Prediction of the risk of adverse outcome for women with preeclampsia
PhD Thesis

Tunde Csoban

Department of Mathematics and Statistics

University of Strathclyde, Glasgow

April 2024

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Abstract

Introduction:

Preeclampsia is a leading cause of potentially life-threatening, -altering, and -ending complications during pregnancy globally. The sole method of initiating maternal recovery from preeclampsia is delivery of the placenta. Hence, to optimise maternal outcomes in preeclampsia, we need objective, time-of-disease maternal risk assessment to inform decision-making. Clinical decisions are made at different points of a woman's journey with preeclampsia, first on admission, then continuously during expectant care, and in vastly different settings with different resources. This thesis describes the development of predictive tools to address these issues: a static model for day of admission, a dynamic model for up to 2 weeks after admission, and hierarchical multistep classification tool for low resource settings with missing data.

Methods:

Using data from published studies, various static and dynamic modelling methods were tested with variable selection to fit a static and a dynamic model, each balancing model complexity and performance, resulting in the static random forest PIERS-ML model, and the binary mixed effects random forest based dynamic PIERS model. Internal validation was carried out using 25% data withheld from model development, assessing performance via area-under-the-receiver-operator characteristic (AUROC) and defining risk strata with likelihood ratios (LR- and LR+). The PIERS-ML was externally validated in an additional cohort. For the panPIERS hierarchical multistep classification tool, predictor variables were divided into variable groups, and models were fitted to each variable group combination using logistic regression, LASSO and random forest methods. Models were ranked on their ability to classify into the very low or very high risk strata. Generative adversarial imputation nets were used to impute missing values within variable groups when making predictions, and the best model with all variables available or imputed was used.

Results:

Of 8843 participants in the static data, 590 (6·7%) developed the composite adverse maternal outcome within 2 days. PIERS-ML was accurate (AUROC 0·80 [95% CI 0·76–

0·84]) and categorised women into very low risk (8, 0% outcome), low risk (321, 2% outcome), moderate risk (979, 5% outcome), high risk (87, 26% outcome), and very high risk (11, 91% outcome). The external validation in a cohort of 2901 women confirmed the model's utility. The dynamic PIERS model exhibited robust performance up to 7 days (daily AUROC>0.70) and acceptable accuracy up to 14 days (daily AUROC>0.60), effectively stratifying for the first 7 days into very low (4-10% of patients, 0-1% outcome rate), low (35-40% of patients, 0.5-3% outcome), moderate (46-52% of patients, 2-7% outcome), high (3-6% of patients, 20-35% outcome) and very high (0.5-1% of patients, 70-100% outcome). Of the 128 possible available variable group combinations, the panPIERS tool could rule in the outcome with a LR+≥10 in 112 combinations using a collection of 12 models, while we could rule out the outcome with LR-≤0.1 in all combinations using a collection of 15 models. Of 11472 patients, the model currently used in clinical practice could be used for only 29.3% of patients, while the panPIERS classification tool could be used for all.

Conclusion:

This thesis presents models and tools optimized for preeclampsia risk assessment, using data from admission, temporal information, and accommodating missing values. These tools provide user-friendly outputs, enhancing the identification of women at varying risks of adverse maternal outcomes within a crucial two-day window. These advancements support informed clinical decision-making, especially in settings with limited resources.

Publications associated with this research

Journal papers

1. Montgomery-Csobán T, Kavanagh K, Murray P, Robertson C, Barry SJE, Vivian Ukah U, Payne BA, Nicolaides KH, Syngelaki A, Ionescu O, Akolekar R, Hutcheon JA, Magee LA, von Dadelszen P; PIERS Consortium. Machine learning-enabled maternal risk assessment for women with preeclampsia (the PIERS-ML model): a modelling study. Lancet Digit Health. 2024 Apr;6(4):e238-e250. doi: 10.1016/S2589-7500(23)00267-4. PMID: 38519152.
2. Yang G, Montgomery-Csobán T, Ganzevoort W, Gordijn SJ, Kavanagh K, Murray P, Magee LA, Groen H, von Dadelszen P; Performance drift detected from consecutive prediction of the adverse maternal outcomes of preeclampsia using the PIERS-ML model: A multi-country prospective observational study. BJOG. (Submitted, Appendix B)

Conference presentations

3. Comparison of modelling methods for prediction of adverse outcomes of preeclampsia at the 44th Research Students' Conference in Probability and Statistics (2021, July 27-29), Online (hosted by Lancaster University).
4. Comparison of modelling methods for prediction of adverse outcomes of preeclampsia at the RSS International Conference 2021 (2021, September 6-9), Manchester, UK.
5. Time-of-disease risk prediction of adverse maternal outcomes of preeclampsia using artificial intelligence at the ISSHP 2022 conference (2022, August 28-31), Montpellier, France.
6. External validation of the PIERS-AI model in a UK cohort of women with preeclampsia at the ISSHP 2023 conference (2023, September 24-27), Bengaluru, India.
7. panPIERS: A tool for risk prediction of adverse maternal outcomes for patients with preeclampsia using incomplete data at the ISSHP 2023 conference (2023, September 24-27), Bengaluru, India.

Acknowledgements

Working on this PhD has been the biggest challenge and the most rewarding experience of my life. The experience of struggles and victories are shared by many, the added challenges of working through the changing world of a global pandemic are shared by much less. Yet, there are some people who have made this shared experience special and uniquely mine.

First, I would like to thank my internal supervisors, Dr. Kim Kavanagh and Dr Paul Murray, who have gone above and beyond in their support. They have not only guided me with their expertise and helped me grow into an independent researcher, but they also accommodated my changing physical and mental health over the past 4 years, and believed in me even when I didn't. I would like to also thank Dr Sarah Barry for her support and encouragement and Prof. Chris Robertson for his continued support despite his busy schedule. Their involvement was a big help to both my progress and confidence. I would like to also acknowledge the funding I received from the University of Strathclyde as part of the STRADDLE DCT which allowed this project to exist in the first place.

I would like to thank my external supervisors, Prof. Laura Magee and Prof. Peter von Dadelszen (Mr Magee). Their support and kindness made a huge difference to my PhD experience, and their infectious enthusiasm made me so excited to be able to contribute to such an important research topic.

I feel very fortunate that I got to work with such fantastic internal and external supervisors, and I feel honoured to call them friends.

Next, I would also like to thank my friends and comrades-in-arms of the virtual office, Natalia Dygas, Dr Matthew Baister, Michael Brown, Connor Watret and Nathan Burns, for keeping my social life alive, and Dr Kate Pyper for making me fall in love with tidyverse and ggplot. It is thanks to them that the periods of isolation during the pandemic did not feel nearly as isolating.

Finally, I would like to thank my husband and my parents and my family, who have supported me every single day of the journey, celebrated the smallest achievement with

me, encouraged me through every challenge and made sure I didn't forget to eat or sleep when I got too focused. It is thanks to them that the end never felt too far to reach.

Table of contents

Table of Contents

List of figures

List of tables

Abbreviations

AI	Artificial Intelligence
AKI	Acute Kidney Injury
ALT	Alanine Transaminase
ANN	Artificial Neural Network
APTT	Activated Partial Thromboplastin Time
AST	Aspartate Transaminase
AUC	Area Under The Curve
AUPRC	Area Under The Precision-Recall Curve
AUROC	Area Under The Roc Curve
BiMM	Binary Mixed Model Random Forest
BJOG	British Journal Of Gynaecology
BMA	Bayesian Model Averaging
CI	Confidence Interval
EDD	Estimated Delivery Date
ELU	Exponential Linear Unit
FGR	Fetal Growth Restriction
FMF	Fetal Medicine Foundation
GAIN	Generative Adversarial Imputation Network
GAN	Generative Adversarial Network
GDP	Gross Domestic Product
GLMM	Generalised Linear Mixed Model
HELLP	Hemolysis, Elevated Liver Enzymes And Low Platelets
HIC	High-Income Countries
INR	International Normalised Ratio
ISSHP	International Society For The Study Of Hypertension In Pregnancy
JMBayes	Bayesian Joint Model
LASSO	Least Absolute Selection And Shrinkage Operator
LB	Live Births
LDH	Lactate Dehydrogenase
LMIC	Low- And Middle-Income Countries

LOCF	Last Observation Carried Forward
LR	Likelihood Ratio
LSTM	Long Short-Term Memory Artificial Neural Network
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MAR	Missing At Random
MCAR	Missing Completely At Random
ML	Machine Learning
MMR	Maternal Mortality Ratio
MNAR	Missing Not At Random
MPV	Mean Platelet Volume
MSE	Mean Squared Error
NHS	National Health Service
NICE	National Institute Of Health And Care Excellence
NICU	Neonatal Intensive Care Unit
NPV	Negative Predictive Value
PIERS	Preeclampsia Integrated Estimate Of Risk
PIH	Pregnancy Induced Hypertension
PPV	Positive Predictive Value
PREP	Prediction Models For Risk Of Complications In Early-Onset Preeclampsia
RFE	Recursive Feature Extraction
RNN	Recurrent Artificial Neural Network
ROC	Receiver Operating Characteristic
SMOTE	Systematic Minority Oversampling Technique
UK	United Kingdom
XGBoost	Extreme Gradient Boosting

Chapter 1: Introduction and literature review

1.1 Background

Pregnancy is a time of uncertainty in a woman's life, with many women experiencing a mixture of excitement, anticipation, and the fear of potential complications. Maternal health is an area of great concern. As stated by the World Health Organisation, maternal mortality is unacceptably high. While there are numerous challenges to a safe pregnancy, preeclampsia is named as one of the leading causes of maternal mortality and life-altering morbidity globally, highlighting the need for effective risk prediction of adverse maternal outcomes for women diagnosed with preeclampsia. For the past 20-30 years, predicting the risk of adverse maternal outcomes of preeclampsia has been a focus of research, and recent advancements in predictive modelling, coupled with increased awareness of machine learning and artificial intelligence in the general population lead to the question: can any of these methods help us stratify women better than existing models, and identify those who need altered care the most?

The research presented in this thesis aims to improve maternal outcomes of preeclampsia by providing clinicians with enhanced tools for individual risk assessment.

1.2 Hypertension and preeclampsia

1.2.1 Definition, diagnosis, and outcomes

Hypertension, or high blood pressure, is the most common medical complication of pregnancies (1). It can be chronic or developed during pregnancy and can lead to preeclampsia and eclampsia. Hypertension developed after 20 weeks gestational age during pregnancy is called Gestational or Pregnancy-induced Hypertension, and generally returns to normal after delivery (2).

Preeclampsia is a state of exaggerated systemic inflammation (3) that complicates 2-8% of all pregnancies (4). It can lead to eclampsia, a severe, life-threatening complication of pregnancy manifested as seizures or coma (5). Eclampsia is a major cause of maternal and perinatal morbidity and death. Causes of eclampsia are not yet known (5). Other complications associated with preeclampsia include fetal growth restriction, stillbirth, preterm delivery, neonatal morbidity (6), stroke and damage to the hepatic and renal organs (7). It also

increases risk of oligohydramnios, placental abruption, fetal distress, and fetal death in utero (8).

A composite adverse maternal outcome was derived by a Delphi consensus on the basis of including outcomes clinicians wished to avoid, rather than waited to occur before altering care. The composite outcome included one or more of maternal mortality, and severe central nervous system, cardiorespiratory, hepatic, renal or haematological morbidity (Table 1.1).

Preeclampsia is diagnosed after 20 weeks gestational age in pregnancy, traditionally the diagnosis criteria is newly onset hypertension (high blood pressure) and proteinuria, however the International Society for the Study of Hypertension in Pregnancy (ISSHP) published a new criteria in 2021 (9), which defines preeclampsia as hypertension accompanied by one or more symptoms of maternal end-organ damage (including neurological complications, pulmonary oedema, haematological complications, AKI, liver involvement or uteroplacental dysfunction). Preeclampsia can also be superimposed on chronic hypertension patients if any of the above symptoms develop after 20 weeks gestational age. While the exact cause of preeclampsia is not yet fully understood, early-onset preeclampsia (diagnosed before 34 weeks gestation) is thought to be caused by shallow invasion of the maternal spiral arteries by the trophoblasts (10, 11) while late-onset preeclampsia (diagnosed after 34 weeks gestation) is thought to be a result of maternal predisposition to arterial disease causing the hyper-inflammatory state (11).

Table 1.1: Composite adverse maternal outcomes of preeclampsia

Outcome	**Definition**
Mortality	Maternal death occurring within six weeks of pregnancy or if later, attributable to complications of preeclampsia
Hepatic dysfunction	International normalised ratio (INR) >1.2 in the absence if disseminated intravascular coagulation (DIC) or treatment of warfarin (DIC is defined as having both: abnormal bleeding and consumptive coagulopathy [i.e., low platelets, abnormal peripheral blood film, or one or more of the following: increased INR, increased prothrombin time (PTT), low fibrinogen, of increased fibrin degradation products that are outside normal non-pregnancy ranges])

Outcome	Definition
Hepatic hematoma or rupture	Blood collection under the hepatic capsule as confirmed by ultrasound or laparotomy
Glasgow coma score (GCS) <13	Based on GCS scoring system: Teasdale G, Jennet B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; **2**:81-83
Stroke	Acute neurological event with deficits lasting longer than 48 hours
Cortical blindness	Loss of visual acuity in the presence of intact papillary response to light
Reversible Ischaemic Neurologic Deficit (RIND)	Cerebral ischaemia lasting longer than 24 hrs but less than 48 hours revealed through clinical examination
Retinal detachment	Separation of the inner layers of the retina from the underlying retinal pigment epithelium (RPE, choroid) and is diagnosed by ophthalmological exam
Acute renal insufficiency	For women with no underlying renal disease, defined as serum creatinine >150 µM
Acute renal failure	For women with an underlying history of renal disease, defined as serum creatinine >200 µM
Dialysis	Including haemodialysis and peritoneal dialysis
Postpartum haemorrhage (PPH) requiring transfusion or hysterectomy	Occurrence of PPH that required transfusion or hysterectomy
Placental abruption	Any occurrence of abruption diagnosed clinically or based on placental pathology report
Platelet count < 50,000 x 10^9/L without blood transfusion	Measurement of platelet count recorded as less than 50,000 x 10^9/L without patient receiving a blood transfusion
Transfusion of blood products	Includes transfusion of any units of blood products: fresh frozen plasma (FFP), platelets, red blood cells (RBCs), cryoprecipitate (cryo) or whole blood
Positive inotropic support	The use of vasopressors to maintain a systolic blood pressure >90 mmHg or mean arterial pressure >70 mmHg
Myocardial ischaemia/infarction	Electrocardiogram (ECG) changes (ST segment elevation or depression) without enzyme changes AND/OR any one of the following: 1) Development of new pathologic Q waves on serial ECGs. The patient may or may not remember previous symptoms. Biochemical markers of myocardial necrosis may have normalised, depending on the length of time that has passed since the infarct developed. 2) Pathological findings of an acute, healed or healing MI 3) Typical rise and gradual fall (troponin) or more rapid rise and fall

Outcome	Definition
	(CK-MB) of biochemical markers of myocardial necrosis with at least one of the following: a) ischaemic symptoms; b) development of pathologic Q waves on the ECG; c) ECG changes indicative of ischaemia (ST segment elevation or depression); or d) coronary artery intervention (e.g., coronary angioplasty)
Eclampsia	Any episode of seizure antepartum, intrapartum or before postpartum discharge as follow-up beyond discharge is not possible
Require >50% oxygen for greater than one hour	Oxygen given at greater than 50% concentration based on local criteria for longer than 1 hour
Intubation other than for Caesarean section	Intubation by endotracheal tube
Severe breathing difficulty	Suspected pulmonary oedema where x-ray confirmation unavailable may be diagnosed by presence of chest pain or dyspnoea, crackles in the lungs and SaO_2 <90%
Pulmonary oedema	Clinical diagnosis with x-ray confirmation or requirement of diuretic treatment and SaO_2 <95%

1.2.2 Management of preeclampsia

The main components of management of preeclampsia are: (i) antihypertensive drugs for the management of high blood pressure, (ii) anticonvulsants for the prevention and treatment of seizures, and (iii) timing of delivery.

Timing of delivery generally depends on gestational age and severity of preeclampsia. Two reviews on the management of preeclampsia advised that interrupting the pregnancy might be appropriate if preeclampsia is diagnosed before 24 weeks gestational age. One of these reviews (8) based this recommendation on only one study (12) that had a small sample size of 31, while the other review (13) referenced the same study and one with a sample size of 39 (14). Both reviews recommended delaying the delivery if possible between 24 and 34 weeks for mild preeclampsia, but (13) suggested delivery for severe.

The World Health Organisation guidelines for managing preeclampsia only suggested induction of labour before 24 weeks gestation if the mother has severe preeclampsia and a fetus is unlikely to achieve viability within 1 or 2 weeks. Expectant management is recommended for mild preeclampsia and for severe preeclampsia with a viable fetus (15).

Other reviews made recommendations for preeclampsia diagnosed up to 34 weeks gestation and did not differentiate between before 24 weeks and 24-34. Most suggested expectant management (16-19) as long as blood pressure and maternal symptoms can be controlled, while one review (17) did not consider delivery before 34 weeks gestation at all. As part of expectant management, keeping systolic and diastolic blood pressure below a target 150mmHg and 80-100mHg respectively was mostly recommended (17, 18), however the NICE (National Institute of Health and Care Excellence) guidelines suggested a target of 135/85mmHg or less (19). Recommendations for when to start antihypertensive treatment was mixed: one review (7) suggested to begin treatment once blood pressure crosses the target threshold, two (16, 18) recommended treatment for blood pressure over 160/110mmHg only while the NICE guidelines (19) suggested antihypertensive treatment for all women with blood pressure over 160/110mmHg and for women with blood pressure consistently above 140/90mmHg. Administration of magnesium sulphate as an anticonvulsant was recommended for women with "severe" preeclampsia (16, 18) or "mild" preeclampsia with maternal symptoms (18), however it should be noted that in 2014 the International Society for the Study of Hypertension in Pregnancy found the classification of preeclampsia as "mild" or "severe" misleading, and updated their recommendation to no longer support clinical distinction (20, 21). One paper (18) recommended corticosteroids as part of management for severe preeclampsia while another added bedrest and fluid management (16).

The review of the literature identified the time between 34- and 37-weeks' gestation the period with the most uncertainty regarding the recommendations on expectant management and timing of birth. One review (17) could not make a recommendation for optimal time of delivery during this period. Two reviews (16, 18) recommended immediate delivery after 34 weeks gestation, as soon as the mother is stable. Most papers (8, 13, 15, 19, 21) recommended expectant management unless there are indications that immediate delivery is required. The indications mentioned included repeated episodes of severe hypertension despite maintenance treatment with three classes of antihypertensives (19, 21), maternal blood oxygen levels under 90% (19), progressive thrombocytopenia (21), progressively abnormal renal or liver enzyme tests (19, 21),"progressive deterioration in haemolysis, or platelet count"(19), "placental abruption, reversed end-diastolic flow in the umbilical artery

doppler velocimetry, a non-reassuring cardiotocograph, or stillbirth"(19), "pulmonary oedema; abnormal neurological features such as severe in-tractable headache, repeated visual scotomata, or convulsions; or non-reassuring fetal status"(21). One paper suggested induction of delivery for all cases of severe preeclampsia (13).

The general recommendation after 37 weeks gestation was delivery (8, 13, 15-19, 21).

In general, most papers suggested magnesium sulphate to be used as an anticonvulsant, while the most commonly used antihypertensives were Labetalol, Nifedipine and Hydralazine (16, 17, 19, 21-23). Other antihypertensives mentioned were Methyldopa (19), Captopril and Clonidine (23). One review (21) suggested not differentiating mild and severe preeclampsia. Other general recommendations included fluid management while an IV line is in place and administration of corticosteroids when delivery is anticipated in the next 7 days (23).

The recommendations on the administration and dosage of magnesium sulphate were not consistent between the review papers. One (16) suggested magnesium sulphate to be administered to patients with delayed delivery for 48 hours, in the form of a loading dose followed by continuous infusion. The review recommended intravenous treatment of 4-6g of $MgSO_4$ followed by 2-3g/h and intramuscular treatment of 10g followed by either 5g or 2.5g every 4 hours for 24 hours (16). Other papers recommended intravenous treatment of 4g $MgSO_4$ loading dose followed by an infusion of 1 g/hour for 24 hours (19, 23). If the mother had eclamptic fits a further 2-4g $MgSO_4$ was recommended to be given over 5 to 15 minutes (19, 23) and the infusions to be continued for 24h after the last fit (19).

Magnesium sulphate was recommended to be used during labour (16, 21-23), however it was not clear how long to continue magnesium sulphate for. One review (21) pointed out that while "one recent study in Latin America found that women who had received at least 8 g of MgSO4 before delivery had no additional benefit of continuing the magnesium for a further 24 h. post-partum" (21), eclampsia can occur after birth, which can be managed/controlled by continued treatment with magnesium sulphate. Others (16, 22, 23) recommended continuing treatment for 24 hours postpartum.

Replacing magnesium sulphate with other anticonvulsants as alternatives was not recommended (19), and one paper (18) advised against using magnesium sulphate for women with non-severe hypertension and no maternal symptoms.

As adverse maternal outcomes can develop over time, continuous monitoring of the severity of the condition is crucial. Women with preeclampsia were recommended to be monitored at each antenatal visit or as inpatients, by recoding any change in symptoms of headache or visual disturbances, right upper quadrant or epigastric pain, nausea or vomiting, and chest pain or dyspnea (shortness of breath), as well as by taking repeated measurements of blood pressure and blood oxygen levels, blood tests and urine tests (19). Routinely performed blood tests include: (i) kidney function tests including serum creatinine and uric acid, which are both waste products (from use of muscles, and from various foods and cells respectively) that would get filtered out from the blood by a healthy kidney, (ii) liver function tests including alanine transaminase (ALT) and aspartate transaminase (AST), which are both enzymes found in the liver that are released into the blood upon damage (iii) full blood count to monitor changes in the blood including red and white blood cell and platelet counts and hematocrit, the proportion of red blood cells in the blood, (iv) blood clotting time (activated partial thromboplastin time), and (v) markers of HELLP syndrome, including lactate dehydrogenase (LDH), an enzyme found in tissues, which indicates tissue or cell damage when found in the blood in high levels, and total bilirubin, which is produced during the breakdown of red blood cells (24). Urinary protein and urinary proterin:creatinine ratio are also measured to assess kidney function (24).

As many systems are affected by preeclampsia, monitoring of different signs, symptoms and laboratory tests are needed to cover the monitoring of all affected areas. The central nervous system is monitored by symptoms of headache or visual disturbances. Cardiorespiratory function is monitored by symptoms of chest pain or dyspnea, blood pressure measurements, and blood oxygen saturation. Symptoms of epigastric pain, and liver function tests are used for monitoring hepatic function, while dipstick proteinuria, serum creatinine and albumin are used for renal monitoring, and total leucocyte count, platelet count and mean platelet volume are useful for haematological monitoring.

1.3 Risk prediction of preeclampsia

Modelling of adverse maternal outcomes of preeclampsia has been attempted previously. Existing methods for prediction of risk of adverse maternal and fetal outcomes for women with preeclampsia can be split into two major categories: univariable and multivariable

prediction models. We have identified 13 studies reporting multivariable models or adverse maternal or maternal and fetal outcomes.

Some of the most pressing concerns for the fetus are the risk of stillbirth, neonatal mortality, fetal growth restriction (FGR) and length of stay in the neonatal intensive care unit (NICU). FGR monitoring starts during pregnancy and can be assessed by measurements on ultrasounds (25). Moreover, risk calculators for both FGR and stillbirth are available from The Fetal Medicine Foundation (26). The biggest risk factors that determine the length of stay in the NICU are birthweight and gestational age at birth (27), and while the risk of neonatal mortality is the most complex issue of all, birthweight and gestational age at birth are both significant risk factors (28). Therefore, *not* inducing the mother for as long as possible is likely lead to better outcomes for the fetus. Maternal risk, however, increases over time, the list of most concerning maternal outcomes is more complex than the list of fetal outcomes, and maternal risk is also more difficult to assess. We chose to focus on adverse maternal outcomes, as an accurate assessment of the maternal risk could be used best to inform timing of birth by balancing the predicted individual risk with the known risks to the fetus.

1.3.1 Multivariable prediction of risk

A review of the literature was conducted to identify multivariable models for the prediction of risk of adverse maternal outcomes of preeclampsia. In January 2018, a systematic review was published on the topic (7). The studies identified in this review included univariable models and multivariable models using logistic or Cox regression. The multivariable models identified for inclusion in the review were the fullPIERS model (3), the miniPIERS model (29), the extended miniPIERS model with SpO2 (30), a combined cardiorespiratory symptom model by Millman et al. for the prediction of the PIERS composite outcome (31), and a model by Chan et al. (32). In addition to the information available on the models in the review paper, the original publications were also accessed where possible. Only minimal information was available on the model by Millman et al. as the article could not be accessed.

To identify models published since the systematic review, on Jan 19, 2024, I searched PubMed using the search terms (“outcome”) and (“preeclampsia” or “preeclampsia”), and (“model” or “risk” or “algorithm”) in the title published since the 2018 review paper (from Jan 1, 2018, to Jan 19, 2024) and found a further seven (33-39) studies presenting models of adverse maternal outcomes of preeclampsia from single country studies in Zimbabwe, Germany, and

China. Full text of Liao et al. was not available. Additionally, a further study was found from 2017 (40) reporting the findings of the PREP (Prediction models for Risks of complications in Early-onset Pre–eclampsia) trial, including the two models developed: the PREP-L logistic model and the PREP-S survival model. This resulted in a total of 13 studies identified for review (Table 1.2). The performance of these models is reported below using area under the receiver-operator curve (AUROC), and positive and negative likelihood ratios. These concepts and their interpretation are introduced and discussed in Chapter 2, Section 2.2.9.1.

Of the 13 studies, four (PREP, Schmidt et al., Zheng et al., and Wang et al.) used multiple modelling methods to create models and compare their performances, while the remaining nine developed only one model each, resulting in 22 models in total. About half of the models (12 of 22) used logistic regression to model the adverse outcome (fullPIERS, miniPIERS, the extended miniPIERS with SpO2, Millman et al., Chan et al., the PREP-L model, Sun et al., Ngwenya et al., Sroka et al., Liao et al., and one each of the models from Zheng et al. and Wang et al.) while the remaining models were split between one survival model (PREP-S) and nine machine learning models (two models by Schmidt et al., six by Zheng et al., and one by Wang et al.). The number of participants in the studies ranged in size from 319 to 2532 with a median 1254, and outcome rate from 5% to 66.9% patients experiencing an outcome at any point, with a mean outcome rate of 27.94% and median of 25.8%. The definition of an adverse outcome also varied, while the three PIERS studies, Millman et al., and Ngwenya et al. all used the PIERS composite maternal outcome, and the PREP study used a version of the PIERS composite outcome with delivery before 34 weeks added, the remaining studies defined adverse outcome differently, and four studies modelled maternal and fetal outcomes together. As the risk of adverse maternal outcomes of preeclampsia increases over time, while the risk on the fetus decreases over time, modelling maternal and fetal outcomes separately would be more useful to aid in the clinical decision-making. The highest outcome rate was observed in the PREP study (66.9%), however it should be noted that the vast majority of outcomes (61.3%) were delivery before 34 weeks, which would only be informative for early-preterm preeclampsia, which represents the minority of diagnoses.

Of the logistic regression models, four models (miniPIERS with SpO2, Millman et al., Chan et al., and Sroka et al.) were not validated either internally or externally, meaning while model performances were reported, it is impossible to tell how the model would perform on unseen

data. Liao et al., Zheng et al. and Wang et al. were validated internally on a dataset held out from model development, and all had a high AUROC value (0.804, 0.958 and 0.832 respectively). The positive likelihood ratio calculated from true positive rate and false positive rate reported by Liao et al. was only 1.79, however, and we could not determine the components of the composite outcome or calculate the negative likelihood ratio due to no full text available in English. The positive and negative likelihood ratios of the Zheng et al. model were excellent (63.92 and 0.17, respectively) and the model had one of the highest AUROC values (0.958, as mentioned above), however the study used a small sample size of 733, which was further divided into development and internal validation. The Wang et al. logistic regression model had poor likelihood ratios (LR+ 2.35, LR- 0.25), and the model was not only developed on the smallest dataset of the found studies of 319 sample size, but also modelled a composite maternal and fetal outcome as mentioned above. Furthermore, none of these models were externally validated.

The fullPIERS, miniPIERS, PREP-L models, and the models presented by Sun et al. and Ngwenya et al. were all validated externally.

The fullPIERS model was first internally validated using bootstrapping, and had a similarly high AUROC value as the previously mentioned models (0.88, 95% CI 0.84–0.92), it was developed in one of the largest datasets at 2023 participants, and had a good positive likelihood ratio, although its negative likelihood ratio was almost 0.3 (LR+ 5.76, LR- 0.282). It was then externally validated on three separate occasions: the primary external validation (41) showed that the model maintained good performance in an entirely unseen dataset (AUROC 0.81, 95% CI 0.76-0.87) with an excellent positive likelihood ratio but poor negative likelihood ratio (LR+ 17.03, LR- 0.65); using the miniPIERS dataset (42), where its performance dropped compared to internal validation (AUROC 0.77, 95% CI 0.72-0.82, LR+ 2.28, LR- 0.33), however the dataset contained a high portion of missing values for the model variables; and finally, in a dataset containing data from the British Columbia Women's Hospital, the Dutch PETRA study and the PREP study (11), showing similar performance to the primary external validation (AUROC 0.8, 95% CI 0.75-0.86, LR+ 10.57, LR- 0.45).

The miniPIERS model was also similarly developed on a large dataset of 2081 patients and internally validated first using bootstrapping, followed by external validation. The results of the internal validation showed that the model had good discriminative performance (AUROC

0.768, 95% CI 0.735–0.801), with a good positive likelihood ratio (LR+ 5.11) but poor negative likelihood ratio (LR- 0.67). On external validation on the fullPIERS data, the model performance was very similar to that in the internal validation (AUROC 0.713, 95% CI 0.658–0.768). (29)

The PREP-L model had good internal discrimination (AUROC 0.82, 95% CI 0.80-0.84) after adjusting for optimism in the original dataset. When applied to data from women with suspected preeclampsia, the model did not perform as well (AUROC 0.68, 95% CI 0.58-0.79). However, the study was smaller than the average of the found studies with a sample size of 946, and as noted above, the main outcome in the dataset the model was developed in was preterm delivery. Sun et al. performed well on external validation (AUROC 0.841), however while the model was developed on a larger than average dataset of 1783 patients, the external validation consisted of only 116 patients. Likelihood ratios could not be calculated of the external validation, but the model did not have good likelihood ratios in the development data (LR+ 4.51, LR- 0.28), and as mentioned above, the model predicted a composite maternal and fetal outcome.

Ngwenya et al. was also externally validated on the miniPIERS dataset, however it could not accurately predict adverse maternal outcomes on external validation, as its AUROC value dropped from 0.796 (development data) to 0.494 (external validation). This is likely an indication of overfitting to the development dataset.

Zheng et al. and Wang et al. both also developed machine learning model(s) on top of a logistic regression model. Zheng et al. tested six machine learning methods (K-nearest neighbour, decision tree, random forest, support vector machine, multi-layer perceptron and linear discriminant analysis), the resulting models each had excellent performance with AUROC>0.9 for all, LR+>10 for all and LR-<0.2 for all except the k-nearest neighbour and random forest models. Wang et al. presented one machine learning model using random forest methodology, which had a high AUROC value and excellent negative likelihood ratio, with a poor positive likelihood ratio (AUROC 0.88, LR+ 2.79, LR- 0.07). These machine learning models appear to be performing better than most logistic regression models, however, it should be noted that machine learning models run the risk of overfitting when the data is small, meaning that they will perform exceptionally well on the data they were developed on but perform poorly on new data. As both of these studies were among the smallest ones with

sample size 733 and 319 respectively, the risk of overfitted machine learning models is exceptionally high, and as none of the models were externally validated, it cannot be determined how reliable the reported performance measures are. Moreover, as mentioned before, Wang et al. modelled a composite maternal and fetal outcome.

Schmidt et al. used random forest and gradient boosted tree to develop two machine learning models, which both performed well to predict a composite maternal and fetal adverse outcome on internal validation, with the gradient boosted tree performing slightly better (AUROC 0.811, LR+ 12.52, LR- 0.34) than the random forest model (AUROC 0.77, LR+ 10.93, LR- 0.43). The models were developed on a slightly bigger than average sample size (1647 patients), however the number of features tested as predictors (114) was quite large relative to the sample size. Schmidt et al. was also the only publication addressing the issue of missing values in clinical practice, and rather than using imputation or complete observations only for their model, they replaced all missing entries with a single value that did not occur within the dataset. While this is a great strength to their study since as a result, the models can used on data with missing values to make predictions, as long as the value used is a numerical value, it can change the summary statistics of variables and potentially skew the models. The models were tested on only 10% of the data as an internal validation and have not been validated externally.

The PREP study was the only study taking a survival modelling approach, creating the PREP-S Cox Proportional Hazards regression survival model. The model had a good AUROC of 0.75 (95% CI 0.73-0.78) on internal validation, however unlike the PREP-L model, it was not externally validated.

The work presented in this thesis compares new models to the model developed in the fullPIERS study (Equation 1.1). This model has been selected as comparator due to its large sample size, robust internal and external validation, and inclusion in clinical guidelines.

$$
\begin{aligned}
logit(pi) = 2.68 \\
&+ [(-5.41 \times 10^{-2}) \\
&\times gestational\ age\ at\ eligibility] \\
&+ [1.23\ \times \text{chest pain or dyspnoea}] \\
&+ [(-2.71 \times 10^{-1}) \times creatinine] \\
&+ [(2.07 \times 10^{-1}) \times platelets] \\
&+ [(4.00 \times 10^{-5}) \times platelets^2] \\
&+ [(1.01 \times 10^{-2}) \\
&\times aspartate\ transaminase] \\
&+ [(-3.05 \times 10^{-6}) \times AST^2] \\
&+ [(2.50 \times 10^{-4}) \times creatinine \times platelet] \\
&+ [(-6.99 \times 10^{-5}) \times platelet \\
&\times aspartate\ transaminase] \\
&+ [(-2.56 \times 10^{-3}) \times platelet \times SpO_2]
\end{aligned}
$$

Equation 1.1

Table 1.2: Previously published multivariable models of maternal outcomes

Study	Publication year	Location(s)	Size	Observed outcome rate	Outcome	Method	AUROC*	Likelihood*	Validation
fullPIERS	2011	Canada, Australia, New Zealand, UK	2023	12.9%	PIERS outcome	Logistic regression	0.81	LR+ 10.57, LR- 0.45	Internal, external
miniPIERS	2014	Fiji, Uganda, South Africa, Brazil, Pakistan	2081	19.6%	PIERS outcome	Logistic regression	0.768	LR+ 5.1, LR- 0.67	Internal, external
miniPIERS with SpO2	2015	South Africa, Pakistan	852	14.0%	PIERS outcome	Logistic regression	0.81	LR+ 10.7, LR- 0.6	None
Millman et al.	2011	Canada, Australia, New Zealand, UK	1534	13.1%	PIERS outcome	Logistic regression	0.73	*NA*	None
Chan et al.	2005	Australia	321	33.6%	Maternal, fetal modelled separately	Logistic regression	0.67	*NA*	None
PREP	2017	UK	946	66.9%	Maternal based on the PIERS outcome + delivery before 34 weeks' gestation	Logistic regression	0.68	*NA*	Internal, external
						Cox regression	0.75	*NA*	
Sun et al.	2021	China	1783	16.0%	Maternal + fetal	Logistic regression	0.841	LR+ 4.51, LR- 0.28	External
Ngwenya et al.	2021	Zimbabwe	549	45.7%	PIERS outcome	Logistic regression	0.49	NA	External
Zheng et al.	2022	China	733	24.8%	Maternal, including admission to ICU	Logistic regression	0.958	LR+ 63.92, LR- 0.17	Internal
						K-Nearest neighbour	0.91	LR+ 22.13, LR- 0.3	
						Decision tree	0.908	LR+ 26.44, LR- 0.16	

Study	Publication year	Location(s)	Size	Observed outcome rate	Outcome	Method	AUROC*	Likelihood*	Validation
						Random forest	0.963	LR+ 23.0, LR- 0.87	
						Support vector machine	0.976	LR+ 11.99, LR- 0.08	
						Multi-layer perceptron	0.973	LR+ 11.99, LR- 0.08	
						Linear discriminant analysis	0.961	LR+ 63.92, LR- 0.17	
Sroka et al.	2023	Germany	655	33.4%	Maternal + fetal	Logistic regression	0.726	LR+ 2.16, LR- 0.40	None
Wang et al.	2024	China	319	29.2%	Maternal + fetal	Logistic regression	0.832	LR+ 2.35, LR- 0.25	Internal
						Random forest	0.88	LR+ 2.79, LR- 0.07	
Liao et al.	2018	China	2532	39.1%	*NA*	Logistic regression	0.804	LR+ 1.79, LR- *NA*	Internal
Schmidt et al.	2022	Germany	1647	23.4%	Maternal + fetal	Random forest	0.77	LR+ 10.93, LR-0.43	None
						Gradient boosted trees	0.81	LR+ 12.52, LR- 0.34	

*on validation, if applicable, on development data otherwise

1.4 Thesis outline

1.4.1 Structure

This thesis is structured as follows: Chapters 2, 3 and 4 are each analysis chapters showing standalone works of modelling the adverse maternal outcomes of preeclampsia from three different angles. Each of these chapters contains an introduction to the specific challenge of risk prediction the chapter aims to address, methodological background for the used statistical and machine learning methods, results, discussion, and conclusion of each modelling approach. Finally, Chapter 5 covers the overall conclusion of the results of the three analysis chapters, as well as discussing the research in context of the existing literature and outlining avenues for future research.

This thesis uses the term “women” to refer to individuals assigned female at birth, and “women” or “patients” to refer to individuals diagnosed with preeclampsia. It should be acknowledged that gender is a complex and fluid spectrum, and not all individuals affected by preeclampsia may identify as women. While this thesis uses specific terms for clarity of scientific communication, we acknowledge that in clinical practice, gender-affirming language should be used instead.

1.4.2 Research questions

This thesis aims to answer three main research questions: (i) can we improve on the performance of the fullPIERS model using machine learning methods, (ii) can we predict the risk of adverse maternal outcomes over time using all up to date information, and (iii) are we able to make informative predictions of risk of adverse maternal outcome in different scenarios when predictor variables are missing?

1.4.3 Data

The analysis presented in this thesis used data collected from previously published studies from multiple sources. The source studies, data format and data cleaning steps for each standalone work of modelling are presented in the corresponding thesis chapter (see Sections 2.3.1, 3.3.1 and 4.3.1). Summary statistics of each resulting dataset are presented in Sections 2.4.2, 3.4.1 and 4.4.1.

Chapter 2: Static modelling of binary adverse outcomes

This chapter is based on our paper, published in The Lancet Digital Health titled The PIERS-ML model: machine learning-enabled maternal risk assessment for women with preeclampsia. Sections of the paper have been incorporated into all sections in this chapter, with additional details, and with the addition of 2.2.1 Predictive modelling, and 2.2.10 Oversampling

2.1 Introduction

Complicating two-to-four percent of pregnancies, preeclampsia (defined as new-onset hypertension at or after 20 weeks' gestation, accompanied by either new-onset proteinuria, other maternal target organ damage, or evidence of uteroplacental dysfunction)(9, 43) remains a leading, global cause of maternal mortality and life-threatening morbidity (9, 29, 43-45). Over 99% of the annual 46,000 preeclampsia-related maternal deaths occur in low- and middle-income countries (LMICs).(46)

In preeclampsia pregnancies, it clear that perinatal survival without major morbidity is largely related to gestational age at birth.(47) However, the burden of adverse maternal outcomes is spread across gestation. While maternal risks are proportionately greater with the earlier onset of preeclampsia,(3, 29, 48) the population-level burden of maternal risk is borne by the 75-80% of cases of preeclampsia that arise at term.(3)

The sole method of initiating recovery from preeclampsia is delivery of the placenta.(43) At term, the focus is initiating birth.(49) Before term, women and their maternity care-providers balance maternal risks from evolving disease with prematurity-related perinatal risks, by subjectively integrating ongoing assessments of symptoms, signs, and laboratory tests.(43) In busy maternity units, considerable experience informs decisions; however, most women with preterm preeclampsia are managed, at least initially, by maternity care-providers with less experience. For many LMIC and disadvantaged high-income country (HIC) populations, access to comprehensive obstetric and newborn care is limited.

To optimise maternal outcomes in preeclampsia, we need objective, time-of-disease maternal risk assessment to inform decision-making over the following 48 hours, wherever that woman lives. Previously, logistic regression was used to develop a model – fullPIERS (Preeclampsia Integrated Estimate of RiSk) – for more-developed countries.(3) Here, we have harnessed the strengths of machine-learning-based classifiers to test the hypothesis that it is

possible to develop and externally validate a novel globally-relevant PIERS-ML model, using information routinely-available at presentation with preeclampsia.

This chapter will introduce the concept of predictive modelling and machine learning, and discuss methods generally used for data cleaning, modelling and model assessment in detail in the methodological background section. We will then discuss the modelling methods used for the specific analyses carried out in the progress of developing the PIERS-ML model in the methods section, and show and discuss the results of these analyses. Finally, we will compare models created with different combination of methods to select the most appropriate model for our purposes and discuss the results in context of previous work and real-world applications.

2.2 Methodological background

2.2.1 Predictive modelling

Predictive modelling is a statistical technique that uses historical data to describe and understand some process with the aim of predicting a future outcome (50). This is done by collecting relevant data either from existing sources or through scientific studies and feeding the data – for which outcomes are known – into some algorithm or formula to train a model, which is then adjusted and optimised (51). Predictive models can be used to predict different types of outcomes, and thus are widely used. Some examples of outcome types and their usage include(50, 52):

- Numerical outcome, e.g. to calculate risk for insurance, or the number of hospital admissions over time,
- Categorical outcome, e.g. for predicting if someone has a disease or not, or if a purchase is likely to be fraud or not,
- Cluster, e.g. to predict groups of customers who have a similar shopping behaviour.

Predictive modelling uses a combination of mathematical and computational techniques to fit and optimise models for the best possible predictions (51). The outcome of interest in predictive modelling is called the outcome variable or sometimes the dependent variable, while the data used to model this is the collection of predictor variables or independent variables.

2.2.2 Introduction to machine learning

"Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed." (53)

Unlike other mathematical modelling techniques, rather than explicitly programming a model which defines the form of the relationship between an outcome variable and potential predictors, machine learning methods are provided with many observations of the potential predictors, and in some cases, the outcome variable, referred to as the label. The machine learning method then learns the relationship from the data rather than from explicitly stated rules. The model's understanding of the relationship increases as the number of observations increases.

Machine learning (ML) can be supervised or unsupervised. In supervised learning the data is labelled and the aim of the model is to correctly identify the label of each observation in the data. The model is then able to compare the predicted label with the observed label to identify errors and make changes. Based on training on data with known labels the machine learning algorithm can predict the label of new, unlabelled data. The accuracy of these predictions depends on the amount of data the ML algorithm was trained on.

In unsupervised learning, the data is unlabelled. Instead of predicting a label, the ML algorithm groups similar data together based on some underlying pattern. Here the algorithm's aim is to minimise the distance between elements within each group while maximising the distance between groups.

Machine learning algorithms can also be semi-supervised and reinforced. The purpose of semi-supervised machine learning is similar to unsupervised learning in that the algorithm aims to group similar data together, however in the case of semi-supervised learning, a small portion of the data is labelled to suggest how the unlabelled data should be grouped (54, 55). In reinforced learning, rather than making predictions or grouping data, the aim of the algorithm is to make decisions in a situation. The reinforcement in the name comes for the step in the learning process where the machine learning model is rewarded for good decisions and penalised for bad decisions (56).

As there is a variable of interest (adverse outcome within 48 hours) to be predicted in this work, only supervised machine learning methods will be used.

2.2.3 Missing data and imputation of missing values

2.2.3.1 Data quality

Machine learning relies heavily on data rather than provided distributions and prespecified relationships, hence correct type and quality of data is especially important for a good model. Machine learning methods can work with (57):

- Quantitative data – discrete or continuous numerical data,
- Qualitative data – categorical or text data,
- Time series data i.e. dates and times.

A factor often affecting the quality of data is completeness. As data is collected, a number of factors can often lead to missing values. Data can be missing completely at random (MCAR) when the missingness is completely by chance, and isn't related to any observed or unobserved data, missing at random (MAR) when missingness is related to some observed data (for example laboratory tests not being done on a Sunday leading to missing measurements on Sundays) but there is no relation between the observed and the unobserved data, or missing not at random (MNAR), where the unobserved data determines the missingness (for example if a machine records only values outside of a normal range, whether a data point is missing is determined by its value) (58). While some machine learning methods have built in ways of dealing with missing data, many do not, hence it is important to deal with missing data before applying machine learning methods.

If data is MNAR, some assumptions can be made about the unobserved data if we know the factor(s) determining whether data was recorded or not. For example, if some test can have a value of integers from 0 to 10, and we know that "negative" tests (test with a zero value) weren't recorded, we can make the assumptions that all missing values are 0. Other methods of dealing with missing data, as seen in Chapter 2.2.3.2, are not advised to use on MNAR data as they are likely to introduce bias (58).

Bias in statistical modelling refers to any flaw in the experiment design or data collection that leads to difference between the sample represented by the data, and the actual population of interest (59). This bias can come from a range of sources, including but not limited to how

the sample for data collection was selected, the methods or questions used to collect the data, not collecting all the relevant variables, inherent bias from the researcher designing the study or collecting the data, and human error or interference (59, 60). The most important tools for avoiding bias are a good study design and randomly sampled data.

As mentioned above, the mechanism of missingness is important to determine the appropriate steps to deal with the missing data, hence, steps should be taken to identify the mechanism of missingness.

Data missing not at random can be identified from information on how the data was collected, for example if the data was collected from/by a machine, were all results recorded or only values outside of an expected range? If the data was collected by a person, was there a protocol for data collection and recording?

The best way to determine if data was missing at random or not is to observe the missing data (61, 62). In some cases, it is possible and appropriate to follow up participants to record the missing data, or we might have longitudinal data available for variables where a later measurement is still applicable. It is important to note, however, that this would only identify the mechanism of missingness if the value of the variable is not expected to change in the period from original data collection to follow-up or the next available data point. For example, it would likely be appropriate to follow up for the number of bedrooms in a house a month later, as it is not likely to change in that timeframe, however measuring the temperature a month later would likely not be useful.

Finally, data missing not at random can be identified from knowledge in the subject area (61, 62). While this does not guarantee that the missing values in our data are missing not at random, knowledge in the area can allow us to make informed assumptions. For example, if we know that young people are less likely to admit to smoking and we have a large number of missing answers to a question on smoking, we could assume based on the prior knowledge that it is missing not at random. Or if we have good knowledge in the area and are not aware of any reason why someone might not be willing to record their hair colour, we might assume that a small number of missing answers to a question on hair colour is missing at or completely at random.

Once we have ruled out missing not at random, next we need to determine if the data is missing at random or completely at random. To test if data in variable A is missing at random, we can create a corresponding missingness variable coded 0 if there is a value observed in variable A and 1 if A is missing. We can them perform a t-test or a chi-square test on the missingness variable with all other predictor variables. If any of these tests is significant, the mechanism of missingness for variable A would be missing at random. (61)

Data missing completely at random can only be identified with a process of elimination. If we are confident that the missing data is not MNAR and none of the tests for MAR showed to be significant, the mechanism of missingness would be missing completely at random.

2.2.3.2 Imputation of missing data

While missing data itself may not always cause bias, it can lead to problems depending on the nature of missingness and how it is handled in the analysis. If data is missing due to a flawed study design, the flaw could be affecting the quality of the gathered data as well, and even the complete data may not be high enough quality for an accurate analysis. If the missingness is not due to the study design, most problems are caused by MNAR data, as it cannot be accurately estimated by the observed data and omitting study subjects with missing observations to use complete data only would make the sample less representative. Data MAR or MCAR can be omitted and still preserve the representativeness of the sample, however this will still lead to loss of information, potentially leaving us with not enough data to create an accurate model (63).

For data that is MAR or MCAR, ways of dealing with missingness include (64):

- Using complete data only (removing rows with missingness),
- Imputation by replacing all missing values in a variable by a single value e.g. mean, median,
- Using imputation methods

Imputation can be single imputation and multiple imputation. As mentioned in Chapter 2.2.2, single imputation can be done by replacing all missing values in a variable by a single value for example the mean for continuous variables, or the median for discrete or categorical variables, or the value from a new data point not yet included in the sample (65). These methods are not very effective for large numbers of missing data, especially if the data is

MCAR, since for MCAR data the unobserved data should have the same distribution and therefore the same variance as the observed data, so replacing a potentially wide range of values with just a single value could lead to severe bias (66). This bias can often be even worse than working with complete data only (65). Single imputation methods also include replacing each missing value separately by either a completely randomly selected value from another study subject, the value from a study subject similar in other variables, or a study subject selected by some sort of pattern or rule (65, 66). Finally, missing values can be imputed from observed data using more complex methods such as regression imputation, random forest, or other machine learning methods.

Multiple imputation repeats the imputation n times creating n datasets, the process is done on each dataset and then combined by either pooling model coefficients to create a new model, or keeping multiple separate models and averaging the model outputs (67). Since multiple imputation follows the same process of imputation every time with some amount of random variability added to the imputed values, the imputations end up similar, but not identical, accounting for the variance observed in the population. All single imputation methods can be used for multiple imputation.

Multiple imputations are recommended over single imputation as single imputation can be more biased, underestimates the standard errors and can't recognise what values are imputed and what values are observed, therefore does not account for the fact that error and therefore bias can come from imputation (65, 68, 69). Single imputation can be useful for small amounts of missing data, otherwise multiple imputations should be used (69).

There are multiple ways of calculating the number of imputations needed. Derived from Bodner's simulation study (70), a rule of thumb for determining the number of imputations is to use at least the percentage of incomplete cases or more (71).

As explained in Chapter 2.2.9, for model fitting the data is split into two datasets, one used for fitting the model while the other used to validate it. In this case, the data is first split, then the two new datasets are imputed separately. This is to make sure that there is no dependency between the two datasets.

To decide when to use complete case analysis, single or multiple imputation, we need to look at the mechanism of missingness, the amount of missing data we have and our computational capacity.

Complete case analysis could be appropriate if we only have a small amount of missingness in our data and the mechanism of missingness is MCAR. Complete case analysis should not be used if there is any data MNAR.

For data MNAR, it might be appropriate to use single imputation if we can make assumptions on the missing values. In this case, we could use imputation of replacement by a single value, replacing the missing data by either an expected value or the expected mean/media. More complex methods of imputation would not be appropriate, as the assumption that the missing data and the observed data come from the same distribution does not hold for data MNAR. If we cannot make assumptions on the missing values, the variable with MNAR missingness cannot be used. If all MNAR variables are removed, complete case analysis could be used on the other variables if the mechanism of missingness for the rest data is MCAR and only a small amount is missing. It is suggested to take up to 5% missing data as negligible (68, 72, 73).

For data MAR or larger amounts of data MCAR, imputation of missing values can be a good way of dealing with missing data (63).

Single imputation is less computationally expensive and easier to use as it only results in one imputed dataset, however it heavily relies on assumptions on the missing data (68). The assumptions depend on the method of imputation used, for example missing values are the same as the last observed value (for the LOCF, Last Observation Carried Forward method), all missing values are the same or are not significantly different (for replacement with a single value), or can be exactly predicted from other observed variables, ignoring the variability of measurements (for all methods attempting to model the variable and predict the missing values). These assumptions only hold in very specific circumstances and are unlikely to be realistic for our data (68).

Single imputation may be appropriate if only small amount of data is missing, if we're confident the assumptions for the method we're using hold for out data, if we're not

expecting large variability or if multiple imputation would be too computationally expensive, however this has the potential of introducing bias.

In cases where we know there is a variety of values missing or when we cannot make assumptions on the missing data, and the mechanism of missingness is not MNAR, multiple imputation is going to provide less biased estimates than single imputation as it takes the variability of the data into account. Multiple imputation, however, is a lot more computationally expensive, which may not be justifiable if the amount of missingness is small. The gain in performance from using multiple imputation should be weighed against the added cost of imputation and analysis on multiple datasets.

Multiple imputation is generally suggested to be used for larger amounts of missing data, however whether there is an amount of missingness that is too much for imputation in general is debatable. Previously it has been suggested that "too large" (such as 40% (68) or 60% (74)) amount of missing data should not be imputed, however recent studies (75, 76) suggest that the decision whether to impute should be based on assumptions of the mechanism of missingness and conditions of the imputation rather than the amount of missing data. These studies imputed up to 80% and 90% data MAR or MCAR with minimal bias under the conditions of:

- Large enough number of imputations
- Using multiple chained equations/random forests
- Well defined imputation
- Informative auxiliary variables.

Given this information, even variables with large proportion of missingness can potentially be imputed if we are confident that all of the above conditions are met. If the imputation cannot be well defined or the squared coefficient of multiple correlation with the observed variables used for imputation is low, imputation cannot be used without introducing bias even at lower proportions of missingness. In this case, the variable should be removed from modelling.

Methods from other fields for imputation of missing values also exist, such as generative adversarial imputation networks, which are discussed in Section 4.2.2 in Chapter 4.

Once all missing data has been addressed and handled appropriately, the now complete dataset(s) can be used for predictive modelling.

The following chapters will discuss predictive modelling methods using machine learning for a binary outcome.

2.2.4 Regression modelling

Regression modelling could be considered a basic form of machine learning, modelling an outcome, called the dependent variable, using one or more predictors that are independent of one another. Models are estimated by optimising the coefficients of an equation such that the squared sum of the prediction errors is minimised.

Regression techniques fall into many categories, including but not limited to (77):

- Linear regression, where a continuous outcome is predicted by some combination of the predictor variables that can include first or higher order terms and a linear or non-linear combination if the predicted outcome is first order and does not appear in the combination of the predictors,
- Logistic regression, where a binary outcome is predicted similarly to linear regression,
- Ridge regression, a form of penalised regression (explained in Chapter 2.2.5),
- LASSO regression, a form of penalised regression and variable selection method (explained in Chapter 2.2.5), and
- Non-linear regression, where linear and non-linear combination of first and higher order terms are used for prediction.

Linear regression models the variable of interest, Y, as

$$E(Y) = \sum_{j=0}^{n} \beta_j X_j$$ *Equation 2.1*

Where β_j are the regression coefficients, $X_0 = 1$ and $X_j, j \neq 0$ are the independent predictor variables. Y and X_j are known from the observed data, while β_j is estimated to minimise the cost/loss function:

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2$$ *Equation 2.2*

Where y_i is the *i*th observed outcome, x_{ij} is the *i*th observation of the *j*th predictor variable and β_j is the coefficient of the *j*th predictor variable.

Logistic regression works in a very different fashion, however as the variable of interest is binary, we are not able to directly predict it using a linear combination of categorical and continuous variables. Hence the regression equation becomes:

$$\text{logit}(p) = \sum_{j=0}^{n} \beta_j X_j \qquad \textit{Equation 2.3}$$

Where β_j and X_j are same as above, $p = \Pr(\text{Y} = 1)$ is the probability that we observe an event and $logit(p) = \log\left(\frac{p}{1-p}\right)$ or the log odds of observing an event. The log odds can be converted back into probability by

$$Pr(Y = 1) = \exp\left(\frac{1 - \text{logit}(p)}{\text{logit}(p)}\right) \qquad \textit{Equation 2.4}$$

The predicted outcome is then estimated as

$$\hat{Y} = \begin{cases} 1, & p \geq C \\ 0, & p < C \end{cases} \qquad \textit{Equation 2.5}$$

for a fixed cut-off value C.

Similarly to linear regression, the beta coefficients are estimated to minimise the cost/loss function which is now

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textit{Equation 2.6}$$

Where $\hat{y}_i$ is the *i*th observation of the predicted outcomes and y_i is as before.

2.2.5 LASSO and Ridge regression

Least absolute shrinkage and selection operator (LASSO) and Ridge regression are both penalised regression methods, while LASSO is also a variable selection method.

When models are fitted, they utilise all predictor variables provided, assess the relationship between each predictor and the outcome variable, and expresses this relationship and the strength of this relationship in either the coefficient of the variable – for models that have observable coefficients – or the structure of the model or both. However, if a predictor variable has no or very weak relationship with the outcome compared to other predictors, it

can lead to having an element in the model that does not contribute to the predicted outcome. Variable selection methods aim to simplify a model by either identifying and keeping the important variables (the variables that contribute the most to the predicted outcome) or identifying and removing the least important variables (the variables that contribute the least to the predicted outcome). Penalisation or regularisation in regression modelling is a constraint that is placed on the coefficients of the model to shrink them towards zero by some amount defined by the constraint.

LASSO selects the most significant variables for a model by performing regularisation to shrink the coefficients, by a degree depending on a tuning parameter. As the coefficients of nonimportant and correlated variables tend to be smaller than that of the more predictive variables, they would be the first to shrink to or close to zero when the tuning parameter is large enough (78, 79). Similarly, Ridge regression is a penalised regression method that regularises models by adding a degree of bias to the regression estimates, however as it uses a different type of regularisation (detailed below), it shrinks larger coefficients by a much larger value, in turn resulting in no coefficients being shrunk to zero. (80, 81)

Regularisation works by adding a regularisation element controlled by a tuning parameter (λ) to the loss function. There are two types of regularisation (82):

- L1 regularisation, where the regularisation element is $\lambda \sum_{j=1}^{p} |\beta_j|$
- L2 regularisation, where the regularisation element is $\lambda \sum_{j=1}^{p} \beta_j^2$

LASSO uses L1 regularisation: for a given lambda, LASSO chooses the beta coefficients to minimise Equation 2.7 (79, 82) for linear regression, where y_i is the i^{th} observation of the outcome variable, x_{ij} is i^{th} observation of the j^{th} predictor variable and β_j is the coefficient corresponding to the j^{th} predictor variable.

$$\sum_{i=1}^{n} (y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \textit{Equation 2.7}$$

For logistic regression L1 regularisation chooses the beta coefficients to minimise Equation 2.8 (83), where $\hat{y}_i$ is the i^{th} observation of the predicted outcome and y_i and β_j are as before.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad \textit{Equation 2.8}$$

Ridge regression uses L2 regularisation, creating the cost function seen in Equation 2.9 (82) for linear regression, and Equation 2.10 for logistic regression.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}{\beta_j}^2 \qquad \textit{Equation 2.9}$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p}{\beta_j}^2 \qquad \textit{Equation 2.10}$$

Setting $\lambda = 0$ gives ordinary regression for both methods, while as $\lambda \to \infty$, all predictor variables are eliminated from the equation giving $y = C$ for some constant C in LASSO (78), and all coefficients are shrunk too small creating underfitting in Ridge. In many implementation algorithms, $\lambda = 1$ is chosen as default, an arbitrary choice which gives a moderate amount of regularisation and a point of reference for interpretation of changing lambda values. $\lambda < 1$ slows the penalty giving more variables and/or larger coefficients in the model, and $\lambda > 1$ accelerates the penalty, having the opposite effect. The default value should be replaced by the optimal value for the model and data.

The optimal λ is found by taking a range of λ values, fitting a model for each value using cross-validation (explained in Chapter 2.2.9) and calculating the cross-validation error, or the average error in prediction on the test sets in cross-validation. The optimal λ value is the one that minimises cross-validation error (79).

In R this can be done by using the `cv.glmnet` function from the `glmnet` package (84). The function estimates the lambda value associated with the lowest cross-validation error (`lambda.min`), and the highest lambda value within 1 standard error of the minimum (`lambda.1se`). Since the aim when using LASSO or Ridge is to have as high accuracy as possible with the simplest possible model, `lambda.1se` is used for the final model. This is due to the fact that lambda is an estimate, not a true minimum and hence will be estimated with some error. Using any lambda value within 1 standard error of the minimum will give us

minimal loss or gain of accuracy, and using the highest lambda from this range will give us the simplest model. (84)

2.2.6 Tree based methods

2.2.6.1 Trees

A regression tree (for continuous outcome) or classification tree for (categorical outcome) is a supervised machine learning method that splits the predictor space into regions and provides a value/category for each region.

Classification and Regression Trees (CARTs) consist of nodes connected by branches, with the final nodes of a branch called leaf nodes. A tree is grown starting with a single node called the root node, that contains all the data points. At each node, the data is split into two branches according to a yes/no test of one of the variables. All possible splits of all variables and all possible conditions are tested at each node and the split with the lowest cost according to some cost function is selected. An example of a classification tree can be seen in Figure 2.1.

For regression trees the cost function is

$$C = \sum (y - \hat{y})^2 \qquad \textit{Equation 2.11}$$

while the cost for classification trees is measured by the Gini index

$$G = \sum p_k * (1 - p_k) \qquad \textit{Equation 2.12}$$

where *y* is the continuous outcome, and p_k are the proportion of the categorical outcome with class k in the current region of the predictor space. (85)

Trees are grown until some stopping criteria is met. This could be:

- Threshold for number of instances in leaf, number of leaf nodes or error of leaf
- All instances in the leaf are in the same class or have the same value (86)
- P-value from a statistical test such as univariate analysis or Bonferroni test, crosses a threshold.

This method of growing trees is, however, an optimistic approach and thus it is easy to overfit the model to the data. The solution to this is pruning.

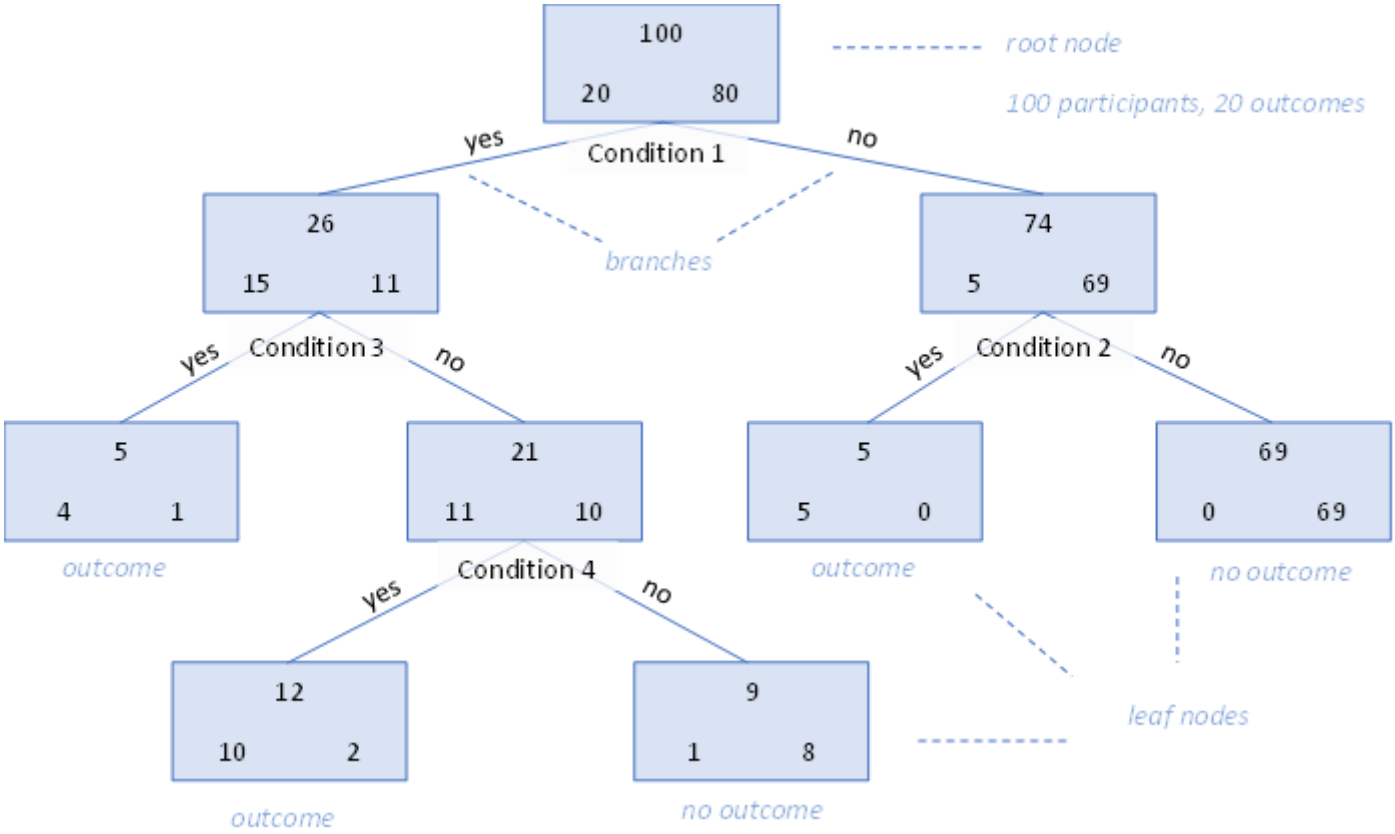

Figure 2.1 General classification tree of 100 participants (80 with no outcome and 20 with an outcome). Each node shows the number of participants in the node (top), number of participants with outcome (bottom left) and number of participants without outcome (bottom right).

The best way to grow a tree with high accuracy that does not overfit the data is to grow a large tree and then prune it back to find the subtree with the lowest error rate. There are multiple ways to prune a tree, such as evaluating the effect of removing a leaf node on the overall cost until the cost cannot be further decreased by the removal of a leaf (85). Another approach is to use a cost complexity pruning algorithm (87).

2.2.6.2 Random forest

Random forest is a supervised machine learning ensemble method that grows several classification/regression trees in order to compensate for overfitting and bias and get a more accurate prediction.

The random forest method takes a random sample of the data, and a random sample of the predictor variables for each tree, grows a tree based on the best splits for the data and variable combination, and repeats this process *m* times (Figure 2.2). This leads to a collection of trees that each fit a subset of the data well, with a variety of different splits. When the

random forest is presented with new data, it is run through on each of the trees, resulting in *m* predictions which are then combined to give the final prediction.

Due to the nature of the growing process of a random forest, while it uses all available variables, it does not overfit. Variables are assigned an importance based on their mean decrease in the Gini index.

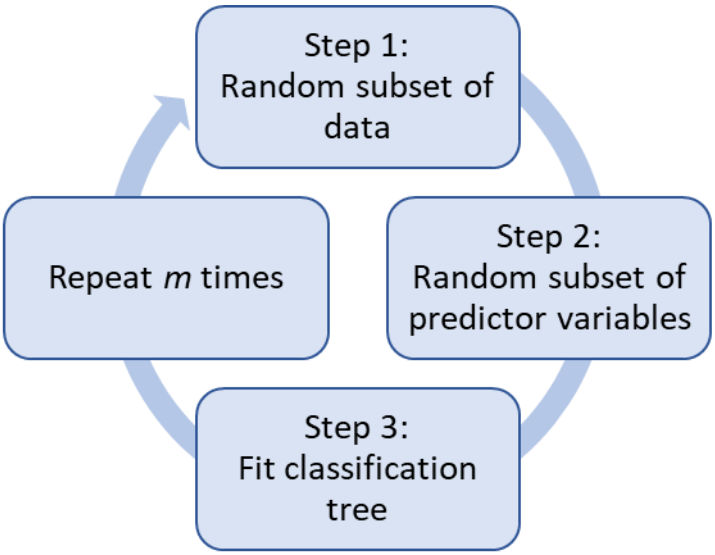

Figure 2.2 Random forest fitting process

For each variable, the Gini index is calculated in each tree with the Equation 2.13. However, since in a random forest each tree is a random subset of the data and the available variables, not the same proportion of observations will reach the node where the variable is used in different trees. For example, variable A could be the first split in one tree, meaning that all observations pass through it, while it could be closer to a leaf node in another tree, where only a much smaller proportion of observations reach. Hence, for random forests we measure the variable's total decrease in node impurity in each tree, which is the difference of the Gini index of the node where the variable is used, and the weighted sum of the Gini index of the two new nodes.

The decrease in the Gini index or node impurity is calculated by

$$\Delta G = G(k) - \sum_{l=1}^{2} p_l G(k_l)$$ *Equation 2.13*

Where $G(k)$ is the Gini index of the k[th] node, k_1 and k_2 are the nodes created by the split at the k[th] node, and $p_l = \frac{n_{k_l}}{n_k}$ for $l = \{1,2\}$ is the proportion of the population of the k[th] node split to the k_1[th] and k_2[th] nodes respectively. (88) The mean of these weighted values is the Mean Decrease in Gini.

Error! Reference source not found. and Figure 2.3 illustrate the calculation of the mean decrease in Gini index (meanDecreaseGini) for variable A in a random forest of two trees.

Variables with a high mean decrease in Gini index have a large effect on the outcome while variables with a mean decrease in Gini index close to zero have little to no effect on the outcome.

Table 2.1: Calculations for decrease in Gini index for variable A in tree1 (left) and tree2 (right) and the mean decrease in Gini index

$G(1) = 1 - \left(\frac{20}{100}\right)^2 - \left(\frac{80}{100}\right)^2 = 1 - 0.04 - 0.64 = 0.32$	$G(1) = 1 - \left(\frac{17}{25}\right)^2 - \left(\frac{8}{25}\right)^2 = 1 - 0.4624 - 0.1024 = 0.4352$
$G(2) = 1 - \left(\frac{5}{70}\right)^2 - \left(\frac{65}{70}\right)^2 = 1 - 0.005 - 0.862 = 0.133$	$G(2) = 1 - \left(\frac{1}{9}\right)^2 - \left(\frac{8}{10}\right)^2 = 1 - 0.0123 - 0.7901 = 0.1976$
$G(3) = 1 - 2 \cdot \left(\frac{15}{30}\right)^2 = 1 - 2 \cdot 0.25 = 0.5$	$G(3) = 1 - \left(\frac{16}{16}\right)^2 - \left(\frac{0}{16}\right)^2 = 1 - 1 - 0 = 0$
$\Delta G_1 = 0.32 - 0.7 \cdot 0.133 - 0.3 \cdot 0.5 = 0.32 - 0.0931 - 0.15 = 0.0769$	$\Delta G_2 = 0.4352 - \frac{9}{25} \cdot 0.1976 - \frac{16}{25} \cdot 0 = 0.4352 - 0.0783 - 0 = 0.3569$
$meanDecreaseGini = \frac{100 \cdot 0.0769 + 25 \cdot 0.3569}{2} = \frac{7.69 + 8.9225}{2} = 8.3063$	

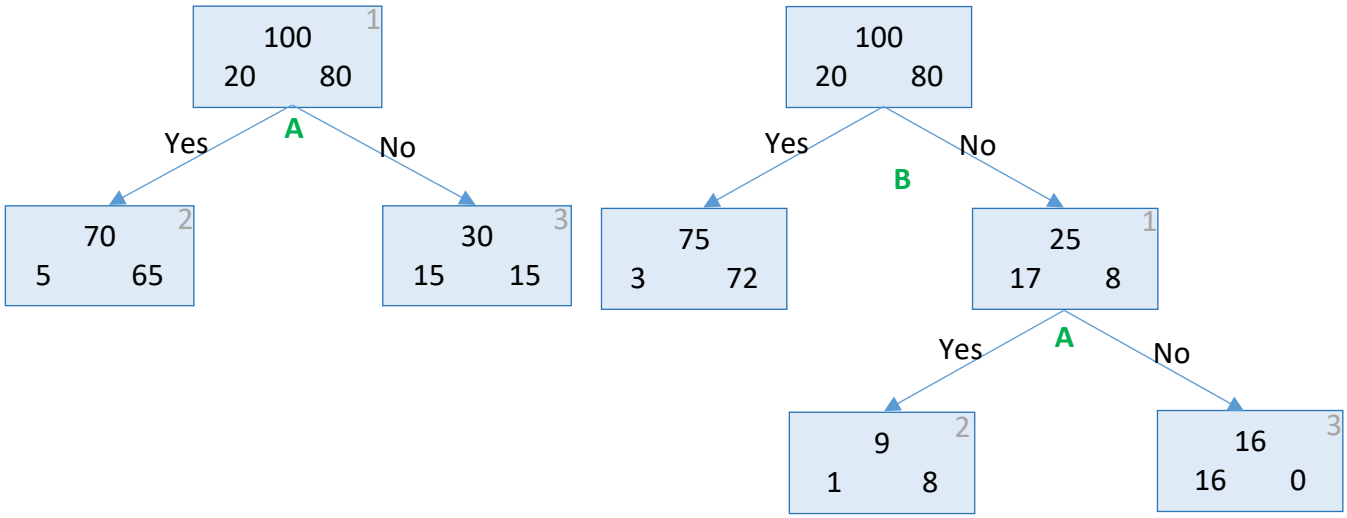

Figure 2.3: Example forest of two trees, tree1 (left) having a condition variable A as the first split, tree2 (right) having a condition on a different variable as the first split and having variable A further down the tree. Each node shows the number of participants in the node (top), number of participants with outcome (bottom left) and number of participants without outcome (bottom right).

This means that while all available variables are used to grow the random forest, some variables may be redundant. When there are only a small number of variables or all are easily obtainable, all available variables could be used. However, it is not always ideal or feasible to collect all variables that were available for model development when we want to use a model for prediction in a real life situation. This is especially true in cases where a model might contain variables that are difficult or expensive to obtain and may not be significant in the created model, or in cases with large numbers of variables. As random forests provide us with a way to assess a variable's importance in a model, we can use variable selection methods and model performance assessment to remove predictor variables and identify the model with the lowest number of variables that still provides accurate predictions.

2.2.6.3 Variable/feature selection in random forests

Once a random forest with all variables is grown, the variables can be ranked based on importance. One variable selection method is to choose only variables with above average importance. This is simply done by calculating the mean of the MeanDecreaseGini (the mean decrease in the Gini index) of all variables and selecting variables for the new forest only if their MeanDecreaseGini is greater than or equal to this mean.

Another method that can be used once a random forest with all variables is grown is recursive feature elimination (RFE)(89). It recursively identifies the least important variable or a specified number of the least important variables, removes them and grows a new random forest with the remaining variables. Each random forest's prediction error rate is assessed. The process is repeated until there is only one variable left and the random forest with the smallest error rate is chosen.

The Boruta method (90) evaluates the relevance of each variable by duplicating each variable and growing a tree with all available variables and their duplicates. The duplicated variables, called shadow variables, are permutations of the values of the original variables, hence they should no longer be important or relevant predictors of the outcome. The Boruta algorithm creates several random forests. For each random forest, shadow variables are created, the forest is grown, and the importance of all variables is calculated. The importance of each real variable is then compared to the highest shadow variable's importance. Variables with importance significantly smaller than this shadow variable's are classed unimportant and removed, along with all the shadow variables and the procedure is repeated until either all variables are classed as important (importance significantly larger) or a specified number of variables is reached.

The final variable selection method considered, Vita, permutes the outcome m times, grows a forest for each permutation and calculates the permutation variable importance for all variables (91). This leads to a vector of m values of permutation variable importance for each variable, which is then used to approximate the null importance distribution of the variable. Given this null importance distribution, the probability of observing a value greater than or equal to the original variable importance is calculated. This probability is the p-value for the variable. A cut-off value is specified, and if the p-value is less than this value, the variable is classed as important.

2.2.6.4 Gradient boosted trees

Gradient boosted trees are also an ensemble method of CARTs which is very similar to random forest, with the distinguishing differences lying in the pruning of the included trees, and the way the ensemble is assembled.

In this section only gradient boosted trees for binary classification will be discussed. A model built with gradient boosted trees is fitted the following way:

Using a training dataset, first all data points are assigned the same prediction. The log odds of the outcome are commonly used for this value. Here let us assume that this prediction is 0.5, and we have coded the binary outcome of the training dataset as 0 for no outcome and 1 for outcome. As our outcome is recorded as a probability, the log odds also need to be converted to this, in our example giving ~0.6 for all data points. The residuals, or the differences between the prediction and the outcome are calculated.

The following steps will detail starting to create tree models to predict this residual probability, rather than the recorded outcome. The objective function is used to select the tree which gives us the best possible prediction while also not being too complex. The objective function (Equation 2.14) consists of two parts, a loss component and a regularization component. The loss, a logistic loss in our case, helps us select the tree with the best discrimination, while the regularisation component prunes all the trees that are grown. The tree for which the objective function generates the lowest value is selected and added to the model. The model is then used to make predictions for all data points, by applying the tree to the predictor variables, converting them to log odds, and then adding the constant (0.5 in our case) to the predicted value.

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i) \qquad \textit{Equation 2.14}$$

The new predictions are converted back to probabilities and then used to calculate new residual probabilities for each data point, and the steps above are repeated using the new residuals as the target y_i. The selected classification tree is added to the model, which at this point consists of a constant for all data points and two classification trees. The process is repeated until either a pre-specified number of iterations is reached, or the residuals become sufficiently small. Equation 2.15 (92) shows the calculation of predicted output in log odds after t iterations for the ith data point, $\hat{y}_i^{(t)}$, where $\hat{y}_i^{(0)}$ is the constant prediction for all data point set at the first step, and $f_k(x_i)$ is the output of the kth classification tree function, f_k based on the predictor variables for the ith data point, x_i.

$$\hat{y}_i^{(0)} = 0.5$$
$$\hat{y}_i^{(1)} = \hat{y}_i^{(0)} + f_1(x_i)$$
$$\hat{y}_i^{(2)} = \hat{y}_i^{(1)} + f_2(x_i)$$
$$\hat{y}_i^{(3)} = \hat{y}_i^{(2)} + f_3(x_i)$$
$$\dots$$
$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) = \hat{y}_i^{(0)} + \sum_{k=1}^{t} f_k(x_i)$$

Equation 2.15

When new data is fed into the model, just like in random forest, all trees in the ensemble create a prediction, however unlike in random forest, these predictions are added together, rather than averaged. Because of this difference, in gradient boosting, the output of each of the trees is converted to a log of the odds, rather than a class or a probability, so that they can be added together, which can be converted back to a probability at the end if desired.

2.2.7 Bayesian Model Averaging

Bayesian model averaging is a Bayesian modelling method that works by estimating a posterior probability of each possible model given a prior probability, observed data and possible weights to be applied to the data. The prior probability can be provided if there is available information from previous studies or prior knowledge. If there is no available information to inform a prior probability, a uniform, uninformative prior is used (0.5). (93)

Data is then provided for the algorithm, with the option to allocate weights to each observation (row of the data), otherwise the weight of 1 is used for all observations. All linear models possible by different combinations of the predictors are fitted on the (weighted) data, and the posterior probability for each of these models is calculated.

