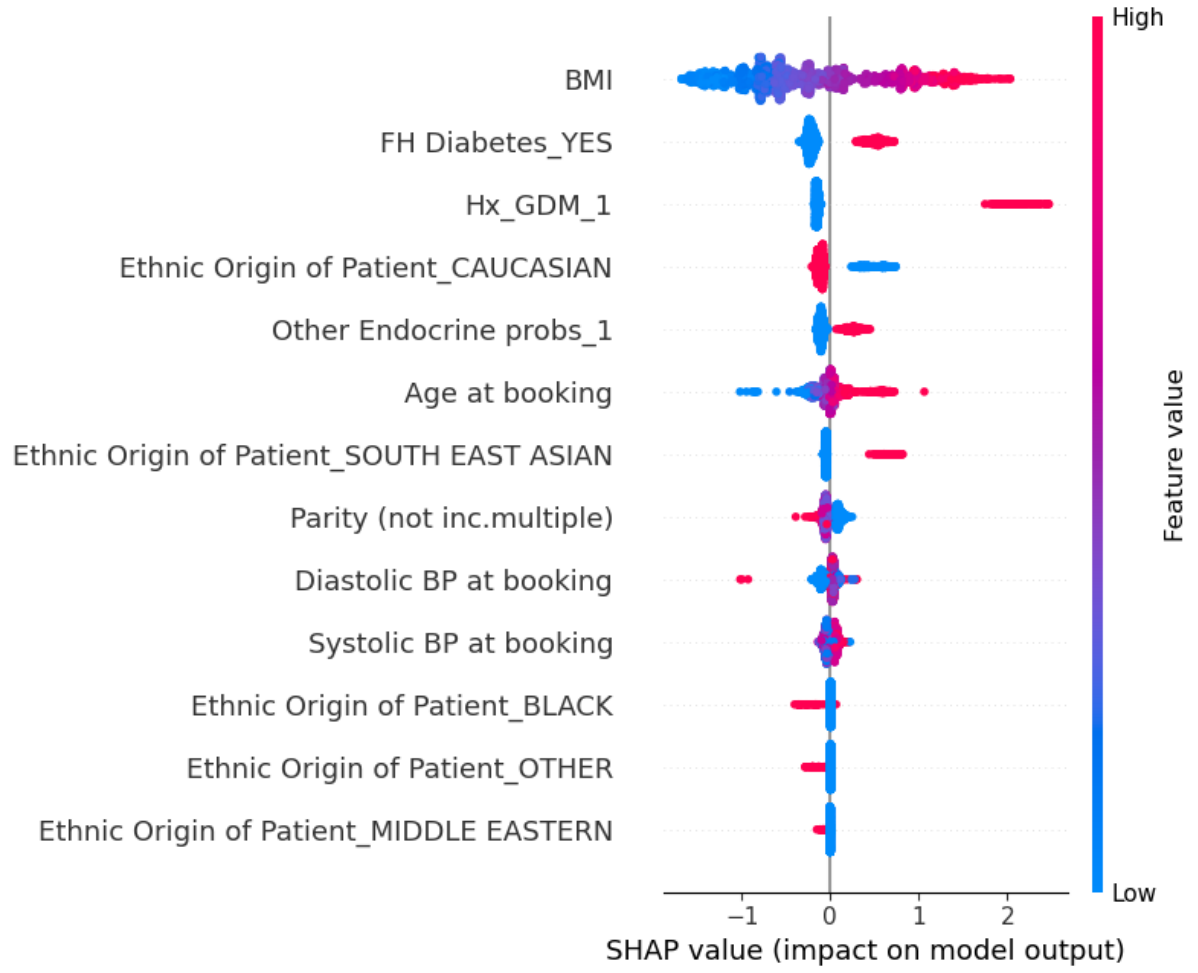
Handling Editors. The journal editors responsible for overseeing the review of the manuscript were Steven E. Kahn and Matthew J. Crowley.

References

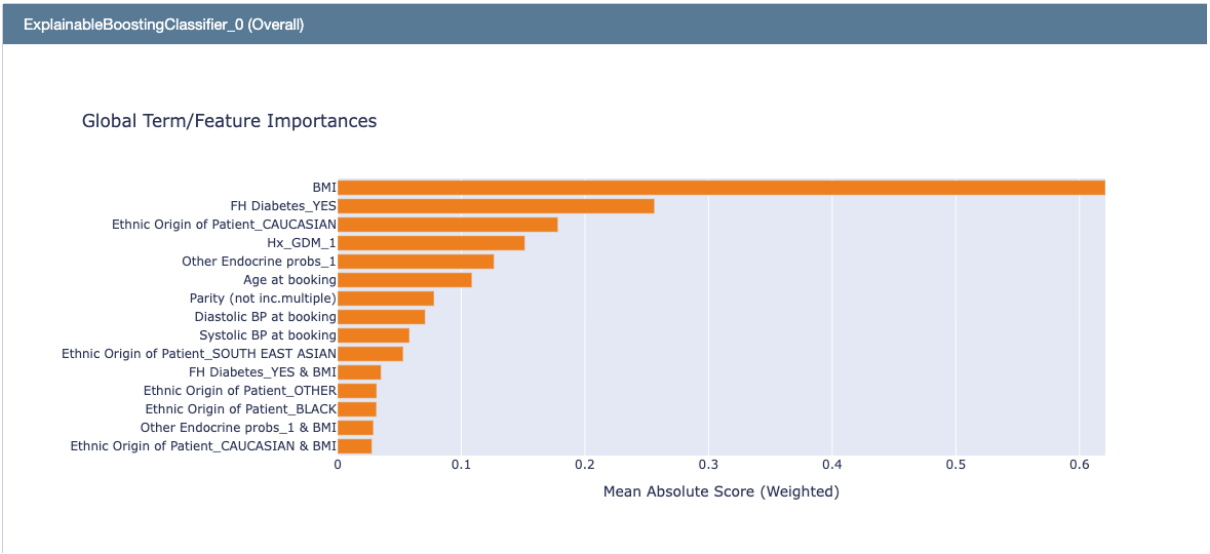
1. Germaine MA, O'Higgins AC, Healy G, Egan B. 1968-LB: Early prediction of gestational diabetes mellitus using electronic health records and machine learning. Diabetes 2024;73(Supplement_1):1968-LB
2. Vines TH, Albert AYK, Andrew RL, et al. The availability of research data declines rapidly with article age. Curr Biol 2014;24:94–97
3. Jaspers GJ, Degraeuwe PLJ. A failed attempt to conduct an individual patient data meta-analysis. Syst Rev 2014;3:97
4. Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open 2016;6:e011784
5. Obels P, Lakens D, Coles NA, Gottfried J, Green SA. Analysis of open data and computational reproducibility in registered reports in psychology. Adv Methods Pract Psychol Sci 2020;3:229–237