For each model M_k, the posterior probability is calculated as seen in Equation 2.16, where $\Pr(M_k)$ is the prior probability of model M_k, and $\Pr(data|M_k)$ is the marginal likelihood of model M_k, defined by Equation 2.17 (94). The posterior probabilities are divided by a normalising factor, which is the sum of the marginal likelihood times the prior probability of all models. After normalisation, the posterior probabilities of all models will add up to 1. Equation 2.17 describes the marginal likelihood of model M_k, calculated as an integral with

respect to the model parameters ($\boldsymbol{\theta}_k$) of the parameter likelihood ($\Pr(\text{data}|\boldsymbol{\theta}_k, M_k)$), and the prior probability of the model parameters ($\Pr(\boldsymbol{\theta}_k|M_k)$).

$$\Pr(M_k|\text{data}) = \frac{\Pr(\text{data}|M_k)\Pr(M_k)}{\sum_{l=1}^{K}\Pr(\text{data}|M_l)\Pr(M_l)} \quad \textit{Equation 2.16}$$

$$\Pr(\text{data}|M_k) = \int \Pr(\text{data}|\boldsymbol{\theta}_k, M_k)\Pr(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k \quad \textit{Equation 2.17}$$

Once the posterior probabilities are calculated, each model is assessed and a decision on whether to include or exclude the model in the averaging process is made using Occam's window where model M_k is excluded if either of the following criteria is met (95, 96):

- $\frac{\max_l\{\Pr(M_l|\text{data})\}}{\Pr(M_k|D)} > C$, where C is a fixed constant, $\Pr(M_k|\text{data})$ is the posterior probability of model M_k, and $\max_l\{\Pr(M_l|\text{data})\}$ is the posterior probability of the model with the highest posterior probability; or
- there exists a submodel of M_k with higher posterior probability.

Using only the models that were not excluded in this step, the posterior distribution of coefficients is estimated by Equation 2.18 (95).

$$\Pr(\Delta|\text{data}) = \sum_{k=1}^{K}\Pr(\Delta|M_k, \text{data})\Pr(M_k|\text{data}) \quad \textit{Equation 2.18}$$

Finally, models are ranked based on their posterior probabilities and the final model parameters are calculated by averaging over the top n models (97)

2.2.8 Artificial Neural Networks

Artificial neural networks are supervised machine learning methods derived from the biological neurons in the brain (98). Biological neurons receive information via dendrites, process the information in the nucleus within the cell and create output signals. Similarly to this, artificial neural networks take information from the predictor variables provided for it in the input layer, the information is processed inside the neural network through hidden layer(s), weights and mathematical functions, and an output is created in the output layer. The output can take on many forms, such as a probability, a numeric outcome, a

classification group, or for multiple group classification, the probability of belonging to a specific group for each of the possible groups of the output variable (99).

Artificial neural networks are created by machine learning algorithms with the goal of creating an output with the lowest amount of error, measured by the cost function (100). The algorithm learns the relationship between the input and the output layer by adjusting the weights, bias, and number of nodes in the hidden layer until finding the combination that can predict the output the most accurately. In some cases, the algorithm can also set the number of hidden layers and the number of nodes in each layer, while in other cases we can provide a fixed number of layers and nodes. The model fitting, or learning process of the algorithm, includes looking at each observation of the data, fitting best weight combination, moving to the next observation and adjusting weights (98). The learning process is controlled by factors such as a list of starting weights, learning rate and weight decay. Learning rate sets the rate for how much the new information changes the weights, the higher the learning rate is, the more the model favours weights that create accurate predictions for the new observation versus the previous observations (101). To prevent overfitting, regularisation can be used in the form of weight decay (101). Initial set of weights before looking at the first observation is randomly generated unless a list of weights is specified.

The model fitting process is similar to a series of connected regression models (100). For each node in the hidden layer, a weighted sum of the input layer is taken where the weights act similarly to coefficients in a regression model, and some bias is added, which could be considered the intercept of a regression model. The last step of the process is different from regression, as an activation function *f* is applied to the weighted sum and bias, as shown in Equation 2.19 (100), where H_m is the *m*th node in the hidden layer, $\beta_{m,i}$ is the weight corresponding to the *i*th input variable, X_i, for the *m*th node in the hidden layer and $b_{m,1}$ is the bias added for the *m*th node in the hidden layer. The activation function can be the identity function, which is equivalent to omitting the last step entirely. Other functions include but are not limited to: the binary step function, the Sigmoid function, tanh, ReKu, Maxout, Ramp function and ELU (98, 102). As each node in the hidden layer is created separately in this fashion, weights are, or can be different for each node. If there are multiple hidden layers, each extra hidden layer is created the same way, by repeating the same steps on the previous hidden layer instead of the input layer. Finally, the output is calculated from

the last hidden layer, shown in Equation 2.20, where Y is the output, γ_j is the weight corresponding to the *j*th node in the last hidden layer, H_j, b_2 is the bias added for the output and *f* is the activation function. Figure 2.4 shows an example neural network with four input nodes, three hidden nodes and a single output node.

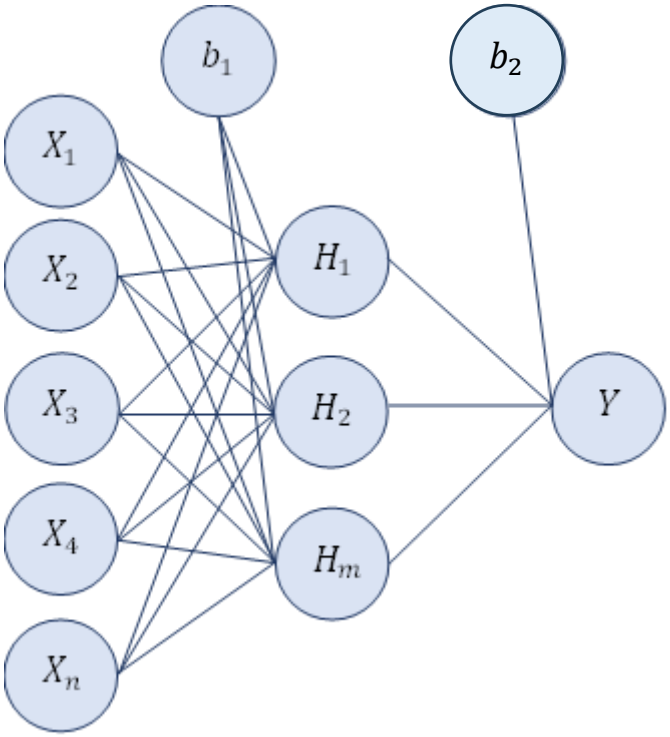

Figure 2.4: Artificial Neural Network

$$H_m = f\left(\sum_{i=1}^{n}(\beta_{m,i} * X_i) + b_{m,1}\right) \qquad \textit{Equation 2.19}$$

$$Y = f\left(\sum_{j=1}^{m}(\gamma_j * H_j) + b_2\right) \qquad \textit{Equation 2.20}$$

2.2.9 Fit and assessment

For model fitting, the data is randomly split into two sets. For example, one-third is selected and reserved for validation of predicted probabilities and the remaining two thirds is used for predictive model development.

In the first step of predictive model build, regardless of the algorithm used, models are estimated on the development data using *k*-fold cross-validation to avoid overfitting. Cross-validation splits the data into *k* sections and creates *k* subsets (folds) by taking a random sample of the predictor variables. Each fold is split into training and test sets by holding out one section for test and combining the other *k-1* sections for training. Models are fitted on the training, tested on the test set and assessed by creating a cross-validation error (Equation 2.21). The final model from cross-validation, chosen as the model with the lowest cross-validation error, is then used to create predictions on the validation dataset. These predictions along with the observed outcomes are used to assess the model's performance. The initial data split and a visual example of cross-validation is shown in Figure 2.5.

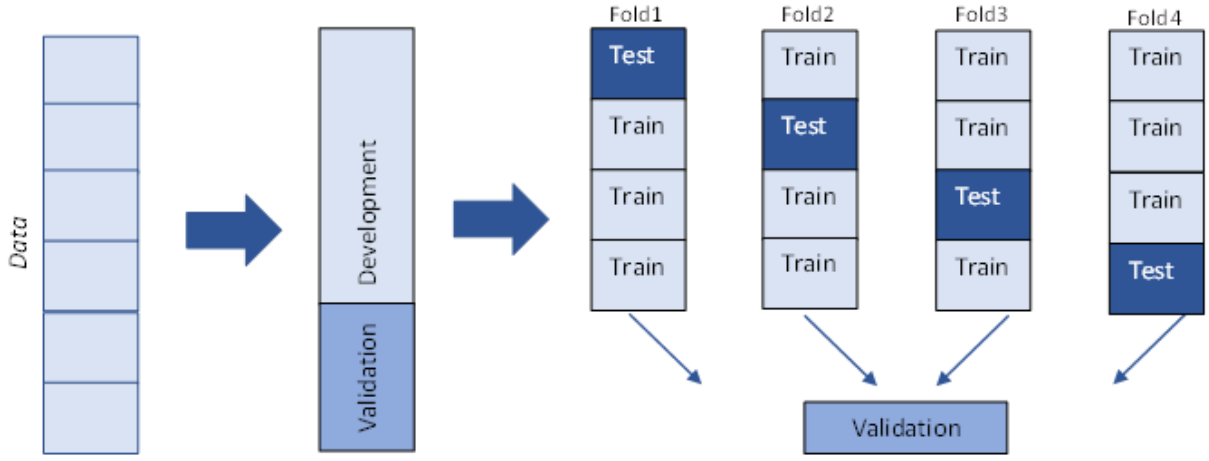

Figure 2.5 Validation-development split and cross-validation

The cross-validation error for k-fold cross-validation is calculated as Equation 2.21 (103), where MSE_i is the Mean Squared Error for predictions made on the testing set of the *i*th fold using the model trained on the training set of the *i*th fold.

$$\text{error}_{CV} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \qquad \textit{Equation 2.21}$$

2.2.9.1 Outcome/no outcome prediction

Models create predicted probabilities of an outcome. *Figure* 2.6 shows examples of predicted probability distributions for 100 observations with an outcome and 100 observations without. Using a threshold, these can be converted into a binary predictor of outcome. Once a threshold is choosen, we can assess the accuracy of the predictions by using confusion matrices and calculating sensitivity, specificity, positive predictive value (PPV), negative

predictive value (NPV) and postive and negative likelihood ratios (LR+ and LR- respectively) using this.

As we can see from Figure 2.6, and **Error! Reference source not found.** and Table 2.3, the *a, b, c* and *d* values in the confusion matrix depends on two factors: the distribution of the predicted probabilities for cases with and without an outcome, and the chosen threshold. Very low thresholds will correctly classify most if not all cases with an observed outcome, but will misclassify a large proportion of the cases with no observed outcome, while high thresholds tend to do the opposite, correctly classify the cases with no outcome, but misclassify the cases with an outcome. What a "low" or "high" threshold is, and what proportion of cases it will classify correctly depend on the distributions and how much they overlap.

Table 2.2: Confusion matrix

		Observed outcome	
		Yes	No
Predicted outcome	Yes	a	b
	No	c	d

Table 2.3: Confusion matrix for the Examples and Thresholds in Figure 2.6 with 100 cases with outcome and 100 cases with no outcome

	Example1			Example2			Example3		
Threshold1		Yes	No		Yes	No		Yes	No
	Yes	100	70	Yes	100	3	Yes	23	0
	No	0	30	No	0	97	No	77	100
Threshold2		Yes	No		Yes	No		Yes	No
	Yes	100	56	Yes	82	14	Yes	3	3
	No	0	44	No	18	86	No	97	97
Threshold3		Yes	No		Yes	No		Yes	No
	Yes	97	76	Yes	62	30	Yes	17	1
	No	3	24	No	38	70	No	83	99

The threshold can be selected based on different criteria, for example to maximise sensitivity or specificity. Most commonly thresholds are selected to optimise both sensitivity and specificity.

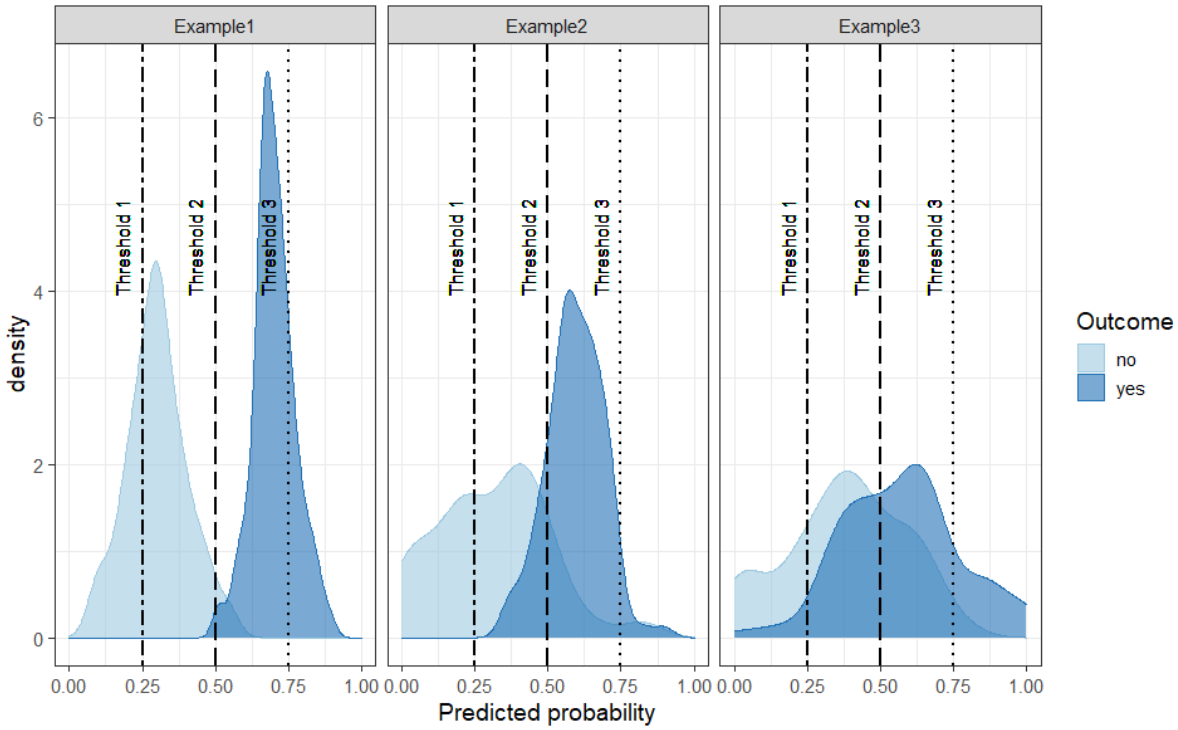

Figure 2.6: Example densities

Sensitivity, calculated from the values in **Error! Reference source not found.** as

$$\text{sensitivity} = \frac{a}{a+c} \qquad \textit{Equation 2.22}$$

is the correctly identified positives, or, from patients who had an outcome, the proportion who tested positive. Specificity is the correctly identified negatives, or, from patients who had no outcome, the proportion of those who tested negative.

$$\text{specificity} = \frac{d}{d+b} \qquad \textit{Equation 2.23}$$

PPV is the proportion of those who had an outcome out of the patients who tested positive, or the probability of having an outcome if you test positive.

$$PPV = \frac{a}{a+b} \qquad \textit{Equation 2.24}$$

NPV is the proportion of those who did not have an outcome out of the patients who tested negative, or the probability of not having an outcome if you test negative.

$$NPV = \frac{d}{c+d} \qquad \textit{Equation 2.25}$$

Likelihood ratios are the probability of the predicted outcome to be observed in a patient with the same observed outcome, compared to a patient with the opposite observed outcome. More specifically, LR+ is the likelihood of a positive outcome predicted for a patient with a positive observed outcome, compared to a patient with a negative observed outcome, and LR- is the likelihood of a negative outcome predicted for a patient with a negative observed outcome, compared to a patient with a positive observed outcome (104).

$$LR+= \frac{\text{sensitivity}}{1-\text{specificity}} \qquad \textit{Equation 2.26}$$

$$LR-= \frac{1-\text{sensitivity}}{\text{specificity}} \qquad \textit{Equation 2.27}$$

Another method used for assessment of model performance is the receiver operating characteristic (ROC) curve, which can also be used to find the optimal threshold. ROC takes the predicted probabilities and estimates the resulting sensitivity and specificity for each possible threshold. The sensitivities are then plotted against 1-specificity using the corresponding specificity value from each treshold to create a curve. The optimal threshold, if it exsists, is the point on the curve with the highest sensitivity and lowest 1-specificity.

The area under the ROC curve (AUC) is used to assess the model performance based on the ROC curve. Plotting sensitivity vs 1-specificity, the area under the ROC curve for a model with

both sensitivity and specificity of 1 defines a square. The AUC value can also be seen as the proportion/percentage of this area which is covered by the area under the ROC curve of the model being assessed. An AUC of 1, or 100% is a model with both sensitivity and specificity of 1, a model like this would be capable of correctly identifying 100% of both positive and negative outcomes and be able to completely separate them. An AUC of 0.5, or 50% is a model describing chance, where the probability of each prediction being correct is 50% with no clear separation between positive and negative outcomes. AUC value between 50% and 100% are usually interpreted in groups of 10 (**Error! Reference source not found.** (105)) Each predicted probability is equally likely to be for a patient with an observed outcome or with no observed outcome.

Table 2.4: Interpretation of AUC values

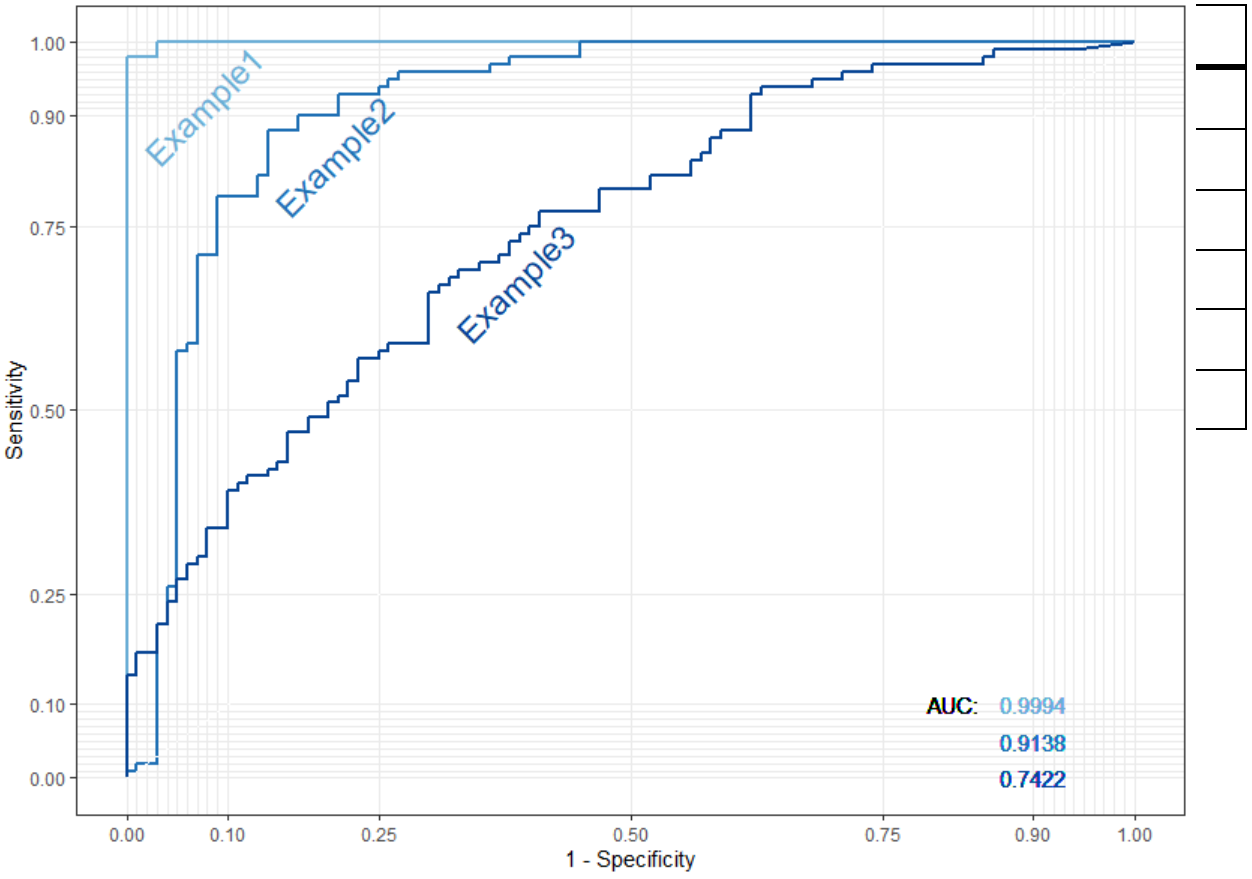

Figure 2.7: ROC curves with area under the curve for Examples in Figure 2.6

For sensitivity, specificity, PPV and NPV, "good" values depend on what we want the model to be able to achieve. Values between 0 and 1, or 0% and 100% are all possible and the higher the values, the better, but any of these values can be maximised at the threshold selection process at the expense of another. The higher sensitivity is, the lower specificity is, and vice versa. The higher PPV is, the lower NPV is, and vice versa. The optimal values depend on the intended use of the model.

Similarly, as likelihood ratios depend on sensitivity and specificity, they will take different values based on the threshold and can be optimised, however there is a general interpretation for LR+ and LR- values. Generally, the bigger LR+ is, the better the performance of the model, and the smaller LR- is, the better the performance of the model. The general interpretation of LR+ and LR- can be seen in Table 2.5 and **Error! Reference source not found.** (106).

Table 2.5: Interpretation of LR+ values

LR+	Increase in probability of outcome
2	Small (15%)
5	Moderate (30%)
10	Large (45%)

Table 2.6: Interpretation of LR- values

LR-	Decrease in probability of outcome
0.5	Small (-15%)
0.2	Moderate (-30%)
0.1	Large (-45%)

While ROC curves plot sensitivity against 1-specificity for each possible predicted probability threshold, precision-recall curves plot precision (number of true positives/number of all positive predictions) against recall/sensitivity (number of true positives/number of all positive observations) for all thresholds. High precision means that the model makes few false positive predictions, while high recall means that the model can identify a large proportion of all positive instances. Similarly to ROC curves, precision-recall curves assess the performance of binary classification, however, unlike ROC curves, precision-recall curves are more accurate and more sensitive to change in performance when the outcome rate is low. This is because when the outcome rate is low, it is possible to get high AUROC values just by predicting no outcome for all patients due to the high number of true negatives and low number of false negatives. In this case, a precision-recall plot may be more informative than an ROC plot as neither precision nor recall depend on the true negative value.

In regression-based modelling methods, coefficients can be used to make inferences of each variable's contribution to the predicted outcome in terms of magnitude and direction of change in the predicted probability. While these inferences only hold as long as underlying assumptions do as well, they still provide us with understanding of what the model is doing to create predictions. So called "black box" methods such as random forest do not provide us with this type of information, however we can use Shapley values to obtain some values which we can interpret similary. Shapley values come from cooperative game theory, created to calculate the contribution of each player in a cooperative game, by recording the pay-out of each different coalition of players(107). In machine learning, Shapley values similarly calculate the contribution of each feature to the prediction by calculating the average difference of the predicted probabilities for "coalitions" or value combinations of the features from the mean predicted probability(108). Shapley values are calculated for each value of each feature. A Shapley value of 0 means that on average, the feature taking on the specified value does not change the predicted probability from the average, while a positive Shapley value indicates an increase in predicted probability, and a negative value indicates a decrease.

The measurements mentioned above mostly assess the model's discrimination – its ability to classify observations into the correct outcome group. Another aspect of a model that can and should be assessed is calibration. Calibration refers to the accuracy of the predicted risk, rather than the predicted outcome, compared to the observed outcome. Similarly to discrimination, there are many tools that can be used to assess calibration.

The first and simplest method to use is calibration-in-the-large (109, 110). This method looks at the mean predicted probability and compares it with the prevalence of the outcome in the dataset. Perfect calibration-in-the-large is an absolute difference of 0, the larger the absolute difference, the worse the calibration is. The direction of difference is also informative - a mean predicted probability greater than the prevalence suggests that the model overestimates the risk, while a mean predicted probability smaller than the prevalence suggests that the model underestimates the risk (110).

The second and most popular method of assessing calibration is estimating the Cox calibration intercept and slope (109-112). This can be calculated by Equation 2.28 (111), where outcome is P(outcome=1), intercept and slope are the Cox calibration intercept and slope, and p is the predicted log odds.

$$logit(\text{outcome}) = \text{intercept} + \text{slope} * p \qquad \textit{Equation 2.28}$$

The Cox calibration intercept is another measure of calibration-in-the-large, representing an overall bias in the predicted risk. The Cox calibration slope is a measure of spread, representing variability in the predicted probabilities. Perfect calibration is a Cox calibration intercept of 0 and slope of 1. It should be noted that while the Cox calibration intercept can be interpreted on its own as a measure of calibration-in-the-large, the slope on its own does not measure calibration (112). As the slope measures the spread of the predicted probabilities, a slope <1 can be interpreted as the predicted probabilities varying too much compared to the observed outcomes, and a slope >1 as the predicted probabilities not varying as much as the observed outcomes. A slope of 1 can be observed along with both good and poor calibration-in-the-large.

Cox calibration fits a straight line through the predicted probability and observed proportion of outcomes pairs, which is useful for overall calibration, but does not show us in what ranges of predicted probabilities our model performs worse. There is also a chance of the model overestimating risk in one area and underestimating risk in another resulting in a good Cox calibration. We are able to assess if the model performs equally well or poorly across all predicted probabilities by obtaining a flexible calibration curve using loess (a nonparametric regression method) or spline (a piecewise function of polynomials) functions (110).

The Hosmer-Lemeshow goodness-of-fit test is another widely used option to measure calibration, however its use is recommended against as it often lacks power and its result is generally difficult to interpret (109-111). The test statistic is calculated by splitting the observations into arbitrary groups either based on equal number of observations per group, or equal increments of predicted probability. This arbitrary choice of grouping is also seen as a disadvantage of using the Hosmer-Lemeshow test. The test statistic is then calculated by Equation 2.29 (111), where g is the number of groups, $E_{s,i}$ is the number of expected outcomes in group *i*, $O_{s,i}$ is the number of observed outcomes in group *i*, and $O_{f,i}$ and $E_{f,i}$ are the observed and expected number of observations in group *i* with no outcome.

$$test\ statistic = \sum_{i=1}^{g} \left[\frac{\left(O_{s,i} - E_{s,i}\right)^2}{E_{s,i}} + \frac{\left(O_{f,i} - E_{f,i}\right)^2}{E_{f,i}} \right] \qquad \textit{Equation 2.29}$$

Finally, derived from the Brier score (a measure of the mean squared error), the Spiegelhalter z test allows us to calculate a z statistic for calibration, which then can be tested to show if the model is statistically significantly improperly calibrated (111). The Spiegelhalter z statistic, Z(E,O), is calculated by Equation 2.30 (111) where N is the number of observations, O_i is the observed outcome for observation *i*, and E_i is the predicted probability for observation *i*. If |Z(E,O)| is greater than the (1-α/2)-quantile of the normal distribution, where α is the significance level, the model is statistically significantly improperly calibrated.

$$Z(E,O) = \frac{\sum_{i=1}^{N}(O_i - E_i)(1-2E_i)}{\sqrt{\sum_{i=1}^{N}(1-2E_i)^2 E_i(1-E_i)}} \qquad \textit{Equation 2.30}$$

2.2.9.2 Risk group prediction

Previously we have discussed assessing the model performance for a two-group outcome prediction, namely, outcome or no outcome. In a clinical setting, however, a prediction like this would only be useful if we could perfectly discriminate between patients with and without an outcome. In a more realistic setting, for a prediction to be clinically useful, we would need to be able to identify patients who are sufficiently high risk to need action or treatment, and patients who are low enough risk that we can be sufficiently reassured that they will not have an outcome. This would require a prediction of at least three risk groups – high enough risk, low enough risk, and moderate risk for those in-between. We could arbitrarily create a definition of what we consider as high and low enough risk for this purpose, or for more robust results, we can use likelihood ratios to determine risk groups.

Table 2.5 and **Error! Reference source not found.** define values for likelihood ratios that can be interpreted as small, moderate and large reduction/increase in risk. Using these, we can define five risk groups:

- Very low risk: the patients classed into this group compared to all other patients have LR-≤0.1, meaning a large reduction in risk
- Low risk: the patients classed in this group compared to all patients not in either the low or very low risk groups have 0.1<LR-≤0.2, meaning a moderate reduction in risk
- Very high risk: the patients classed in this group compared to all other patients have LR+≥10, meaning a large increase in risk

- High risk: the patients classed in this group compared to all patients not in either the high or very high risk groups have 5≤LR+<10, meaning a moderate increase in risk
- Moderate risk: all other patients

The probability threshold values of the predicted probability for each group should be defined on a testing dataset that is not used either for the model fitting or the validation.

To define the probability threshold:

1. All probabilities in 0.1% increments should be tested as a cut-off to split the predicted probabilities into a yes/no prediction, and LR- and LR+ should be calculated for each.
2. Very low risk probability threshold is defined as the highest probability for which LR-≤0.1, and very high risk probability threshold is defined as the lowest probability for which LR+≥10.
3. Next, all patients with predicted probability less than or equal to the probability threshold for very low risk should be removed and the negative likelihood ratios should be calculated in 0.1% increments from the threshold again. Low risk probability threshold is defined as the highest probability for which LR-≤0.2.
4. Similarly, patients with predicted probability greater than the probability threshold for very high risk should be removed from the whole testing set, and positive likelihood ratios should be calculated in 0.1% increments up to the threshold. High risk probability threshold is defined as the lowest probability for which LR+≥5.
5. Moderate risk is defined as predicted probability greater than the probability threshold for low risk and less than or equal to the probability threshold for high risk.

Once the thresholds are determined on a testing set, performance should be tested on a separate validation set not used either in development or testing.

2.2.9.3 Clinical utility – decision curve analysis

Decision curve analysis assesses clinical utility of a predictive model by calculating the Net Benefit of treating patients above each predicted probability threshold, compared to treating all patients or none. The net benefit is calculated for each probability threshold $p_{treshold}$ *by*

$$Net\ Benefit = \frac{TruePositiveCount}{n} - \frac{FlasePositiveCount}{n} * \frac{p_{treshold}}{1 - p_{threshold}}$$

where n is the total number of patients in the data, and the true and false positive counts for the model are determined by classing those with predicted probability greater than $p_{treshold}$ as predicted positive, and the rest as predicted negative 19. To calculate the Net Benefit for treating all, all patients are classed as predicted positive for every threshold, while no patients are classed as predicted positive for the Net Benefit of treat none. Decision curve analysis assesses clinical utility by showing at which predicted probability thresholds would the model be more beneficial than treating all patients or none, if we were to treat only the patients with predicted probability above the specified threshold.

2.2.10 Oversampling

Modelling an outcome is often easiest on a dataset where the number of observations with the outcome of interest is as close to perfectly balanced with the number of observations without the outcome as possible. In datasets like this, we have a 50% chance of guessing which outcome group an observation belongs to by just randomly choosing one group or the other. This 50% chance is what we then compare our model's performance to. Often statistical software is even programmed in such a way that if we do not specify a threshold to classify predicted probabilities into one of the outcome and no outcome groups, the software will use 0.5 as this threshold.

In the real world, however, datasets are often imbalanced, in some cases due to the fact that the outcome of interest is very rare in the population, and thus occurs only in a small percentage of the collected data points. In these cases, trying to gather more data points with the outcome could be very time consuming and costly, and discarding data points without the outcome until we achieve a balance might not leave us with enough data for the method(s) we intend to use. We could also run into problems when deciding which observations to discard, creating problems that could affect our models.

This is where systematic oversampling of the minority class comes in to play. Systematic minority oversampling technique (SMOTE) (113) is an algorithm that creates synthetic samples of the minority class. For each real sample of the minority class provided to the algorithm, it selects the *k* nearest neighbours of the sample and stores their indices. To generate a synthetic sample, a real sample and one from its *k* nearest neighbours is selected randomly. The difference between the sample value and the neighbour value is calculated for each variable in the sample, this difference is then multiplied by a random number between

0 and 1 and added to the original sample value. This process essentially shifts the original sample creates a new sample that is between the original and one of its nearest neighbours. The process is repeated until a sufficiently large dataset is created.

2.3 Methods

2.3.1 Data

I used prospectively-collected data from women with preeclampsia, broadly-defined according to the 2021 International Society for the Study of Hypertension in Pregnancy (ISSHP) criteria,(9) as women presented for initial facility-based assessment at centres with general policies of expectant management of preeclampsia remote from term.

The primary study outcome was a composite outcome developed by Delphi consensus,(114) and defined as the first occurrence of one or more of: maternal mortality or severe maternal morbidity (listed in Table 2.12 and defined in Table 2.11), within two days of first assessment for preeclampsia.

To maximise the sample size for machine learning, data were collated from published model development and validation studies for: (i) miniPIERS (2008-2012): N=2126 women from Brazil, Fiji, Pakistan, South Africa, and Uganda;(29) and (ii) fullPIERS development and validation (2003-2016): N=6717; Australia, Canada, Finland, New Zealand, United Kingdom, and United States (3, 41). More details of the studies are presented in Table *2.7*.

The data from each independent study was cleaned and variables were cross-referenced between studies. In case of difference in the coding or recording of variables, variables were modified or combined into new variables to create matches across all datasets. For example, some datasets used values such as -99 to indicate missing values while others used “NA”, some recorded headaches and visual disturbances in separate variables while some recorded them in a single variable for headache and/or visual disturbances, and some lab measurements were recorded with different units of measurement. In these cases, -99 was replaced with “NA” values, a new variable for headache and/or visual disturbances was created, and units of measurements were converted. The outcome variable also required cleaning as some studies recorded the components of the composite outcome in separate variables, some in individual patient notes as free text, and some included further outcomes which needed to be removed. The variables corresponding to the elements of the composite

outcome in Table 2.12 were selected in each dataset, along with any variables that recorded date and time of occurrence, and combined into two variables – one showing outcomes and one for dates. Outcome within two days was defined based on the first outcome for each patient. Following these initial steps of data cleaning, data from the studies were combined into a single dataset. Dates in the combined dataset were checked for typos to ensure only existing dates were used and there were no pregnancies recorded to last longer than physically possible – for example the pregnancy recorded to last 3 years due to an error in the year of one of the dates was corrected.

In addition, a second dataset was collected from a prospective observational cohort study, using the electronic health records of 2901 women with singleton pregnancies who were admitted with a diagnosis of preeclampsia (ISSHP definition) to King's College Hospital, London, and Medway Maritime Hospital, Gillingham, United Kingdom, between December 2013 and December 2021. Data cleaning steps were repeated for the external validation dataset to ensure the structure and coding of variables matched the internal data.

All women gave written informed consent to participate in the study, which was conducted according to the guidelines of the Declaration of Helsinki, and approved by the NHS Research Ethics Committee (REC reference: 02-03- 033 on 11 March 2003).

Data was split into a development and a validation dataset by randomly selecting 75% of the data for development. Missing data imputations was handled separately for the two datasets, using the same methods for each variable in both cases, as seen in Chapter 2.3.2. After imputation, validation was further split into two datasets, testing and validation for risk group prediction.

Models were fitted on the imputed development data using logistic regression, random forests, Least Absolute Shrinkage and Selection Operator (LASSO) and Bayesian Model Averaging (BMA). Variables were considered for modelling only if assessed prior to the occurrence of any component of the combined adverse maternal outcome. Variables detailed the woman's health system, and her demographics, past and current medical and obstetric history, and relevant symptoms, signs, and laboratory tests. A full list of predictor variables can be seen in Table *2.8*.

Table 2.7 Summary of the studies included

Variable	fullPIERS	miniPIERS	BCW	FINNPEC	Oxford	PETRA	PREP
Number of participants	2011	2126	1956	199	291	1244	1096
Number of participants with outcome in 2 days	111 (5.5%)	277 (13%)	101 (5.2%)	15 (7.5%)	20 (6.9%)	30 (2.4%)	36 (3.3%)
Number of participants with outcome at any point	226 (11.2%)	383 (18%)	158 (8.1%)	49 (24.6%)	46 (15.8%)	60 (4.8%)	161 (14.7%)

Variable	fullPIERS	miniPIERS	BCW	FINNPEC	Oxford	PETRA	PREP
Aim	"Identifying the risk of fatal or life-threatening complications in women with preeclampsia within 48h of hospital admission"	Reduce triage delays for women with any hypertensive disorders of pregnancy in low- and middle-income countries	Validation of the fullPIERS model	Set up a nationwide clinical and DNA database on women with and without preeclampsia to identify genetic risk factors for preeclampsia	Validation of the fullPIERS model	Clarify the temporal relationship between the measured angiogenic factors and the time to delivery in women with suspected preeclampsia at <35 weeks gestation	Provide individual risks of adverse maternal outcomes, including delivery of preterm infant before 34 weeks, in women with early-onset preeclampsia in the UK, by 48 hours and by discharge
Inclusion criteria	Admitted with or developed preeclampsia after admission	Any hypertensive disorder of pregnancy	Unselected cohort of women admitted with any form of pregnancy hypertension	Singleton pregnancy, diagnosis of preeclampsia	Unselected cohort of women who presented with preeclampsia before 36^{+0} weeks of gestation	18-45 years old, any signs or symptoms of preeclampsia, between 20 and 41 weeks gestation	Suspected or confirmed preeclampsia before 34 weeks' gestation
Exclusion criteria	Admitted in spontaneous labour, or had	Admitted in spontaneous labour, had		Multiple pregnancy, maternal age			Occurrence of outcome prior to the

Variable	fullPIERS	miniPIERS	BCW	FINNPEC	Oxford	PETRA	PREP
	any component of the outcome before eligibility or data collection	any component of the outcome before eligibility or data collection, or had confirmed positive HIV/AIDS status with CD4 count ,250 cells/ml or AIDS-defining illness		under 18, or inability to provide informed consent in Finnish or Swedish			assessment of predictors, insufficient time to obtain informed consent or lack of translator for non-English speakers
Countries	Australia, Canada, New Zealand, UK	Brazil, Fiji, Pakistan, South Africa, Uganda	Canada	Finland	UK	North America	UK

2.3.2 Imputation of missing values

Variables were considered for modelling only if assessed prior to the occurrence of any component of the combined adverse maternal outcome. Variables detailed the woman's health system, and her demographics, past and current medical and obstetric history, and relevant symptoms, signs, and laboratory tests (Table *2.14* and Table *2.15*). For face validity, at least one objective variable was required for each of cardiorespiratory, renal, hepatic, and haematological organ systems.

To deal with missing data, we need to know the type of missingness (MCAR, MAR or MNAR) and the percentage of missingness. For our data, the type of missingness was determined based on clinical input on how the data was or was likely to have been collected. As there was a protocol in place for data collection, all data would have been measured regularly if it was possible and all variables would have been recorded, including values within the normal range. Hence, we assume that all missing observations were due to lack of access to equipment to measure (such as fingertip oximeters or laboratory equipment), or data was missing by chance, and data MNAR was ruled out. It should be noted that this is an assumption only and not a statistical test for mechanism of missingness. To test if data in any given variable were missing completely at random, I created a corresponding missingness variable coded 0 if there is a value observed in the given variable and 1 if missing. A t-test or a chi-square test (dependent on data type) was then performed on the missingness variable with all other predictor variables. If any of these tests is significant, the mechanism of missingness for the given variable would be missing at random, and missing completely at random if none is significant. (61) All variables had at least one significant t- or chi-square test, meaning that all variables were missing at random.

Table *2.8* shows all potential predictor variables, a summary of missingness, and the action taken. A more detailed overview of the missingness can be seen in Table *2.13*.

Table 2.8 Predictor variables, summary of missingness and action taken

Variable	Missingness	Action
National *per capita* gross domestic product – USD	MAR or MCAR, <60% missing	Multiple Imputation
National MMR– maternal deaths per 100,000 live births		
Ethnicity – no. (%)		
Maternal age at expected date of delivery – years		
Nulliparous – no. (%)		
Multiple pregnancy – no. (%)		
Gestational age at eligibility – weeks		
Cigarette smoking		
Chronic hypertension		
Pre-gestational renal disease		
Pre-gestational diabetes		
Gestational diabetes		
Nausea or vomiting		
Headache or visual disturbance		
Right upper quadrant or epigastric pain		
Chest pain or dyspnea		
Height on admission – cm		
Weight on admission – kg		
Systolic blood pressure – mm Hg		
Diastolic blood pressure – mm Hg		
Oxygen saturation – %		
Dipstick proteinuria – '+'		
Hematocrit – %		
Total leucocyte count – x 10^9 per L		
Platelet count – x 10^9 per L		
Mean platelet volume – fL		
Fibrinogen – g per L		
Activated partial thromboplastin time – sec		
Serum creatinine – sec		
Uric acid – mmol per L		
Aspartate transaminase – units per L		
Alanine transaminase – units per L		
Albumin – g per L		
Lactate dehydrogenase – units per L	≥60% missing	Dropped
International normalized ratio		
Total bilirubin		
Protein:creatinine ratio		

For all variables collected multiple times over the pregnancy, only one value per day used. If multiple values were available for any day for a patient, the clinically worst value was taken. Table 2.9 shows the list of variables that were collected repeatedly (not including variables that were dropped), and the definition of "worst" for each. This step was done based on clinical input to reflect clinical practice as some levels and measurements can vary throughout a day.

We opted for imputing variables with missing values only if the level of missingness was 60% or under. Hence, for multiple imputation, Total bilirubin, Protein:creatinine ratio, International normalised ratio and Lactate dehydrogenase were dropped. Data from the day of first admission was selected for each patient and the worst measurement was taken for each variable as per Table 2.9. As we had a mixture of numeric and categorical predictor variables, multiple chained random forests was selected as the method of imputation. I used the `missRanger` R package for imputation, imputing all remaining predictor variables, using all remaining predictor variables as auxiliary variables.

Table 2.9: Repeated variables

Variables	**Worst value**
Headache or visual disturbance Nausea or vomiting Right upper quadrant or epigastric pain Chest pain or dyspnoea	"Yes"
Oxygen saturation – % Platelet count – x 10^9 per L Fibrinogen – g per L Serum albumin – g per L Random glucose	Minimum
Systolic blood pressure – mm Hg Diastolic blood pressure – mm Hg Total leucocyte count – x 10^9 per L Mean platelet volume – fL	Maximum

Haematocrit Activated partial thromboplastin time – sec Serum creatinine – sec Uric acid – mmol per L ALT– units per L AST– units per L Dipstick proteinuria – '+'	

Missing values in the development and validation sets were imputed separately using all other predictor variables except Patient ID and Outcome. Missing values were estimated with 20 iterations of a random forest of 100 trees from the non-missing values of the data. Each random forest used weighted observations where the weights were proportionate to the completeness of the observation, i.e. rows with an observed value for all variables will have the highest weight, while rows with all variables missing will have the lowest weight so that more complete data has a bigger influence on the imputation. Predictive mean matching from a random sample of 5 non-missing values for each imputed data point is used to make sure it is in the same format and range as observed in the data.

In our case, as ≈19% of all data were missing, development and validation datasets were imputed 20 times each. To assess the impact of imputation on our model performance, I conducted a complete case analysis of the final model.

I compared summary measurements of the predictor variables between women with and without an outcome before imputation and after imputation to make sure that multiple imputation did not fundamentally change the relationships between the two patient groups.

For each method, models were fitted on each of the complete static development sets. To make predictions on the multiply imputed validation dataset, the 20 datasets were fed into each of the 20 models created on the development dataset, creating a total of 400 predictions per patient, the mean of which was taken to create the output prediction for the patient.

2.3.3 Logistic regression

For comparison with the rest of the models, first the original fullPIERS logistic regression model (O_fP) was recreated by taking the formula from the publication. The model was not fitted on a dataset and no changes were made to it, the variables, interactions, and exact coefficients were as seen in Equation 1.1.

As the dataset the original model was fitted on was a lot less diverse by being limited to high-income countries, and for the fitting the worst value in the first 48 hours rather than day of admission only was used, I created a second logistic regression model, the refitted fulPIERS model (R_fP). This model also took the variables and interactions from the original fullPIERS model; however, it was fitted on the development datasets using the `caret` package in R for 10-fold cross-validation using the ROC value to assess models, the "glm" method and the `caretList` function from the `caretEnsemble` package to create a list of models from the multiply imputed datasets.

2.3.4 Random forest

All random forest models were fitted using the `caret` package in R with 10-fold cross-validation using the ROC value to assess models and the "rf" method. For model fitting on multiply imputed datasets, the `caretList` function was used from the `caretEnsemble` package.

First, a random forest model was created using all variables, labelled as RF1.

Using variable importance values from RF1, a model using variables with above average importance only (RF2) was fitted. This was followed by models using RFE (RF3), Boruta (RF4) and Vita (RF5). Finally, a random forest model using the fullPIERS variables only (RF6) was also created. For RF2, RF3 and RF4, all variables that were selected as important in at least one of the 13 datasets were retained.

2.3.5 Gradient boosted trees

Gradient boosted trees were fitted the same way as random forest, except the method used was "xgbTree". A model with all variables, XGB1, was created first, variable importance values were calculated, and a model using variables with above average importance only (XGB2) was fitted. No other feature selection methods were used.

2.3.6 LASSO and Ridge regression

Similarly to random forest, LASSO and ridge regression models were fitted using the `caret` package with 10-fold cross-validation using the ROC value to assess models and the "glmnet" method. For model fitting on multiply imputed datasets, the `caretList` function was used from the `caretEnsemble` package.

For both LASSO and Ridge models, a list of lambdas was tested between 10^{-3} and 10^{3} by taking a 100 elements long equally distanced sequence from -3 to 3 and raising 10 to the power of each element. To create a LASSO model using `glmnet`, alpha was set to be 1, while to create Ridge, alpha was set to be 0. Two LASSO models were created, L1 fitted using all variables, allowing the method to shrink the coefficients of the unimportant variables to 0, and L2 using fullPIERS variables. There was only one Ridge model created, using all variables, as all fullPIERS variables were important in L2 meaning no variable selection.

2.3.7 Bayesian Model Averaging

A BMA model fitted using the `bic.glm` function in R. From this, variables with probability of having a non-negative coefficient greater than 0% were selected and the model was fitted again using these variables only. As this is a Bayesian method, there was no need for cross-validation as the method estimates the underlying probability distributions as opposed to fitting to the observed data only (115). With dfactor.type set to "false", `bic.glm` turns all factors into dummy variables and treats them independently for variable selection meaning that for factors with more than 2 levels, it is possible for some levels to be selected as important while some levels as not important.

2.3.8 Artificial Neural Networks

For artificial neural networks, similarly to before, the `caret` package with 10-fold cross-validation using the ROC value to assess models and the "nnet" method was used. For model fitting on multiply imputed datasets, the `caretList` function was used from the `caretEnsemble` package. The number of nodes in the hidden layer was provided in the size option as an integer between 1 and 10, and weight decay was given in the decay option as a list of a sequence from 0.1 to 1 in steps of 0.1, and 10 to the power of -7 to -2. All possible combination of size and decay was tested on all imputed datasets and the combination with the lowest cross-validation error per dataset was selected.

Three neural networks models were created. First, a model using all available variables (NN1). Using importance measures from NN1, a model using only variables with above average importance (NN2) was fitted. Finally, a model using only fullPIERS variables was created (NN3).

2.3.9 Fit and assessment

A randomly-selected 75% of the combined cohort was used for model development, 12.5% were used to select thresholds for risk strata, and 12.5% reserved for model validation, which was completely unseen by the model during training. The model was externally validated on the second dataset of 2901 women.

I assessed the models' ability to classify women into outcome/no outcome groups, using the area-under-the-receiver-operator characteristic (AUROC), calibration, and precision-recall curves. Decision curve analysis was also carried out to assess clinical utility. Likelihood ratios used to determine risk strata were data-defined, as follows: very-low risk (by a negative likelihood ratio <0.10), low risk (negative likelihood ratio of 0.1 to 0.2), high risk (positive likelihood ratio of 5.0 to 10.0), very-high risk (positive likelihood ratio greater than 10.0), and moderate risk.(116) The number and percentage of patients classified into each group was presented along with the number and percentage patients with outcome within 2 days within each group. Using the same threshold values and predicted probabilities, the likelihood ratios were calculated for outcome within 7 days and at any point with a 95% confidence interval, and presented along with the above results, as well as the number and percentage of outcomes within 7 days and at any point within each group.

2.3.9.1 Sensitivity analyses

In sensitivity analyses, new datasets were created where coagulation-related variables were excluded (as they are costly in all health systems), and mean platelet volume was excluded (as it is not routinely reported by all haematology laboratories). Modelling steps were repeated on these datasets and the results were compared.

Secondary analyses were undertaken to predict outcomes: (i) at either seven days or at any time following admission until primary hospital discharge (outcomes within seven days included patients who had outcomes within two days; outcomes at any time included outcomes within two and seven days), and (ii) limited to eclampsia or stillbirth. The first

secondary analysis was carried out for all models and were reported along with the primary analysis results, while secondary analysis of eclampsia and stillbirths was only carried out for the selected final model (referred to as the PIERS-ML model).

As some renal and haematological measures were included in both the candidate variables and the definitions of components of the combined maternal outcomes, I assessed the performance of the final model in women who experienced no renal, no haematological, or neither renal nor haematological outcomes. The influence of multiple imputation on predictions was assessed by complete case analysis and mean imputation.

2.3.9.2 External validation

I assessed the stratification accuracy of the PIERS-ML model in the external validation cohort. Following selection of the variables in the PIERS-ML model, the external validation dataset had ≈19% missingness. These missing values were imputed 20 times, independently of the existing combined PIERS-ML dataset. The PIERS-ML model was applied to each of the 20 imputed datasets and the mean prediction per woman was taken. Oxygen saturation was not collected from women in the external validation cohort, so it was assumed to be normal and replaced uniformly by 97%, the expected normal measurement. Data for some less-common components of the combined adverse maternal outcome were not available in women's electronic health records, and some other components were conflated into summary measures.

2.3.9.3 Systematic oversampling

As the outcome rate in our combined dataset was very low, I also assessed the performance of the final model using systematic oversampling of the data points with an outcome. This was done using the `smote` function in the `performanceEstimation` package (i) on the validation dataset, to which the PIERS-ML model was applied and performance measures were recorded, and (ii) on the development dataset, which was used to refit the PIERS-ML model on a more balanced dataset, following which the original validation dataset was used to obtain performance measurements.

2.4 Results

2.4.1 Outcome choice

The outcomes included in the outcome list were derived by the Delphi consensus. Experienced midwives, obstetric physicians and obstetric anaesthetists compiled a list of

outcomes they wished to avoid, rather than wait for to start treatment. These outcomes form the components of the composite outcome. Because of this criteria, HELLP syndrome is not a component of the outcome as between 24 and ≈34 weeks gestation, clinicians will await the occurrence of HELLP to warrant a response, rather than trying to prevent it. Therefore, the composite outcome includes hepatic failure, haematoma and rupture instead. These outcomes represent a range of severity, from blood transfusion to maternal death. This combined outcome has been used in other clinical trials as well (2-4). The composite outcome used for PIERS-ML is also very similar to the iHOPE core outcome set, which was developed subsequent to the PIERS Delphi consensus.

Table 2.10: Occurrence of outcomes by outcome group

Outcome group	Outcome in 2 days	Outcome in 7 days	Outcome at any point
Maternal mortality	0 (0%)	1 (0.01%)	2 (0.02%)
Central nervous system	65 (0.74%)	85 (0.96%)	102 (1.15%)
Cardiorespiratory	140 (1.58%)	207 (2.34%)	269 (3.04%)
Haematological	293 (3.31%)	413 (4.67%)	531 (6%)
Hepatic	24 (0.27%)	36 (0.41%)	53 (0.6%)
Renal	39 (0.44%)	50 (0.57%)	61 (0.69%)
Other	149 (1.68%)	218 (2.47%)	333 (3.77%)

The incidence of each outcome separately is quite low, as seen in Table 2.10. Separating them into the outcome groups, most groups still do not reach even 1% incidence, therefore would be very difficult to create an accurate model for as if a model predicted no outcome for all participants, it would be accurate for over 99% of patients. Hence, for our investigation, the target outcome was an occurrence of an outcome belonging to any outcome group. Table *2.12* shows the occurrence of the adverse maternal outcomes (defined in Table *2.11*) for the internal data used for development, testing and validation, and the external validation dataset.

Table 2.12 shows that 590 women (6.7%) had an adverse maternal outcome within two days of first assessment, 813 (9.2%) within seven days, and 1083 (12.2%) at any time prior to primary discharge. Most adverse outcomes were cardiorespiratory, haematological, hepatic, or placental. There were two maternal deaths, neither within 48 hours of first assessment. In the external validation cohort, 83 women (2.9%) had an adverse maternal outcome within two days, 99 (3.4%) within seven days, and 121 (4.2%) at any time prior to primary discharge. Most adverse outcomes were central nervous system, haematological or placental.

Table 2.11: Definitions of Race and Adverse Maternal Outcomes

Outcome	**Definition**
Race	
White	Origins in Europe, Middle East, North Africa [Arabic origins], Western Russia [including Afghanistan and South Russia] and Hispanics of European origin
Asian	Origins in the Indian sub-continent [e.g., India, Pakistan, Bangladesh, and Sri Lanka], or in the Far East and Southeast Asia [e.g., China, Japan, Korea, Philippines, Thailand, Eastern Russia]
Black	Origins in any of the original peoples of Africa
Other	Including mixed ancestry
Mortality	Maternal death occurring within six weeks of pregnancy or if later, attributable to complications of preeclampsia
Hepatic dysfunction	International normalised ratio (INR) >1.2 in the absence if disseminated intravascular coagulation (DIC) or treatment of warfarin (DIC is defined as having both: abnormal bleeding and consumptive coagulopathy [i.e., low platelets, abnormal peripheral blood film, or one or more of the following: increased INR, increased prothrombin time (PTT), low fibrinogen, of increased fibrin degradation products that are outside normal non-pregnancy ranges])
Hepatic hematoma or rupture	Blood collection under the hepatic capsule as confirmed by ultrasound or laparotomy
Glasgow coma score (GCS) <13	Based on GCS scoring system: Teasdale G, Jennet B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; **2**:81-83
Stroke	Acute neurological event with deficits lasting longer than 48 hours

Outcome	Definition
Cortical blindness	Loss of visual acuity in the presence of intact papillary response to light
Reversible Ischaemic Neurologic Deficit (RIND)	Cerebral ischaemia lasting longer than 24 hrs but less than 48 hours revealed through clinical examination
Retinal detachment	Separation of the inner layers of the retina from the underlying retinal pigment epithelium (RPE, choroid) and is diagnosed by ophthalmological exam
Acute renal insufficiency	For women with no underlying renal disease, defined as serum creatinine >150 µM
Acute renal failure	For women with an underlying history of renal disease, defined as serum creatinine >200 µM
Dialysis	Including haemodialysis and peritoneal dialysis
Postpartum haemorrhage (PPH) requiring transfusion or hysterectomy	Occurrence of PPH that required transfusion or hysterectomy
Placental abruption	Any occurrence of abruption diagnosed clinically or based on placental pathology report
Platelet count < 50,000 x 10^9/L without blood transfusion	Measurement of platelet count recorded as less than 50,000 x 10^9/L without patient receiving a blood transfusion
Transfusion of blood products	Includes transfusion of any units of blood products: fresh frozen plasma (FFP), platelets, red blood cells (RBCs), cryoprecipitate (cryo) or whole blood
Positive inotropic support	The use of vasopressors to maintain a systolic blood pressure >90 mmHg or mean arterial pressure >70 mmHg
Myocardial ischaemia/infarction	Electrocardiogram (ECG) changes (ST segment elevation or depression) without enzyme changes AND/OR any one of the following: 1) Development of new pathologic Q waves on serial ECGs. The patient may or may not remember previous symptoms. Biochemical markers of myocardial necrosis may have normalised, depending on the length of time

Outcome	Definition
	that has passed since the infarct developed. 2) Pathological findings of an acute, healed or healing MI 3) Typical rise and gradual fall (troponin) or more rapid rise and fall (CK-MB) of biochemical markers of myocardial necrosis with at least one of the following: a) ischaemic symptoms; b) development of pathologic Q waves on the ECG; c) ECG changes indicative of ischaemia (ST segment elevation or depression); or d) coronary artery intervention (e.g., coronary angioplasty)
Eclampsia	Any episode of seizure antepartum, intrapartum or before postpartum discharge as follow-up beyond discharge is not possible
Require >50% oxygen for greater than one hour	Oxygen given at greater than 50% concentration based on local criteria for longer than 1 hour
Intubation other than for Caesarean section	Intubation by endotracheal tube
Severe breathing difficulty	Suspected pulmonary oedema where x-ray confirmation unavailable may be diagnosed by presence of chest pain or dyspnoea, crackles in the lungs and SaO_2 <90%
Pulmonary oedema	Clinical diagnosis with x-ray confirmation or requirement of diuretic treatment and SaO_2 <95%

Table 2.12: Occurrence of Adverse Maternal Outcomes Following First Assessment, by Mortality or Morbidity Event

Outcome	**Internal combined data (N=8843)**			**External validation (N=2901)**		
	Within two days (N=590)	**Within seven days (N=813)**	**At any time (N=1083)**	**Within two days (N=83)**	**Within seven days (N=99)**	**At any time (N=121)**
Maternal death	0	1	2	1	1	1
Central nervous system						
Eclamptic seizure(s)	49	63	75	22	22	25
Glasgow coma score less than 13	14	15	20	Not collected	Not collected	Not collected
Stroke or reversible ischaemic neurological deficit	4	5	6	0	0	0
Transient ischaemic attack	0	1	1	Not collected	Not collected	Not collected
Cortical blindness	3	5	6	1	1	1
Posterior reversible encephalopathy	4	5	5			
Cardiorespiratory						
Positive inotropic support required	3	4	7	Not collected	Not collected	Not collected
Infusion of a third injectable antihypertensive	17	28	33	Not collected	Not collected	Not collected
Myocardial ischaemia or infarction	4	5	6	2	2	2

Outcome	**Internal combined data (N=8843)**			**External validation (N=2901)**		
	Within two days (N=590)	**Within seven days (N=813)**	**At any time (N=1083)**	**Within two days (N=83)**	**Within seven days (N=99)**	**At any time (N=121)**
Oxygen saturation less than 90%	52	77	103	2*	2*	3*
At least 50% fractional inspired oxygen for at least one hour	29	50	72			
Intubation other than for Caesarean birth	23	34	47			
Pulmonary oedema	55	82	96	2	3	3
Haematological						
Blood transfusion	242	347	460	29	36	53
Platelet count less than $50x10^9$ per L, without transfusion	83	103	111	5	8	9
Hepatic						
Dysfunction	23	30	44	5	5	5
Haematoma or rupture	0	0	0	2	2	2
Renal						
Acute renal insufficiency in women without	3	5	9	8†	8†	8†

Outcome	Internal combined data (N=8843)			External validation (N=2901)		
	Within two days (N=590)	**Within seven days (N=813)**	**At any time (N=1083)**	**Within two days (N=83)**	**Within seven days (N=99)**	**At any time (N=121)**
chronic kidney disease						
Acute renal failure in women with chronic kidney disease	36	45	52			
Dialysis	2	7	11	0	0	0
Other						
Placental abruption	75	98	129	17	25	29
Severe ascites	30	50	65	0	0	0
Bell's Palsy	3	3	6	0	0	0

***** Respiratory failure (pulmonary oedema accompanied by severe hypoxaemia with need for intubation or mechanical ventilation); † acute kidney injury (serum creatinine >2 mg/dL [>176.8 mM]). All other outcomes defined in Table S2 in the Supplementary Appendix

2.4.2 Data, missingness and imputation

Table 2.13 shows the breakdown of missing data between women with and without adverse outcomes. Women who had an outcome had higher rates of missingness of GDP per capita, MMR, maternal age, parity, multiple pregnancy, all variables for past and current medical and obstetrical history, all variables for signs on day of admission bar dipstick proteinuria, haematocrit, platelet count, mean platelet volume, uric acid, AST, ALT, and albumin. They also had lower rates of missingness of right upper quadrant or epigastric pain, chest pain or dyspnoea, and dipstick proteinuria.

Data were not missing equally between studies (Table 2.13). Race was more likely to be missing in the Vancouver cohort, the fullPIERS cohort, and the Oxford cohort. Symptoms were

rarely recorded within a day of first admission in the PETRA dataset and, except for right upper quadrant pain, in the PREP dataset. Blood pressure and oxygen saturation were missing for >70% on day of admission in the PREP data, however blood pressure was often recorded after day of admission, with 90% of women in the PREP data having blood pressure recorded on or the day after of admission. The highest rate of missingness was reported for the laboratory tests, some tests missing entirely from some datasets. Fibrinogen, activated partial thromboplastin time, aspartate transaminase and albumin were rarely measured in the FINNPEC data. Mean platelet volume and albumin were rarely measured in the miniPIERS data. In the Oxford data, haematocrit was not measured; fibrinogen and activated partial thromboplastin time were rarely measured. Total leucocyte count, mean platelet volume, and albumin were not measured, and fibrinogen and activated partial thromboplastin time were rarely measured in the PETRA data. Haematocrit and mean platelet volume were not measured in the PREP data. The fullPIERS data had the least amount of missing data.

While rates of missingness were different between datasets and not all datasets recorded all variables, we assumed that patients within all cohorts were similar enough in their presentation that their observed and missing values for each variable would not be significantly different, or any potential difference can be accounted for by the other observed variables (such as gestational age). Datasets were combined on this assumption to create the data for model development and internal validation.

Data was available for 8843 eligible women with preeclampsia, recruited from 53 institutions in 11 countries. Other cohort details have been published previously (3, 29, 41). For external validation from two new institutions, data were available from an additional 2901 women, and collected as part of a prospective observational study.

Table 2.15 shows that the health system and individual-level characteristics after imputation of women who experienced an adverse maternal outcome differed from women who did not. Women who experienced adverse outcomes were more often cared for in countries with lower per capita gross domestic products. These women were younger, more likely to have a multiple pregnancy, and presented at an earlier gestational age; less often, their past history was complicated by chronic hypertension and their pregnancies by gestational diabetes. The study population had similar proportions of women from White, Asian, or Black ethnic backgrounds, regardless of complications.

At presentation with preeclampsia, women who subsequently developed an adverse maternal outcome (vs. those who did not) differed in their symptom profile, signs, and laboratory results, although results largely overlapped. These women were more often symptomatic, had lower weight, higher blood pressure, lower oxygen saturation, higher dipstick proteinuria, and more perturbed laboratory results; differences in the latter were not changed significantly by imputation. Table 2.14 shows the summary of predictor variables for women with and without an adverse outcome at any point, with summary measures calculated before imputation ignoring missing values. In the internal validation data, the difference between the two groups in total leucocyte count was not significant (p-value of 0.593), however this difference became highly significant (p-value <0.001) after imputation. All other variables that were significantly different between the groups remained significantly different, and those that were not significantly different remained so after imputation. Symptom variables, fibrinogen and activated partial thromboplastin time were not imputed in the external validation dataset as they were not included in the final chosen model. Of the remaining variables, National maternal mortality ratio was not significantly different between the two groups (p-value 0.128), but became significant after imputation (p-value 0.02), while haematocrit and uric acid were both only close to being significant before imputation (p-values 0.068 and 0.076 respectively), after imputation haematocrit became significantly different but close to the 0.05 significance level (p-value 0.043) and uric acid became clearly significant (p-value <0.001). Other variables did not change from significant to not significant or vice versa.

Looking at treatments received and other pregnancy outcomes shown in Table 2.15, women who subsequently developed an adverse outcome more often received antenatal corticosteroids, antihypertensives, and magnesium sulphate. Babies of women who experienced adverse outcomes were born earlier and of lower birthweight, and more often died.

Table 2.13: Breakdown of Missing Data

Variable	**Missing (%)**									
	Vancouver N=1956	**FINNPEC N=119**	**fullPIERS N=2011**	**miniPIERS N=2126**	**Oxford N=291**	**PETRA N=1244**	**PREP N=1096**	**Women with adverse outcome at any time after first assessment (N=1083)**	**Women without an adverse outcome (N=7760)**	**Total**
Health system										
National per capita gross domestic product (USD)	0.00	0.00	0.40	0.47	1.03	0.00	0.00	1.94	0.00	0.24
National maternal mortality ratio (maternal deaths per 100,000 live births)	0.00	0.00	0.40	0.47	1.03	0.00	0.00	1.94	0.00	0.24
Demographics										

Variable	Missing (%)									
	Vancouver N=1956	FINNPEC N=119	fullPIERS N=2011	miniPIERS N=2126	Oxford N=291	PETRA N=1244	PREP N=1096	Women with adverse outcome at any time after first assessment (N=1083)	Women without an adverse outcome (N=7760)	Total
Race	26.12	0.84	19.69	1.69	30.58	0.64	0.55	10.53	12.02	11.84
Maternal age at expected date of delivery (years)	0.10	0.84	0.40	1.18	1.72	0.00	0.18	2.31	0.23	0.49
Nulliparous	0.15	0.84	0.40	0.52	1.03	0.00	0.00	2.12	0.04	0.29
Multiple pregnancy	0.00	0.84	0.40	0.52	1.03	0.24	0.00	2.03	0.05	0.29
Gestational age at eligibility (weeks)	0.10	0.00	0.00	0.61	0.69	0.00	0.00	0.18	0.19	0.19
Past and current medical and obstetrical history – no. (%)										
Cigarette smoking	1.48	2.52	5.47	2.45	2.75	0.40	0.00	4.06	2.10	2.34
Chronic hypertension	0.87	0.84	0.65	4.23	2.75	0.80	1.09	3.51	1.46	1.71

Variable	Missing (%)									
	Vancouver N=1956	FINNPEC N=119	fullPIERS N=2011	miniPIERS N=2126	Oxford N=291	PETRA N=1244	PREP N=1096	Women with adverse outcome at any time after first assessment (N=1083)	Women without an adverse outcome (N=7760)	Total
Pre-gestational renal disease	0.92	0.84	0.90	4.42	3.09	0.40	1.00	3.60	1.51	1.76
Pre-gestational diabetes	1.02	0.84	0.80	3.86	2.75	0.72	0.55	3.42	1.35	1.61
Gestational diabetes	1.64	0.84	0.99	2.54	2.06	0.24	0.00	3.14	1.06	1.31
Symptoms on day of first assessment										
Nausea or vomiting	3.68	22.69	0.90	2.96	13.06	97.27	77.28	24.93	25.84	25.73
Headache or visual disturbance	3.73	21.85	0.90	2.78	13.06	89.39	77.28	24.10	24.64	24.57
Right upper quadrant or epigastric pain	3.68	22.69	0.90	2.87	12.37	97.35	7.03	14.87	17.28	16.99
Chest pain or dyspnoea	3.73	22.69	61.16	2.78	13.06	91.96	77.28	35.27	39.12	38.65
Signs on day of first assessment										

Variable	Missing (%)									
	Vancouver N=1956	FINNPEC N=119	fullPIERS N=2011	miniPIERS N=2126	Oxford N=291	PETRA N=1244	PREP N=1096	Women with adverse outcome at any time after first assessment (N=1083)	Women without an adverse outcome (N=7760)	Total
Height(cm)	11.04	0.84	13.08	17.92	8.59	0.08	1.64	14.77	9.60	10.23
Weight (kg)	6.34	0.84	3.93	12.51	34.02	0.08	1.64	12.19	5.88	6.65
Systolic blood pressure (mm Hg)	8.08	21.85	2.83	3.25	14.43	13.67	77.37	20.87	14.74	15.49
Diastolic blood pressure (mm Hg)	8.08	21.01	2.83	3.34	14.43	13.67	77.37	20.78	14.77	15.50
Oxygen saturation (%)	0.00	0.00	0.00	0.00	0.00	0.00	90.15	12.93	10.93	11.17
Dipstick proteinuria (number of '+')	72.75	53.78	32.67	23.00	32.30	48.39	19.34	32.96	41.03	40.04
Laboratory tests – worst values on day of first assessment										
Haematocrit (%)	10.79	39.50	68.37	39.42	99.66	23.71	100.00	50.97	46.39	46.95

Variable	Missing (%)									
	Vancouver N=1956	FINNPEC N=119	fullPIERS N=2011	miniPIERS N=2126	Oxford N=291	PETRA N=1244	PREP N=1096	Women with adverse outcome at any time after first assessment (N=1083)	Women without an adverse outcome (N=7760)	Total
Total leucocyte count (x 10^9 per L)	9.92	39.50	10.84	58.00	13.40	100.00	23.27	37.21	36.43	36.53
Platelet count (x 10^9 per L)	10.33	39.50	10.99	37.82	13.40	24.04	23.18	24.01	20.70	21.10
Mean platelet volume (fL)	14.11	40.34	15.32	96.43	47.77	100.00	100.00	66.76	57.19	58.36
Fibrinogen (g/L)	22.44	92.44	29.04	75.16	89.69	86.01	61.68	56.23	53.21	53.58
Activated partial thromboplastin time (sec)	22.44	92.44	29.04	75.35	89.69	86.01	61.68	56.23	53.26	53.62
Serum creatinine (µmol/L)	18.35	77.31	13.72	52.54	12.71	31.19	22.90	30.29	28.25	28.50
Uric acid (mmol/L)	19.58	60.50	14.37	65.05	14.43	51.13	34.03	43.67	34.86	35.94
Aspartate transaminase (U/L)	18.66	99.16	16.41	66.60	72.85	31.03	76.19	51.62	39.99	41.41

Variable	Missing (%)									
	Vancouver N=1956	FINNPEC N=119	fullPIERS N=2011	miniPIERS N=2126	Oxford N=291	PETRA N=1244	PREP N=1096	Women with adverse outcome at any time after first assessment (N=1083)	Women without an adverse outcome (N=7760)	Total
Alanine transaminase (U/L)	18.40	41.18	12.78	54.09	35.74	31.35	25.73	36.84	28.26	29.31
Albumin (g/L)	26.12	79.83	32.02	95.91	15.81	100.00	25.09	57.99	54.46	54.89

Table 2.14: Baseline Characteristics, Co-interventions, and Pregnancy Outcomes of the Study Cohort before imputation

Variable	**Internal combined data**			**External validation**		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Health system						
National *per capita* gross domestic product (USD)	40,773 [73,28-46,594]	43,585 [28,205-50,114]	<0.001	42,330 [41,064-43,043]	42,330 [40,361-43,043]	0.043
National maternal mortality ratio (maternal deaths per 100,000 live births)	11 [10-171]	11 [11-15]	0.069	8 [7.5-8]	8 [7.0-8]	0.128
Demographics						
Race			0.026	Not collected	Not collected	NA
White	379 (34.8%)	2485 (31.8%)				
Asian	323 (29.7%)	2365 (30.2%)				
Black	303 (27.8%)	2001 (25.6%)				
Other	119 (10.9%)	909 (11.6%)				
Maternal age at expected date of delivery (years)	30 [25.25-35]	31 [27-36]	0.004	29 [26-34]	31 [27-35]	0.201
Nulliparous	653 (60%)	4542 (59%)	0.277	Not collected	Not collected	NA

Variable	Internal combined data			External validation		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Multiple pregnancy	112 (10%)	527 (7%)	<0.001	Not collected	Not collected	NA
Gestational age at eligibility (weeks)	33.24 [30-37.13]	36 [32-38.36]	<0.001	35.6 [31.4-38.9]	37.9 [35.6-39.7]	<0.001
Past and current medical and obstetrical history – no. (%)						
Cigarette smoking	139 (13%)	1011 (13%)	0.885	Not collected	Not collected	NA
Chronic hypertension	146 (13%)	1324 (17%)	0.003	Not collected	Not collected	NA
Pre-gestational renal disease	63 (6%)	519 (7%)	0.296	Not collected	Not collected	NA
Pre-gestational diabetes	61 (6%)	409 (5%)	0.613	Not collected	Not collected	NA
Gestational diabetes	75 (7%)	973 (13%)	<0.001	Not collected	Not collected	NA
Symptoms on day of first assessment						
Nausea or vomiting	103 (10%)	351 (5%)	<0.001	20 (16.5%)	175 (6.3%)	<0.001
Headache or visual disturbance	328 (30%)	1426 (18%)	<0.001	54 (44.6%)	760 (27.4%)	<0.001
Right upper quadrant or epigastric pain	168 (16%)	493 (6%)	<0.001	27 (22.3%)	138 (5.0%)	<0.001
Chest pain or dyspnoea	62 (6%)	93 (1%)	<0.001	17 (14.1%)	29 (1.0%)	<0.001
Signs on day of first assessment						

Variable	Internal combined data			External validation		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Height(cm)	162.56 [157.48-167]	162.56 [157.48-167.64]	0.475	164 [160-168]	165 [161-169.6]	0.069
Weight (kg)	78.0 [68.5-87.0]	81.0 [70.3-93.4]	<0.001	77.0 [67.5-88]	85.9 [75-100]	<0.001
Systolic blood pressure (mm Hg)	155 [143-170]	150 [140-162]	<0.001	150 [143-164]	148 [142-156]	0.007
Diastolic blood pressure (mm Hg)	100 [90-110]	95 [90-101]	<0.001	94 [90-100]	94 [90-99]	0.526
Oxygen saturation less than 93%	97 [97-97]	97 [97-97]	<0.001	Not collected	Not collected	NA
Dipstick proteinuria (number of '+')	2 [1-3]	1 [1-2]	<0.001	1 [1-1]	1 [0-1]	<0.001
Laboratory tests – worst values on day of first assessment						
Haematocrit (%)	0.35 [0.31-0.38]	0.36 [0.34-0.38]	<0.001	0.35 [0.32-0.38]	0.36 [0.33-0.38]	0.068
Total leucocyte count (x 10^9 per L)	10.4 [8.4-13]	10.4 [8.6-12.6]	0.593	9.66 [7.62-13.57]	9.9 [8.2-12]	0.733
Platelet count (x 10^9 per L)	189 [143.5-244]	208 [166-253.75]	<0.001	189 [126-237]	219 [182-265]	<0.001
Mean platelet volume (fL)	10.7 [9.17-11.83]	11.1 [10.2-12]	<0.001	9.5 [8.7-10.7]	9.8 [8.9-10.9]	0.28
Fibrinogen (g/L)	27.7 [24.82-31]	26 [24-28]	<0.001	6.2 [5.4-7.4]	6.5 [5.7-7.2]	0.186
Activated partial thromboplastin time (sec)	28 [25-31]	26 [24.2-28.1]	<0.001	26.2 [24.37-28.85]	26.4 [24.35-27.8]	0.691

Variable	Internal combined data			External validation		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Serum creatinine (µmol/L)	63 [52-76.5]	59 [50-69]	<0.001	61 [50-74.5]	54 [47-63]	<0.001
Uric acid (mmol/)L	367 [298.5-440]	339.04 [282-398]	<0.001	350 [297-435]	339.5 [281.25-396]	0.076
Aspartate transaminase (U/L	32 [22-55.25]	25 [20-34]	<0.001	28 [22.5-49.5]	23 [18-31]	<0.001
Alanine transaminase (U/L)	24 [15-42.25]	18 [12-29]	<0.001	17 [11-33.25]	14 [10-20]	0.018
Albumin (g/L)	28.2 [3.8-33]	31 [26-34]	<0.001	34 [31-36]	35 [33-37]	<0.001
Interventions						
Corticosteroid received	377 (34.8%)	1695 (21.8%)	<0.001	Not collected	Not collected	NA
Antihypertensive medication received	714 (65.9%)	4097 (52.7%)	<0.001	53 (43.8%)	849 (30.6%)	0.003
Magnesium sulphate received	556 (51.%)	2257 (29.0%)	<0.001	4 (3.3%)	15 (0.5%)	0.007
Pregnancy outcomes						
Gestational age at delivery (weeks)	35.0 [31.5-37.6]	37.0 [34.1-38.7]	<0.001	36.43 [32.86-39.43]	39 [37.29-40.14]	<0.001
Birthweight (g)	1900 [965-2750]	2560 [1556-3200]	<0.001	Not collected	Not collected	NA
Intrauterine fetal death, ≥20^{+0} weeks or ≥500 g(117)	65 (6.0%)	163 (2.1%)	<0.001	6 (4.96%)	18 (0.65%)	<0.001

Variable	Internal combined data			External validation		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Neonatal death, within 28 days	40 (3.7%)	102 (1.3%)	<0.001	2 (1.65%)	1 (0.04%)	0.005

Values expressed as number (%) or median [interquartile range]; * Chi-squared, Fisher's exact, or Mann-Whitney U test

Race defined in Supplemental Table S2 in the Supplementary Appendix

Table 2.15: Baseline Characteristics, Co-interventions, and Pregnancy Outcomes of the Study Cohort after imputation

Variable	Internal combined data			External validation		
	Women with adverse outcome at any time after first assessment N=1083, 12.2%	**Women without an adverse outcome N=7760, 87.8%**	**p-value***	**Women with adverse outcome at any time after first assessment N=121, 4.2%**	**Women without an adverse outcome N=2780, 95.4%**	**p-value***
Health system						
National *per capita* gross domestic product [USD]	41064.13 [7501.47-46594.45]	43585.51 [28205.73-50114.18]	<0.001	42356.6 [42140.91-43043.23]	42330.12 [41064.13-43043.23]	0.037
National maternal mortality ratio [maternal deaths per 100,000 live births]	11 [10-161]	11 [11-15]	0.069	7.4 [7-8]	7.05 [7-8]	0.02
Demographics						
Race			0.026	Not collected	Not collected	NA
White	379 [34.8%]	2485 [31.8%]				
Asian	323 [29.7%]	2365 [30.2%]				
Black	303 [27.8%]	2001 [25.6%]				
Other	119 [10.9%]	909 [11.6%]				
Maternal age at expected date of delivery [years]	31 [26-35]	31 [27-36]	0.038	29 [26-34]	31 [27-35]	0.201
Nulliparous	653 [60.0%]	4542 [58.1%]	0.251	Not collected	Not collected	NA

Multiple pregnancy	128 [11.8%]	529 [6.8%]	<0.001	Not collected	Not collected	NA
Gestational age at eligibility [weeks]	33.4 [30.0-37.1]	36.0 [32.0-38.4]	<0.001	35.57 [31.43-38.86]	37.86 [35.57-39.71]	<0.001
Past and current medical and obstetrical history – no. [%]						
Cigarette smoking	139 [12.8%]	1011 [12.9%]	0.923	Not collected	Not collected	NA
Chronic hypertension	146 [13.4%]	1324 [16.9]	0.003	Not collected	Not collected	NA
Pre-gestational renal disease	63 [5.8%]	520 [6.6%]	0.296	Not collected	Not collected	NA
Pre-gestational diabetes	61 [5.6%]	409 [5.2%]	0.612	Not collected	Not collected	NA
Gestational diabetes	75 [6.9%]	974 [12.5%]	<0.001	Not collected	Not collected	NA
Symptoms on day of first assessment						
Nausea or vomiting	107 [9.8%]	443 [5.7%]	<0.001	Not imputed	Not imputed	NA
Headache or visual disturbance	406 [37.3%]	2132 [27.3%]	<0.001	Not imputed	Not imputed	NA
Right upper quadrant or epigastric pain	205 [18.8%]	754 [9.6%]	<0.001	Not imputed	Not imputed	NA
Chest pain or dyspnoea	73 [6.7%]	117 [1.5%]	<0.001	Not imputed	Not imputed	NA
Signs on day of first assessment						
Height [cm]	162 [157-166]	163 [157-168]	0.244	164 [160-168]	165 [161-169.6]	0.069
Weight [kg]	78.0 [68.5-87.0]	81.0 [70.3-93.4]	<0.001	77.0 [67.5-88]	85.9 [75-100]	<0.001
Systolic blood pressure [mm Hg]	156 [147-166]	151 [140-161]	<0.001	150 [143-164]	148 [142-156]	0.007
Diastolic blood pressure [mm Hg]	99 [91-105]	96 [90-100]	<0.001	94 [90-100]	94 [90-99]	0.526
Oxygen saturation less than 93%	43 [3.9%]	28 [0.4%]	<0.001	Not collected	Not collected	NA
Dipstick proteinuria [number of '+']	2 [1-3]	1 [1-2]	<0.001	Not imputed	Not imputed	NA
Laboratory tests – worst values on day of first assessment						

Haematocrit [%]	0.36 [0.34-0.38]	0.36 [0.34-0.38]	0.002	0.35 [0.33-0.37]	0.36 [0.34-0.37]	0.043
Total leucocyte count [x 10^9 per L]	10.9 [9.5-12.3]	10.6 [9.3-12.0]	<0.001	9.7 [7.7-13.3]	10 [8.4-11.9]	0.689
Platelet count [x 10^9 per L]	198 [157-235]	212 [175-245]	<0.001	194 [127-237]	219 [183-264]	<0.001
Mean platelet volume [fL]	9.6 [8.9-10.8]	10.0 [9.1-11.2]	<0.001	9.89 [9.16-11.1]	9.9 [9.2-10.79]	0.479
Fibrinogen [g/L]	27.1 [25.1-29.3]	26.2 [24.6-27.9]	<0.001	Not imputed	Not imputed	NA
Activated partial thromboplastin time [sec]	27.5 [25.5-30.0]	26.6 [24.8-28.2]	<0.001	Not imputed	Not imputed	NA
Serum creatinine [µmol/L]	64 [54-75]	60 [52-70]	<0.001	60 [50-74]	54 [48-63]	<0.001
Uric acid [mmol/]L	365 [321-418]	339 [297-385]	<0.001	360 [303.7-435]	330 [293.61-376.38]	<0.001
Aspartate transaminase [U/L	39 [29-65]	29 [22-41]	<0.001	32 [24-64.05]	25.05 [20.35-34.83]	<0.001
Alanine transaminase [U/L]	32 [20-55]	23 [14-35]	<0.001	22.3 [12-41]	16 [12-22.45]	<0.001
Albumin [g/L]	27 [18-30]	29 [23-32]	<0.001	34 [32-36]	35 [33-37]	<0.001

Values expressed as number (%) or median [interquartile range]; ***** Chi-squared, Fisher's exact, or Mann-Whitney U test

Race defined in Supplemental Table S2 in the Supplementary Appendix

2.4.3 Modelling

Following imputation, the set of imputed development datasets were used to fit models using methods described in section 0. The description of the fitting process for each model, along with the corresponding method and model code can be seen in Table 2.16. The fullPIERS variables were gestational age on eligibility, symptom of chest pain or dyspnoea, serum creatinine, platelet count AST and oxygen saturation, based on Equation 1.1.

O_fP, R_fP, R6, L2 and NN3 only used these variables for modelling. The remaining models used either all available predictor variables, or some variable selection methods. The final variables used by each of the original set of models can be seen in Table 2.17, and variables of the models created during sensitivity analyses are shown in Table 2.18.

The formula for the original fullPIERS (O_fP) model was taken from the original publication (3). As it is suggested that the model needs to be re-calibrated to be used in a new setting, we choose to refit the model on our dataset (R_fP), and compare the models created with our proposed methods to this version. In some secondary analyses I also compared to the original model. Models were then fitted as described in the Methods section. After considering the results presented in Table *2.19* and Table *0.1* (Appendix C), along with model complexity as seen by the number of variables in the models, we decided to carry out the s1 sensitivity analysis for all models created using random forest methodology. As the performance of the random forest using only variables with above average importance (RF2) in this sensitivity analysis was most liked by the clinicians advising on the project, s2 sensitivity analysis was carried out by only following the model fitting steps of RF2.

Table 2.16 Description of model fitting processes and the corresponding model code names

Model code	**Method**	**Fitting**
O_fP	Original fullPIERS model	*NA*
R_fP	Refitted fullPIERS model	Each model was fitted using the `caret` function, the formula defined according to Equation **, using "glm" method with 10-fold cross-validation, combined to a caretList object.
RF1	Random forest using all variables	Each model was fitted using the `caret` function, the formula defined as outcome modelled by all variables, using "rf" method with 10-fold cross-validation, combined to a caretList object.
RF2	Random forest using variables with above average importance in RF1	Importance was extracted from RF1, mean importance M calculated. Variables with importance >M were selected. Each was model fitted using the `caret` function, the formula defined as outcome modelled by selected variables, using "rf" method with 10-fold cross-validation, combined to a caretList object.
RF3	Random forest using Recursive Feature Elimination	The `rfe` function was used on each dataset with 10-fold cross-validation to order variables on importance. The top 8 most important variables were extracted from each. Each was model fitted using the `caret` function, the formula defined as outcome modelled by all variables appearing in any top 8, using "rf" method with 10-fold cross-validation, combined to a caretList object.
RF4	Random forest using Boruta variable selection method	The `Boruta` function with 600 trees and maximum 250 runs was used to class variables into Confirmed, Rejected or Tenatative groups. Variables classed as Confirmed for any of the datasets were selected. Each was model fitted using the `caret` function, the formula defined as outcome modelled by selected variables, using "rf" method with 10-fold cross-validation, combined to a caretList object.
RF5	Random forest using Vita variable selection method	The `CVPVI` function with 16 variables randomly sampled at each split for the *i*th forest was used for variable selection with cross-validation on each dataset. Variables with p-value <0.05 were selected. Each was model fitted using the `caret` function, the formula defined as outcome modelled by selected

Model code	Method	Fitting
		variables, using "rf" method with 10-fold cross-validation, combined to a caretList object.
RF6	Random forest using fullPIERS variables only	Each model fitted using the `caret` function, the formula defined according to Equation 1.1, using "rf" method with 10-fold cross-validation, combined to a caretList object.
L1	LASSO regression fitted using all available variables	Each was model fitted using the `caret` function, the formula defined as outcome modelled by all variables, using "glmnet" method with 10-fold cross-validation and using a tuning grid with alpha=1 and a range of lambda values from 10^{-3} to 10^{3}. The models were combined into a caretList object.
L2	LASSO regression fitted using fullPIERS variables	Each model fitted using the `caret` function, the formula defined according to Equation **, using "glmnet" method with 10-fold cross-validation and using a tuning grid with alpha=1 and a range of lambda values from 10^{-3} to 10^{3}. The models were combined into a caretList object.
Ridge	Ridge regression fitted using all available variables	Each was model fitted using the `caret` function, the formula defined as outcome modelled by all variables, using "glmnet" method with 10-fold cross-validation and using a tuning grid with alpha=0 and a range of lambda values from 10^{-3} to 10^{3}. The models were combined into a caretList object.
NN1	Artificial neural network using all available variables	Each was model fitted using the `caret` function, the formula defined as outcome modelled by all variables, using "nnet" method with 10-fold cross-validation and using a tuning grid with a range of hidden layer sizes from 1 to 10 and a range of decay values from 10^{-7} to 1. The models were combined into a caretList object.
NN2	Artificial neural network using variables with above average importance from NN1	Importance was extracted from NN1, mean importance M calculated. Variables with importance >M were selected. Each was model fitted using the `caret` function, the formula defined as outcome modelled by selected variables, using "nnet" method with 10-fold cross-validation and using a tuning grid with a range of hidden layer sizes from 1 to 10 and a range of decay values from 10^{-7} to 1. The models were combined into a caretList object.
NN3	Artificial neural network using fullPIERS variables	Each model fitted using the `caret` function, the formula defined according to Equation 1.1, using

Model code	Method	Fitting
		"nnet" method with 10-fold cross-validation and using a tuning grid with a range of hidden layer sizes from 1 to 10 and a range of decay values from 10^{-7} to 1. The models were combined into a caretList object.
BMA	Bayesian model averaging fitted using all available variables	Variable importance was determined by using the `bic.glm` function to fit a model on each dataset using all variables and calculate the probability of the coefficient of each variable being non-zero. Variables with probability of being non-zero >0 on any dataset were selected. Each model was then fitted using `bic.glm` with the formula defined as outcome modelled by selected variables.
XGB1	Gradient boosted trees using all predictor variables	Each model was fitted using the `caret` function, the formula defined as outcome modelled by all variables, using "xgboost" method with 10-fold cross-validation, combined to a caretList object.
XGB2	Gradient boosted trees using variables with above average importance	Importance was extracted from XGB1, mean importance M calculated. Variables with importance >M were selected. Each was model fitted using the `caret` function, the formula defined as outcome modelled by selected variables, using "xgboost" method with 10-fold cross-validation, combined to a caretList object.
RF1_s1, RF2_s1, RF3_s1, RF4_s1	RF1:4, coagulation variables excluded from potential predictor variables	Coagulation variables were removed from the dataset before repeating the same model fitting steps as used for RF1 to RF4
RF2_s2	RF2 fitted with coagulation variables and mean platelet volume excluded from potential predictor variables	Coagulation variables and mean platelet volume were removed from the dataset before the fitting steps for RF1 and RF2 were repeated.

Table 2.17: Variables per model for models including coagulation variables and MPV. Models using fullPIERS variables only not included.

Variable	RF1	RF2	RF3	RF4	RF5	L1	R1	NN1	NN2	XGB1	XGB2	BMA
Health system												
National per capita gross domestic product [USD]	✓	✓		✓	✓	✓	✓			✓	✓	
National maternal mortality ratio [maternal deaths per 100,000 live births]	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
Demographics												
Race	✓			✓	✓	✓	✓	✓	✓	✓		✓
Maternal age at expected date of delivery [years]	✓	✓			✓	✓	✓			✓		
Nulliparous	✓			✓	✓	✓	✓			✓		
Multiple pregnancy	✓			✓	✓			✓				
Gestational age at eligibility [weeks]	✓	✓		✓	✓	✓	✓			✓		
Past and current medical and obstetrical history – no. [%]												
Cigarette smoking	✓			✓	✓	✓	✓			✓		
Chronic hypertension	✓				✓	✓	✓	✓		✓		✓
Pre-gestational renal disease	✓			✓	✓	✓	✓	✓	✓	✓		✓
Pre-gestational diabetes	✓				✓	✓	✓	✓		✓		✓
Gestational diabetes	✓				✓	✓	✓	✓		✓		✓
Symptoms on day of first assessment												
Nausea or vomiting	✓		✓	✓	✓	✓	✓	✓	✓	✓		
Headache or visual disturbance	✓			✓	✓	✓	✓	✓		✓		✓
Right upper quadrant or epigastric pain	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓
Chest pain or dyspnoea	✓		✓	✓	✓	✓	✓	✓	✓	✓		✓
Signs on day of first assessment												
Height [cm]	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
Weight [kg]	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
Systolic blood pressure [mm Hg]	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓

Variable	**RF1**	**RF2**	**RF3**	**RF4**	**RF5**	**L1**	**R1**	**NN1**	**NN2**	**XGB1**	**XGB2**	**BMA**
Diastolic blood pressure [mm Hg]	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
Oxygen saturation less than 93%	✓	✓	✓	✓	✓	✓	✓			✓	✓	
Dipstick proteinuria [number of '+']	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
Laboratory tests – worst values on day of first assessment												
Haematocrit [%]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Total leucocyte count [x 10^9 per L]	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
Platelet count [x 10^9 per L]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mean platelet volume [fL]	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓
Fibrinogen [g/L]	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Activated partial thromboplastin time [sec]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Serum creatinine [µmol/L]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Uric acid [mmol/]L	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Aspartate transaminase [U/L]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Alanine transaminase [U/L]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Albumin [g/L]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Total number of variables	33	19	12	29	33	32	32	27	21	32	12	25

Table 2.18: Variables per model for sensitivity analysis models

Variable	**RF1_s**	**RF2_s**	**RF3_s**	**RF4_s**	**RF5_s**	**RF2_s2**
Health system						
National *per capita* gross domestic product [USD]	✓	✓		✓	✓	✓
National maternal mortality ratio [maternal deaths per 100,000 live births]	✓	✓		✓	✓	✓
Demographics						
Race	✓			✓	✓	
Maternal age at expected date of delivery [years]	✓	✓		✓	✓	✓
Nulliparous	✓			✓	✓	
Multiple pregnancy	✓			✓	✓	
Gestational age at eligibility [weeks]	✓	✓		✓	✓	✓
Past and current medical and obstetrical history – no. [%]						
Cigarette smoking	✓			✓	✓	
Chronic hypertension	✓				✓	
Pre-gestational renal disease	✓			✓	✓	
Pre-gestational diabetes	✓				✓	
Gestational diabetes	✓				✓	
Symptoms on day of first assessment						
Nausea or vomiting	✓		✓	✓	✓	
Headache or visual disturbance	✓			✓	✓	
Right upper quadrant or epigastric pain	✓		✓	✓	✓	
Chest pain or dyspnoea	✓		✓	✓	✓	
Signs on day of first assessment						
Height [cm]	✓	✓		✓	✓	✓
Weight [kg]	✓	✓		✓	✓	✓
Systolic blood pressure [mm Hg]	✓	✓		✓	✓	✓

Variable	**RF1_s**	**RF2_s**	**RF3_s**	**RF4_s**	**RF5_s**	**RF2_s2**
Diastolic blood pressure [mm Hg]	✓	✓		✓	✓	✓
Oxygen saturation less than 93%	✓	✓	✓	✓	✓	✓
Dipstick proteinuria [number of '+']	✓			✓	✓	
Laboratory tests – worst values on day of first assessment						
Haematocrit [%]	✓	✓	✓	✓	✓	✓
Total leucocyte count [x 10^9 per L]	✓	✓		✓	✓	✓
Platelet count [x 10^9 per L]	✓	✓	✓	✓	✓	✓
Mean platelet volume [fL]	✓	✓		✓	✓	
Fibrinogen [g/L]						
Activated partial thromboplastin time [sec]						
Serum creatinine [µmol/L]	✓	✓	✓	✓	✓	✓
Uric acid [mmol/]L	✓	✓	✓	✓	✓	✓
Aspartate transaminase [U/L]	✓	✓	✓	✓	✓	✓
Alanine transaminase [U/L]	✓	✓	✓	✓	✓	✓
Albumin [g/L]	✓	✓	✓	✓	✓	✓
Total number of variables	31	18	11	28	31	17

Out of the models not using only the fullPIERS variables, RF3_s1 used the smallest number of variables at 11, and RF1 used the largest number, including all 33 predictor variables. The variables belonging to the demographics and past medical history groups were used in the least models, while variables in the signs and laboratory test groups were used the most. Five variables, namely platelet count, uric acid, aspartate transaminase, alanine transaminase and albumin were, on the other hand, included in all models except for the ones using fullPIERS variables only.

2.4.3.1 Outcome/no outcome prediction

All models were first assessed on their ability to classify into outcome and no outcome groups. This was assessed by the ROC, sensitivity, specificity, PPV, NPV and LR+ and LR-. Calibration was assessed using Cox calibration slope and intercept. These measures are presented in Table 2.19.

The AUROC was very similar for all methods, since the outcome rate in the dataset is quite low, however there was much bigger differences between the predicted risk groups.

In terms of the measurements of discrimination, all models performed very similarly, with 95% confidence intervals largely overlapping. This suggests that models separate patients with an outcome from those without an outcome equally well or equally poorly. The AUC values were all at or above 70%, indicating fair discrimination, with most confidence intervals going above 80%, indicating possible good discrimination. This is not very surprising, as even classifying all patients into the no outcome group would equal to over 90% correctly classified patients, and the performance would only decrease drastically if a large number of patients were falsely classified as outcomes. Area under the precision-recall curve was poor for all models, however, with a 40% maximum achieved by RF1, RF4 and XGB1, and a 25% minimum for R_fP and L2.

At the optimal probability threshold balancing sensitivity and specificity, RF6 and BMA had the highest sensitivity at 77% & 76%, R_fP and NN2 had the lowest at 44% and 58%, while all other models had a sensitivity between 60% and 75% except for R_fP, however all confidence intervals except for R_fP overlapped. R_fP and RF3 had the highest specificity at 86%, with confidence intervals only overlapping with NN2. The models with the lowest specificity were RF6 and L2 (both models using fullPIERS variables) at 66%, with confidence intervals only

overlapping with BMA. The positive predictive value was poor, around 20% for all models. This shows that from the patients who are predicted in the outcome group, only around 20% go on to having an outcome, indicating that the models are not suitable for ruling in outcome. This is reinforced by the positive likelihood ratios being between 2 and 5 for all models, showing that there is only a small increase in probability of outcome in the group predicted to have an outcome. This suggests that the high AUC values of the models indeed mainly come from correctly classifying patients into the no outcome group, which is also reinforced by the low area under the precision-recall curve for all models.

The negative predictive value was around 97% for all models, showing that 97% of those who were predicted as no outcome did not develop an outcome. This indicates that the model could be used for ruling out an outcome, however all negative likelihood ratios are either around 0.4 or greater than 0.5, indicating only a small or no reduction in risk of outcome for the group predicted no outcome compared to the outcome group.

Calibration-in-the-large, measured by the Cox calibration intercept, was not great. Only RF6 had a Cox calibration intercept close to 0, all other models having 0.5 or greater, with L2 having a Cox calibration intercept as high as 9.25. Cox calibration slope was good for RF6 and acceptable for R-fP and L1, all other models having values greater than 1.2. L2 performed worst in this case as well, with a slope of 4.59. There was no difference in the Brier score between models.

The Spiegelhalter z statistic and p-value indicated good calibration for R_fP, RF6, L1 (which was shown to be the closest to perfect calibration out of the models), L1, Ridge, BMA, XGB1 and XGB2. RF2, 4 and 5 received the worst scores using this method.

Table 2.19: Performance measurements of outcome/no outcome classification per model

Performance Measure	**R_fP**	**RF1**	**RF2**	**RF3**	**RF4**	**RF5**	**RF6**	**L1**	**L2**	**R1**	**NN1**	**NN2**	**NN3**	**BMA**	**XGB 1**	**XGB 2**
Area under ROC	0.7 (0.65 - 0.75)	0.81 (0.77 - 0.85)	0.8 (0.76 - 0.84)	0.8 (0.76 - 0.84)	0.81 (0.77 - 0.85)	0.81 (0.77 - 0.85)	0.78 (0.74 - 0.82)	0.78 (0.74 - 0.83)	0.74 (0.69 - 0.78)	0.78 (0.74 - 0.83)	0.78 (0.74 - 0.82)	0.78 (0.74 - 0.82)	0.77 (0.73 - 0.81)	0.77 (0.73 - 0.82)	0.82 (0.78 - 0.86)	0.8 (0.76 - 0.85)
Area under precision-recall curve	0.25	0.40	0.38	0.37	0.40	0.41	0.34	0.32	0.25	0.31	0.26	0.28	0.27	0.29	0.40	0.37
Cox Calibration Intercept	0.36	0.74	0.36	0.42	0.66	0.73	0.09	0.42	9.25	1.29	0.87	0.78	0.66	0.59	0.66	0.5
Cox Calibration Slope	1.16	1.42	1.28	1.23	1.39	1.42	1.07	1.18	4.59	1.55	1.42	1.38	1.33	1.28	1.29	1.24
Brier Score	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Spiegelhalter z statistic	-0.7	-2.71	-3.39	-1.89	-2.92	-2.85	-1.17	-0.61	-1.11	-1.25	-1.62	-1.74	-1.67	-1.24	-1.18	-1.33

Performance Measure	**R_fP**	**RF1**	**RF2**	**RF3**	**RF4**	**RF5**	**RF6**	**L1**	**L2**	**R1**	**NN1**	**NN2**	**NN3**	**BMA**	**XGB 1**	**XGB 2**
Spiegelhalter p-value	0.49	0.01	0	0.06	0	0	0.24	0.54	0.27	0.21	0.1	0.08	0.1	0.21	0.24	0.18
Sensitivity	0.44 (0.36 - 0.53)	0.69 (0.61 - 0.77)	0.74 (0.65 - 0.81)	0.62 (0.54 -0.7)	0.68 (0.59 - 0.75)	0.7 (0.62 - 0.77)	0.77 (0.69 - 0.84)	0.69 (0.61 - 0.77)	0.71 (0.62 - 0.78)	0.61 (0.53 -0.7)	0.63 (0.54 - 0.71)	0.58 (0.49 - 0.66)	0.67 (0.59 - 0.75)	0.76 (0.69 - 0.83)	0.66 (0.58 - 0.74)	0.69 (0.6- 0.76)
Specificity	0.86 (0.85 - 0.88)	0.81 (0.79 - 0.82)	0.72 (0.7- 0.74)	0.86 (0.85 - 0.88)	0.83 (0.81 - 0.84)	0.8 (0.78 - 0.81)	0.66 (0.64 - 0.68)	0.75 (0.73 - 0.77)	0.66 (0.64 - 0.68)	0.82 (0.81 - 0.84)	0.81 (0.79 - 0.83)	0.84 (0.83 - 0.86)	0.75 (0.73 - 0.77)	0.68 (0.66 -0.7)	0.84 (0.82 - 0.85)	0.79 (0.77 - 0.81)
Positive Predictive Value	0.18 (0.14 - 0.22)	0.2 (0.16 - 0.23)	0.15 (0.13 - 0.18)	0.23 (0.19 - 0.28)	0.21 (0.17 - 0.25)	0.19 (0.16 - 0.23)	0.13 (0.11 - 0.16)	0.16 (0.13 - 0.19)	0.12 (0.1- 0.15)	0.19 (0.16 - 0.23)	0.18 (0.15 - 0.22)	0.2 (0.16 - 0.24)	0.15 (0.13 - 0.19)	0.14 (0.11 - 0.16)	0.21 (0.18 - 0.26)	0.18 (0.15 - 0.22)
Negative Predictive Value	0.96 (0.95 - 0.97)	0.97 (0.97 - 0.98)	0.98 (0.97 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.97 - 0.98)	0.98 (0.97 - 0.98)	0.98 (0.97 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.96 - 0.98)	0.98 (0.97 - 0.98)	0.97 (0.96 - 0.98)	0.97 (0.97 - 0.98)

Performance Measure	**R_fP**	**RF1**	**RF2**	**RF3**	**RF4**	**RF5**	**RF6**	**L1**	**L2**	**R1**	**NN1**	**NN2**	**NN3**	**BMA**	**XGB 1**	**XGB 2**
Positive likelihood ratio	3.24 (2.61 - 4.02)	3.59 (3.12 - 4.14)	2.64 (2.34 - 2.98)	4.47 (3.78 - 5.28)	3.95 (3.4-4.58)	3.44 (3-3.95)	2.29 (2.05 - 2.55)	2.8 (2.45 - 3.19)	2.1 (1.86 - 2.38)	3.47 (2.96 - 4.08)	3.31 (2.83 - 3.87)	3.71 (3.12 - 4.41)	2.7 (2.36 -3.1)	2.38 (2.13 - 2.65)	4.03 (3.46 -4.7)	3.31 (2.88 - 3.81)
Negative likelihood ratio	0.65 (0.56 - 0.75)	0.38 (0.3-0.49)	0.37 (0.28 - 0.48)	0.44 (0.36 - 0.54)	0.39 (0.3-0.49)	0.38 (0.29 - 0.49)	0.34 (0.25 - 0.47)	0.41 (0.32 - 0.52)	0.44 (0.34 - 0.57)	0.47 (0.38 - 0.58)	0.46 (0.37 - 0.57)	0.5 (0.41 - 0.61)	0.44 (0.34 - 0.55)	0.35 (0.26 - 0.47)	0.4 (0.32 - 0.51)	0.4 (0.31 - 0.51)

2.4.3.2 Risk group prediction

After assessment for an outcome/no outcome prediction, the validation set was divided 50-50 into a testing set and a new validation set, risk groups were defined on the test set as described in Chapter 2.2.9.2 and the validation group was classified accordingly. Table 0.1 (Appendix C) shows the risk intervals for each group, the number of patients per risk group, the number of outcomes within 2 days, 7 days and at any time after first admission per group and likelihood ratios for each.

All models were able to define a predicted probability threshold for which the positive likelihood ratio was 10 or greater and thus could classify into the very high-risk stratum. In the internal validation dataset all models except for XGB1 had a positive likelihood ratio of at least 10 in this stratum, with most models having at least a 50% outcome rate (with the exception of RF1, RF4, RF5, XGB1 and XGB2). The proportion of people classified into this stratum varied between as little as 0.2% to as much as 7.2% of the 1105 patients. The positive likelihood ratio remained above 10 for outcomes within 7 days for R_fP, RF1, RF2, RF3, RF6, L1, Ridge and NN3, and apart from RF1, remained above 10 for outcomes at any time after first admission as well. In terms of balancing the number of patients classed into this actionable risk strata with the outcome rate and likelihood ratio, random forest models performed best.

Classification into the high-risk stratum was also achieved by all models. In the internal validation dataset, most models classed between 5-10% of patients into this group and an outcome rate of 15-20%. Only NN2 had a positive likelihood ratio of 5 or above in this stratum and it did not remain above 5 for outcomes in 7 days or at any time. As such, all models performed similarly for this risk stratum, and no model or method performed best.

As the aim of the moderate-risk stratum was not to meet a specified negative or positive likelihood ratio, but simply to have as few of the patients in this group as possible, all models were able to set predicted probability thresholds for this stratum. L2, NN1 and NN2 classified all remaining patients into this risk group, as they could not specify low- and very low-risk thresholds. While R_fP, NN3 and BMA did have risk groups lower than moderate, they still classified over 80% of patients into this stratum. XGB1 had the smallest percentage of patients classified into this stratum at 32.5%, followed by RF5 at 38.3% and RF1 and RF4 each at 41.4%.

RF1, 2, 4, 5 and 6 all had less than 50% of patients in this stratum. Models using random forest and gradient boosted trees methods performed best in this stratum.

As mentioned above, L2, NN1 and NN2 could not classify into either low- or very low-risk strata. R_fP could set a predicted probability threshold with negative likelihood ratio under 0.1 for the very low-risk stratum, but not for low-risk. All remaining models could classify into both. In the internal validation dataset, RF1 and 3, L1, Ridge and NN3 all had negative likelihood ratios of 0.2 or less in the low-risk stratum, which only remained at or under 0.2 for L1 and Ridge for both outcome in 7 days and at any point. Number of patients classed into this stratum varied a lot, from as little as 4.1% for BMA, to 52% for XGB1. While L1 and Ridge performed very well in terms of likelihood ratios, they had only 5.6% and 8.6% of patients in this group, respectively. Random forest models except for RF3 each had over 40% of patients in this stratum with outcomes rates between 2-3% and negative likelihood ratios of 0.2-0.4. While not all random forest models had the desired negative likelihood ratio, we deemed these models to have performed best in terms of balancing number of patients in the group with outcome rate and negative likelihood ratio,

In the very low-risk stratum in the internal validation dataset, R_fP, RF2, 3 and 4, L1, Ridge, NN3, BMA and both XGB models had an outcome rate of 0, resulting in a negative likelihood ratio of 0. Of these models, the outcome rate remained 0 for outcome within 7 days for all except, and for R_fP, RF3 and 4, NN3, and the XGB models for outcome at any time. Of these 6 models with no outcomes in the very low-risk stratum, R_fP had 1 person (0.1% of patients) in this stratum, RF3 had 0.7%, RF4 1.7%, NN3 2%, and XGB1 and XGB2 0.5% and 0.6% respectively. Based on these measures, NN3 was decided to have performed best in this stratum, followed by RF4.

Overall, random forest methods seemed to have performed best for five group classification on our internal validation dataset.

2.4.3.3 Sensitivity analyses

As coagulation variables (fibrinogen, APTT and INR) can be difficult to collect in clinical setting as not all locations have the necessary equipment, the inclusion of these variables in the model can limit its usefulness. Hence, a sensitivity analysis was carried out by removing fibrinogen and APTT from the list of predictor variables (as INR was already excluded due to >60% missingness) and fitting models on the data without these variables using the same methods as RF1, 2, 3, 4 and 5. These models have been selected as random forest models showed the best performance on risk prediction, however the variable selection methods could behave differently after removing some variables. Table 2.18 shows the variables per model for these 5 models when fitted without fibrinogen or APTT. Comparing to Table 2.17, we can see that originally all five models contained APTT, and only RF3 did not contain fibrinogen. RF1 used 33 variables, RF2 used 19, RF3 12, RF4 29 and RF5 used all 33 variables. After removing APTT and fibrinogen from the predictor variables, RF1_s1 used all remaining 31 variables, RF2_s1 used 18, RF3_s1 11, RF4_s1 28, and RF5_s1 used all remaining 31 variables. Models used all the same variables as before minus APTT and fibrinogen. RF2_s1 included maternal mortality ratio after the removal of coagulation variables, but not before, and similarly, RF4_s1 gained maternal age as a variable.

Similarly to before, the models were tested both for outcome/no outcome prediction and risk group classification. *Table 2.20* shows the model performances for outcome/no outcome classification for the models without fibrinogen or APTT. Comparing these results to Table *2.19*, we can see that removing the variables did not change the area under the curve for either the ROC or the precision-recall curve for any of the models, while calibration stayed extremely close to the original models. There was no significant difference in the Cox calibration measurements or the brier score, however the Spiegerhalter p-value became significant for RF1_s1 as the z statistic changed from -0.7 of the original model to -2.09 of the sensitivity analysis model. The Spiegelhalter p-value changed in the opposite direction for RF3_s1, becoming non-significant as the z statistic change from -1.89 of the original model to -1.08 of the sensitivity analysis model. There was no significant change in the sensitivity, specificity, PPV, NPV and positive and negative likelihood ratios for any model.

Table *0.2* (Appendix C) shows the model performances for risk group prediction for outcome within 2 days, 7 days and at any point. All models were able define a predicted probability threshold for all 5 risk strata, and all had a positive likelihood ratio of 10 or above in the very high-risk stratum in the internal validation data. RF1_s1, 4_s1 and 5_s1 had under 50% outcomer ate in this stratum, while RF2_s1 had an outcome rate of over 90%, although only 1% of patients were predicted into the stratum. Of the models made during s1 sensitivity analysis, only RF2_s1 had a positive likelihood ratio of 5 or above in the high-risk stratum. This modelling method was also used to fit the only model in s2 sensitivity analysis, which also had a positive likelihood ratio of 5 in the high-risk stratum, but its performance worsened. Models had ~35-60% of patients in the moderate risk group, which was deemed an acceptable range. In the low-risk stratum, RF3_s1 had a negative likelihood ratio of 0.2, while all other models' was 0.3, however RF3_s1 was the only model not to have an outcome rate of 0 in the very low-risk stratum. Overall, RF2_s1 was selected as the best model.

Table 2.20: Outcome/no outcome classification performance of models from sensitivity analyses

Performance Measure	**RF1_s1**	**RF2_s1**	**RF3_s1**	**RF4_s1**	**RF5_s1**	**RF2_s2**
Area under ROC	0.81 (0.76-0.85)	0.8 (0.76-0.84)	0.79 (0.75-0.83)	0.81 (0.76-0.85)	0.81 (0.77-0.85)	0.8 (0.76-0.84)
Area under precision-recall curve	0.40	0.38	0.37	0.40	0.40	0.39
Cox Calibration Intercept	0.8	0.4	0.36	0.73	0.79	0.35
Cox Calibration Slope	1.41	1.28	1.17	1.39	1.41	1.26
Brier Score	0.05	0.05	0.05	0.05	0.05	0.05
Spiegelhalter z statistic	-2.09	-3.02	-1.08	-2.4	-2.26	-2.94
Spiegelhalter p-value	0.04	0	0.28	0.02	0.02	0
Sensitivity	0.66 (0.57-0.74)	0.71 (0.62-0.78)	0.61 (0.53-0.7)	0.61 (0.53-0.7)	0.64 (0.55-0.72)	0.71 (0.63-0.79)
Specificity	0.83 (0.81-0.84)	0.76 (0.74-0.78)	0.85 (0.83-0.87)	0.87 (0.86-0.89)	0.85 (0.84-0.87)	0.76 (0.74-0.78)
Positive Predictive Value	0.2 (0.17-0.25)	0.17 (0.14-0.2)	0.22 (0.18-0.26)	0.25 (0.2-0.3)	0.22 (0.18-0.27)	0.17 (0.14-0.2)

Performance Measure	**RF1_s1**	**RF2_s1**	**RF3_s1**	**RF4_s1**	**RF5_s1**	**RF2_s2**
Negative Predictive Value	0.97 (0.96-0.98)	0.97 (0.97-0.98)	0.97 (0.96-0.98)	0.97 (0.96-0.98)	0.97 (0.96-0.98)	0.98 (0.97-0.98)
Positive likelihood ratio	3.81 (3.27-4.44)	2.98 (2.61-3.4)	4.1 (3.47-4.85)	4.87 (4.1-5.79)	4.29 (3.64-5.04)	2.95 (2.59-3.36)
Negative likelihood ratio	0.41 (0.33-0.52)	0.38 (0.3-0.5)	0.45 (0.37-0.56)	0.44 (0.36-0.54)	0.43 (0.34-0.53)	0.38 (0.29-0.49)

2.4.4 Final model

The 18-variable RF2_s1 random forest model, where, for parsimony, variables with above average importance using the Gini index were used, was chosen as final model, henceforth referred to as the PIERS-ML model.

Figure 2.8 contains a plot of the Shapley values for a single random forest in the final model ensemble on its corresponding validation dataset. Shapley values represent the feature value's contribution to the individual's predicted probability. Positive Shapley values increase the predicted probability, negative values decrease the predicted probability and values of zero show no change to the predicted probability. The colour of each point represents the corresponding value of the feature, dark colour indicating a high value and light colour indicating a low value, while the height of the point clusters indicates the density of the points. The plot of these Shapley values shows any possible pattern between feature values and predicted probabilities. As the average predicted probability was low due to the low event rate, and most women had a low predicted probability, most values for each feature had a Shapley value close to zero. High values of serum creatinine, national MMR, aspartate transaminase and alanine transaminase, and low values of platelet count, oxygen saturation, national per capita GDP and haematocrit increased predicted probability. The biggest Shapley values were produced by a few observations of low oxygen saturation, while platelet count had the most non-zero Shapley values.

Figure 2.9 shows the relative importance of the 18 PIERS-ML model variables, based on the mean decrease of the Gini Index. This figure rates the features from 0–100, with 100 being the most important. Platelet count and oxygen saturation were of greatest and least importance, respectively. All target organ systems were included, with following numbers of covariates: cardiorespiratory (N=3), renal (N=3), hepatic (N=2), and haematological (N=4). In addition, there were variables representing health systems (N=2), demographic characteristics (N=2), and anthropometry (N=2).

From the remaining 2210 women, 1105 informed selection of cut-off points for the risk strata according to likelihood ratios, and the remaining 1105 women informed PIERS-ML model validation according to the selected cut-off points. Calibration was assessed on the 1105 women in the internal validation (Figure 2.11). Predictions on the validation and testing datasets, in increasing order, were binned together into ten groups of 221 predictions, event

rates (observed risk) were calculated along with confidence intervals and plotted against the mean predicted probability per group to create the dot and whisker plot. Smooth lines were plotted using the individual predicted probabilities and yes/no outcomes, with Linear (red) and Loess (grey) methods. Cox calibration intercept and slope, Brier score and Spiegelhalter z scores were calculated. While the linear calibration curve appeared to be close to an intercept of 0 and slope of 1, the Spiegelhalter p-value<0.05 showed the model to not be optimally calibrated. As calibration did not improve upon recalibration (Figure 2.13), we chose to prioritise stratification into risk classification groups to inform clinical decision-making. PIERS-ML accurately stratified risk for adverse maternal outcomes within two days, with an AUROC of 0.80 [95% CI 0.76 to 0.84] (Figure 2.10). As our data had a low outcome rate, a precision-recall plot was also included, with the minimum thresholds for each risk group indicated in the plot (Figure 2.12). The precision-recall curve of the PIERS-ML model on the validation data is very close to a straight line, meaning that there is no one probability threshold with both good precision and good recall. This was expected however, which is why we were not looking for a single cut-off point to create a treatment and a no treatment group, rather we looked to create multiple risk groups, where the high and very high risk groups have high precision to rule in an outcome, and the low and very low risk groups have high recall to rule out. Recall did not need to be high in the high and very high risk groups as not being predicted in these groups was not necessarily ruling out an outcome, and recall did not need to be low in the low and very low risk groups as not being predicted into these groups was not necessarily ruling in an outcome due to the inclusion of a non-informative moderate risk group. As expected, the low risk groups covered low precision and high recall thresholds, and the very high group had high precision and low recall, however the high risk group covered moderate to high precision and moderate to low recall.

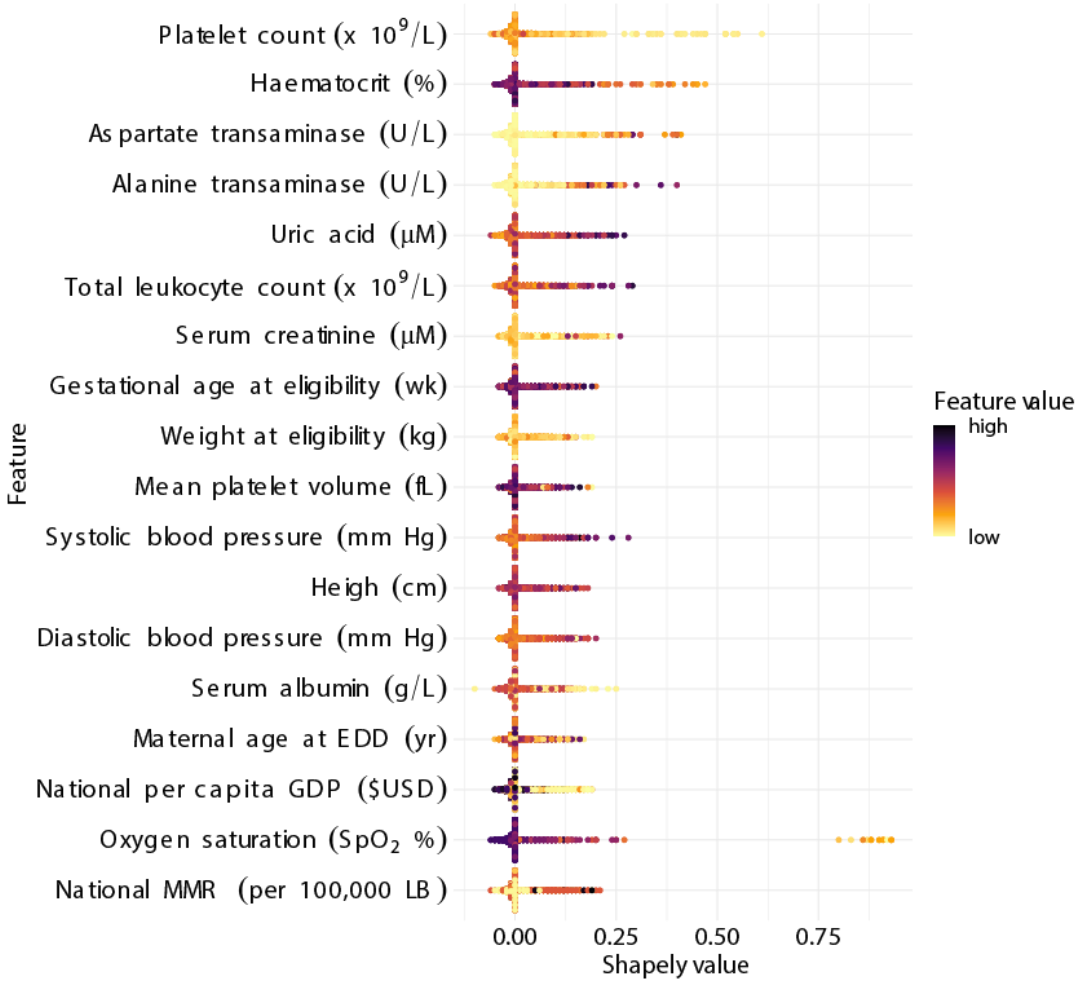

Figure 2.8: Shapley values for the PIERS-ML variables in a single random forest on the corresponding development dataset

* EDD, expected date of delivery; GDP, gross domestic product; LB, live births; MMR, maternal mortality ratio; SpO2, oxygen saturation by pulse oximetry; USD, United States dollar

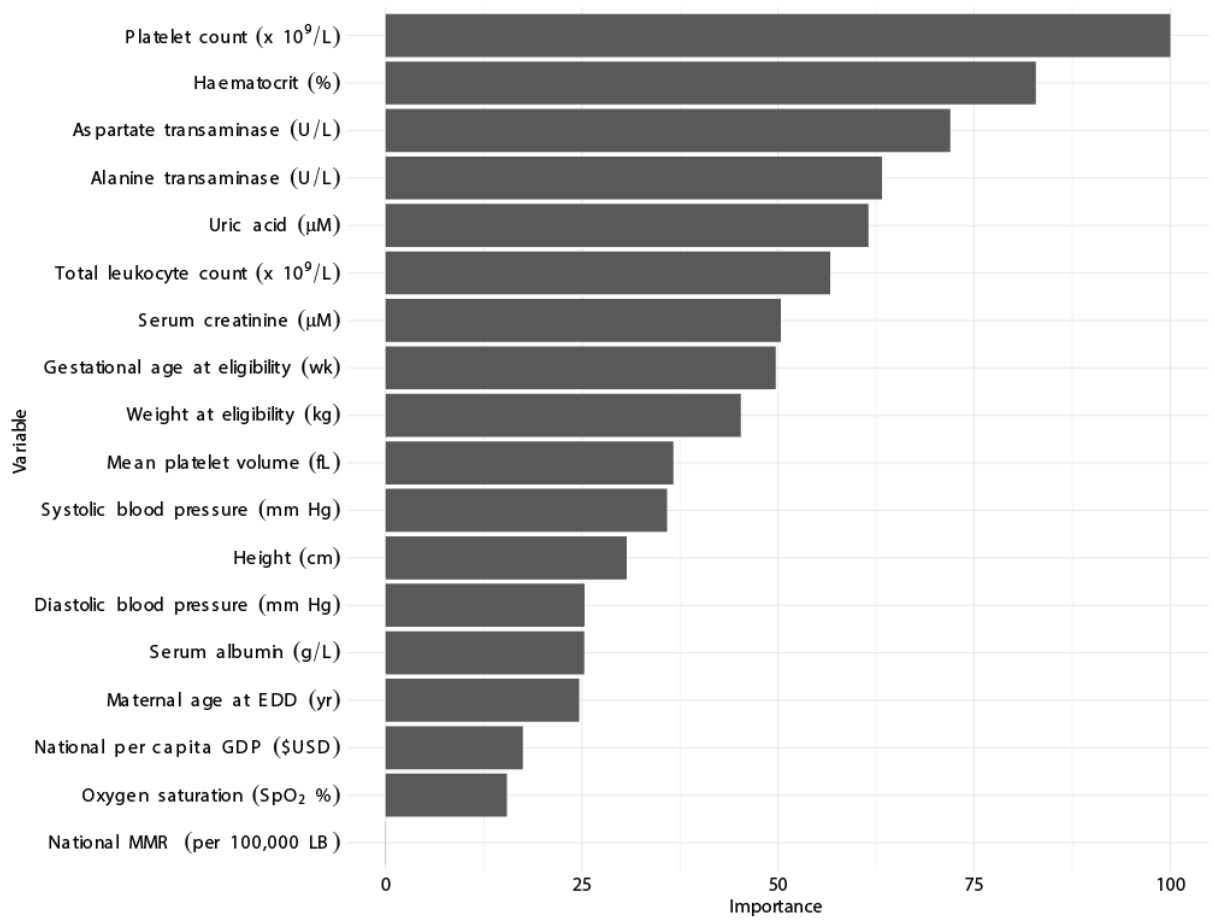

Figure 2.9: PIERS-ML variables ranked by importance within the random forest model based on Gini index, compared with the least important variable (National MMR).

* EDD, expected date of delivery; GDP, gross domestic product; LB, live births; MMR, maternal mortality ratio; SpO2, oxygen saturation by pulse oximetry; USD, United States dollar

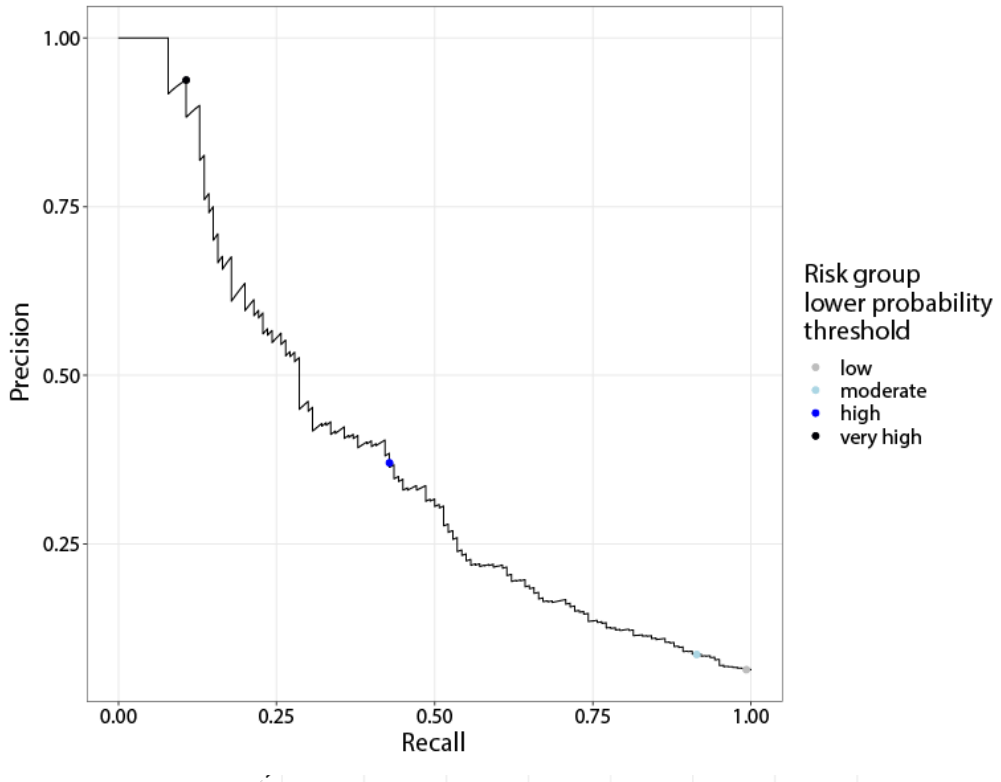

Figure 2.12: PIERS-ML precision-recall curve for adverse maternal outcomes within two days of initial assessment, using data within one day of initial assessment and prior to the occurrence of any outcome

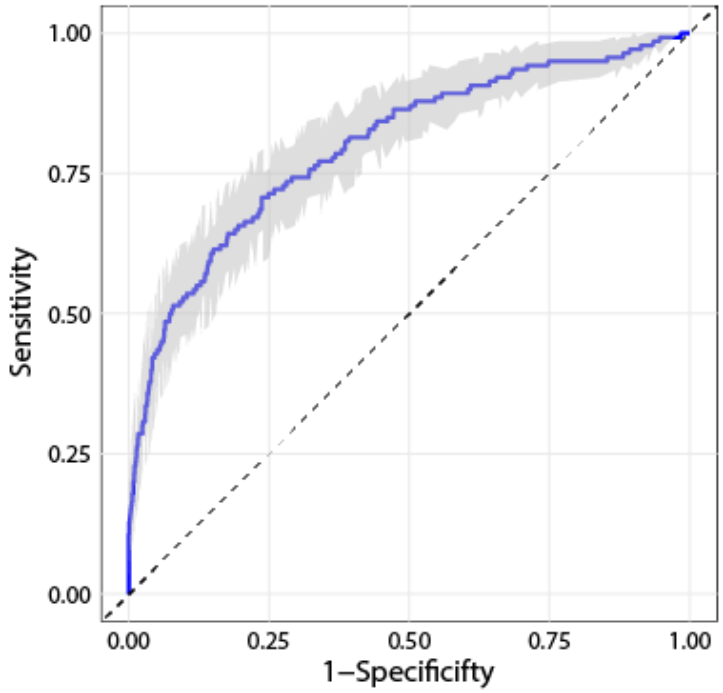

Figure 2.10: PIERS-ML receiver-operator characteristic curve for adverse maternal outcomes within two days of initial assessment, using data within one day of initial assessment and prior to the occurrence of any outcome

2.4.4.1 Secondary analyses

As part of the secondary analyses, predictive ability of the models of outcome within 7 days and at any point after first admission was also assessed. This has been presented along with the result of the primary analysis of predictive ability of outcome within 2 days in sections 2.4.3.2 and 2.4.3.3, with the results presented in Table 0.1 and in Appendix C. The remaining secondary analyses, namely assessment of predictive ability of occurrence of eclampsia and stillbirth are shown in Table 0.3 (Appendix C). There was no eclamptic seizures in the low- and very low-risk strata within 2 days, and no eclamptic seizures in the very low-risk stratum within 7 days or any point. Eclamptic seizures occurred in the low-risk stratum within 7 days, all in patients whose data was used for model development. Of the 9 patients who had eclampsia at any time in the low-risk group, 8 patients were in the development dataset, none were in the testing set and 1 was in the internal validation dataset. The highest rate of eclampsia was in the very high-risk, and the highest number in the high-risk group. This shows that the model also stratifies patients well in terms of risk of eclampsia.

No patient had a stillbirth in either the very low-risk or the very high-risk strata. No stillbirths in the very high-risk stratum could be explained by these patients delivering soon after admission as delivery is induced for those who seem high risk. The highest rate of stillbirth was observed in the high-risk stratum, while the highest number in the moderate-risk.

While not specified as part of the secondary analyses, I have also looked at improving model calibration, and the clinical utility of the PIERS-ML model, both for the original set of adverse outcomes. As seen in *Table 2.20*, the model chosen as the final PIERS-ML model (RF2_s1) did not have perfect calibration. Despite calibration being sub-optimal, we had chosen to prioritise risk classification groups to give a more accurate prediction of risk. As the risk categories were determined based on likelihood ratios from the observed risk in the testing dataset, the selected threshold values account for the overestimation of small risk by moving the thresholds for very-low, low and moderate risk close together, and the underestimation of large risk by having wider ranges for the high and very-high risk groups.

To try to improve calibration, Platt scaling and isotonic regression was tested for re-calibration, but calibration did not improve (*Figure 2.13*). This could be due to not enough data.

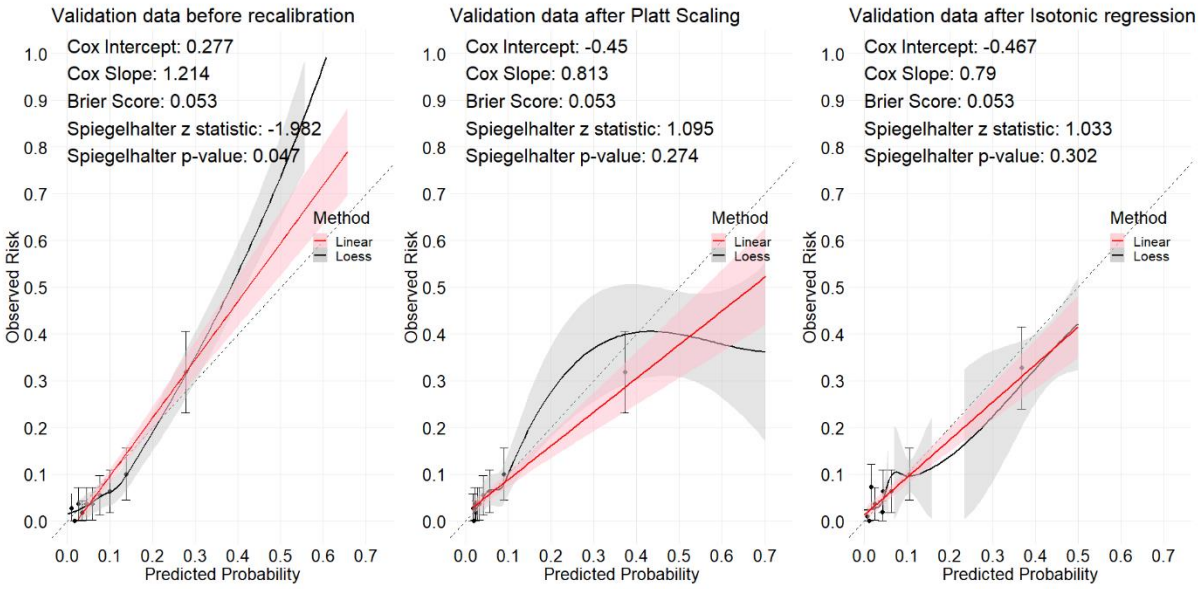

Figure 2.13: Calibration plots of original model (left), after Platt scaling (middle) and after Isotonic regression (right)

While our PIERS-ML model was not intended to be used with a single threshold value to sort patients into a treatment or no treatment group, decision curve analysis was carried out to visualise the clinical utility of our method (Figure 2.14). The model had greater Net Benefit than treating all and treating none between the ~3% and 75% predicted probability thresholds. Risk groups are marked on the plot with dashed lines. Treating very high risk only, high and very high, and moderate, high and very high all have a greater net benefit than treating all or treating none, while treating all has a greater net benefit than including low, or low and very low groups.

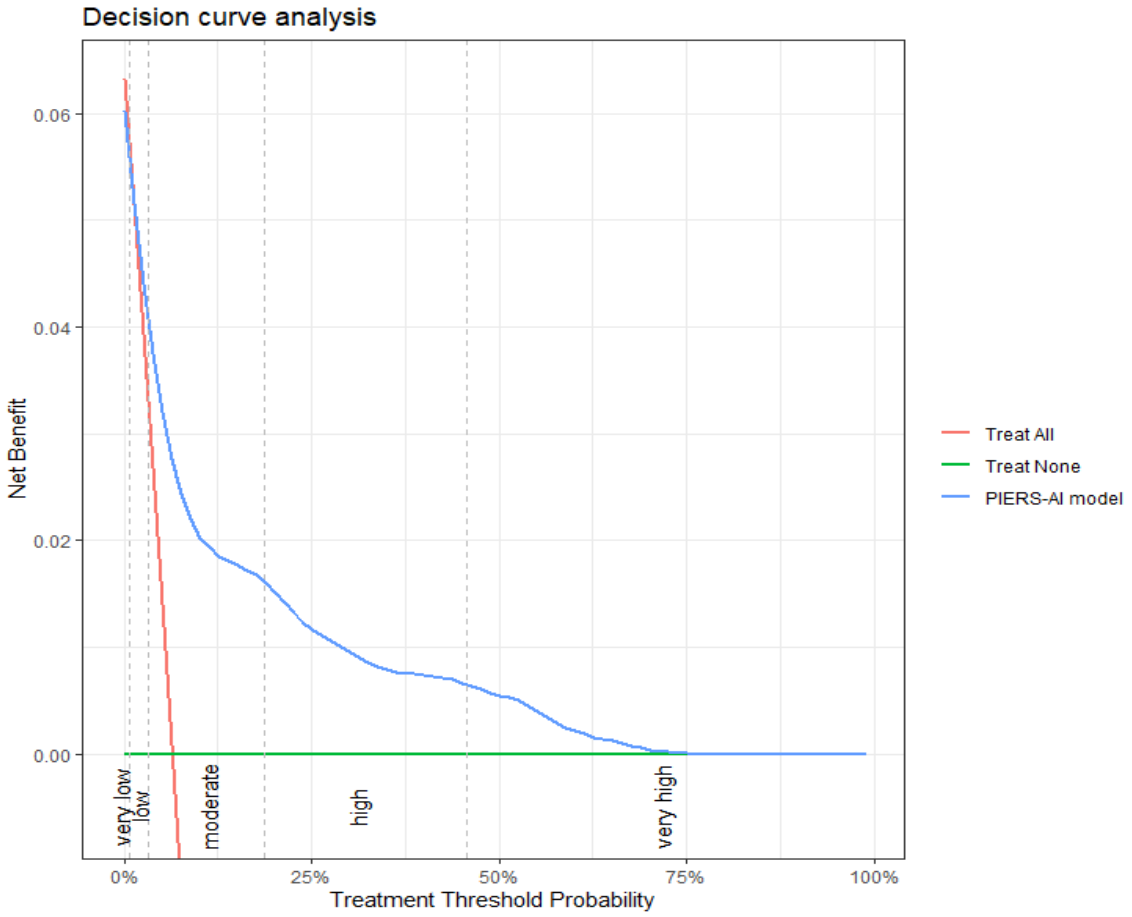

Figure 2.14: Decision curve analysis

2.4.4.2 External validation

The PIERS-ML model was also externally validated. The event rate in the external validation was 2.9% in 2 days, 3.4% in 7 days and 4.2% at any point, compared with 6.7%, 9.2% and 12.2% in the combined dataset, respectively. From the 2901 women in the external validation cohort, nine (0.3%), 1512 (52.1%), 1324 (45.7%), 52 (1.8%), and three (0.1%) were classified as being at very-low, low, moderate, high, or very-high risk, respectively. PIERS-ML accurately stratified risk for adverse maternal outcomes within two days, with women assigned to each stratum experiencing an outcome rate within the predicted range, and an AUROC of 0.76 [95% CI 0.71 to 0.82) (Table 2.21). While the event rates were lower in the external validation set than the development data, the model still correctly classified women at very-low and low risk, as well as those at high and very-high risk.

Table 2.21: PIERS-ML final model five group classification performance on external validation. (AUROC 0.76, AUPRC 0.17)

Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
			Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
Very high	45.6 - 100%	3 (0.1)	2 (66.7)	LR+ 67.9 [95% CI 6.2, 741.2]	2 (66.7)	LR+ 56.6 [95% CI 5.2, 618.8]	2 (66.7)	LR+ 45.9 [95% CI 4.2, 503.1]
High	18.7 - 45.5%	52 (1.8)	17 (32.7)	LR+ 16.9 [95% CI 9.9, 28.9]	20 (38.5)	LR+ 18 [95% CI 10.7, 30.4]	20 (38.5)	LR+ 14.6 [95% CI 8.6, 24.7]
Moderate	3.1 - 18.6%	1324 (45.7)	47 (3.5)	-	55 (4.2)	-	65 (4.9)	-
Low	0.6 - 3%	1512 (52.1)	17 (1.1)	LR- 0.4 [95% CI 0.3, 0.6]	22 (1.5)	LR- 0.4 [95% CI 0.3, 0.6]	34 (2.2)	LR- 0.5 [95% CI 0.4, 0.7]

Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
			Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
Very low	0 - 0.5%	9 (0.3)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]

2.4.4.3 Complete case analysis

Of the 2210 women in the internal validation dataset, 351 had no missing values for any of the PIERS-ML or fullPIERS variables. To test model performance with no imputations, the PIERS-ML model, the original fullPIERS model and the refitted fullPIERS model were all tested on these complete cases. The complete cases had a lower outcome rate (4.3% in 2 days, 5.4% in 7 days, 6.6% at any point) than the whole of the dataset (6.65, 9.1% and 12.2% respectively). This is in line with the previous observation that patients with an outcome were more likely to have missing values. The PIERS-ML model performed better on the complete observations than the imputed validation datasets.

While PIERS-ML model is intended and expected to be used with all variables available for a patient, to simulate a practical example of the use of the model for a patient with missing data, I applied mean imputation to the validation dataset. The mean of each variable was calculated in the development dataset, and missing values in the validation data were replaced with the corresponding mean values. Both the number of patients predicted into each group and the number of outcomes per group were very similar between the original validation and the validation with mean imputation.

The results of both complete case analysis and mean imputation are presented in Table 0.4 (Appendix C).

2.4.4.4 Performance in high- vs low- and middle-income countries

The reported performance of our PIERS-ML model is an average across countries, however the included country-specific features are intended to correct for the potential differences between countries. We assumed that low- and middle-countries (LMICs) have a higher base risk of adverse maternal outcomes than high income countries (HICs). The data appear to align with this assumption as among the participants in LMICs, the outcome rate in two days was 13.5%, while in the HICs it was 4.8%. Since this difference may or may not be reflected in the clinical predictor variables, a categorical variable for country of study site was included for the multiple imputation along with GDP per capita and maternal mortality ratio. The country variable was not used for modelling so that the model can be applied to data from countries not included in the model development.

To test our hypothesis that the country-specific variables correct for these potential differences, the model's performance measurements were also calculated in the LMIC and HIC subsets of the validation data separately (Appendix C, Table 0.5). The model performance reflected the difference in outcome rates as in the LMIC subset much lower percentage of the participants was predicted in the low risk group and none in the very low risk group, while much higher percentages were predicted in the high and very high risk groups than in the HIC subset.

2.4.4.5 Systematic minority oversampling

The PIERS-ML model's five group classification performance was tested in a dataset created by applying SMOTE to the combined testing and internal validation dataset (Appendix C, Table 0.6). The dataset had 2210 patients before applying the method, 140 (6.3%) of whom had an outcome within 2 days, and 2070 (94.7%) did not. The dataset created by SMOTE had 1820 data points, with 980 (53%) outcomes and 840 (46.2%) no outcomes. In this more balanced dataset, the model achieved the desired likelihood ratios in every risk strata for outcome in 2 days, 7 days and at any time. The outcome rate in the very low-risk stratum was 0%, and 7.3% in the low-risk stratum, which was much lower than the outcome rate in the dataset. Moderate risk was an uninformative class as expected, having an outcome rate very close to the outcome rate of the whole dataset, while high and very high-risk strata had an outcome rate of 90.2% and 99.2%, respectively.

I have also refitted the model in a SMOTE'd development dataset, which, instead of the 6633 patients with 450 (6.8%) outcomes and 6183 (93.2%) no outcomes, had 5850 data points with 3150 (53.9%) outcomes and 2700 (46.1%) no outcome. The model was then tested in the original combined testing and internal validation datasets, however the model performed worse than the PIERS-ML model. The refitted model could not classify any patients into the very low-risk stratum, and none of the other strata had the desired likelihood ratios or outcome rates.

2.5 Discussion

2.5.1 Summary of findings

The analysis in this chapter included data from 8843 women from 11 countries presenting for first assessment of preeclampsia, and used multiple machine learning and feature selection methods to develop and internally validate models for maternal risk stratification, applicable for LMICs and HICs. I used five machine learning enabled modelling methods – gradient boosted trees, random forest, LASSO and ridge regression, and artificial neural network – and a Bayesian modelling method – Bayesian model averaging. These methods were combined with variable selection methods where appropriate to find the best performing model using the smallest number of variables. A further 2901 UK-resident women contributed data to a fully external validation dataset, which was used to validate the model selected as the final model.

Initially 16 models were created, with a further 6 fitted during sensitivity analyses and 1 more during systematic oversampling. When using the models to classify patients into an outcome and a no outcome group, all models had good discrimination with an AUC value close to 0.8, and were slightly better at ruling out an outcome with specificity >0.7 and negative predictive value at 0.97, than ruling in. However, the difference in probability of outcome between the two groups was very small for all models, with a positive likelihood ration between 2 and 5, meaning only a small increase in probability of outcome for the outcome group, and a negative likelihood ratio ~0.5. The area under the precision-recall curve was also low for all models. The sensitivity analysis removing fibrinogen and APTT from the predictor variables did not significantly change this performance. Based on these results, the two-group outcome/no outcome prediction should not be used to inform clinical care.

Random forest based models performed best in the five group classification, balancing number of patients in the very high-risk stratum with the outcome rate and likelihood ratio well, and having no outcomes in the very low-risk stratum. While the models identified predicted probability thresholds for the high- and low-risk strata, these strata generally did not reach the expected likelihood ratios in the validation data, but still managed to have an acceptable outcome rate and a good proportion of patients in the strata. This performance was maintained after sensitivity analysis.

The final model selected for modelling of the composite adverse outcome was the random forest model using above average importance variables, without fibrinogen or APTT, including MPV (RF2_s1), further referred to as PIERS-ML. The PIERS-ML model identified nearly 40% of women with preeclampsia for whom care should be altered. The 29.8% of women (and their families and maternity care-providers) identified as being at very-low (0.7%) or low risk (29.1%) can be reassured that it is very unlikely that adverse maternal events will occur within two days. However, for the 8.9% of women identified to be a high (7.9%) or very-high risk (1.0%), a timely clinical response can be justified, based on a substantial risk of an adverse maternal event within two days, or for women at very-high risk, at any time. Identifying these women can inform discussions about place of care, transfer of care, antenatal and postnatal surveillance, co-interventions, and timed birth.

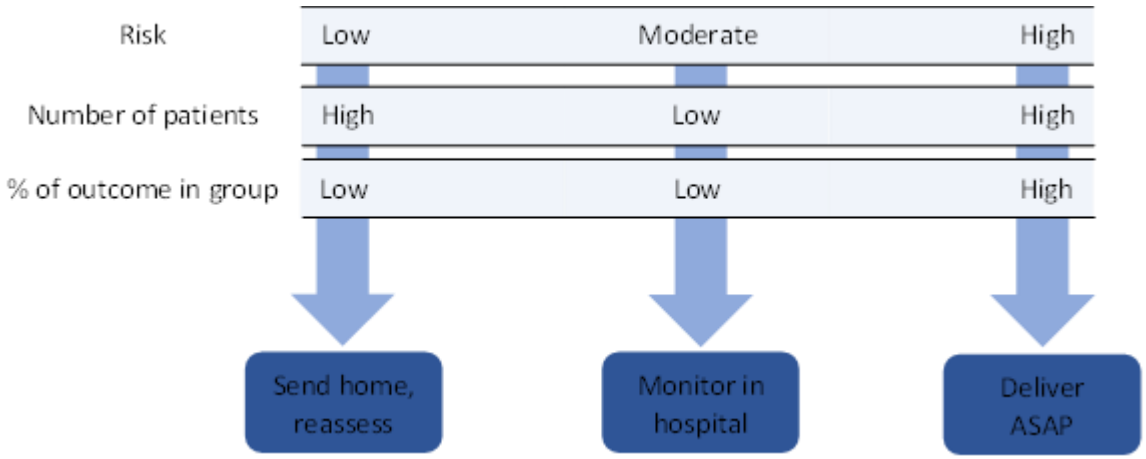

Figure 2.15 Risk group classification

2.5.2 Strengths and limitations

Strengths of our study include the large sample size – much larger than any previous study modelling adverse outcomes in women with preeclampsia. (3, 29, 41, 117, 118)

I tested a list of variables with clinical external validity and availability, including all target organ systems, with the exception of the central nervous system. Clinical central nervous system predictors rely on either the subjectivity of symptoms or the questionable reproducibility of deep tendon reflexes/clonus, particularly in pregnancy. Also, the PIERS-ML model does not include direct measures of coagulation; these are not routinely performed, and their addition neither altered model performance nor

warranted related costs in women with preeclampsia. Machine learning is suitable for managing a large number of variables, without assumption with respect to interactions and mediation, and addresses concerns regarding collinearity. Trade-offs were considered between model performance, complexity, and face validity. In addition, machine learning algorithms can use all available predictor variables as inputs, as variables that are not predictive of the outcome will have little to no impact on the predicted probabilities. Therefore, a machine learning-based model with more variables than necessary should not have degraded performance in the context of this application. However, this means that included variables may not contribute significantly to the predictions. Using feature selection to remove unimportant or less significant variables can help to balance model performance and complexity by identifying redundant variables while retaining model performance as less-informative variables are censored. A less-complex model will be easier to use in clinical practice and may reduce the costs associated with laboratory-based variable collection.

There are significant differences between our PIERS-ML model and the recently-published Charité machine learning-based model.(33) Schmidt and colleagues using data from 1647 women with preeclampsia admitted to a single HIC institution, and modelled against a composite maternal and fetal outcome, despite rates of adverse maternal and perinatal outcomes not tracking together.(119, 120) Although they applied different machine learning methods, they used approximately 24 observations (internally normalised to multiples of the median) per candidate variable, compared with approximately 230 observations per variable in the PIERS-ML dataset. They undertook no imputations, thereby assuming that missingness is predictive of either outcome or no outcome. In addition, I fitted models using 10-fold cross-validation to minimise overfitting, and validated the model in an independent quarter (vs 10% [Charité]) of the dataset. Importantly, our risk-stratification was data-driven in ruling-out or ruling-in the risk of adverse maternal outcomes. The main strength of their study was the inclusion of angiogenic markers, while the main strengths of ours are the diversity of data, the use of multiple imputation, and external validation.

Missing values were common in our dataset; however, where appropriate, missingness was handled by multiple imputation, with chained random forest to minimise bias. Of note, development and validation datasets were imputed separately, thereby creating

independent model development and validation datasets. Some variables (namely bilirubin, urinary protein to creatinine ratio, international normalised ratio, and lactate dehydrogenase had to be excluded due to high levels of missingness. This limitation reflects current clinical practice. Modelling on complete data from a representative and diverse sample population will always give a more accurate result than modelling on a dataset with a significant amount of missingness. However, obtaining such a dataset would be very difficult as many of these variables are not regularly collected in clinical practice from patients with preeclampsia, especially close to term. Additionally, international normalised ratio and lactate dehydrogenase were not expected to be strong predictors, and while urinary protein to creatinine ratio would be anticipated to outperform the readily-available dipstick proteinuria, it requires a 24 hour urine collection which is often not available.

The composite adverse maternal outcome for PIERS was Delphi-derived, similarly to the core maternal outcome list (iHOPE) for preeclampsia.(121) However, there are differences: only PIERS includes uncontrolled hypertension, inotropic support, myocardial ischaemia or infarction, hepatic dysfunction, or transfusion, and only the iHOPE outcome set contains elevated liver enzymes, postpartum haemorrhage, and admission to intensive care. To confirm model performance, as some factors are both predictors and components of the combined outcome (e.g., serum creatinine, platelet count), I assessed PIERS-ML excluding renal or haematological components of the outcome, or both (Table S5).

The PIERS-ML model also proved to be highly effective on external validation in a UK-resident cohort, achieving the predicted outcome rates in the risk classification strata expected from internal validation. The model could identify women for whom a timely response can be justified very well, while reassuring women predicted very-low risk who had no outcomes at any point. However, the full range of the components of the combined maternal outcome were not readily available, and some outcomes (e.g., renal) were conflated within the dataset as those data had been collected previously. While this introduces the possibility of underestimating the outcome rate as defined by the original PIERS combined outcome, we remain confident in the model's ability to classify into risk strata, especially for ruling in outcomes in the high and very-high risk strata.

2.5.3 Interpretation in light of existing literature

Multivariable model-based risk stratification of women with preeclampsia is recommended by national and international clinical practice guidelines.(9, 43, 122) With access to full laboratory facilities, models have been based on either logistic regression or a survival model for time-to-adverse event. (3, 118) Women with a hypertensive disorder of pregnancy (including preeclampsia) in LMICs, without ready access to laboratory tests, benefit from the demographics-, symptom-, and signed-based miniPIERS model, with model performance improved by pulse oximetry.(29, 30)

The PIERS-ML model improves on prior models in a number of ways.

First, PIERS-ML is the first of model for risk stratification in preeclampsia that has been developed using machine learning from women with preeclampsia living in LMICs (sub-Saharan Africa, South America, South Asia, and Oceania – areas of the world where more than 99% of preeclampsia-related maternal mortality occurs)(46) and HICs; we are unaware of another model that has included data from 11 less- and more-developed countries.

Uniquely, the model includes national per capita gross domestic product and maternal mortality ratio; variables that adjust for location, avoiding adaptation of models to local outcome rates, particularly where information governance and research resources are absent or limited. Adjusting for local settings was required for original fullPIERS model validation in LMICs.(42) Therefore, we were not surprised by the poor performance of fullPIERS within this combined dataset. This approach to create auto-adjustment for setting should be validated in further geographies.

Second, PIERS-ML does not include maternal symptoms, the inclusion of which was criticised as a weakness of prior models,(3) given the subjective nature, variable definitions, and inconsistent documentation of symptoms in health records. However, mean platelet volume, a marker of platelet consumption and release of immature platelet forms, is important within the PIERS-ML model.(123) Haematology analysers routinely measure, but many laboratories do not report, mean platelet volume; our findings suggest mean platelet volume should be reported for all hypertensive pregnant women.

Third, although PIERS-ML does not include symptoms of central nervous system involvement, the model has clinically-relevant performance in identifying women at both least (very-low and low risk strata) and greatest (high and very-high risk strata) risk of developing eclampsia, especially within 7 days of initial assessment (Table 4). These results could guide the targeted use of magnesium sulphate, and reduce the number-needed-to-treat, for the prevention of eclampsia.(124) Depending on health system resilience, women in the moderate risk strata may or may not be considered for magnesium sulphate prophylaxis. For women with disease onset before 34+0 weeks' gestation, a loading dose of magnesium sulphate should be administered to reduce the risk of prematurity-related cerebral palsy.(125)

Fourth, women in the very-low and low risk strata were very unlikely to suffer a stillbirth (Table 4), and we believe that such women can be appropriately reassured by our model. All intrauterine fetal deaths were noted within two days of admission with preeclampsia. However, it was notable that the women in the moderate risk and high risk strata bore the greatest risk of stillbirth, presumably as maternity care providers were sufficiently concerned by the condition of women in the very-high risk stratum to intervene for either maternal or fetal indications, or both, or in response to the woman experiencing an adverse maternal event. In addition, the very-high risk stratum had a limited sample size; therefore, performance in predicting stillbirth could not be accurately assessed.

Finally, the PIERS-ML model has been externally validated with very good performance in an independent cohort of women with preeclampsia admitted to hospitals that did not participate in the fullPIERS development and validation projects, and where neither fullPIERS nor miniPIERS were in routine clinical practice.

There are three important lines of investigation that follow from our work.

The first is that use of this method offers the potential for the accuracy of the model to improve over time as data accumulate and the model 'learns' with regularly scheduled manual updates provided to model users. Amassing such data is feasible, as all individual-level variables in PIERS-ML are part of routine clinical and laboratory assessment of women with preeclampsia in well-resourced settings, and available from electronic health records in real-time.

Second, to explore whether addition of new markers may improve PIERS-ML model performance. As markers of uteroplacental dysfunction of preeclampsia, angiogenic markers are used increasingly in the investigation of women with suspected preeclampsia, at first presentation, for ongoing surveillance(126) (although performance in trials has been variable(127, 128)), and within the Schmidt et al. model.(33) If independently informative, angiogenic markers could be incorporated into the model during clinical implementation. Ophthalmic artery Doppler may provide useful information about the less-accessible intracranial circulation.(129)

Third, to determine how best to assess evolving risk among women with preeclampsia, especially those women at moderate risk who will require ongoing, close surveillance, particularly during the first seven days when most adverse outcomes occur. PIERS-ML, and other models, perform best over the first two-to-seven days; women with preeclampsia may be expectantly managed for up to four weeks.(130) While repeated risk stratification has been recommended,(122) future work should examine replacing this 'serial static model' approach with a new 'dynamic' approach that accounts for changes over time.

2.6 Conclusion

In conclusion, we used the power of machine learning to develop a new effective risk stratification tool for international use, PIERS-ML, which can accurately advise care for patients with preeclampsia on first admission and provide an assessment of risk of adverse outcomes within 2 days, and in some cases up to 7 days. The model significantly improved on the fullPIERS model, recommended by the National Institute of Health and Care Excellence (NICE) for use of risk assessment of adverse outcomes of preeclampsia. While this model works well for short term risk stratification, able to classify about half of all patients into actionable groups, a dynamic model is needed to assess longer term risk.

Chapter 3: Modelling binary outcome with longitudinal data

3.1 Introduction

As Identified previously (Section 2.5.3 of Chapter 2), there is a need for a dynamic model to assess long term risk of maternal outcome, as neither the PIERS-ML model presented in Chapter 2, nor any of the previously published models have been developed to use longitudinal observations and address the evolving risk over time. Moreover, our recently completed work currently under review with BJOG: An International Journal of Obstetrics and Gynaecology proved that the performance of fullPIERS and PIERS-ML models, developed on data from day of admission, declines rapidly beyond two days from admission. The analysis in this chapter aims to address this need for a dynamic model.

This chapter is structured as follows. Section 3.2 presents the methodological background to dynamic modelling and dynamic modelling methods. Section 3.3 covers the methods used to apply these techniques to our data of preeclampsia patients. In Section 3.4, the performance of the models created in Section 3.3 is compared on their calibration, discrimination, and stratification ability. Finally, Section 3.5 discusses the implications of our study, including its strengths and limitations, its interpretation in light of the existing literature, and the avenues for further research.

3.2 Methodological background

3.2.1 `Dynamic` modelling and consecutive prediction

The previous chapter explored modelling methods to predict a binary outcome using observations from a single point in time. We have referred to this method of modelling as `static` modelling as it uses only one static point in time to predict all future outcomes.

In many cases of real-life applications of predictive modelling, multiple observations may be available of the predictor variables over time. Static models, while only trained and tested on a single set of observations, can be used repeatedly on multiple observations to create new predictions using the latest available set. We will refer to this as consecutive prediction. There are many drawbacks of using static models for consecutive prediction, as while it may be very simple to use, the model will only use one set of observations from the available data and won't be able to account for the magnitude and rate of change in the predictor variables for an individual over time. Moreover, as the models are usually only tested for the set of

observations from a pre-defined timepoint (for example first or last observations), a maintained performance over time is not guaranteed. Static models should be tested for consecutive prediction and recalibrated at different timepoints as appropriate, or other methods should be considered where including multiple observations at a time is a possibility.

We define `dynamic` modelling in this chapter as a statistical or machine learning method using longitudinal data to predict an outcome.

3.2.2 Mixed effects models

Before moving on to dynamic modelling methods, fixed and random effects, and the use of mixed effects models should be discussed. In the methods discussed in Section 2.3 in Chapter 2, all variables and/or interactions had a fixed effect – seeing the same values in two patients would always result in the same prediction. When a variable has the same effect across patients and over time, and all possible values of the variable are observed in our data, we call this a fixed effect (131, 132). However, this may not always be the case. If we have repeated measures of a variable, the effect could change over time. We could have variability between individuals, or predictor variables that don't contain all the values or levels present in the population (132, 133). Variability between individuals is likely in many cases, as the observations from a single individual are likely to be correlated with each other in a way that the same observation between different individuals are not. In these cases, we cannot accurately model the effect the variable has on the outcome by a single fixed parameter, and therefore we need methods that allow us to model this correlation within individual that is different from the between-individual correlation. When the effect a variable has on the outcome is not a single fixed value, but the effect itself varies, we call this a random effect (131). Random effects can be coefficients of variables or interaction terms - these are called random slopes - or the intercept/baseline risk - called a random intercept (133).