Appendix F. Chapter 5 Supplementary Figures

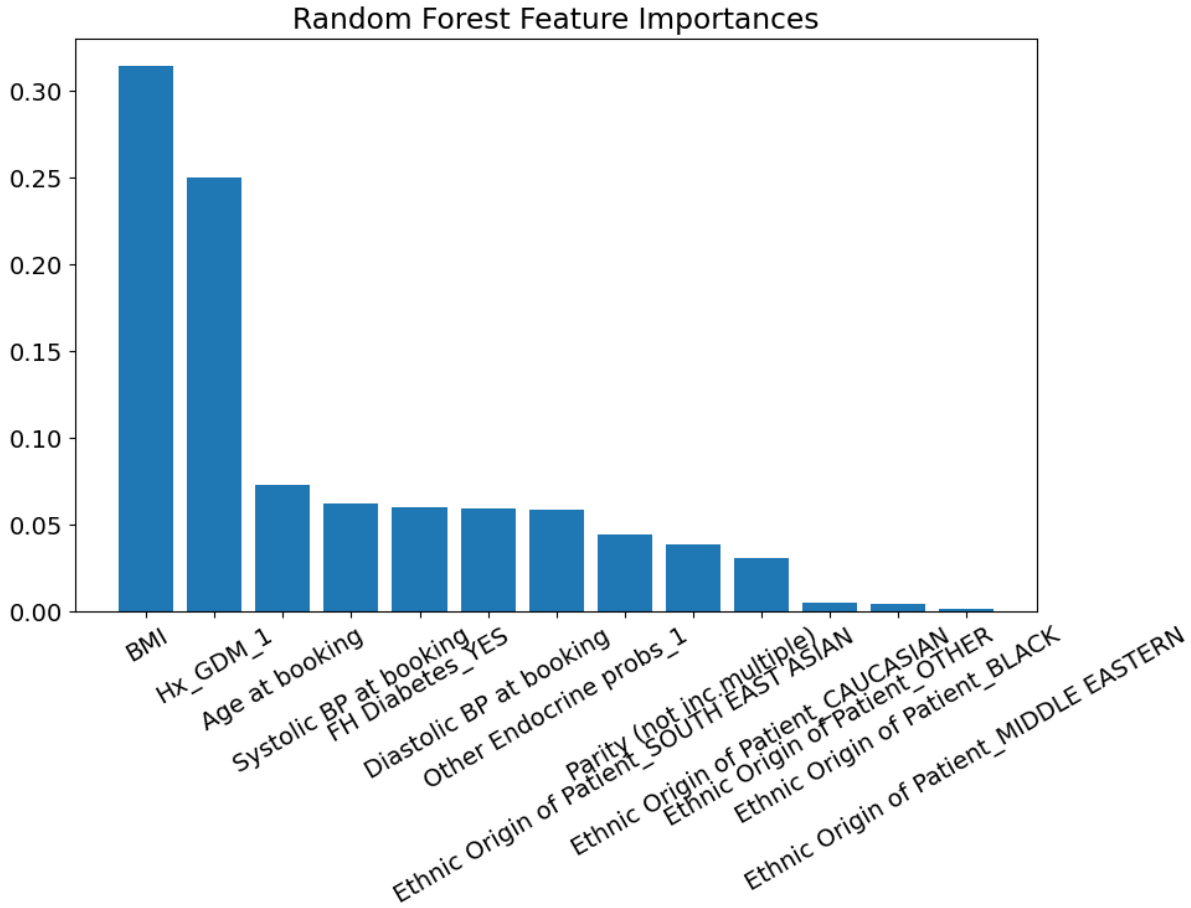
First Trimester XGBoost Model



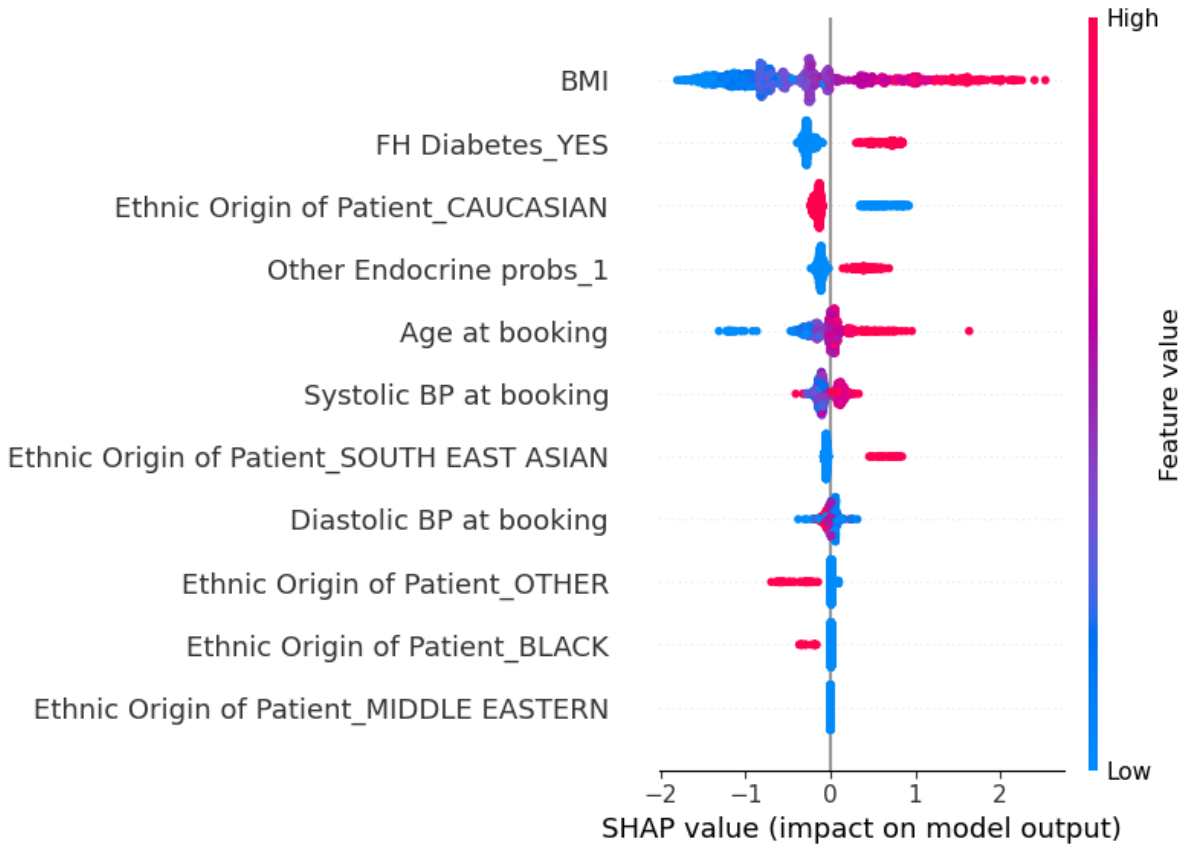
First Trimester EBM Model



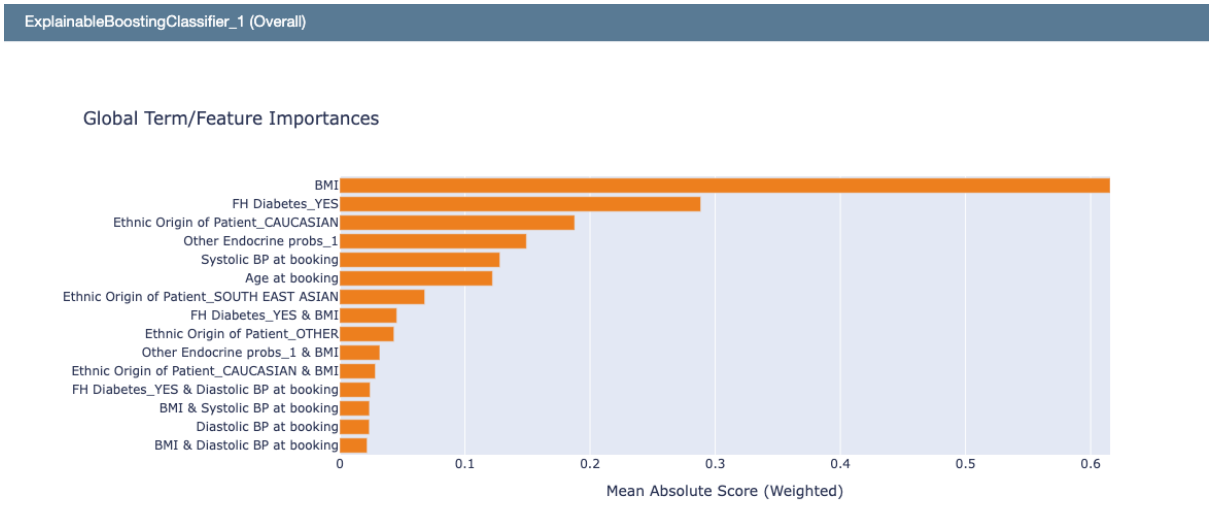
First Trimester RF Model



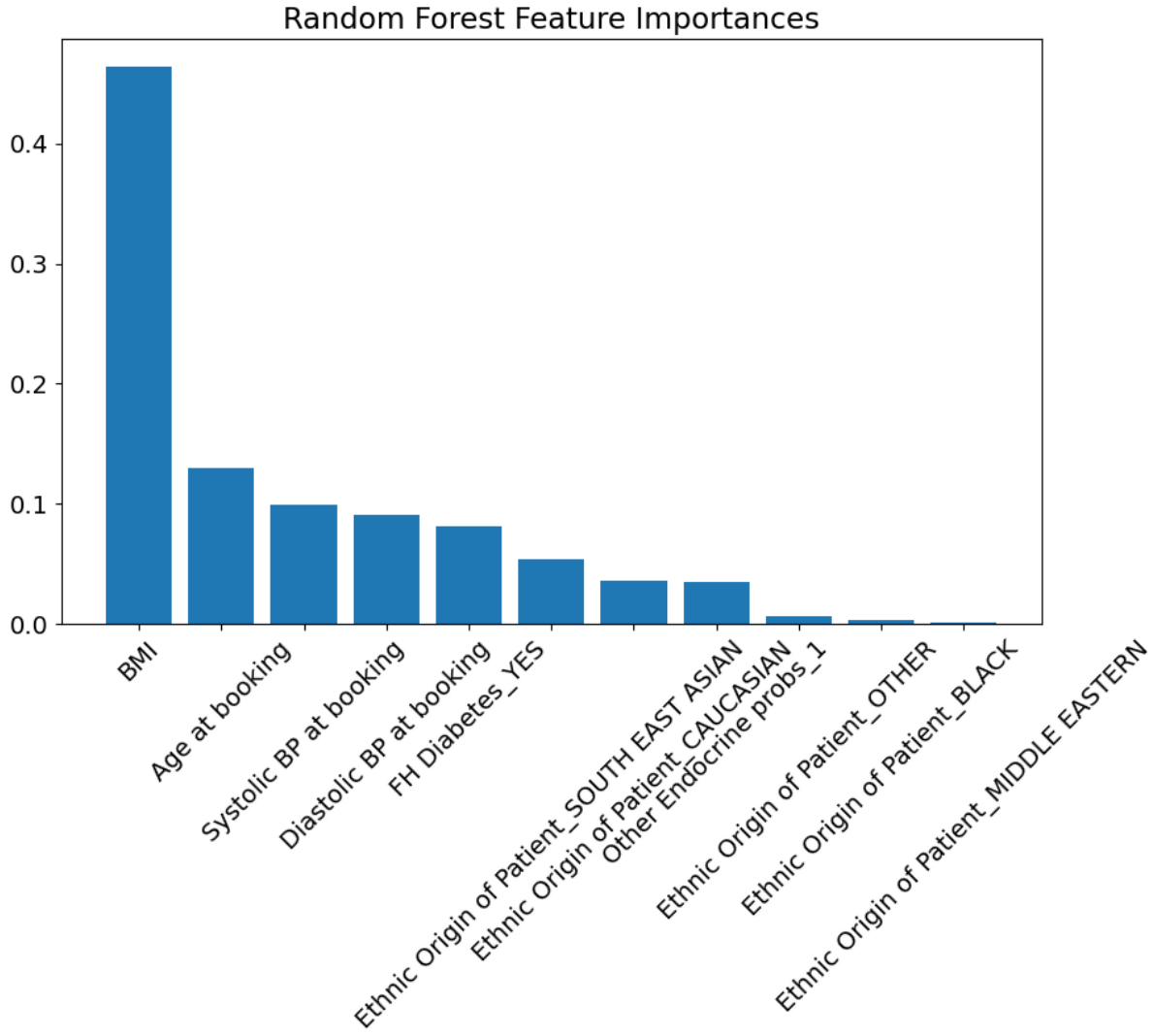
Nulliparous XGB Model



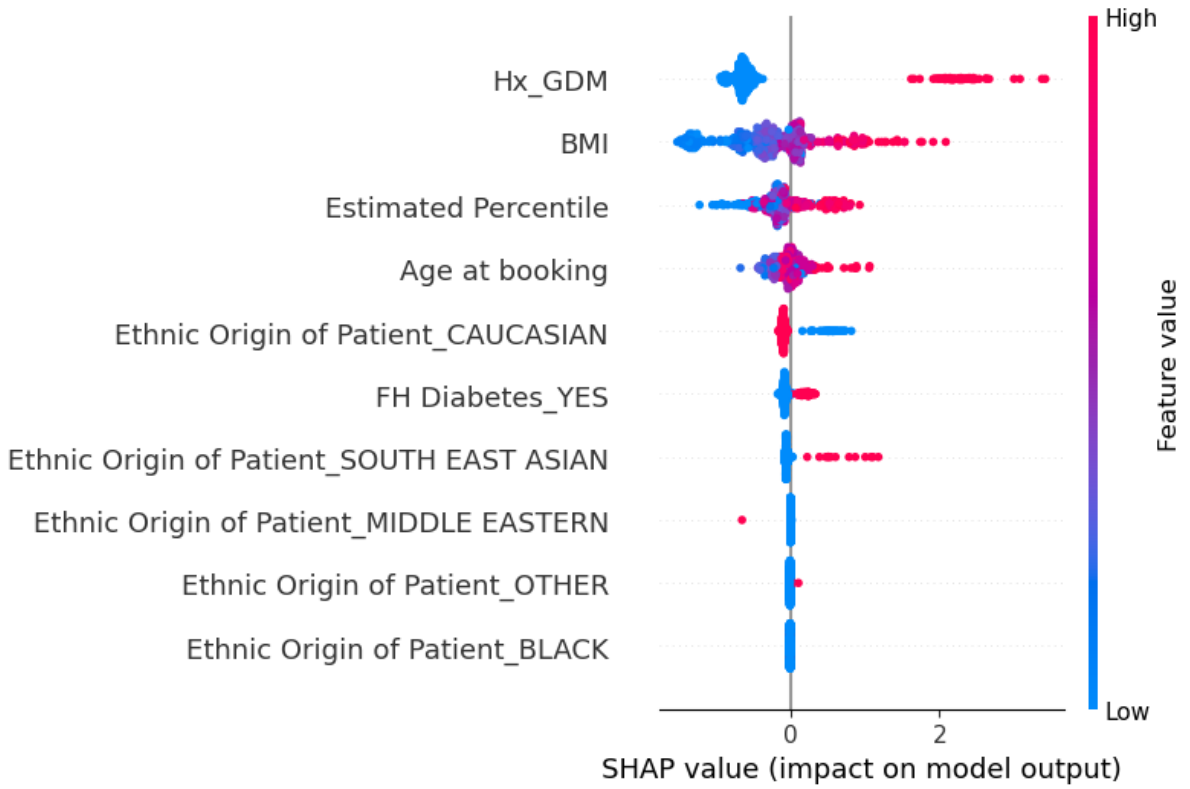
Nulliparous EBM Model



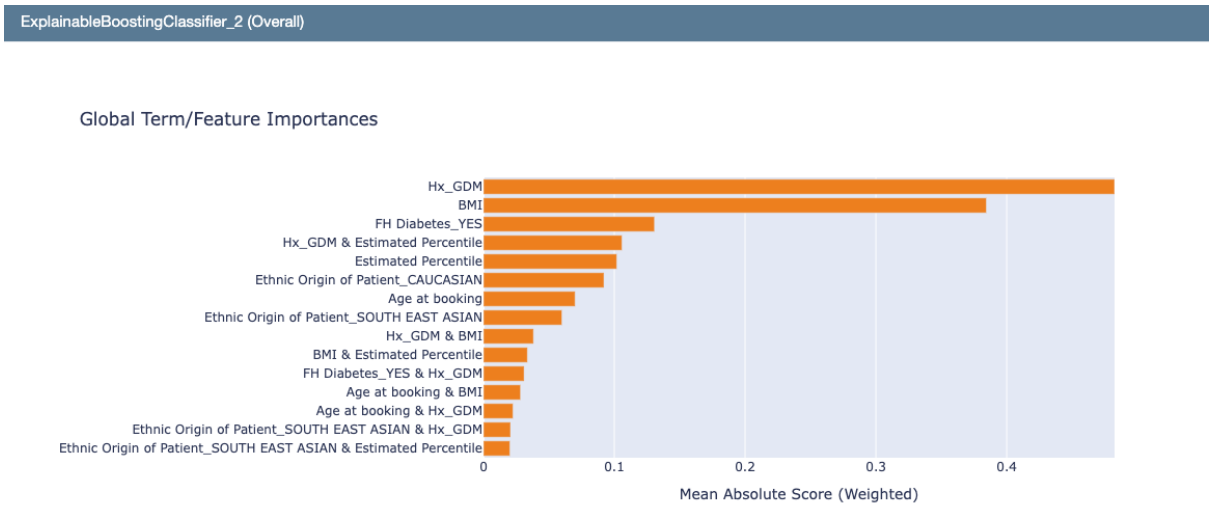
Nulliparous RF Model



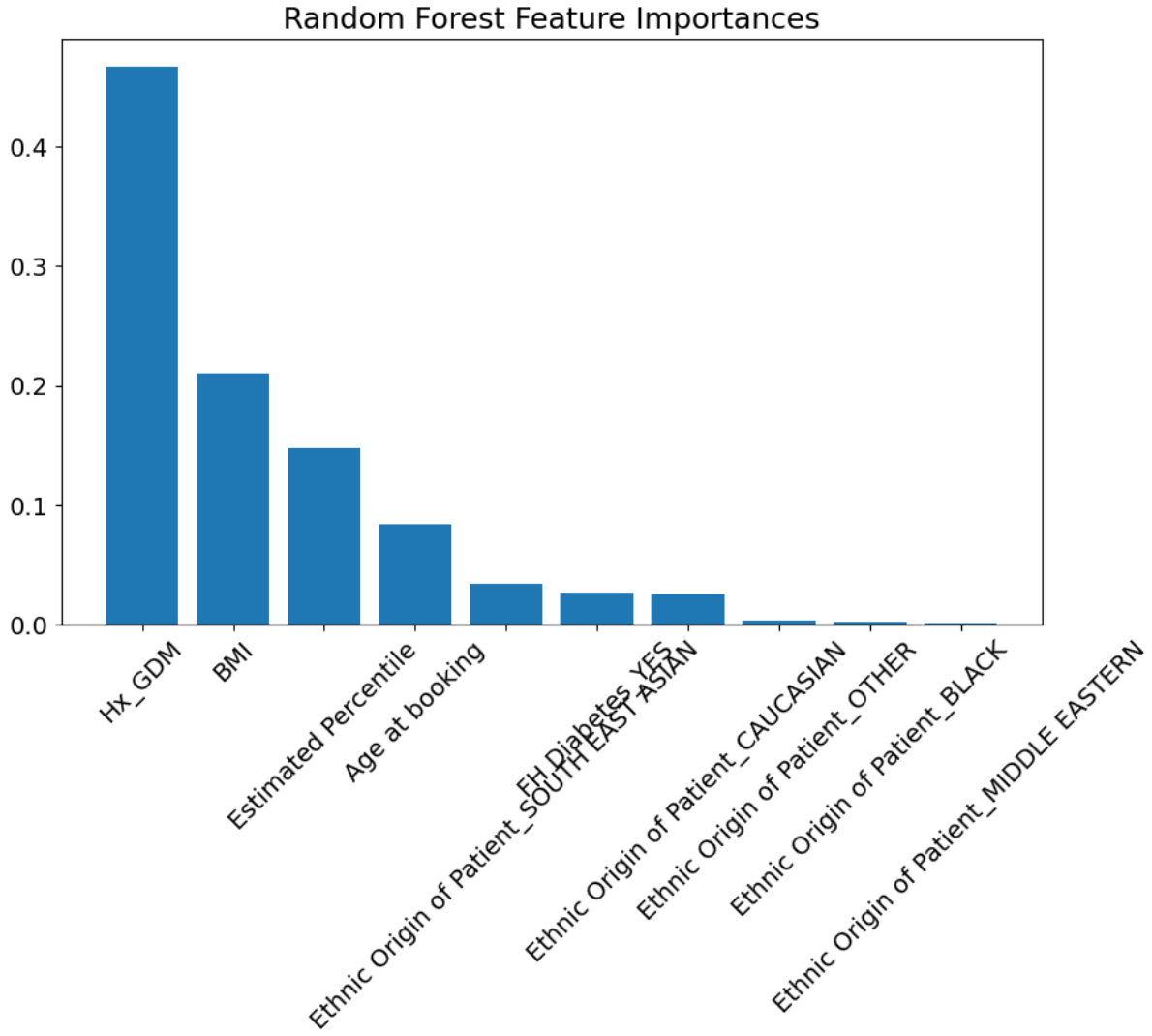
Past Pregnancy XGB Model



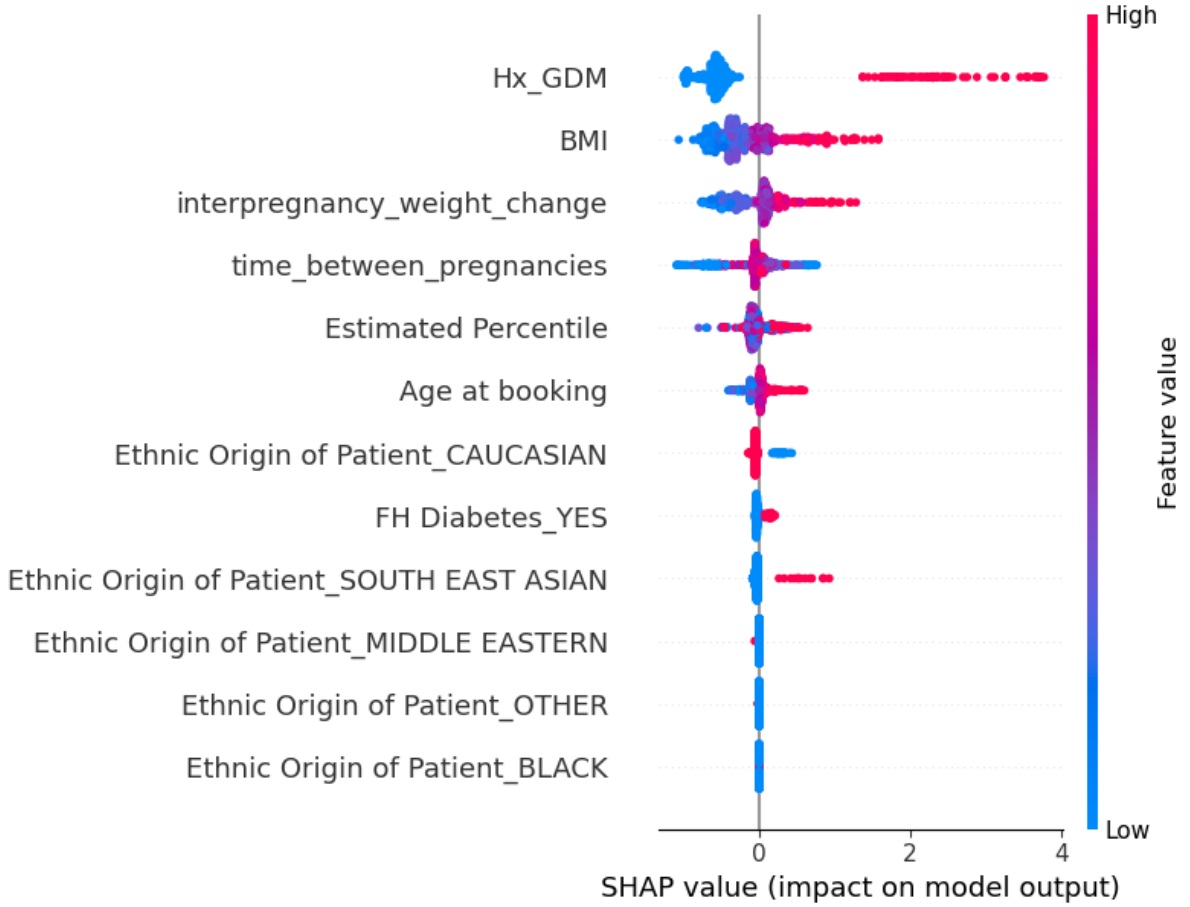
Past Pregnancy EBM Model



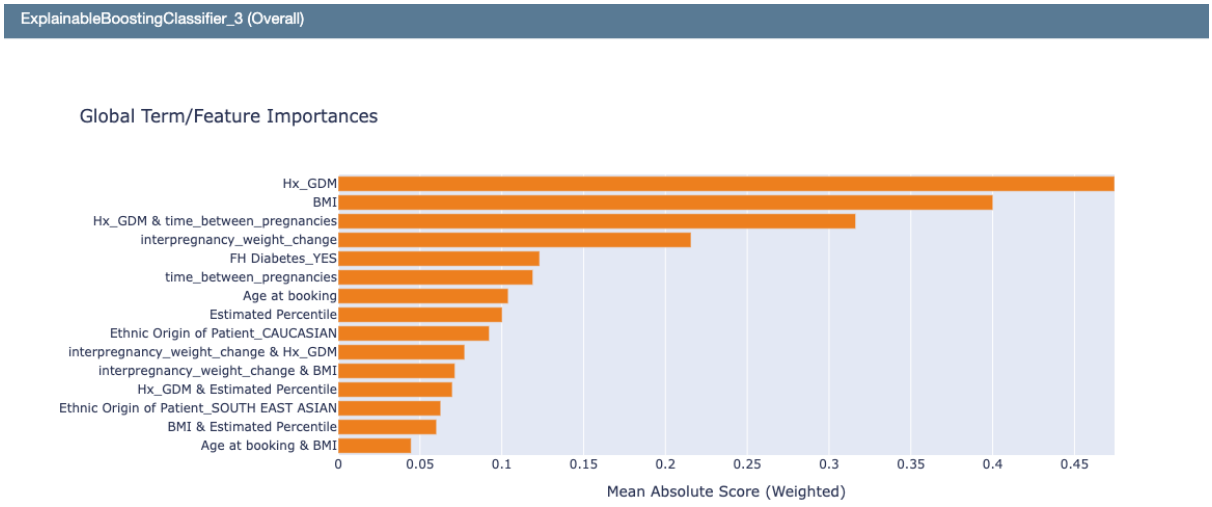
Past Pregnancy RF Model



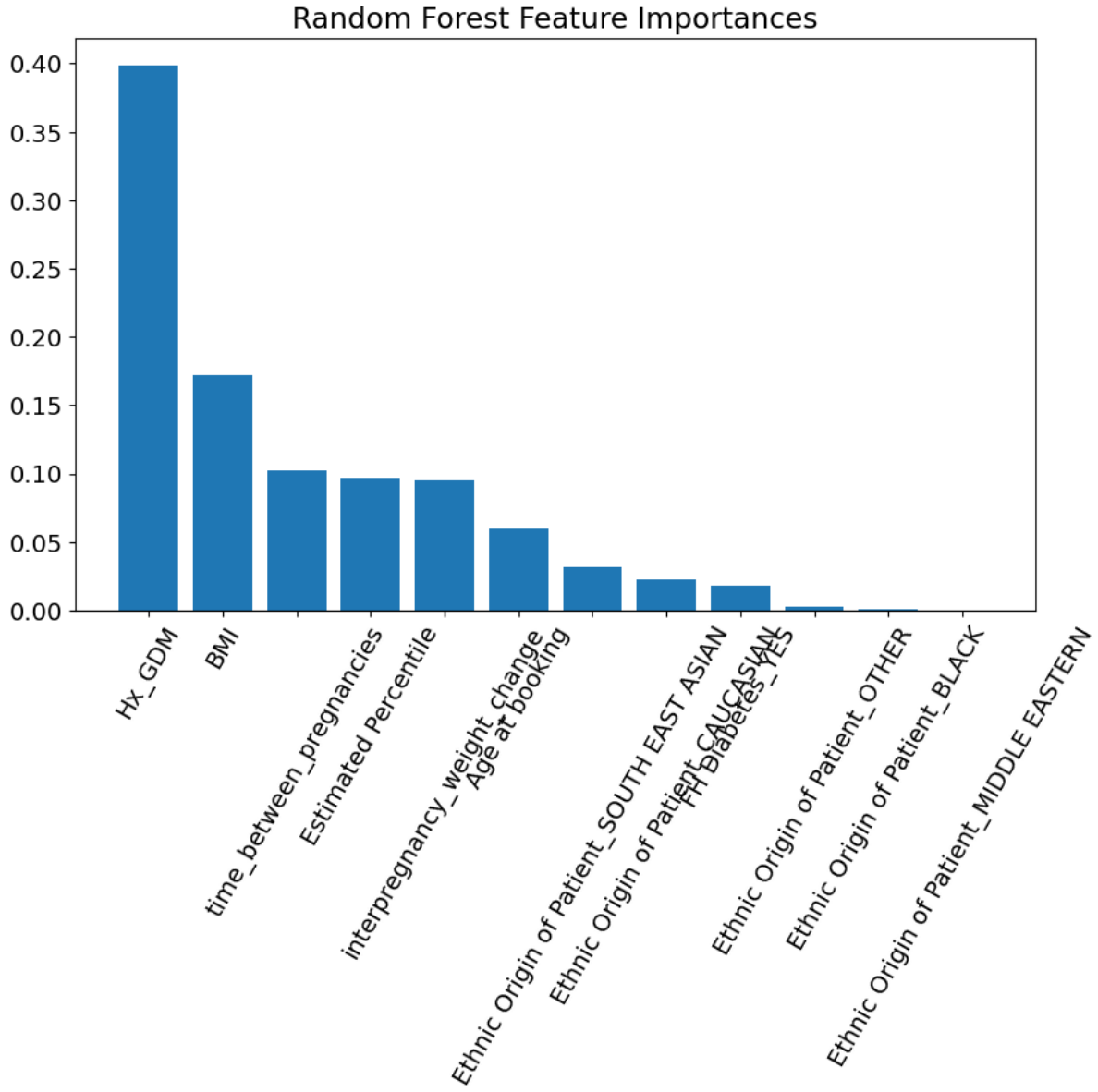
Multiparous XGB Model



Multiparous EBM Model



Multiparous RF Model



Appendix G. Chapter 5 Supplementary Tables

S Table 2. Patient characteristics grouped by ethnicity expressed as number of occurrences (%)

Characteristic	Caucasian n=24,180	Southeast Asian n=1,360	Black n=554	Asian n=489	Middle Eastern n=154	Other n=824
FH Diabetes*						
No	19,161 (79.2)	603 (44.3)	390 (70.4)	343 (70.1)	83 (53.9)	574 (69.7)
Yes	5,019 (20.8)	757 (55.7)	164 (29.6)	146 (29.9)	71 (46.1)	250 (30.3)
Hx of GDM*						
No	23,385 (96.7)	1,187 (87.3)	518 (93.5)	465 (95.1)	140 (90.1)	788 (95.6)
Yes	795 (3.3)	173 (12.7)	36 (6.5)	24 (4.9)	14 (9.9)	36 (4.4)
GDM						
No	21,772 (90.0)	899 (66.1)	475 (85.7)	383 (78.3)	133 (86.4)	711 (86.3)
Yes	2,408 (10.0)	461 (33.9)	79 (14.3)	106 (21.7)	21 (13.6)	113 (13.7)
Parity+						
0	10,022 (41.4)	511 (37.6)	229 (41.3)	222 (45.4)	54 (35.1)	406 (49.3)