For example, let us consider a case when the outcome, *y*, is predicted by a baseline risk and a continuous predictor variable, *x*. Figure 3.1 shows the four possible combinations of random and fixed effects for the intercept and slope, based on data for three patients. In scenario A, the baseline risk (y(x=0)) is 2 for all patients, and y increases by 1 for every unit of 1 increase in x for all patients. As there is no variability in the intercept, this is a fixed effect. The effect of x also does not vary between the patients, and we can assume that all values of the variable have been observed, therefore this is also a fixed effect. In scenario B, the slope of y(x) stays

constant among the patients, hence the effect of x is still a fixed effect, however patients have different baseline risks. This means that scenario B shows a random intercept, which we can model as a random variable from a specified distribution (for example normal distribution with mean μ and standard deviation σ).

In scenario C, all patients start with the same baseline risk (fixed intercept), but y increases at different rates per person from 1 unit of change in x, suggesting a random slope. Finally, scenario D shows a random slope and a random intercept, where patients start with a different baseline risk, and y increases at a different rate per person.

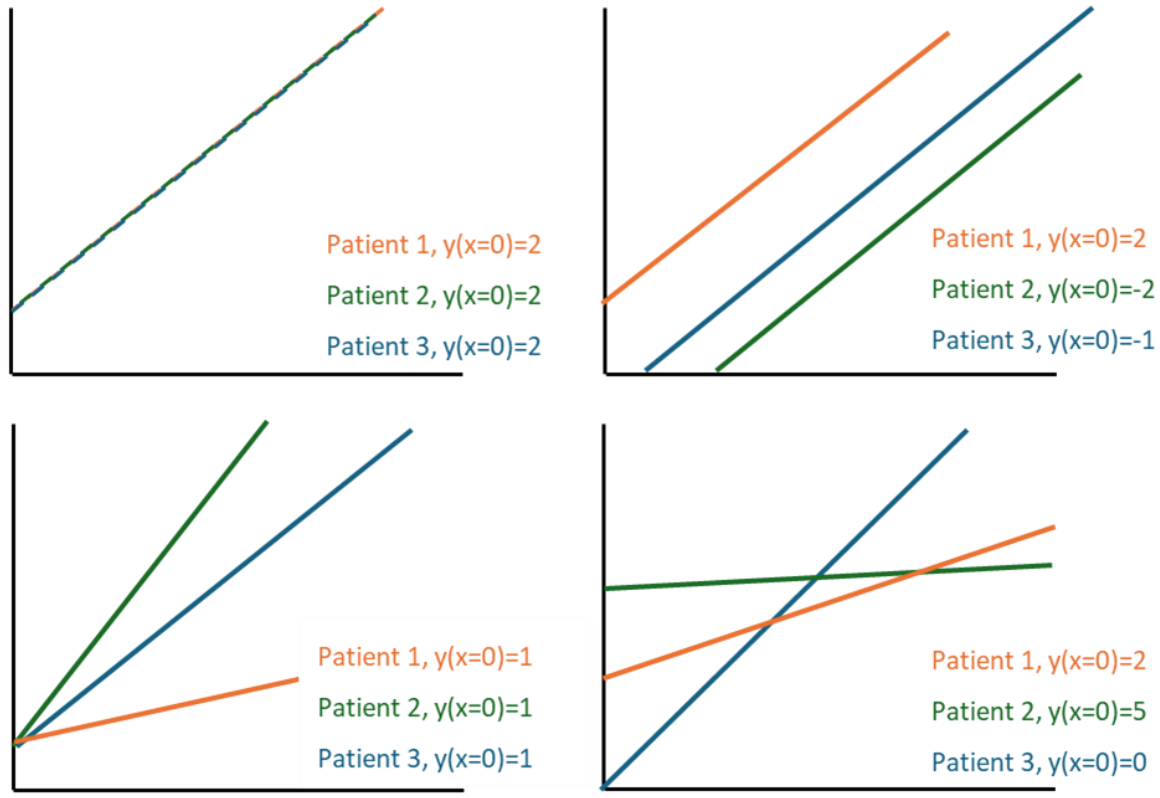

Figure 3.1: Visualisation of random intercept and slope

However, the illustration in Figure 3.1 is a very simplified view of an ideal case of modelling with no random error term. In reality, even when working with fixed effects, we also have to consider noise in our data, meaning that data points often don't fit on a simple line, and in most cases, we will be working with more than one predictor variable, making visualisation more difficult. Hence, in most cases, when to consider fixed vs random effects will have to be based on expected variability in the data.

A variety of modelling methods make use of fixed and random effects to model outcomes based on longitudinal predictors.

3.2.3 Two-stage model

Two-stage models separately model the change in predictors variables over time, and the outcome variable (134). In the first stage, a mixed effects model is created for each longitudinal predictor variable, using a random intercept and time as a random effect, in addition to any other relevant variables as fixed effects. The random effects are extracted from these models for each patient and each time point, where the random effect for any time point is calculated by using data up to that time point for the individual patient. These random effects are saved along with any static predictor variables.

In the second stage, fixed effects logistic regression is used to model the binary outcome using the intercept and time effect from the stage 1 models and any non-longitudinal predictors as the variables of the model.

Equation 3.1, Equation 3.2, and Equation 3.3 – where $X_{1_{it}}$ and $X_{3_{it}}$ are longitudinal predictors of the *i*th patient at time *t*, X_{2_i} and X_{4_i} are static predictors of the *i*th patient, β_j and α_j are fixed effects, b_{ji} are random effects, and ε_{it_j} are error terms – demonstrate the process of creating a two-stage model of a binary outcome using two longitudinal variables and two static variables (134).

$$X_{1_{it}} = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) * time + \beta_2 * X_{2_i} + \varepsilon_{it_1} \quad \textit{Equation 3.1}$$

$$X_{3_{it}} = \beta_3 + b_{3i} + (\beta_4 + b_{4i}) * time + \varepsilon_{it_2} \quad \textit{Equation 3.2}$$

$$logit\big(P(outcome_{it} = 1)\big) = \alpha_0 + \alpha_1 * \hat{b}_{0it} + \alpha_2 * \hat{b}_{1it} + \alpha_3 * \hat{b}_{2i}t + \alpha_4 * \hat{b}_{3it} + \alpha_5 * X_{4_i} + \varepsilon_{it_{333}} \quad \textit{Equation 3.3}$$

To make predictions for a new patient with a two-stage model, longitudinal data from the new patient has to be added to the development dataset, which is then used to repeat the fitting process of the mixed effects models. The random effects for the new patient only are extracted from the mixed effects models and combined with the patient's static data. The logistic regression model of stage 2 is then used to create a predicted probability from the new patient's data.

3.2.4 Bayesian joint modelling

3.2.4.1 Bayesian joint modelling of a non-longitudinal binary outcome

Similarly to the two-stage modelling method, Bayesian joint modelling is based on using a combination of mixed effects models to describe the longitudinal predictors, and a logistic regression model for the binary outcome variable (134). In joint modelling, however, the random effects in the mixed effect models are estimated simultaneously in the two models they are used in, rather than estimating one model after the other.

As the name indicates, this method of modelling is a Bayesian method, hence the distribution of the parameters in both models is estimated by using some prior distributions and combining them with the likelihood of the observed data to create the posterior distribution. The prior distributions can be inferred from previous studies or expert opinion, but if no useful prior information is available, uninformative prior distributions should be used. In particular, we need to specify a prior distribution of the model parameters of each model, the residual error variance, σ_e^2, and the covariance matrix of the random effects, *D*.

To obtain the posterior distribution, we first need a likelihood, which is calculated by the joint likelihood of the models that share parameters (Equation 3.4), where *L* is the likelihood function of the logistic regression model (Equation 3.5), p_i is the predicted *probability* of outcome from the logistic regression model, and φ is the density function of a normal distribution, with mean of 0 and variance σ_e^2 for the error term of the logistic regression model, ε_{it}, and mean of 0 and covariance matrix *D* for the vector of random effects, $\boldsymbol{b}_i$ (134).

$$likelihood = \int L * \prod_{i=1}^{n} \prod_{t=1}^{T} \varphi(\varepsilon_{it})\, \varphi(\boldsymbol{b}_i)\, db_i \qquad \textit{Equation 3.4}$$

$$L = \prod_{i=1}^{n} (p_i{}^{outcome_i}(1 - p_i)^{1-outcome_i}) \qquad \textit{Equation 3.5}$$

The posterior distribution is proportional to the joint likelihood multiplied by the prior distributions of the mixed effects and logistic regression models, meaning that the random effects are estimated simultaneously from the logistic and mixed effects models.

3.2.4.2 Bayesian joint modelling of a time-to-event outcome

When outcomes occur over time, rather than modelling a binary outcome, we can use models predicting the expected time until an outcome occurs. This is called a time-to-event outcome,

and we can use Bayesian joint modelling to model this outcome by using a Cox regression model instead of logistic regression.

Following the same Bayesian methodology, the posterior distribution is obtained from a combination of the prior distribution of model parameters, and the joint likelihood of the Cox regression model predicting the outcome, and the mixed effects models predicting the longitudinal variables. This joint likelihood has been derived by various studies (135, 136), and implemented in the JMBayes2 R package (137).

3.2.5 Mixed effects random forest

The previously described methods combined generalised linear mixed effects models (GLMMs) with regression to predict a binary or time-to-event outcome using longitudinal variables. Binary Mixed Model (BiMM) forests (138) use Bayesian GLMMs and combine them with random forests instead by modelling the fixed components with the random forest, and the random components with the Bayesian GLMM part of the model.

The BiMM forest is fitted through an iterative process.

First, a random forest is fitted using the fixed predictors ($\boldsymbol{X_{it}}$) for each patient (*i = 1, 2, …, N*) at each timepoint (*t = 1, 2, …, T_i*), using y_{it}, the binary outcome, as the outcome of the model. This stage of the model fitting considers each observation per patient as a separate data point and therefore does not take previous observations per patient into account.

Next, the predicted probability for each patient and time point is extracted from the random forest ($p_{RF}(\boldsymbol{X_{it}})$) and a Bayesian GLMM is fitted as shown in Equation 3.6, where Z_{it} is a clustered variable and b_{it} is the random effect.

$$logit(y_{it}) = \beta_0 + \beta_1 * p_{RF}(\boldsymbol{X_{it}}) + Z_{it} * b_{it} \qquad \textit{Equation 3.6}$$

The predicted probability for each patient and time point is, again, extracted from the Bayesian GLMM ($p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})$). As a final step, a new target outcome y_{it}^* is created by adding $p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})$ to the previous target outcome y_{it} for each *i* and *t*. As y_{it} is a 0/1 binary value and $p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})$ is a probability between 0 and 1, adding the two together creates a continuous numeric value between 0 and 2. To turn this into a binary value again, we apply a predefined split function *h*().

$$y_{it}^{*} = h(y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})) \qquad \textit{Equation 3.7}$$

y_{it} is replaced by y_{it}^{*} and the whole fitting process is repeated until the posterior log likelihood from the Bayesian GLMM is less than a prespecified value. Once this condition is met, the iterative process is stopped, and no new target outcome is created.

The split function can be a simple form such as Equation 3.8, which simply uses a threshold *k* to separate the values by classifying number greater than *k* as 1s and less than or equal to *k* as 0s. It can also take a slightly more complicated form, such as Equation 3.9, where we classify as 1s or 0s for values greater than 1.5 or less than 0.5, respectively, while for all other values, we assign the probability of $p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})$ that $y_{it}^{*} = 1$ (and, as it is a binary value, the probability that $y_{it}^{*} = 0$ is $1 - p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})$) (138).

$$h\big(y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})\big) = \begin{cases} 1 \; if \; y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it}) > k \\ 0 \; otherwise \end{cases} \qquad \textit{Equation 3.8}$$

$$h\big(y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it})\big) = \begin{cases} 1 \; if \; y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it}) > 1.5 \\ p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it}) \; if \; 0.5 < p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it}) < 1.5 \\ 0 \; if \; y_{it} + p_{BGLMM}(\boldsymbol{X_{it}}, Z_{it}) < 0.5 \end{cases} \qquad \textit{Equation 3.9}$$

3.2.6 Neural network

We have previously introduced artificial neural networks in Chapter 2, Section 2.2.8. While we referred to the method discussed in this section as an "artificial neural network" in general, what we have presented so far was only one type of ANN, called a feedforward ANN.

A feedforward artificial neural network is the simplest form of artificial neural network, where nodes of a layer are connected to some number of nodes in the next layer, information is fed into the first layer of the network, which is the input layer, and it is fed forward layer by layer until it reaches the last layer, where an output is created.

Feedforward neural networks are more suited for static modelling than dynamic as they process sequences of observations as independent – the predicted probability for each observation is not affected by the order in which the observations are provided.

3.2.6.1 Recurrent ANN

Recurrent artificial neural networks (RNNs) are better suited for sequences of inputs than feedforward neural networks, as they contain a unique component that passes some information from previous inputs to the next one. The name comes from the RNN's defining

feature, the recurrent unit, which combines the input to the neural network with past information. The recurrent unit can be as simple as a single node or layer, but it can take on much more complex forms too (139). RNNs use sequences or lists of inputs – for each input, inside the recurrent unit, past information is combined with the new one according to a set of rules ("gates"), and an output is created, which is not only used as the output for the current input, but it also becomes part of the past information for any future inputs (139).

Due to this structure of the recurrent unit, input order is important for RNNs, as changes in the order of the sequence of inputs will result in different outputs. Simple RNN architecture, including simply adding the new input to the past information and using only a hyperbolic tangent activation function, can be very effective at modelling associations between inputs and outputs that are close to each other in the sequence. However, the simple architecture is less likely to pick up on the long-term effects of inputs further away in the sequence and thus might miss out on valuable information (139). There are many applications of RNN that rely on the long-term memory just as much or more than the short-term, however, for example in language processing, where the information needed to predict the most likely next word may be in prior sentences, or in the case of tracking a patient's status in healthcare, some important information might be among test results from weeks, months or even years prior. In these scenarios, the simple RNN would likely miss key information and thus might not be appropriate to use.

We need a different RNN structure to be able to handle these scenarios. Long short-term memory ANNs (LSTMs) have a much more complex recurrent unit that allows information to carry forward from even the first input, with gates specifying if it is ever to be replaced with new information, and whether it is to be replaced completely or not (139-141). This allows LSTM models to "remember" information both from the most recent previous inputs (the short-term memory) and anything going back to the first input, if still relevant (long-term memory).

In a simple RNN, the output of the previous recurrent unit is added to the new input with a bias, and a hyperbolic tangent activation function is applied, creating the new output. The recurrent unit in a LSTM network is a lot more complex than this. In a LSTM not only do we have the output of the unit, but also a memory component (cell state), which is modified and

passed forward in addition to the output. LSTM also makes use of three gates (Figure 3.2) (139-141):

1. The "forget gate", created by adding together the memory and output from the previous input with the new input, and a bias, and applying a sigmoid function. The output of this gate is the same size as the memory component and consists of values between 0 and 1. This controls how much of the old memory is forgotten, as it is multiplied with the memory component elementwise.
2. The "input gate", created the same way as the previous gate. The output of this gate also contains values between 0 and 1, and controls how much of the processed input is added to the old memory. The processed input here does not refer to the new input, but the new input added together with the old memory, the previous output and some bias, followed by a hyperbolic tangent activation function.
3. The "output gate", which controls how much of the *new* memory, created by adding together the outputs of the forget gate and the input gate together, is used in the output of the LSTM unit. The output gate is very similar to the forget gate, except using the new memory instead of the old.

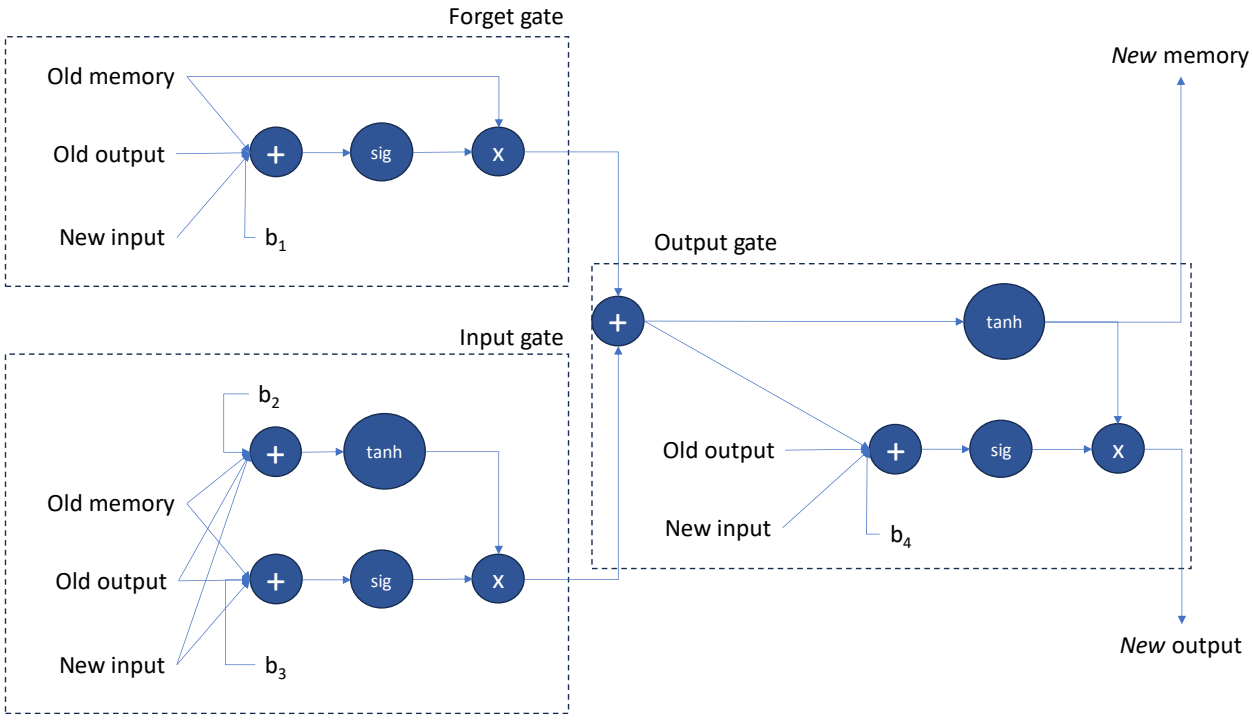

Figure 3.2: Structure of the recurrent unit of a LSTM

Parameters are tuned using backpropagation. Backpropagation is the technique used to train Artificial Neural Networks (ANNs). It uses the observed outcome and the output of the ANN to help us understand how adjusting the parameters of an ANN affects the overall cost, which represents the sum of losses across all training samples. (142)

During backpropagation, we fix the input and observed outcome by selecting a specific sample, while treating the ANN's output as a variable that we can modify by making small adjustments to the ANN itself. The ANN is applied to the input data, in each layer calculating a weighted sum of each node's inputs, adding bias, applying activation functions, and passing the information forward to the next layer until reaching the output layer. These weights and biases of the ANN are trained through the backpropagation process. As we have the observed outcome of the selected sample, we can calculate the error between it and the output of the ANN. By calculating the gradient of the cost function (also called loss function) and the error for the output layer's nodes, we can propagate this error backward, determining the error and its gradient for nodes in the previous layers. This complex process, known as backpropagation, involves intricate mathematical calculations, which will not be discussed in detail in this chapter. Once the error gradients are calculated, the weights of the ANN are adjusted in the opposite direction of the gradient, allowing the network to learn and minimize the overall error. This process is repeated for each sample, updating the weights iteratively. (142)

The most commonly used loss/cost functions can be divided into two main categories: probabilistic losses and regression losses. Probabilistic losses are used for labelled data, which may be binary, categorical, or sparse categorical. Probabilistic loss can be calculated by cross-entropy loss function (Equation 3.10) (143), for data with $n \geq 2$ classes, where t_i is a 0/1 variable indicating if the true label of the data belongs to the ith class, and p_i is output probability that the data belongs to the ith class.

$$L_{cross-entropy} = -\sum_{i=1}^{n} t_i log(p_i) \qquad \textit{Equation 3.10}$$

Regression losses are used for numerical data and include losses such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) among others. Equation 3.11,Equation 3.12, and Equation 3.13 (144), respectively, describe these

loss functions for n datapoints where y_{true} is the true value of the variable of interest, y, and y_{pred} is the predicted value.

$$L_{MSE} = \frac{1}{n}\sum(y_{true} - y_{pred})^2 \quad \textit{Equation 3.11}$$

$$L_{MAE} = \frac{1}{n}\sum|y_{true} - y_{pred}| \quad \textit{Equation 3.12}$$

$$L_{MAPE} = \frac{1}{n}\sum 100 * |(y_{true} - y_{pred})/y_{true}| \quad \textit{Equation 3.13}$$

In the case of mean absolute percentage error, a very small value is added to the true y values to avoid dividing by zero.

This is by no means a complete list of all loss functions as not only there are other, less commonly used loss functions, they can also be combined, and custom loss functions can also be created. The loss function most appropriate to the aim of the model and data available should be selected or created.

3.2.7 Model assessment

To assess the performance over time, predictions have to be made for patients for each day where data is available, and as the models predict outcome within two days of the current day, an indicator variable has to be created for each patient and day to show whether an outcome occurred within two days of the current day. We can then use the same performance measurements per day as indicated in Section 2.2.9 in Chapter 2.

3.3 Methods

3.3.1 The data

Prospectively-collected data was used from women with preeclampsia - broadly-defined according to the 2021 International Society for the Study of Hypertension in Pregnancy (ISSHP) criteria (9) - as women presented for initial facility-based assessment at centres with general policies of expectant management of preeclampsia remote from term. To maximise the sample size , data were collated from the same published model development and validation studies used in the static modelling, presented in Table 2.7 and additional data from the Dutch PETRA (Preeclampsia Trial Amsterdam) study (145, 146).

The Dutch PETRA study was a had 216 participants, with 65 (30.1%) of them having an adverse outcome within two days of admission, and 73 (33.8%) at any point. The aim of the study was to "investigate the effect of plasma volume expansion on the pulsatility indices of the fetal umbilical and middle cerebral arteries". The Dutch PETRA study included patients who had HELLP syndrome, hypertension-related fetal growth restriction with gestational hypertension, `severe` preeclampsia, or eclampsia, and were between 24 and 34 weeks' gestation. Patients were excluded if informed consent could not be obtained, or at the time of eligibility showed signs of fetal distress or had a diagnosis of lethal fetal congenital abnormalities.

Data from this study was initially cleaned following the same steps as presented in Section 2.3.1 to ensure compatibility with the previously used data, and combined with the other study data.

For the dynamic modelling, data was collected on first assessment, and any further date when assessments were carried out until occurrence of any components of the outcome, or final discharge after delivery.

The primary study outcome was the same composite outcome developed by Delphi consensus (114) as used for static modelling, and defined as the first occurrence of one or more of: maternal mortality or severe maternal morbidity (listed in Table 2.12 and defined in Table 2.11), within two days of any day when data was available. This resulted in an outcome variable that could change over time for each uniquely identified patient.

Data was split into a development and a validation dataset by randomly selecting 75% of the patients and extracting all of their data for development, removing any variables with over 60% missing values. Missing data imputations was handled separately for the two datasets, using the same methods for each variable in both cases, with multiple imputation of chained random forests as seen in Chapter 2, Section 2.3.2. All observations for all individuals in each dataset were used for imputation, however only observations within 14 days of admission were kept for the modelling and validation steps as less than 10% of patients had observations beyond this point.

Models were fitted on the imputed development data using the two-stage method, Bayesian joint modelling, LSTM artificial neural networks and binary mixed random forests. Variables

were considered for modelling only if assessed prior to the occurrence of any component of the combined adverse maternal outcome. As in the previous chapter, variables detailed the woman's health system, and her demographics, past and current medical and obstetric history, and relevant symptoms, signs, and laboratory tests. A full list of predictor variables can be seen in Table 2.8.

3.3.2 Two-stage model

First all numerical variables were centred and scaled to normalise the dataset using the preprocess function from the caret package on all imputed development datasets, as many variables had a large difference in scale, which can cause issues with convergence of linear mixed effects models (147).

Of the 18 variables found to be the most important predictors of the outcome in the PIERS-ML model, only platelet count, serum creatinine, gestational age at eligibility, weight and height at eligibility, systolic and diastolic blood pressure, oxygen saturation, and national per capita GDP and maternal mortality rate were still present in the dynamic dataset after removing variables with over 60% missing values over time. Of these variables, platelet count, serum creatinine, oxygen saturation and the two blood pressure variables were longitudinal, while the rest did not change over time. Hence, models in the first stage were fitted for these five variables, using an intercept, the number of days since first admission, the static variables, and additional variables of Ethnicity and an indicator for whether the pregnancy was a singleton or multiple pregnancy. These last variables were added as while they were not found to be among the most important predictors of the outcome, they could be predictive of some of the longitudinal values, as for example there is a proven connection between ethnicity and blood pressure (148), and women with a multiple pregnancy are also known to be at a higher risk of hypertension (149). Not all variables were used in each equation, as some needed to be removed due to causing issues with convergence.

Each model in the first stage included a random intercept, to allow for natural variation of baseline values between individuals that may not be accounted for by the predictor variables, and a random effect for time, which allowed the longitudinal variables to have different rates and magnitudes of change over time between individuals.

Using the lmer function with the option REML=TRUE, the following models were fitted on all 5 datasets:

- SpO2 ~ 1 + days + Ethnicity + heightcm + MaternalMortality + GDP + (1 + days | ID)
- Plat ~ 1 + days + Ethnicity + fetusNum + heightcm + ad_wgtkg + GAonelig + MaternalMortality + GDP + (1 + days | ID)
- creatinine ~ 1 + days + Ethnicity + fetusNum + ad_wgtkg + GAonelig + MaternalMortality + GDP + (1 + days | ID)
- sBP ~ 1 + days + Ethnicity + fetusNum + heightcm + ad_wgtkg + GAonelig + MaternalMortality + GDP + (1 + days | ID)
- dBP ~ 1 + days + Ethnicity + fetusNum + ad_wgtkg + GAonelig + MaternalMortality + (1 + days | ID)

where heightcm is the height of the patient on admission in units of cm, MaternalMortality is the national maternal mortality ratio of the country the patient was treated in, at the year of admission, plat is the patient's platelet count, sBP is the patient's systolic blood pressure, dBP is the patient's diastolic blood pressure and (1 + days | ID) is the random intercept and random effect of time (in days) per patient.

The random effects (int – intercept and sl – slope/random effect of time) were extracted from each equation for each patient and each imputed dataset and added to the corresponding dataset and patient. A logistic regression model was created using ethnicity, the indicators for singleton/multiple pregnancy, chronic hypertension, pre-gestational diabetes, right upper quadrant or epigastric pain, headache or visual disturbances and whether the patient already delivered, national maternal mortality rate, national per capita GDP and the extracted intercepts and slopes as the predictor variables.

3.3.3 Bayesian joint model

3.3.3.1 Binary outcome

Following the example in Dandis et al (134), the uninformed prior distributions were: normal distribution with mean 0 and variance 100 for the parameters of the mixed effects and logistic regression models, inverse gamma distribution with both shape and scale of 0.01 for the residual error variance, and multivariate Wishart with scale matrix of a 2x2 identity matrix and 3 degrees of freedom was used for the covariance matrix of the random effects.

However, this model could only predict outcome at any point in time and not outcome within a two day rolling period and thus was excluded from further analysis.

3.3.3.2 Time-to-event outcome

The data for the time-to-event Bayesian joint model (JMB) was split into two parts: the status data and the longitudinal data. The status data recorded if the patient was outcome free (0) or experienced an element of the composite outcome (1) on each day up to the occurrence of the first outcome or final discharge, while the longitudinal data contained the predictor variables in the same format as previous models used. A Cox regression model was fitted on the status data using ethnicity, gestational age on first admission, and maternal mortality rate and national per capita GDP of the country where the patient was treated as predictors. On the longitudinal dataset, 5 mixed effects linear regression models were fitted to predict the daily levels of blood oxygen, platelet count, serum creatinine, and systolic and diastolic blood pressures using the lme function using (1 + days) | ID as the random effect, and the same formulas for the fixed effects as the two-stage model:

- SpO2 ~ 1 + days + Ethnicity + heightcm + MaternalMortality + GDP
- Plat ~ 1 + days + Ethnicity + fetusNum + heightcm + ad_wgtkg + GAonelig + MaternalMortality
- creatinine ~ 1 + days + Ethnicity + fetusNum + ad_wgtkg + GAonelig + MaternalMortality + GDP
- sBP ~ 1 + days + Ethnicity + fetusNum + heightcm + ad_wgtkg + GAonelig + MaternalMortality
- dBP ~ 1 + days + Ethnicity + fetusNum + ad_wgtkg + GAonelig + MaternalMortality

where heightcm is the height of the patient on admission in units of cm, MaternalMortality is the national maternal mortality ratio of the country the patient was treated in, at the year of admission, plat is the patient's platelet count, and sBP and dBP are is the patient's and diastolic blood pressures, respectively.

The joint model was fitted using the jm function from the JMBayes2 R package using the models specified above, 3 chains, 2000 burn-in iterations, 12000 iterations per chain, and thinning the chains to n_thin=5.

3.3.4 Recurrent ANN (long-short term memory)

Data was pre-processed and its format changed from data frame to a three-dimensional array. For each patient, data up to 14 days since first admissions, and the 5 longitudinal predictor variables identified in the PIERS-ML model and present in the dynamic dataset (SpO2, platelet count, serum creatinine, systolic and diastolic blood pressure) were selected. For each patient, data was converted to a 15 by 5 matrix, where each row corresponded to number of days since first admission (0 to 14), and the columns were the five predictor variables. This resulted in a matrix per patient. From these matrices, matrices for each patient and day (referring to days since first admission) combination were created next by taking the full matrix, and replacing the values in rows beyond the day in the patient-day combination with zeros. For example, the matrix for a specific patient 5 days after first admission would have values in the first 6 rows (corresponding to 0, 1, 2, 3, 4 and 5 days after first admission), while rows 7 to 15 would have only zeros. Not all patients had data for 2 weeks after first admission, as patients were often delivered or had an outcome within the first week after first admission. If a patient had data for less than 14 days after first admission, then the patient-day combination matrices were only created up to the last day the patient had data available, however the matrix dimensions were still 15 by 5. Finally, these matrices were combined together into a three-dimensional array, where the first dimension was the total number of patient-day combinations, and the other two dimensions were 15 and 5. This was the data used for model development (in batches) and validation.

The model was created in R using Keras (150, 151) and Tensorflow (152, 153), using the `keras_model_sequential` R function. Model input was a 3 dimensional array of size [batch size, 15, 5]. The model had three layers: a LSTM layer of N_1 nodes, a dense layer of N_2 nodes and α activation function, and a final dense layer of 1 node and a sigmoid activation function. Models were compiled with adam optimizer algorithm – a stochastic gradient descent method used for optimising the training of machine learning models (154), binary cross-entropy loss and the precision value at 0.8 recall was used as the performance metric. Batches of 400 datapoints were sampled from the development dataset at each epoch for training, which is the period of the training process during which the entire dataset passes through the model, and models were trained for 100 epochs each. This essentially means that each model have "seen" the entire dataset 100 times, however only seeing 400 datapoints at

any given time by dividing the data into batches, cycling through the batches one by one, then splitting the data into new batches and repeating the process 100 times. Using batches of 400 data points rather than each data point individually allows us to change the parameters of our models in smaller increments. LSTM models were fitted on each batch with a development:validation split of 80:20, meaning that 320 data points were used to update model parameters and calculate a development performance, and 80 data points were used to calculate a validation performance of each batch in each epoch. In total, 90 models were created from combinations of $N_1 \in \{5,10,15\}$, $N_2 \in \{1:15\}$, and $\alpha \in \{rectified\ linear\ unit, sigmoid\}$. The performance measurements on the development portion and validation portion of the sampled batch at each epoch was saved for each resulting models. The model with the highest precision value at 0.8 recall on validation at the final epoch was selected as the final model.

The final model was tested on the separate validation data, which was held out from the model training process.

3.3.5 Binary mixed random forest

The binary mixed random forest (BiMM) model was developed using the bimm R package, which is still under development but available from github at https://github.com/bcjaeger/bimm.

Similarly to the data processing for two-stage modelling, first all numerical variables were centered and scaled to normalise the dataset using the preprocess function from the caret package on all imputed development datasets.

The model was specified with all available predictor variables as fixed effects, and a random intercept and random effect for days since admission, predicting outcome in 2 days from current day, and running a standard maximum of 10 iterations.

3.3.6 Model assessment

For each model, predicted values and outcomes were grouped by days since first admission. For each model on each day, area under the ROC curve, area under the precision-recall curve, and calibration measurements were calculated. Using the thresholds determined by the PIERS-ML model on day of admission, patients were also classified into risk strata on each day, and likelihood ratios were calculated for the risk groups per day for each model. Mean

predicted probabilities for patients with and without an outcome in 2 days per day was also plotted.

3.4 Results

3.4.1 The data

Table 3.1 shows that after the addition of the Dutch PETRA data, 610 women (6.7%) had an adverse maternal outcome within two days of admission, 853 (9.4%) within seven days, and 1138 (12.6%) at any time prior to primary discharge. In the new addition to the data, oxygen saturation under 90% was the most common outcome, with 16 women experiencing this outcome. There were three additional eclamptic seizures, two additional women required positive inotropic support, one woman had infusion of a third injectable antihypertensive, three women had at least 50% fractional inspired oxygen for at least one hour, two required intubation, seven experienced pulmonary oedema, four required blood transfusion while an additional five women had platelet count less than $50x10^9$ per L without transfusion. In addition, five women had a placental abruption.

While the aim was to model the occurrence of an adverse maternal outcome in a two day rolling window for as long as possible, modelling was restricted to the first 14 days after admission, as over 90% of women experienced an outcome or reached primary discharge by this point (Figure 3.3).

Table 3.2 describes the cohort characteristics on day of admission based on the imputed data. Variables with at least 60% of observations missing were excluded from imputation and analysis, which were most of laboratory tests. Women who experienced an outcome were younger, lived in countries with higher per capita GDP, they were more likely to have more than one fetus and were admitted earlier in gestation. They were less likely to have chronic hypertension or gestational diabetes, but more likely to experience symptoms of headache, visual disturbances, chest pain or dyspnoea. They generally weight less, had higher blood pressure (systolic and diastolic), and were more likely to have oxygen saturation under 93%. They had lower platelet counts and more serum creatinine.

In terms of interventions and pregnancy outcomes, women who experienced at least one component of the composite maternal outcome received more interventions, delivered at a

lower gestational age, their babies had a lower birthweight and were more likely to be stillborn or die within 28 days of birth.

Table 3.1: Outcomes within 2 days and 7 days from admission, and at any time

Outcome	**Within two days of admission (N=610)**	**Within seven days of admission (N=853)**	**At any time (N=1138)**
Maternal death	0	1	2
Central nervous system			
Eclamptic seizure(s)	50 (+1)	65 (+2)	78 (+3)
Glasgow coma score less than 13	14	15	20
Stroke or reversible ischaemic neurological deficit	4	5	6
Transient ischaemic attack	0	1	1
Cortical blindness	3	5	6
Posterior reversible encephalopathy	4	5	5
Cardiorespiratory			
Positive inotropic support required	4 (+1)	6 (+2)	9 (+2)
Infusion of a third injectable antihypertensive	17	29 (+1)	34 (+1)
Myocardial ischaemia or infarction	4	5	6

Outcome	Within two days of admission (N=610)	Within seven days of admission (N=853)	At any time (N=1138)
Oxygen saturation less than 90%	55 (+3)	90 (+13)	119 (+16)
At least 50% fractional inspired oxygen for at least one hour	230 (+1)	53 (+3)	75 (+3)
Intubation other than for Caesarean birth	24 (+1)	36 (+2)	49 (+2)
Pulmonary oedema	56 (+1)	85 (+3)	103 (+7)
Haematological			
Blood transfusion	244 (+2)	351 (+4)	464 (+4)
Platelet count less than $50x10^9$ per L, without transfusion	88 (+5)	103 (+5)	111 (+5)
Hepatic			
Dysfunction	23	30	44
Haematoma or rupture	0	0	0
Renal			
Acute renal insufficiency in women without chronic kidney disease	3	5	9

Outcome	Within two days of admission (N=610)	Within seven days of admission (N=853)	At any time (N=1138)
Acute renal failure in women with chronic kidney disease	36	45	52
Dialysis	2	7	11
Other			
Placental abruption	76 (+1)	101 (+3)	134 (+5)
Severe ascites	30	50	65
Bell's Palsy	3	3	6

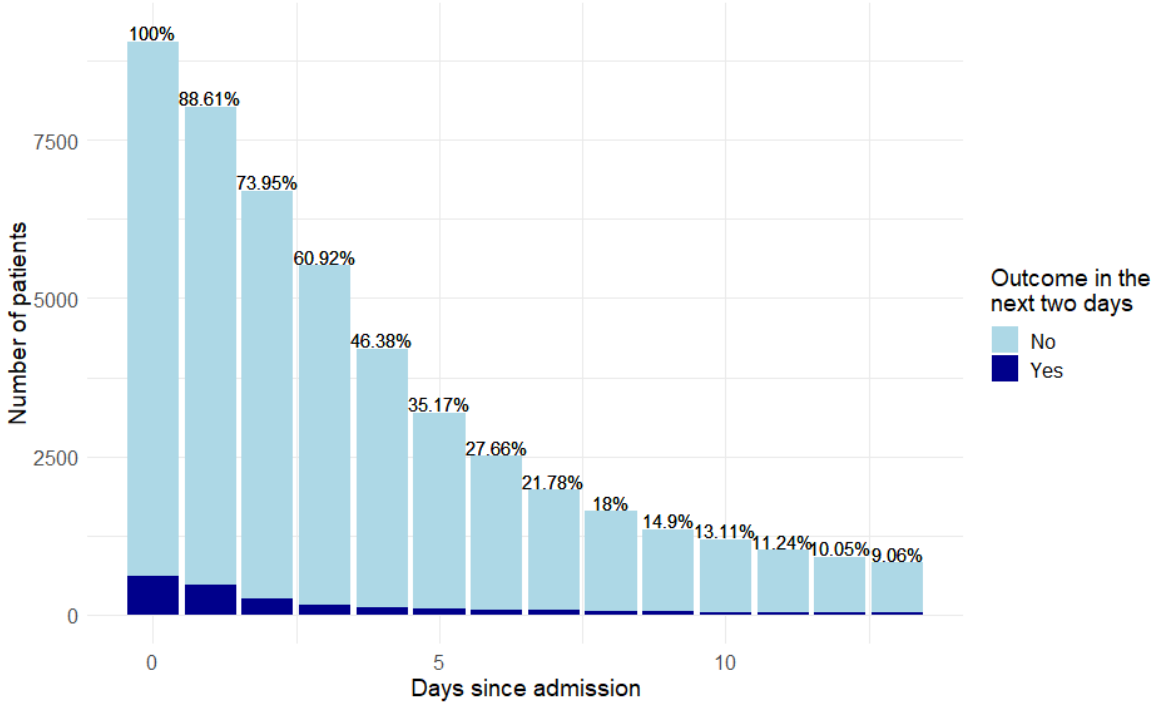

Figure 3.3: Number of patients in the data over time

Table 3.2: Proportion of missing values from all observations of the entire cohort of predictor variables and pregnancy outcomes and intervention, and cohort characteristics on day of admission after imputation for women with and without an adverse outcome at any time

Variable	Proportion missing in the total cohort (N=9059)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
Health system				
National *per capita* gross domestic product (USD)	21 (0.03%)	40385.87 [7501.47-46425.04]	43585.51 [28200.66-50114.18]	<0.001
National maternal mortality ratio (maternal deaths per 100,000 live births)	2565 (3.97%)	11 [8-161]	11 [11-15]	0.087
Demographics				
Race	7204 (11.14%)	-	-	0.214
White	-	386 (33.92%)	2665 (33.64%)	-
Asian	-	308 (27.07%)	2207 (27.86%)	-
Black	-	338 (29.7%)	2180 (27.52%)	-
Other	-	106 (9.31%)	869 (10.97%)	-
Maternal age at expected date of delivery (years)	120 (0.19%)	30 [26-35]	31 [27-36]	<0.001
Nulliparous	2590 (4.01%)	662 (58.17%)	4553 (57.48%)	0.682
Multiple pregnancy	43 (0.07%)	112 (9.84%)	527 (6.65%)	<0.001
Gestational age at eligibility (weeks)	89 (0.14%)	33 [29.96-36.99]	35.92 [31.94-38.29]	<0.001
Past and current medical and obstetrical history – no. (%)				
Cigarette smoking	3947 (6.11%)	140 (12.3%)	973 (12.28%)	1.000
Chronic hypertension	996 (1.54%)	149 (13.09%)	1347 (17.01%)	0.001
Pre-gestational renal disease	1050 (1.62%)	63 (5.54%)	524 (6.62%)	0.187
Pre-gestational diabetes	887 (1.37%)	61 (5.36%)	410 (5.18%)	0.849
Gestational diabetes	3203 (4.95%)	75 (6.59%)	974 (12.3%)	<0.001
Symptoms on day of first assessment				
Nausea or vomiting	16563 (25.62%)	175 (15.38%)	1388 (17.52%)	0.08
Headache or visual disturbance	16214 (25.08%)	408 (35.85%)	2484 (31.36%)	0.003
Right upper quadrant or epigastric pain	14558 (22.52%)	264 (23.2%)	1675 (21.15%)	0.124
Chest pain or dyspnoea	29552 (45.72%)	86 (7.56%)	192 (2.42%)	<0.001

Variable	Proportion missing in the total cohort (N=9059)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
Signs on day of first assessment				
Height(cm)	6245 (9.66%)	162 [157.48-167]	162.56 [157.48-167.64]	0.398
Weight (kg)	2997 (4.64%)	78 [68-89.07]	81.5 [70.31-94]	<0.001
Systolic blood pressure (mm Hg)	7499 (11.6%)	155 [145-169]	150 [140-163]	<0.001
Diastolic blood pressure (mm Hg)	7503 (11.61%)	100 [90-107]	96 [90-102]	<0.001
Oxygen saturation less than 93%	5370 (8.31%)	234 (20.56%)	968 (12.22%)	<0.001
Dipstick proteinuria (number of '+')	39941 (61.79%)*	162 [157.48-167]	162.56 [157.48-167.64]	0.398
Laboratory tests – worst values on day of first assessment				
Haematocrit (%)	**41415 (64.07%)**			
Total leucocyte count (x 10^9 per L)	**41911 (64.84%)**			
Platelet count (x 10^9 per L)	32910 (50.91%)	197 [150-243]	211.8 [171-252]	<0.001
Mean platelet volume (fL)	**45645 (70.61%)**			
Fibrinogen (g/L)	**44458 (68.78%)**			
Activated partial thromboplastin time (sec)	**44463 (68.78%)**			
Serum creatinine (µmol/L)	35793 (55.37%)	63 [52-76]	59.29 [50-69]	<0.001
Uric acid (mmol/)L	**39560 (61.2%)**			
Aspartate transaminase (U/L	**38577 (59.68%)**			
Alanine transaminase (U/L)	36762 (56.87%)†			
Albumin (g/L)	**45373 (70.19%)**			
LDH	**44991 (69.6%)**			
International Normalised Ratio	**47549 (73.56%)**			
Total Bilirubin	**55063 (85.18%)**			
Urinary protein:creatinine ratio	**61564 (95.24%)**			
Interventions per pregnancy				
Corticosteroid received	1643 (18.14%)	377 (33.13%)	1695 (21.4%)	<0.001

Variable	Proportion missing in the total cohort (N=9059)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
Antihypertensive medication received	1629 (17.98%)	714 (62.74%)	4097 (51.72%)	<0.001
Magnesium sulphate received	1655 (18.27%)	556 (48.86%)	2257 (28.49%)	<0.001
Pregnancy outcomes per fetus				
Gestational age at delivery (weeks)	1405 (14.63%)	35.57 [31.86-38]	37.29 [34.86-39]	<0.001
Birthweight (g)	1396 (14.54%)	2200 [1410-2830]	2719.5 [2000-3280]	<0.001
Intrauterine fetal death, ≥20⁺⁰ weeks or ≥500 g[37]	1361 (14.17%)	61 (4.96%)	144 (1.72%)	<0.001
Neonatal death, within 28 days	1340 (13.95%)	30 (2.44%)	87 (1.04%)	<0.001

** Dipstick proteinuria had <60% missingness on initial assessment prior to the addition of further datasets, and hence was included in the imputed variables. † Alanine transaminase had ≥60% missingness on initial assessment prior to the addition of further datasets, and hence was not included in the imputed variables.*

3.4.2 Mean predicted probability per day per outcome group

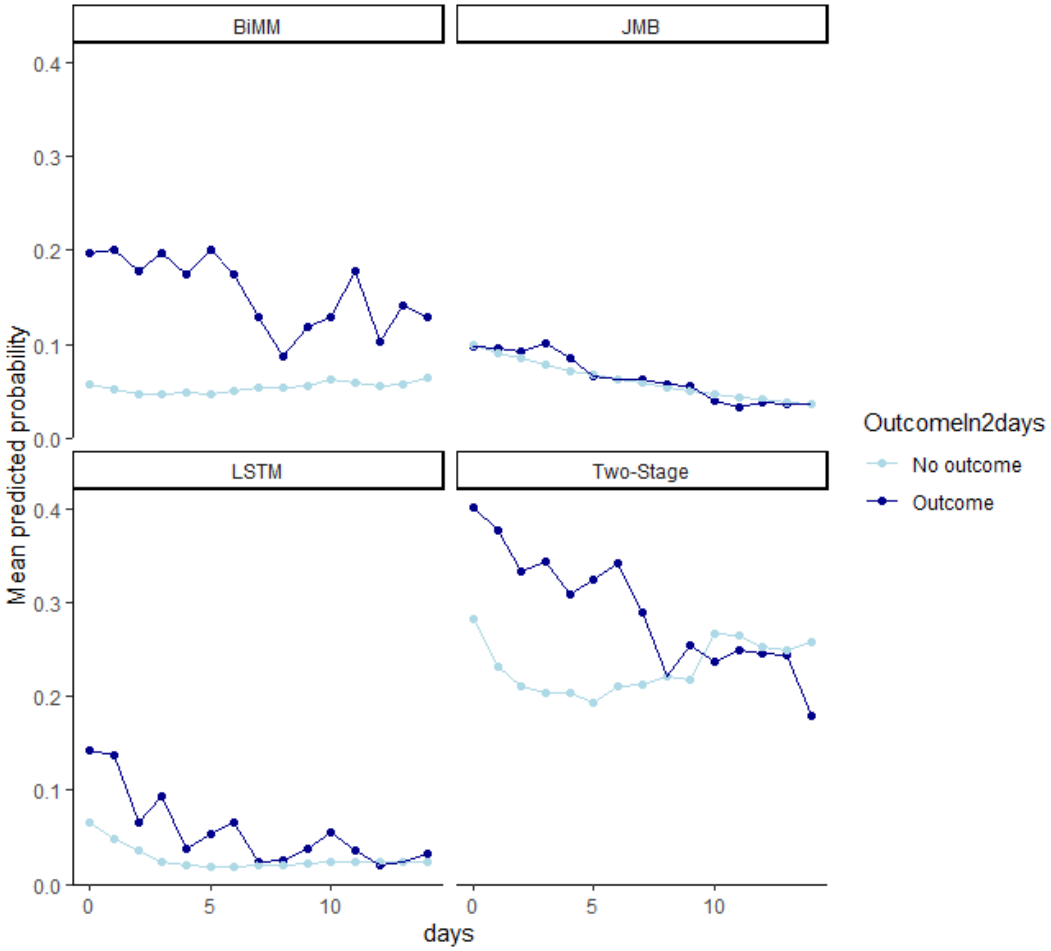

Figure 3.4: Mean predicted probability per outcome group, model and day

Figure 3.4 shows the mean predicted probability per day for the outcome group and the no outcome group for each dynamic model. Of the four models, the outcome group generally had higher mean predicted probabilities for all except the JMB model. The BiMM showed the best discrimination of the outcome and no outcome groups of the four models, the patients who had an outcome in the next two days consistently had a higher mean predicted probability than those who did not, for the whole two weeks. We can observe a dip in the mean predicted probability of the outcome group between days 6-8, however the no outcome group's mean predicted probability stayed at a lower value.

The JMB model's performance was on the other end of the scale, performing very poorly. On all days except day 3, the model had the same or very similar mean predicted probabilities for patients with and without an outcome in two days. Moreover, the mean predicted probabilities for both groups on all days were at or below 10%. The LSTM model has similarly

low levels of mean predicted probabilities, however the outcome group had generally higher values for the first 11 days than the no outcome group, with a pattern of "spikes" in the mean predicted probabilities of the outcome group. A similar pattern is also observable in the BiMM and two-stage models, but less pronounced. After day 11, the model performs similarly poorly to the JMB model.

Aside from the difference between the outcome and the no outcome groups, there was also a visible difference in the scale of predicted probabilities between the models. This was likely due to a collection of reasons, such as regression-based methods assuming linear relationship between the log odds of the outcome and predictor variables and calculating probabilities via an equation, while random forests and neural networks make no assumption of linearity, and random forests calculate probability using a voting system. The models may not all be equally well calibrated (see Section 3.4.5), and different methods may not be as sensitive to a low outcome rate as others.

The TS (Two-Stage) model had the highest mean predicted probability for both groups. The model performed well in the first week, with the outcome group having a higher mean predicted probability, however the difference quickly dropped after day 7, and the two groups had either similar mean predicted probabilities, or the no outcome group was predicted higher.

3.4.3 Area under the ROC

The ROC curve plots the sensitivity against the corresponding 1-specificity from each possible predicted probability threshold that could be used to divide the predictions into a "Predicted Outcome" and "Predicted No Outcome" group. We can calculate the area under this curve as a measure of discrimination ability of a model.

The AUROC values of each model over time are shown in Figure 3.5. As suggested by the mean predicted probabilities, the BiMM model performed better than other models on most days, and the JMB model performed worst. The AUROC of the BiMM model was generally between 70% and 80%, which corresponds to fair performance, and even higher than 80% on days 5, 6 and 14, showing good performance. The model had no major drops in AUROC values, although confidence intervals became large from day 10 onwards. Importantly, the performance of the model was never worse than random allocation. The JMB model on the

other hand, achieved its best performance on day 3 at not much higher than 60%, while most days it had an AUROC value of around 50%, making it no better than random allocation. The model had a large drop in performance on days 10 and 11, when its AUROC values were significantly worse than 50%, the representation of random allocation.

The performance of the LSTM model was similar to the BiMM model, although somewhat worse in later time periods. The model had AUROC values between 70% and 80%, corresponding to fair performance, and its confidence intervals remained smaller than other models until after day 11. The model had a large drop in AUROC values starting at 11 days, however it only dropped to the level of random allocation on day 12. The TS model consistently had AUROC values around 60% for the first 7 days, which, while not corresponding to a good performance, was the most consistent of the models. After this period of stable performance, however, the model started to perform worse over time, and by day 10 it had an AUROC value lower than random allocation.

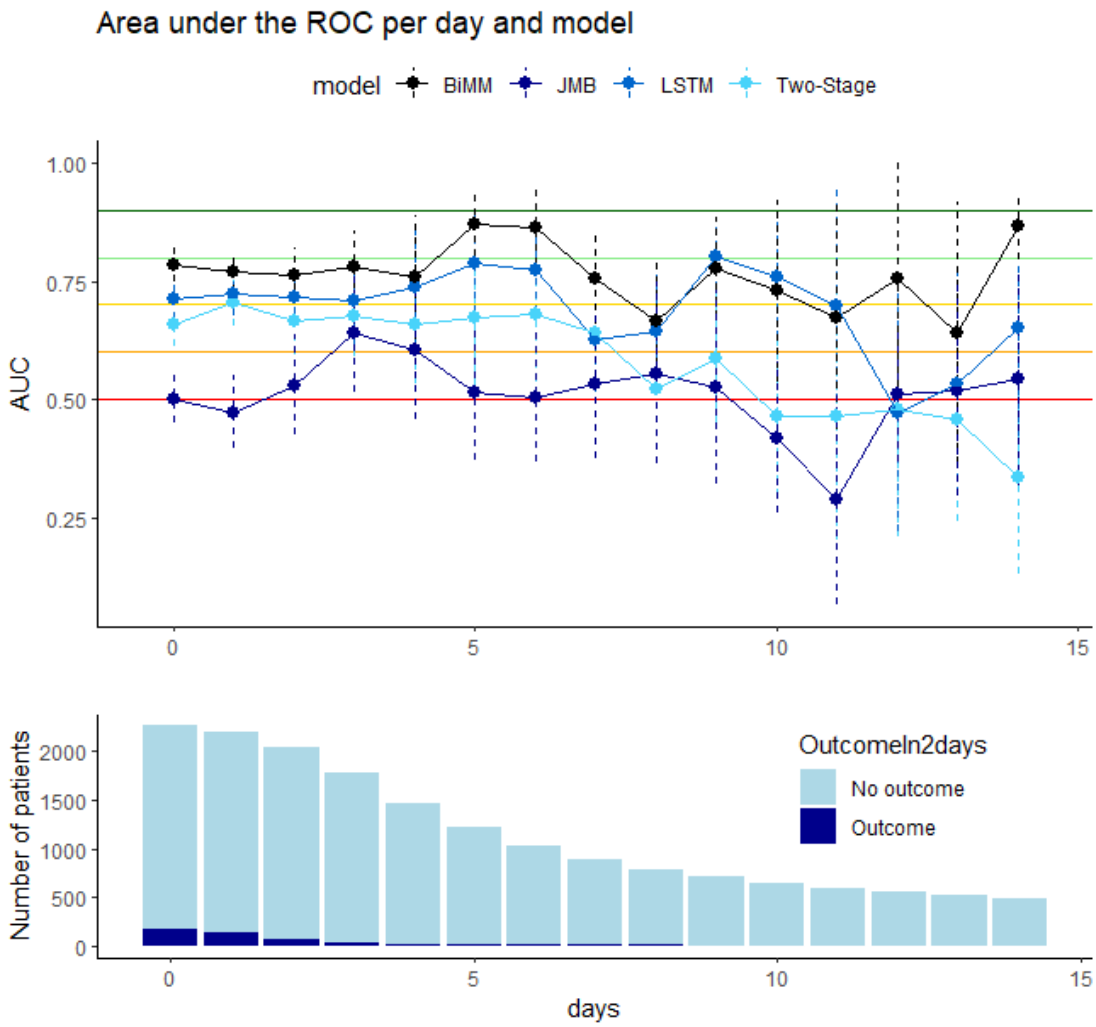

Figure 3.5: Area under the ROC curve with 95% confidence interval per model over time

3.4.4 Area under the precision-recall curve

A precision-recall curve plots the precision (number of true positives/number of all positive predictions) against recall/sensitivity (number of true positives/number of all positive observations) for all predicted probability thresholds threshold that could be used to divide the predictions into a "Predicted Outcome" and "Predicted No Outcome" group. We can calculate the area under this curve as a measure of discrimination ability of a model, especially when the rate of outcomes is low.

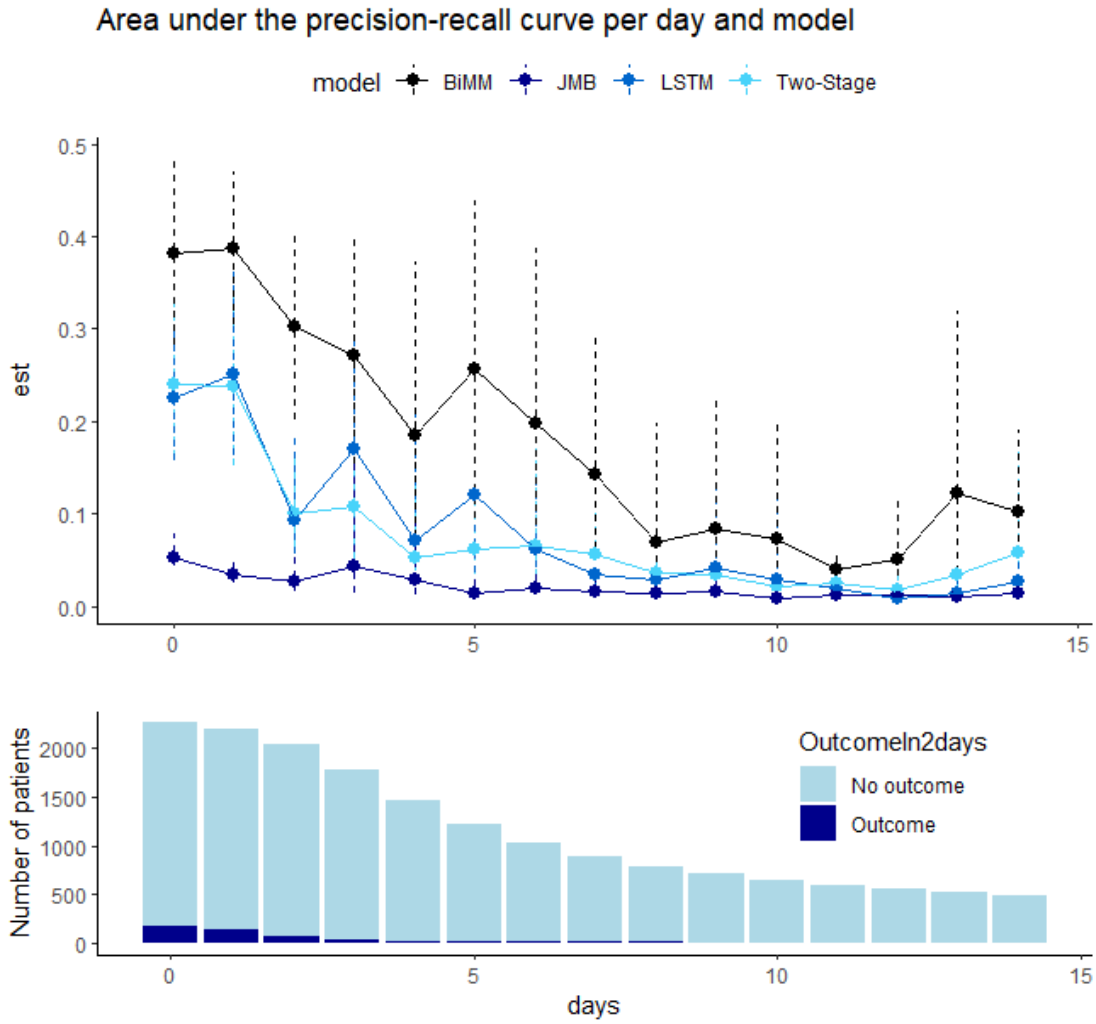

Figure 3.6: Area under the precision-recall curve with 95% confidence interval per model over time

Figure 3.6 shows the AUPRC values of the four models over time. The BiMM model again had the highest values out of the models, this time performing better than the other models on all days without exception. All models, except for the JMB model performed best on day 0 and the performances gradually declined over time, with the AUPRCs getting close to 0. The exception, the JMB model, unfortunately had an AUPRC near 0 from day 0 and did not improve on any day. This suggests that the model cannot differentiate between patients who did have an outcome and those who did not. The BiMM model had an AUPRC near 40% on day 0, while the LSTM and TS models both had an AUPRC around 25%, and the three models' performance dropped over time at roughly the same rate. The LSTM and TS models had very similar AUPRC values on most days, with the LSTM model performing slightly better on a few occasions.

3.4.5 Calibration

Model calibration values for each model over time are displayed in Table 3.3. Calibration refers to the accuracy of the predicted risk, rather than the predicted outcome, compared to the observed outcome. The BiMM model had a cox calibration intercept fairly close to 0 for the first 7 days (although some days were better than others), after which it began to decrease, showing a consistent overall bias for the first week after admission. Its cox calibration slope close to 1 on most days. The Spiegelhalter z-value for this model had a p-value less than 0.05 for about half of the days (days 3, 4, 5, 6, 10, 11, 12), meaning that the model was significantly improperly calibrated the majority of the time after 2 days since admission.

The time-to-event JMB model had poor calibration-in-the-large: the cox calibration intercept shows that the predicted values vary too much compared to the observed data. The calibration slope also flips from negative to positive and back, and many values of the slope are close to zero, showing no pattern in the predicted probabilities compared to the observed risk. The Spiegelhalter z-value had a p-value of 0 for all days, meaning that the model was significantly improperly calibrated at all times after admission.

The LSTM model had very inconsistent calibration over time. Its cox calibration intercept was close to 0 on days 0 and 1, but showed consistent bias in either the positive or negative direction on the remaining days, while the slope covered the entire range of “close to 1 and similarly varied as the observed data”, “more varied than the observed data” and “less varied than the observed data”. The p-value of its Spiegelhalter z-score showed significantly inappropriate calibration on days 1, 10, 11, 12.

The TS model had similarly poor calibration as the JMB model. The cox calibration intercept started at -2 and only got lower over time, showing a consistent overall bias, while the cox calibration slope started close to 1, meaning an acceptable spread of values for the first 2 days, but values became lower and lower after 2 days, showing that the values vary too much compared to the observed data. The Spiegelhalter z-value had a p-value of 0 for all days, meaning that the model was significantly improperly calibrated at all times after admission.

Table 3.3: Model calibration measures per model over time

Model	Day	Cox Calibration Intercept	Cox Calibration Slope	Brier Score	Spiegelhalter z statistic	Spiegelhalter p-value
BiMM	0	0.24	1.05	0.06	0.78	0.44
	1	0.07	0.96	0.05	1.02	0.31
	2	-0.22	0.96	0.04	-1.57	0.12
	3	-0.32	1.08	0.03	-3.61	0
	4	-0.8	1.03	0.02	-4.43	0
	5	0.3	1.47	0.02	-3.59	0
	6	0.22	1.4	0.03	-2.88	0
	7	-0.4	0.97	0.04	-1.41	0.16
	8	-1.33	0.65	0.04	-1.49	0.14
	9	-1.02	0.85	0.03	-1.8	0.07
	10	-1.33	0.81	0.03	-2.55	0.01
	11	-1.48	0.83	0.02	-2.62	0.01
	12	-1.1	0.99	0.02	-2.43	0.02
	13	-1.26	0.72	0.03	-1.67	0.1
	14	-0.03	1.34	0.03	-1.89	0.06
JMB	0	-3.07	-0.08	0.05	-6.73	0
	1	-3.35	0.03	0.04	-9	0
	2	-2.95	0.36	0.03	-9.54	0
	3	-0.79	1.39	0.02	-8.96	0
	4	-1.51	1.04	0.02	-7.66	0
	5	-4.31	-0.06	0.02	-6.58	0
	6	-3.92	0.01	0.02	-5.32	0
	7	-3.17	0.35	0.02	-5.16	0
	8	-2.7	0.57	0.01	-4.87	0
	9	-3.07	0.4	0.02	-4.27	0
	10	-7.43	-0.85	0.01	-4.48	0
	11	-11.56	-2.03	0.01	-4.25	0
	12	-5.2	-0.27	0.01	-3.22	0
	13	-5.18	-0.18	0.01	-3.27	0
	14	-3.82	0.15	0.01	-2.64	0.01
LSTM	0	-0.22	0.87	0.07	1.59	0.11
	1	0.18	0.99	0.05	2.02	0.04
	2	-0.64	0.8	0.03	0.21	0.83
	3	0.4	1.13	0.02	-0.37	0.71
	4	0.81	1.27	0.02	-0.81	0.42
	5	2.71	1.77	0.02	-0.43	0.67

Model	Day	Cox Calibration Intercept	Cox Calibration Slope	Brier Score	Spiegelhalter z statistic	Spiegelhalter p-value
	6	1.77	1.5	0.02	-0.4	0.69
	7	-1.32	0.62	0.02	0.6	0.55
	8	-1.25	0.67	0.02	-0.09	0.93
	9	0.98	1.39	0.02	-1.21	0.23
	10	0.85	1.5	0.01	-2.27	0.02
	11	-0.7	1.05	0.01	-2.18	0.03
	12	-5.6	-0.22	0.01	-2.24	0.03
	13	-3.6	0.18	0.01	-1.51	0.13
	14	-1.54	0.71	0.01	-1.33	0.18
TS	0	-1.85	0.82	0.12	-15.47	0
	1	-1.74	0.8	0.1	-13.42	0
	2	-2.42	0.52	0.09	-11.53	0
	3	-2.51	0.69	0.08	-11.86	0
	4	-3.1	0.45	0.08	-10.49	0
	5	-2.91	0.42	0.08	-7	0
	6	-2.82	0.48	0.09	-6.37	0
	7	-2.72	0.35	0.09	-5.03	0
	8	-3.1	0.07	0.1	-3.99	0
	9	-3.07	0.25	0.1	-4.07	0
	10	-3.73	0.01	0.11	-6	0
	11	-3.76	-0.03	0.11	-4.91	0
	12	-3.8	0.02	0.1	-4.98	0
	13	-3.55	-0.1	0.1	-4.53	0
	14	-3.97	-0.4	0.11	-4	0

3.4.6 Risk groups

The model performances at stratifying patients in to risk groups using the PIERS-ML thresholds over time are displayed in Table 0.7 in Appendix C. The groups were expected to stratify women such that the very high risk group had a positive likelihood ratio of at least 10, the high risk group had a positive likelihood ratio of at least 5, while the very low risk group had a negative likelihood ratio of at most 0.1 and the low risk group had a negative likelihood ratio of at most 0.2. The likelihood ratios where any of these condition was met are highlighted in the table in blue. Looking at the risk classification performance of the BiMM model first, the model performed very well for the first week after admission: it achieved the

desired likelihood ratios on all days in the high and very high risk groups, and it was only on days 2 and 4 when the very low risk group did not have the desired negative likelihood ratio. Beyond the first week, the very low risk group had a negative likelihood ratio of 0.1 or less on all remaining days. The model's week point was the classification into the low risk group, which only achieved the negative likelihood ratio of 0.2 or less on five days (days 5, 6, 7, 9, and 14). Beyond day 7, the high and very high risk groups had an inconsistent performance, with the high risk group achieving the desired likelihood ratio on days 8, 10, and 13 only, and the very high risk group succeeding on day 11 only.

The JMB model had a perfect 100% outcome rate in the very high risk group on day 1, however this 100% comes from only a single patient, which is not a reliable measure. Similarly, the high risk group on day 3 and the low risk group on days 3, 4 and 6 show good likelihood ratios, but all of these groups had very small numbers (~1% of the patients or less). On most days, this model is only stratified patients into low or moderate risk groups, and while the low risk group met the expected negative likelihood ratio, the majority of patients were still classified as moderate risk, which suggests that these risk thresholds are not useable for this model.

Looking at classification into risk groups of the LSTM model, this model did not classify any patients into the very low risk group at any point, and the low risk group, while included patients, did not achieve the required negative likelihood ratio at any point either. This is likely a good indication that the risk group thresholds developed for the PIERS-ML model do not work well for this model either and we should pick new group thresholds, however there was not enough data available to do this. The model performed well for the other groups on days 0 , 1 and 3, the high risk group had a positive likelihood ratio above 5 on two more days (days 2 and 5), and the very high risk group had a positive likelihood ratio above 10 on day 6 also. This model also had inconsistent performance.

Finally, the TS model could not rule in an outcome within two days at any time in the first 2 weeks after admission as the very high risk group always had a positive likelihood ratio less than 10, and the high risk group always had a positive likelihood ratio less than 5. On four days (days 2, 9, 10, and 12) the model had a negative likelihood ratio of 0.2 or less in the low risk group, and on all days except for day 1, 10 and 11, the model had a negative likelihood ratio of 0 in the very low risk group, showing that it could successfully rule out the outcome for a small proportion of the patients. However, the very low risk group had less than 1% of

patients on each day, and thus the sample size for the likelihood ratio calculation was very small in this group.

3.5 Discussion

3.5.1 Summary of findings

The analysis in this chapter included data from 9059 women from 11 countries presenting for first assessment of preeclampsia and up to two weeks after, and used multiple dynamic modelling methods to develop and internally validate models for maternal risk stratification, applicable for LMICs and HICs. I used a statistical modelling method – two-stage modelling – a Bayesian modelling method – Bayesian joint modelling – and two machine learning enabled modelling methods – binary mixed model random forest and long short-term memory artificial neural network.

Four models, one using each modelling method, were fitted. When using the models to classify patients into an outcome and a no outcome group, only the BiMM model had good discrimination with an AUC value close to 0.8, while one model using Bayesian joint modelling performed no better than a random classifier. However, the area under the precision-recall curve was also low for all models. Based on these results, the two-group outcome/no outcome prediction should not be used to inform clinical care, and the five risk strata should be used instead, as not only is it more accurate, but it also provides better interpretability in the care setting.

The pattern of random forest-based models performing best in risk stratification that we saw in static modelling appears to hold for this analysis as well, as the BiMM model achieved the desired likelihood ratios in the classification for the most risk groups out of the models. However, even the best performing model's performance dropped after the first week, which could be attributed to over 75% of patients either delivering or experiencing an outcome by that point. The number of outcomes within 2 days is also drastically lower in the second week after admission, with less than 10 outcomes within 2 days per day 9 days after admission. While this lack of data is most likely the main cause of the drop in performance of all models, it is also possible that patients who remain pregnant without any outcomes for over a week after first admission with preeclampsia are fundamentally different from those who suffer an outcome or deliver within the first week.

The final model selected for modelling the composite adverse outcome over time was the binary mixed model random forest using all predictor variables, further referred to as the PIERSdynamic model. For the first 11 days after admission, the model identified approximately 50% of patients for whom care should be altered. During the first 7 days after admission, approximately 45% of women were identified as being at very-low (~4-10%) or low risk (~35-40%). These women could be reassured that it is very unlikely they would suffer an adverse maternal outcome within the following two days, as the very-low risk strata observed ~0-1% outcome rate, and the low risk ~0.5-3%. However, for the approximately 5% of women identified to be at high (~3-6%) or very-high risk (~0.5-1%) at some point within the first 7 days after admission, a timely clinical response can be justified based on a substantial risk (~20-35% and ~70-100% respectively) of an adverse maternal outcome occurring within two days of the day of risk assessment. Using the PIERSdynamic model allows clinicians to expectantly manage women diagnosed with preeclampsia for up to a week after admission as the model output can be confidently used to identify women at high or very high risk of outcome and inform discussion about place of care, transfer of care, antenatal and postnatal surveillance, co-interventions, and timed birth.

3.5.2 Strengths and limitations

As mentioned for the studies in the previous chapters, one of the strengths of this study is the large sample size and diverse demography. The cohort, which is over three times the size of any previous study, was also followed until the occurrence of an outcome or final discharge after delivery, with predictor variables regularly collected as specified in study protocols.

The study is one of a kind, as to our knowledge the PIERSdynamic model is the only model currently designed for and tested to model the risk of adverse maternal outcome of preeclampsia using all available information for a patient. The model uses clinically significant and regularly collected (where possible) predictor variables covering all organ systems. A limitation of this study is the lack of variable selection, which may make this model less accessible in low resource settings. While the model is a random forest-based method, and machine learning is suitable for managing a large number of variables, and addresses concerns regarding collinearity, as mentioned previously this means that included variables may not contribute significantly to the predictions. Using feature selection to remove unimportant or less significant variables can help to balance model performance and

complexity by identifying redundant variables while retaining model performance as less-informative variables are censored. A less-complex model will be easier to use in clinical practice and may reduce the costs associated with laboratory-based variable collection. In real world application of the PIERSdynamic model, a last observation carried forward approach could be a temporary solution to manage the cost of using the model, however we suggest the testing of variable selection methods to create a less-complex model to be carried out as future work.

Another strength of this study is the testing and comparison of various modelling methods, including machine learning and Bayesian methods, to select the most suitable model for the purpose.

Missing values were common in our dataset, even more so than they were in the static data; however, where appropriate, missingness was handled by multiple imputation, with chained random forest to minimise bias. Of note, development and validation datasets were imputed separately, thereby creating independent model development and validation datasets. Some variables in addition to those excluded in the static modelling (namely aspartate transaminase, uric acid, dipstick proteinuria, haematocrit, white blood count, fibrinogen, active partial thromboplastin time, serum albumin, mean platelet volume) had to be excluded due to high levels of missingness in the dynamic data. This limitation reflects current clinical practice, as while these variables are routinely collected, collection will be carried out at different frequencies.

The predictive ability of the model was also limited by the number of patients declining over time, in particular there was a drastically smaller number of patients in the second week after admission. This limitation also reflects current clinical practice, as most patients with preeclampsia are only expectantly managed for up to two weeks, and guidelines suggest delivery within two weeks from admission. This is due to most outcomes occurring within the first week of admission, which we have also observed in our data.

A limitation in comparing the risk stratification of our models was the use of risk stratification thresholds derived for use with the PIERS-ML model, rather than for our dynamic models. Ideally, we would want to calculate risk stratification thresholds for each model and each day,

as the same predicted probability could justify different clinical decisions at different times in the pregnancy, however this was not possible due to the amount of available data over time.

Finally, the model has not been externally validated, however this is suggested to be undertaken as future work.

3.5.3 Interpretation in light of existing literature

National and international guidelines recommend using predictive models to assess the risk of adverse maternal outcomes of women with preeclampsia, as mentioned previously. In particular, the NICE guidelines suggest the repeated use of models developed to be used on admission, to predict risk over time with the latest available information. We have recently shown (Appendix B) that the performance of static models when used for serial static prediction is not reliable, and in fact very quickly deteriorates after only a couple days from admission.

The PIERSdynamic model improves on both the currently recommended model (fullPIERS) and the PIERS-ML model when used repeatedly.

The dynamic model shares many of the same improvements over the fullPIERS model as the PIERS-ML model, described in the previous chapter: the model was created using a diverse dataset of 11 countries, including LMICs; the model uses routinely collected variables that cover all organ systems and include per capita GDP and national maternal mortality ratio to auto adjust to location while the fullPIERS model has to be re-calibrated to be used in different settings.

The PIERSdynamic model also improves over both the fullPIERS and the PIERS-ML model by using multiple measurements per patient to see the change in the patient's condition over time, being able to assess their risk of an adverse maternal outcome within two days of the current day using all previous observations. There are currently no other models capable of using longitudinal variables to monitor maternal risk of adverse outcomes over time.

There are three important lines of investigation that follow from our work.

The first is that use of this method offers the potential for the accuracy of the model to improve to extend accurate prediction to 14 days after admission, and more by obtaining data from pregnancies with early pre-term preeclampsia diagnoses, which are more likely to be

expectantly managed for over a week after diagnosis. This should be feasible, as all individual-level variables in the PIERSdynamic are part of routine clinical and laboratory assessment of women with preeclampsia in well-resourced settings, and approximately 30% of cases of preeclampsia are diagnosed away from term.

Second, to explore the use of different thresholds over time to improve the stratification ability of the model. This line of investigation also relies on the collection of further data, with particular focus on women with preeclampsia who were in expectant management beyond the first week after their diagnosis.

Third, to externally validate the PIERSdynamic model in an unseen cohort, to confirm its performance and gather evidence to suggest changes to clinical guidelines over serial static use of currently recommended models.

3.6 Conclusion

In conclusion, we have created a new effective risk stratification tool for repeated international use, PIERSdynamic, which can accurately advise care for patients with preeclampsia not only on admission, but up to 7 days after, on the risk of adverse maternal outcomes of preeclampsia. The model significantly improved on the fullPIERS and PIERS-ML models when used repeatedly over time, which is suggested by the National Institute of Health and Care Excellence (NICE). While we have shown that the model work well for at least 7 days after admission, more data is needed to assess model performance longer term.

Chapter 4: Prediction on missing data

4.1 Introduction

In Chapters 2 and 3, we have explored multiple methods for creating static and dynamic models for prediction of adverse outcomes of preeclampsia. In the development of these models, imputation methods were used to create multiple complete datasets. While we were able to manage the missing data for the development, the created models all require complete data to be able to make predictions – that is, data either needs to be available or imputed when the model is used in real world application, otherwise the model is unusable for patients with missing information.

Missing data in the development phase of predictive modelling is a widely discussed topic (155), however there is far less information and guidance available on the issue of missing data at the model deployment stage (156). Many predictive modelling methods, especially commonly used regression-based methods, require complete observations with no missing values for the purpose of prediction. The currently available software for the gold standard method for handling missing data during model development, multiple imputation, does not allow for the imputation method to be trained on the development dataset and deployed on a single new observation in most cases (156). The few cases that do allow this require the development dataset to be available for the new observation to be appended to, as the parameters of the imputation are estimated from the development data each time (156, 157). Moreover, as data is often collected in the controlled environment of a study for model development, it may not be appropriate to use the same method of imputation as the development stage when the models are deployed in much less controlled real-world settings, where a range of additional factors may influence data availability (155). In these settings, missing values could be indicative of outcome even if they were not during the development stage, missing values could be indicative of the available resources, or of personal bias of the decision-maker (158, 159).

Data collected for modelling adverse maternal outcomes of preeclampsia in the studies included in this analysis falls into categories of baseline patient information, medical history, signs, symptoms, clinical assessment, and laboratory tests. Data in different categories are not all available at the same time, some are available immediately on admission while others become available over time. Generally, baseline patient information and medical history are

available immediately on admission through patient records, although access to patient records is not always available in low resource settings. Signs and symptoms can be recorded along with a clinical assessment given that the patient is conscious and a healthcare professional is available to perform the assessment. Hence variables in these groups will be available at different times depending on waiting times, and the condition of the patient. Variables in the laboratory test groups are expected to become available later than all other groups, as they require a sample to be taken and analysed. Additionally, patients with preeclampsia are treated in low-, middle- and high-income countries and are treated both as in- and outpatients in each setting, leading to limited access to certain tests in some cases and different waiting times for results.

Multiple imputation, the method used to prepare the data for modelling in the previous chapters, is a method that can minimise bias introduced from imputation very efficiently as long as there is a sufficiently large amount of data, strong relationships between variables, and the data is imputed a large number of times (at least the percentage of missingness). When dealing with large datasets with missing data, multiple imputations is easy to use and can be very effective, but this is not the case when dealing with a single observation with missing values.

Thus, there is a need for a tool that can make the best use of the currently available data, handle missing information without heavily relying on multiple imputation to be deployable without the development dataset, make a risk prediction using this subset of data and quantify the uncertainty around the prediction.

To address this need, I aim to build a multistep hierarchical classification tool using machine learning, minimise the amount of imputation used in the process, and find the best method of handling missing data.

4.2 Methodological background

4.2.1 Multistep modelling

Modelling methods covered in Chapter 3 could be referred to as single step models. A single step model can be applied to new data by just supplying the new data to the trained model and obtaining the desired output from the model in a single step. Multistep models, as the name applies, consist of multiple steps, one or more of which could be applying a trained

model to the data. We will refer to models that take the new data as the input to the first step, and produce the desired output as the output of the last step, however have steps in-between that may take different inputs and generate different outputs as multistep models.

An example of this in the literature is stacked generalization, a process in which models are "stacked" by the output of one model being passed to the next, and so on until the final output is created (160). Artificial neural networks could also be considered multistep models, as each layer could be thought of as a model or a step, the output of which is passed onto the next until the output layer is reached.

Figure 4.1 illustrates the basic structure of a single step model (A) and a multistep model (B).