1	8,741 (36.1)	485 (35.7)	167 (30.1)	173 (35.4)	51 (33.1)	264 (32.0)
>=2	5,417 (22.4)	364 (26.8)	158 (28.5)	94 (19.2)	49 (31.8)	154 (18.7)
Age#						
Age (Mean ± SD)	33 ± 5	32 ± 5	31 ± 6	33 ± 5	31 ± 5	33 ± 5
>=40	1,934 (8.0)	50 (3.7)	50 (9.0)	51 (10.4)	4 (2.6)	89 (10.8)
<40	22,246 (92.0)	1,310 (96.3)	504 (91.0)	438 (89.6)	150 (97.4)	735 (89.2)
BMI						
BMI (Mean ± SD)	26.2 ± 4.3	26.1 ± 4.6	28.5 ± 6.1	24.2 ± 4.3	25.8 ± 4.0	25.6 ± 5.1
<25	12,092 (50.0)	601 (44.2)	182 (32.9)	316 (64.6)	76 (49.4)	439 (53.3)
25 to <30	7,298 (30.2)	540 (39.0)	173 (31.2)	117 (23.9)	56 (37.0)	237 (28.8)
30 to <35	3,081 (12.7)	176 (12.9)	116 (20.9)	44 (9.0)	18 (11.7)	102 (12.4)
35 to <40	1,149 (4.8)	36 (2.6)	52 (9.4)	10 (2.0)	1 (0.6)	32 (3.9)
>=40	560 (2.3)	17 (1.3)	31 (5.6)	2 (0.4)	2 (1.3)	14 (1.7)
Systolic BP (Mean ± SD)	111 ± 11	109 ± 10	113 ± 11	109 ± 11	107 ± 12	111 ± 10
Diastolic BP (Mean ± SD)	67 ± 8	67 ± 8	69 ± 9	66 ± 8	65 ± 8	67 ± 8

Dataset and Model	AUC	AP	Sensitivity	Specificity	F1 Score	Brier Score
Nulliparous Model Subset						
Random Forest Classifier	0.822	0.377	0.028	**0.997**	0.053	0.078
Logistic Regression	**0.825**	0.388	**0.106**	0.990	**0.178**	0.078
XGBoost Classifier	0.824	0.359	0.050	0.993	0.091	0.078
Explainable Boosting Machine	0.825	**0.390**	0.084	0.995	0.078	0.078
Dummy Classifier	0.503	0.098	0.108	0.899	0.106	0.178

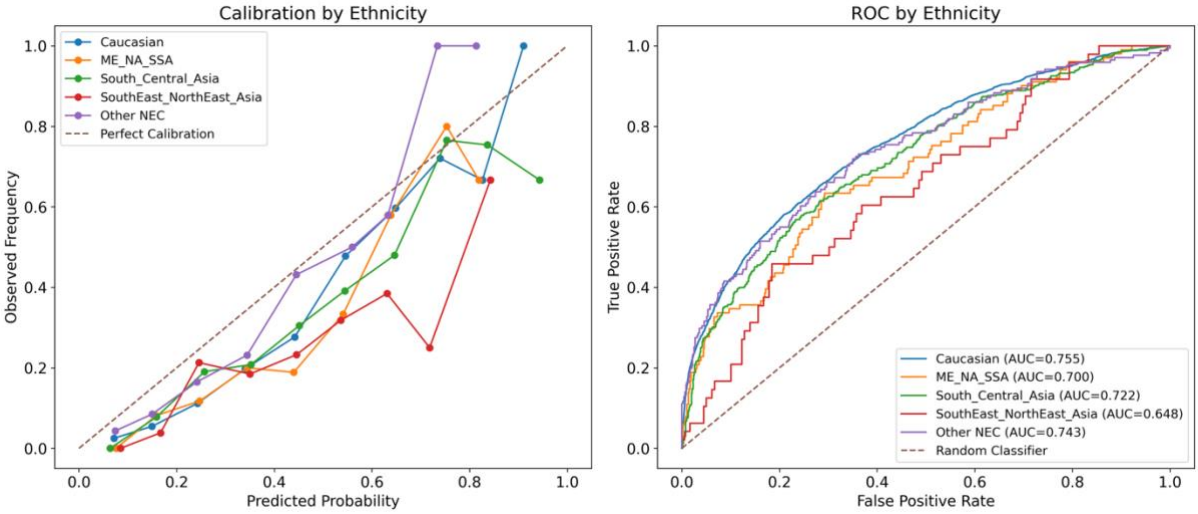
S Table 3. Final hyperparameters chosen for all models as a result of hyperparameter tuning. Only the tuned hyperparameters are displayed, all other parameters are thus default values.

Dataset and Model	Hyperparameters
First-Trimester Models	
Random Forest Classifier	RandomForestClassifier(max_depth=10, min_samples_leaf=4)
Logistic Regression	LogisticRegression(C=0.08858667904100823, max_iter=1000)
XGBoost Classifier	XGBClassifier(max_depth=3, n_estimators=100, learning_rate=1)
Explainable Boosting Machine	ExplainableBoostingClassifier()
FTP-9 Models	
Random Forest Classifier	RandomForestClassifier(max_depth=10, min_samples_leaf=4)
Logistic Regression	LogisticRegression(C=0.23357214690901212, max_iter=1000, solver='liblinear')
XGBoost Classifier	XGBClassifier(learning_rate=0.01, max_depth=5,n_estimators=300)
Explainable Boosting Machine	ExplainableBoostingClassifier(learning_rate=0.1, max_bins=256)
Nulliparous Model	
Random Forest Classifier	RandomForestClassifier(max_depth=10, min_samples_leaf=4)
Logistic Regression	LogisticRegression(C=0.03359818286283781, max_iter=1000, solver='liblinear')
XGBoost Classifier	XGBClassifier(max_depth=3, n_estimators=100, learning_rate=0.1)
Explainable Boosting Machine	ExplainableBoostingClassifier(learning_rate=0.1, max_bins=64)
Sequential Model (Previous Pregnancy Variables)	
Random Forest Classifier	RandomForestClassifier(max_depth=20, min_samples_leaf=4, min_samples_split=5,n_estimators=300)
Logistic Regression	LogisticRegression(C=0.03359818286283781, max_iter=1000, solver='liblinear')
XGBoost Classifier	XGBClassifier(max_depth=3, n_estimators=100, learning_rate=0.1)
Explainable Boosting Machine	ExplainableBoostingClassifier(learning_rate=1)
Sequential Model (Including First Trimester Data)	
Random Forest Classifier	RandomForestClassifier(min_samples_leaf=2, n_estimators=300)
Logistic Regression	LogisticRegression(C=0.03359818286283781, max_iter=1000, solver='liblinear')
XGBoost Classifier	XGBClassifier(max_depth=10, n_estimators=100, learning_rate=0.1)
Explainable Boosting Machine	ExplainableBoostingClassifier()
Sequential Model (Top 8 Features)	
Random Forest Classifier	RandomForestClassifier(max_depth=10, min_samples_split=5)
Logistic Regression	LogisticRegression(C=0.08858667904100823, max_iter=1000)
XGBoost Classifier	XGBClassifier(max_depth=5, n_estimators=300, learning_rate=0.01)
Explainable Boosting Machine	ExplainableBoostingClassifier()

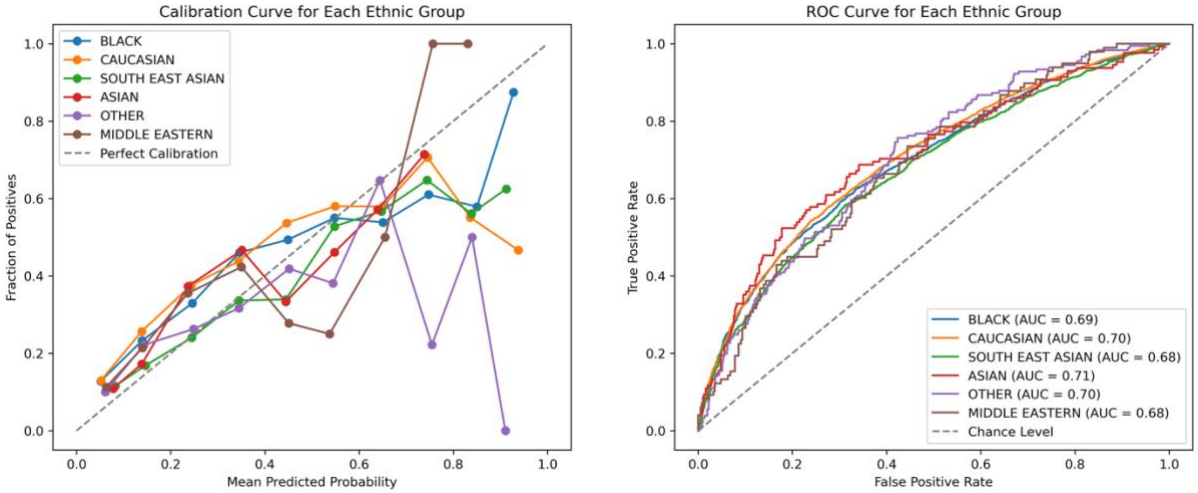
S Table 4. Performance of machine learning models trained on entirety of data available for each population.

Dataset and Model	AUC (95% CI)	Calibration Slope	Calibration Intercept	O:E Ratio	AP	Sensitivity	Specificity	F1 Score	Brier Score
First-Trimester Models									
Random Forest Classifier	0.819 (0.810–0.827)	1.194	0.288	0.986	0.440	0.141	0.993	0.236	0.083
Logistic Regression	0.823 (0.815–0.831)	0.998	-0.004	0.999	0.443	0.218	0.983	0.324	0.082
XGBoost Classifier	0.824 (0.816–0.832)	1.034	0.057	1.001	0.442	0.192	0.987	0.298	0.082
Explainable Boosting Machine	0.822 (0.814–0.830)	0.961	-0.055	1.006	0.443	0.212	0.985	0.320	0.082
Nulliparous Models									
Random Forest Classifier	0.810 (0.796–0.824)	1.311	0.568	1.000	0.326	0.005	0.999	0.011	0.077
Logistic Regression	0.818 (0.804–0.831)	1.036	0.052	0.989	0.347	0.105	0.990	0.176	0.076
XGBoost Classifier	0.816 (0.803–0.829)	1.004	-0.018	1.009	0.337	0.066	0.993	0.117	0.076
Explainable Boosting Machine	0.812 (0.799–0.825)	0.878	-0.134	1.067	0.343	0.135	0.987	0.215	0.076
Past Pregnancy Models									
Random Forest Classifier	0.854 (0.833–0.872)	1.480	0.679	0.980	0.556	0.163	0.993	0.268	0.077
Logistic Regression	0.848 (0.826–0.866)	1.018	-0.011	0.976	0.548	0.425	0.969	0.514	0.073
XGBoost Classifier	0.861 (0.841–0.879)	1.053	0.059	0.986	0.591	0.423	0.976	0.530	0.071
Explainable Boosting Machine	0.855 (0.835–0.873)	0.961	-0.030	1.019	0.570	0.445	0.971	0.570	0.072
Multiparous Models									
Random Forest Classifier	0.875 (0.858–0.891)	1.445	0.547	0.962	0.600	0.229	0.989	0.351	0.073
Logistic Regression	0.868 (0.849–0.884)	0.980	-0.051	0.988	0.558	0.437	0.971	0.530	0.072
XGBoost Classifier	0.887 (0.870–0.902)	0.606	-0.050	1.318	0.637	0.460	0.976	0.637	0.070
Explainable Boosting Machine	0.890 (0.873–0.905)	0.824	-0.113	1.074	0.638	0.487	0.976	0.586	0.066

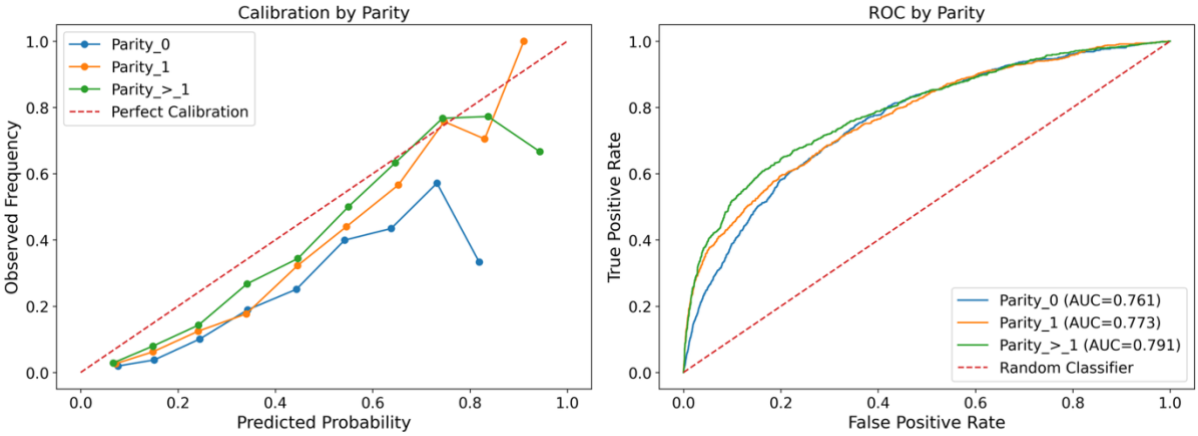
Appendix H. Chapter 6 Supplementary Figures



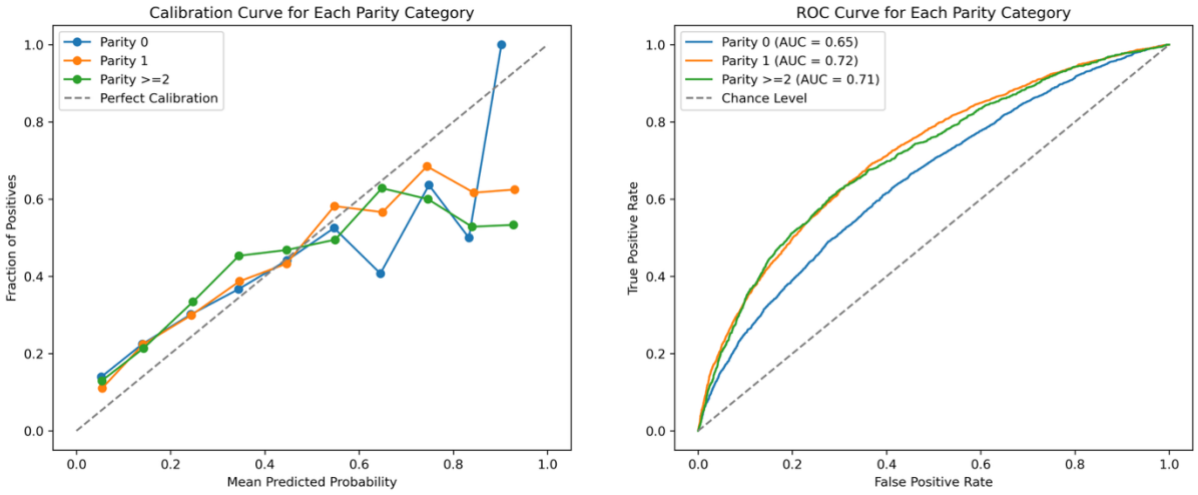
1. Monash model validated on Irish Cohort, by ethnicity.



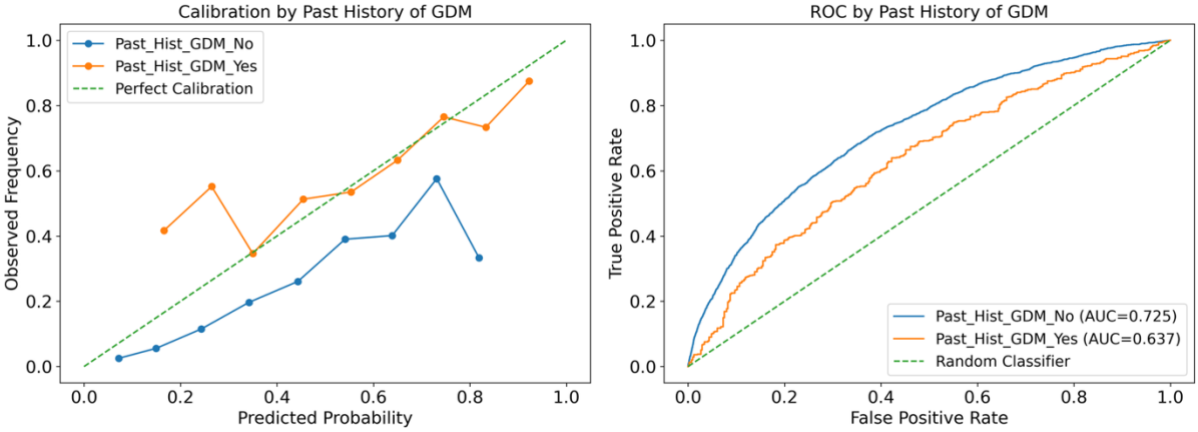
2. DCU model validated on Australian Cohort, by ethnicity.



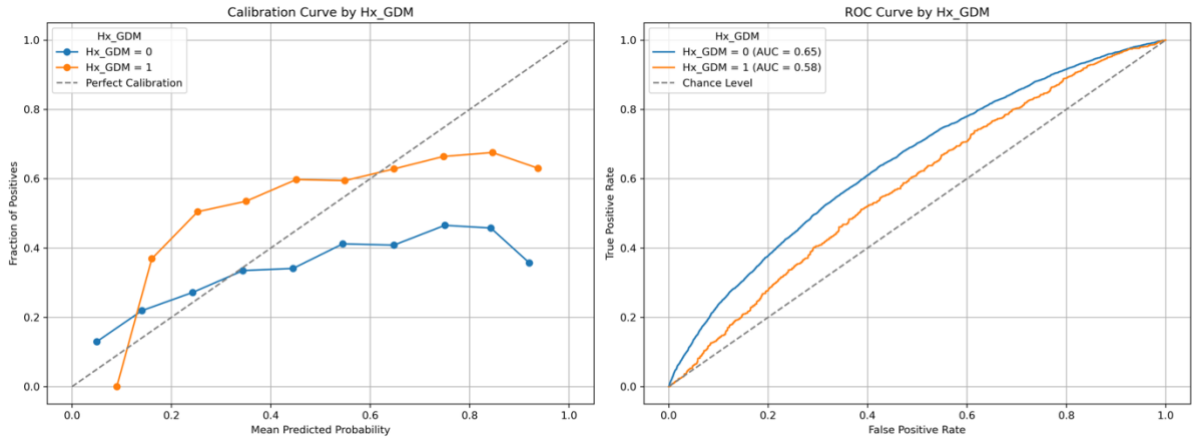
3. Monash model validated on Irish Cohort, by parity.



4. DCU model validated on Australian Cohort, by parity.



5. Monash model validated on Irish Cohort, by history of GDM.



6. DCU model validated on Australian Cohort, by parity.

Appendix I. Predicting GDM Treatment

TITLE

Evaluation of Machine Learning Models for Prediction of the Treatment Gestational Diabetes Using Electronic Health Records

SHORT RUNNING TITLE

Early prediction of GDM Treatment

AUTHOR LIST

Mark Germaine[1,2,5], Amy C O'Higgins[3], Brendan Egan[2,4], Graham Healy[1]

AFFILIATIONS

[1] School of Computing, Dublin City University, Dublin 9, Ireland

[2] School of Health and Human Performance, Dublin City University, Dublin 9, Ireland

[3] UCD Centre for Human Reproduction, The Coombe Hospital, Dublin 8, Ireland

[4] Florida Institute for Human and Machine Cognition, Pensacola FL, USA

[5] SFI Centre for Research Training in Machine Learning, Dublin City University, Dublin 9, Ireland

ORCIDs

Mark Germaine 0000-0002-7862-7714

Amy O'Higgins 0000-0002-2020-1585

Brendan Egan 0000-0001-8327-9016

Graham Healy 0000-0001-6429-6339

CORRESPONDING AUTHOR

Mark Germaine, Research Ireland Centre for Research Training in Machine Learning, Dublin City University, Glasnevin, Dublin 9, Ireland

e: mark.germaine2@mail.dcu.ie

INTRODUCTION

Several groups have modelled treatment requirements in gestational diabetes mellitus (GDM), but existing tools differ in scope and performance. Liao et al.^1 analysed >30 000 GDM pregnancies in Kaiser Permanente, comparing interpretable models (e.g. LASSO) with an ensemble "super learner" that combined decision trees, random forests and XGBoost. Using electronic health-record (EHR) data available up to one week after diagnosis, their ensemble achieved excellent discrimination in development (cross-validated AUC ~0.93) and good temporal validation (AUC ~0.81), indicating that rich EHR features can predict the need for any pharmacotherapy.