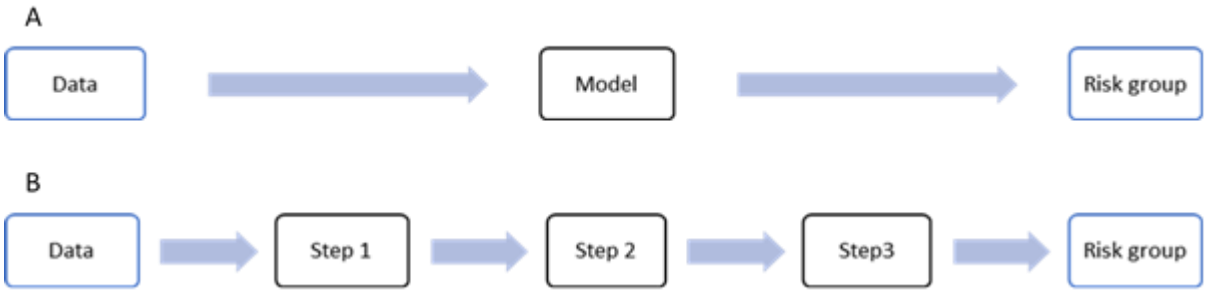

Figure 4.1 Single step model (A) application vs multistep model (B) application process

Similarly to single step modelling, one or more of the steps in a multistep model will include a predictive model, however thanks to the structure of a multistep model, imputation models and model selection processes can also be included, allowing for the model to predict on missing data and select from multiple predictive models, if required. Methodology for predictive models tested as part of any steps in our multistep model is covered in Chapter 3.2.

In our case of multiple predictive models being included in the multistep model, we can rank these models to create a hierarchy such that if the provided model selection process could equally choose multiple models in a specific scenario, the model ranked highest in the hierarchy can be chosen. This combines the concept of multistep modelling with simple model selection methods. Models can be ranked based on one of the performance measures covered in Chapter 3.2.9, a non-statistical measurement such as smallest number of variables used or fastest runtime, or they can be ranked manually.

In the case of the multistep model being provided with data including missing values, the model can contain a step to handle missing data so that it will still create a prediction. An example of this can be seen in artificial neural networks, using an imputer as step 1 of the multistep modelling process, followed by prediction models and a combiner (161). Some methods of imputation mentioned in Chapter 3.2.3, such as replacing missing values with a single predetermined value, can be used as an imputation step, while methods that require a minimum number of observations, for example multiple imputation, cannot as the data may just be a single observation. While there are some multiple imputation methods that can be trained in advance on training data and then applied even on a single new observation, such as the mixgb package in R (162), using gradient boosted trees for imputation, they are still difficult to use when applying predictive models rather than when developing them as the number of imputations needed depends on the proportion of missing data in the new observation(s), which is an unknown when creating the multistep model.

There exists, however, a method of imputation that lends itself very well to multistep modelling which is commonly used in engineering and image analysis, while very rarely used for analysis of non-image clinical data. Generative Adversarial Imputation Nets (GAINs) were developed from Generative Adversarial Networks (GANs) and are capable of taking data with

missing values of any length and predetermined width and creating an output with no missing values (163, 164).

4.2.2 GANs

Generative Adversarial Networks, or GANs, refer to machine learning models of pairs of neural networks that are trained in competition with each other to generate data indistinguishable from some sample of real data. GANs contain a Generator – an artificial neural network that takes random noise as its input and outputs generated fake data proportionate to the size of the inputs – and a Discriminator – another artificial neural network that takes a mixture of fake and real data sample as the input and outputs labels for whether the data is real or fake. These two ANN models both receive feedback based on the output of the discriminator: the discriminator is penalised for accepting fake data as real, or rejecting real data as fake, while the generator is penalised only if the generated fake data is rejected by the discriminator. The two models are trained simultaneously, meaning that at each step of the training process, the generator generates fake data, to which a sample of real data is added, and labels are created by the discriminator. The labels are then used to provide the feedback to both the discriminator and the generator, and the algorithm makes the appropriate changes to both models to optimise them based on the feedback received. Figure 4.2 displays the structure of a GAN model and the training process. As the figure shows, both the discriminator and the generator receive feedback from the discriminator output. In practice, this feedback is a loss function (Section 3.2.6 in Chapter 3), evaluated comparing the generated labels to the real labels. (165, 166) The specific loss functions used for the GAINs in this chapter are described in Section 4.3.3.

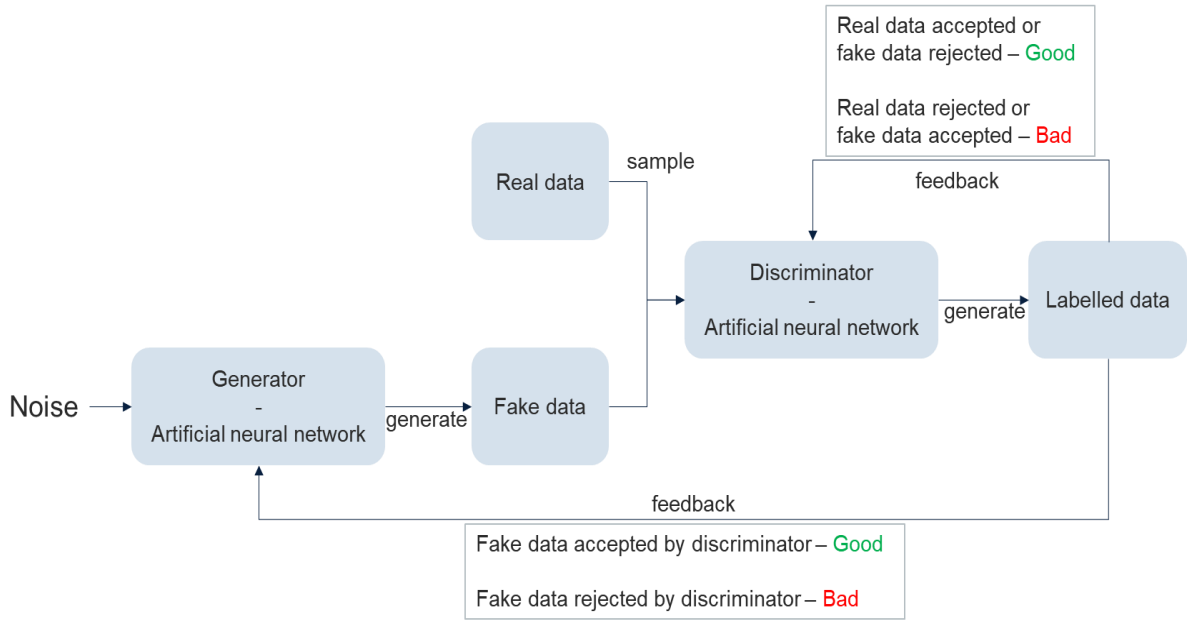

Figure 4.2: GAN structure

In practice, training GANs involves:

1. Specifying structure and layers for both generator and discriminator
2. Specifying the loss functions for both generator and discriminator which is used for adjusting model parameters based on the feedback from the labelled data during training
3. Training loop consisting of taking random noise as input, generating fake data, mixing fake data with a sample of real data, labelling data and using backpropagation (Section 4.2.2.2 and Section 3.2.6 in Chapter 3) to adjust model parameters based on the loss functions

4.2.2.1 Artificial neural network structure and layers

Chapter 3.2.8 describes the basic principles of artificial neural networks and shows the example of a simple fully connected ANN. A fully connected ANN consists of an input layer, and a number of dense hidden layers that generate the final output. Dense layers have a specified number of nodes, each of which are connected to every node in the previous layer. The ANNs used to create GANs often have more complicated structures. These ANNs often include other types of layers on top of dense layers, such as layers for normalisation and activation layers that apply the specified activation function to all nodes in the previous layer.

Some GANs often used for image analysis can be made even more complex with their specialised layers such as conditional GANs, deep convolutional GANs or recurrent GANs. These are, however, outside the scope of this work, and thus not covered in this chapter.

4.2.2.2 Backpropagation

We have already covered the concept of backpropagation for ANNs in Section 3.2.6 in Chapter 3. In the case of GANs, the two ANNs (generator and discriminator) are trained simultaneously. The costs to both networks are dependent on the other – when generator is improved, discriminator performance worsens, when discriminator is improved, generator performance worsens – and both are trained to minimise respective costs, which results in a competing performance until a balance is reached and neither can improve further.

4.2.2.3 GAINs (163, 164)

GAINs (163, 164), or Generative Adversarial Imputation Networks, are an extension of GANs used specifically for missing data imputation. While GANs are primarily used to generate new complete data samples or datasets, GAINs focus on imputing the missing values within existing incomplete datasets. Thus, GANs and GAINs differ in their data labelling process – GANs generating one label for an entire data sample or dataset, while GAINs apply labels to each individual value within the data. In other words, GAINs consider the missing values themselves as the targets of generation and discrimination, allowing us to have a combination of real and generated data within a dataset.

As Figure 4.3 shows, GAINs are trained on complete datasets with generated missing values created by removal of data points, rather than using already incomplete data for the training. This is because GAINs aim to generate imputations that are as close as possible to the real data. The loss function for the generator in GAINs includes a term (usually a MSE, on top of the binary cross-entropy loss from the discriminator) that measures the difference between the generated data and the actual complete data, ensuring that the imputed values follow the observed patterns in the dataset.

A notable distinction between the GAIN and GAN structure, as seen in the figures, is the input of the generator. While the GAN generator is provided with noise, the generator of a GAIN has access to further information in the form of 0/1 indicator values of missingness, and the

observed values in the dataset with missing data. By considering this extra information, GAINs can better understand the underlying patterns and dependencies in the dataset, enabling imputed values more appropriate to each individual sample, rather than values which could belong to any sample in the dataset.

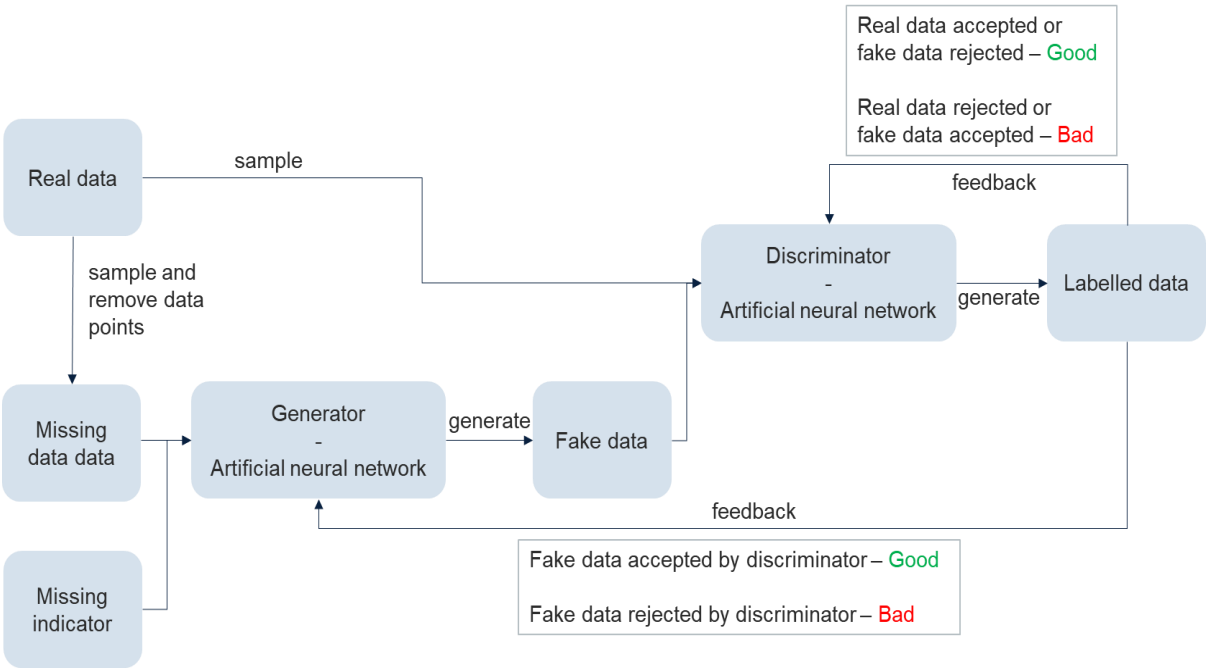

Figure 4.3 GAIN structure

4.3 Methods

4.3.1 Data, variable groups and imputation

Data was used from 11472 patients combined from 9 studies, made up from the 7 studies described in Chapters 2 and 3, data of 2901 patients from the Fetal Medicine Foundation, and a small dataset of 208 patients from Brazil. The Fetal Medicine Foundation (FMF) data was collected from electronic health records as part of a prospective observational cohort of UK women with singleton pregnancies, diagnosed with preeclampsia. Data was collected between December 2013 and December 2021 at King's College Hospital, London, and Medway Maritime Hospital, Gillingham. Data was collected for the validation of the PIERS-ML model, after which it was combined into the panPIERS dataset. A breakdown of the adverse maternal outcomes in this dataset can be found in Table *2.12*.

The Brazil dataset came from a validation study of the fullPIERS model, a cross-sectional study of women diagnosed with preeclampsia admitted to the Women's Hospital at the University

of Campinas, Brazil between January 2017 and February 2018 (167). The aim of the study was to validate the fullPIERS model in a new, unseen cohort, the study included all women admitted for childbirth who were diagnosed with preeclampsia. Only 27 patients had an adverse maternal outcome in this dataset, all within 2 days of admission. The most severe outcome recorded was one case of maternal death, while the most common adverse outcomes were eclamptic seizure, blood transfusion and acute renal failure (8 occurrences each), followed by 5 placental abruptions, 2 cases of hepatic dysfunction.

Similarly to the static modelling, data from the studies was cleaned as described in Section 2.3.1 and combined into a single dataset. Only data from day of or day after admission was used. For each patient, the worst value of each variable with repeated measurements was taken. Data was split into development (2/3 data), testing (1/6 data) and validation (1/6 data) data.

The 35 predictor variables were grouped into baseline always available information plus 7 variable groups: patient data, medical history, signs, symptoms, blood tests part 1, blood test part 2 (the measures of coagulation) and urine (Table 4.1). These variables were grouped together as they were expected to be mostly likely to be available at the same time – i.e. if one symptom variable was available, it is likely all symptom variables would be available, and collected simultaneously. The coagulation variables were separated from the other blood test variables as these variables are less likely to be collected routinely in low resource settings.

Table 4.1: Variable groups and the corresponding predictor variables

Group	**Variables**
Baseline	Gestational age on admission, maternal age at expected delivery date, systolic and diastolic blood pressure, national maternal mortality rate and national per capita GDP
Patient data	singleton/multiple pregnancy, ethnicity, parity
Medical history	renal disease, chronic hypertension, pre-gestational diabetes, gestational diabetes in previous pregnancy, history of smoking
Signs	SpO_2, height on admission, weight on admission
Symptoms	vomiting or nausea, right upper quadrant or epigastric pain, headache or visual disturbances, chest pain or dyspnoea

Blood tests part 1	total leucocyte count, platelet count, mean platelet volume, uric acid, hematocrit, serum creatinine, aspartate transaminase, alanine transaminase, lactase dehydrogenase, serum albumin
Blood tests part 2 (coagulation)	International normalised ratio, fibrinogen, activated partial thromboplastin time
Urine	Dipstick proteinuria

For each possible combination of the 7 variable groups (128 combinations in total), a dataset was created for development, testing and validation. For each combination, the variables belonging to the selected variable group(s), the baseline variables and the outcome variable were selected, and all other variables were treated as missing, thereby artificially creating every possible scenario in our data. After the selection of variables, for each combination in the development dataset, M_i *= proportion of missing in the dataset corresponding to the ith variable group combination, rounded to the nearest percent* was calculated. Each development dataset was imputed M_i times using multiple chained random forests. Multiple imputation was not used on the testing and validation datasets.

4.3.2 Modelling

Logistic regression, random forest and LASSO methods were used for modelling. Neural networks have been the most computationally expensive of the modelling methods covered in Chapter 2, however performed poorly in predicting the adverse maternal outcome, therefore the method was not considered in this analysis. Models were fitted using tidymodels in R. Tidymodels are specified using a "model recipe", which describes the steps of pre-processing and model formula; model specification, including the computational engine to be used for modelling (a package or software), the mode of modelling (regression or classification) and any additional model options or parameters; and workflow consisting of adding the model specification to the recipe and fitting the model.

The following modelling specifications and steps were completed for each variable group combination.

Model recipe was the same for logistic regression and LASSO methods, consisting of the steps of updating the ID variable to have an ID role and thus not be used as a predictor variable,

creating dummy variables with 0 and 1 values of categorical variables, normalising numerical variables and finally modelling the outcome variable by all variables except ID. While other functions used in R to create logistic regression models, such as lm or glm, do not require this step of creating dummy variables as part of the data pre-processing as the functions identify any categorical variables and create the dummy variables as part of the model fitting process, tidymodels includes this step in the pre-processing stage to allow for greater flexibility in modelling. An extra step was also included to remove any variables with zero variance (in other words, variables that have the same value for all patients), however no variables met this condition.

The model recipe for random forest was much simpler since the method could easily handle categorical variables and variables with very different variances. The model recipe for random forest consisted of updating the ID variable to have an ID role and thus not be used as a predictor variable in modelling, followed by modelling the outcome variable by all other variables.

All model specifications included "classification" mode, which indicates that the outcome is a categorical variable rather than continuous. For logistic regression the glm engine was used, random forest was fitted with randomForest engine and the glmnet engine was used for LASSO.

With each method, steps were taken to simplify models and create a final model per method that only included variables that were significant predictors of the outcome to minimise the number of variables required to make predictions, as discussed in Section 2.5.2. First, logistic regression models were fitted using all predictor variables with the recipe and specifications above. A model was fitted on each imputed dataset. For all resulting models, the p-value for each predictor variable was obtained, associated with the observed T-statistic used to test the hypothesis that the corresponding regression term is non-zero. Variables that had a p-value<0.05 in at least one of the imputed datasets were kept as predictor variables for the final model, while the rest were removed. A new model was then fitted on each imputed dataset using only the remaining predictor variables, with the same specifications and recipe as before. For each variable group combination i, if M_i was greater than 1, and thus multiple imputed datasets were created for the combination, the resulting final model was a list of the final model from each imputed dataset. If M_i was not greater than 1, and thus the

combination was imputed only once, the resulting final model was the single final model created on the imputed dataset.

Variable selection for random forest was carried out similarly. First, random forest models were fitted on each imputed dataset using all predictor variables, with the recipe and specifications above. For all resulting models, variable importance was calculated for each variable using mean decrease in Gini index. The average importance was also calculated per model. Variables that had an importance greater than the corresponding average importance in at least one of the imputed datasets were kept as predictor variables for the final model, while the rest were removed. A new model was then fitted on each imputed dataset using only the remaining predictor variables, with the same specifications and recipe as before. If the variable groups combination had more than one imputed dataset, the resulting final models were combined into a list.

LASSO automatically carries out variable selection as part of its fitting process as the coefficients of less significant predictors are shrunk to or close to zero, hence the extra step of variable selection was not necessary for this method. However, this method had to be tuned to find the optimal penalty value. Here extra information was provided in the model specification: within the glmnet engine, mixture was set the value of 1 as that specifies LASSO method (rather than ridge regression or elastic net), and at first penalty was set as 0. Bootstrapping was used to tune the penalty on a tuning grid with 50 levels. Best penalty was chosen based on highest AUC. Once penalty was tuned, model specification was changed to include the new penalty (using the same penalty on all imputed datasets) and models were refitted. If the variable groups combination had more than one imputed dataset, the resulting final models were combined into a list.

4.3.3 GAINs

GAINs were created for each variable group using a combination of R and python software. For the GAIN for the base variable group, only variables belonging to this group were included in the GAIN, while for all other variable groups, variables belonging to the specified variable group or the base variable group were utilised.

To create the GAINs, first training datasets for each were created in R by selecting variables as specified above and creating a subset of the data by keeping complete observations only.

A second dataset was also created for each by simulating missing values for each variable by randomly deleting 20% of values per variable.

Creating the GAIN structure, the generator and the discriminator, and the training process were all carried out in python. The layers and loss functions of the models were based on the implementation of GAIN published by Dong et al (163), by updating the code available on Github (https://github.com/dongdongdongdwn/GAIN-Dovey) to use the newest version of tensorflow. During training, samples of 200 observations (a batch) were taken from the dataset with generated missingness, which was the batch size of the original GAIN code, then the generator was used to replace the missing values. The completed dataset was then supplied to the discriminator which labelled the data points as real or fake data. The loss of the generator (Gloss) and of the discriminator (Dloss) was calculated using Equation 4.1 and Equation 4.5 (163).

$$L_{M1} = \sum_{j=1}^{N_{col}} \frac{1}{N_{row}} \sum_{i=1}^{N_{row}} ((1 - M_{i,j}) X_{i,j} I_{cont_j} - (1 - M_{i,j}) \hat{X}_{i,j} I_{cont_j})^2 \qquad \textit{Equation 4.1}$$

$$L_{M2} = \sum_{j=1}^{N_{col}} -\frac{1}{N_{row}} \sum_{i=1}^{N_{row}} \left(1 - I_{cont_j}\right) X_{i,j} M_{i,j} \log(\hat{X}_{i,j}) \qquad \textit{Equation 4.2}$$
$$+ \left(1 - X_{i,j}\right)\left(1 - I_{cont_j}\right)(1 - M_{i,j}) \log\left(1 - \hat{X}_{i,j}\right)$$

$$L_G = -\frac{1}{N_{row} * N_{col}} \sum_{i=1}^{N_{row}} \sum_{j=1}^{N_{col}} (1 - M_{i,j}) \log\left(D(\hat{X}, H)_{i,j}\right) \qquad \textit{Equation 4.3}$$

$$G_{loss} = L_G + \alpha L_{M1} + \beta L_{M2} \qquad \textit{Equation 4.4}$$

In Equation 4.1 to Equation 4.5, X is the original, complete dataset of size N_{row}xN_{col}, M is a matrix the same size of X with values of 0 and 1 marking missing values, I_{cont} is a vector of 0s and 1s indicating continuous variables, $\hat{X}$ is the generated dataset and $D(\hat{X}, H)$ is the matrix of labels generated by the discriminator for $\hat{X}$ based on the hint matrix *H*. L_{M1} is the sum of MSE of missing values in numeric columns and L_{M2} is the sum of cross entropy of missing values in the categorical variables.

$$D_{loss} = \frac{1}{N_{row} * N_{col}} \sum_{i=1}^{N_{row}} \sum_{j=1}^{N_{col}} M_{i,j} log D\left(\hat{X}, H\right)_{i,j} + \left(1 - M_{i,j}\right) \log\left(1 - D\left(\hat{X}, H\right)_{i,j}\right)$$

Equation 4.5

The generator and the discriminator were then both backpropagated using the calculated losses, the generator is trained to minimise G_{loss}, while the discriminator is trained to maximise D_{loss}, and losses were recorded to to track performance over training instances. The process was then repeated with a new sample 3000 times to allow the loss functions to converge.

To use the trained GAINs for imputation of new data, an R function was created (Appendix A), which, moving through each variable group one by one, would first select the rows of data with either all variables observed or all variables missing and return these rows without any changes, then take the remaining data and apply the corresponding GAIN to impute missing value(s), thus only imputing variable groups if at least one variable of the group was observed, otherwise imputation would be based on random noise only.

4.3.4 Model assessment

The testing and validation datasets were used for model assessment. For each variable group combination, GAINs per variable group were used on the testing and validation datasets to impute missing data. These GAINs were only used for their corresponding variable groups if at least one variable in the group was observed, thus not all missing values were imputed. Moreover, using the same GAIN on the same data will always create the same prediction, therefore multiple imputation of the training and validation datasets was not needed, and only one imputed dataset was created per variable group combination. Models corresponding to the variable group combination were used on the complete cases in the training and validation datasets. As models were created on multiple imputed datasets, in most cases of variable group combinations, more than 1 model was created, so to make predictions on a new, unseen dataset, each of the corresponding models were used to make predictions, the mean of which was returned as the final predicted probability.

Using the methods described in Section 2.2.9 in Chapter 2, thresholds for the risk classification groups were determined on the testing dataset, and patients in the validation dataset were classified into risk groups accordingly. On the validation dataset the positive likelihood ratio,

and the number and percentage of outcomes for the very high risk group was recorded and later used to assess the ability of the models to rule in outcome. Similarly, the negative likelihood ratio as well as the number and percentage of outcomes was recorded for the very low risk group for assessment of ability to rule out the outcome.

This resulted in a list of performance measurements corresponding to the list of 128 models (some of which were lists of models themselves). This process was repeated for each modelling method.

For each modelling method, tables of summaries of model performance in the very low and very high risk groups was created.

Since many variable group combinations are subsets of others, it is possible to use not only the corresponding model on a variable group combination but also models corresponding to combinations that are subsets of it. For example, in the combination consisting of the signs and symptoms groups with the baseline group, naturally it is possible to use the model created on the signs and symptoms combination, the model created on the signs group, the model created on the symptoms group, and the model made for the baseline group. However, it is also possible that after variable selection, a model created on a completely different variable group combination, for example symptoms and medical history, would not use variables from all groups in the combination. In case of a model created on symptoms and medical history, a model could be made only using symptom variables, and thus would be useable on the combination of signs and symptoms.

Hence, for each variable group combination, a list of available variables was created and compared to the variables used in each of the 128 models of each method. Models for which all required variables were available were listed for each variable group combination. When aiming to rule in an outcome, the listed models were ranked based on decreasing positive likelihood ratio and decreasing percentage of patients classed in the very high risk group, and the highest ranking model was selected. This ensured that we selected models that could identify the most patients at the highest risk of an adverse outcome.

For ruling out, ranking was based on increasing percentage of outcomes in the very low risk group and decreasing percentage of patients in the very low risk group, and the highest ranking model was selected. This method was chosen over the negative likelihood ratio as

models could have very low negative likelihood ratios by only classifying 1 patient with no outcome into the very low risk group. Ensuring that a high percentage of patients were classified into the group with a very low outcome rate, while still making sure that the negative likelihood ratio was acceptably low gave us better confidence in the system's utility for ruling out the outcome.

4.3.5 panPIERS process and application

To apply the models created in this chapter to new, unseen data, a process, shown in Figure 4.4, was created and embedded in an RShiny application.

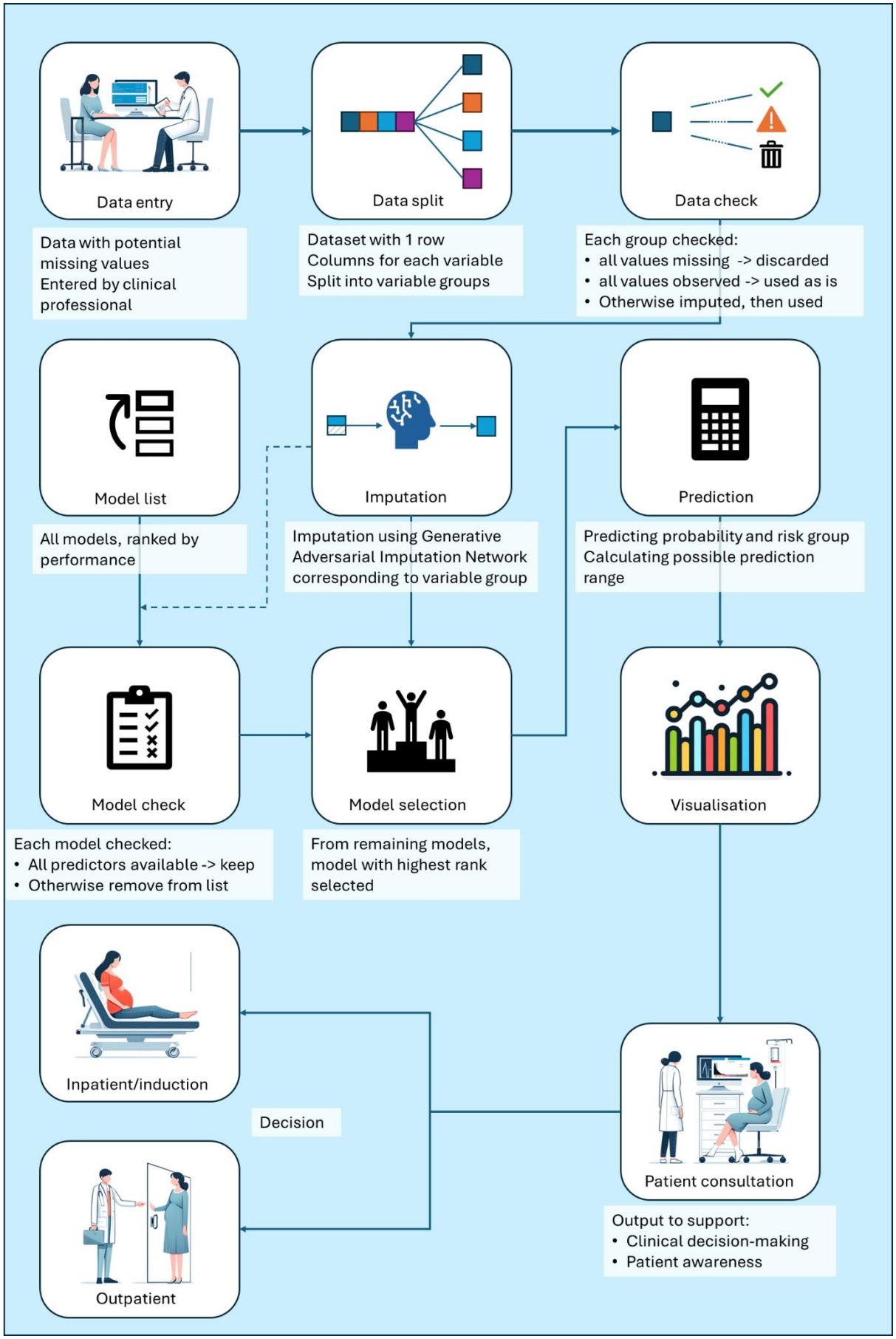

Figure 4.4: Prediction process on new data with the panPIERs tool

The process begins with data entry into the application by a user, where missing values may be present in the dataset. To address this, the GAINs are applied on a per variable group basis, where if at least one variable in a group has observed values, the corresponding GAIN is used to impute missing values for that group. The imputed values are recorded, and information about the available variables is stored.

Next, the models are filtered to retain only those groups in which all variables are available, ensuring that no missing values exist. From the filtered models, the highest ranked model is selected based on the ruling in or ruling out criteria, described above, as selected by the user. This model is then utilized to predict outcomes on the GAIN-imputed dataset, incorporating the imputed values.

In cases where the data entered by the user contained missing values, the next step returns to the original data with missing values, and the variables used in the previously selected model are isolated for uncertainty quantification. If any of the variables used in the model contain missing values, we calculate a lower and upper bound for the predicted probability by replacing the missing values with the 2.5 and 97.5 quantile values calculated from the observed training data. If more than one variable is missing a value, a data grid is constructed, considering all possible combinations of replacements for the missing data. Predictions are generated for each combination, and the lowest and highest predicted probabilities are selected and reported. This reporting strategy provides insights into the range of potential outcomes, accounting for both conservative and optimistic estimates, thereby supporting informed decision-making.

Visualisations of the predicted risk, as well as comparisons with average predicted risks of patients at the same gestational age in weeks are created and returned on the user interface of the application. Text summaries of the risk, missing values and ranking of models are also displayed for the user.

An example of the layout and outputs of panPIERS app for a set of random variable values and missing data can be found at https://tundecsoban.shinyapps.io/Scenario3/ .

4.3.6 Usability testing

A study was conducted to gather input from clinicians, who would be the target users, on the usability and usefulness of the panPIERS app. A 15-20 minutes long interviews were designed,

including questions on background information on the participant, a randomly selected scenario out of four possibilities, where observed data and the generated output was showed to the participant, and questions interpreting the outputs were asked, and some general feedback questions on the app.

The interviews were conducted in person at the International Society for the Study of Hypertension in Pregnancy (ISSHP) 2023 world congress in Bangalore, India to include responses from clinicians working in low- and middle-income countries who would be difficult to reach for online interviews.

Results were analysed using RStudio, bar charts and pie charts were created from questions with categorical answers, and table summaries and word clouds were created of free text answers, presented in Section 4.4.3.

4.4 Results

4.4.1 The data

Table 4.2 shows the characteristics of patients with and without outcome at any point in the total combined dataset with no imputation. Of the observed values, there was a significant difference between the group of patients in all baseline variables, patient information and symptom variables, variables in the second part of blood tests, and urine dipstick. All but one blood test variables were significantly different between patients with and without an outcome at any point.

In the combined dataset of 11472 patients, all variables in the patient information variable group were missing for 3057 (26.7%) patients, symptoms were all missing for 1930 (8.6%), at least one variable for medical history was recorded for all patients, blood tests part 1 was all missing for 1265 (11.0%), blood tests part 2 for 6364 (55.5%), and urine for 3486 (30.4%).

For 994 (8.7%) patients some, but not all variables in the patient information group were missing, similarly some but not all variables were missing in the symptoms, medical history, signs, and bloods part 1 and part 2 were missing for 1930 (16.8%), 3444 (30.0%), 4611 (40.2%), 8258 (72.0%), and 1414 (12.3%) patients, respectively.

Table 4.2: Baseline characteristics and proportion of missingness per variable group in the panPIERS data

Variable group	Variable	Proportion of women with no observation for any variables in the group (of N=11472)	Proportion of women with observations for all variables in the group (of N=11472)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
Baseline	Gestational age at eligibility (weeks)	0 (0%)	10917 (95.162%)	33 [29.98-36.64]	36.57 [32.86-38.78]	<0.001
	Maternal age at expected date of delivery (years)			30 [26-35]	31 [27-35]	0.026
	Systolic blood pressure (mm Hg)			154 [143-168]	150 [140-160]	<0.001
	Diastolic blood pressure (mm Hg)			99 [90-105]	95 [90-100]	<0.001
	National per capita gross domestic product (USD)			42030.29 [9925.39-44974.83]	42330.12 [36266.19-47425.61]	<0.001
	National maternal mortality ratio (maternal deaths per 100,000 live births)			11 [8-62]	11 [7-15]	<0.001
Patient information	Multiple pregnancy	3057 (26.647%)	7421 (64.688%)	92 (11.23%)	522 (4.9%)	<0.001
	Race			-	-	0.005
	White			233 (28.45%)	2185 (20.51%)	-
	Asian			154 (18.8%)	1974 (18.53%)	-
	Black			198 (24.18%)	1756 (16.48%)	-
	Other			81 (9.89%)	843 (7.91%)	-
	Nulliparous			543 (66.3%)	7476 (70.18%)	0.022
	Pre-gestational renal disease	0 (0%)	8028 (69.979%)	51 (6.23%)	515 (4.83%)	0.762

Variable group	Variable	Proportion of women with no observation for any variables in the group (of N=11472)	Proportion of women with observations for all variables in the group (of N=11472)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
Medical history	Chronic hypertension			106 (12.94%)	1308 (12.28%)	0.124
	Pre-gestational diabetes			666 (81.32%)	7195 (67.54%)	0.221
	Gestational diabetes			54 (6.59%)	965 (9.06%)	<0.001
	Cigarette smoking			105 (12.82%)	950 (8.92%)	0.102
Signs	Oxygen saturation less than 93%	6 (0.052%)	6855 (59.754%)	129 (15.75%)	1039 (9.75%)	<0.001
	Height(cm)			162.56 [158-167.63]	163 [158-168]	0.094
	Weight (kg)			77.85 [68-89.85]	82 [71.21-96]	<0.001
Symptoms	Nausea or vomiting	985 (8.586%)	8557 (74.59%)	71 (8.67%)	596 (5.59%)	<0.001
	Headache or visual disturbance			176 (21.49%)	1578 (14.81%)	<0.001
	Right upper quadrant or epigastric pain			104 (12.7%)	642 (6.03%)	<0.001
	Chest pain or dyspnoea			59 (7.2%)	194 (1.82%)	<0.001
Bloods part 1	Total leucocyte count (x 10^9 per L)	1265 (11.027%)	1949 (16.989%)	10.7 [8.4-13.8]	11.1 [9-14.1]	0.003
	Platelet count (x 10^9 per L)			189 [143-241.75]	205 [163-250]	<0.001
	Mean platelet volume (fL)			10.4 [9-11.87]	10.8 [9.7-11.8]	0.001
	Uric acid (mmol/)L			365 [301-438]	350 [290-410]	<0.001
	Haematocrit (%)			0.35 [0.32-0.38]	0.36 [0.34-0.38]	<0.001

Variable group	Variable	Proportion of women with no observation for any variables in the group (of N=11472)	Proportion of women with observations for all variables in the group (of N=11472)	Women with adverse outcome at any time after first assessment (N=1138, 12·6%)	Women without an adverse outcome (N=7921, 87·4%)	p-value
	Serum creatinine (µmol/L)			65 [53-79]	58.41 [50-69]	<0.001
	Aspartate transaminase (U/L			30 [21-44.5]	26 [20-36]	<0.001
	Alanine transaminase (U/L)			21 [13-36]	18 [12-28]	<0.001
	LDH			540 [375.75-741.5]	548 [434-696.75]	0.988
	Albumin (g/L)			28 [17-32]	31 [26-35]	<0.001
Bloods part 2	International Normalised Ratio	6364 (55.474%)	3694 (32.2%)	0.92 [0.9-1]	0.91 [0.9-0.94]	<0.001
	Fibrinogen (g/L)			26 [23.3-29.9]	25.1 [23.1-27.5]	<0.001
	Activated partial thromboplastin time (sec)			27.6 [25-30.5]	26.7 [24.8-29]	<0.001
Urine	Dipstick proteinuria (number of '+')	3486 (30.387%)	7986 (69.613%)	2 [1-3]	1 [0-2]	<0.001

4.4.2 Modelling

Results in this section will be presented in detailed tables. For each modelling method, a table presents the model performance and ranking of each model in the testing dataset corresponding to the dataset the model was trained on. For example, the performance of model 13, which was built on the development dataset corresponding to scenario 13 (baseline data, patient information and the first part of blood test results), when applied to the testing dataset corresponding to scenario 13. A missing ruling in or ruling out performance in this table indicates that no threshold could be determined for that model to classify patients into a very high or a very low risk stratum, respectively. Models reaching the desirable performance on ruling in and/or ruling out are also highlighted in the table, ruling in performance of a model is highlighted in green if the positive likelihood ratio was at least 10, and highlighted in yellow if the positive likelihood ratio was less than 10 but no less than 9. Similarly, the ruling out performance of a model is highlighted in green if the negative likelihood ratio was no more than 0.1, and highlighted in yellow for values greater 0.1 but no more than 0.2.

A second table, for each modelling method, shows the performance of models selected based on the hierarchy for ruling in and ruling out. This table lists the models that would be used for each scenario, determined by availability of all model variables in the variable group combination and the existence of ranking for ruling in/out of the model. The table also shows the selected model from the useable model for ruling in/out, and its performance on the testing dataset corresponding to the given scenario. For example, if models 1, 2, 5 and 13 could be used for scenario 13, but model 5 was selected for ruling in, then the table would show the performance of model 5, which was built on the development dataset corresponding to scenario 5, when applied to the testing dataset corresponding to scenario 13. The scenarios where the positive likelihood ratio on ruing in is at least 10, and/or the negative likelihood ratio on ruling out is at most 0.1 are highlighted in green. If no models could be found for ruling in and/or ruling out, the corresponding ruling in and/or ruling out performance is highlighted in red.

4.4.2.1 Logistic regression

Table 0.8 in Appendix C shows the performance of logistic regression models to rule in and rule out the composite maternal outcome in each variable group combination. Risk strata thresholds were selected on a testing dataset to achieve desired likelihood ratios,

as previously described. If a threshold for the very high risk group could be found such that the positive likelihood ratio on the testing dataset was at least 10, this threshold was used to determine the ruling in performance on the validation dataset, shown in the table. If a threshold could not be found, the ruling in performance for that model is left blank in the table. The ruling out performance was determined similarly. I was able to classify patients into a very high risk group (Appendix C, Table 0.8) with 46.1% (59 of 128) models. In the validation datasets, 33.9% (20 out of the 59) models achieved a positive likelihood ratio (LR+) greater than 10, while 10.2% (6 of 59) had LR+ between 9 and 10. Variable group combinations with larger number of variable groups had less models which could rule in the outcome, however a bigger proportion of the models had LR+≥10 It should be noted that the top 5 ranked models each classified only a small proportion of the population into the very high risk groups, at or under 1% of the population each.

I was able to classify patients into a very low risk group (Appendix C, Table 0.8) with only 61.7% (79 of 128) of the models. In the validation datasets, 10.1% (8 out of the 79) models achieved a negative likelihood ratio (LR-) less than 0.1, while 2.5% (2 of 79) had LR- between 0.1 and 0.2. All models classified a larger proportion of the population into the very low risk group than into the very high risk group, and there were less than 5 outcomes in the very low risk group in the validation dataset for most models, even if the LR- was greater than 0.2.

Based on the ruling in criterion for choosing any model for which all variables were available in each of the variable group combinations, 13 models were selected (models 1, 8, 14, 15, 16, 18, 19, 23, 29, 32, 45, 51, and 103, Table 0.9 Appendix C). Of the chosen models, only one model had LR+<9 (model 1), four models showed LR+ values between 9 and 10, and the remaining 7 demonstrated LR+ values exceeding 10, indicating their effectiveness in ruling in the composite maternal outcome.

One of the selected 13 models was chosen in all 128 variable group combination, meaning that a prediction to rule in outcome could be made regardless of what information was provided, with the confidence of LR+>10 48.4% (62 of 128) of the time.

Seven models (1, 18, 29, 44, 48, 60, and 90, Table 0.9 Appendix C) were needed to cover all variable group combinations when selecting models based on the ruling out criterion

for choosing any model for which all variables were available. Importantly, all of these models achieved LR-≤0.1 on their corresponding validation dataset at model ranking, indicating that they can be used effectively to rule out the composite maternal outcome with LR-<0.1.

4.4.2.2 Random Forest

Table 0.10 in Appendix C shows the performance of random forest models to rule in and rule out the composite maternal outcome in each variable group combination. I was able to classify patients into a very high risk group with 50% (64 of 128) of the models. In the validation datasets, 12.5% (8 out of the 64) models achieved LR+≥10, while 3.1% (2 of 64) had LR+ between 9 and 10. It should be noted that the top 5 ranked models each classified only a small proportion of the population into the very high risk groups, under 1% of the population each.

I was able to classify patients into a very low risk group (Appendix C, Table 0.10) with only 8.6% (11 of 128) of the models. In the validation datasets, 10 of the 11 models achieved LR-≤0.1, while the last model had LR->0.2. There were not enough models able to classify into the very low risk group to look for any relationship between the number of models with desirable LR- and the number of variable groups in the combination.

Based on the ruling in criterion for choosing any model for which all variables were available in each of the variable group combinations, only four models were selected (models 34, 45, 73, and 68, Appendix C Table 0.11). Of these models, only three (45, 73 and 68) were also among the top 5 ranked model, which were also the only models of the chosen four with LR+≥10 on their corresponding validation dataset. When tested based on the ruling in criterion, choosing the best model for each variable group combination, I could rule in an outcome for all combinations, however 9.4% (12 of 128) of combinations did not reach the LR+ of 10. As one of the selected 3 models was chosen in all 128 variable group combination, a prediction to rule in outcome could be made regardless of what information was provided, with the confidence of LR+>10 in most cases.

Only two models (models 45 and 105, Appendix C Table 0.11) were used in 75% (96 of 128) of variable group combinations when selecting models based on the ruling out criterion for choosing any model for which all variables were available. The outcome

could not be ruled out in the remaining 32 variable group combinations. The models used were ranked number 1 (model 45) and number 12 (model 105), and only model 45 achieved LR-≤0.1 on its corresponding validation dataset at model ranking. When tested using the ruling out criterion, I could achieve the desired LR- for 26% (25 of 96) of the variable group combinations that had a selected model, meaning that while I could rule out the outcome for the majority of variable group combinations, I did not have the desired level of confidence.

4.4.2.3 LASSO

Table 0.12 in Appendix C shows the performance of LASSO models to rule in and rule out the composite maternal outcome in each variable group combination. I was able to classify patients into a very high risk group with 39.1% (50 of 128) models. In the validation datasets, 24% (12 out of the 50) models achieved a positive likelihood ratio (LR+) greater than 10, while 2% (1 of 50) had LR+ between 9 and 10. Variable group combinations with larger number of variable groups had less models which could rule in the outcome. Similarly to before, the top 5 ranked models each classified only a small proportion of the population into the very high risk groups, under 1% of the population each.

I was able to classify patients into a very low risk group (Appendix C, Table 0.12) with the most models, with 85.9% (110 of 128). In the validation datasets, 10.1% (50 out of the 110) models achieved a negative likelihood ratio (LR-) less than 0.1, while 2.5% (3 of 110) had LR- between 0.1 and 0.2. Unlike with the other two methods, most but not all models classified a larger proportion of the population into the very low risk group than into the very high risk group, and and all but 4 models had less than 5 outcomes in the very low risk group, even if the LR- was greater than 0.2.

Based on the ruling in criterion for choosing any model for which all variables were available in each of the variable group combinations, 11 models were selected (models 1, 3, 7, 17, 29, 34, 54, 71, 74, 90, and 105, Appendix C Table 0.13). Of the chosen models, only five had LR+<9 (models 17, 34, 71, 74, and 105), one model (model 90) showed LR+ values between 9 and 10, and the remaining five demonstrated LR+ values exceeding 10, indicating their effectiveness in ruling in the composite maternal outcome. One of the selected 11 models was chosen in all 128 variable group combination, meaning that

a prediction to rule in outcome could be made regardless of what information was provided, with the confidence of LR+>10 31.3% (40 of 128) of the time.

Ten models (models 1, 3, 17, 40, 41, 51, 69, 73, 100, and 122, Appendix C Table 0.13) were needed to cover all variable group combinations when selecting models based on the ruling out criterion for choosing any model for which all variables were available. One of the models were chosen in all 128 variable groups and achieved LR-<0.1, meaning that a prediction to rule out outcome could be made regardless of what information was provided.

4.4.2.4 Combined ruling in

I also tested the selection of models from all available models including across the different methods used. As the best model for each scenario was previously selected for each modelling method, only the performance of the best model per method was compared for each scenario. The results of selection of models for ruling in is shown in Table 4.3. Altogether 12 models were needed to rule in the outcome in all 128 scenarios, eight logistic regression models (model 1, 14, 15, 188, 23, 26, 51, and 103), three random forest models (models 45, 68, and 73) and one LASSO model (model 3). As a model was selected in all 128 scenarios, the panPIERS system was able to rule in an outcome within two days, regardless of what information was available. Moreover, nearly all scenarios, 87.5% (112 of 128), showed a positive likelihood ratio of at least 10, of the remaining 16 scenarios, in the six scenarios where logistic regression models 15 or 18 were used, the positive likelihood ratio was between 9 and 10, and only the 10 scenarios using logistic regression model 1, or LASSO model 3, which were the simplest models, did I see positive likelihood ratios under 9. Generally, I saw better performance of models in scenarios with more variable groups available, with no scenarios with more than four available variable groups having a positive likelihood ratio under 10.

Table 4.3: Ruling in performance of models selected from all three methods

Scenario	Models	Selected method	Selected model	LR+
1	Logistic regression: 1, Random forest: 34, LASSO: 1	LR	1	LR+ 8.4 [95% CI 1-73]

Scenario	Models	Selected method	Selected model	LR+
2	Logistic regression: 1, Random forest: 34, LASSO: 1	LR	1	LR+ 6.7 [95% CI 0.8-55.9]
3	Logistic regression: 1, Random forest: 34, LASSO: 3	LASSO	3	LR+ 8.4 [95% CI 2.9-23.8]
4	Logistic regression: 1, Random forest: 34, LASSO: 1	LR	1	
5	Logistic regression: 1, Random forest: 45, LASSO: 1	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
6	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
7	Logistic regression: 1, Random forest: 34, LASSO: 7	LR	1	LR+ 8.4 [95% CI 1-73]
8	Logistic regression: 8, Random forest: 68, LASSO: 1	RF	68	LR+ 10.9 [95% CI 3.1-38]
9	Logistic regression: 1, Random forest: 34, LASSO: 3	LASSO	3	LR+ 8.4 [95% CI 2.9-23.8]
10	Logistic regression: 1, Random forest: 34, LASSO: 1	LR	1	LR+ 6.7 [95% CI 0.8-55.9]
11	Logistic regression: 1, Random forest: 45, LASSO: 1	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
12	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
13	Logistic regression: 1, Random forest: 34, LASSO: 7	LR	1	LR+ 6.7 [95% CI 0.8-55.9]
14	Logistic regression: 14, Random forest: 68, LASSO: 1	LR	14	LR+ 16.4 [95% CI 3.1-85.5]
15	Logistic regression: 15, Random forest: 34, LASSO: 3	LR	15	LR+ 9.3 [95% CI 3.6-23.7]

Scenario	Models	Selected method	Selected model	LR+
16	Logistic regression: 16, Random forest: 45, LASSO: 3	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
17	Logistic regression: 1, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
18	Logistic regression: 18, Random forest: 34, LASSO: 3	LR	18	LR+ 9.6 [95% CI 2.1-44.5]
19	Logistic regression: 19, Random forest: 68, LASSO: 3	RF	68	LR+ 10.9 [95% CI 3.1-38]
20	Logistic regression: 1, Random forest: 45, LASSO: 1	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
21	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
22	Logistic regression: 1, Random forest: 34, LASSO: 7	LR	1	LR+ 8.4 [95% CI 1-73]
23	Logistic regression: 23, Random forest: 68, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
24	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
25	Logistic regression: 1, Random forest: 45, LASSO: 7	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
26	Logistic regression: 8, Random forest: 45, LASSO: 1	RF	45	
27	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
28	Logistic regression: 8, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]

Scenario	Models	Selected method	Selected model	LR+
29	Logistic regression: 29, Random forest: 68, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]
30	Logistic regression: 15, Random forest: 34, LASSO: 3	LR	15	LR+ 9.3 [95% CI 3.6-23.7]
31	Logistic regression: 16, Random forest: 45, LASSO: 3	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
32	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
33	Logistic regression: 18, Random forest: 34, LASSO: 3	LR	18	LR+ 9.6 [95% CI 2.1-44.5]
34	Logistic regression: 14, Random forest: 68, LASSO: 34	LR	14	LR+ 16.4 [95% CI 3.1-85.5]
35	Logistic regression: 1, Random forest: 45, LASSO: 1	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
36	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
37	Logistic regression: 1, Random forest: 34, LASSO: 7	LR	1	LR+ 6.7 [95% CI 0.8-55.9]
38	Logistic regression: 23, Random forest: 68, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
39	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
40	Logistic regression: 1, Random forest: 45, LASSO: 7	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
41	Logistic regression: 14, Random forest: 45, LASSO: 1	LR	14	LR+ 16.4 [95% CI 3.1-85.5]

Scenario	Models	Selected method	Selected model	LR+
42	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
43	Logistic regression: 14, Random forest: 73, LASSO: 1	RF	73	
44	Logistic regression: 29, Random forest: 68, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]
45	Logistic regression: 45, Random forest: 45, LASSO: 3	LR	45	LR+ 12.3 [95% CI 5.5-27.5]
46	Logistic regression: 15, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
47	Logistic regression: 18, Random forest: 34, LASSO: 3	LR	18	LR+ 9.6 [95% CI 2.1-44.5]
48	Logistic regression: 23, Random forest: 68, LASSO: 3	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
49	Logistic regression: 16, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
50	Logistic regression: 16, Random forest: 45, LASSO: 3	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
51	Logistic regression: 51, Random forest: 45, LASSO: 3	LR	51	LR+ 16.4 [95% CI 6.6-40.3]
52	Logistic regression: 18, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
53	Logistic regression: 19, Random forest: 73, LASSO: 17	RF	73	

Scenario	Models	Selected method	Selected model	LR+
54	Logistic regression: 29, Random forest: 68, LASSO: 54	LR	29	LR+ 26.2 [95% CI 7.4-92]
55	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
56	Logistic regression: 1, Random forest: 45, LASSO: 7	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
57	Logistic regression: 23, Random forest: 45, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
58	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
59	Logistic regression: 23, Random forest: 73, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
60	Logistic regression: 23, Random forest: 68, LASSO: 29	LR	23	
61	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
62	Logistic regression: 8, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
63	Logistic regression: 29, Random forest: 45, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]
64	Logistic regression: 29, Random forest: 73, LASSO: 29	LR	29	
65	Logistic regression: 45, Random forest: 45, LASSO: 3	LR	45	LR+ 12.3 [95% CI 5.5-27.5]
66	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]

Scenario	Models	Selected method	Selected model	LR+
67	Logistic regression: 18, Random forest: 34, LASSO: 3	LR	18	LR+ 9.6 [95% CI 2.1-44.5]
68	Logistic regression: 23, Random forest: 68, LASSO: 34	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
69	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
70	Logistic regression: 16, Random forest: 45, LASSO: 3	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
71	Logistic regression: 51, Random forest: 45, LASSO: 71	LR	51	LR+ 16.4 [95% CI 6.6-40.3]
72	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
73	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	
74	Logistic regression: 29, Random forest: 68, LASSO: 74	LR	29	LR+ 26.2 [95% CI 7.4-92]
75	Logistic regression: 1, Random forest: 73, LASSO: 1	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
76	Logistic regression: 1, Random forest: 45, LASSO: 7	RF	45	LR+ 11.2 [95% CI 1.2-104.7]
77	Logistic regression: 23, Random forest: 45, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
78	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]

Scenario	Models	Selected method	Selected model	LR+
79	Logistic regression: 23, Random forest: 73, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
80	Logistic regression: 23, Random forest: 68, LASSO: 29	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
81	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
82	Logistic regression: 14, Random forest: 73, LASSO: 1	RF	73	
83	Logistic regression: 29, Random forest: 45, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]
84	Logistic regression: 29, Random forest: 73, LASSO: 29	LR	29	
85	Logistic regression: 45, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
86	Logistic regression: 45, Random forest: 45, LASSO: 3	LR	45	LR+ 12.3 [95% CI 5.5-27.5]
87	Logistic regression: 23, Random forest: 45, LASSO: 3	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
88	Logistic regression: 18, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
89	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
90	Logistic regression: 23, Random forest: 68, LASSO: 90	LR	23	

Scenario	Models	Selected method	Selected model	LR+
91	Logistic regression: 16, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
92	Logistic regression: 51, Random forest: 73, LASSO: 17	RF	73	
93	Logistic regression: 29, Random forest: 45, LASSO: 54	LR	29	LR+ 26.2 [95% CI 7.4-92]
94	Logistic regression: 29, Random forest: 73, LASSO: 17	LR	29	
95	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
96	Logistic regression: 23, Random forest: 73, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
97	Logistic regression: 23, Random forest: 45, LASSO: 29	LR	23	
98	Logistic regression: 23, Random forest: 73, LASSO: 29	LR	23	
99	Logistic regression: 29, Random forest: 73, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]
100	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
101	Logistic regression: 45, Random forest: 45, LASSO: 3	LR	45	LR+ 12.3 [95% CI 5.5-27.5]
102	Logistic regression: 23, Random forest: 45, LASSO: 71	LR	23	LR+ 32.7 [95% CI 6.9-154.7]

Scenario	Models	Selected method	Selected model	LR+
103	Logistic regression: 103, Random forest: 73, LASSO: 17	LR	103	LR+ 22.3 [95% CI 3.8-129.5]
104	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
105	Logistic regression: 23, Random forest: 68, LASSO: 105	LR	23	
106	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
107	Logistic regression: 32, Random forest: 73, LASSO: 17	RF	73	
108	Logistic regression: 29, Random forest: 45, LASSO: 71	LR	29	LR+ 26.2 [95% CI 7.4-92]
109	Logistic regression: 29, Random forest: 73, LASSO: 17	LR	29	
110	Logistic regression: 1, Random forest: 73, LASSO: 7	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
111	Logistic regression: 23, Random forest: 73, LASSO: 1	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
112	Logistic regression: 23, Random forest: 45, LASSO: 29	LR	23	
113	Logistic regression: 23, Random forest: 73, LASSO: 29	LR	23	
114	Logistic regression: 29, Random forest: 73, LASSO: 29	LR	29	LR+ 26.2 [95% CI 7.4-92]

Scenario	Models	Selected method	Selected model	LR+
115	Logistic regression: 45, Random forest: 73, LASSO: 17	RF	73	LR+ 16.7 [95% CI 3.2-88.5]
116	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
117	Logistic regression: 23, Random forest: 45, LASSO: 90	LR	23	
118	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	
119	Logistic regression: 29, Random forest: 73, LASSO: 17	LR	29	LR+ 26.2 [95% CI 7.4-92]
120	Logistic regression: 23, Random forest: 73, LASSO: 29	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
121	Logistic regression: 103, Random forest: 73, LASSO: 17	LR	103	LR+ 22.3 [95% CI 3.8-129.5]
122	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	LR+ 32.7 [95% CI 6.9-154.7]
123	Logistic regression: 23, Random forest: 45, LASSO: 71	LR	23	
124	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	
125	Logistic regression: 29, Random forest: 73, LASSO: 17	LR	29	LR+ 26.2 [95% CI 7.4-92]
126	Logistic regression: 23, Random forest: 73, LASSO: 29	LR	23	LR+ 32.7 [95% CI 6.9-154.7]

Scenario	Models	Selected method	Selected model	LR+
127	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	
128	Logistic regression: 23, Random forest: 73, LASSO: 17	LR	23	

4.4.2.5 Combined ruling out

Similarly, I also tested the selection of models from all available models including across the different methods used for the purpose of ruling out the outcome (Table 4.4). Altogether 15 models were needed to rule out the outcome in all 128 scenarios, which were seven logistic regression models (model 1, 18, 29, 44, 48, 60, and 90), and eight LASSO models (models 3, 17, 41, 51, 69, 73, 100, and 122). No random forest models were selected. As a model was selected in all 128 scenarios, the panPIERS system was also able to rule out an outcome within two days, regardless of what information was available. Importantly, all scenarios showed a negative likelihood ratio less than 0.1 as there were no outcomes within two days in the very low risk group for any scenario.

Table 4.4: Ruling out performance of models selected from all three methods

Scenario	Models	Selected method	Selected model	LR-
1	Logistic regression: 1, LASSO: 1	LR	1	LR- 0 [95% CI 0-NaN]
2	Logistic regression: 1, LASSO: 1	LR	1	
3	Logistic regression: 1, LASSO: 3	LASSO	3	
4	Logistic regression: 1, LASSO: 1	LR	1	
5	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
6	Logistic regression: 1, LASSO: 1	LR	1	
7	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	

Scenario	Models	Selected method	Selected model	LR-
8	Logistic regression: 1, LASSO: 1	LR	1	
9	Logistic regression: 1, LASSO: 3	LASSO	3	
10	Logistic regression: 1, LASSO: 1	LR	1	
11	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
12	Logistic regression: 1, LASSO: 1	LR	1	
13	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	
14	Logistic regression: 1, LASSO: 1	LR	1	
15	Logistic regression: 1, LASSO: 3	LASSO	3	
16	Logistic regression: 1, Random forest: 45, LASSO: 3	LASSO	3	
17	Logistic regression: 1, LASSO: 17	LASSO	17	
18	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	
19	Logistic regression: 1, LASSO: 3	LASSO	3	
20	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
21	Logistic regression: 1, LASSO: 1	LR	1	
22	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	
23	Logistic regression: 1, LASSO: 1	LR	1	
24	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
25	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	

Scenario	Models	Selected method	Selected model	LR-
26	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
27	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	
28	Logistic regression: 1, LASSO: 1	LR	1	
29	Logistic regression: 29, Random forest: 105, LASSO: 1	LR	29	
30	Logistic regression: 1, LASSO: 3	LASSO	3	
31	Logistic regression: 1, Random forest: 45, LASSO: 3	LASSO	3	
32	Logistic regression: 1, LASSO: 17	LASSO	17	
33	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	
34	Logistic regression: 1, LASSO: 3	LASSO	3	
35	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
36	Logistic regression: 1, LASSO: 1	LR	1	
37	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	
38	Logistic regression: 1, LASSO: 1	LR	1	
39	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
40	Logistic regression: 1, Random forest: 45, LASSO: 40	LR	1	
41	Logistic regression: 1, Random forest: 45, LASSO: 41	LASSO	41	
42	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	

Scenario	Models	Selected method	Selected model	LR-
43	Logistic regression: 1, LASSO: 1	LR	1	
44	Logistic regression: 44, Random forest: 105, LASSO: 1	LR	44	
45	Logistic regression: 1, Random forest: 45, LASSO: 3	LASSO	3	
46	Logistic regression: 1, LASSO: 17	LASSO	17	
47	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	
48	Logistic regression: 48, LASSO: 3	LR	48	
49	Logistic regression: 1, Random forest: 45, LASSO: 17	LASSO	17	
50	Logistic regression: 18, Random forest: 45, LASSO: 3	LR	18	
51	Logistic regression: 1, Random forest: 45, LASSO: 51	LASSO	51	
52	Logistic regression: 18, Random forest: 105, LASSO: 17	LASSO	17	
53	Logistic regression: 1, LASSO: 17	LASSO	17	
54	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	
55	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
56	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
57	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
58	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	

Scenario	Models	Selected method	Selected model	LR-
59	Logistic regression: 1, LASSO: 1	LR	1	
60	Logistic regression: 60, Random forest: 105, LASSO: 1	LR	60	
61	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
62	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
63	Logistic regression: 29, Random forest: 45, LASSO: 1	LR	29	
64	Logistic regression: 29, Random forest: 105, LASSO: 1	LR	29	
65	Logistic regression: 1, Random forest: 45, LASSO: 3	LASSO	3	
66	Logistic regression: 1, LASSO: 17	LASSO	17	
67	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	
68	Logistic regression: 48, LASSO: 3	LR	48	
69	Logistic regression: 1, Random forest: 45, LASSO: 69	LASSO	69	
70	Logistic regression: 18, Random forest: 45, LASSO: 3	LR	18	
71	Logistic regression: 1, Random forest: 45, LASSO: 41	LASSO	41	
72	Logistic regression: 18, Random forest: 105, LASSO: 17	LASSO	17	
73	Logistic regression: 1, LASSO: 73	LASSO	73	
74	Logistic regression: 18, Random forest: 105, LASSO: 3	LR	18	

Scenario	Models	Selected method	Selected model	LR-
75	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
76	Logistic regression: 1, Random forest: 45, LASSO: 40	LR	1	
77	Logistic regression: 1, Random forest: 45, LASSO: 41	LASSO	41	
78	Logistic regression: 1, Random forest: 105, LASSO: 1	LR	1	
79	Logistic regression: 1, LASSO: 1	LR	1	
80	Logistic regression: 44, Random forest: 105, LASSO: 1	LR	44	
81	Logistic regression: 1, Random forest: 45, LASSO: 40	LR	1	
82	Logistic regression: 1, Random forest: 45, LASSO: 41	LASSO	41	
83	Logistic regression: 44, Random forest: 45, LASSO: 41	LR	44	
84	Logistic regression: 44, Random forest: 105, LASSO: 1	LR	44	
85	Logistic regression: 1, Random forest: 45, LASSO: 17	LASSO	17	
86	Logistic regression: 18, Random forest: 45, LASSO: 3	LR	18	
87	Logistic regression: 48, Random forest: 45, LASSO: 51	LR	48	
88	Logistic regression: 18, Random forest: 105, LASSO: 17	LASSO	17	
89	Logistic regression: 48, LASSO: 17	LR	48	

Scenario	Models	Selected method	Selected model	LR-
90	Logistic regression: 90, Random forest: 105, LASSO: 3	LR	90	
91	Logistic regression: 18, Random forest: 45, LASSO: 17	LASSO	17	
92	Logistic regression: 1, Random forest: 45, LASSO: 17	LASSO	17	
93	Logistic regression: 18, Random forest: 45, LASSO: 51	LR	18	
94	Logistic regression: 18, Random forest: 105, LASSO: 17	LASSO	17	
95	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
96	Logistic regression: 1, Random forest: 45, LASSO: 1	LR	1	
97	Logistic regression: 60, Random forest: 45, LASSO: 1	LR	60	
98	Logistic regression: 60, Random forest: 105, LASSO: 1	LR	60	
99	Logistic regression: 29, Random forest: 45, LASSO: 1	LR	29	
100	Logistic regression: 1, Random forest: 45, LASSO: 100	LASSO	100	
101	Logistic regression: 18, Random forest: 45, LASSO: 3	LR	18	
102	Logistic regression: 48, Random forest: 45, LASSO: 41	LR	48	
103	Logistic regression: 18, Random forest: 105, LASSO: 17	LASSO	17	
104	Logistic regression: 48, LASSO: 73	LR	48	

Scenario	Models	Selected method	Selected model	LR-
105	Logistic regression: 90, Random forest: 105, LASSO: 3	LR	90	
106	Logistic regression: 18, Random forest: 45, LASSO: 69	LASSO	69	
107	Logistic regression: 1, Random forest: 45, LASSO: 73	LASSO	73	
108	Logistic regression: 18, Random forest: 45, LASSO: 41	LR	18	
109	Logistic regression: 18, Random forest: 105, LASSO: 73	LASSO	73	
110	Logistic regression: 1, Random forest: 45, LASSO: 40	LR	1	
111	Logistic regression: 1, Random forest: 45, LASSO: 41	LASSO	41	
112	Logistic regression: 44, Random forest: 45, LASSO: 41	LR	44	
113	Logistic regression: 44, Random forest: 105, LASSO: 1	LR	44	
114	Logistic regression: 44, Random forest: 45, LASSO: 41	LR	44	
115	Logistic regression: 18, Random forest: 45, LASSO: 17	LASSO	17	
116	Logistic regression: 48, Random forest: 45, LASSO: 17	LR	48	
117	Logistic regression: 90, Random forest: 45, LASSO: 51	LR	90	
118	Logistic regression: 90, Random forest: 105, LASSO: 17	LR	90	
119	Logistic regression: 18, Random forest: 45, LASSO: 17	LASSO	17	

Scenario	Models	Selected method	Selected model	LR-
120	Logistic regression: 60, Random forest: 45, LASSO: 1	LR	60	
121	Logistic regression: 18, Random forest: 45, LASSO: 100	LASSO	100	
122	Logistic regression: 48, Random forest: 45, LASSO: 122	LASSO	122	
123	Logistic regression: 90, Random forest: 45, LASSO: 41	LR	90	
124	Logistic regression: 90, Random forest: 105, LASSO: 73	LR	90	
125	Logistic regression: 18, Random forest: 45, LASSO: 73	LASSO	73	
126	Logistic regression: 44, Random forest: 45, LASSO: 41	LR	44	
127	Logistic regression: 90, Random forest: 45, LASSO: 17	LR	90	
128	Logistic regression: 90, Random forest: 45, LASSO: 122	LASSO	122	

4.4.3 PanPIERS app and usability test

10 individuals with a medical background and experience in the clinical management of preeclampsia participated in the study, covering a range of ages and years of experience (Figure 4.6) Majority of the participants fell in the 25-34 age group and graduated from medical or midwifery school within the last 10 years. Eighty percent of the participants specialised in obstetric medicine, and majority of them worked in a referral hospital (Figure 4.5). Fourty percent of the participants did not use any tool for risk prediction in their clinical practice, with an equal amount using one of the PIERS models, as seen in Figure 4.7. The remaining 20% while did not use a predictive model of risk, they classified patients based on gestosis score (168).

When presented with options for using predictive risk assessment tools for ruling in (focusing on accurately predicting patients into the high and very high risk groups, even if it decreases the accuracy of the low and very low risk groups) and ruling out (focusing on accurately predicting patients into the low and very low risk groups, even if it decreases the accuracy of the high and very high risk groups) the outcome, participants were more likely to be interested in ruling in the outcome or looking at both options for patients they were concerned about or patients they were unsure of (Figure 4.7). Participants were the least likely to want to use a risk assessment tool for patients they were not concerned about, and of those who still considered using one, similar numbers would use it to rule in or rule out.

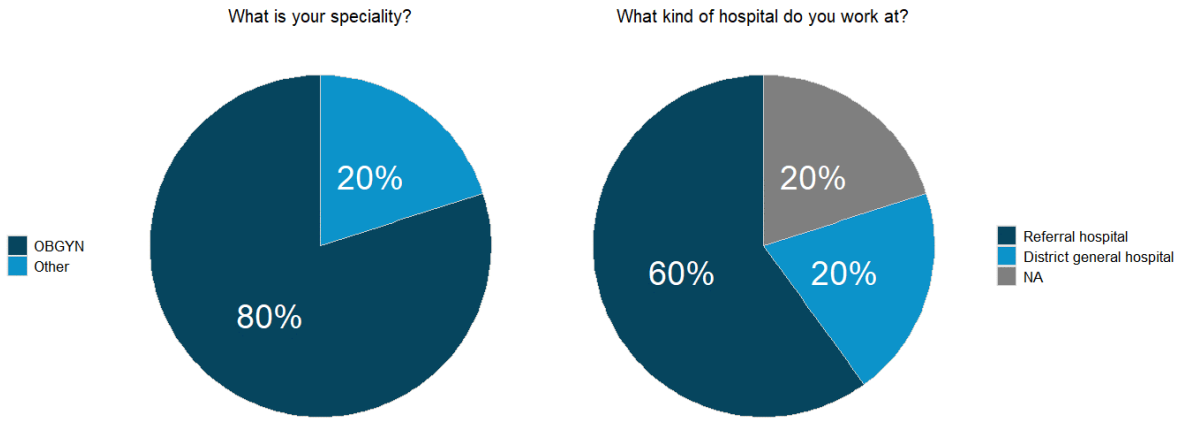

Figure 4.5: Speciality and workplace type

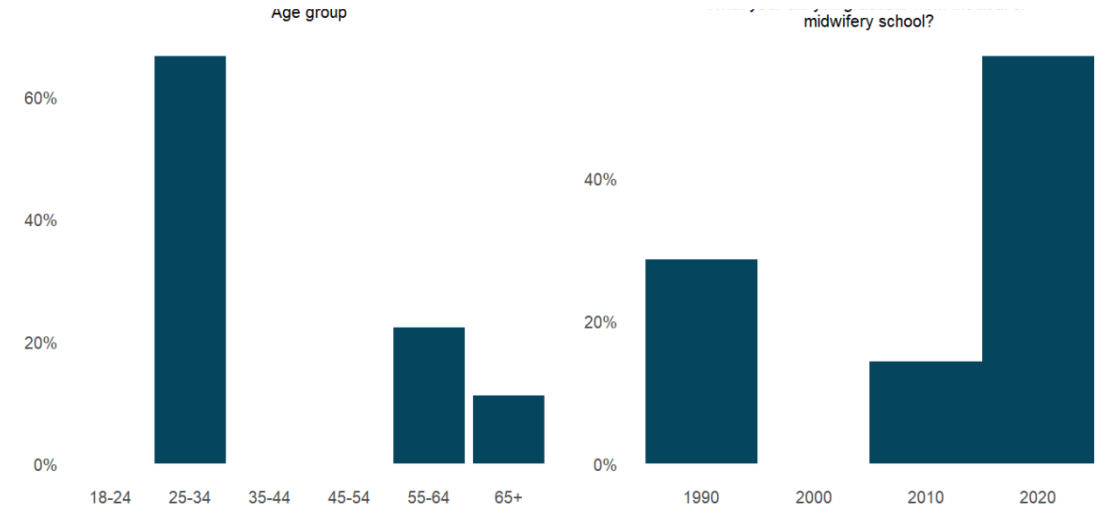

Figure 4.6: Participant age groups and graduation years

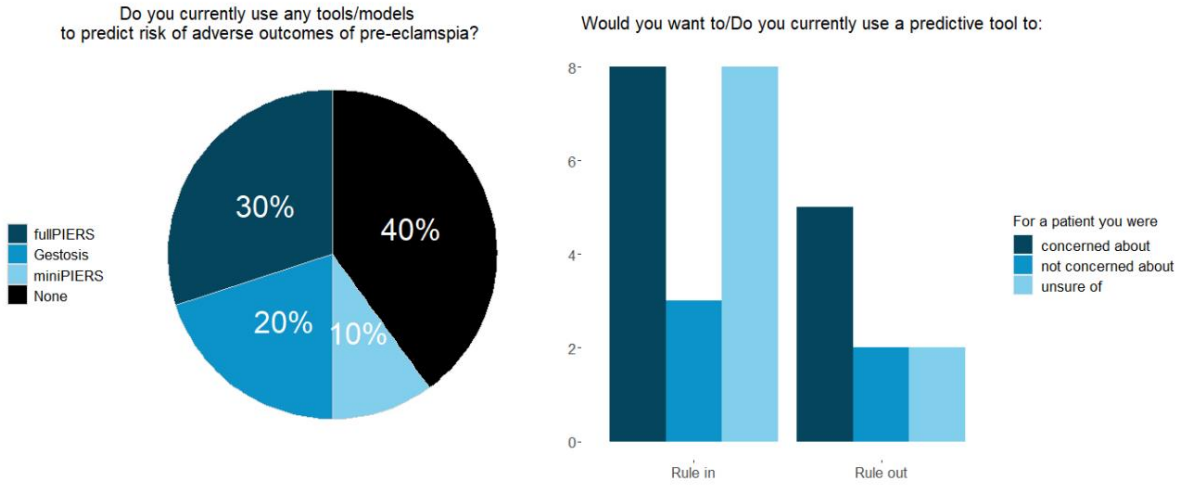

Figure 4.7: Predictive tools and ruling in/out.

Following these questions, the data input section was showed to the participants first, starting with the information box which would be the first thing an user of the app would see upon opening to use it. The information box (Figure 4.8) describes the purpose of the tool, including who it is meant to be used by and for, explanations of ruling in and ruling out and instructions on how to use it. Participants were asked to read it first before moving on to further questions. Only one participant found any of the information in this section confusing or unclear (Figure 4.10), and this participant suggested slight re-wording of the paragraph discussing ruling in and ruling out to make the sentence easier to read.

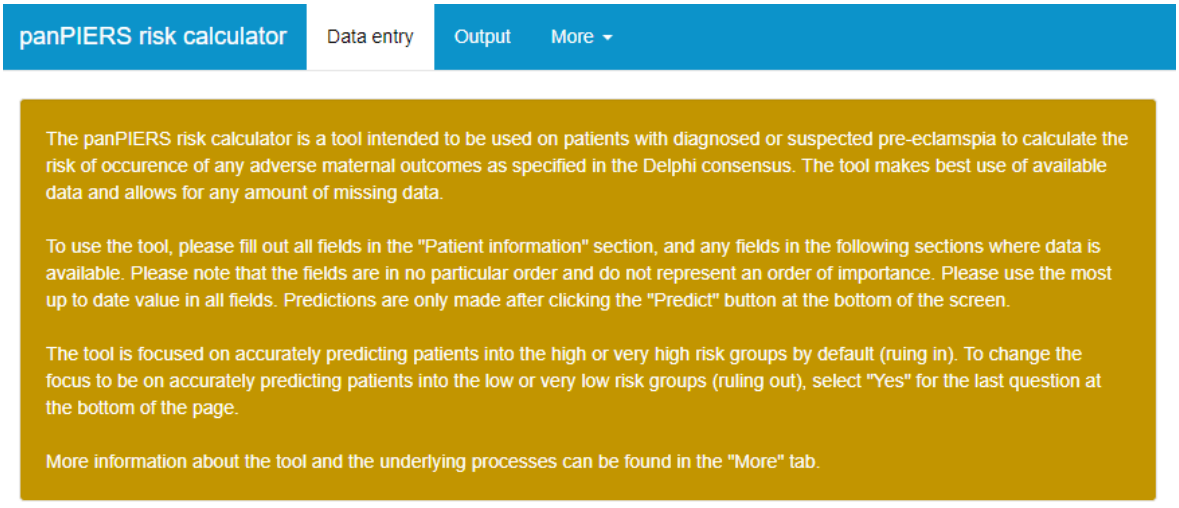

Figure 4.8: App information box

Following this, the participants were given time to look through the rest of the data entry section, where data fields were grouped into panels by variable groups. The Patient information panel (Figure 4.9) was highlighted to be required in the information section, and also indicated by a different background colour of the panel than other, optional panels (Figure 4.11and by * symbols beside the field names. Despite these indicators, only 20% of the participants could correctly identify all of these fields as mandatory (Figure 4.10).