Other studies focus on insulin alone. Eleftheriades et al.^2 prospectively followed 775 women and used logistic regression and CART to predict insulin initiation. Higher pre-pregnancy BMI and fasting OGTT glucose > 98 mg/dL were independent predictors (OR 2.21 and 4.04, respectively); the model reached AUC ~0.74 in both internal and external validation. Rostin et al.^3 derived the CHANGED score in 1 611 women using stepwise regression on variables available at diagnosis (age, BMI, parity, foetal sex, and OGTT values). The score stratified insulin risk with AUC 0.77 (95% CI 0.75–0.80), 72% sensitivity and 69% specificity at the optimal cut-off. Despite these advances, limitations persist: moderate predictive power, infrequent external validation, and binary outcomes that fail to distinguish oral agents from insulin. Metformin is now a common first-line therapy, yet most models treat all pharmacotherapy as a single category.

Therefore, the aim of this study is to develop and evaluate ML models that classify GDM treatment into three categories: diet only, metformin, or insulin, using variables available at diagnosis (demographics, medical history, and OGTT results). We will also assess two binary tasks: (i) any pharmacotherapy versus diet alone and (ii) insulin versus no insulin. Multiple algorithms will be compared, and the incremental value of OGTT glucose concentrations will be examined. The goal is a high-accuracy, interpretable model suitable for decision support, enabling earlier, tailored management of GDM.

METHODS

Study Design and Ethical Approval

A retrospective cohort study of pregnant women diagnosed with GDM at the Coombe Hospital (Dublin, Ireland). Ethical approved granted by the Coombe Hospital Research Ethics Committee (Study No. 06–2023). Patient data were de-identified and handled in compliance

with GDPR and hospital data protection policies. Given the retrospective design using existing clinical records, the ethics committee granted a waiver of informed consent.

Data Sources and Participants

Data for this study were obtained from two primary sources, the hospital's Diabetes in Pregnancy clinic database and the main maternity EHR system. All women who were diagnosed with GDM between 2018 and 2022 and who gave birth at the study hospital. GDM was defined according to IADPSG criteria (fasting glucose ≥5.1 mmol/L, 1-hour ≥10.0 mmol/L, or 2-hour ≥8.5 mmol/L). To be included, participants needed to have a recorded OGTT result from 24–28 weeks' gestation (or earlier if high risk) and complete follow-up data on their GDM treatment course. We excluded pregnancies with pre-gestational diabetes (Type 1 or Type 2 diabetes diagnosed before pregnancy) and those with multifetal gestations (twins or higher-order pregnancies), as these conditions require different management. After applying inclusion and exclusion criteria, the remaining eligible GDM cases were compiled into the analysis dataset. Data preparation was described in Chapter 3.

Outcome Definition

The primary outcome was the GDM treatment modality, categorised into three groups: Diet-only, Metformin, or Insulin. Diet-only indicated that the patient's glucose concentrations were managed through diet and lifestyle measures alone, with no pharmacological treatment needed throughout the pregnancy. Metformin indicated that the patient required oral metformin to control GDM, with no insulin use. Insulin indicated that the patient required insulin injections (with or without concomitant metformin) to achieve glycaemic control. For analysis purposes, patients who needed insulin at any point in pregnancy were classified in the Insulin group (even if they also were on metformin or diet modifications), since insulin represents the most intensive therapy. In addition to this three-class outcome, I defined two binary outcomes for secondary analyses: (1) Medication vs. Diet, distinguishing patients who required any pharmacotherapy (either metformin or insulin) from those managed by diet alone; and (2) Insulin vs. No Insulin, distinguishing those who required insulin from those who did not (the latter includes diet-only and metformin-managed GDM). These binary classifications allow focused evaluation of predicting any escalation of care and specifically the need for insulin, respectively.

Machine Learning Model Development

We developed four supervised models to predict GDM treatment: logistic regression, random forest, XGBoost, and CatBoost Classifier. Data were split 80%/20% into training and hold-out test sets. Within the training set we ran stratified five-fold cross-validation to preserve the diet, metformin and insulin class proportions. Hyper-parameters were tuned by random search, selecting the configuration with the highest mean balanced accuracy. We trained separate sets of models for the two feature sets described above. Models without OGTT data used all 3,185 cases and included only the non-OGTT predictors. Models with OGTT data were developed using the subset of 486 cases with available OGTT glucose values, incorporating those values as additional features. In total, we built six primary classification models: three outcome predictions (multiclass, any treatment vs diet, insulin vs no insulin) × two feature sets (without vs with OGTT). All models were developed in Python using scikit-learn and related libraries (CatBoost, XGBoost).

In a separate exploratory analysis, we also attempted to predict the magnitude of the OGTT results themselves from the non-OGTT features. For the subset of 486 patients, we trained regression models to predict each of the three OGTT glucose concentrations (fasting, 1-hour, 2-hour) based on the other baseline variables. This was done to investigate whether one could approximate the OGTT values from routine clinical data. We fit a random forest regressor, an XGBoost regressor, and a linear ridge regression model for each glucose time-point. Model performance for these regression tasks was evaluated with R-squared and Pearson correlation to actual values, as well as mean absolute error (MAE) and root mean square error (RMSE).

Evaluation of Models

We evaluated model performance using multiple metrics to capture different aspects of predictive ability. For the multiclass classification (Diet vs Metformin vs Insulin), we calculated overall balanced accuracy (the average of sensitivity/recall across all classes, to account for class imbalance) as well as the per-class precision, recall, and F1-score. We also examined the confusion matrix to see where misclassifications were occurring (e.g., predicting Metformin when the true class was Insulin, etc.). For the binary classification tasks, we focused on the positive class (pharmacotherapy needed, and insulin needed, respectively) and, AUC, sensitivity (recall), specificity, precision, F1-score, and balanced accuracy. In addition, we computed the Area Under the ROC Curve (ROC AUC) for each model as a threshold-independent measure of discrimination. Calibration of the predictions was assessed with the

Brier score, which is the mean squared error of the probability predictions; a lower Brier score indicates better calibrated and more accurate probabilistic predictions.

Results

Participant Characteristics

A total of 3,185 pregnant women with GDM were included. Table 1 summarises their baseline demographic and clinical characteristics overall and stratified by final treatment modality (diet only, metformin, or insulin). The mean age of the cohort was 33 ± 5 years, and the mean pre-pregnancy BMI was 30.4 ± 6.4 kg/m². Women who eventually required insulin had higher BMI and higher glucose concentrations on the diagnostic OGTT compared to those managed with diet or metformin. A history of prior GDM and a family history of diabetes were also more frequent among women needing pharmacotherapy, especially insulin.

Table 1. Characteristics of the two cohorts.

Category	Full (n=3,185)	Diet %	Met %	Ins %	OGTT (n=487)	Diet %	Met %	Ins %
Ethnicity								
Asian	106 (3.3)	52.8	27.4	19.8	17 (3.5)	41.2	17.6	41.2
Black African	79 (2.5)	48.1	24.1	27.8	13 (2.7)	69.2	15.4	15.4
Caucasian	2,405 (75.5)	46.5	31.0	22.5	359 (73.7)	49.3	31.2	19.5
Middle-Eastern	21 (0.7)	33.3	38.1	28.6	1 (0.2)	0.0	100.0	0.0
Other	113 (3.5)	44.2	31.0	24.8	20 (4.1)	45.0	25.0	30.0
South-East Asian	461 (14.5)	34.1	33.2	32.8	77 (15.8)	32.5	31.2	36.4
FH Diabetes								
No	1,783 (56.0)	46.4	32.9	20.8	257 (52.8)	48.2	30.4	21.4
Yes	1,402 (44.0)	42.7	28.7	28.5	230 (47.2)	44.8	30.0	25.2
Hx GDM								
0	2,511 (78.8)	49.9	31.5	18.6	389 (79.9)	51.4	29.6	19.0
1	674 (21.2)	25.5	29.4	45.1	98 (20.1)	27.6	32.7	39.8
Endocrine Disorders								
No	1,849 (58.1)	52.9	30.9	16.1	298 (61.2)	53.4	29.5	17.1
Yes	1,336 (41.9)	33.5	31.2	35.3	189 (38.8)	36.0	31.2	32.8
Age								
< 40 y	2,812 (88.3)	44.5	31.8	23.8	420 (86.2)	46.4	29.8	23.8
≥ 40 y	373 (11.7)	47.2	25.5	27.3	67 (13.8)	47.8	32.8	19.4
BMI								
< 25	674 (21.2)	52.1	30.0	18.0	119 (24.4)	54.6	26.9	18.5
25 – < 30	973 (30.5)	47.2	31.2	21.6	137 (28.1)	44.5	32.1	23.4
30 – < 35	853 (26.8)	43.7	31.3	25.0	134 (27.5)	44.8	32.8	22.4

Table 1. Characteristics of the two cohorts.

Category	Full (n=3,185)	Diet %	Met %	Ins %	OGTT (n=487)	Diet %	Met %	Ins %
35 – < 40	423 (13.3)	36.4	32.9	30.7	61 (12.5)	44.3	26.2	29.5
≥ 40	262 (8.2)	34.0	29.4	36.6	36 (7.4)	38.9	30.6	30.6
Parity								
0	1,121 (35.2)	52.7	29.0	18.3	169 (34.7)	55.0	29.6	15.4
1	1,122 (35.2)	40.6	32.9	26.6	174 (35.7)	43.1	32.2	24.7
≥ 2	942 (29.6)	40.3	31.3	28.3	142 (29.6)	40.8	31.0	28.2

Multiclass Prediction: Diet v Metformin v Insulin

Multiclass discrimination was limited without OGTT data (best = logistic regression, balanced accuracy 0.45, ROC-AUC 0.64, macro-F1 0.42), but adding OGTT values in the 486-patient subset raised performance: XGBoost yielded the highest balanced accuracy (0.57) while CatBoost delivered the strongest overall discrimination (ROC-AUC 0.79, macro-F1 0.51) and a lower Brier score (~0.20), indicating clearly improved but still moderate accuracy in separating diet, metformin-only and insulin groups (Table 2).

Table 2. Evaluation of ML models for multiclass prediction of GDM treatment.

Model (algorithm)	Balanced accuracy	ROC-AUC	Macro-F1	Brier
Without OGTT (n=3,185)				
Logistic Regression	0.445	0.639	0.421	0.226
Random Forest	0.422	0.620	0.409	0.229
XGBoost	0.409	0.610	0.402	0.231
CatBoost	0.425	0.639	0.416	0.228
With OGTT (n=486)				
Decision Tree	0.496	0.669	0.435	0.247
Random Forest	0.502	0.742	0.461	0.220
XGBoost	0.565	0.735	0.489	0.200
CatBoost	0.551	0.786	0.514	0.203

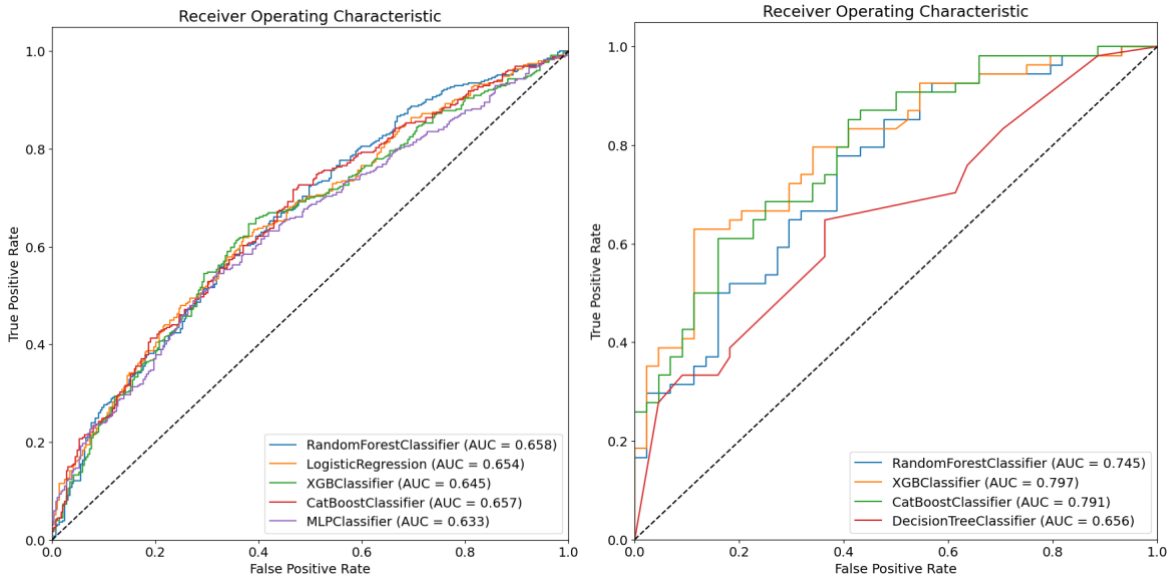
Binary Prediction: Medication vs Diet

Including OGTT values modestly but consistently strengthened prediction of any pharmacotherapy: the best non-OGTT model (Random Forest) reached AUC of 0.68 (Figure 1) with sensitivity = 0.75/specificity = 0.46, whereas the OGTT-enhanced XGBoost improved

discrimination to AUC = 0.80 (AP = 0.84, Figure 1) and achieved a more balanced operating point (sens = 0.80, spec = 0.66). CatBoost with OGTT delivered the highest sensitivity (0.87) but at the cost of lower specificity (0.52). Full results are reported in Table 3.

Table 3. Evaluation of models in predicting medication vs no medication.