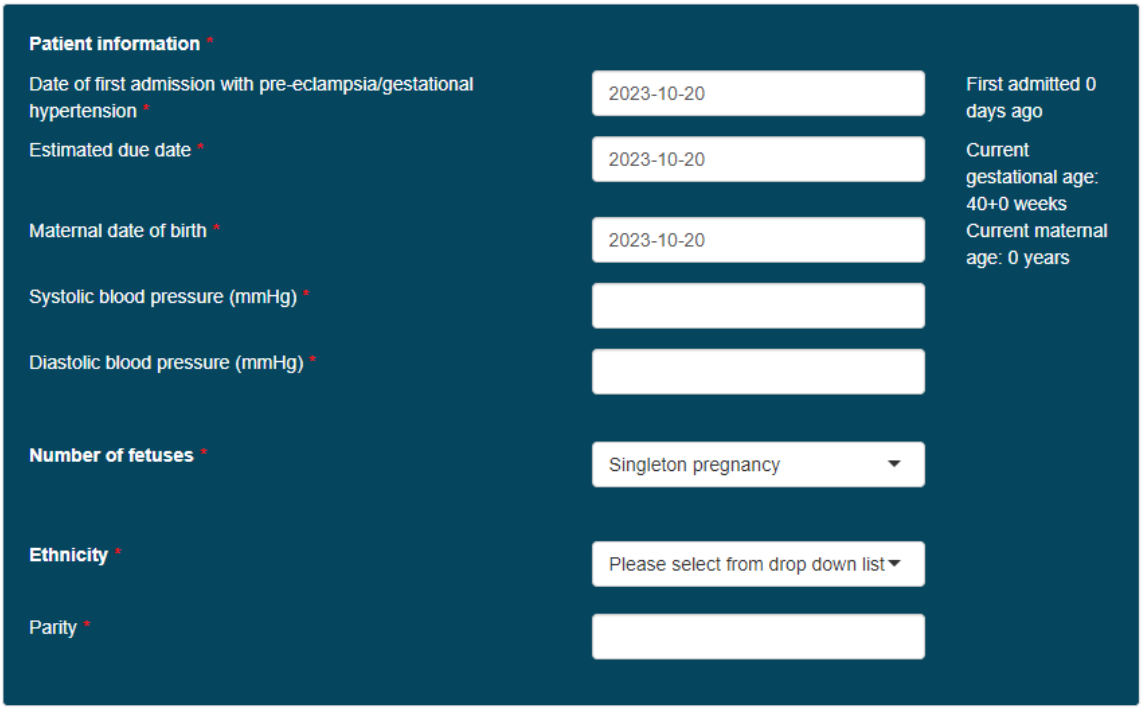

Figure 4.9: App required information section

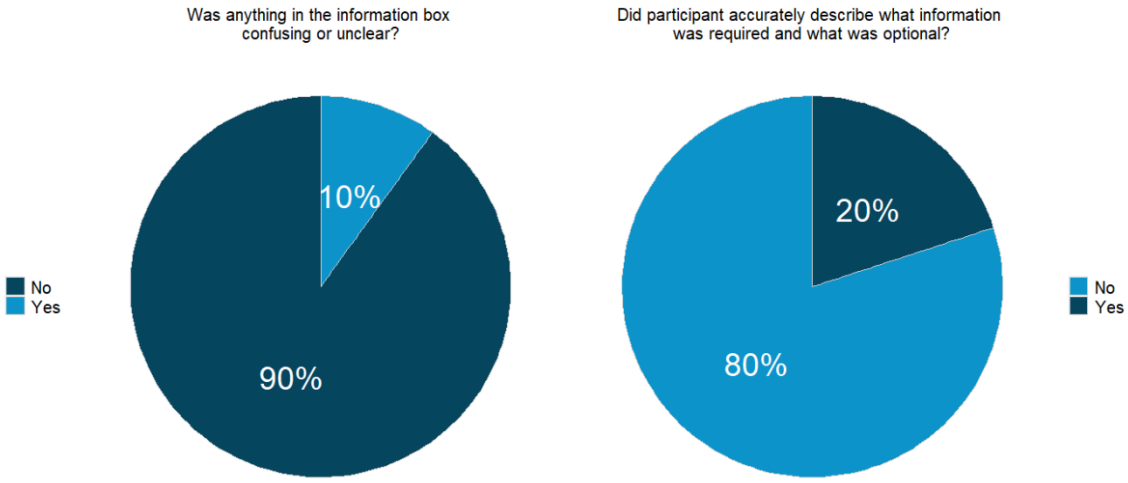

Figure 4.10: Understanding of app information structure

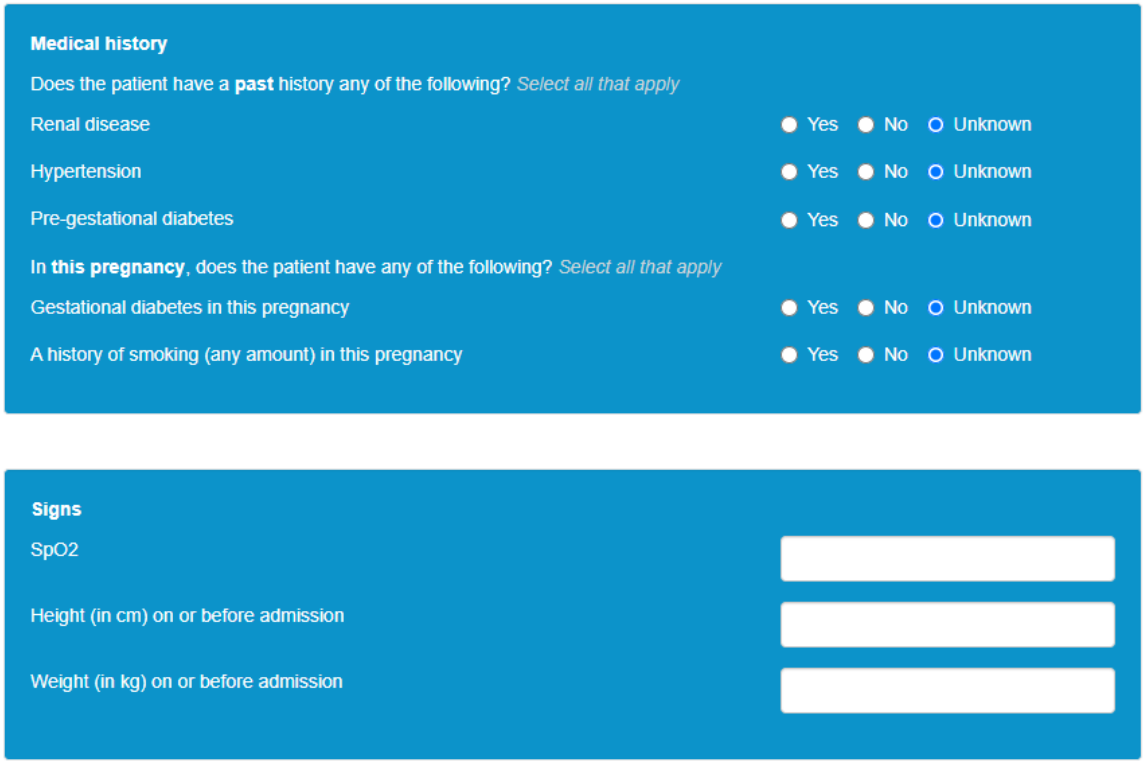

Figure 4.11: App example of additional information sections

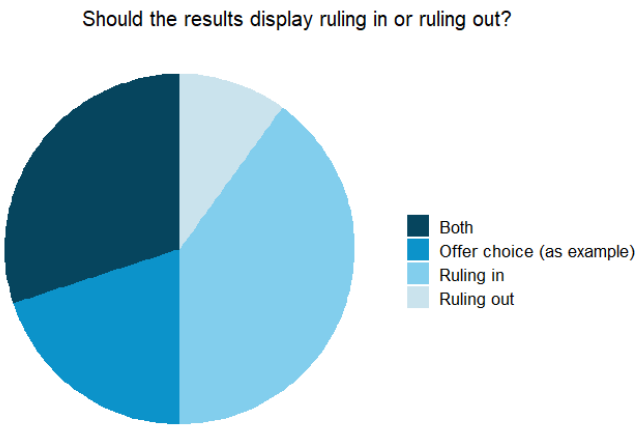

Figure 4.12: Ruling in/out preference

The example app showed to the participants offered a choice between focusing on ruling in or ruling out, however it was explained that it would be possible to show both outputs as a default, or only have access to the ruling in or the ruling out output. Half of the participants preferred to either have a choice between the outputs or be shown both outputs as default, while the majority of the remaining half was only interested in seeing the output focusing on ruling in the outcome.

Participants were asked to rate the data entry page on a scale of 1 to 5, where 1 was confusing and difficult to use, and 5 was clear and straightforward. The lowest response given was 4 out of 5, the highest was 5 out of 5, and both the mean and median were 4.5.

Participants made the following suggestions to improve this part of the app:

- Have the option to provide age as years instead of date of birth (in LMIC women might not know their date of birth, just approximate age),
- Have the ability to select units of measurement for lab results,
- Preference in wording: use estimated date of delivery, instead of estimated due date,
- Suggestions of additional predictors: medication, autoimmune diseases, type of conception (IVF), knee jerk, facial puffiness, weight gain, reduced fetal movements, urinary output, gravidity, family history on both parents' side.

We will update the data entry section to reflect suggestions 1-3. Adding the suggested additional predictors is not feasible at the moment, as we did not have data on them and therefore we are unable to test their relationship with the predicted outcome. Gravidity was considered as a predictor, however was excluded from the modelling due to unreliability, as early miscarriages might not be known to either parent, and elective termination would be known to the mother, but might be hidden from partners or other family members, and therefore not reported to the clinician. We have also limited our use of symptoms, as they are not objectively measured, predictors related to reflexes are not reliably reproduced, and changes in the physical appearance of the patient are more or less likely to be noticed by a clinician depending on their familiarity with the patient.

Most of the new predictors suggested by the clinicians are generally considered as predictors of preeclampsia itself (169, 170), not of its outcomes, however clinicians may be interested in whether these predictors of disease are also predictors of outcome. While there was no capacity to address this question in this study, it could be part of potential future work, similarly to fetal variables, which have only been proven to be indicative of fetal outcomes, rather than maternal.

Finally, medication, especially antihypertensive medication, is notoriously difficult to include in modelling maternal outcomes of preeclampsia, as the dose is regularly adjusted for each patient to achieve a goal, which varies between medical professionals and locations. For future work, we would suggest considering using number of medications, and the percentage of the maximum prescribable dose as predictors on medications.

As the next part of the interview process, participants were shown 1 or 2 scenarios, depending on time constraints. For each scenario, participants would see the data entry section filled in by a hypothetical patient's data, with varying amount of missing information, after which the app output was loaded, and the participants were asked to provide their interpretation of each element of the output.

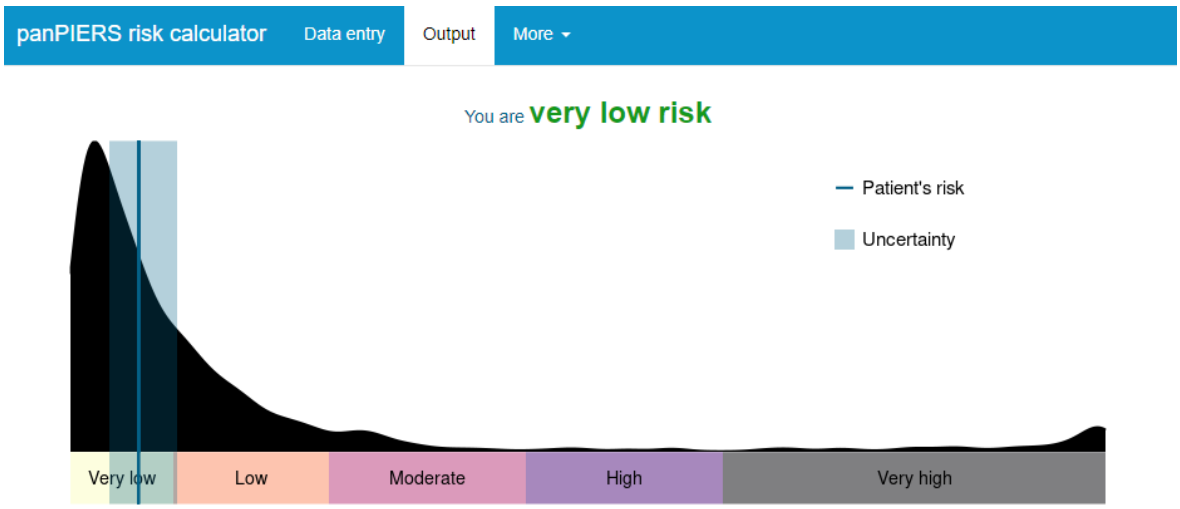

Figure 4.13: App output 1 example

Table 4.5: Summary of interpretations of first app output

Scenario	Risk	Uncertainty	Seen by	Answers
1	Very low	Yes	5	Participants all concluded that patient is at low or very low risk, and unlikely to have an outcome, no comments on uncertainty or distribution of other patients
2	Very high	No	3	Two participants concluded the woman is at very high risk (one specifically stating because her blood counts showed features of severity), and one interpreted the output as showing a high risk case and disagreed with the prediction (note that participant works with different units of measurement and did not know the reference values for the blood tests – an issue which could be resolved by allowing users to change units) No comments on distribution of other patients
3	Very low	Yes	2	Both participants interpreted as the patient being in the very low risk (noting normal

				blood tests or no symptoms as the reason), one patient commented on the uncertainty but concluded that the patient will be safe as the uncertainty is still within very low risk No comments on distribution of other patients
4	Very low	No	2	Both identified patient as very low risk and highly unlikely to have an outcome. No comment on distribution of other patients

The first output of the app (Figure 4.13) was the patient's predicted risk group in text form and visualised along with the distribution of predicted probabilities of other patients. Table 4.5 shows the summary of participants' interpretation of this output. Overall, participants interpreted the patient's risk group correctly and easily, however the uncertainty was not noted by many, and no-one discussed the distribution in the visualisation. Therefore, we do not plan to make any large changes to this output, however we will include more text to help with interpretation of the visuals (such as "uncertainty includes low risk", and explanation and context of the distribution of predicted probabilities)

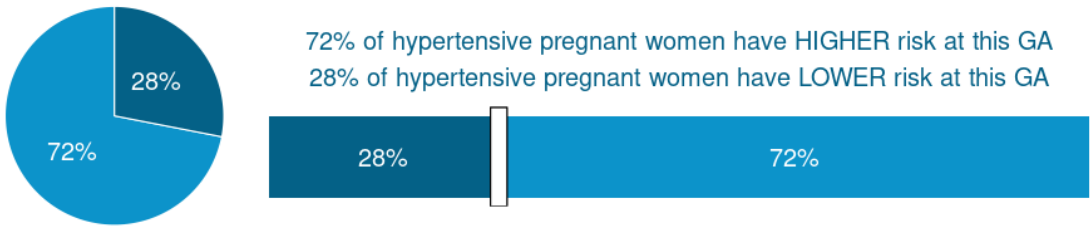

Figure 4.14: App output 2 example

Table 4.6: Summary of interpretations of second app output

Scenario	Risk	% at higher risk	Seen by	Answers
1	Very low	90%	5	Majority of participants understood that at this gestational age, more women were at a higher risk, however it was not a quick interpretation, and while the general understanding was correct, some also made further assumptions (such as all 90% were at increased risk, or the 10% could also be at a low risk)
2	Very high	2%	3	Two did not understand without further explanation, one participant concluded that the patient is very unusual and at the extreme end
3	Very low	72%	2	Neither understood without further explanation
4	Very low	84%	2	One participant interpreted the output incorrectly, concluding that patient is in the minority at this gestational age who are in the lower risk categories, while the other patient interpreted the output correctly

The second output, as seen in Figure 4.14, was a comparison of the current patient's risk with the risk of other patients at the same gestational age. More explanation is needed for this output, as it was not well understood by the participants (Table 4.6). More detailed wording would be useful, for example the example could be re-worded as:

"Comparing this patient's risk to other patients we have data on:
72% of hypertensive pregnant women who were at this gestational age had a HIGHER predicted risk than this patient, and
28% had a LOWER predicted risk than this patient."

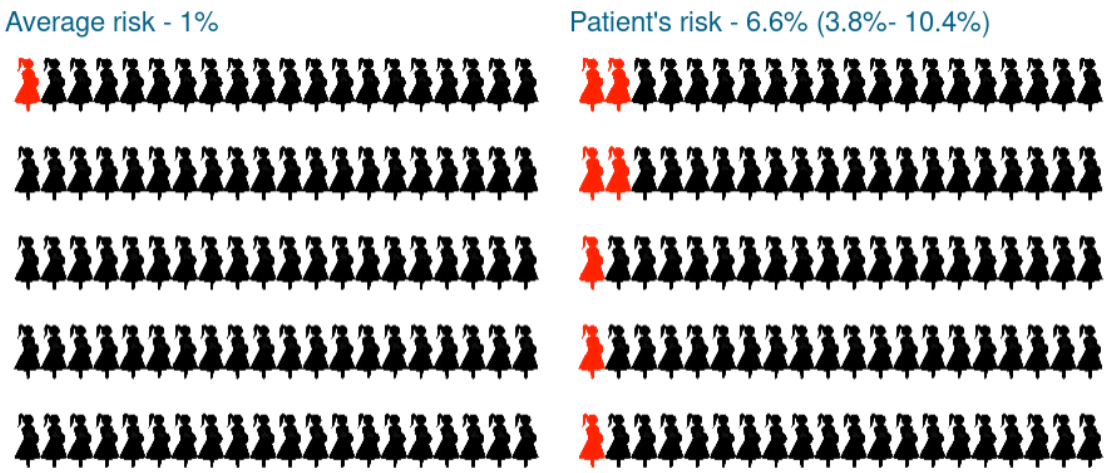

Figure 4.15: App output 3 example

Moreover, this output may only be useful at certain gestational ages, or in certain resource setting, and hence it could be an optional output within the app, which is not displayed by default, but could be toggled in a section titled "Comparison with other patients at the same gestational age".

The final visual output of the app (Figure 4.15) was the patient's risk with uncertainty where applicable, compared with the overall average risk. This output was also overall clearly and easily understood by the participants, as their responses in Table 4.7 suggests, with only a few minor modification suggestions. We will take these into account by changing the comparison to be with the average (median, indicated in small print) risk of patients at this gestational age, while also reporting the overall average risk for any gestational age. We will consider using a different shade of red on the two sides of the output to make the separation clear, and we will include a figure caption of "visual representation of average risk" and "visual representation of patient risk" for clarity.

Table 4.7: Summary of interpretations of third app output

Scenario	Patient's risk	Seen by	Answers
1	2.6% (2.4%-5.4%)	5	Correctly and easily interpreted by all as the calculated risk for the patient vs the average risk, and noted that the patient's risk is higher than the average
2	82% (no confidence interval)	3	Correctly interpreted by all, easily by two participants, while the other noted that it would be clearer if the two sides used different colour or shades
3	6.6% (3.8%-10.4%)	2	Correctly and easily interpreted by both, one participant asked for clarification if this is also comparison of patients at the same gestational age
4	4.6% (no confidence interval)	2	Not clear at first glance for one participant, but understood after some time, understood clearly and easily by the other

The last part of the output of the app was a section of text containing information on missing variables and the uncertainty caused by them, as well as advice for the clinician of what variable or variable group could be obtained to see a more accurate prediction for the patient. An example of this is shown in Figure 4.16. Participant interpretations of this output (Table 4.8) were mixed, while the majority found it clear and easy to understand, a few asked for further explanation or misunderstood the displayed variable name. Additionally, one participant noted the lack of clarity on how results were arrived at. As the most commonly known and used models to predict maternal outcomes of preeclampsia are logistic regression models (3, 29, 118), and the relationship between the predictors and the resulting outcome are known to those who are using the tool, our app is indeed a lot less transparent due to the nature of machine

learning models. While we cannot provide the same kind of information as it does not exist, we would instead consider including a “More information” tab in the final product, that provides detailed description of how these machine learning models work and how they arrive at a result. As the information in the output was otherwise generally well understood, we do not plan to make any further changes to it, aside from correcting spelling mistakes.

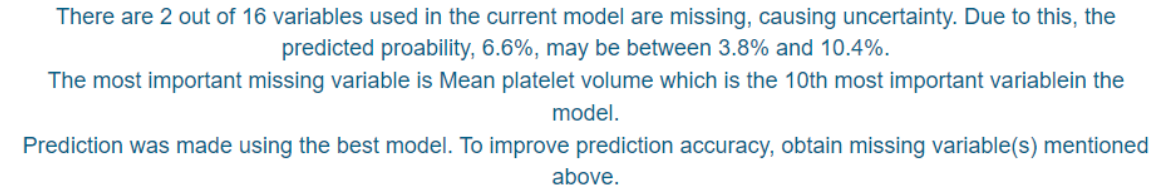

Figure 4.16: App output 4 example

Table 4.8: Summary of interpretations of fourth app output

Scenario	Missing model variables	Model rank	Seen by	Answers
1	1 missing (LDH)	7th, get Urine for better	5	One participant didn't understand without further explanation, the other four found it clear and concluded that urine sample should be taken
2	0	9th, get Signs for better	2	Participants found it clear and easy to understand
3	2 missing (most important MPV)	best	3	One participant interpreted the uncertainty caused by the missing value correctly and linked it to the first output that also displayed the uncertainty, but did not comment on the missing variable itself. One participant misinterpreted mean platelet volume as platelet count and stated it should be added as it is the most important predictor. The third participant did not fully understand, stated that 2 variables are missing but it is not shown what they are or how important they are. Interpreted risk correctly but did not understand how predicted probability was arrived at.
4	0	6th, get Bloods (part 1) for better	2	Participants found it clear and easy to understand and interpreted it correctly

For each scenario, I then assessed how the output provided by the tool contributed to the participant's decision-making process for the hypothetical patient. The patients in the scenarios were largely at a very low risk, with only 4 cases of very high risk patients. When asked if their level of concern for the specific hypothetical patient would change after seeing the app output (Figure 4.17), all participants assessing a very high risk

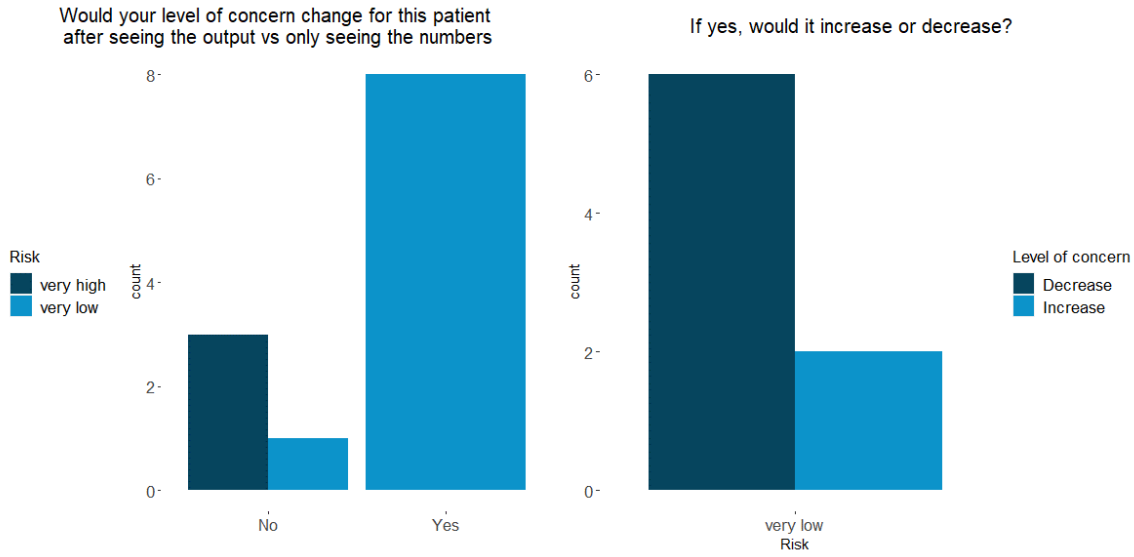

Figure 4.17: Change in level of concern

patient stated it would not, while all but one participant assessing a very low risk patient stated that their level of concern would change. Most participants' level of concern decreased for their patient after seeing them classed into the very low risk group, while two participants' level of concern increased despite their patients being classed into the very low risk group, from seeing the comparison of risk with other patients at the same gestational age.

Participants generally stated they would show all or parts of the output to the patient, with each visual component of the output chosen roughly equally to show to the patient, displayed in Figure 4.18. Only one participant stated they would not show any of the output to the patient, but still explain their risk level as reason for their clinical recommendation, and one participant, seeing a patient with a large number of missing predictors, chose to obtain more information first, then use the app again, and only show the output to the patient once more predictors were available.

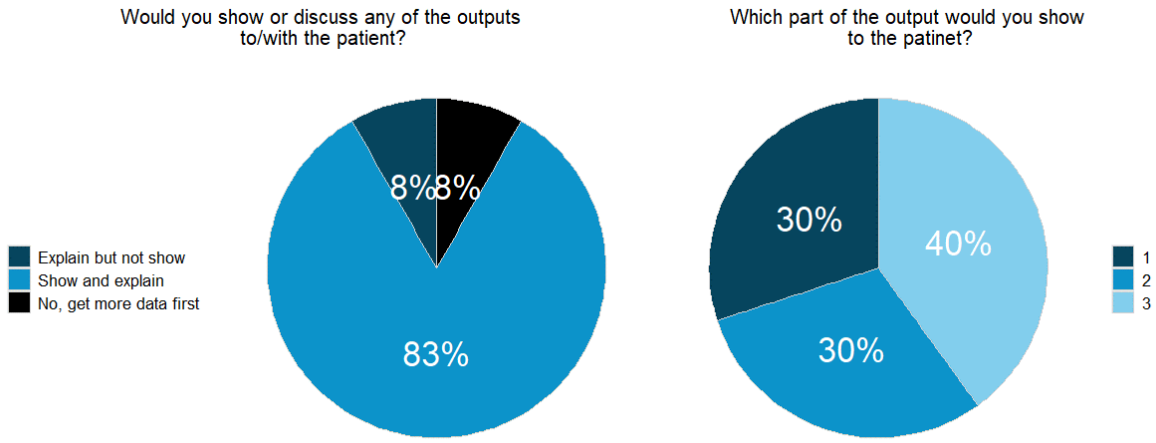

Figure 4.18: Information sharing with patients.

For each scenario, participants were also asked what they would do next for their hypothetical patient. A wordcloud of these answers can be seen in Figure 4.20 Admission to hospital was a common answer for both very low risk and very high risk patients, while for very high risk patients induction of labour was a frequent answer, and participants suggested to counsel, and reassure very low risk patients, and obtain blood test results if they were missing.

After questions related to the scenarios, I also asked participants questions about their experience with the app as a whole and any general feedback. All participants stated they would use this app once it became available (Figure 4.19), one participant choosing to replace the existing tool they were using with it, while the other participants opting to use it in addition to what they are already using, or not currently using a risk prediction tool in their clinical practice. The preference was to use the app in either a mobile app form or integrated into the healthcare system, only a few participants being

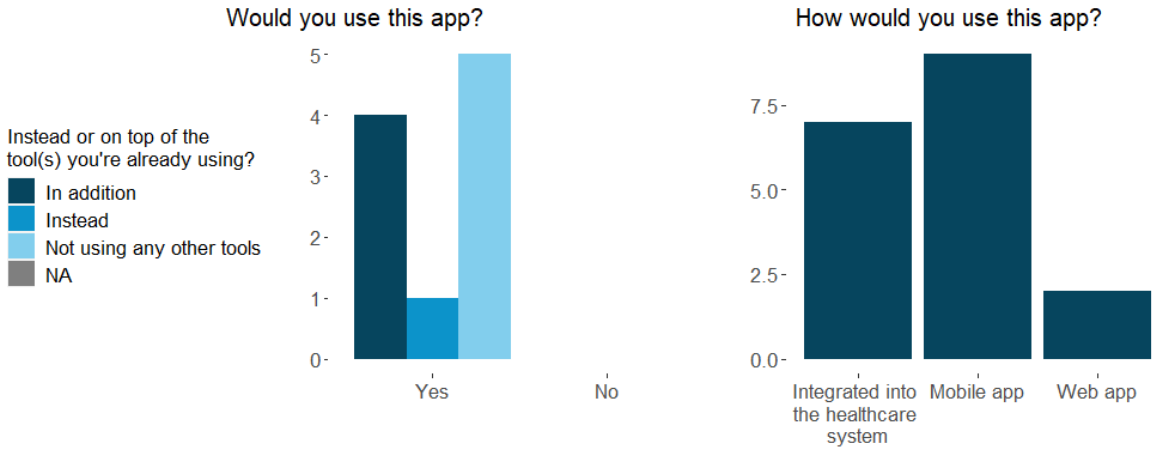

Figure 4.19: App feedback questions

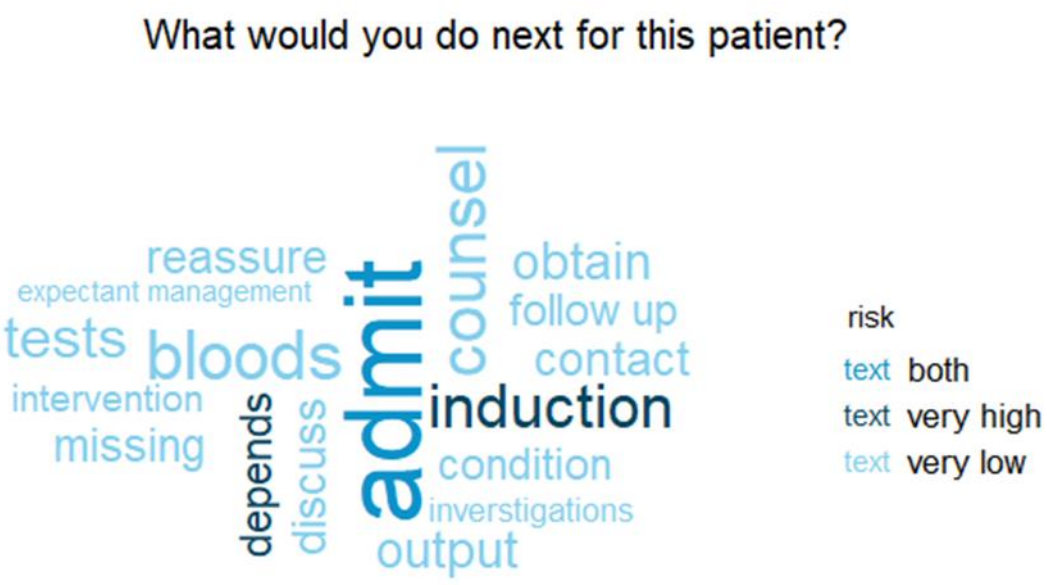

Figure 4.20: Next action wordcloud

willing to use it in website form mainly due to access issues such as not always having a computer at hand when seeing patients with preeclampsia, or having restrictions on what could be accessed on work computers.

Participants thought the tool would benefit the patients by helping them understand their risk better, and having visual outputs they could take away and show to their families to help communicate the information they received in usually a high stress situation. Participants also mentioned that it would help clinicians make more informed decisions and sensitise them to the patient's condition, which would in turn reduce morbidity, and benefit the patients. Participants expected that the use of this tool would change how they understood and communicated risk, while some noted that their decisions would still depend on gestational age, and that the tool may be more impactful for those with less experience treating patients with preeclampsia. When asked about potential different impact at different gestational ages, participants stated that they would likely use the tool regardless of gestation, but it would have an impact on the decisions made, as there were different implications of the same action at different gestational ages, and the threshold for acting on risk would be higher earlier in the pregnancy.

40% of participants expected the use of the tool to increase money and resources spent on treating patients with preeclampsia in a low resource setting, as they expected that resources would need to be used to provide access to the tool and maintain equipment. Another 40% expected that the use of the tool would decrease the money and resources spent on patients with preeclampsia, as it would help manage resources better, prevent outcomes and help identify patients who did not need to be admitted. The remaining 20% either could not make an estimate, or expected no change in the resources, for a combination of the reasons stated above.

Participants also made the following suggestions to further improve the tool:

- Have patient profiles within the app to store patient data,

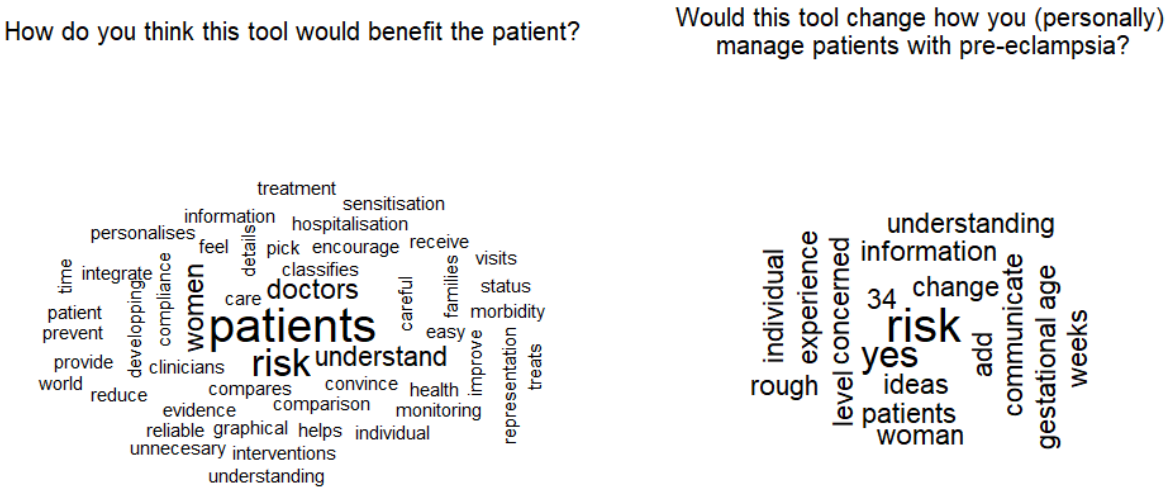

Figure 4.21: Benefits and clinical practice

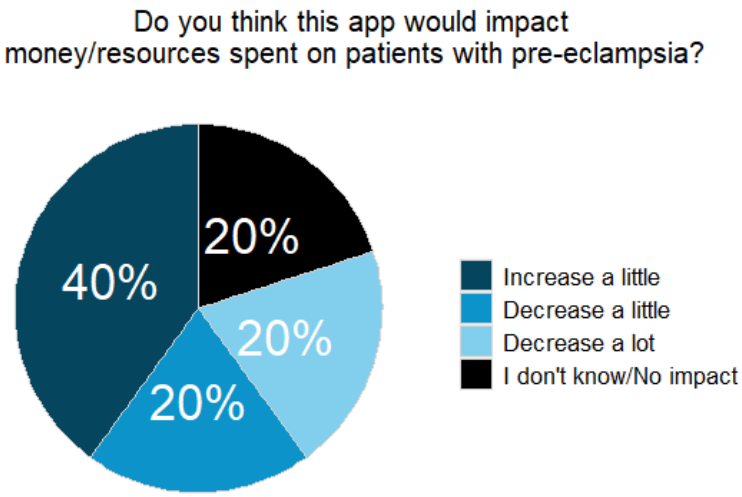

Figure 4.22: Impact of app on resources

- Consider including biomarkers as predictors, and
- Consider an option for the clinician to provide what they think the patient's risk group is going to be before the prediction is made, mainly as a feedback loop to use as a training tool for clinicians with less experience managing patients with pre-eclampsia.

There is currently ongoing work to test the predictive ability of biomarkers and include them in the modelling process if results suggest that it would be beneficial. We will also include the option to provide the expected output by the user, so that the tool would not only aid in clinical decision making, but also function as a useful learning tool. While it would be ideal to include patient profiles in the app and be able to store patient data, so that new predictions could be easily made once new information is available, we are aware that patient data is sensitive information, and storing this sensitive data with sufficient security and according to guidelines that may differ by region would be a challenge.

4.5 Discussion

4.5.1 Summary of findings

This study analysed data from 11472 patients, 327 (2.9%) of whom had an adverse outcome within 2 days, and 819 (7.1%) at any point. Dividing the data into base variables and seven variable groups, all variable groups had some amount of missingness, and all but the medical history variable group had cases where all variables within the group were missing for a patient. The second part of blood tests, the coagulation variables, was the most frequently completely missing variable group, with over half of the patients lacking all variables in this group. Medical history was the least frequently completely missing, with at least one variable observed for all patients, and patient information was the least frequently partially missing.

Although blood tests were found to be the most indicative of the adverse maternal outcome in the analysis in previous chapters, they were most often partially or completely missing in our dataset. Therefore, the development of tools, such as the one presented in this chapter, that can generate predictions of risk of adverse outcome on data with missing values is likely to be particularly valuable in practical clinical settings. In this chapter, logistic regression, LASSO, and random forest modelling methods were

employed, along with multiple imputations during model fitting and GANs during model testing, to address missing data. Risk classification groups were created similarly to previous chapters and for the same aims, focusing on the ruling out with the very-low and ruling in with the very-high risk strata in this case to reassure or justify treatment of patients, respectively.

Logistic regression models performed well in ruling in the outcome using most combinations of variable groups, creating a large list of possible models, however a small selection was sufficient to rule in outcome for all combinations. Surprisingly, even models with only urine dipstick or base variables achieved high predictive performance, few of which were included in the final selection. Logistic regression performed better for ruling out the outcome as it did for ruling in, using a slightly smaller set of models and successfully ruling out the outcome for all scenarios.

Random forest models showed poor performance in ruling out the outcome, only being able to determine probability thresholds for the very-low risk strata for 11 out of 128 models. Although two of these models could make predictions on 75% of the variable group combinations, we saw LR-<0.1 in only 25 scenarios. We suspect this would be due to the fact that the variables picked up as the most significant predictors of the outcome by the random forest in the previous chapters, namely the blood test results, were not available for the majority of variable group combinations. Ruling in the outcome with random forest was a lot more successful only four models could cover all 128 variable group combinations, and all but 12 scenarios achieved LR+>10.

LASSO models demonstrated the worst performance in terms of ruling in the outcome out of all three methods, it could determine risk thresholds for only 39% of the models, and despite having a similar number of models chosen to cover all 128 scenarios, only 30% of these scenarios achieved LR+>10. Its ruling out performance was similar to that of logistic regression, while more LASSO models were needed to rule out the outcome in two days for all scenarios, the LR-<0.1 goal was achieved in all cases.

A combined system that integrated all modelling methods and selected the best performing model, regardless of the method used, yielded excellent results. This system was able to rule in and out the outcome for all variable groups. When ruling in the outcome, nearly 90% of scenarios achieved the desired LR+>10, and the remaining

scenarios had a positive likelihood ratio near 9. Additionally, it successfully ruled out the outcome for all variable group combinations, resulting in no outcomes in the very-low risk strata.

As a result of the modelling methods working well on all variable group combinations, the resulting tool can allow for missing values in the provided data for prediction. Moreover, any uncertainty caused by the potential missing values of the variables used in the selected model is communicated to the clinician, along with information on the most important missing value. The development of such a risk prediction too carries real-life implications. By providing clinicians with actionable insights, such as predicted risk, uncertainty and most important missing variable, the tool can support decision-making and enhance patient care.

Furthermore, the tool also facilitates comparisons between patients at the same gestational age, enabling clinicians to contextualize individual patient risk within a broader reference group. This feature enhances risk communication and aids in shared decision-making between clinicians and patients. Clinicians can utilize this information to explain the patient's risk profile, discuss potential interventions or treatment options, and provide more personalized care.

4.5.2 Strengths and limitations

The risk prediction tool developed in this chapter exhibits several strengths. One of it's key strengths lies in its flexibility to handle many different possibilities of missing data effectively. This is particularly important because blood tests, which are highly predictive of outcomes, are frequently missing in the dataset. This flexibility enhances the practical applicability of the tool in real-life clinical scenarios. This novel approach is also strengthened by employing innovative techniques, such as Generative Adversarial Imputation Networks (GAINs), which helps the tool to generate accurate predictions while often being less computationally expensive than multiple imputation.

Furthermore, the risk prediction tool demonstrates good performance in ruling in and ruling out the risk of adverse outcomes. Leveraging multiple modelling methods and selecting the most appropriate model based on the available data, the tool takes advantage of different strengths of different modelling methods to achieve reliable

predictions. This comprehensive approach ensures that the tool can effectively handle different data scenarios and provides clinicians with dependable risk assessments.

Another strength of the risk prediction tool is its interpretability. While the actual insights of each model used are not displayed, the tool offers features that allow clinicians to gain insights into the contributing variables that drive the risk predictions. This interpretability enhances the understanding of the underlying mechanisms and helps clinicians make informed decisions based on the provided risk assessment.

Importantly, the risk prediction tool goes beyond providing predictions and incorporates mechanisms to communicate the impact of missing values to clinicians both verbally and visually. By specifying the missing variable that contributes the most to the uncertainty, the tool highlights the importance of gathering complete data for accurate risk assessment. This information empowers clinicians to prioritize the collection of missing variables, thereby improving the overall accuracy and reliability of the risk prediction.

Moreover, the tool has been developed and internally tested on a diverse population of multiple healthcare settings and thus shows promise to be widely generalisable, although this should be confirmed by further research in more-, and less-developed areas.

On the other hand, there are several limitations that should be acknowledged. Firstly, external validation of the risk prediction tool may be necessary to assess its performance on independent datasets. While the tool showed to be promising both in terms of generalisability and reliability on internal validation, any concerns regarding these performances can only truly be answered by external validation. Moreover, while the data used for internal validation was held out from the model development process and thus was unseen data for the models, the data was included in the testing of the selected predicted probability thresholds and was used to create the ranking of models. This leaves a possibility for the final performance summaries to be overly optimistic, which should be tested on external data.

Additionally, GAINs typically require large training datasets for optimal performance. While this study utilised the largest available combined dataset of women with preeclampsia, the sample size compared to that of image analysis datasets, where GAINs are most often used, is still small. Sampling with replacement was employed during

training of each GAIN, however, the impact of this limitation on the accuracy of imputations should be considered.

Finally, I only utilised data from the day of or the day after admission during the development and testing of this tool. Clinicians may wish to use such a model for longer periods during a patient's pregnancy, such as up to 14 days after admission, which is not accounted for in the current approach.

While the tool performed well on internal validation, awareness of these limitations is crucial when implementing the tool in clinical practice.

4.5.3 Usability test

By and large, the feedback I received from clinicians with experience managing patients with preeclampsia was positive, already showing interest in the current version of the app. Taking on board the constructive feedback provided by the participants of the study from a range of different backgrounds, I will be able to create a version of the tool that is even more beneficial to the clinicians and the patients than the responses already suggest this version would be.

4.5.4 Interpretation in light of existing literature

While national and international guidelines are in place that recommend the use of predictive models for preeclampsia patients, namely recommending the PIERS and PREP models (19, 21), the guidelines do not offer suggestions on what to do if the required variables for these models are not available. The models recommended in guidelines are regression-based models, which are not able to make predictions on data with missing values.

To our knowledge, of all the available models, only the model created by Schmidt et al. (33) is capable of making predictions on data with missing values, as the missing data can be easily replaced with the single value they used to indicate missingness. This method, however, does not provide any information on the impact the missing values have on the predicted probability, and could lead to predictions being made with too much uncertainty without the user's knowledge.

The panPIERS tool improves on both the models currently recommended in clinical guidelines, and the Schmidt et al. model, by not only developing a system that ensures

that predictions are made with a model most suitable to the available data, rather than attempting to make a single model work on a large variety of potentially available data, but also quantifying the loss in prediction accuracy due to unavailable predictors and presenting it all in an easy to understand format. The tool shows great potential for clinical utility, as it showed excellent performance on a 16.6% subset of our data, and received helpful, positive feedback on its design from clinicians in both high-, and low- and middle-income countries.

There are two important lines of investigation that follow from our work.

First, while all models were tested internally on a dataset held out from the development step, the tool as a whole including the model selection step has not been validated yet. Therefore, the most important line of further investigation would be to externally validate the tool in a diverse dataset, ideally with a larger variety of observed scenarios. This would be particularly important as our testing dataset contained only two different observed scenarios (scenarios 121 and 128), and all other scenarios were generated by removing variables.

Second, further app development needs to be undertaken to implement the suggestions from the panPIERS usability study, and develop a secure platform that complies with regulations.

4.6 Conclusion

We have successfully developed a multistep hierarchical predictive tool to predict adverse maternal outcomes of preeclampsia. Overall, our tool, which can effectively rule in and rule out the risk of adverse outcomes, while accounting for missing values and providing information on uncertainty, offers tangible benefits for clinicians and patients alike. It enhances risk assessment, supports clinical decision-making, and fosters patient-centred care by improving the understanding of individual patient risk profiles and addressing the challenges posed by missing data in real-life clinical scenarios.

Chapter 5: Conclusions and further work

In this thesis I presented three new tools for the prediction of adverse maternal outcomes of preeclampsia: the PIERS-ML model, for prediction of outcome within 2 days of admission, using data from day of admission; the PIERSdynamic model, for prediction of outcome in a rolling 2 day window from longitudinal data up to 14 days after admission, and the panPIERS tool to predict outcome in 2 days of admission making best use of any available data.

From the results of the comparison of 16 models created using 6 modelling methods (random forest, LASSO, ridge regression, artificial neural network, Bayesian model averaging and gradient boosted trees), the 18-variable random forest model selected as the PIERS-ML model could rule in and rule out an outcome within 2 days. The PIERS-ML model had excellent classification ability both on the internal and the external validation and significantly improved on the models currently recommended for use in clinical practice.

The binary mixed effects random forest PIERSdynamic model extended this classification ability for 7 days after admission with excellent performance, and 14 days with good performance. This model was also selected from a comparison of modelling methods, assessing the predictive and classification abilities of four models using a different method each (two-stage model, Bayesian joint model, long-short term memory artificial neural network and the binary mixed random forest model). The model selected as the PIERSdynamic model had both the highest area under the ROC and precision-recall curves, and the best likelihood ratios when classifying patients into the five risk strata.

Additionally, the new panPIERS hierarchical classification tool addressed the challenge of missing data and resource constraints, presenting a unique contribution to preeclampsia risk stratification in diverse healthcare settings. For all 128 possible available variable group combinations I tested, the tool could provide a model to rule in or rule out the outcome from a small collection of logistic regression, random forest and LASSO models.

5.1 Research in context

Multivariable model-based risk stratification is recommended for women with preeclampsia in clinical guidelines, and several models using logistic and Cox regression,

and recently using machine learning methods have been developed to facilitate this. The most notable of these models are the PIERS models (fullPIERS, developed for use in high-income countries, and miniPIERS for use in low- and middle-income countries), and the PREP models, which are named in clinical guidelines. While there have been a number of machine learning models published recently for risk prediction of adverse outcomes of preeclampsia, most of them were developed on very small sample sizes of less than 1000 participants. A publication by Schmidt et al presented two machine learning models, which were developed on a somewhat larger dataset of 1647 participants, these models predicted a combined fetal and maternal outcome occurring at any point after admission. The models presented in this research have several advantages over all previously published models.

The PIERS-ML and PIERSdynamic models, and the panPIERS tool are the first predictive tools for the outcomes of preeclampsia that stratify women into clinically relevant risk strata rather than simply providing a predicted probability. All three tools were developed on the largest dataset used for predictive modelling in preeclampsia to date, using highly diverse data including high-, low- and middle-income countries, and including variables representative of the effectiveness of the local healthcare system to adjust the baseline risk.

The PIERS-ML and panPIERS tools can be used to inform risk of adverse maternal outcome of preeclampsia, while the PIERSdynamic model could confidently inform of risk of outcome within a rolling two-day window up to 7 days after admission, and up to 14 days with some uncertainty. Unlike models predicting outcome at any point, this output can help inform timing of birth as action to prevent an outcome can be taken in a timely manner.

The PIERS-ML model included laboratory test-based predictor variables which have been found to be highly predictive in previous models as well, and excluded more subjective variables such as maternal symptoms, which have previously received criticism for their inclusion in predictive modelling. While this could raise the concern that the PIERS-ML model would then be only useable in high-income, well-resources settings, the panPIERS tool assures that an informative predictive tool is available in any resource setting.

The PIERSdynamic model is the only model to date to use multiple observations per woman to account for the rate and magnitude of change in a woman's risk of adverse outcomes of preeclampsia over time. While the PREP study developed a survival model, which, rather than providing a predicted probability of outcome within a certain time period such as two days or any point, can predict time to outcome, the prediction is based only on one set of observations from a patient. This means that the PIERSdynamic model is the only currently existing model that is able to detect a change in a patient's condition after the patient has been admitted. The model also achieved a better AUROC value on admission than the PREP-S model, and maintained excellent performance for 7 days after admission, while it has been shown that repeated use of static predictive models, such as the fullPIERS or PIERS-ML models results in a large drop in accuracy in just 2 days after admission.

Finally, unlike any other previously published model, the panPIERS tool is able to make predictions on data with missing values, provide information on the uncertainty caused by the missing data, recommend variables to be collected next if available, compare the predicted risk to other patients at the same gestational age, and all the while show excellent performance for both ruling in and ruling out the adverse maternal outcome within two days.

5.2 Further work

There are seven important lines of investigation that follow from our work.

The first is that use of random forest for the PIERS-ML model offers the potential for the accuracy of the model to improve over time as data accumulate and the model 'learns' with regularly scheduled manual updates provided to model users. Amassing such data is feasible, as all individual-level variables in PIERS-ML are part of routine clinical and laboratory assessment of women with preeclampsia in well-resourced settings, and available from electronic health records in real-time.

Second, to explore whether addition of new markers may improve PIERS-ML model performance. As markers of uteroplacental dysfunction of preeclampsia, angiogenic markers are used increasingly in the investigation of women with suspected preeclampsia, at first presentation, for ongoing surveillance (126) (although performance in trials has been variable34,35), and within the Schmidt et al. model.(33)

If independently informative, angiogenic markers could be incorporated into the model during clinical implementation. Ophthalmic artery Doppler may provide useful information about the less-accessible intracranial circulation.(129)

Third, the PIERSdynamic model also offers the potential for the accuracy of the model to improve to extend accurate prediction to 14 days after admission, and more by obtaining data from pregnancies with early pre-term preeclampsia diagnoses, which are more likely to be expectantly managed for over a week after diagnosis. This should be feasible, as all individual-level variables in the PIERSdynamic model are part of routine clinical and laboratory assessment of women with preeclampsia in well-resourced settings, and approximately 30% of cases of preeclampsia are diagnosed away from term.

Fourth, to explore the use of different thresholds over time to improve the stratification ability of the model. This line of investigation also relies on the collection of further data, with particular focus on women with preeclampsia who were in expectant management beyond the first week after their diagnosis.

Fifth, to externally validate the PIERSdynamic model in an unseen cohort, to confirm its performance and gather evidence to suggest changes to clinical guidelines over serial static use of currently recommended models.

Sixth, while all models in the panPIERS tool were tested internally on a dataset held out from the development step, the tool as a whole including the model selection step has not been validated yet. Therefore, the most important line of further investigation would be to externally validate the tool in a diverse dataset, ideally with a larger variety of observed scenarios. This would be particularly important as our testing dataset contained only two different observed scenarios (scenarios 121 and 128), and all other scenarios were generated by removing variables.

And finally, further app development needs to be undertaken to implement the suggestions from the panPIERS usability study, and develop a secure platform that complies with regulations.

References

1. Mammaro A, Carrara S, Cavaliere A, Ermito S, Dinatale A, Pappalardo EM, et al. Hypertensive disorders of pregnancy. J Prenat Med. 2009;3(1):1-5.
2. Association AP. Gestational Hypertension: Pregnancy Induced Hypertension (PIH) 2015 [updated 08/2015. Available from: https://americanpregnancy.org/pregnancy-complications/pregnancy-induced-hypertension/.
3. Von Dadelszen P, Payne B, Li J, Ansermino JM, Pipkin FB, Côté A-M, et al. Prediction of adverse maternal outcomes in Preeclampsia: development and validation of the fullPIERS model. The Lancet. 2011;377(9761):219-27.
4. Altman D, Carroli G, Duley L, Farrell B, Moodley J, Neilson J, et al. Do women with Preeclampsia, and their babies, benefit from magnesium sulphate? The Magpie Trial: a randomised placebo-controlled trial. Lancet. 2002;359(9321):1877-90.
5. Stöppler MC, MD. Eclampsia 2018 [Available from: https://www.emedicinehealth.com/eclampsia/article_em.htm.
6. Ukah UV, Hutcheon JA, Payne B, Haslam MD, Vatish M, Ansermino JM, et al. Placental Growth Factor as a Prognostic Tool in Women With Hypertensive Disorders of Pregnancy: A Systematic Review. Hypertension. 2017;70(6):1228-37.
7. Ukah UV, De Silva DA, Payne B, Magee LA, Hutcheon JA, Brown H, et al. Prediction of adverse maternal outcomes from Preeclampsia and other hypertensive disorders of pregnancy: A systematic review. Pregnancy Hypertens. 2018;11:115-23.
8. Fox R, Kitt J, Leeson P, Aye CYL, Lewandowski AJ. Preeclampsia: Risk Factors, Diagnosis, Management, and the Cardiovascular Impact on the Offspring. J Clin Med. 2019;8(10).
9. Magee LA, Brown MA, Hall DR, Gupte S, Hennessy A, Karumanchi SA, et al. The 2021 International Society for the Study of Hypertension in Pregnancy classification, diagnosis & management recommendations for international practice. Pregnancy Hypertension. 2022;27:148-69.
10. Gregg AR. Preeclampsia. Reference Module in Biomedical Sciences: Elsevier; 2014.
11. Ukah UV, Payne B, Hutcheon JA, Ansermino JM, Ganzevoort W, Thangaratinam S, et al. Assessment of the fullPIERS risk prediction model in women with early-onset preeclampsia. Hypertension. 2018;71(4):659-65.
12. Budden A, Wilkinson L, Buksh MJ, McCowan L. Pregnancy outcome in women presenting with Preeclampsia at less than 25 weeks gestation. Aust N Z J Obstet Gynaecol. 2006;46(5):407-12.
13. Malik A, Jee B, Gupta SK. Preeclampsia: Disease biology and burden, its management strategies with reference to India. Pregnancy Hypertens. 2019;15:23-31.
14. Jenkins SM, Head BB, Hauth JC. Severe preeclampsia at <25 weeks of gestation: maternal and neonatal outcomes. Am J Obstet Gynecol. 2002;186(4):790-5.
15. WHO Guidelines Approved by the Guidelines Review Committee. WHO recommendations: Policy of interventionist versus expectant management of severe Preeclampsia before term. Geneva: World Health Organization

(c) World Health Organization 2018.; 2018.
16. Upadya M, Rao ST. Hypertensive disorders in pregnancy. Indian J Anaesth. 2018;62(9):675-81.

17. Duhig K, Vandermolen B, Shennan A. Recent advances in the diagnosis and management of Preeclampsia. F1000Res. 2018;7:242.
18. Folk DM. Hypertensive Disorders of Pregnancy: Overview and Current Recommendations. J Midwifery Womens Health. 2018;63(3):289-300.
19. National Guideline A. National Institute for Health and Care Excellence: Clinical Guidelines. Hypertension in pregnancy: diagnosis and management. London: National Institute for Health and Care Excellence (UK)

20. Tranquilli AL, Dekker G, Magee L, Roberts J, Sibai BM, Steyn W, et al. The classification, diagnosis and management of the hypertensive disorders of pregnancy: A revised statement from the ISSHP. Pregnancy Hypertens. 2014;4(2):97-104.
21. Brown MA, Magee LA, Kenny LC, Karumanchi SA, McCarthy FP, Saito S, et al. The hypertensive disorders of pregnancy: ISSHP classification, diagnosis & management recommendations for international practice. Pregnancy Hypertens. 2018;13:291-310.
22. Bortolotto MR, Francisco RPV, Zugaib M. Resistant Hypertension in Pregnancy: How to Manage? Curr Hypertens Rep. 2018;20(8):63.
23. Magee LA, von Dadelszen P. State-of-the-Art Diagnosis and Treatment of Hypertension in Pregnancy. Mayo Clin Proc. 2018;93(11):1664-77.
24. LabTestsOnlineUK. Preeclampsia. 2019.
25. Osuchukwu OO RD. Small for Gestational Age. 2022.
26. Foundation TFM. The Fetal Medicine Foundation 2024 [Available from: https://fetalmedicine.org/research/assess/stillbirth.
27. Fu M, Song W, Yu G, Yu Y, Yang Q. Risk factors for length of NICU stay of newborns: A systematic review. Front Pediatr. 2023;11:1121406.
28. Tingting Y, Luke CM, Seema S, Elizabeth AH, Subarna KK, Diwakar M, et al. Risk factors for neonatal mortality: an observational cohort study in Sarlahi district of rural southern Nepal. BMJ Open. 2023;13(9):e066931.
29. Payne BA, Hutcheon JA, Ansermino JM, Hall DR, Bhutta ZA, Bhutta SZ, et al. A Risk Prediction Model for the Assessment and Triage of Women with Hypertensive Disorders of Pregnancy in Low-Resourced Settings: The miniPIERS (Preeclampsia Integrated Estimate of RiSk) Multi-country Prospective Cohort Study. PLoS Medicine. 2014;11(1).
30. Payne BA, Hutcheon JA, Dunsmuir D, Cloete G, Dumont G, Hall D, et al. Assessing the Incremental Value of Blood Oxygen Saturation (SpO<inf>2</inf>) in the miniPIERS (Preeclampsia Integrated Estimate of RiSk) Risk Prediction Model. Journal of Obstetrics and Gynaecology Canada. 2015;37(1):16-24.
31. Millman AL, Payne B, Qu Z, Joanne Douglas M, Hutcheon JA, Lee T, et al. Oxygen saturation as a predictor of adverse maternal outcomes in women with preeclampsia. J Obstet Gynaecol Can. 2011;33(7):705-14.
32. Chan P, Brown M, Simpson JM, Davis G. Proteinuria in Preeclampsia: how much matters? Bjog. 2005;112(3):280-5.
33. Schmidt LJ, Rieger O, Neznansky M, Hackelöer M, Dröge LA, Henrich W, et al. A machine-learning-based algorithm improves prediction of preeclampsia-associated adverse outcomes. Am J Obstet Gynecol. 2022;227(1):77.e1-.e30.
34. Liao Y, Liu XH, Tan J, He GL, Yang HM, Chen M. [Development of a Predictive Model for Adverse Outcomes of Preeclampsia]. Sichuan Da Xue Xue Bao Yi Xue Ban. 2018;49(5):797-802.

35. Ngwenya S, Jones B, Mwembe D, Nare H, Heazell AEP. Development and validation of risk prediction models for adverse maternal and neonatal outcomes in severe preeclampsia in a low-resource setting, Mpilo Central Hospital, Bulawayo, Zimbabwe. Pregnancy Hypertens. 2021;23:18-26.
36. Sun F, Han B, Chen Y, Gao Y, Shen M. Development and external validation of a model for predicting adverse outcomes in women with preeclampsia: A retrospective study from two trans-regional centers in China. Pregnancy Hypertens. 2021;26:133-40.
37. Sroka D, Lorenz-Meyer LA, Scherfeld V, Thoma J, Busjahn A, Henrich W, et al. Comparison of the Soluble fms-Like Tyrosine Kinase 1/Placental Growth Factor Ratio Alone versus a Multi-Marker Regression Model for the Prediction of Preeclampsia-Related Adverse Outcomes after 34 Weeks of Gestation. Fetal Diagn Ther. 2023;50(3):215-24.
38. Zheng D, Hao X, Khan M, Wang L, Li F, Xiang N, et al. Comparison of machine learning and logistic regression as predictive models for adverse maternal and neonatal outcomes of preeclampsia: A retrospective study. Front Cardiovasc Med. 2022;9:959649.
39. Wang L, Mo Y, Wang P, Shen W, Xu L, Zhao G, et al. Prediction Model of Adverse Pregnancy Outcome in Preeclampsia Based on Logistic Regression and Random Forest Algorithm. Altern Ther Health Med. 2024;30(1):142-7.
40. Thangaratinam S, Allotey J, Marlin N, Mol BW, Von Dadelszen P, Ganzevoort W, et al. Development and validation of Prediction models for Risks of complications in Early-onset Preeclampsia (PREP): a prospective cohort study. Health Technol Assess. 2017;21(18):1-100.
41. Ukah UV, Payne B, Karjalainen H, Kortelainen E, Seed PT, Conti-Ramsden FI, et al. Temporal and external validation of the fullPIERS model for the prediction of adverse maternal outcomes in women with Preeclampsia. Pregnancy Hypertension. 2019;15:42-50.
42. Vivian Ukah U, Payne B, Lee T, Magee LA, Von Dadelszen P. External Validation of the fullPIERS Model for Predicting Adverse Maternal Outcomes in Pregnancy Hypertension in Low- and Middle-Income Countries. Hypertension. 2017;69(4):705-11.
43. Magee LA, Smith GN, Bloch C, Côté AM, Jain V, Nerenberg K, et al. Guideline No. 426: Hypertensive Disorders of Pregnancy: Diagnosis, Prediction, Prevention, and Management. J Obstet Gynaecol Can. 2022;44(5):547-71.e1.
44. Heitkamp A, Meulenbroek A, van Roosmalen J, Gebhardt S, Vollmer L, de Vries JI, et al. Maternal mortality: near-miss events in middle-income countries, a systematic review. (1564-0604 (Electronic)).
45. Magee LA, Sharma S, Nathan HL, Adetoro OO, Bellad MB, Goudar S, et al. The incidence of pregnancy hypertension in India, Pakistan, Mozambique, and Nigeria: A prospective population-level analysis. PLoS Med. 2019;16(4):e1002783.
46. Global, regional, and national levels of maternal mortality, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1775-812.
47. Payne BA, Groen H, Ukah UV, Ansermino JM, Bhutta Z, Grobman W, et al. Development and internal validation of a multivariable model to predict perinatal death in pregnancy hypertension. Pregnancy Hypertension. 2015;5(4):315-21.
48. Wright D, Wright A, Nicolaides KH. The competing risk approach for prediction of preeclampsia. Am J Obstet Gynecol. 2020;223(1):12-23.e7.

49. Thangaratinam S, Koopmans CM, Iyengar S, Zamora J, Ismail KM, Mol BW, et al. Accuracy of liver function tests for predicting adverse maternal and fetal outcomes in women with preeclampsia: a systematic review. Acta Obstet Gynecol Scand. 2011;90(6):574-85.
50. Gartner_Inc. Definition of Predictive Modeling - Gartner Information Technology Glossary: Gartner_Inc; 2021 [Available from: https://www.gartner.com/en/information-technology/glossary/predictive-modeling.
51. Carew JM, Lawton G, Burns E. What is Predictive Modeling? : @TTBusinessTech; 2021 [Available from: https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling.
52. Ali R. Predictive Modeling: Types, Benefits, and Algorithms. Oracle NetSuite; 2020.
53. What is Machine Learning? A definition - Expert System 2017 [updated March 2020. Available from: https://expertsystem.com/machine-learning-definition/.
54. Brownlee J. What Is Semi-Supervised Learning - Machine Learning Mastery. 2021.
55. Maxime. Introduction to semi-supervised learning and adversarial training: @IBMData; 2019 [updated 2019-05-24. Available from: https://medium.com/inside-machine-learning/placeholder-3557ebb3d470.
56. Osiński B, Budek K. What is reinforcement learning? The complete guide - deepsense.ai. 2018.
57. @algorithmia. The importance of machine learning data | Algorithmia Blog. 2020.
58. Mack C, Su Z, Westreich D. AHRQ Methods for Effective Health Care. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018.
59. Gutbezahl J. 5 Types of Statistical Biases to Avoid in Your Analyses 2017 [updated 2017-06-13. Available from: https://online.hbs.edu/blog/post/types-of-statistical-bias.
60. Sires E. 9 Ways to Prevent Data Bias in Predictive Models - Rapid Insight. 2020.
61. Grace-Martin K. How to Diagnose the Missing Data Mechanism - The Analysis Factor: @analysis_factor; 2013 [updated 2013-05-20. Available from: https://www.theanalysisfactor.com/missing-data-mechanism/.
62. Leavitt C. Uncovering Missing Not at Random Data: @TDataScience; 2019 [updated 2019-10-24. Available from: https://towardsdatascience.com/uncovering-missing-not-at-random-data-8d2cd3eda31a.
63. Groenwold RHH, Dekkers OM. Missing data: the impact of what is not there. European Journal of Endocrinology. 2020;183(4):E7-E9.
64. Kumar S. 7 Ways to Handle Missing Values in Machine Learning. 2020.
65. Grace-Martin K. Seven Ways to Make up Data: Common Methods to Imputing Missing Data - The Analysis Factor. 2009.
66. Jamshidian M, Mata M. 2 - Advances in Analysis of Mean and Covariance Structure when Data are Incomplete**This research was supported in part by the National Science Foundation Grant DMS-0437258. In: Lee S-Y, editor. Handbook of Latent Variable and Related Models. Amsterdam: North-Holland; 2007. p. 21-44.
67. Karen G-M. Two Recommended Solutions for Missing Data: Multiple Imputation and Maximum Likelihood - The Analysis Factor. 2012.

68. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17(1):162.
69. SurveyMethods. When to Use Single Imputation or Multiple Imputation: @SurveyMethods; 2011 [updated 2011-10-07. Available from: https://surveymethods.com/when-to-use-single-imputation-or-multiple-imputation/.
70. Bodner TE. What improves with increased missing data imputations? Structural equation modeling: a multidisciplinary journal. 2008;15(4):651-75.
71. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99.
72. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus. 2013;2(1):222-.
73. Montelpare WJ, Read E, McComber T, Mahar A, Ritchie K. Working with Missing Data. Applied Statistics in Healthcare Research2020.
74. Swalin A. How to Handle Missing Data - Towards Data Science. Medium. 2018.
75. Lee JH, Huber JC, Jr. Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much? Iranian journal of public health. 2021;50(7):1372-80.
76. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 2019;110:63-73.
77. Vadapalli P. upGrad blog. 2020 2020-07-27. Available from: https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/.
78. Glen S. Lasso Regression: Simple Definition. StatisticsHowTocom: Elementary Statistics for the rest of us! 2015.
79. Health CUMSoP. Least Absolute Shrinkage and Selection Operator (LASSO) [Available from: https://www.publichealth.columbia.edu/research/population-health-methods/least-absolute-shrinkage-and-selection-operator-lasso.
80. Glen S. StatisticsHowTo.com: Elementary Statistics for the rest of us". 2017. [cited 2021]. Available from: https://www.statisticshowto.com/ridge-regression/.
81. NCSS. Ridge Regression. NCSS Documentation. Kaysville, Utah, USA: NCSS, LLC; 2021.
82. Nagpal A. L1 and L2 Regularization Methods. 2017.
83. Miller M. The Basics: Logistic Regression and Regularization. 2019.
84. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010. 2010;33(1):22.
85. Brownlee J. Classification And Regression Trees for Machine Learning 2016 [updated August 15, 2020. Available from: https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.
86. van der Schaar M. Classification and Regression Trees. University of Oxford; 2017.
87. Vala K. Tree-Based Methods: Regression Trees - Towards Data Science. Medium. 2019.
88. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10(213).

89. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Statistics and Computing. 2017;27(3):659-78.
90. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11):1-13.
91. Celik E. vita: Variable Importance Testing Approaches. 2015.
92. developers x. Introduction to Boosted Trees — xgboost 2.0.3 documentation. Readthedocsio. 2022.
93. Raftery A, Hoeting J, Volinsky C, Painter I, Yeung KY. BMA: Bayesian Model Averaging. 2020.
94. Madigan D, Raftery AE. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. Journal of the American Statistical Association. 1994;89(428):1535-46.
95. Raftery AE, Madigan D, Hoeting JA. Bayesian Model Averaging for Linear Regression Models. https://doiorg/101080/01621459199710473615. 2012.
96. Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodology. 1995;25:111-63.
97. Hinne M, Gronau QF, van den Bergh D, Wagenmakers E-J. A Conceptual Introduction to Bayesian Model Averaging. Advances in Methods and Practices in Psychological Science. 2020;3(2):200-15.
98. Naviani A. ANN (Artificial Neural Network) Models in R: Code & Examples on How to Build Your NN. 2021.
99. Nicholson C. A Beginner's Guide to Neural Networks and Deep Learning. 2021.
100. Education IC. What are Neural Networks? 2021 [updated 2021-08-03. Available from: https://www.ibm.com/cloud/learn/neural-networks.
101. (https://stats.stackexchange.com/users/248160/ketan-vishwakarma) KV. Difference between neural net weight decay and learning rate: Cross Validated; 2021 [Available from: https://stats.stackexchange.com/q/408702.
102. Verma A. Building A Neural Net from Scratch Using R - Part 1. 2020.
103. Khandelwal R. K fold and other cross-validation techniques: @dd_invest; 2019 [updated 2019-01-25. Available from: https://medium.datadriveninvestor.com/k-fold-and-other-cross-validation-techniques-6c03a2563f1e.
104. Likelihood Ratios — Centre for Evidence-Based Medicine (CEBM), University of Oxford. 2020.
105. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology. 2010;5(9):1315-6.
106. McGee S. Simplifying likelihood ratios. Journal of general internal medicine. 2002;17(8):646-9.
107. Shapley LS. A Value for N-Person Games. Santa Monica, CA: RAND Corporation; 1952.
108. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2 ed2022.
109. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass). 2010;21(1):128-38.
110. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.

111. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020;27(4):621-33.
112. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? Journal of Clinical Epidemiology. 2020;118:93-9.
113. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
114. Milholland AV, Wheeler SG, Heieck JJ. Medical assessment by a Delphi group opinion technic. N Engl J Med. 1973;288(24):1272-5.
115. johnswentworth. 2019. Available from: https://www.lesswrong.com/posts/vFom2WsZkqCJPWqvZ/cross-validation-vs-bayesian-model-comparison.
116. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. Jama. 1994;271(9):703-7.
117. Executive B. Expert Committee on Health Statistics: report on second session. Geneva: World Health Organization; 1950.
118. Thangaratinam S, Allotey J, Marlin N, Dodds J, Cheong-See F, von Dadelszen P, et al. Prediction of complications in early-onset Preeclampsia (PREP): development and external multinational validation of prognostic models. BMC medicine. 2017;15(1):68.
119. Magee LA, von Dadelszen P, Rey E, Ross S, Asztalos E, Murphy KE, et al. Less-tight versus tight control of hypertension in pregnancy. N Engl J Med. 2015;372(5):407-17.
120. von Dadelszen P, Bhutta ZA, Sharma S, Bone J, Singer J, Wong H, et al. The Community-Level Interventions for Preeclampsia (CLIP) cluster randomised trials in Mozambique, Pakistan, and India: an individual participant-level meta-analysis. Lancet. 2020;396(10250):553-63.
121. Duffy J, Cairns AE, Richards-Doran D, van 't Hooft J, Gale C, Brown M, et al. A core outcome set for Preeclampsia research: an international consensus development study. Bjog. 2020;127(12):1516-26.
122. National Guideline Alliance and National Institute for Health and Care E. Hypertension in pregnancy : diagnosis and management.
123. Korniluk A, Koper-Lenkiewicz OM, Kamińska J, Kemona H, Dymicka-Piekarska V. Mean Platelet Volume (MPV): New Perspectives for an Old Marker in the Course and Prognosis of Inflammatory Conditions. Mediators Inflamm. 2019;2019:9213074.
124. Duley L, Gülmezoglu AM, Henderson-Smart DJ, Chou D. Magnesium sulphate and other anticonvulsants for women with Preeclampsia. Cochrane Database Syst Rev. 2010;2010(11):Cd000025.
125. Magee LA, De Silva DA, Sawchuck D, Synnes A, von Dadelszen P. No. 376-Magnesium Sulphate for Fetal Neuroprotection. J Obstet Gynaecol Can. 2019;41(4):505-22.
126. Rana S, Burke SD, Karumanchi SA. Imbalances in circulating angiogenic factors in the pathophysiology of preeclampsia and related disorders. Am J Obstet Gynecol. 2022;226(2s):S1019-s34.
127. Duhig KE, Myers J, Seed PT, Sparkes J, Lowe J, Hunter RM, et al. Placental growth factor testing to assess women with suspected Preeclampsia: a multicentre,

pragmatic, stepped-wedge cluster-randomised controlled trial. Lancet. 2019;393(10183):1807-18.
128. Hayes-Ryan D, Khashan AS, Hemming K, Easter C, Devane D, Murphy DJ, et al. Placental growth factor in assessment of women with suspected Preeclampsia to reduce maternal morbidity: a stepped wedge cluster randomised control trial (PARROT Ireland). Bmj. 2021;374:n1857.
129. Nicolaides KH, Sarno M, Wright A. Ophthalmic artery Doppler in the prediction of preeclampsia. Am J Obstet Gynecol. 2022;226(2s):S1098-s101.
130. Magee LA, Nicolaides KH, von Dadelszen P. Preeclampsia. N Engl J Med. 2022;386(19):1817-32.
131. Group USC. Introduction to Linear Mixed Models. Uclaedu. 2021.
132. Magezi DA. Linear mixed-effects models for within-participant psychology experiments: an introductory tutorial and free, graphical user interface (LMMgui). Frontiers in Psychology. 2015;6.
133. Sushmita S. Chapter 17: Mixed Effects Modeling. Uicedu. 2024.
134. Dandis R, Teerenstra S, Massuger L, Sweep F, Eysbouts Y, IntHout J. A tutorial on dynamic risk prediction of a binary outcome based on a longitudinal biomarker. Biom J. 2020;62(2):398-413.
135. McHunu NN, Mwambi HG, Reddy T, Yende-Zuma N, Naidoo K. Joint modelling of longitudinal and time-to-event data: an illustration using CD4 count and mortality in a cohort of patients initiated on antiretroviral therapy. BMC Infect Dis. 2020;20(1):256.
136. Zhang D, Chen MH, Ibrahim JG, Boye ME, Shen W. Bayesian Model Assessment in Joint Modeling of Longitudinal and Survival Data with Applications to Cancer Clinical Trials. J Comput Graph Stat. 2017;26(1):121-33.
137. Rizopoulos D, Papageorgiou G, Miranda Afonso P. JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data. 2023.
138. Speiser JL, Wolf BJ, Chung D, Karvellas CJ, Koch DG, Durkalski VL. BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. Chemometrics and Intelligent Laboratory Systems. 2019;185:122-34.
139. Olah C. Understanding LSTM Networks -- colah's blog. Githubio. 2015.
140. Yan S. Understanding LSTM and its diagrams - ML Review. Medium. 2016.
141. Jiang C, Chen Y, Chen S, Bo Y, Li W, Wenxin T, et al. A Mixed Deep Recurrent Neural Network for MEMS Gyroscope Noise Suppressing. Electronics. 2019;8:181.
142. Nielsen MA. Neural Networks and Deep Learning: Determination Press; 2015.
143. Koech KE. Cross-Entropy Loss Function. 2022.
144. Team K. 2023. Available from: https://keras.io/api/losses/.
145. Ganzevoort W, Rep A, Bonsel GJ, De Vries JI, Wolf H. A randomized trial of plasma volume expansion in hypertensive disorders of pregnancy: influence on the pulsatility indices of the fetal umbilical artery and middle cerebral artery. Am J Obstet Gynecol. 2005;192(1):233-9.
146. Akkermans J, Payne B, Dadelszen Pv, Groen H, Vries Jd, Magee LA, et al. Predicting complications in Preeclampsia: external validation of the fullPIERS model using the PETRA trial dataset. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2014;179:58-62.
147. Clark M. Convergence Problems. Michael Clark. 2020.
148. Lackland DT. Racial differences in hypertension: implications for high blood pressure management. Am J Med Sci. 2014;348(2):135-8.

149. Wang Y, Wu N, Shen H. A Review of Research Progress of Pregnancy with Twins with Preeclampsia. Risk Manag Healthc Policy. 2021;14:1999-2010.
150. Chollet F, et al. Keras. 2015.
151. Allaire J, Chollet F. R Interface to 'Keras' [R package keras version 2.13.0]. R-projectorg. 2023.
152. Allaire J, Tang Y. R Interface to 'TensorFlow' [R package tensorflow version 2.15.0]. R-projectorg. 2024.
153. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
154. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
155. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. Journal of Clinical Epidemiology. 2020;125:183-7.
156. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. Stat Med. 2020;39(25):3591-607.
157. Buuren S, Groothuis-Oudshoorn C. MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45.
158. Marcelin JR, Siraj DS, Victor R, Kotadia S, Maldonado YA. The Impact of Unconscious Bias in Healthcare: How to Recognize and Mitigate It. The Journal of Infectious Diseases. 2019;220(Supplement_2):S62-S73.
159. Gopal DP, Chetty U, O'Donnell P, Gajria C, Blackadder-Weinstein J. Implicit bias in healthcare: clinical practice, research and decision making. Future Healthc J. 2021;8(1):40-8.
160. Wolpert DH. Stacked generalization. Neural Networks. 1992;5(2):241-59.
161. Han J, Kang S. Optimization of missing value imputation for neural networks. Information Sciences. 2023;649:119668.
162. Deng Y. mixgb: Multiple Imputation Through 'XGBoost'. 2023.
163. Dong W, Fong DYT, Yoon J-s, Wan EYF, Bedford LE, Tang EHM, et al. Generative adversarial networks for imputing missing data for big data clinical research. BMC Medical Research Methodology. 2021;21(1):78.
164. Yoon J, Jordon J, van der Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets. 2018.
165. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine. 2018;35(1):53-65.
166. Goodfellow IaP-AJaMMaXBaW-FDaOSaCAaBY. Generative Adversarial Nets. In: Weinberger ZGaMWaCCaNLaKQ, editor.2014.
167. Guida JP, Cralcev C, Costa Santos J, Marangoni-Junior M, Sanchez MP, Laura Costa M. Validation of the fullPIERS model for prediction of adverse outcomes in preeclampsia at a referral center. Pregnancy Hypertens. 2021;23:112-5.
168. Gupta M, Yadav P, Yaqoob F. A Prospective Study to Determine the Predictive Ability of HDP-Gestosis Score for the Development of Preeclampsia. J Obstet Gynaecol India. 2022;72(6):485-91.
169. Romundstad LB, Romundstad Pr Fau - Sunde A, Sunde A Fau - von Düring V, von Düring V Fau - Skjaerven R, Skjaerven R Fau - Vatten LJ, Vatten LJ. Increased risk of

placenta previa in pregnancies following IVF/ICSI; a comparison of ART and non-ART pregnancies in the same mother. Hum Reprod. 2006;21(9):2353-8.
170. Clowse ME, Jamison M Fau - Myers E, Myers E Fau - James AH, James AH. A national study of the complications of lupus in pregnancy. Am J Obstet Gynecol. 2008;199(2):127.e1-6.