Model (algorithm)	Sens	Spec	ROC-AUC	AP	F1	Brier
Without OGTT (n=3,185)						
Random Forest	0.746	0.460	0.676	0.685	0.682	0.230
Logistic Regression	0.703	0.505	0.654	0.706	0.668	0.230
XGBoost	0.740	0.426	0.645	0.687	0.670	0.233
CatBoost	0.751	0.484	0.657	0.699	0.692	0.232
With OGTT (n=486)						
Random Forest	0.759	0.614	0.745	0.790	0.732	0.205
Decision Tree	0.648	0.636	0.656	0.709	0.667	0.269
XGBoost	0.796	0.659	0.797	0.838	0.768	0.187
CatBoost	0.870	0.523	0.791	0.831	0.770	0.205



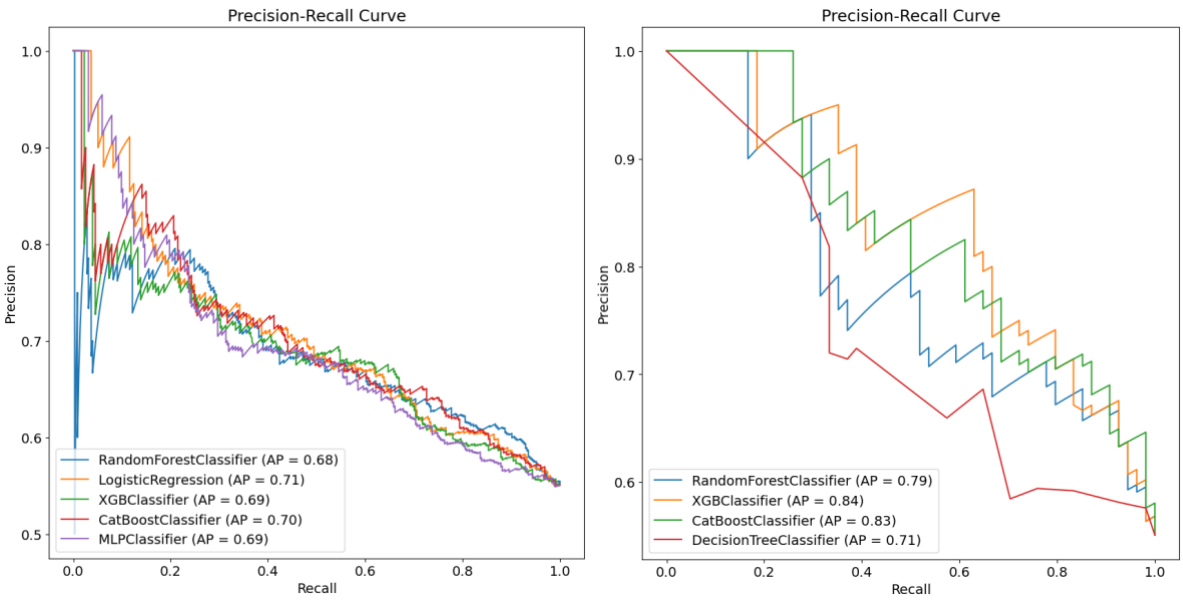


Figure 1. Top left: ROC curve for diet vs medication using full cohort (n=3,185). Top right: ROC curve for diet vs medication using OGTT cohort (n=486). Bottom left: Average precision for diet vs medication using full cohort (n=3,185). Average precision for diet vs medication using OGTT cohort (n=486).

Binary Prediction: Insulin vs No-Insulin

Including OGTT glucose values markedly improved discrimination and recall for predicting insulin requirement: the best non-OGTT model (logistic regression) achieved an AUC = 0.69 (Figure 2) with very low sensitivity (≤ 0.08), whereas the OGTT-enhanced Random Forest raised AUC to 0.87 and sensitivity to 0.52 while preserving high specificity (0.97) and halving the Brier score. CatBoost with OGTT offered the strongest overall discrimination (AUC = 0.90) and perfect specificity, but at a modest sensitivity of 0.44. Full results are reported in Table 4.

Table 4. Evaluation of models in predicting insulin vs no insulin.

Model (algorithm)	Sens	Spec	ROC-AUC	AP	F1	Brier
Without OGTT (n=3,185)						
Random Forest	0.046	0.992	0.676	0.407	0.085	0.168
Logistic Regression	0.078	0.990	0.692	0.461	0.141	0.163
XGBoost	0.039	0.986	0.679	0.398	0.072	0.167
CatBoost	0.013	0.998	0.684	0.406	0.026	0.168
With OGTT (n=486)						
Random Forest	0.522	0.973	0.874	0.786	0.649	0.113
Decision Tree	0.435	0.947	0.837	0.619	0.541	0.130

Table 4. Evaluation of models in predicting insulin vs no insulin.

Model (algorithm)	Sens	Spec	ROC-AUC	AP	F1	Brier
XGBoost	0.348	0.973	0.879	0.758	0.485	0.128
CatBoost	0.435	1.000	0.903	0.827	0.606	0.134

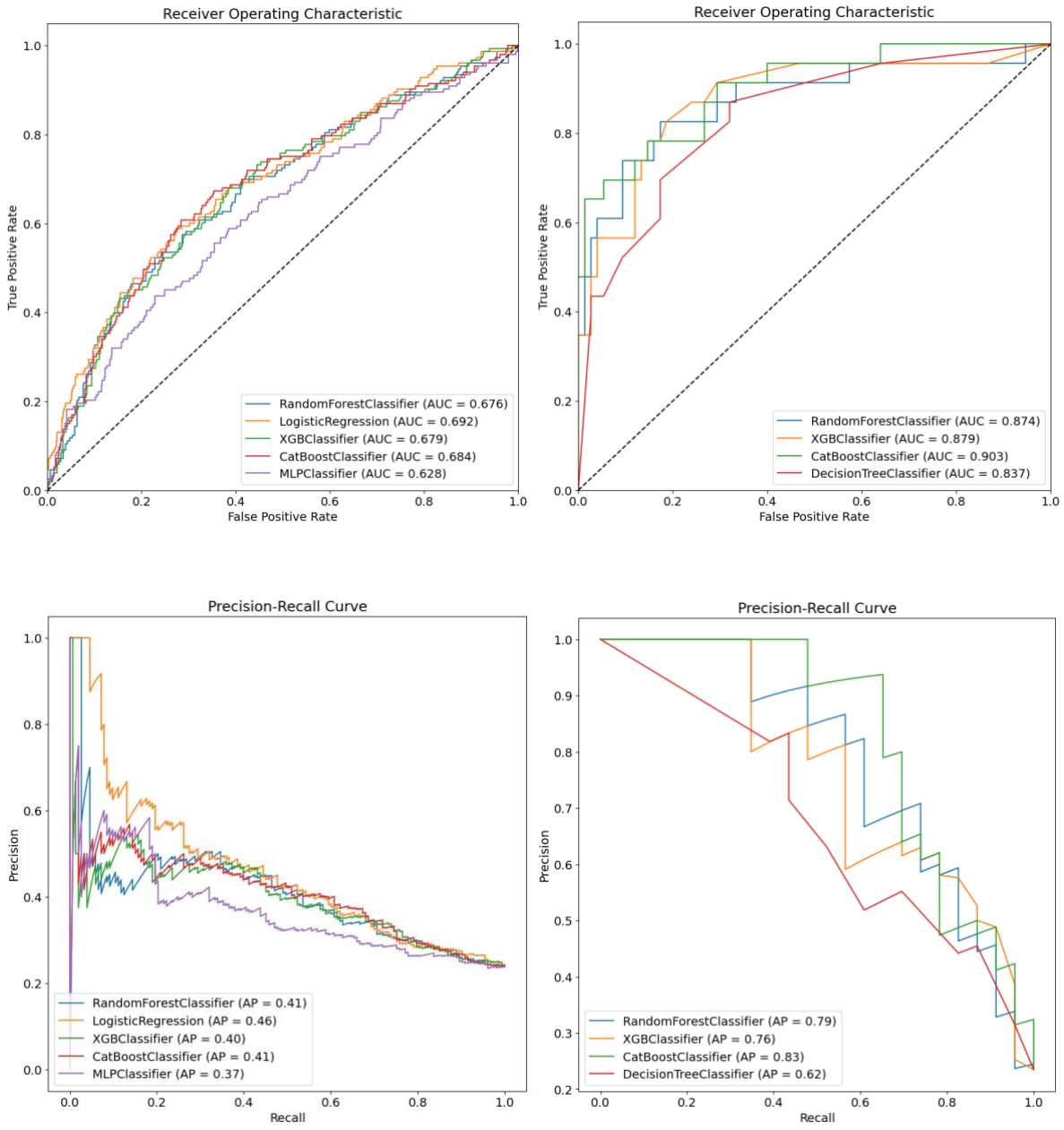


Figure 2. Top left: ROC curve for insulin vs no insulin using full cohort (n=3,185). Top right: ROC curve for insulin vs no insulin using OGTT cohort (n=486). Bottom left: Average precision for insulin vs no insulin using full cohort (n=3,185). Average precision for insulin vs no insulin using OGTT cohort (n=486).

OGTT value regression (n = 486)

Table 4 gives test-set performance predicting fasting (GTT-0 h), 1-h and 2-h glucose values from non-OGTT features.

Model	GTT-0 h R^2	GTT-1 h R^2	GTT-2 h R^2	r (0 h / 1 h / 2 h)	MAE (mmol L^{-1})	RMSE (mmol L^{-1})
Random-forest	0.07	–0.07	0.08	0.26 / 0.20 / 0.34	0.37 / 0.41 / 1.27	0.52 / 0.55 / 1.76
Ridge	0.11	–0.02	0.01	0.37 / 0.08 / 0.14	0.35 / 0.41 / 1.32	0.51 / 0.55 / 1.83
XGB-regressor	–0.07	–0.12	0.09	0.20 / 0.10 / 0.34	0.41 / 0.41 / 1.26	0.55 / 0.55 / 1.75

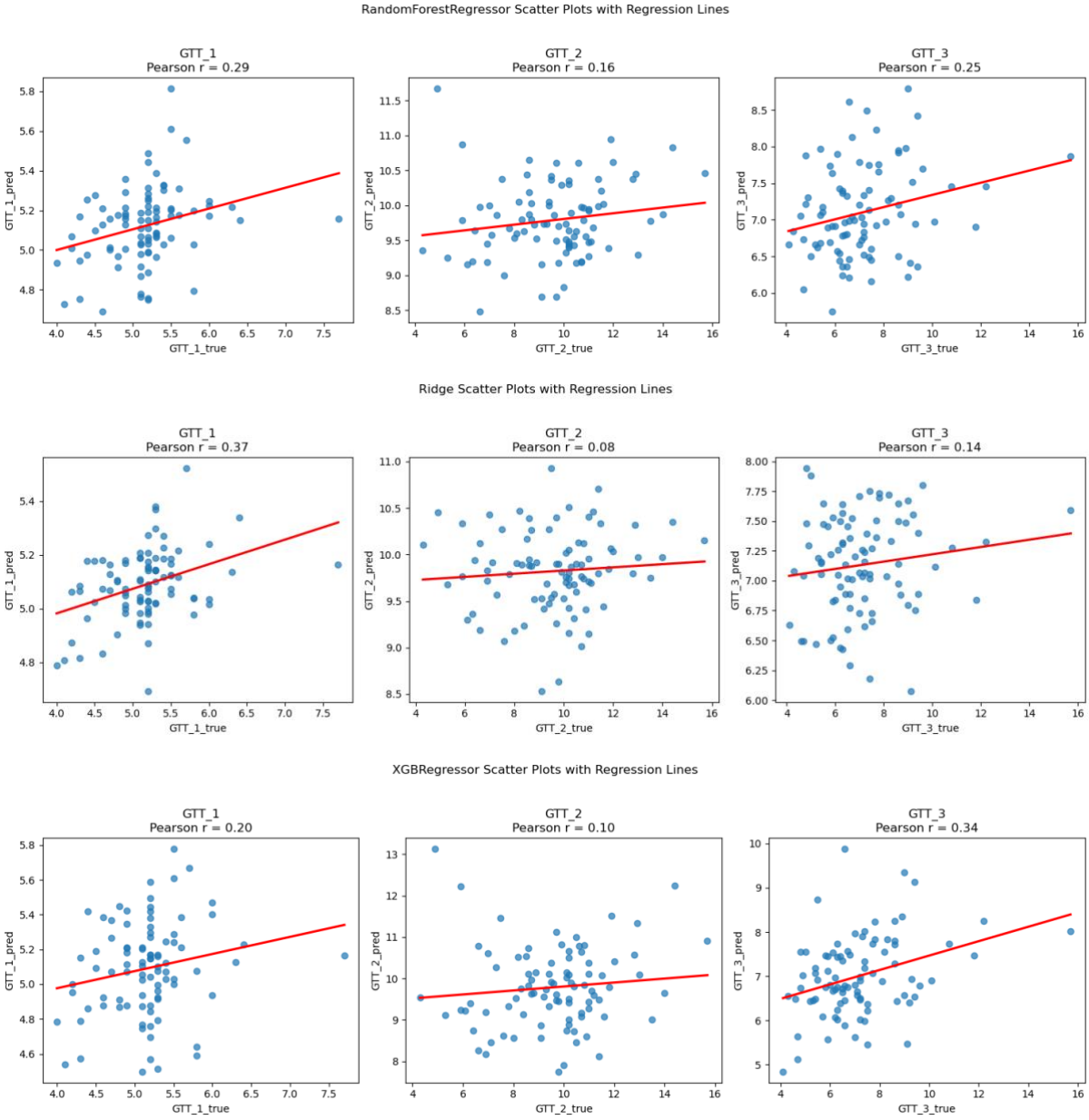
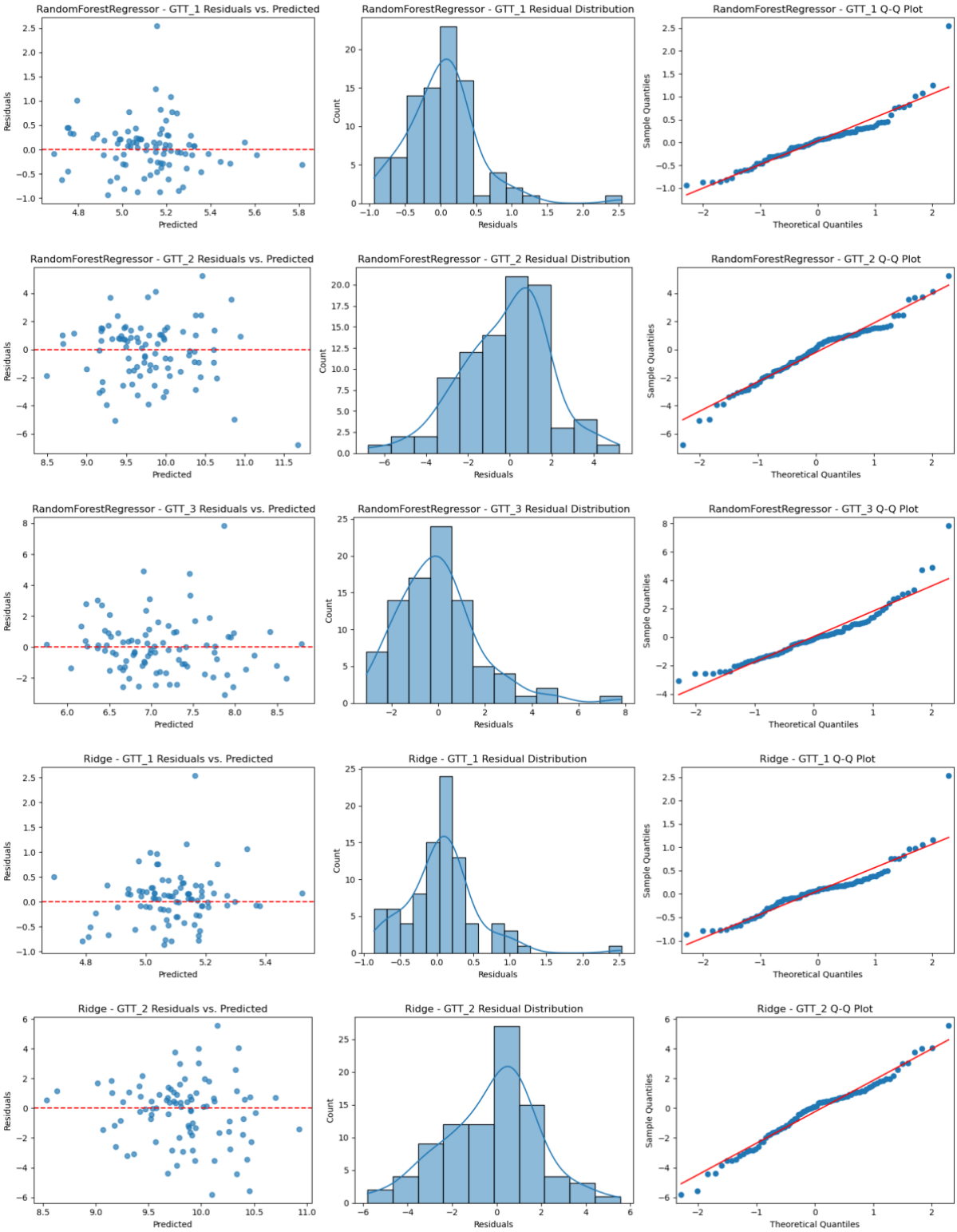


Figure 3. Scatter plots of predicted versus observed OGTT glucose values at 0 h, 1 h and 2 h

for the Random-Forest, Ridge and XGBoost regressors. Each panel displays a best-fit regression line (red) and the corresponding Pearson correlation (r).