Appendix A: panPIERS prediction code

R component of code

```
if(!"reticulate" %in% (.packages())){
  require(reticulate)
}
if(!"sjlabelled" %in% (.packages())){
  require(sjlabelled)
}
```

```
minmax <- read_rds("minmax.rds")
mean <- read_rds("mean.rds")
```

```
group_other <- function(data){
  names <- colnames(data)
  xx = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% as.matrix(
) %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% mutate_a
t(vars(-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unna
me()
  min = minmax$other$min
  max = minmax$other$max
  mean = mean$other
  full = py$group_other(data,xx,Train_No,min,max,mean) %>% as_tibble()
%>% set_names(names)
  return(full)
}
```

```
group_patient <- function(data){
  ethncols <- c(Ethnicity0=0,Ethnicity1=0,Ethnicity2=0,Ethnicity3=0)
  xx = data %>% mutate(dummy_v=1,r=1:nrow(.)) %>%
    pivot_wider(names_from = Ethnicity, names_prefix = "Ethnicity",val
ues_from = dummy_v,values_fill=0) %>% select(-r) %>%
    mutate_if(~is.factor(.),~as_numeric(.)-1) %>%
    add_column(.,!!!ethncols[setdiff(paste0("Ethnicity",0:3),colnames(
.))]) %>%
    mutate_at(vars(Ethnicity0,Ethnicity1,Ethnicity2,Ethnicity3),~case_
when(EthnicityNA==0~.)) %>% select(-EthnicityNA) %>%
    as.matrix() %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate(dummy_v=1,r=1:nrow(.)) %>%
    pivot_wider(names_from = Ethnicity, names_prefix = "Ethnicity",val
ues_from = dummy_v,values_fill=0) %>% select(-r) %>%
    mutate_if(~is.factor(.),~as_numeric(.)-1) %>%
    add_column(.,!!!ethncols[setdiff(paste0("Ethnicity",0:3),colnames(
.))]) %>%
```

```
    mutate_at(vars(Ethnicity0,Ethnicity1,Ethnicity2,Ethnicity3),~case_
when(EthnicityNA==0~.)) %>% select(-EthnicityNA) %>%
    mutate_at(vars(-days),~case_when(is.na(.)~-1,TRUE~.))
  names <- colnames(data)
  data %<>% as.matrix() %>% unname()
  min = minmax$patient$min
  max = minmax$patient$max
  mean = mean$patient
  full = py$group_patient(data,xx,Train_No,min,max,mean) %>% as_tibble
() %>% set_names(names)
  return(full)
}
```

```
group_symptoms <- function(data){
  names <- colnames(data)
  xx = data %>% mutate(headOrvis=factor(headOrvis,levels=c("No","Yes")
)) %>% mutate_if(~is.factor(.),~as_numeric(.)-1) %>% as.matrix() %>% u
nname()
  Train_No = nrow(xx)
  data = data %>% mutate(headOrvis=factor(headOrvis,levels=c("No","Yes
"))) %>% mutate_if(~is.factor(.),~as_numeric(.)-1) %>% mutate_at(vars(
-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unname()
  min = minmax$symptoms$min
  max = minmax$symptoms$max
  mean = mean$symptoms
  full = py$group_symptoms(data,xx,Train_No,min,max,mean) %>% as_tibbl
e() %>% set_names(names)
  return(full)
}
```

```
group_medhist <- function(data){
  names <- colnames(data)
  xx = data %>%
    mutate_at(vars(d_renal,d_htn,d_pgd,ad_smoke),~as_numeric(.)-1) %>%
    mutate_at(vars(multipar,ad_gd),~as_numeric(.)) %>% as.matrix() %>%
unname()
  Train_No = nrow(xx)
  data = data %>%
    mutate_at(vars(d_renal,d_htn,d_pgd,ad_smoke),~as_numeric(.)-1) %>%
    mutate_at(vars(multipar,ad_gd),~as_numeric(.)) %>% mutate_at(vars(
-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unname()
  min = minmax$mhist$min
  max = minmax$mhist$max
  mean = mean$mhist
  full = py$group_medhist(data,xx,Train_No,min,max,mean) %>% as_tibble
() %>% set_names(names)
  return(full)
}
```

```
group_signs <- function(data){
  names <- colnames(data)
  xx = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% as.matrix(
) %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% mutate_a
t(vars(-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unna
me()
  min = minmax$signs$min
  max = minmax$signs$max
  mean = mean$signs
  full = py$group_signs(data,xx,Train_No,min,max,mean) %>% as_tibble()
%>% set_names(names)
  return(full)
}
```

```
group_blood1 <- function(data){
  names <- colnames(data)
  xx = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% as.matrix(
) %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% mutate_a
t(vars(-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unna
me()
  min = minmax$blood1$min
  max = minmax$blood1$max
  mean = mean$blood1
  full = py$group_blood1(data,xx,Train_No,min,max,mean) %>% as_tibble(
) %>% set_names(names)
  return(full)
}
```

```
group_blood2 <- function(data){
  names <- colnames(data)
  xx = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% as.matrix(
) %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>%mutate_at
(vars(-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unnam
e()
  min = minmax$blood2$min
  max = minmax$blood2$max
  mean = mean$blood2
  full = py$group_blood2(data,xx,Train_No,min,max,mean) %>% as_tibble(
) %>% set_names(names)
  return(full)
}
```

```
group_dip <- function(data){
  names <- colnames(data)
```

```
  xx = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% as.matrix(
) %>% unname()
  Train_No = nrow(xx)
  data = data %>% mutate_if(~is.factor(.),~as_numeric(.)) %>% mutate_a
t(vars(-days),~case_when(is.na(.)~-1,TRUE~.)) %>% as.matrix() %>% unna
me()
  min = minmax$dipstick$min
  max = minmax$dipstick$max
  mean = mean$dipstick
  full = py$group_dip(data,xx,Train_No,min,max,mean) %>% as_tibble() %
>% set_names(names)
  return(full)
}
```

```
GAIN <- function(data){
  if (!is.data.frame(data)) {stop("Data is not data.frame or tibble fo
rmat")}
  order <- colnames(data)
  data %<>% rownames_to_column("row")
  patient_info <- c("fetusNum","Ethnicity")
  if (any(!patient_info %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(pati
ent_info, colnames(data)),collapse=", ")))}
  symptoms <- c("vomitting","ruq","headOrvis","chestOrdysp")
  if (any(!symptoms %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(symp
toms, colnames(data)),collapse=", ")))}
  medical_history <- c("multipar","d_renal","d_htn","d_pgd","ad_gd","a
d_smoke")
  if (any(!medical_history %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(medi
cal_history, colnames(data)),collapse=", ")))}
  signs <- c("SpO2","heightcm","ad_wgtkg")
  if (any(!signs %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(sign
s, colnames(data)),collapse=", ")))}
  blood1 <- c("uricacid","wbc","plat","mpv","hematocrit","creatinine",
"ast","alt","ldh","albumin")
  if (any(!blood1 %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(bloo
d1, colnames(data)),collapse=", ")))}
  blood2 <- c("inr","fibrinogen","aptt")
  if (any(!blood2 %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(bloo
d2, colnames(data)),collapse=", ")))}
  urine <- c("dipstick")
  if (any(!urine %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(urin
e, colnames(data)),collapse=", ")))}
  other <- c("days","GA","sBP","dBP","MaternalMortality","GDP","AgeAtE
DD")
  if (any(!other %in% colnames(data))) {
    stop(paste("The following columns are missing:",paste(setdiff(othe
```

```
r, colnames(data)),collapse=", "))}
  if (data %>%
      select(row,all_of(patient_info),
             all_of(symptoms),
             all_of(medical_history),
             all_of(signs),
             all_of(blood1),
             all_of(blood2),
             all_of(urine)) %>%
      filter_at(vars(-row), any_vars(!is.na(.))) %>% filter_at(vars(-r
ow), any_vars(is.na(.))) %>%
      nrow() != 0){
    if (data %>% select(row,all_of(patient_info)) %>% filter_at(vars(-
row), any_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.)
)) %>% nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(patient_info)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      factors <- data %>%
        select(row,all_of(other),all_of(patient_info)) %>%
        select_if(~is.factor(.)) %>%
        colnames()
      data_pat <- data %>%
        select(all_of(other),all_of(patient_info)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_patient(.) %>% mutate(row=rows) %>%
        mutate_at(vars(Ethnicity0,Ethnicity1,Ethnicity2,Ethnicity3),~r
ound(.)) %>%
        pivot_longer(c(Ethnicity0,Ethnicity1,Ethnicity2,Ethnicity3),na
mes_to=c(NA,"Ethnicity"),names_sep = 9,values_to="dummy") %>%
        group_by(row) %>%
        mutate(dummy=case_when(all(dummy==0)&Ethnicity==0~1,TRUE~dummy
)) %>%
        filter(dummy==1) %>%
        mutate(Ethnicity=as.numeric(Ethnicity),Ethnicity=case_when(0 %
in% Ethnicity~0,TRUE~max(Ethnicity))) %>%
        ungroup() %>%
        distinct() %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(patient_info)))

      data_pat %<>% mutate(fetusNum=factor(fetusNum,levels=c(0,1),labe
ls = c("Singleton pregnancy","Multiple pregnancy")),Ethnicity=as.facto
r(Ethnicity))


      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_pat)
    }
    if (data %>% select(row,all_of(symptoms)) %>% filter_at(vars(-row)
```

```
, any_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.))) %
>% nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(symptoms)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      factors <- data %>%
        select(row,all_of(other),all_of(symptoms)) %>%
        select_if(~is.factor(.)) %>%
        colnames()
      data_sym <- data %>%
        select(all_of(other),all_of(symptoms)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_symptoms(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(symptoms)))

        data_sym %<>% mutate_at(factors[-3],~factor(.,levels = c(0,1),
labels = c("No","Yes"))) %>%
          mutate(headOrvis=factor(headOrvis,levels = c(1,0),labels = c
("Yes","No")))

      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_sym)
    }
    if (data %>% select(row,all_of(medical_history)) %>% filter_at(var
s(-row), any_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na
(.))) %>% nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(medical_history)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      factors <- data %>%
        select(row,all_of(other),all_of(medical_history)) %>%
        select_if(~is.factor(.)) %>%
        colnames()
      data_mhist <- data %>%
        select(all_of(other),all_of(medical_history)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_medhist(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(medical_history
)))

        data_mhist %<>% mutate_at(factors[-c(1,5)],~factor(.,levels=c(
0,1),labels=c("No","Yes"))) %>%
          mutate_at(factors[c(1,5)],~as.factor(.))
```

```
      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_mhist)
    }
    if (data %>% select(row,all_of(signs)) %>% filter_at(vars(-row), a
ny_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.))) %>%
nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(signs)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      data_sig <- data %>%
        select(all_of(other),all_of(signs)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_signs(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(signs)))

      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_sig)
    }
    if (data %>% select(row,all_of(blood1)) %>% filter_at(vars(-row),
any_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.))) %>%
nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(blood1)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      data_bl1 <- data %>%
        select(all_of(other),all_of(blood1)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_blood1(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(blood1)))

      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_bl1)
    }
    if (data %>% select(row,all_of(blood2)) %>% filter_at(vars(-row),
any_vars(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.))) %>%
nrow() != 0) {
      rows <- data %>%
        select(row,all_of(other),all_of(blood2)) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      data_bl2 <- data %>%
```

```
        select(all_of(other),all_of(blood2)) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_blood2(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-all_of(blood2)))

      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_bl2)
    }
    if (data %>% select(row,dipstick) %>% filter_at(vars(-row), any_va
rs(!is.na(.))) %>% filter_at(vars(-row), any_vars(is.na(.))) %>% nrow(
) != 0) {
      rows <- data %>%
        select(row,all_of(other),wbc,plat, creatinine,alt,albumin,dips
tick) %>%
        filter_at(vars(-row), any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        pull(row)
      data_dip <- data %>%
        select(all_of(other),wbc,plat, creatinine,alt,albumin,dipstick
) %>%
        filter_all(any_vars(!is.na(.))) %>%
        filter_all(any_vars(is.na(.))) %>%
        group_dip(.) %>% mutate(row=rows) %>%
        bind_cols(data %>%
                    filter(row %in% rows) %>%
                    select(-row,-all_of(other),-c(wbc,plat, creatinine
,alt,albumin,dipstick)))

      data %<>%
        filter(!row %in% rows) %>%
        bind_rows(data_dip)
    }
  }

  if (data %>%
      select(row,all_of(other)) %>%
      filter_if(~is.numeric(.), any_vars(!is.na(.))) %>% filter_if(~is
.numeric(.), any_vars(is.na(.))) %>%
      nrow() != 0) {
    rows <- data %>%
      select(row,all_of(other)) %>%
      filter_at(vars(-row), any_vars(!is.na(.))) %>%
      filter_all(any_vars(is.na(.))) %>%
      pull(row)
    data_other <- data %>%
      select(all_of(other)) %>%
      filter_all(any_vars(!is.na(.))) %>%
      filter_all(any_vars(is.na(.))) %>%
      group_other(.) %>% mutate(row=rows) %>%
      bind_cols(data %>%
                  filter(row %in% rows) %>%
```

```
                  select(-row,-all_of(other)))

    data %<>%
      filter(!row %in% rows) %>%
      bind_rows(data_other)
  }

  data %<>% select(all_of(order))

  return(data)

}
```

Python component of code

```
import tensorflow as tf
import pandas as pd
import numpy as np
import os
from sklearn.model_selection import train_test_split
from tqdm import tqdm
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
from sklearn.preprocessing import MinMaxScaler
import openpyxl
from itertools import product
import xlrd
from xlutils.copy import copy
import time
```

```
generator_other = tf.keras.models.load_model("GAIN models new/other/ge
nerator")

def group_other(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1]*len(data[0]))
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
```

```
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_other([X_final, M_final, finalX, 1-M_final
], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6]] = np.round(sample_final[:,[0,2,3,4,6]],
0)
  sample_final[:,[5]] = np.round(sample_final[:,[5]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

```
generator_patient = tf.keras.models.load_model("GAIN models new/patien
t/generator")

def group_patient(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0])
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con[0,:]
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_patient([X_final, M_final, finalX, 1-M_fin
al], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,7,8]] = np.round(sample_final[:,[0,2,3,4,6
,7,8]], 0)
  sample_final[:,[5]] = np.round(sample_final[:,[5]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

```
generator_symptoms = tf.keras.models.load_model("GAIN models new/sympt
om/generator")

def group_symptoms(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0])
```

```
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_symptoms([X_final, M_final, finalX, 1-M_fi
nal], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,7,8,9,10]] = np.round(sample_final[:,[0,2,
3,4,6,7,8,9,10]], 0)
  sample_final[:,[5]] = np.round(sample_final[:,[5]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

```
generator_mhist = tf.keras.models.load_model("GAIN models new/mhist/ge
nerator")

def group_medhist(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0])
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_mhist([X_final, M_final, finalX, 1-M_final
], training=False)
```

```
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,7,8,9,10,11,12]] = np.round(sample_final[:
,[0,2,3,4,6,7,8,9,10,11,12]], 0)
  sample_final[:,[5]] = np.round(sample_final[:,[5]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

```
generator_signs = tf.keras.models.load_model("GAIN models new/signs/ge
nerator")

def group_signs(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1]*len(data[0]))
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_signs([X_final, M_final, finalX, 1-M_final
], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,7]] = np.round(sample_final[:,[0,2,3,4,6,7
]], 0)
  sample_final[:,[5,8]] = np.round(sample_final[:,[5,8]], 2)
  sample_final[:,[1,9]] = np.round(sample_final[:,[1,9]], 5)
  return sample_final
```

```
generator_blood1 = tf.keras.models.load_model("GAIN models new/blood1/
generator")

def group_blood1(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1]*len(data[0]))
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
```

```
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_blood1([X_final, M_final, finalX, 1-M_fina
l], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,7,9,12,13,14,15,16]] = np.round(sample_fin
al[:,[0,2,3,4,6,7,9,12,13,14,15,16]], 0)
  sample_final[:,[5,8,10]] = np.round(sample_final[:,[5,8,10]], 2)
  sample_final[:,[1,11]] = np.round(sample_final[:,[1,11]], 5)
  return sample_final
```

```
generator_blood2 = tf.keras.models.load_model("GAIN models new/blood2/
generator")

def group_blood2(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1]*len(data[0]))
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_blood2([X_final, M_final, finalX, 1-M_fina
l], training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
```

```
  sample_final[:,[0,2,3,4,6]] = np.round(sample_final[:,[0,2,3,4,6]],
0)
  sample_final[:,[5,7,8,9]] = np.round(sample_final[:,[5,7,8,9]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

```
generator_dip = tf.keras.models.load_model("GAIN models new/dipstick/g
enerator")

def group_dip(data, xx, Train_No, xmin, xmax, meanvalue):
  con = np.array([1]*len(data[0]))
  con = [con, ] * Train_No
  con = np.asarray(con)
  mask = np.isnan(xx)
  mask = mask + 0
  mask = 1. - mask
  max_min = xmax - xmin
  max_min = max_min + 1e-8
  mydata = ((data - xmin) / max_min) * con + (1 - con) * data
  mydata = np.asarray(mydata)
  meanZ = [meanvalue, ] * Train_No
  meanZ = np.asarray(meanZ)