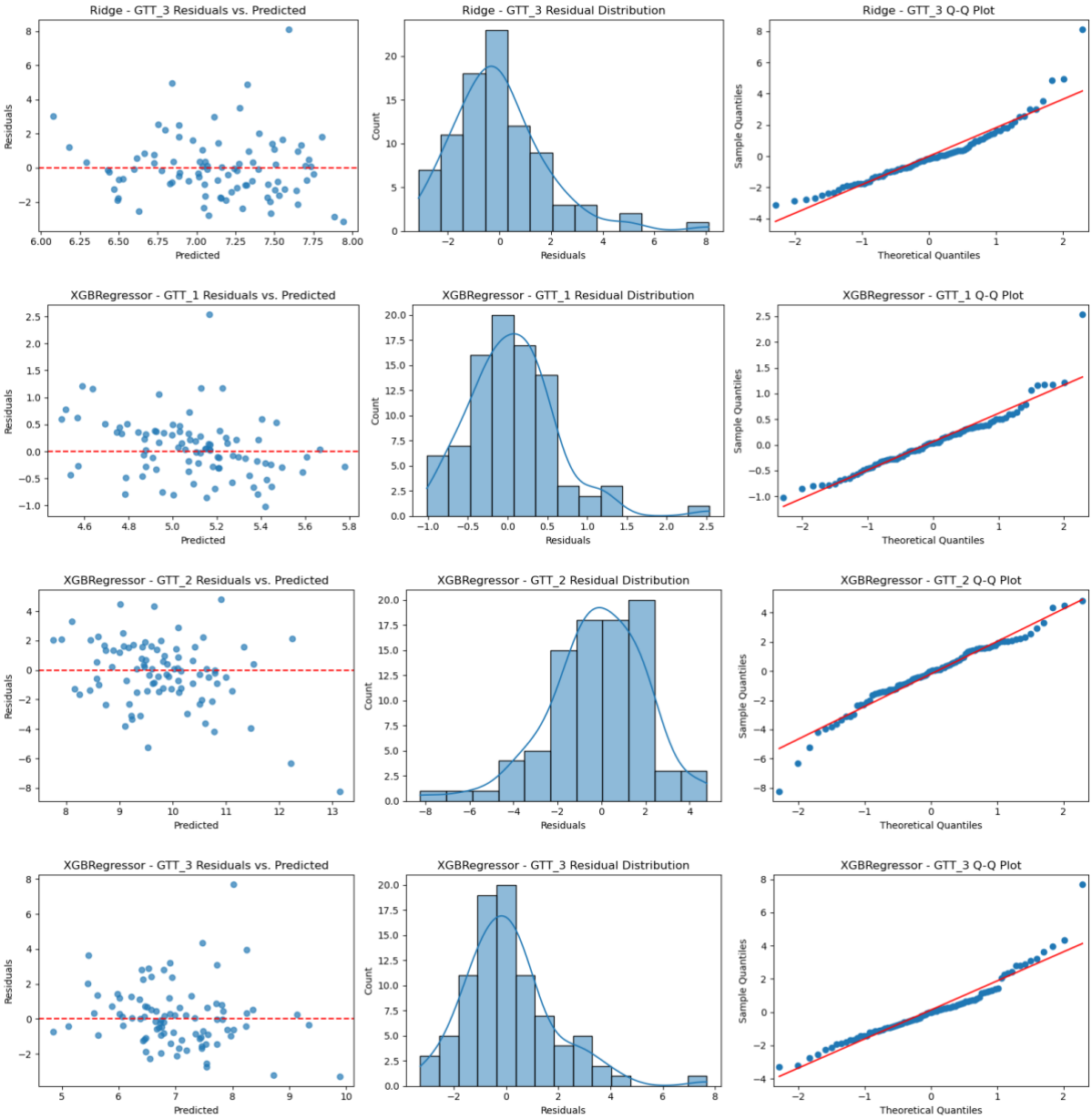


Figure 4. Residual diagnostics for Random-Forest (upper panels) and Ridge (middle panels) and XGB (lower panels) regressors in predicting OGTT glucose at 0 h, 1 h and 2 h. For each time-point, the left plot shows residuals versus predicted values, the centre plot shows the residual distribution with kernel density overlay, and the right plot presents a normal Q–Q comparison; dashed red lines mark zero residuals and the theoretical normal reference, respectively.

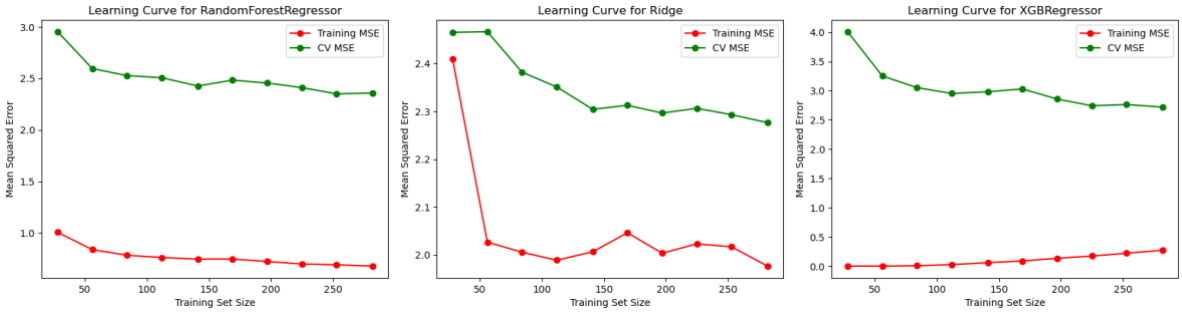
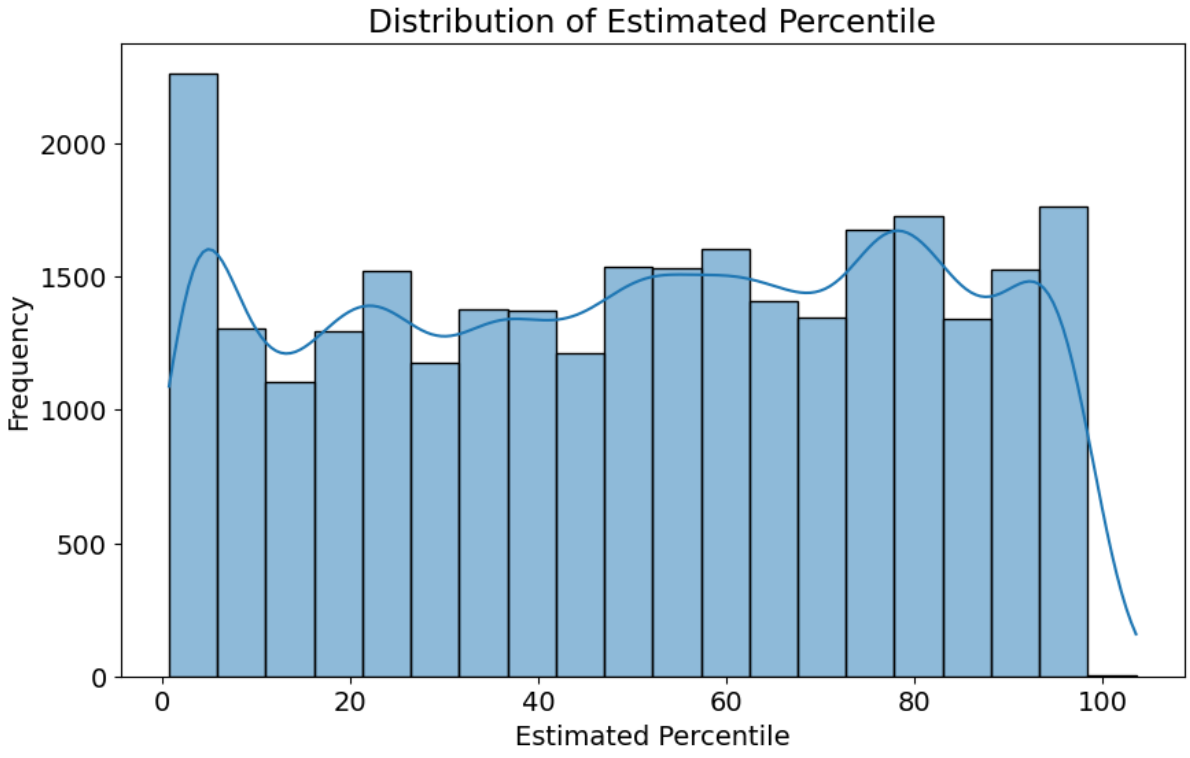
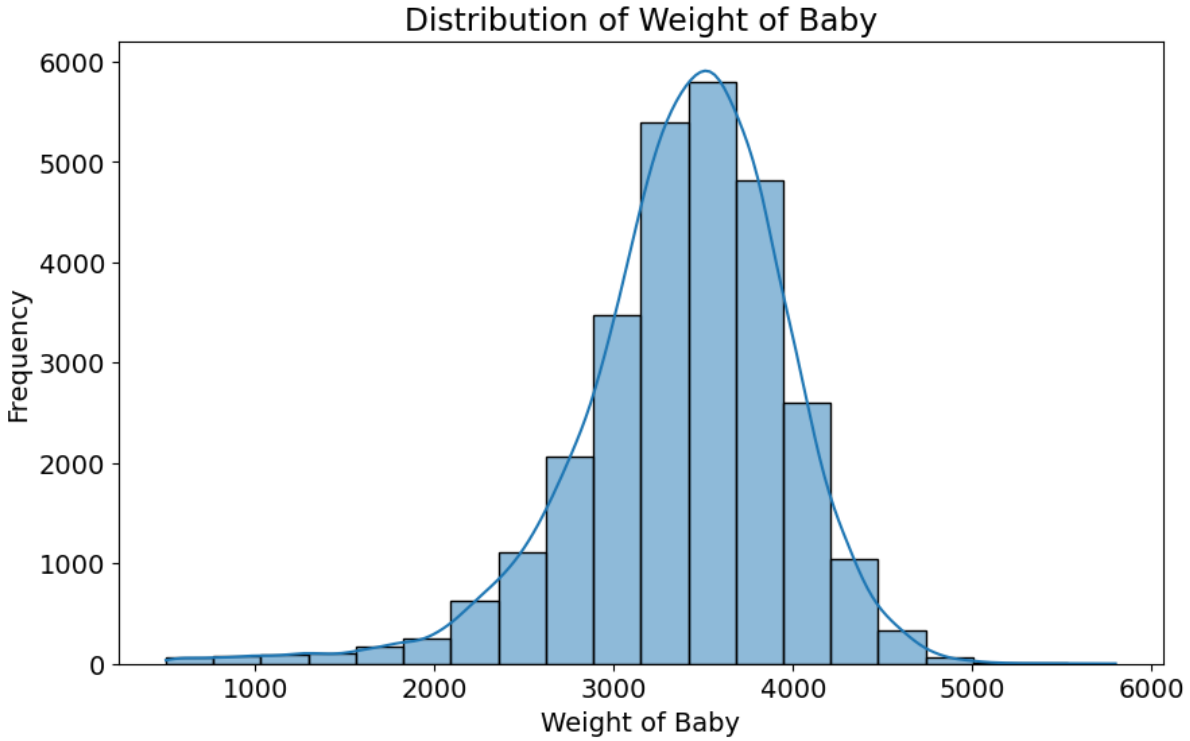
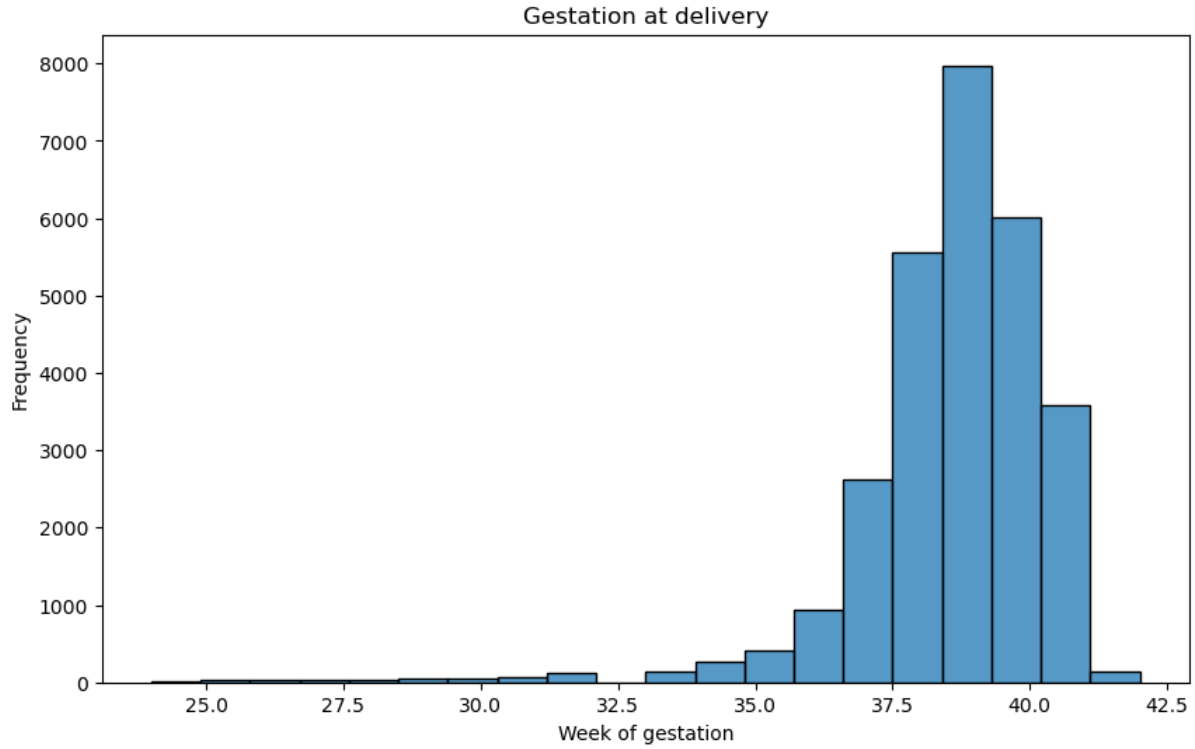


Figure 5. Learning curves for Random-Forest, Ridge and XGBoost regressors showing mean-squared-error (MSE) on the training data (red) and 5-fold cross-validation (green) as a function of training-set size; curves illustrate steady generalisation improvement with more data and the persistent performance gap between linear and tree-based models.

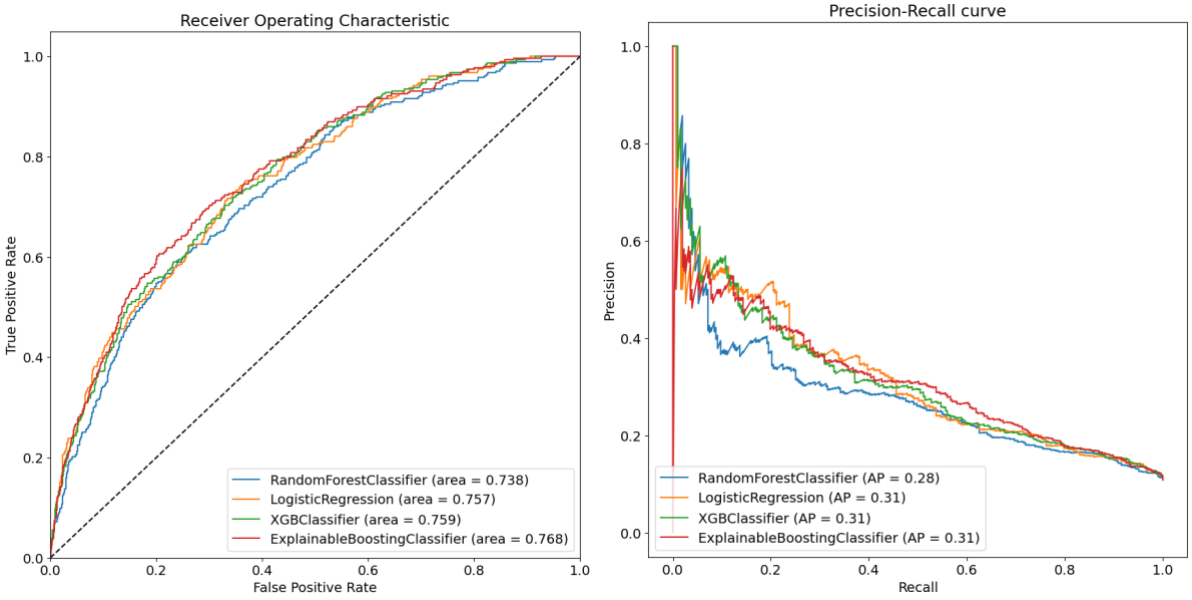
Appendix J. Birth Outcomes Preliminary Analysis



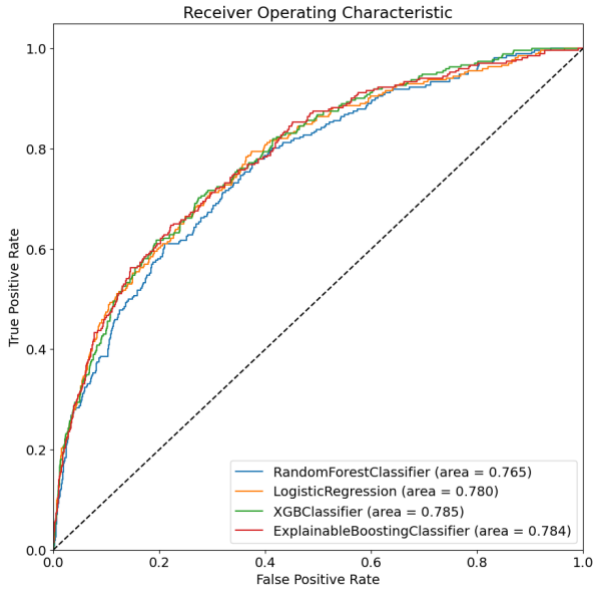


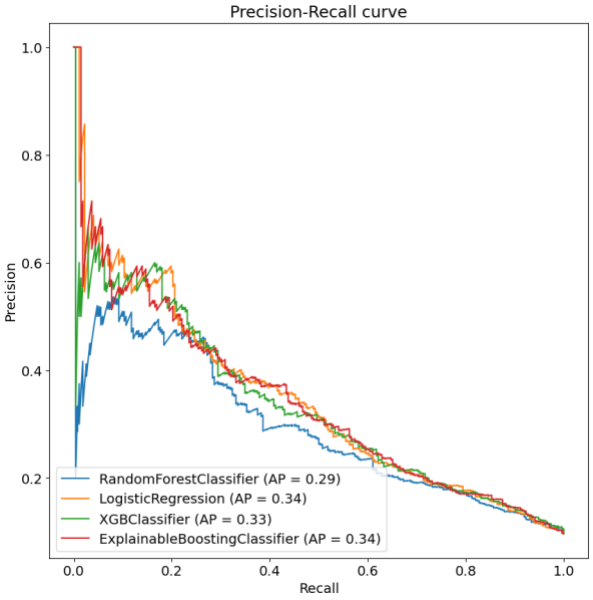


Predicting Macrosomia

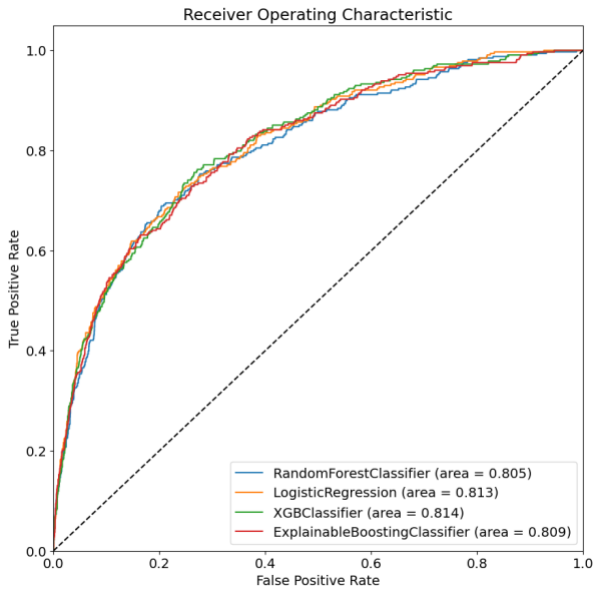


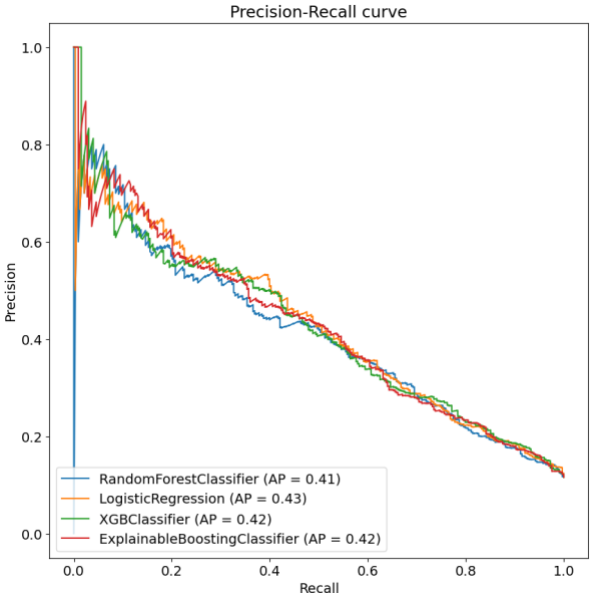
Predicting Large-for-gestational-age



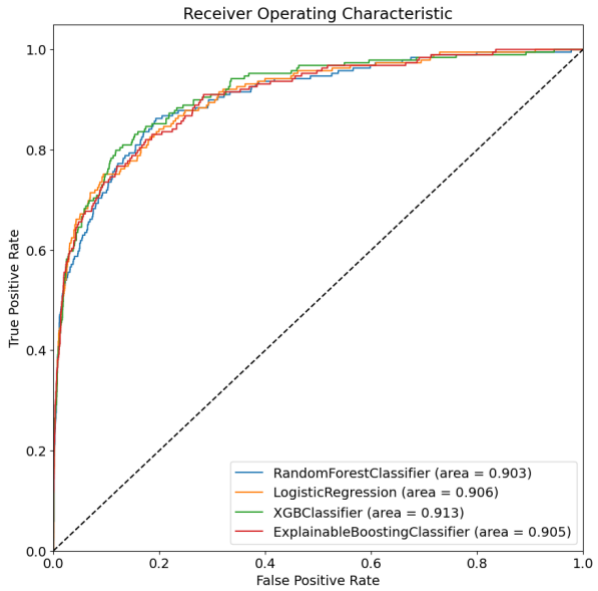


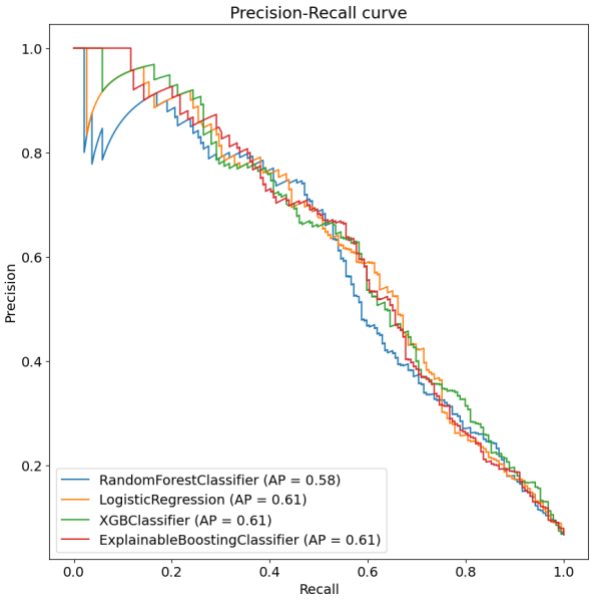
Predicting Small-for-gestational-age



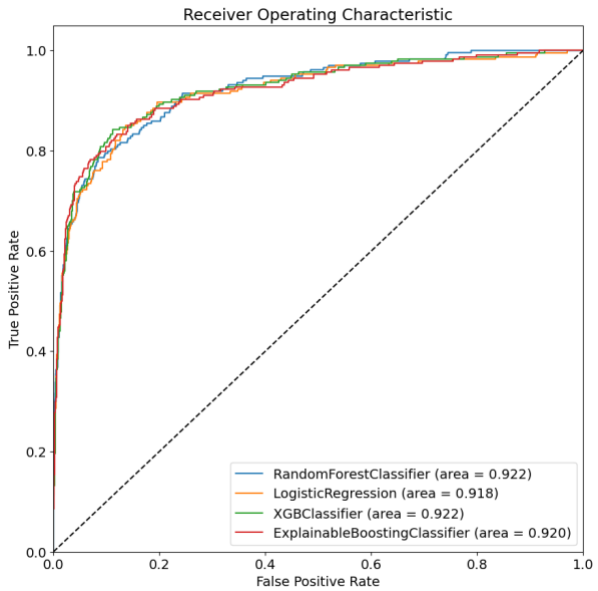


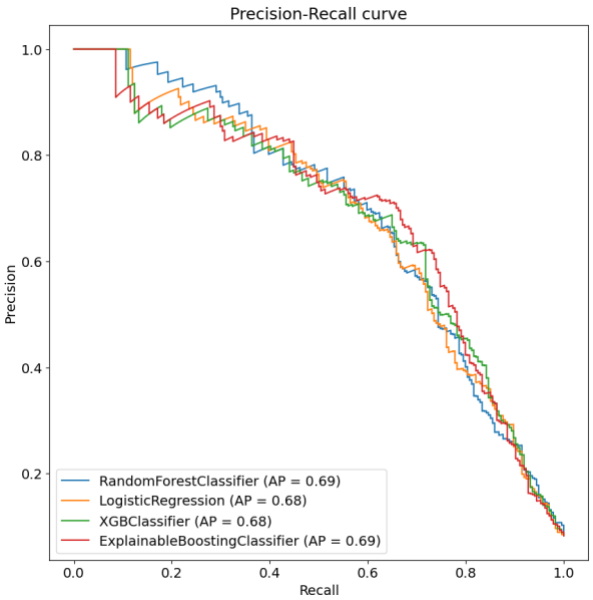
Predicting Low Birthweight





Predicting Preterm Birth





Appendix K. Coombe OGTT Referral Form

The Coombe Hospital

OGTT/FPP Referral Form

Gestational Diabetes Screening

Addressograph

Gestation: / 40

BMI _____ Parity _____ EDD _____

Reason for referral:

1. History of Gestational Diabetes during a previous pregnancy ☐
   Any woman with a history of GDM in a previous pregnancy should have a random plasma glucose done at booking visit and subsequent fasting/postprandial plasma glucose (FPP) booked.
2. First-degree relative with Type 1 or 2 Diabetes *(ie Parent, Sibling, Child)* ☐
3. Previous baby weighing 4.5kgs or above ☐
4. Body Mass Index of 30 or above ☐
5. Aged 40 years or more ☐
6. History of Polycystic Ovarian Syndrome ☐
7. Previous unexplained Intrauterine Death (IUD) ☐
8. Other:_____

Test required:

1) **FPP** (as soon as possible then again at 18 & 24 weeks) *Plus* **OGTT** (at 26-28 weeks, no later) ☐
2) **OGTT** *only* (26-28 weeks gestation, no later) ☐

Referred by: _____ Department_____ Date: _____

Author: AMcCarthy & G Sullivan Date: 23/4/2024