  X_final = mydata
  M_final = mask
  Z_final = meanZ
  CO_final = con
  finalX = M_final * X_final + (1 - M_final) * Z_final
  sample_final1 = generator_dip([X_final, M_final, finalX, 1-M_final],
training=False)
  sample_final2 = X_final * M_final + (1 - M_final) * sample_final1
  sample_final = sample_final2 * max_min + xmin
  sample_final = sample_final.numpy()
  sample_final[:,[0,2,3,4,6,8,9,10,11,12]] = np.round(sample_final[:,[
0,2,3,4,6,8,9,10,11,12]], 0)
  sample_final[:,[5,7]] = np.round(sample_final[:,[5,7]], 2)
  sample_final[:,[1]] = np.round(sample_final[:,[1]], 5)
  return sample_final
```

Appendix B: Performance drift detected from consecutive prediction of the adverse maternal outcomes of preeclampsia using the PIERS-ML model: A multi-country prospective observational study

Guiyou Yang[1$], Tünde Montgomery-Csobán[2$], Wessel Ganzevoort[3,4], Sanne J Gordijn[5], Kimberley Kavanagh[2], Paul Murray[6], Laura A. Magee[7], Henk Groen[1#], Peter von Dadelszen[7#]

Affiliations

1 Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

2 Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom

3 Department of Obstetrics and Gynaecology, Amsterdam University Medical Centers, location AMC, Amsterdam, The Netherlands.

4 Amsterdam Reproduction and Development Research Institute, Amsterdam, The Netherlands..

5 Department of Obstetrics and Gynaecology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

6 Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom.

7 Institute of Women and Children's Health, School of Life Course and Population Sciences, King's College London, London, United Kingdom

$ Co-first authors

Co-last authors

* Correspondence

Henk Groen, University Medical Center Groningen, University of Groningen, Groningen, 9700

RB, The Netherlands.

Email: h.groen01@umcg.nl

Abstract

Objective: The PIERS-ML model predicts adverse maternal outcomes within 48 hours following admission with preeclampsia. This study evaluated the model's performance employed for consecutive prediction as suggested by National Institute for Health and Care Excellence (UK), using new predictor measurements.

Design: A multi-country prospective observational study.

Setting: Maternity units in the Americas, sub-Saharan Africa, South Asia, Europe, and Oceania.

Population: Women admitted with a diagnosis of preeclampsia.

Methods: The risk differentiation performance of the PIERS-ML model was assessed for each day within a two-week post-admission window, in the model development and internal validation database.

Main Outcome Measures: Trajectory of the mean risk of each of the uncomplicated course and adverse outcome groups, the daily area-under-the-precision-recall-curve (AUC-PRC), clinical impact, dynamic shifts of multiple risk groups, and daily likelihood ratios.

Results: Consistently higher mean risk was observed in the adverse outcome (vs. uncomplicated course) group. The model's AUC-PRC peaked (at around 0.68) in the initial 48 hours, and notably decreased thereafter. When categorising women into multiple risk groups over time, the model generally showed good rule-in capacity for

the 'very high' risk group, and good rule-out capacity for the 'very low' risk group. However, performance declined notably for other risk groups beyond 48 hours post-admission. Decision curve analysis revealed a diminishing relative advantage for treatment guided by the model over time.

Conclusions: Performance drift was detected when the PIERS-ML model was used beyond 48 hours post-admission. For clinical practice, models should be adapted to retain accuracy when deployed consecutively.

Introduction

Of women with preeclampsia 5-20% will develop severe complications, particularly if the syndrome has an early onset.[1, 2] {Garovic, 2020 #6} Severe complications can have a high potential impact and therefore overtreatment, iatrogenic harm from prematurity and increased healthcare costs can result if adequate prediction of adverse outcomes is lacking. Some prediction models for severe complications of early and/or late preeclampsia have been developed.[3-5] The most recent Preeclampsia Integrated Estimate of RiSk-Machine Learning (PIERS-ML) model was trained with data of women from 30 to 38.4 weeks of gestation to predict the risk of the adverse maternal outcomes of preeclampsia within 48 hours following admission.[6] This interval is clinically useful as it reflects the opportunity to arrange in utero transfers, induce labour, and to achieve the full benefit of antenatal corticosteroids for fetal lung maturation.[7]

To be useful during ongoing clinical care, risk prediction needs to be updated regularly to monitor disease risk progression. We are unaware of any models that can accommodate repeated measurements to guide joint decision-making for maternity care providers and women with preeclampsia. Without a better alternative, clinicians could be tempted to use iteratively for consecutive prediction, prediction models that have been developed and validated only using baseline data, as suggested by the UK's National Institute of Health and Care Excellence.[8]

This study was designed to evaluate the performance of the recently developed and validated PIERS-ML model when it is used for consecutive prediction using the latest known measurements of predictors (not their change).

Methods

Data source

For this study we used the PIERS-ML model and the pooled database which was used for the development of the model. The studies comprising the pooled database have been introduced elsewhere,[6] and were the development and validation cohorts for the miniPIERS (low- and middle-income countries) and fullPIERS (high-income countries) models.[9-13] This study was conducted according to the guidelines of the Declaration of Helsinki.

Inclusion and exclusion criteria

We used prospectively-collected data from women with preeclampsia, broadly-defined according to the 2021 International Society for the Study of Hypertension in Pregnancy (ISSHP) criteria,[14] as women presented for initial facility-based assessment at centres with general policies of expectant management of preeclampsia remote from term. No data were used once a woman had developed any component of the combined adverse maternal outcome.

The PIERS combined adverse maternal outcome

The primary study outcome was a composite developed by Delphi consensus,[15] and defined as the first occurrence of one or more of: maternal mortality or severe maternal morbidity (listed in Table S1, Supplementary Appendix), within two days of first assessment for preeclampsia. This combined outcome is similar to, but not completely consistent with, the more recently Delphi-derived iHOPE core maternal

outcome set for women with preeclampsia.[16] Women who did not develop the composite adverse maternal outcome were defined as having an uncomplicated course.

Prediction model

The PIERS-ML model was employed for consecutive daily prediction in this study. The components of the composite outcome and the predictors of this model are shown in Table 1.[6]

Statistical analyses

The predicted probability of the PIERS-ML model may be used to dichotomise women into high/low-risk groups or stratify them into multiple risk groups. Thus, we explored the performance of the model when used for consecutive prediction for both risk stratification scenarios.

The date of first admission with preeclampsia was set as day 0 in this study. Using the latest measurements available each day, we re-calculated the predicted probability of developing adverse maternal outcome within 48 hours based on the PIERS-ML model. This enables us to evaluate the two-week trajectory (i.e. day 0 to day 13) of the predicted probabilities. A duration of two weeks was chosen because over 98% of the adverse maternal outcome events occurred within two weeks, and it was considered a sufficient time window to demonstrate the performance of the model over an extended period.

We plotted the trajectory of the mean predicted probabilities and the number of patients with measurements for the uncomplicated course and adverse outcome groups, respectively. This step was conducted to show the change in the mean predicted probabilities in each group as well as the difference between the two groups.

Use of the PIERS-ML model with patients dichotomised into high versus low-risk

The difference in the mean predicted probabilities between the uncomplicated course versus the adverse outcome group does not reliably indicate the discriminative capacity of the model. Typically, to assess this aspect, a receiver operating characteristic curve (ROC curve) is employed. However, due to the relatively low number of adverse outcomes, the prevalence of negative cases significantly impacts the area under the ROC curve. Consequently, to mitigate this influence, we opted to calculate the area under the precision-recall curve (AUC-PRC) value for each day (day 0 to day 13) to show how the model's discriminative ability changed over time.[17]

Use of the PIERS-ML model with patients stratified into multiple risk groups

Based on the magnitude of positive likelihood ratios (used for the high-risk groups) and negative likelihood ratios (used for the low-risk groups) of the predicted risks, the PIERS-ML model categorises women into five risk groups: very low-risk, low-risk, moderate risk, high-risk and very high-risk with the following cut-off values of the predicted risks: < 0.6%, 0.6% to 3.1% (inclusive), 3.1% to 18.8% (inclusive),

18.8% to 45.6% (inclusive), and > 45.6%. For example, the cut-off value for the very high-risk group was determined by finding the smallest predicted risk that resulted in a positive likelihood ratio greater than 10 when used to dichotomise women.[6] To evaluate the change of the clinical impact of the model over an extended period, a series of decision curve analyses (on day 0, day 4, day 8, and day 13) were carried out. In such analyses, net benefit is the measure of clinical impact, the assessment of which is recommended in the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guideline,[18] and threshold probability (cut-off value) is assumed to be informative of how a clinician or patient weighs the relative harm of a false positive versus false negative prediction.[19] A decision curve plots net benefit on the y-axis against threshold probability on the x-axis, and by default compares the model under study to the "treat all" and "treat none" policies.[19-21] In the context of this study, the "treat all" policy refers to immediate delivery for all patients, whereas "treat none" means expectant management for all patients. The x-axis was limited to the range of 0 to 60%, as this range was considered sufficient to cover the clinically plausible threshold probabilities at which patients or clinicians would opt for intervention in this study, and the threshold probabilities for the risk stratification of the PIERS-ML model, i.e. 0.006, 0.031, 0.188, and 0.456, were marked on the x-axis.

A Sankey diagram was utilised to provide an overview of the dynamic shifts in risk groups and their respective contributions to the endpoint of adverse outcome or uncomplicated course when the model was used for consecutive prediction.

Functioning as a type of flow diagram, the width of "paths" in a Sankey diagram is proportional to the number of subjects.[22]

Lastly, we quantified the daily positive likelihood ratio by applying predefined thresholds for both the high-risk and very high-risk groups. This analysis aimed to illustrate how the model's "rule-in" ability evolves over time. Similarly, we computed the daily negative likelihood ratio using thresholds for both the low-risk and very low-risk groups. This analysis was conducted to demonstrate how the model's "rule-out" capacity changes over time.

Performance of the fullPIERS model deployed for consecutive prediction

The fullPIERS model was developed earlier for a similar purpose of the PIERS-ML model but in the setting of developed countries.[5] It is a far-reaching model with a history of over 10 years. Thus as a secondary analysis we evaluated its performance of dichotomizing patients into "high-risk" and "not high-risk" when deployed for consecutive prediction by quantifying the change of AUC-PRC, plotting the trajectory of mean predicted risks and decision curves in this study. The results can be found in the supplementary file.

Analyses were performed using R Statistical Software (version 4.0.5, R Foundation for Statistical Computing, Vienna, Austria.), R Studio (version 1.4.1106), and the following packages: tidyverse (version 2.0.0), gganimate (version 1.0.8), easyalluvial (version 0.3.1), zoo (version 1.8-11), ggsankey (version 1.0), magrittr (version 2.0.3), epiR (version 2.0.63), pROC (version 1.18.0)[23] and dcurves (0.4.0).

Results

Trajectory of mean predicted probabilities and number of patients with measurements

A consistent decline was seen in the number of patients with measurements from day 0 to day 13 since admission (Figure 1). The mean predicted probability of the adverse outcome (in the next 48 hours) group peaked on day 0, steadily decreased until day 3, and then fluctuated between 0.1 and 0.2. Conversely, the mean predicted probability of the uncomplicated course group was at its lowest on day 0, experienced an upward trend until day 2, and subsequently remained stable. The adverse outcome group consistently exhibited a higher mean predicted probability than the uncomplicated course group.

Area under the precision-recall curve per day

The AUC-PRC values on day 0 and day 1 were around 0.68, respectively (Figure 2). From day 2 onwards almost all the AUC-PRC values were within 0.1-0.5, except for day 12 when the value was less than 0.1. The 95% confidence interval (95% CI) tended to expand from day 0 to day 13 as the number of daily outcome events decreased from over 200 to around 10.

Decision curve analyses

Figure 3 presents the results of decision curve analysis on day 0, day 4, day 8 and day 13, respectively. The dashed vertical lines represent four threshold probabilities from

the PIERS ML model (REF), i.e. 0.006, 0.031, 0.188, 0.456, that stratify patients into very low-risk, low-risk, moderate risk, high-risk and very high-risk groups.

On each day, when the threshold probability was 0.006, treatment guided by the model was as good as treating all patients; when the threshold probability was 0.031, treatment guided by the model only yielded higher benefit than treating all patients on day 0; when the threshold probability was 0.187 or 0.456, treatment guided by the model resulted in higher net benefit compared with treating all patients; however, the relative advantage of treatment guided by the model tended to decrease over time.

Dynamic shifts of risk groups under consecutive prediction

The Sankey diagram (Figure 4) provides a descriptive overview of the dynamic shifts of the five risk groups when the PIERS-ML model was used repeatedly each day. It reveals that patients of moderate risk constituted the largest daily proportion, followed by those in the low-risk, very low-risk, high-risk and very high-risk group, in descending order. In general, the top three highest risk groups were the primary contributors to adverse events, while adverse events were rare in the low-risk group and very low-risk groups.

In the very high-risk group, a significant portion of patients experienced adverse outcomes within 2 days following their admission, and a minority transitioned to other risk categories, primarily the high-risk group. The prevalence of patients predicted to be in the very high-risk category beyond the first day was notably low.

Within the high-risk group, comparable fractions of patients had adverse outcomes or delivery on the next day from admission to day 12; the majority of other patients in this group were retained in the high-risk group on the next day; very few patients were predicted to have high risk after day 8.

Regarding the moderate risk group at admission, nearly half of the patients delivered within 2 days following admission; after day 1, most of the patients remained in the moderate risk group on the next day; and on each day there were similar proportions of patients switching to the high-risk or low-risk group.

For the low-risk group, the majority of the patients continued to be in the low-risk group on the next day; a small subset of patients would switch to the very low-risk group, and others mainly delivered or switched to the moderate risk group. In the case of the very low-risk group, within 2 days following admission around half of the patients had uncomplicated course , nearly one third of them shifted to the low-risk group on the next day and the other patients stayed in the very low-risk group; for the rest of the days, most patients remained in the very low-risk group.

Change of likelihood ratios

Table 2 shows the change of positive/negative likelihood ratios across different risk groups from day 0 to day 13. In the very high-risk group, the positive likelihood ratio consistently exhibited high values ranging from 70.99 to infinity. The only exception to this trend was observed on day 12, when no women were identified as being at very high risk. For the high-risk group, the positive likelihood ratio exceeded 10

during the initial 48 hours. Subsequently, it displayed fluctuations ranging from 2.01 to 7.92. The negative likelihood ratios for the very low-risk group predominantly remained at 0, and for the low-risk group, the negative likelihood ratio was lower than 0.05 on the first 48 hours, then fluctuated between 0.15 and 0.80.

Performance of the fullPIERS model used for consecutive prediction

Within two weeks following admission, the adverse maternal outcome group consistently had a higher mean predicted risk than the uncomplicated course group (Figure S1); the AUC-PRC values remained quite low and mainly fluctuated between 0.1 to 0.4 without a clear pattern of deteriorating (Figure S2); and the decision curve analyses showed that the fullPIERS model guided treatment barely had any advantage over the "treat all" policy even at admission (Figure S3).

Discussion

Main findings

When the PIERS-ML model was used for consecutively dichotomising patients, the performance of the model was best within 48 hours following admission. This finding aligns with the development methodology of PIERS ML which used baseline data from the first 24 hours following admission.[6] When the PIERS-ML model was used for stratifying patients into several risk groups consecutively, generally the model displayed good "rule-in" and "rule-out" capacity for the very high-risk and very low-risk groups. However, the performance of the high-risk and low-risk groups

deteriorated substantially after 48 hours following admission, even while in this analysis the latest available measurements were used.

Interpretation

Consecutive predictions revealed disparity in the trajectories of the mean predicted probabilities in the adverse outcome group and the uncomplicated course group, which can be attributable to various external factors. Temporal dynamics of the disease severity, number of remaining patients and the subsequent treatment effect may play important roles in favour of such trajectory patterns. As depicted in Figure 2, the adverse outcome group experienced a swift occurrence of many adverse events after admission, suggesting severe conditions leading to higher predicted probabilities in the beginning. Subsequently, early interventions in the sickest patients, and less severe conditions in the remaining patients may have resulted in relatively lower predicted probabilities. The rapidly decreasing and small number of patients contributed to the notable variations in the estimated mean predicted probability. Conversely, the uncomplicated course group may have initially presented milder conditions and, thus, lower predicted probabilities. However, their conditions deteriorated later, resulting in an increased mean predicted probability and eventual stabilization due to medical intervention. Furthermore, the model's inability to capture the change or non-change of predictors over time could be a pivotal factor in the relatively steady pattern of the mean predicted probability for both groups after 48 hours following admission.

AUC-PRC exceeded 0.6 within the initial 48 hours but consistently dropped below 0.5 thereafter, indicating poor discriminative capacity. This decline may be attributed to the model's reliance on baseline, rather than repeated, measurements.[6] Nonetheless, it is intriguing why the model loses accuracy over time. An argument may be posited that the change in gestational age over time could potentially signify a clinical state of different severity, given that a longer duration allows for greater progression towards a maternal high-risk scenario, but gestational age was included as a predictor in the model. A plausible explanation lies in the static nature of the PIERS-ML model, which essentially captures a singular "snapshot" of a patient's condition. In contrast, clinical practice involves regular assessments to monitor disease progression. The individual trajectory of predictors may play a major role in patients' outcome.

It is noteworthy that some positive likelihood ratios of the very high risk group were infinite, which probably resulted from small number of women in that group and thus not necessarily indicated optimal discriminative capacity. In general, both the descriptive Sankey diagram and the daily likelihood ratios demonstrated that the very low-risk group and the very high-risk group maintained good "rule-out" and "rule-in" capacity over time. But the very high and low-risk group only accounted for a small proportion of the patients; in contrast, the low-risk group and high-risk group had many more patients and the predictive accuracy decreased substantially after 48 hours following admission. Therefore, the overall performance of the model used for stratifying patients into different risk groups deteriorated over time. The results of

decision curve analyses, which incorporates how we weigh the harm of a false positive prediction against that of a false negative prediction, aligned with previous findings. Moreover, the time and effort needed to gather data for and implement a model are not considered in decision curve analysis. Thus, if predictors of the model take non-trivial efforts, the model would not be considered truly useful if it just brings a slight increase in net benefit.[20] This means the model under study might be not helpful when employed e.g., on day 8 or day 13 considering its small advantage in terms of net benefit. The change of the net benefit of the model could be explained by similar reasons responsible for the trajectories of the mean risk and the change of AUC-PRC values described above.

Performance drift

Clinical periodic assessments are often necessary, rendering consecutive prediction important and potentially valuable. We could not identify any existing that assessed static prediction models being employed for consecutive predictions; thus, it remains unknown whether the alternative static prediction models would outperform PIERS-ML model for this purpose. However, in general the static clinical prediction models may well provide unreliable predictions when population characteristics, clinical practice, disease prevalence or the whole healthcare system changes, which is usually termed as performance/calibration drift and is of growing concern.[24-26] In this study, for example, when the PIERS-ML model was deployed at a later time for consecutive prediction, the characteristics of the remaining patients may well be

different from those in the beginning; moreover, probably the patients had a different treatment plan at a later time, and thus led to less accurate predictions.

Potential solutions

Performance drift suggests necessity for updating the model under study. Updating prediction models, instead of developing another model, is usually preferred as it uses the potentially valuable historical data.[27] The main approaches to updating prediction models include recalibrating intercept or both intercept and joint effects of predictors, merging previous prediction models in a meta-model, and dynamic modelling.[27, 28] Among them dynamic modelling is the most complicated approach.

In this study, dynamic models refer to models using the information of individual repeated measurements (the change or non-change) and offering time-varying predictions.[25] The following methodological frameworks were proposed for harnessing repeated measurements for dynamic prediction: landmark prediction, time-dependent covariate modelling, joint modelling, trajectory classification, machine learning, etc.[29] Landmark prediction and joint modelling are the most commonly employed approaches.[30, 31] To date, there have been attempts to construct dynamic models for predicting preeclampsia,[32-34] but not for the adverse outcomes of preeclampsia.

To minimise the data-action lag, a new trend in prediction models is the so-called living prediction models which intends to update models whenever new data become available instead of waiting until evidence of performance drift

accumulates.[35] Lifelong machine learning (LML)-based method may be one way to achieve this goal. Model performance and the distribution of disease risk features are supervised simultaneously in LML so it can address the performance drift quickly and provide sensible explanation.[36]

Currently, there is no optimal dynamic modelling method . Generally, a large sample size enables more methods, while for smaller subgroups, using a Bayesian method and data of multiple years may yield the best performance.[28]

Strengths and Limitations

All analyses were performed on the model development and internal validation dataset. Ideally an external dataset would be preferred. However, if the most optimistic performance deteriorates considerably, it is strong evidence of performance drift. Some limitations of the PIERS-ML model should be noted as well. The model did not encompass fetal interests which in clinical practice are not fully separable from maternal interests. Moreover, unplanned caesarean deliveries and severity classification were not incorporated in the outcome.

Conclusion

The use of the PIERS-ML model for consecutive prediction cannot be recommended due to performance drift; a dynamic PIERS-ML model is required.

AUTHOR CONTRIBUTIONS

PvD, WG, HG, and LAM conceptualised this project. PvD and LAM obtained the data. TM-C cleaned the data and wrote the R scripts for the analyses under the supervision of KK and PM. TM-C and GY performed the data analyses. GY drafted and addressed feedback on the manuscript; all authors contributed to revisions and approved the submitted version.

ACKNOWLEDGEMENTS

The authors sincerely thank all the participants who contributed to the PIERS-ML dataset, and the relevant co-investigators for their collaborative spirt. Also, we thank Joost Akkermans for his preliminary work on this project.

FUNDING INFORMATION

TM-C is funded by the University of Strathclyde, through the STRADDLE (University of Strathclyde Diversity in Data Linkage) Centre for Doctoral Training. The PIERS datasets were primarily funded by operating grants from the Canadian Institutes of Health Research and the Bill & Melinda Gates Foundation. Funders had no role in the design, analyses, interpretation, or manuscript preparation for this study.

CONFLICT OF INTEREST STATEMENT

All authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Access to the data for this study may be granted upon reasonable negotiation with TM-C.

ETHICS STATEMENT

This study was approved by the NHS Research Ethics Committee (REC reference: 02-03- 033 on 11 March 2003), and conducted according to the guidelines of the Declaration of Helsinki.

References

1. Garovic VD, White WM, Vaughan L, Saiki M, Parashuram S, Garcia-Valencia O, et al. Incidence and Long-Term Outcomes of Hypertensive Disorders of Pregnancy. J Am Coll Cardiol. 2020;75(18):2323-34.
2. Magee LA, Nicolaides KH, von Dadelszen P. Preeclampsia. N Engl J Med. 2022;386(19):1817-32.
3. Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One. 2019;14(8):e0221202.
4. Thangaratinam S, Allotey J, Marlin N, Dodds J, Cheong-See F, von Dadelszen P, et al. Prediction of complications in early-onset preeclampsia (PREP): development and external multinational validation of prognostic models. BMC Med. 2017;15(1):68.
5. von Dadelszen P, Payne B, Li J, Ansermino JM, Pipkin FB, Côté A-M, et al. Prediction of adverse maternal outcomes in preeclampsia: development and validation of the fullPIERS model. The Lancet. 2011;377(9761):219-27.
6. Mongomery-Csoban T, Kavanagh K, Murray P, Robertson C, Barry SJE, Ukah UV, et al. Artificial intelligence-enabled maternal risk assessment in preeclampsia. Lancet Digit Health. [in press].
7. Roberts D, Dalziel S. Antenatal corticosteroids for accelerating fetal lung maturation for women at risk of preterm birth. The Cochrane database of systematic reviews. 2006(3):Cd004454.
8. National Institute for Health and Care Excellence: Guidelines. Hypertension in pregnancy: diagnosis and management. London: National Institute for Health and Care Excellence (NICE)

Copyright © NICE 2019.; 2019.
9. Payne B, Hodgson S, Hutcheon JA, Joseph KS, Li J, Lee T, et al. Performance of the fullPIERS model in predicting adverse maternal outcomes in preeclampsia using patient data from the PIERS (Preeclampsia Integrated Estimate of RiSk) cohort, collected on admission. BJOG. 2013;120(1):113-8.
10. Payne BA, Hutcheon JA, Ansermino JM, Hall DR, Bhutta ZA, Bhutta SZ, et al. A risk prediction model for the assessment and triage of women with hypertensive disorders of pregnancy in low-resourced settings: the miniPIERS (Preeclampsia Integrated Estimate of RiSk) multi-country prospective cohort study. PLoS Med. 2014;11(1):e1001589.
11. Ukah UV, Payne B, Karjalainen H, Kortelainen E, Seed PT, Conti-Ramsden FI, et al. Temporal and external validation of the fullPIERS model for the prediction of adverse maternal outcomes in women with preeclampsia. Pregnancy Hypertens. 2019;15:42-50.
12. Ukah UV, Payne B, Lee T, Magee LA, von Dadelszen P, fullPiers, mini PWG. External validation of the fullPIERS model for predicting adverse maternal

outcomes in pregnancy hypertension in low- and middle-income countries. Hypertension. 2017;69(4):705-11.
13. von Dadelszen P, Payne B, Li J, Ansermino JM, Broughton Pipkin F, Cote AM, et al. Prediction of adverse maternal outcomes in preeclampsia: development and validation of the fullPIERS model. Lancet. 2011;377(9761):219-27.
14. Magee LA, Brown MA, Hall DR, Gupte S, Hennessy A, Karumanchi SA, et al. The hypertensive disorders of pregnancy: the 2021 International Society for the Study of Hypertension in Pregnancy classification, diagnosis and management recommendations for international practice. Pregnancy Hypertens. 2022;27:148-69.
15. Milholland AV, Wheeler SG, Heieck JJ. Medical assessment by a Delphi group opinion technic. N Engl J Med. 1973;288(24):1272-5.
16. Duffy J, Cairns AE, Richards-Doran D, van 't Hooft J, Gale C, Brown M, et al. A core outcome set for preeclampsia research: an international consensus development study. BJOG. 2020;127(12):1516-26.
17. Davis J, Goadrich M, editors. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; 2006.
18. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.
20. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.
21. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol. 2018;74(6):796-804.
22. Sankey diagram [updated 11 July 2023. Available from: https://en.wikipedia.org/wiki/Sankey_diagram.
23. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S plus to analyze and compare ROC curves. Bmc Bioinformatics. 2011;12.
24. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61.
25. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. Diagn Progn Res. 2018;2:23.
26. Parikh RB, Zhang YC, Kolla L, Chivers C, Courtright KR, Zhu JS, et al. Performance drift in a mortality prediction algorithm among patients with cancer during the SARS-CoV-2 pandemic. J Am Med Inform Assn. 2022.

27. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. Stat Methods Med Res. 2018;27(1):185-97.
28. Siregar S, Nieboer D, Vergouwe Y, Versteegh MI, Noyez L, Vonk AB, et al. Improved Prediction by Dynamic Modeling: An Exploratory Study in the Adult Cardiac Surgery Database of the Netherlands Association for Cardio-Thoracic Surgery. Circ Cardiovasc Qual Outcomes. 2016;9(2):171-81.
29. Bull LM, Lunt M, Martin GP, Hyrich K, Sergeant JC. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. Diagn Progn Res. 2020;4:9.
30. Ferrer L, Putter H, Proust-Lima C. Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment. Statistical Methods in Medical Research. 2019;28(12):3649-66.
31. Finelli A, Beer TM, Chowdhury S, Evans CP, Fizazi K, Higano CS, et al. Comparison of Joint and Landmark Modeling for Predicting Cancer Progression in Men With Castration-Resistant Prostate Cancer A Secondary Post Hoc Analysis of the PREVAIL Randomized Clinical Trial. Jama Netw Open. 2021;4(6).
32. Li SL, Wang ZC, Vieira LA, Zheutlin AB, Ru BS, Schadt E, et al. Improving preeclampsia risk prediction by modeling pregnancy trajectories from routinely collected electronic medical record data. Npj Digit Med. 2022;5(1).
33. Zhang XY, Xu Q, Yang L, Sun G, Liu GL, Lian CT, et al. Dynamic risk prediction models for different subtypes of hypertensive disorders in pregnancy. Front Surg. 2022;9.
34. Li ZW, Xu Q, Sun G, Jia RQ, Yang L, Liu GL, et al. Dynamic gestational week prediction model for preeclampsia based on ID3 algorithm. Front Physiol. 2022;13.
35. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, Peek N. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Progn Res. 2021;5(1):1.
36. Chi S, Tian Y, Wang F, Zhou T, Jin S, Li J. A novel lifelong machine learning-based method to eliminate calibration drift in clinical prediction models. Artif Intell Med. 2022;125:102256.

Appendices

Table 1. Predictors and outcome components of the PIERS-ML model

Predictor	Outcome component (N = 1083)[#]
Characteristics: gestational age on admission (weeks), maternal age at expected delivery date (years), height on admission (cm); weight on admission (kg), national maternal mortality ratio (per 100,000 live births), national *per capita* GDP ($USD) **Biomarkers**: Oxygen saturation (%), total leucocyte count; platelet count (10^9/L), uric acid (mmol/L); serum creatinine (µmol/L), aspartate transaminase (U/L), alanine transaminase (U/L), mean platelet volume (fL), haematocrit (%), serum albumin (g/L), systolic blood pressure (mm Hg), diastolic blood pressure (mm Hg)	maternal death (2) **Central nervous system**: eclamptic seizure(s) (75); Glasgow coma score less than 13 (20); stroke or reversible ischaemic neurological deficit (6); transient ischaemic attack (1); cortical blindness (6); posterior reversible encephalopathy (5) **Cardiorespiratory**: positive inotropic support required (7); infusion of a third injectable antihypertensive (33); myocardial ischaemia or infarction (6); oxygen saturation less than 90% (103); at least 50% fractional inspired oxygen for at least one hour (72); intubation other than for Caesarean birth (47); pulmonary oedema (96) **Haematological**: blood transfusion (460); platelet count less than 50x109 per L, without transfusion (111) **Hepatic**: Dysfunction (44); haematoma or rupture (0) **Renal**: acute renal insufficiency in women without chronic kidney disease (9); acute renal failure in women with chronic kidney disease (52); dialysis (11) **Other**: severe ascites (65); Bell's palsy (6); placental abruption (129)

#: Total number of outcome events that occurred at any time following admission

Table 2. Daily positive/negative likelihood ratios calculated with thresholds of different risk groups

days	very high-risk$	high-risk$	low-risk*	very low-risk*
0	70.99 (51.55, 97.77)	16.44 (13.89, 19.46)	0.04 (0.02, 0.08)	0.00
1	173.99 (85.77, 352.95)	11.73 (10.16, 13.53)	0.03 (0.01, 0.07)	0.00
2	196.70 (47.59, 813.01)	7.92 (6.36, 9.88)	0.19 (0.12, 0.31)	0.00
3	174.23 (23.07, 1315.71)	6.68 (5.01, 8.91)	0.29 (0.18, 0.48)	0.54 (0.07, 3.98)
4	Inf	5.99 (4.03, 8.90)	0.45 (0.27, 0.75)	0.00
5	96.18 (11.99, 771.80)	3.62 (2.03, 6.48)	0.54 (0.33, 0.87)	0.00
6	Inf	4.14 (2.25, 7.63)	0.49 (0.27, 0.87)	0.00
7	Inf	4.56 (2.49, 8.37)	0.30 (0.14, 0.64)	1.56 (0.20, 12.12)
8	Inf	3.92 (1.93, 7.99)	0.19 (0.06, 0.56)	0.00
9	Inf	6.00 (2.96, 12.17)	0.15 (0.05, 0.46)	0.00
10	Inf	3.32 (1.30, 8.45)	0.21 (0.08, 0.54)	0.00
11	Inf	5.42 (2.37, 12.38)	0.24 (0.08, 0.71)	0.00
12	NA	2.01 (0.29, 13.98)	0.80 (0.35, 1.82)	0.00
13	Inf	3.63 (0.86, 15.22)	0.49 (0.20, 1.20)	0.00

$: positive likelihood ratio; *: negative likelihood ratio; Inf: infinite; NA: not applicable

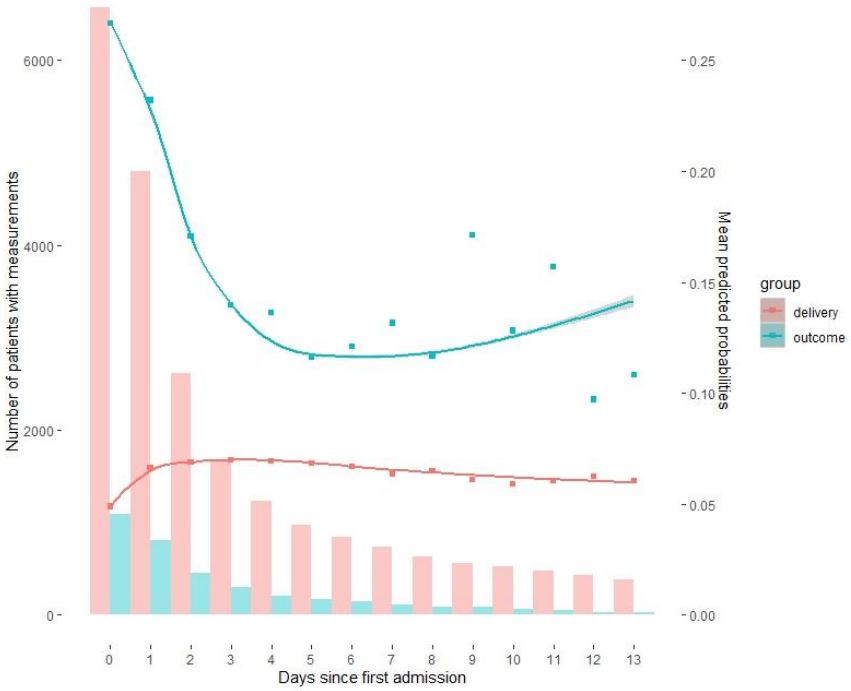

Figure 1. Mean predicted probabilities of complications in the next 48 hours (lines) and number of patients with measurements (bars) in uncomplicated course (pink) and adverse outcome group (blue) per day since admission

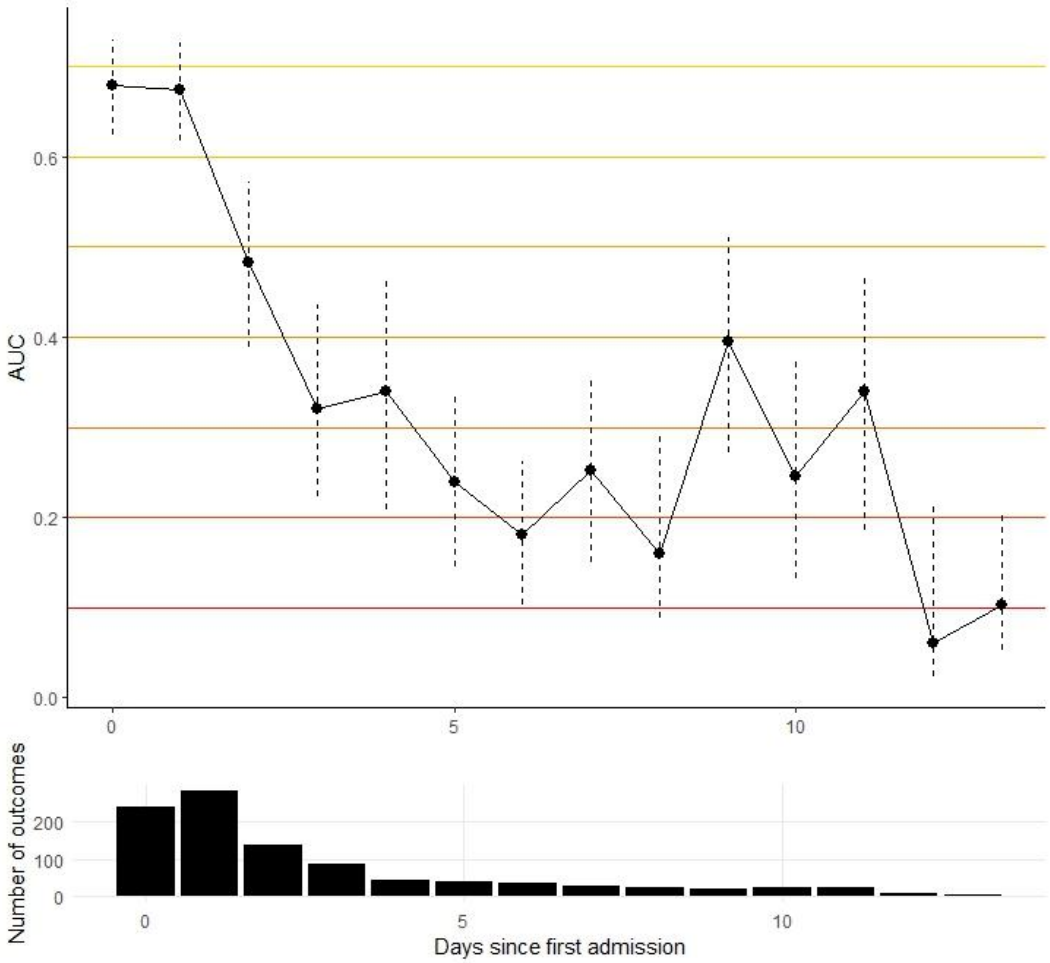

Figure 2. Daily number of adverse outcomes (bars) and area under the precision-recall curve per day. Dashed vertical lines indicate 95% confidence intervals.

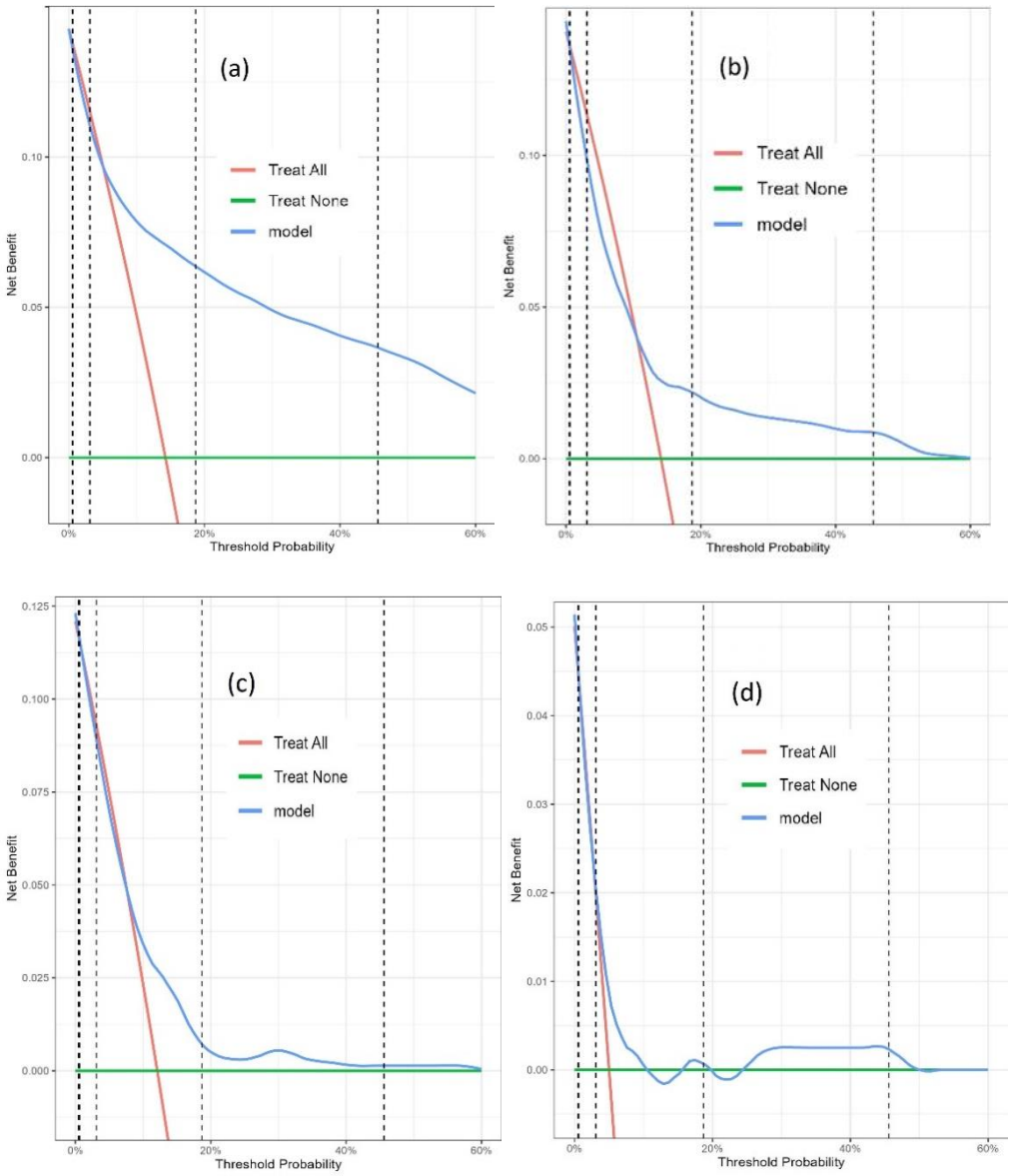

Figure 3. Decision curve analysis on day 0 (a), day 4 (b), day 8 (c) and day 13 (d)

Dashed vertical lines (from left to right) represent four threshold probabilities, i.e. 0.6%, 3.1%, 18.8%, 45.6%, that stratify patients into very low-risk, low-risk, moderate risk, high-risk and very high-risk groups

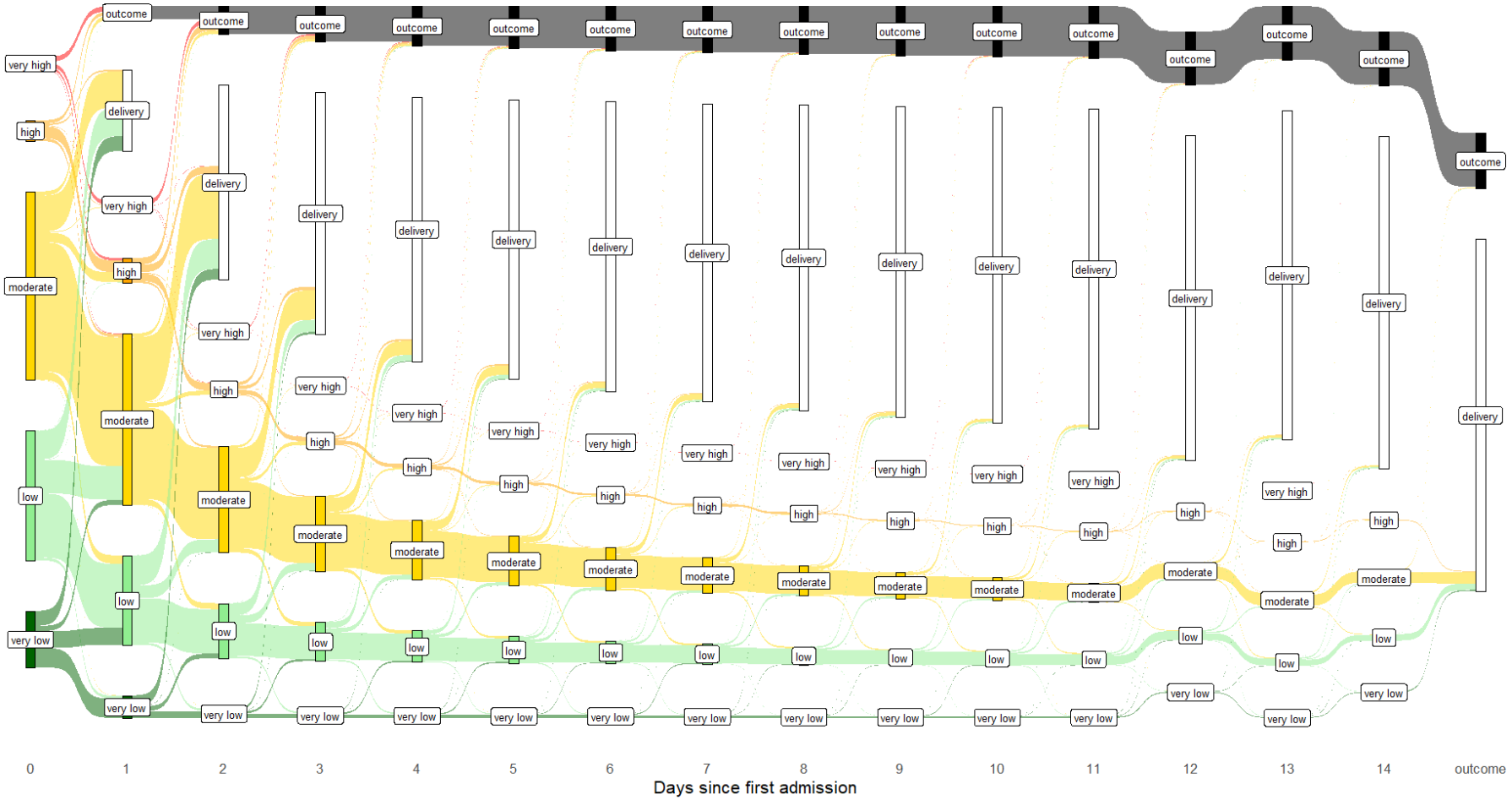

Figure 4. Sankey diagram showing an overview of the dynamic shifts of the five risk groups from day 0 to day 14 after admission

Table S1. Components and definitions of adverse maternal outcomes

Outcome	Definition
Mortality	Maternal death occurring within six weeks of pregnancy or if later, attributable to complications of preeclampsia
Hepatic dysfunction	International normalised ratio (INR) >1.2 in the absence if disseminated intravascular coagulation (DIC) or treatment of warfarin (DIC is defined as having both: abnormal bleeding and consumptive coagulopathy [i.e., low platelets, abnormal peripheral blood film, or one or more of the following: increased INR, increased prothrombin time (PTT), low fibrinogen, of increased fibrin degradation products that are outside normal non-pregnancy ranges])
Hepatic haematoma or rupture	Blood collection under the hepatic capsule as confirmed by ultrasound or laparotomy
Glasgow coma score (GCS) <13	Based on GCS scoring system: Teasdale G, Jennet B. Assessment of coma and impaired consciousness: a practical scale. *Lancet* 1974; **2**:81-83
Stroke	Acute neurological event with deficits lasting longer than 48 hours
Cortical blindness	Loss of visual acuity in the presence of intact papillary response to light
Reversible Ischaemic Neurologic Deficit (RIND)	Cerebral ischaemia lasting longer than 24 hrs but less than 48 hours revealed through clinical examination
Retinal detachment	Separation of the inner layers of the retina from the underlying retinal pigment epithelium (RPE, choroid) and is diagnosed by ophthalmological exam
Acute renal insufficiency	For women with no underlying renal disease, defined as serum creatinine >150 µM
Acute renal failure	For women with an underlying history of renal disease, defined as serum creatinine >200 µM
Dialysis	Including haemodialysis and peritoneal dialysis
Postpartum haemorrhage (PPH) requiring transfusion or hysterectomy	Occurrence of PPH that required transfusion or hysterectomy
Placental abruption	Any occurrence of abruption diagnosed clinically or based on placental pathology report

Platelet count < 50,000 x 10^9/L without blood transfusion	Measurement of platelet count recorded as less than 50,000 x 10^9/L without patient receiving a blood transfusion
Transfusion of blood products	Includes transfusion of any units of blood products: fresh frozen plasma (FFP), platelets, red blood cells (RBCs), cryoprecipitate (cryo) or whole blood
Positive inotropic support	The use of vasopressors to maintain a systolic blood pressure >90 mmHg or mean arterial pressure >70 mmHg
Myocardial ischaemia/infarction	Electrocardiogram (ECG) changes (ST segment elevation or depression) without enzyme changes AND/OR any one of the following: 1) Development of new pathologic Q waves on serial ECGs. The patient may or may not remember previous symptoms. Biochemical markers of myocardial necrosis may have normalised, depending on the length of time that has passed since the infarct developed. 2) Pathological findings of an acute, healed or healing MI 3) Typical rise and gradual fall (troponin) or more rapid rise and fall (CK-MB) of biochemical markers of myocardial necrosis with at least one of the following: a) ischaemic symptoms; b) development of pathologic Q waves on the ECG; c) ECG changes indicative of ischaemia (ST segment elevation or depression); or d) coronary artery intervention (e.g., coronary angioplasty)
Eclampsia	Any episode of seizure antepartum, intrapartum or before postpartum discharge as follow-up beyond discharge is not possible
Require >50% oxygen for greater than one hour	Oxygen given at greater than 50% concentration based on local criteria for longer than 1 hour
Intubation other than for Caesarean section	Intubation by endotracheal tube
Severe breathing difficulty	Suspected pulmonary oedema where x-ray confirmation unavailable may be diagnosed by presence of chest pain or dyspnoea, crackles in the lungs and SaO_2 <90%
Pulmonary oedema	Clinical diagnosis with x-ray confirmation or requirement of diuretic treatment and SaO_2 <95%

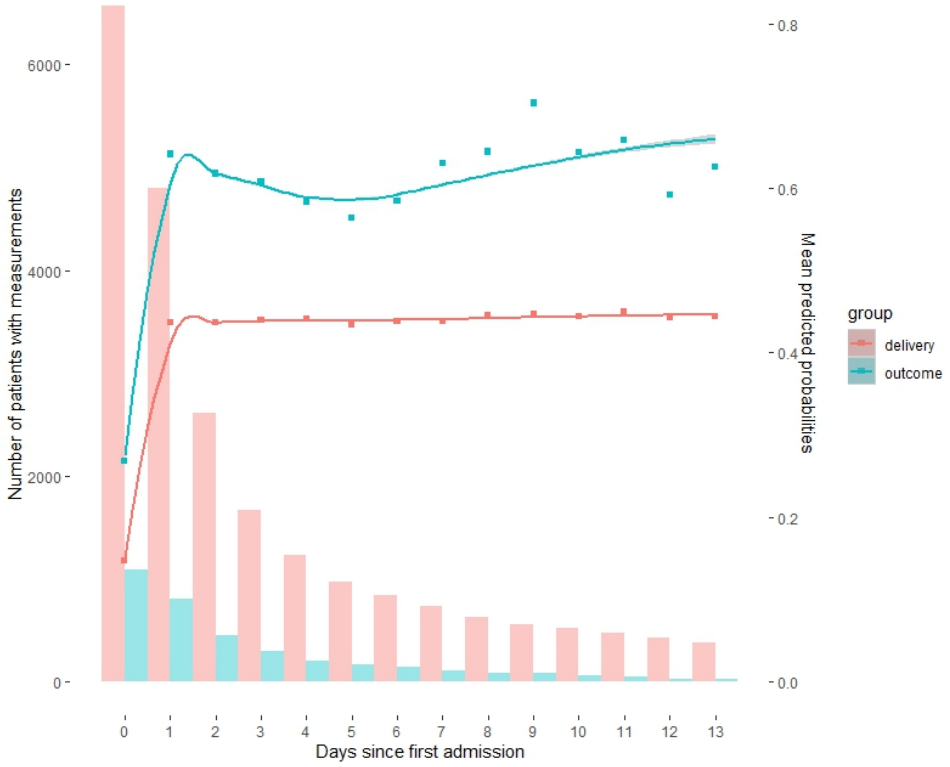

Figure S1. Mean predicted probabilities of complications in the next 48 hours (lines) and number of patients with measurements (bars) in uncomplicated course (pink) and adverse outcome group (blue) per day since admission (fullPIERS model)

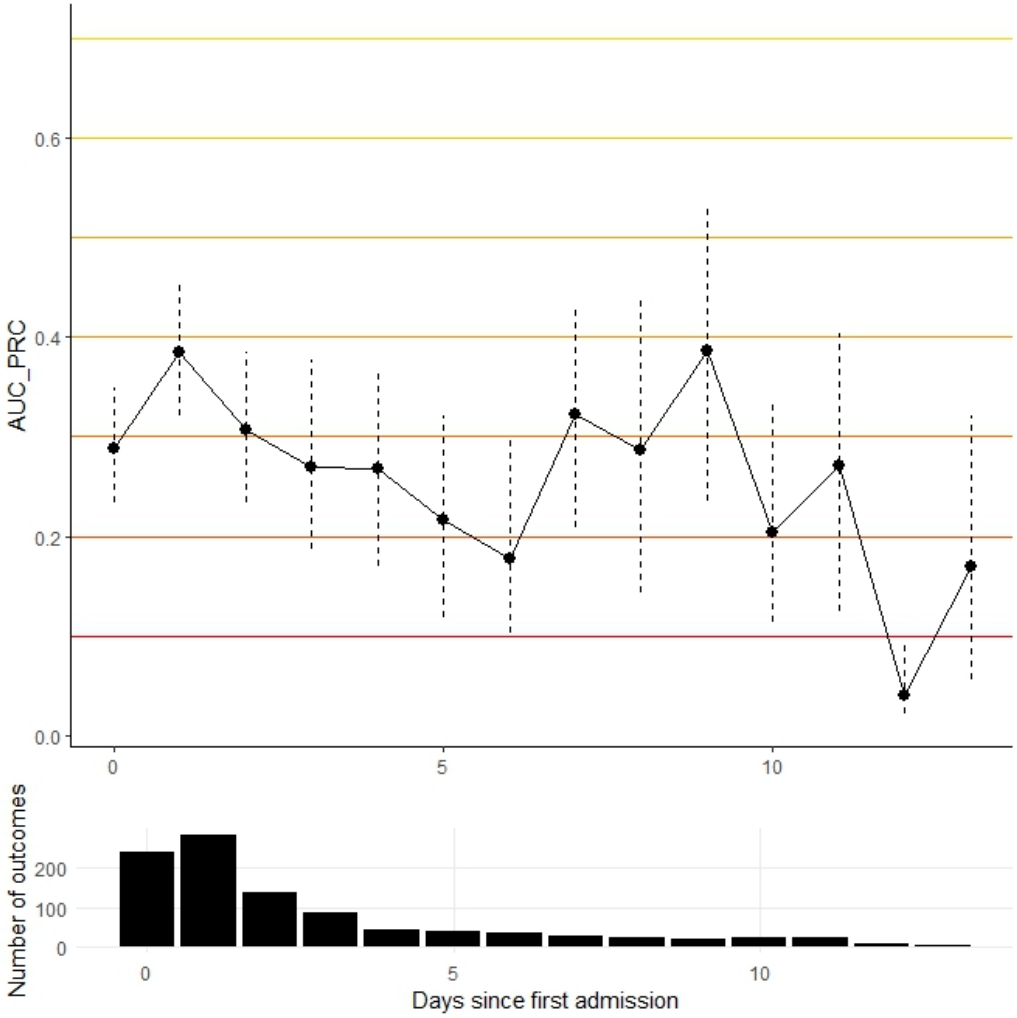

Figure S2. Daily number of adverse outcomes (bars) and area under the precision-recall curve per day (fullPIERS model). Dashed vertical lines indicate 95% confidence intervals.

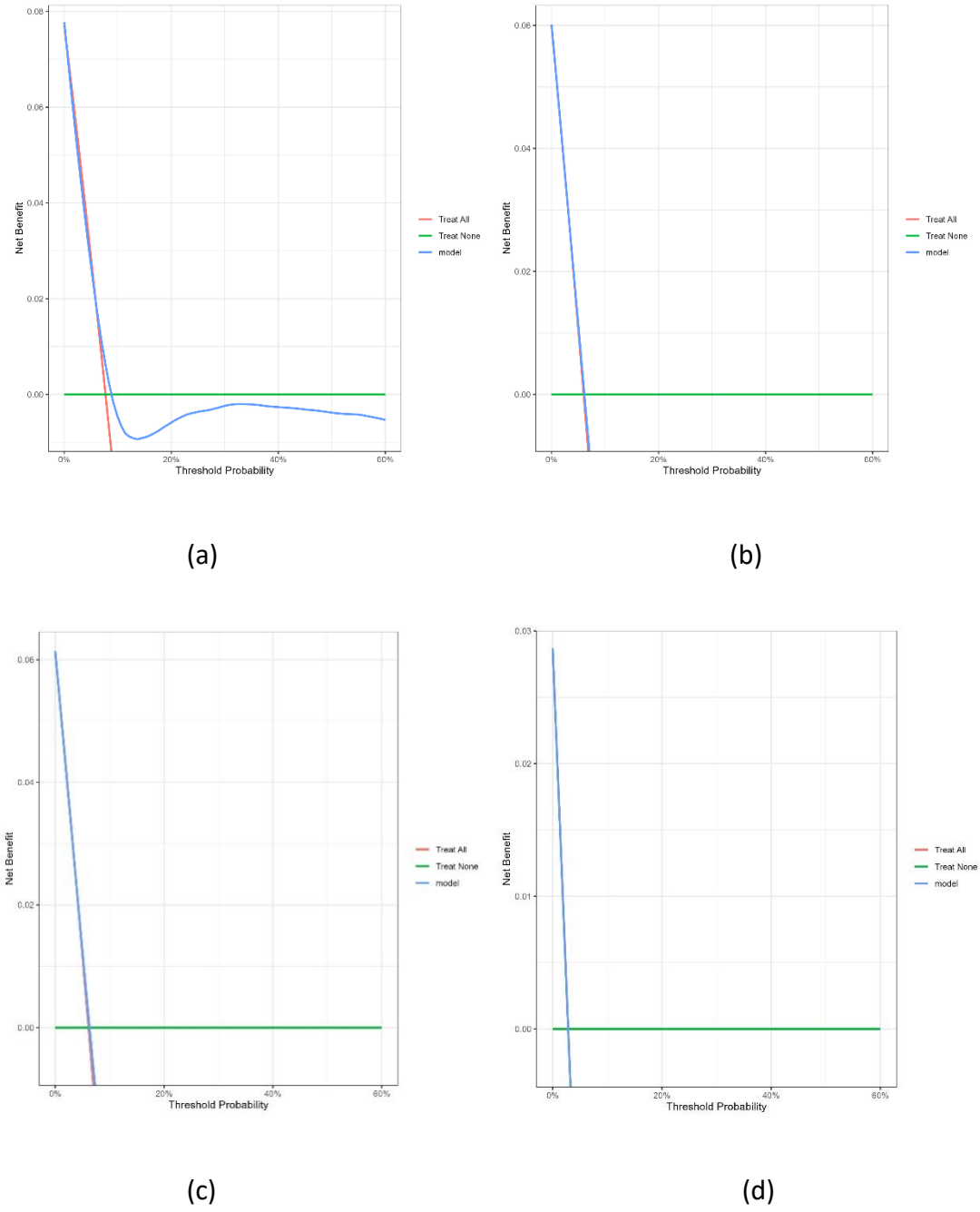

Figure S3. Decision curve analysis on day 0 (a), day 4 (b), day 8 (c) and day 13 (d) (fullPIERS model)

Appendix C: Tables

Table 0.1: Five group classification performance per model

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
R_fP	very high	38.1 - 100%	9 (0.8)	6 (66.7)	LR+ 27.1 [95% CI 6.9, 106.2]	6 (66.7)	LR+ 18.7 [95% CI 4.7, 73.5]	6 (66.7)	LR+ 14.2 [95% CI 3.6, 56.3]
	high	11.9 - 38%	67 (6.1)	15 (22.4)	LR+ 4.2 [95% CI 2.5, 7.1]	19 (28.4)	LR+ 3.9 [95% CI 2.4, 6.4]	20 (29.9)	LR+ 3.2 [95% CI 1.9, 5.2]
	moderate	1 - 11.8%	1028 (93)	55 (5.4)	-	82 (8)	-	110 (10.7)	-
	low	NA	NA	NA	NA	NA	NA	NA	NA

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 1%	1 (0.1)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
RF1	very high	22.8 - 100%	52 (4.7)	25 (48.1)	LR+ 12.5 [95% CI 7.7, 20.5]	27 (51.9)	LR+ 10.1 [95% CI 6.1, 16.7]	30 (57.7)	LR+ 9.7 [95% CI 5.8, 16.3]
	high	13.9 - 22.7%	92 (8.3)	16 (17.4)	LR+ 4.1 [95% CI 2.6, 6.6]	19 (20.7)	LR+ 3.2 [95% CI 2, 5]	20 (21.7)	LR+ 2.5 [95% CI 1.6, 3.9]
	moderate	4.6 - 13.8%	458 (41.4)	26 (5.7)	-	44 (9.6)	-	59 (12.9)	-
	low	1 - 4.5%	478 (43.3)	8 (1.7)	LR- 0.2 [95% CI 0.1, 0.4]	15 (3.1)	LR- 0.3 [95% CI 0.2, 0.5]	25 (5.2)	LR- 0.4 [95% CI 0.3, 0.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.9%	25 (2.3)	1 (4)	LR- 0.6 [95% CI 0.1, 4.1]	2 (8)	LR- 0.8 [95% CI 0.2, 3.4]	2 (8)	LR- 0.6 [95% CI 0.1, 2.6]
RF2	very high	32.6 - 100%	28 (2.5)	18 (64.3)	LR+ 24.4 [95% CI 11.7, 50.9]	19 (67.9)	LR+ 19.7 [95% CI 9.1, 42.4]	20 (71.4)	LR+ 17.8 [95% CI 8, 39.6]
	high	17.5 - 32.5%	86 (7.8)	18 (20.9)	LR+ 4.7 [95% CI 3, 7.3]	20 (23.3)	LR+ 3.4 [95% CI 2.2, 5.3]	22 (25.6)	LR+ 2.8 [95% CI 1.8, 4.4]
	moderate	5.1 - 17.4%	468 (42.4)	29 (6.2)	-	48 (10.3)	-	62 (13.2)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	low	0.7 - 5%	512 (46.3)	11 (2.1)	LR- 0.3 [95% CI 0.2, 0.5]	19 (3.7)	LR- 0.4 [95% CI 0.2, 0.5]	31 (6.1)	LR- 0.5 [95% CI 0.3, 0.6]
	very low	0 - 0.6%	11 (1)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (9.1)	LR- 0.9 [95% CI 0.1, 7.2]	1 (9.1)	LR- 0.7 [95% CI 0.1, 5.5]
RF3	very high	27.7 - 100%	32 (2.9)	19 (59.4)	LR+ 19.8 [95% CI 10.2, 38.5]	20 (62.5)	LR+ 15.5 [95% CI 7.8, 30.9]	21 (65.6)	LR+ 13.6 [95% CI 6.7, 27.6]
	high	14.7 - 27.6%	81 (7.3)	13 (16)	LR+ 3.4 [95% CI 2, 5.8]	16 (19.8)	LR+ 2.8 [95% CI 1.7, 4.6]	19 (23.5)	LR+ 2.6 [95% CI 1.6, 4.1]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	moderate	3 - 14.6%	671 (60.7)	40 (6)	-	60 (8.9)	-	81 (12.1)	-
	low	0.5 - 2.9%	313 (28.3)	4 (1.3)	LR- 0.2 [95% CI 0.1, 0.5]	11 (3.5)	LR- 0.3 [95% CI 0.2, 0.6]	15 (4.8)	LR- 0.4 [95% CI 0.2, 0.6]
	very low	0 - 0.4%	8 (0.7)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
RF4	very high	21.8 - 100%	62 (5.6)	28 (45.2)	LR+ 11.2 [95% CI 7.2, 17.4]	30 (48.4)	LR+ 8.7 [95% CI 5.5, 13.8]	33 (53.2)	LR+ 8.1 [95% CI 5.1, 12.9]
	high	14.3 - 21.7%	81 (7.3)	12 (14.8)	LR+ 3.6 [95% CI 2.1, 6.2]	15 (18.5)	LR+ 2.9 [95% CI 1.7, 4.7]	16 (19.8)	LR+ 2.2 [95% CI 1.4, 3.7]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	moderate	4.6 - 14.2%	458 (41.4)	26 (5.7)	-	43 (9.4)	-	59 (12.9)	-
	low	0.8 - 4.5%	485 (43.9)	10 (2.1)	LR- 0.3 [95% CI 0.2, 0.5]	19 (3.9)	LR- 0.4 [95% CI 0.2, 0.6]	28 (5.8)	LR- 0.4 [95% CI 0.3, 0.6]
	very low	0 - 0.7%	19 (1.7)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
RF5	very high	22.2 - 100%	62 (5.6)	28 (45.2)	LR+ 11.2 [95% CI 7.2, 17.4]	30 (48.4)	LR+ 8.7 [95% CI 5.5, 13.8]	33 (53.2)	LR+ 8.1 [95% CI 5.1, 12.9]
	high	13.8 - 22.1%	87 (7.9)	12 (13.8)	LR+ 3.3 [95% CI 1.9, 5.7]	16 (18.4)	LR+ 2.8 [95% CI 1.7, 4.6]	17 (19.5)	LR+ 2.2 [95% CI 1.4, 3.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	moderate	4.9 - 13.7%	423 (38.3)	25 (5.9)	-	41 (9.7)	-	55 (13)	-
	low	1 - 4.8%	508 (46)	10 (2)	LR- 0.3 [95% CI 0.2, 0.5]	18 (3.5)	LR- 0.3 [95% CI 0.2, 0.5]	29 (5.7)	LR- 0.4 [95% CI 0.3, 0.6]
	very low	0 - 0.9%	25 (2.3)	1 (4)	LR- 0.6 [95% CI 0.1, 4.1]	2 (8)	LR- 0.8 [95% CI 0.2, 3.4]	2 (8)	LR- 0.6 [95% CI 0.1, 2.6]
RF6	very high	32.9 - 100%	21 (1.9)	14 (66.7)	LR+ 27.1 [95% CI 11.3, 65.1]	15 (71.4)	LR+ 23.3 [95% CI 9.2, 58.8]	16 (76.2)	LR+ 22.8 [95% CI 8.5, 61.2]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	high	16.4 - 32.8%	62 (5.6)	11 (17.7)	LR+ 3.6 [95% CI 2, 6.5]	12 (19.4)	LR+ 2.6 [95% CI 1.4, 4.7]	13 (21)	LR+ 2.1 [95% CI 1.2, 3.8]
	moderate	4.7 - 16.3%	461 (41.7)	36 (7.8)	-	51 (11.1)	-	64 (13.9)	-
	low	0.6 - 4.6%	527 (47.7)	14 (2.7)	LR- 0.4 [95% CI 0.2, 0.6]	28 (5.3)	LR- 0.5 [95% CI 0.4, 0.7]	42 (8)	LR- 0.6 [95% CI 0.5, 0.8]
	very low	0 - 0.5%	34 (3.1)	1 (2.9)	LR- 0.4 [95% CI 0.1, 3]	1 (2.9)	LR- 0.3 [95% CI 0, 2]	1 (2.9)	LR- 0.2 [95% CI 0, 1.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
L1	very high	48.8 - 100%	6 (0.5)	4 (66.7)	LR+ 27.1 [95% CI 5, 145.5]	4 (66.7)	LR+ 18.7 [95% CI 3.5, 100.7]	4 (66.7)	LR+ 14.3 [95% CI 2.6, 77.1]
	high	11.1 - 48.7%	140 (12.7)	31 (22.1)	LR+ 4.1 [95% CI 2.9, 5.6]	36 (25.7)	LR+ 3.3 [95% CI 2.4, 4.6]	40 (28.6)	LR+ 2.9 [95% CI 2.1, 4]
	moderate	2.1 - 11%	877 (79.4)	41 (4.7)	-	65 (7.4)	-	88 (10)	-
	low	1.5 - 2%	62 (5.6)	0 (0)	LR- 0 [95% CI 0, NaN]	2 (3.2)	LR- 0.3 [95% CI 0.1, 1.2]	3 (4.8)	LR- 0.4 [95% CI 0.1, 1.1]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 1.4%	20 (1.8)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	1 (5)	LR- 0.4 [95% CI 0.1, 2.8]
L2	very high	9.9 - 100%	20 (1.8)	10 (50)	LR+ 13.5 [95% CI 5.8, 31.5]	10 (50)	LR+ 9.3 [95% CI 4, 21.9]	10 (50)	LR+ 7.1 [95% CI 3, 16.8]
	high	8.3 - 9.8%	25 (2.3)	3 (12)	LR+ 2.1 [95% CI 0.6, 6.9]	4 (16)	LR+ 1.9 [95% CI 0.7, 5.5]	4 (16)	LR+ 1.4 [95% CI 0.5, 4.2]
	moderate	0 - 8.2%	1060 (95.9)	63 (5.9)	-	93 (8.8)	-	122 (11.5)	-
	low	NA	NA	NA	NA	NA	NA	NA	NA

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	NA	NA	NA	NA	NA	NA	NA	NA
R1	very high	33.5 - 100%	7 (0.6)	5 (71.4)	LR+ 33.8 [95% CI 6.7, 171.6]	5 (71.4)	LR+ 23.3 [95% CI 4.6, 118.7]	5 (71.4)	LR+ 17.8 [95% CI 3.5, 90.9]
	high	10.9 - 33.4%	118 (10.7)	29 (24.6)	LR+ 4.7 [95% CI 3.3, 6.6]	31 (26.3)	LR+ 3.5 [95% CI 2.4, 5]	34 (28.8)	LR+ 3 [95% CI 2.1, 4.3]
	moderate	3.2 - 10.8%	870 (78.7)	41 (4.7)	-	69 (7.9)	-	93 (10.7)	-
	low	2.4 - 3.1%	95 (8.6)	1 (1.1)	LR- 0.1 [95% CI 0, 1]	2 (2.1)	LR- 0.2 [95% CI 0, 0.8]	3 (3.2)	LR- 0.2 [95% CI 0.1, 0.7]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 2.3%	15 (1.4)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	1 (6.7)	LR- 0.5 [95% CI 0.1, 3.8]
NN1	very high	47.2 - 100%	2 (0.2)	1 (50)	LR+ 13.5 [95% CI 0.9, 214.3]	1 (50)	LR+ 9.3 [95% CI 0.6, 148.1]	1 (50)	LR+ 7.1 [95% CI 0.4, 113.3]
	high	15.1 - 47.1%	76 (6.9)	19 (25)	LR+ 4.6 [95% CI 2.9, 7.3]	21 (27.6)	LR+ 3.6 [95% CI 2.3, 5.7]	22 (28.9)	LR+ 2.9 [95% CI 1.8, 4.6]
	moderate	0 - 15%	1027 (92.9)	56 (5.5)	-	85 (8.3)	-	113 (11)	-
	low	NA	NA	NA	NA	NA	NA	NA	NA

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	NA	NA	NA	NA	NA	NA	NA	NA
NN2	very high	45.2 - 100%	4 (0.4)	2 (50)	LR+ 13.5 [95% CI 1.9, 94.8]	2 (50)	LR+ 9.3 [95% CI 1.3, 65.5]	2 (50)	LR+ 7.1 [95% CI 1, 50.2]
	high	15.9 - 45.1%	67 (6.1)	18 (26.9)	LR+ 5.1 [95% CI 3.1, 8.3]	20 (29.9)	LR+ 4 [95% CI 2.5, 6.5]	21 (31.3)	LR+ 3.3 [95% CI 2, 5.3]
	moderate	0 - 15.8%	1034 (93.6)	56 (5.4)	-	85 (8.2)	-	113 (10.9)	-
	low	NA	NA	NA	NA	NA	NA	NA	NA
	very low	NA	NA	NA	NA	NA	NA	NA	NA

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
NN3	very high	32.9 - 100%	20 (1.8)	10 (50)	LR+ 13.5 [95% CI 5.8, 31.5]	11 (55)	LR+ 11.4 [95% CI 4.8, 26.9]	12 (60)	LR+ 10.7 [95% CI 4.4, 25.7]
	high	14 - 32.8%	64 (5.8)	13 (20.3)	LR+ 3.9 [95% CI 2.3, 6.9]	13 (20.3)	LR+ 2.6 [95% CI 1.5, 4.7]	13 (20.3)	LR+ 2 [95% CI 1.1, 3.5]
	moderate	2.7 - 13.9%	944 (85.4)	53 (5.6)	-	83 (8.8)	-	111 (11.8)	-
	low	2.6 - 2.6%	55 (5)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
	very low	0 - 2.5%	22 (2)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
BMA	very high	22.9 - 100%	28 (2.5)	14 (50)	LR+ 13.5 [95% CI 6.7, 27.4]	14 (50)	LR+ 9.3 [95% CI 4.6, 19]	14 (50)	LR+ 7.1 [95% CI 3.5, 14.6]
	high	12.6 - 22.8%	75 (6.8)	13 (17.3)	LR+ 3.4 [95% CI 2, 5.9]	15 (20)	LR+ 2.6 [95% CI 1.6, 4.5]	18 (24)	LR+ 2.5 [95% CI 1.5, 4.1]
	moderate	2.4 - 12.5%	924 (83.6)	48 (5.2)	-	76 (8.2)	-	99 (10.7)	-
	low	2 - 2.3%	45 (4.1)	1 (2.2)	LR- 0.3 [95% CI 0, 2.1]	2 (4.4)	LR- 0.4 [95% CI 0.1, 1.7]	3 (6.7)	LR- 0.5 [95% CI 0.2, 1.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 1.9%	33 (3)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	2 (6.1)	LR- 0.5 [95% CI 0.1, 1.9]
XGB1	very high	17.3 - 100%	80 (7.2)	31 (38.8)	LR+ 8.6 [95% CI 5.8, 12.6]	34 (42.5)	LR+ 6.9 [95% CI 4.6, 10.2]	37 (46.2)	LR+ 6.1 [95% CI 4.1, 9.2]
	high	10.9 - 17.2%	85 (7.7)	12 (14.1)	LR+ 3.6 [95% CI 2.1, 6.1]	15 (17.6)	LR+ 2.8 [95% CI 1.7, 4.6]	15 (17.6)	LR+ 2 [95% CI 1.2, 3.4]
	moderate	4.4 - 10.8%	359 (32.5)	20 (5.6)	-	36 (10)	-	49 (13.6)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	low	1.4 - 4.3%	575 (52)	13 (2.3)	LR- 0.3 [95% CI 0.2, 0.5]	22 (3.8)	LR- 0.4 [95% CI 0.3, 0.5]	35 (6.1)	LR- 0.5 [95% CI 0.3, 0.6]
	very low	0 - 1.3%	6 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
XGB2	very high	23.9 - 100%	46 (4.2)	20 (43.5)	LR+ 10.4 [95% CI 6.1, 17.8]	22 (47.8)	LR+ 8.5 [95% CI 5, 14.7]	24 (52.2)	LR+ 7.8 [95% CI 4.5, 13.5]
	high	12.3 - 23.8%	94 (8.5)	19 (20.2)	LR+ 4.5 [95% CI 3, 6.9]	23 (24.5)	LR+ 3.7 [95% CI 2.5, 5.6]	24 (25.5)	LR+ 2.9 [95% CI 1.9, 4.4]
	moderate	3 - 12.2%	631 (57.1)	31 (4.9)	-	51 (8.1)	-	70 (11.1)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	low	1.5 - 2.9%	327 (29.6)	6 (1.8)	LR- 0.3 [95% CI 0.1, 0.5]	11 (3.4)	LR- 0.3 [95% CI 0.2, 0.6]	18 (5.5)	LR- 0.4 [95% CI 0.3, 0.6]
	very low	0 - 1.4%	7 (0.6)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]

Table 0.2: Five group classification performance of sensitivity analysis models

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
RF1_s1	very high	21.4 - 100%	56 (5.1)	24 (42.9)	LR+ 10.2 [95% CI 6.3, 16.3]	26 (46.4)	LR+ 8.1 [95% CI 5, 13.1]	29 (51.8)	LR+ 7.7 [95% CI 4.7, 12.5]
	high	12.8 - 21.3%	91 (8.2)	16 (17.6)	LR+ 4.1 [95% CI 2.6, 6.5]	20 (22)	LR+ 3.4 [95% CI 2.2, 5.2]	21 (23.1)	LR+ 2.6 [95% CI 1.7, 4.1]
	moderate	4.7 - 12.7%	419 (37.9)	24 (5.7)	-	40 (9.5)	-	52 (12.4)	-
	low	0.9 - 4.6%	518 (46.9)	12 (2.3)	LR- 0.3 [95% CI 0.2, 0.5]	20 (3.9)	LR- 0.4 [95% CI 0.2, 0.6]	33 (6.4)	LR- 0.5 [95% CI 0.4, 0.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.8%	21 (1.9)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (4.8)	LR- 0.5 [95% CI 0.1, 3.4]	1 (4.8)	LR- 0.4 [95% CI 0, 2.6]
RF2_s1	very high	45.6 - 100%	11 (1)	10 (90.9)	LR+ 135.4 [95% CI 17.6, 1043.8]	10 (90.9)	LR+ 93.3 [95% CI 12.1, 721.6]	10 (90.9)	LR+ 71.2 [95% CI 9.2, 552.2]
	high	18.8 - 45.5%	85 (7.7)	23 (27.1)	LR+ 5.8 [95% CI 3.8, 8.7]	25 (29.4)	LR+ 4.3 [95% CI 2.8, 6.5]	28 (32.9)	LR+ 3.8 [95% CI 2.5, 5.7]
	moderate	3.1 - 18.7%	680 (61.5)	36 (5.3)	-	58 (8.5)	-	76 (11.2)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	low	0.6 - 3%	321 (29)	7 (2.2)	LR- 0.3 [95% CI 0.1, 0.6]	13 (4)	LR- 0.4 [95% CI 0.2, 0.7]	21 (6.5)	LR- 0.5 [95% CI 0.3, 0.7]
	very low	0 - 0.5%	8 (0.7)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (12.5)	LR- 1.3 [95% CI 0.2, 10.7]	1 (12.5)	LR- 1 [95% CI 0.1, 8.
RF3_s1	very high	29.1 - 100%	28 (2.5)	16 (57.1)	LR+ 18.1 [95% CI 8.9, 36.8]	17 (60.7)	LR+ 14.4 [95% CI 6.9, 30]	18 (64.3)	LR+ 12.8 [95% CI 6, 27.2]
	high	12.9 - 29%	100 (9)	19 (19)	LR+ 4 [95% CI 2.6, 6.1]	22 (22)	LR+ 3.1 [95% CI 2, 4.7]	25 (25)	LR+ 2.7 [95% CI 1.8, 4.1]
	moderate	3.2 - 12.8%	605 (54.8)	35 (5.8)	-	53 (8.8)	-	70 (11.6)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	low	0.5 - 3.1%	361 (32.7)	5 (1.4)	LR- 0.2 [95% CI 0.1, 0.4]	14 (3.9)	LR- 0.4 [95% CI 0.2, 0.6]	22 (6.1)	LR- 0.5 [95% CI 0.3, 0.7]
	very low	0 - 0.4%	11 (1)	1 (9.1)	LR- 1.4 [95% CI 0.2, 10.4]	1 (9.1)	LR- 0.9 [95% CI 0.1, 7.2]	1 (9.1)	LR- 0.7 [95% CI 0.1, 5.5]
RF4_s1	very high	22.9 - 100%	53 (4.8)	24 (45.3)	LR+ 11.2 [95% CI 6.9, 18.3]	26 (49.1)	LR+ 9 [95% CI 5.4, 14.8]	29 (54.7)	LR+ 8.6 [95% CI 5.2, 14.3]
	high	12.9 - 22.8%	101 (9.1)	17 (16.8)	LR+ 3.9 [95% CI 2.5, 6]	21 (20.8)	LR+ 3.1 [95% CI 2.1, 4.8]	22 (21.8)	LR+ 2.5 [95% CI 1.6, 3.8]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	moderate	4.8 - 12.8%	406 (36.7)	23 (5.7)	-	39 (9.6)	-	51 (12.6)	-
	low	0.8 - 4.7%	527 (47.7)	12 (2.3)	LR- 0.3 [95% CI 0.2, 0.5]	19 (3.6)	LR- 0.3 [95% CI 0.2, 0.5]	32 (6.1)	LR- 0.5 [95% CI 0.3, 0.6]
	very low	0 - 0.7%	18 (1.6)	0 (0)	LR- 0 [95% CI 0, NaN]	2 (11.1)	LR- 1.2 [95% CI 0.3, 5]	2 (11.1)	LR- 0.9 [95% CI 0.2, 3.8]
RF5_s1	very high	21.6 - 100%	56 (5.1)	25 (44.6)	LR+ 10.9 [95% CI 6.8, 17.5]	27 (48.2)	LR+ 8.7 [95% CI 5.3, 14.1]	30 (53.6)	LR+ 8.2 [95% CI 5, 13.5]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	high	12.9 - 21.5%	96 (8.7)	16 (16.7)	LR+ 3.9 [95% CI 2.5, 6.2]	19 (19.8)	LR+ 3 [95% CI 1.9, 4.7]	20 (20.8)	LR+ 2.3 [95% CI 1.5, 3.7]
	moderate	4.8 - 12.8%	409 (37)	23 (5.6)	-	40 (9.8)	-	52 (12.7)	-
	low	0.8 - 4.7%	530 (48)	12 (2.3)	LR- 0.3 [95% CI 0.2, 0.5]	20 (3.8)	LR- 0.4 [95% CI 0.2, 0.5]	33 (6.2)	LR- 0.5 [95% CI 0.3, 0.6]
	very low	0 - 0.7%	14 (1.3)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (7.1)	LR- 0.7 [95% CI 0.1, 5.4]	1 (7.1)	LR- 0.5 [95% CI 0.1, 4.2]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
RF2_s2	very high	27.2 - 100%	34 (3.1)	20 (58.8)	LR+ 19.3 [95% CI 10.2, 36.8]	22 (64.7)	LR+ 17.1 [95% CI 8.7, 33.6]	24 (70.6)	LR+ 17.1 [95% CI 8.4, 35]
	high	16.3 - 27.1%	97 (8.8)	21 (21.6)	LR+ 5 [95% CI 3.4, 7.5]	23 (23.7)	LR+ 3.6 [95% CI 2.4, 5.4]	25 (25.8)	LR+ 3 [95% CI 2, 4.5]
	moderate	4 - 16.2%	527 (47.7)	25 (4.7)	-	43 (8.2)	-	56 (10.6)	-
	low	0.9 - 3.9%	405 (36.7)	9 (2.2)	LR- 0.3 [95% CI 0.2, 0.6]	16 (4)	LR- 0.4 [95% CI 0.2, 0.6]	28 (6.9)	LR- 0.5 [95% CI 0.4, 0.7]
	very low	0 - 0.8%	42 (3.8)	1 (2.4)	LR- 0.3 [95% CI 0, 2.4]	3 (7.1)	LR- 0.7 [95% CI 0.2, 2.3]	3 (7.1)	LR- 0.5 [95% CI 0.2, 1.7]

Table 0.3: Final model performance on secondary analyses

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
Occurrence of eclampsia in the whole internal dataset [AUROC 0.94, AUPRC 0.06]	very high	45.6 - 100%	238 (2.7)	15 (6.3)	LR+ 12.1 [95% CI 7.8, 18.8]	16 (6.7)	LR+ 10 [95% CI 6.5, 15.6]	16 (6.7)	LR+ 8.4 [95% CI 5.4, 13.3]
	high	18.7 - 45.5%	412 (4.7)	25 (6.1)	LR+ 16.3 [95% CI 13, 20.4]	25 (6.1)	LR+ 11.8 [95% CI 8.8, 15.6]	26 (6.3)	LR+ 9.8 [95% CI 7.2, 13.2]
	moderate	3.1 - 18.6%	4268 (48.3)	9 (0.2)	-	17 (0.4)	-	24 (0.6)	-
	low	0.6 - 3%	3501 (39.6)	0 (0)	LR- 0 [95% CI 0, NaN]	5 (0.1)	LR- 0.2 [95% CI 0.1, 0.4]	9 (0.3)	LR- 0.3 [95% CI 0.2, 0.5]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.5%	424 (4.8)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
Occurrence of eclampsia in the testing and internal validation datasets [AUROC	very high	45.6 - 100%	16 (0.7)	1 (6.2)	LR+ 10.5 [95% CI 1.5, 73.9]	1 (6.2)	LR+ 8.1 [95% CI 1.1, 58.2]	1 (6.2)	LR+ 7.3 [95% CI 1, 52.6]
	high	18.7 - 45.5%	147 (6.7)	5 (3.4)	LR+ 5.9 [95% CI 2.9, 12]	5 (3.4)	LR+ 4.5 [95% CI 2.1, 9.6]	5 (3.4)	LR+ 4 [95% CI 1.9, 8.7]
	moderate	3.1 - 18.6%	1340 (60.6)	8 (0.6)	-	12 (0.9)	-	13 (1)	-
	low	0.6 - 3%	684 (31)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	1 (0.1)	LR- 0.2 [95% CI 0, 1.1]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
0.84, AUPRC 0.04]	very low	0 - 0.5%	23 (1)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
Occurrence of eclampsia in the internal validation dataset [AUROC 0.89, AUPRC 0.05]	very high	45.6 - 100%	11 (1)	1 (9.1)	LR+ 18.3 [95% CI 2.8, 121.6]	1 (9.1)	LR+ 18.3 [95% CI 2.8, 121.6]	1 (9.1)	LR+ 13.7 [95% CI 2, 94.9]
	high	18.7 - 45.5%	87 (7.9)	2 (2.3)	LR+ 5.1 [95% CI 1.7, 15.3]	2 (2.3)	LR+ 5.1 [95% CI 1.7, 15.3]	2 (2.3)	LR+ 3.7 [95% CI 1.1, 12]
	moderate	3.1 - 18.6%	678 (61.4)	3 (0.4)	-	3 (0.4)	-	4 (0.6)	-
	low	0.6 - 3%	321 (29)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	1 (0.3)	LR- 0.4 [95% CI 0.1, 2.7]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.5%	8 (0.7)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
Occurrence of stillbirth in the testing and internal validation datasets where birth outcome	very high	45.6 - 100%	15 (0.8)	0 (0)	LR+ 0 [95% CI 0, NaN]	0 (0)	LR+ 0 [95% CI 0, NaN]	0 (0)	LR+ 0 [95% CI 0, NaN]
	high	18.7 - 45.5%	138 (7.2)	16 (11.6)	LR+ 5 [95% CI 3.2, 7.7]	16 (11.6)	LR+ 5 [95% CI 3.2, 7.7]	16 (11.6)	LR+ 4.9 [95% CI 3.1, 7.6]
	moderate	3.1 - 18.6%	1148 (59.6)	32 (2.8)	-	32 (2.8)	-	33 (2.9)	-
	low	0.6 - 3%	601 (31.2)	1 (0.2)	LR- 0.1 [95% CI 0, 0.4]	1 (0.2)	LR- 0.1 [95% CI 0, 0.4]	1 (0.2)	LR- 0.1 [95% CI 0, 0.4]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
was recorded (N=1925) [AUROC 0.81, AUPRC 0.09]	very low	0 - 0.5%	23 (1.2)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]

Table 0.4: Five group classification performance of the final PIERS-ML, the original and refitted fullPIERS models on complete case and mean imputed internal validation datasets

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
Original fullPIERS model on complete case internal validation [AUROC 0.76,	very high	45.6 - 100%	14 (4)	5 (35.7)	LR+ 12.4 [95% CI 4.8, 32.6]	5 (35.7)	LR+ 9.7 [95% CI 3.6, 26.1]	5 (35.7)	LR+ 7.9 [95% CI 2.9, 21.7]
	high	18.7 - 45.5%	56 (16)	2 (3.6)	LR+ 1.2 [95% CI 0.3, 4.3]	2 (3.6)	LR+ 0.9 [95% CI 0.2, 3.2]	3 (5.4)	LR+ 1 [95% CI 0.3, 2.9]
	moderate	3.1 - 18.6%	276 (78.6)	8 (2.9)	-	12 (4.3)	-	15 (5.4)	-
	low	0.6 - 3%	3 (0.9)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
AUPRC 0.33]	very low	0 - 0.5%	2 (0.6)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
Refitted fullPIERS model on complete case internal validation [AUROC 0.64,	very high	45.6 - 100%	3 (0.9)	2 (66.7)	LR+ 44.8 [95% CI 4.3, 467]	2 (66.7)	LR+ 34.9 [95% CI 3.3, 368.5]	2 (66.7)	LR+ 28.5 [95% CI 2.7, 303]
	high	18.7 - 45.5%	13 (3.7)	3 (23.1)	LR+ 7.7 [95% CI 2.4, 24.8]	3 (23.1)	LR+ 5.8 [95% CI 1.8, 19.3]	3 (23.1)	LR+ 4.7 [95% CI 1.4, 15.7]
	moderate	3.1 - 18.6%	318 (90.6)	9 (2.8)	-	13 (4.1)	-	17 (5.3)	-
	low	0.6 - 3%	17 (4.8)	1 (5.9)	LR- 1.4 [95% CI 0.2, 9.9]	1 (5.9)	LR- 1.1 [95% CI 0.2, 7.8]	1 (5.9)	LR- 0.9 [95% CI 0.1, 6.4]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
AURPC 0.26]	very low	0 - 0.5%	0 (0)	NA	NA	NA	NA	NA	NA
PIERS-ML model on complete case internal validation [AUROC 0.62, AURPC 0.32]	very high	45.6 - 100%	3 (0.9)	3 (100)	LR+ Inf [95% CI NaN, Inf]	3 (100)	LR+ Inf [95% CI NaN, Inf]	3 (100)	LR+ Inf [95% CI NaN, Inf]
	high	18.7 - 45.5%	12 (3.4)	1 (8.3)	LR+ 2.5 [95% CI 0.4, 18.2]	1 (8.3)	LR+ 1.9 [95% CI 0.3, 13.7]	1 (8.3)	LR+ 1.5 [95% CI 0.2, 11]
	moderate	3.1 - 18.6%	172 (49)	6 (3.5)	-	9 (5.2)	-	12 (7)	-
	low	0.6 - 3%	151 (43)	5 (3.3)	LR- 0.7 [95% CI 0.4, 1.5]	5 (3.3)	LR- 0.6 [95% CI 0.3, 1.3]	6 (4)	LR- 0.6 [95% CI 0.3, 1.2]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.5%	13 (3.7)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (7.7)	LR- 1.5 [95% CI 0.2, 10.6]	1 (7.7)	LR- 1.2 [95% CI 0.2, 8.7]
Original fullPIERS model on mean imputed internal validation [AUROC 0.62,	very high	45.6 - 100%	81 (3.7)	31 (38.3)	LR+ 9.2 [95% CI 6.1, 13.9]	34 (42)	LR+ 7.4 [95% CI 4.9, 11.2]	37 (45.7)	LR+ 6.4 [95% CI 4.2, 9.7]
	high	18.7 - 45.5%	286 (12.9)	24 (8.4)	LR+ 1.7 [95% CI 1.2, 2.5]	30 (10.5)	LR+ 1.4 [95% CI 1, 2]	40 (14)	LR+ 1.4 [95% CI 1, 1.9]
	moderate	3.1 - 18.6%	1794 (81.2)	76 (4.2)	-	123 (6.9)	-	170 (9.5)	-
	low	0.6 - 3%	45 (2)	9 (20)	LR- 3.7 [95% CI 1.8, 7.5]	10 (22.2)	LR- 2.9 [95% CI 1.5, 5.8]	10 (22.2)	LR- 2.2 [95% CI 1.1, 4.3]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
AUPRC 0.24]	very low	0 - 0.5%	4 (0.2)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
Refitted fullPIERS model on mean imputed internal validation [AUROC 0.68,	very high	45.6 - 100%	14 (0.6)	11 (78.6)	LR+ 54.2 [95% CI 15.3, 192.1]	11 (78.6)	LR+ 37.5 [95% CI 10.5, 133.2]	11 (78.6)	LR+ 27.9 [95% CI 7.8, 99.2]
	high	18.7 - 45.5%	40 (1.8)	13 (32.5)	LR+ 7.7 [95% CI 4.1, 14.6]	14 (35)	LR+ 5.8 [95% CI 3.1, 10.9]	14 (35)	LR+ 4.3 [95% CI 2.3, 8.1]
	moderate	3.1 - 18.6%	2026 (91.7)	111 (5.5)	-	165 (8.1)	-	222 (11)	-
	low	0.6 - 3%	130 (5.9)	5 (3.8)	LR- 0.6 [95% CI 0.2, 1.4]	7 (5.4)	LR- 0.6 [95% CI 0.3, 1.2]	10 (7.7)	LR- 0.6 [95% CI 0.3, 1.2]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
AUPRC 0.24]	very low	0 - 0.5%	NA	NA	NA	NA	NA	NA	NA
PIERS-ML model on mean imputed internal validation [AUROC 0.79, AUPRC 0.38]	very high	45.6 - 100%	16 (0.7)	16 (100)	LR+ Inf [95% CI NaN, Inf]	16 (100)	LR+ Inf [95% CI NaN, Inf]	16 (100)	LR+ Inf [95% CI NaN, Inf]
	high	18.7 - 45.5%	122 (5.5)	41 (33.6)	LR+ 8.4 [95% CI 6.1, 11.7]	45 (36.9)	LR+ 6.5 [95% CI 4.7, 9.1]	47 (38.5)	LR+ 5.1 [95% CI 3.6, 7.1]
	moderate	3.1 - 18.6%	1261 (57.1)	68 (5.4)	-	108 (8.6)	-	149 (11.8)	-
	low	0.6 - 3%	781 (35.3)	15 (1.9)	LR- 0.3 [95% CI 0.2, 0.5]	27 (3.5)	LR- 0.4 [95% CI 0.3, 0.5]	44 (5.6)	LR- 0.4 [95% CI 0.3, 0.6]
	very low	0 - 0.5%	30 (1.4)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (3.3)	LR- 0.4 [95% CI 0.1, 2.8]	1 (3.3)	LR- 0.3 [95% CI 0, 2.1]

Table 0.5: Five group classification performance of the PIERS-ML model in high- vs low- and middle-income countries

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
Testing and internal validation dataset entries from high-income countries (N=1690)	very high	45.6 - 100%	11 (0.7)	10 (90.9)	LR+ 209.5 [95% CI 27.2, 1615.7]	10 (90.9)	LR+ 134.4 [95% CI 17.4, 1041.3]	10 (90.9)	LR+ 90 [95% CI 11.6, 698.7]
	high	18.7 - 45.5%	51 (3)	16 (31.4)	LR+ 11 [95% CI 6.4, 18.8]	18 (35.3)	LR+ 8 [95% CI 4.7, 13.7]	20 (39.2)	LR+ 6.2 [95% CI 3.6, 10.6]
	moderate	3.1 - 18.6%	933 (55.2)	40 (4.3)	NA	71 (7.6)	NA	107 (11.5)	NA
	low	0.6 - 3%	672 (39.8)	11 (1.6)	LR- 0.3 [95% CI 0.2, 0.6]	17 (2.5)	LR- 0.3 [95% CI 0.2, 0.5]	31 (4.6)	LR- 0.4 [95% CI 0.3, 0.6]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
[AUROC 0.77, AUPRC 0.35]	very low	0 - 0.5%	23 (1.4)	0 (0)	LR- 0 [95% CI 0, NaN]	1 (4.3)	LR- 0.6 [95% CI 0.1, 4.5]	1 (4.3)	LR- 0.4 [95% CI 0.1, 3]
Testing and internal validation dataset entries from low-	very high	45.6 - 100%	5 (1)	5 (100)	LR+ Inf [95% CI NaN, Inf]	5 (100)	LR+ Inf [95% CI NaN, Inf]	5 (100)	LR+ Inf [95% CI NaN, Inf]
	high	18.7 - 45.5%	96 (18.5)	29 (30.2)	LR+ 3.4 [95% CI 2.4, 4.8]	31 (32.3)	LR+ 2.8 [95% CI 2, 4]	32 (33.3)	LR+ 2.6 [95% CI 1.8, 3.7]
	moderate	3.1 - 18.6%	407 (78.3)	29 (7.1)	-	44 (10.8)	-	51 (12.5)	-

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
and middle-income countries (N=520) [AUROC 0.77, AUPRC 0.41]	low	0.6 - 3%	12 (2.3)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
	very low	0 - 0.5%	NA	NA	NA	NA	NA	NA	NA

Table 0.6: Model performances with systematic minority oversampling

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
PIERS-ML on SMOTE'd validation and testing datasets [AUROC 0.86, AUPRC 0.87]	very high	45.6 - 100%	129 (7.1)	128 (99.2)	LR+ 109.7 [95% CI 15.4, 783.2]	128 (99.2)	LR+ 101.5 [95% CI 14.2, 724.6]	128 (99.2)	LR+ 96.4 [95% CI 13.5, 688.3]
	high	18.7 - 45.5%	266 (14.6)	240 (90.2)	LR+ 9.1 [95% CI 6.1, 13.5]	241 (90.6)	LR+ 8.7 [95% CI 5.9, 13]	241 (90.6)	LR+ 8.3 [95% CI 5.5, 12.3]
	moderate	3.1 - 18.6%	1083 (59.5)	588 (54.3)	-	617 (57)	-	632 (58.4)	-
	low	0.6 - 3%	328 (18)	24 (7.3)	LR- 0.1 [95% CI 0, 0.1]	29 (8.8)	LR- 0.1 [95% CI 0.1, 0.1]	37 (11.3)	LR- 0.1 [95% CI 0.1, 0.1]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.5%	14 (0.8)	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]	0 (0)	LR- 0 [95% CI 0, NaN]
PIERS-ML refitted on SMOTE'd development, on original validation and testing datasets [AUROC 0.72, AUPRC 0.2]	very high	45.6 - 100%	239 (10.8)	55 (23)	LR+ 4.4 [95% CI 3.4, 5.7]	67 (28)	LR+ 4 [95% CI 3.1, 5.1]	73 (30.5)	LR+ 3.3 [95% CI 2.6, 4.3]
	high	18.7 - 45.5%	482 (21.8)	40 (8.3)	R+ 2 [95% CI 1.6, 2.6]	56 (11.6)	LR+ 1.9 [95% CI 1.5, 2.3]	71 (14.7)	LR+ 1.7 [95% CI 1.4, 2.1]
	moderate	3.1 - 18.6%	1424 (64.4)	43 (3)	-	70 (4.9)	-	109 (7.7)	-
	low	0.6 - 3%	65 (2.9)	2 (3.1)	LR- 0.5 [95% CI 0.1, 1.9]	4 (6.2)	LR- 0.7 [95% CI 0.2, 1.8]	4 (6.2)	LR- 0.5 [95% CI 0.2, 1.4]

Model	Risk Stratum	Predicted Probability	Hypertensive pregnant women in stratum N (%)	Within 2 days		Within 7 days		At any time	
				Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome	Hypertensive pregnant women within risk stratum with outcome N (%)	Likelihood ratio of outcome
	very low	0 - 0.5%	0 (0)	NA	NA	NA	NA	NA	NA

Table 0.7: Performance measures of risk stratification of models over time

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
BiMM	0	very high	45.6 - 100%	23 (0.9)	18 (78.3)	LR+ 44.5 [95% CI 16.7, 118.6]
		high	18.7 - 45.5%	166 (6.3)	58 (34.9)	LR+ 7.3 [95% CI 5.5, 9.7]
		moderate	3.1 - 18.6%	1359 (51.6)	95 (7)	-
		low	0.6 - 3%	962 (36.5)	26 (2.7)	LR- 0.3 [95% CI 0.2, 0.5]
		very low	0 - 0.5%	123 (4.7)	0 (0)	LR- 0 [95% CI 0, NaN]
	1	very high	45.6 - 100%	24 (1)	24 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	120 (5.1)	30 (25)	LR+ 5.1 [95% CI 3.5, 7.5]
		moderate	3.1 - 18.6%	1157 (49.3)	82 (7.1)	-
		low	0.6 - 3%	869 (37)	29 (3.3)	LR- 0.4 [95% CI 0.3, 0.6]
		very low	0 - 0.5%	179 (7.6)	1 (0.6)	LR- 0.1 [95% CI 0, 0.5]
	2	very high	45.6 - 100%	11 (0.6)	10 (90.9)	LR+ 198.6 [95% CI 25.7, 1534.1]
		high	18.7 - 45.5%	64 (3.5)	18 (28.1)	LR+ 8.8 [95% CI 5.4, 14.4]
		moderate	3.1 - 18.6%	851 (46.9)	41 (4.8)	-
		low	0.6 - 3%	721 (39.7)	16 (2.2)	LR- 0.4 [95% CI 0.3, 0.7]
		very low	0 - 0.5%	168 (9.3)	2 (1.2)	LR- 0.2 [95% CI 0.1, 0.9]
	3	very high	45.6 - 100%	9 (0.6)	8 (88.9)	LR+ 229.2 [95% CI 29.2, 1795.9]
		high	18.7 - 45.5%	51 (3.6)	10 (19.6)	LR+ 8.4 [95% CI 4.5, 15.5]
		moderate	3.1 - 18.6%	667 (46.9)	23 (3.4)	-
		low	0.6 - 3%	588 (41.3)	7 (1.2)	LR- 0.3 [95% CI 0.2, 0.6]
		very low	0 - 0.5%	108 (7.6)	0 (0)	LR- 0 [95% CI 0, NaN]
	4	very high	45.6 - 100%	3 (0.3)	2 (66.7)	LR+ 84.7 [95% CI 7.9, 904]
		high	18.7 - 45.5%	34 (3.1)	6 (17.6)	LR+ 9.9 [95% CI 4.5, 21.5]
		moderate	3.1 - 18.6%	511 (47.1)	11 (2.2)	-
		low	0.6 - 3%	463 (42.7)	5 (1.1)	LR- 0.4 [95% CI 0.2, 1]
		very low	0 - 0.5%	73 (6.7)	1 (1.4)	LR- 0.6 [95% CI 0.1, 4.1]

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
	5	very high	45.6 - 100%	3 (0.4)	3 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	26 (3.2)	6 (23.1)	LR+ 12.4 [95% CI 5.6, 27.4]
		moderate	3.1 - 18.6%	381 (47)	12 (3.1)	-
		low	0.6 - 3%	347 (42.8)	1 (0.3)	LR- 0.1 [95% CI 0, 0.7]
		very low	0 - 0.5%	53 (6.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	6	very high	45.6 - 100%	2 (0.3)	1 (50)	LR+ 31.3 [95% CI 2, 482.7]
		high	18.7 - 45.5%	31 (4.8)	7 (22.6)	LR+ 9.6 [95% CI 4.7, 19.5]
		moderate	3.1 - 18.6%	315 (48.8)	10 (3.2)	-
		low	0.6 - 3%	264 (40.9)	2 (0.8)	LR- 0.2 [95% CI 0.1, 0.8]
		very low	0 - 0.5%	34 (5.3)	0 (0)	LR- 0 [95% CI 0, NaN]
	7	very high	45.6 - 100%	1 (0.2)	1 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	26 (5)	6 (23.1)	LR+ 7.1 [95% CI 3.2, 15.8]
		moderate	3.1 - 18.6%	258 (49.8)	13 (5)	-
		low	0.6 - 3%	212 (40.9)	2 (0.9)	LR- 0.2 [95% CI 0.1, 0.8]
		very low	0 - 0.5%	21 (4.1)	0 (0)	LR- 0 [95% CI 0, NaN]
	8	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	16 (3.7)	2 (12.5)	LR+ 3.7 [95% CI 0.9, 15]
		moderate	3.1 - 18.6%	243 (56.2)	11 (4.5)	-
		low	0.6 - 3%	156 (36.1)	3 (1.9)	LR- 0.5 [95% CI 0.2, 1.4]
		very low	0 - 0.5%	17 (3.9)	0 (0)	LR- 0 [95% CI 0, NaN]
	9	very high	45.6 - 100%	2 (0.5)	0 (0)	LR+ 0 [95% CI 0, NaN]
		high	18.7 - 45.5%	9 (2.4)	1 (11.1)	LR+ 3.7 [95% CI 0.5, 27.6]
		moderate	3.1 - 18.6%	204 (54.5)	10 (4.9)	-
		low	0.6 - 3%	140 (37.4)	1 (0.7)	LR- 0.2 [95% CI 0, 1.3]
		very low	0 - 0.5%	19 (5.1)	0 (0)	LR- 0 [95% CI 0, NaN]
	10	very high	45.6 - 100%	0 (0)	0 (0)	-

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		high	18.7 - 45.5%	22 (7)	3 (13.6)	LR+ 5.4 [95% CI 1.9, 14.9]
		moderate	3.1 - 18.6%	169 (53.7)	4 (2.4)	-
		low	0.6 - 3%	111 (35.2)	2 (1.8)	LR- 0.6 [95% CI 0.2, 2]
		very low	0 - 0.5%	13 (4.1)	0 (0)	LR- 0 [95% CI 0, NaN]
	11	very high	45.6 - 100%	1 (0.4)	1 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	10 (3.8)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	159 (60)	4 (2.5)	-
		low	0.6 - 3%	84 (31.7)	1 (1.2)	LR- 0.5 [95% CI 0.1, 3]
		very low	0 - 0.5%	11 (4.2)	0 (0)	LR- 0 [95% CI 0, NaN]
	12	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	10 (4)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	146 (59.1)	4 (2.7)	-
		low	0.6 - 3%	82 (33.2)	1 (1.2)	LR- 0.6 [95% CI 0.1, 3.4]
		very low	0 - 0.5%	9 (3.6)	0 (0)	LR- 0 [95% CI 0, NaN]
	13	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	10 (4.4)	3 (30)	LR+ 11.8 [95% CI 3.7, 37.5]
		moderate	3.1 - 18.6%	131 (57.2)	2 (1.5)	-
		low	0.6 - 3%	80 (34.9)	3 (3.8)	LR- 1 [95% CI 0.4, 2.6]
		very low	0 - 0.5%	8 (3.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	14	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	10 (4.8)	1 (10)	LR+ 3.2 [95% CI 0.5, 22.1]
		moderate	3.1 - 18.6%	125 (59.5)	6 (4.8)	-
		low	0.6 - 3%	68 (32.4)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	7 (3.3)	0 (0)	LR- 0 [95% CI 0, NaN]
JMB	0	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	118 (5.4)	7 (5.9)	LR+ 1.1 [95% CI 0.5, 2.3]
		moderate	3.1 - 18.6%	2077 (94.6)	110 (5.3)	-
		low	0.6 - 3%	0 (0)	0 (0)	-
		very low	0 - 0.5%	0 (0)	0 (0)	-

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
	1	very high	45.6 - 100%	1 (0)	1 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	66 (3.3)	3 (4.5)	LR+ 1.5 [95% CI 0.5, 4.5]
		moderate	3.1 - 18.6%	1963 (96.7)	61 (3.1)	-
		low	0.6 - 3%	0 (0)	0 (0)	-
		very low	0 - 0.5%	0 (0)	0 (0)	-
	2	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	33 (1.8)	2 (6.1)	LR+ 3 [95% CI 0.7, 12]
		moderate	3.1 - 18.6%	1753 (98.2)	36 (2.1)	-
		low	0.6 - 3%	0 (0)	0 (0)	-
		very low	0 - 0.5%	0 (0)	0 (0)	-
	3	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	9 (0.6)	2 (22.2)	LR+ 18 [95% CI 3.9, 81.9]
		moderate	3.1 - 18.6%	1456 (99)	21 (1.4)	-
		low	0.6 - 3%	5 (0.3)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	4	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	1211 (99.3)	19 (1.6)	-
		low	0.6 - 3%	8 (0.7)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	5	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	1023 (98.9)	15 (1.5)	-
		low	0.6 - 3%	10 (1)	1 (10)	LR- 7.1 [95% CI 1, 52.6]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	6	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	877 (98.9)	17 (1.9)	-

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		low	0.6 - 3%	10 (1.1)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	7	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	763 (97)	12 (1.6)	-
		low	0.6 - 3%	23 (2.9)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	8	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	678 (95.4)	9 (1.3)	-
		low	0.6 - 3%	33 (4.6)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	9	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	599 (92.6)	9 (1.5)	-
		low	0.6 - 3%	48 (7.4)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	10	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	522 (87.1)	5 (1)	-
		low	0.6 - 3%	77 (12.9)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	11	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	457 (81.3)	3 (0.7)	-
		low	0.6 - 3%	105 (18.7)	1 (1)	LR- 1.3 [95% CI 0.2, 7.4]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	12	very high	45.6 - 100%	0 (0)	0 (0)	-

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	402 (75.6)	6 (1.5)	-
		low	0.6 - 3%	130 (24.4)	1 (0.8)	LR- 0.6 [95% CI 0.1, 3.6]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	13	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	331 (66.6)	4 (1.2)	-
		low	0.6 - 3%	166 (33.4)	1 (0.6)	LR- 0.6 [95% CI 0.1, 3.5]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	14	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	263 (56.7)	4 (1.5)	-
		low	0.6 - 3%	201 (43.3)	2 (1)	LR- 0.8 [95% CI 0.2, 2.4]
		very low	0 - 0.5%	0 (0)	0 (0)	-
LSTM	0	very high	45.6 - 100%	24 (1.1)	11 (45.8)	LR+ 10.3 [95% CI 4.7, 22.6]
		high	18.7 - 45.5%	58 (2.6)	17 (29.3)	LR+ 5.3 [95% CI 3.1, 9.2]
		moderate	3.1 - 18.6%	1922 (85.1)	136 (7.1)	-
		low	0.6 - 3%	254 (11.2)	8 (3.1)	LR- 0.4 [95% CI 0.2, 0.8]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	1	very high	45.6 - 100%	13 (0.6)	10 (76.9)	LR+ 48.5 [95% CI 13.5, 174.3]
		high	18.7 - 45.5%	43 (2)	16 (37.2)	LR+ 9.3 [95% CI 5.1, 16.8]
		moderate	3.1 - 18.6%	1209 (55.1)	90 (7.4)	-
		low	0.6 - 3%	928 (42.3)	25 (2.7)	LR- 0.4 [95% CI 0.3, 0.6]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	2	very high	45.6 - 100%	2 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		high	18.7 - 45.5%	18 (0.9)	4 (22.2)	LR+ 7.9 [95% CI 2.7, 23.3]

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		moderate	3.1 - 18.6%	888 (43.8)	45 (5.1)	-
		low	0.6 - 3%	1119 (55.2)	22 (2)	LR- 0.6 [95% CI 0.4, 0.8]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	3	very high	45.6 - 100%	4 (0.2)	3 (75)	LR+ 118.5 [95% CI 12.6, 1116.8]
		high	18.7 - 45.5%	3 (0.2)	2 (66.7)	LR+ 84.7 [95% CI 7.8, 915.9]
		moderate	3.1 - 18.6%	543 (30.5)	21 (3.9)	-
		low	0.6 - 3%	1232 (69.1)	18 (1.5)	LR- 0.6 [95% CI 0.4, 0.8]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	4	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	3 (0.2)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	284 (19.3)	15 (5.3)	-
		low	0.6 - 3%	1181 (80.4)	10 (0.8)	LR- 0.5 [95% CI 0.3, 0.8]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	5	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	2 (0.2)	1 (50)	LR+ 54.4 [95% CI 3.5, 842.2]
		moderate	3.1 - 18.6%	185 (15.2)	10 (5.4)	-
		low	0.6 - 3%	1032 (84.7)	11 (1.1)	LR- 0.6 [95% CI 0.4, 0.9]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	6	very high	45.6 - 100%	1 (0.1)	1 (100)	LR+ Inf [95% CI NaN, Inf]
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	190 (18.4)	8 (4.2)	-
		low	0.6 - 3%	843 (81.5)	10 (1.2)	LR- 0.6 [95% CI 0.4, 1]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	7	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	2 (0.2)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	182 (20.5)	6 (3.3)	-

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		low	0.6 - 3%	702 (79.2)	14 (2)	LR- 0.9 [95% CI 0.7, 1.2]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	8	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	183 (23.3)	5 (2.7)	-
		low	0.6 - 3%	603 (76.6)	11 (1.8)	LR- 0.9 [95% CI 0.6, 1.2]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	9	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.1)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	180 (25.3)	8 (4.4)	-
		low	0.6 - 3%	530 (74.5)	3 (0.6)	LR- 0.4 [95% CI 0.1, 1]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	10	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	1 (0.2)	0 (0)	LR+ 0 [95% CI 0, NaN]
		moderate	3.1 - 18.6%	191 (29.5)	5 (2.6)	-
		low	0.6 - 3%	455 (70.3)	2 (0.4)	LR- 0.4 [95% CI 0.1, 1.3]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	11	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	179 (29.9)	4 (2.2)	-
		low	0.6 - 3%	420 (70.1)	2 (0.5)	LR- 0.5 [95% CI 0.2, 1.5]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	12	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	164 (29.2)	1 (0.6)	-
		low	0.6 - 3%	398 (70.8)	4 (1)	LR- 1.1 [95% CI 0.7, 1.8]

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		very low	0 - 0.5%	0 (0)	0 (0)	-
	13	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	158 (29.7)	3 (1.9)	-
		low	0.6 - 3%	374 (70.3)	4 (1.1)	LR- 0.8 [95% CI 0.4, 1.5]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	14	very high	45.6 - 100%	0 (0)	0 (0)	-
		high	18.7 - 45.5%	0 (0)	0 (0)	-
		moderate	3.1 - 18.6%	151 (30.4)	4 (2.6)	-
		low	0.6 - 3%	346 (69.6)	3 (0.9)	LR- 0.6 [95% CI 0.3, 1.4]
		very low	0 - 0.5%	0 (0)	0 (0)	-
TS	0	very high	45.6 - 100%	277 (12.3)	59 (21.3)	LR+ 3.3 [95% CI 2.6, 4.2]
		high	18.7 - 45.5%	1434 (63.5)	89 (6.2)	LR+ 1.1 [95% CI 1, 1.2]
		moderate	3.1 - 18.6%	465 (20.6)	21 (4.5)	-
		low	0.6 - 3%	81 (3.6)	3 (3.7)	LR- 0.5 [95% CI 0.1, 1.5]
		very low	0 - 0.5%	1 (0)	0 (0)	LR- 0 [95% CI 0, NaN]
	1	very high	45.6 - 100%	187 (9.4)	44 (23.5)	LR+ 4.1 [95% CI 3.1, 5.5]
		high	18.7 - 45.5%	1114 (56.2)	70 (6.3)	LR+ 1.2 [95% CI 1.1, 1.4]
		moderate	3.1 - 18.6%	514 (25.9)	22 (4.3)	-
		low	0.6 - 3%	165 (8.3)	2 (1.2)	LR- 0.2 [95% CI 0, 0.7]
		very low	0 - 0.5%	3 (0.2)	1 (33.3)	LR- 6.6 [95% CI 0.6, 72.7]
	2	very high	45.6 - 100%	149 (9)	22 (14.8)	LR+ 3.9 [95% CI 2.6, 5.7]
		high	18.7 - 45.5%	831 (50.3)	30 (3.6)	LR+ 1.1 [95% CI 0.9, 1.4]
		moderate	3.1 - 18.6%	437 (26.4)	14 (3.2)	-
		low	0.6 - 3%	227 (13.7)	5 (2.2)	LR- 0.5 [95% CI 0.2, 1.2]
		very low	0 - 0.5%	9 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	3	very high	45.6 - 100%	98 (7.2)	13 (13.3)	LR+ 4.6 [95% CI 2.8, 7.5]

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
		high	18.7 - 45.5%	681 (50.2)	19 (2.8)	LR+ 1.1 [95% CI 0.9, 1.5]
		moderate	3.1 - 18.6%	439 (32.4)	11 (2.5)	-
		low	0.6 - 3%	135 (9.9)	1 (0.7)	LR- 0.2 [95% CI 0, 1.6]
		very low	0 - 0.5%	4 (0.3)	0 (0)	LR- 0 [95% CI 0, NaN]
	4	very high	45.6 - 100%	83 (8)	6 (7.2)	LR+ 3.3 [95% CI 1.6, 6.8]
		high	18.7 - 45.5%	489 (47.3)	12 (2.5)	LR+ 1.3 [95% CI 0.9, 1.8]
		moderate	3.1 - 18.6%	359 (34.7)	4 (1.1)	-
		low	0.6 - 3%	100 (9.7)	2 (2)	LR- 0.9 [95% CI 0.2, 3.3]
		very low	0 - 0.5%	3 (0.3)	0 (0)	LR- 0 [95% CI 0, NaN]
	5	very high	45.6 - 100%	75 (9.5)	7 (9.3)	LR+ 3.6 [95% CI 1.9, 6.9]
		high	18.7 - 45.5%	274 (34.8)	8 (2.9)	LR+ 1.4 [95% CI 0.9, 2.3]
		moderate	3.1 - 18.6%	343 (43.6)	5 (1.5)	-
		low	0.6 - 3%	91 (11.6)	2 (2.2)	LR- 0.8 [95% CI 0.2, 3]
		very low	0 - 0.5%	4 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	6	very high	45.6 - 100%	59 (9.5)	6 (10.2)	LR+ 3.6 [95% CI 1.8, 7.3]
		high	18.7 - 45.5%	267 (43)	9 (3.4)	LR+ 1.5 [95% CI 1, 2.1]
		moderate	3.1 - 18.6%	226 (36.4)	2 (0.9)	-
		low	0.6 - 3%	66 (10.6)	2 (3)	LR- 1 [95% CI 0.3, 3.7]
		very low	0 - 0.5%	3 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	7	very high	45.6 - 100%	45 (9.2)	2 (4.4)	LR+ 1.2 [95% CI 0.3, 4.4]
		high	18.7 - 45.5%	212 (43.2)	12 (5.7)	LR+ 1.5 [95% CI 1.1, 2.1]
		moderate	3.1 - 18.6%	177 (36)	3 (1.7)	-
		low	0.6 - 3%	55 (11.2)	2 (3.6)	LR- 0.9 [95% CI 0.2, 3.5]
		very low	0 - 0.5%	2 (0.4)	0 (0)	LR- 0 [95% CI 0, NaN]
	8	very high	45.6 - 100%	40 (9.7)	1 (2.5)	LR+ 0.6 [95% CI 0.1, 4.3]
		high	18.7 - 45.5%	182 (44.2)	10 (5.5)	LR+ 1.4 [95% CI 1, 2]
		moderate	3.1 - 18.6%	147 (35.7)	3 (2)	-
		low	0.6 - 3%	40 (9.7)	2 (5)	LR- 1.3 [95% CI 0.3, 4.9]
		very low	0 - 0.5%	3 (0.7)	0 (0)	LR- 0 [95% CI 0, NaN]

Model	Day	Risk group	Probability	N (%)	Yes (%)	Likelihood ratios
	9	very high	45.6 - 100%	36 (10.1)	1 (2.8)	LR+ 0.9 [95% CI 0.1, 6]
		high	18.7 - 45.5%	154 (43.3)	6 (3.9)	LR+ 1.3 [95% CI 0.7, 2.1]
		moderate	3.1 - 18.6%	122 (34.3)	4 (3.3)	-
		low	0.6 - 3%	41 (11.5)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	3 (0.8)	0 (0)	LR- 0 [95% CI 0, NaN]
	10	very high	45.6 - 100%	34 (11.2)	0 (0)	LR+ 0 [95% CI 0, NaN]
		high	18.7 - 45.5%	197 (65)	5 (2.5)	LR+ 1 [95% CI 0.6, 1.6]
		moderate	3.1 - 18.6%	58 (19.1)	2 (3.4)	-
		low	0.6 - 3%	14 (4.6)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	11	very high	45.6 - 100%	33 (12.9)	1 (3)	LR+ 1.3 [95% CI 0.2, 8]
		high	18.7 - 45.5%	155 (60.8)	3 (1.9)	LR+ 0.9 [95% CI 0.4, 1.8]
		moderate	3.1 - 18.6%	52 (20.4)	2 (3.8)	-
		low	0.6 - 3%	15 (5.9)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	0 (0)	0 (0)	-
	12	very high	45.6 - 100%	26 (11)	1 (3.8)	LR+ 1.8 [95% CI 0.3, 11.1]
		high	18.7 - 45.5%	142 (60.2)	3 (2.1)	LR+ 1.1 [95% CI 0.6, 2]
		moderate	3.1 - 18.6%	52 (22)	1 (1.9)	-
		low	0.6 - 3%	15 (6.4)	0 (0)	LR- 0 [95% CI 0, NaN]
		very low	0 - 0.5%	1 (0.4)	0 (0)	LR- 0 [95% CI 0, NaN]
	13	very high	45.6 - 100%	22 (10)	1 (4.5)	LR+ 1.4 [95% CI 0.2, 9.3]
		high	18.7 - 45.5%	134 (61.2)	5 (3.7)	LR+ 1.2 [95% CI 0.9, 1.8]
		moderate	3.1 - 18.6%	44 (20.1)	0 (0)	-
		low	0.6 - 3%	18 (8.2)	1 (5.6)	LR- 1.8 [95% CI 0.3, 11.5]
		very low	0 - 0.5%	1 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]
	14	very high	45.6 - 100%	24 (11.9)	0 (0)	LR+ 0 [95% CI 0, NaN]
		high	18.7 - 45.5%	125 (62.2)	4 (3.2)	LR+ 0.8 [95% CI 0.4, 1.5]
		moderate	3.1 - 18.6%	40 (19.9)	2 (5)	-
		low	0.6 - 3%	11 (5.5)	1 (9.1)	LR- 2.8 [95% CI 0.4, 18.7]
		very low	0 - 0.5%	1 (0.5)	0 (0)	LR- 0 [95% CI 0, NaN]

Table 0.8 Logistic regression ruling in and ruling out model performances on their corresponding testing dataset

	Ruling in				Ruling out				Variable group						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
1	30	5 (0.4)	1 (20)	8.4	7	45 (3.5)	0 (0)	0.0							
2	42	6 (0.5)	0 (0)	0.0	46	136 (10.7)	2 (1.5)	0.5	✓						
3	35	13 (1)	2 (15.4)	6.1						✓					
4	37	8 (0.6)	1 (12.5)	4.8							✓				
5					47	66 (5.2)	1 (1.5)	0.5				✓			
6	53	3 (0.2)	0 (0)	0.0									✓		
7	54	2 (0.2)	0 (0)	0.0	75	31 (2.4)	1 (3.2)	1.1						✓	
8	22	13 (1.4)	3 (23.1)	9.8											✓
9	33	19 (1.5)	3 (15.8)	6.3					✓	✓					
10	39	11 (0.9)	1 (9.1)	3.3					✓		✓				
11					22	226 (17.7)	2 (0.9)	0.3	✓			✓			
12	58	1 (0.1)	0 (0)	0.0					✓				✓		
13	40	9 (0.7)	0 (0)	0.0	25	116 (9.1)	1 (0.9)	0.3	✓					✓	
14	12	6 (0.7)	2 (33.3)	16.4	20	225 (24.7)	2 (0.9)	0.3	✓						✓
15	27	23 (1.8)	5 (21.7)	9.3	65	134 (10.5)	3 (2.2)	0.8		✓	✓				
16	17	34 (2.7)	8 (23.5)	10.3	79	158 (12.4)	6 (3.8)	1.3		✓		✓			
17					51	350 (27.5)	6 (1.7)	0.6		✓			✓		
18	23	9 (0.7)	2 (22.2)	9.6	3	92 (7.2)	0 (0)	0.0		✓				✓	
19	21	26 (2.9)	6 (23.1)	9.8	54	119 (13.1)	2 (1.7)	0.6		✓					✓
20					70	108 (8.5)	3 (2.8)	1.0			✓	✓			
21	55	2 (0.2)	0 (0)	0.0							✓		✓		
22	38	9 (0.7)	1 (11.1)	4.2	72	69 (5.4)	2 (2.9)	1.0			✓			✓	
23	1	6 (0.7)	3 (50)	32.7							✓				✓
24	46	4 (0.3)	0 (0)	0.0	13	262 (20.5)	2 (0.8)	0.3				✓	✓		

	Ruling in				Ruling out				Variable group						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
25					73	101 (7.9)	3 (3)	1.0				✓		✓	
26					64	94 (10.3)	2 (2.1)	0.7				✓			✓
27	59	1 (0.1)	0 (0)	0.0									✓	✓	
28													✓		✓
29	2	9 (1)	4 (44.4)	26.2	6	33 (3.6)	0 (0)	0.0						✓	✓
30	32	29 (2.3)	5 (17.2)	7.0					✓	✓	✓				
31					57	225 (17.6)	4 (1.8)	0.6	✓	✓		✓			
32	7	3 (0.2)	1 (33.3)	16.7	17	440 (34.5)	4 (0.9)	0.3	✓	✓			✓		
33	24	9 (0.7)	2 (22.2)	9.6	23	212 (16.6)	2 (0.9)	0.3	✓	✓				✓	
34	31	20 (2.2)	4 (20)	8.2	53	237 (26)	4 (1.7)	0.6	✓	✓					✓
35					62	154 (12.1)	3 (1.9)	0.7	✓		✓	✓			
36	41	8 (0.6)	0 (0)	0.0					✓		✓		✓		
37	36	14 (1.1)	2 (14.3)	5.6	74	157 (12.3)	5 (3.2)	1.1	✓		✓			✓	
38	5	5 (0.5)	2 (40)	21.8					✓		✓				✓
39	47	4 (0.3)	0 (0)	0.0	8	356 (27.9)	1 (0.3)	0.1	✓			✓	✓		
40					30	204 (16)	2 (1)	0.3	✓			✓		✓	
41					21	224 (24.6)	2 (0.9)	0.3	✓			✓			✓
42	43	5 (0.4)	0 (0)	0.0					✓				✓	✓	
43									✓				✓		✓
44	44	4 (0.4)	0 (0)	0.0	4	58 (6.4)	0 (0)	0.0	✓					✓	✓
45	14	26 (2)	7 (26.9)	12.3	78	169 (13.3)	6 (3.6)	1.2		✓	✓	✓			
46										✓	✓		✓		
47	25	9 (0.7)	2 (22.2)	9.6	68	126 (9.9)	3 (2.4)	0.8		✓	✓			✓	
48	19	30 (3.3)	7 (23.3)	10.0	2	103 (11.3)	0 (0)	0.0		✓	✓				✓
49					43	400 (31.4)	6 (1.5)	0.5		✓		✓	✓		
50					77	148 (11.6)	5 (3.4)	1.2		✓		✓		✓	

	Ruling in				Ruling out				Variable group						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
51	10	18 (2)	6 (33.3)	16.4	61	154 (16.9)	3 (1.9)	0.6		✓		✓			✓
52					28	288 (22.6)	3 (1)	0.4		✓			✓	✓	
53										✓			✓		✓
54	20	30 (3.3)	7 (23.3)	10.0	24	108 (11.9)	1 (0.9)	0.3		✓				✓	✓
55	48	4 (0.3)	0 (0)	0.0							✓	✓	✓		
56					71	106 (8.3)	3 (2.8)	1.0			✓	✓		✓	
57					66	91 (10)	2 (2.2)	0.7			✓	✓			✓
58	34	19 (1.5)	3 (15.8)	6.3							✓		✓	✓	
59											✓		✓		✓
60	6	11 (1.2)	4 (36.4)	18.7	5	35 (3.8)	0 (0)	0.0			✓			✓	✓
61	49	4 (0.3)	0 (0)	0.0								✓	✓	✓	
62					39	228 (25.1)	3 (1.3)	0.4				✓	✓		✓
63					58	112 (12.3)	2 (1.8)	0.6				✓		✓	✓
64													✓	✓	✓
65					59	262 (20.5)	5 (1.9)	0.7	✓	✓	✓	✓			
66	8	3 (0.2)	1 (33.3)	16.7					✓	✓	✓		✓		
67	26	9 (0.7)	2 (22.2)	9.6	40	235 (18.4)	3 (1.3)	0.4	✓	✓	✓			✓	
68	29	24 (2.6)	5 (20.8)	8.6	36	240 (26.4)	3 (1.2)	0.4	✓	✓	✓				✓
69					18	438 (34.4)	4 (0.9)	0.3	✓	✓		✓	✓		
70					56	283 (22.2)	5 (1.8)	0.6	✓	✓		✓		✓	
71	28	19 (2.1)	4 (21.1)	8.7	44	267 (29.3)	4 (1.5)	0.5	✓	✓		✓			✓
72	9	3 (0.2)	1 (33.3)	16.7	11	396 (31.1)	3 (0.8)	0.3	✓	✓			✓	✓	
73									✓	✓			✓		✓
74					42	209 (23)	3 (1.4)	0.5	✓	✓				✓	✓
75	56	3 (0.2)	0 (0)	0.0					✓		✓	✓	✓		
76					31	183 (14.4)	2 (1.1)	0.4	✓		✓	✓		✓	

	Ruling in				Ruling out				Variable group						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
77	50	3 (0.3)	0 (0)	0.0	32	91 (10)	1 (1.1)	0.4	✓		✓	✓			✓
78	45	5 (0.4)	0 (0)	0.0					✓		✓		✓	✓	
79									✓		✓		✓		✓
80	11	9 (1)	3 (33.3)	16.4					✓		✓			✓	✓
81	51	4 (0.3)	0 (0)	0.0					✓			✓	✓	✓	
82					27	294 (32.3)	3 (1)	0.3	✓			✓	✓		✓
83					19	230 (25.3)	2 (0.9)	0.3	✓			✓		✓	✓
84									✓				✓	✓	✓
85					38	396 (31.1)	5 (1.3)	0.4		✓	✓	✓	✓		
86					76	181 (14.2)	6 (3.3)	1.1		✓	✓	✓		✓	
87	15	30 (3.3)	8 (26.7)	11.9	69	154 (16.9)	4 (2.6)	0.9		✓	✓	✓			✓
88										✓	✓		✓	✓	
89										✓	✓		✓		✓
90	18	34 (3.7)	8 (23.5)	10.1	1	108 (11.9)	0 (0)	0.0		✓	✓			✓	✓
91					37	323 (25.3)	4 (1.2)	0.4		✓		✓	✓	✓	
92					9	278 (30.5)	2 (0.7)	0.2		✓		✓	✓		✓
93					60	162 (17.8)	3 (1.9)	0.6		✓		✓		✓	✓
94					12	244 (26.8)	2 (0.8)	0.3		✓			✓	✓	✓
95	57	3 (0.2)	0 (0)	0.0							✓	✓	✓	✓	
96											✓	✓	✓		✓
97					63	107 (11.8)	2 (1.9)	0.6			✓	✓		✓	✓
98											✓		✓	✓	✓
99												✓	✓	✓	✓
100					55	442 (34.7)	8 (1.8)	0.6	✓	✓	✓	✓	✓		
101					67	310 (24.3)	7 (2.3)	0.8	✓	✓	✓	✓		✓	
102					52	242 (26.6)	4 (1.7)	0.5	✓	✓	✓	✓			✓

	Ruling in				Ruling out				Variable group						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
103	4	5 (0.4)	2 (40)	22.3					✓	✓	✓		✓	✓	
104									✓	✓	✓		✓		✓
105	3	7 (0.8)	3 (42.9)	24.5	41	217 (23.8)	3 (1.4)	0.5	✓	✓	✓			✓	✓
106					33	420 (32.9)	5 (1.2)	0.4	✓	✓		✓	✓	✓	
107					15	339 (37.3)	3 (0.9)	0.3	✓	✓		✓	✓		✓
108					29	198 (21.8)	2 (1)	0.3	✓	✓		✓		✓	✓
109									✓	✓			✓	✓	✓
110	52	4 (0.3)	0 (0)	0.0					✓		✓	✓	✓	✓	
111									✓		✓	✓	✓		✓
112					49	187 (20.5)	3 (1.6)	0.5	✓		✓	✓		✓	✓
113									✓		✓		✓	✓	✓
114									✓			✓	✓	✓	✓
115					48	315 (24.7)	5 (1.6)	0.5		✓	✓	✓	✓	✓	
116					10	276 (30.3)	2 (0.7)	0.2		✓	✓	✓	✓		✓
117	13	28 (3.1)	8 (28.6)	13.1	50	182 (20)	3 (1.6)	0.5		✓	✓	✓		✓	✓
118										✓	✓		✓	✓	✓
119					35	257 (28.2)	3 (1.2)	0.4		✓		✓	✓	✓	✓
120											✓	✓	✓	✓	✓
121					34	419 (32.9)	5 (1.2)	0.4	✓	✓	✓	✓	✓	✓	
122					14	344 (37.8)	3 (0.9)	0.3	✓	✓	✓	✓	✓		✓
123	16	24 (2.6)	6 (25)	10.9	45	262 (28.8)	4 (1.5)	0.5	✓	✓	✓	✓		✓	✓
124									✓	✓	✓		✓	✓	✓
125					26	312 (34.3)	3 (1)	0.3	✓	✓		✓	✓	✓	✓
126									✓		✓	✓	✓	✓	✓
127										✓	✓	✓	✓	✓	✓
128					16	323 (35.5)	3 (0.9)	0.3	✓	✓	✓	✓	✓	✓	✓

Table 0.9: Selected logistic regression models for ruling in and ruling out for each scenario

Scenario	Usable models	Ruling in		Ruling out	
		Model	LR+	Model	LR-
1	1	1	LR+ 8.4 [95% CI 1-73]	1	LR- 0 [95% CI 0-NaN]
2	1-2	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
3	1, 3	1	LR+ 8.4 [95% CI 1-73]	1	
4	1, 4	1		1	
5	1, 5	1		1	
6	1, 6	1		1	
7	1, 7	1		1	
8	1, 8	8	LR+ 9.8 [95% CI 2.9-33.6]	1	
9	1-3, 9	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
10	1-2, 4, 10	1		1	
11	1-2, 5, 11	1		1	
12	1-2, 6, 12	1		1	
13	1-2, 7, 13	1		1	
14	1-2, 8, 14	14	LR+ 16.4 [95% CI 3.1-85.5]	1	
15	1, 3-4, 15	15	LR+ 9.3 [95% CI 3.6-23.7]	1	
16	1, 3, 5, 16	16	LR+ 10.3 [95% CI 5-21.2]	1	
17	1, 3, 6, 17	1	LR+ 8.4 [95% CI 1-73]	1	
18	1, 3, 7, 18	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
19	1, 3, 8, 19	19	LR+ 9.8 [95% CI 4.3-22.5]	1	
20	1, 4-5, 20	1	LR+ 8.4 [95% CI 1-73]	1	
21	1, 4, 6, 21	1		1	
22	1, 4, 7, 22	1		1	
23	1, 4, 8, 23	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
24	1, 5-6, 24	1	LR+ 8.4 [95% CI 1-73]	1	
25	1, 5, 7, 25	1		1	
26	1, 5, 8, 26	8	LR+ 9.8 [95% CI 2.9-33.6]	1	
27	1, 6-7, 27	1	LR+ 8.4 [95% CI 1-73]	1	

28	1, 6, 8, 28	8	LR+ 9.8 [95% CI 2.9-33.6]	1	
29	1, 7-8, 29	29	LR+ 26.2 [95% CI 7.4-92]	29	
30	1-4, 9-10, 15, 30	15	LR+ 9.3 [95% CI 3.6-23.7]	1	
31	1-3, 5, 9, 11, 16, 31	16	LR+ 10.3 [95% CI 5-21.2]	1	
32	1-3, 6, 9, 12, 17, 32, 66	32	LR+ 16.7 [95% CI 1.6-180.4]	1	
33	1-3, 7, 9, 13, 18, 33	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
34	1-3, 8-9, 14, 19, 34	14	LR+ 16.4 [95% CI 3.1-85.5]	1	
35	1-2, 4-5, 10-11, 20, 35	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
36	1-2, 4, 6, 10, 12, 21, 36	1		1	
37	1-2, 4, 7, 10, 13, 22, 37	1		1	
38	1-2, 4, 8, 10, 14, 23, 38	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
39	1-2, 5-6, 11-12, 24, 39	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
40	1-2, 5, 7, 11, 13, 25, 40	1		1	
41	1-2, 5, 8, 11, 14, 26, 41	14	LR+ 16.4 [95% CI 3.1-85.5]	1	
42	1-2, 6-7, 12-13, 27, 42	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
43	1-2, 6, 8, 12, 14, 28, 43, 79	14	LR+ 16.4 [95% CI 3.1-85.5]	1	
44	1-2, 7-8, 13-14, 29, 44	29	LR+ 26.2 [95% CI 7.4-92]	44	
45	1, 3-5, 15-16, 20, 45	45	LR+ 12.3 [95% CI 5.5-27.5]	1	
46	1, 3-4, 6, 15, 17, 21, 46	15	LR+ 9.3 [95% CI 3.6-23.7]	1	
47	1, 3-4, 7, 15, 18, 22, 47	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
48	1, 3-4, 8, 15, 19, 23, 48	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
49	1, 3, 5-6, 16-17, 24, 49, 85	16	LR+ 10.3 [95% CI 5-21.2]	1	
50	1, 3, 5, 7, 16, 18, 25, 50	16		18	
51	1, 3, 5, 8, 16, 19, 26, 51	51	LR+ 16.4 [95% CI 6.6-40.3]	1	
52	1, 3, 6-7, 17-18, 27, 52	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
53	1, 3, 6, 8, 17, 19, 28, 53	19	LR+ 9.8 [95% CI 4.3-22.5]	1	
54	1, 3, 7-8, 18-19, 29, 54	29	LR+ 26.2 [95% CI 7.4-92]	18	
55	1, 4-6, 20-21, 24, 55	1	LR+ 8.4 [95% CI 1-73]	1	
56	1, 4-5, 7, 20, 22, 25, 56	1		1	
57	1, 4-5, 8, 20, 23, 26, 57	23	LR+ 32.7 [95% CI 6.9-154.7]	1	

58	1, 4, 6-7, 21-22, 27, 58	1	LR+ 8.4 [95% CI 1-73]	1	
59	1, 4, 6, 8, 21, 23, 28, 59	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
60	1, 4, 7-8, 22-23, 29, 60	23		60	
61	1, 5-7, 24-25, 27, 61	1	LR+ 8.4 [95% CI 1-73]	1	
62	1, 5-6, 8, 24, 26, 28, 62	8	LR+ 9.8 [95% CI 2.9-33.6]	1	
63	1, 5, 7-8, 25-26, 29, 63	29	LR+ 26.2 [95% CI 7.4-92]	29	
64	1, 6-8, 27-29, 64	29		29	
65	1-5, 9-11, 15-16, 20, 30-31, 35, 45, 65	45	LR+ 12.3 [95% CI 5.5-27.5]	1	
66	1-4, 6, 9-10, 12, 15, 17, 21, 30, 32, 36, 46, 66	32	LR+ 16.7 [95% CI 1.6-180.4]	1	
67	1-4, 7, 9-10, 13, 15, 18, 22, 30, 33, 37, 47, 67	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
68	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
69	1-3, 5-6, 9, 11-12, 16-17, 24, 31-32, 39, 49, 66, 69, 85	32	LR+ 16.7 [95% CI 1.6-180.4]	1	
70	1-3, 5, 7, 9, 11, 13, 16, 18, 25, 31, 33, 40, 50, 70	16	LR+ 10.3 [95% CI 5-21.2]	18	
71	1-3, 5, 8-9, 11, 14, 16, 19, 26, 31, 34, 41, 51, 71	51	LR+ 16.4 [95% CI 6.6-40.3]	1	
72	1-3, 6-7, 9, 12-13, 17-18, 27, 32-33, 42, 52, 66, 72	32	LR+ 16.7 [95% CI 1.6-180.4]	18	
73	1-3, 6, 8-9, 12, 14, 17, 19, 28, 32, 34, 43, 53, 66, 73, 79, 104	32		1	
74	1-3, 7-9, 13-14, 18-19, 29, 33-34, 44, 54, 74	29	LR+ 26.2 [95% CI 7.4-92]	18	
75	1-2, 4-6, 10-12, 20-21, 24, 35-36, 39, 55, 75	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
76	1-2, 4-5, 7, 10-11, 13, 20, 22, 25, 35, 37, 40, 56, 76	1		1	
77	1-2, 4-5, 8, 10-11, 14, 20, 23, 26, 35, 38, 41, 57, 77	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
78	1-2, 4, 6-7, 10, 12-13, 21-22, 27, 36-37, 42, 58, 78	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
79	1-2, 4, 6, 8, 10, 12, 14, 21, 23, 28, 36, 38, 43, 59, 79	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
80	1-2, 4, 7-8, 10, 13-14, 22-23, 29, 37-38, 44, 60, 80	23		44	

81	1-2, 5-7, 11-13, 24-25, 27, 39-40, 42, 61, 81	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
82	1-2, 5-6, 8, 11-12, 14, 24, 26, 28, 39, 41, 43, 62, 79, 82, 111	14	LR+ 16.4 [95% CI 3.1-85.5]	1	
83	1-2, 5, 7-8, 11, 13-14, 25-26, 29, 40-41, 44, 63, 83	29	LR+ 26.2 [95% CI 7.4-92]	44	
84	1-2, 6-8, 12-14, 27-29, 42-44, 64, 79, 84	29		44	
85	1, 3-6, 15-17, 20-21, 24, 45-46, 49, 55, 85	45	LR+ 12.3 [95% CI 5.5-27.5]	1	
86	1, 3-5, 7, 15-16, 18, 20, 22, 25, 45, 47, 50, 56, 86	45		18	
87	1, 3-5, 8, 15-16, 19-20, 23, 26, 45, 48, 51, 57, 87	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
88	1, 3-4, 6-7, 15, 17-18, 21-22, 27, 46-47, 52, 58, 88	18	LR+ 9.6 [95% CI 2.1-44.5]	18	
89	1, 3-4, 6, 8, 15, 17, 19, 21, 23, 28, 46, 48, 53, 59, 89	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
90	1, 3-4, 7-8, 15, 18-19, 22-23, 29, 47-48, 54, 60, 90	23		90	
91	1, 3, 5-7, 16-18, 24-25, 27, 49-50, 52, 61, 85, 91, 115	16	LR+ 10.3 [95% CI 5-21.2]	18	
92	1, 3, 5-6, 8, 16-17, 19, 24, 26, 28, 49, 51, 53, 62, 85, 92	51	LR+ 16.4 [95% CI 6.6-40.3]	1	
93	1, 3, 5, 7-8, 16, 18-19, 25-26, 29, 50-51, 54, 63, 93	29	LR+ 26.2 [95% CI 7.4-92]	18	
94	1, 3, 6-8, 17-19, 27-29, 52-54, 64, 94	29		18	
95	1, 4-7, 20-22, 24-25, 27, 55-56, 58, 61, 95	1	LR+ 8.4 [95% CI 1-73]	1	
96	1, 4-6, 8, 20-21, 23-24, 26, 28, 55, 57, 59, 62, 96	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
97	1, 4-5, 7-8, 20, 22-23, 25-26, 29, 56-57, 60, 63, 97	23		60	
98	1, 4, 6-8, 21-23, 27-29, 58-60, 64, 98	23		60	
99	1, 5-8, 24-29, 61-64, 99	29	LR+ 26.2 [95% CI 7.4-92]	29	
100	1-6, 9-12, 15-17, 20-21, 24, 30-32, 35-36, 39, 45-46, 49, 55, 65-66, 69, 75, 85, 100	32	LR+ 16.7 [95% CI 1.6-180.4]	1	
101	1-5, 7, 9-11, 13, 15-16, 18, 20, 22, 25, 30-31, 33, 35, 37, 40, 45, 47, 50, 56, 65, 67, 70, 76, 86, 101	45	LR+ 12.3 [95% CI 5.5-27.5]	18	

102	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
103	1-4, 6-7, 9-10, 12-13, 15, 17-18, 21-22, 27, 30, 32-33, 36-37, 42, 46-47, 52, 58, 66-67, 72, 78, 88, 103	103	LR+ 22.3 [95% CI 3.8-129.5]	18	
104	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
105	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	23		90	
106	1-3, 5-7, 9, 11-13, 16-18, 24-25, 27, 31-33, 39-40, 42, 49-50, 52, 61, 66, 69-70, 72, 81, 85, 91, 106, 115	32	LR+ 16.7 [95% CI 1.6-180.4]	18	
107	1-3, 5-6, 8-9, 11-12, 14, 16-17, 19, 24, 26, 28, 31-32, 34, 39, 41, 43, 49, 51, 53, 62, 66, 69, 71, 73, 79, 82, 85, 92, 104, 107, 111, 122	32		1	
108	1-3, 5, 7-9, 11, 13-14, 16, 18-19, 25-26, 29, 31, 33-34, 40-41, 44, 50-51, 54, 63, 70-71, 74, 83, 93, 108	29	LR+ 26.2 [95% CI 7.4-92]	18	
109	1-3, 6-9, 12-14, 17-19, 27-29, 32-34, 42-44, 52-54, 64, 66, 72-74, 79, 84, 94, 104, 109, 124	29		18	
110	1-2, 4-7, 10-13, 20-22, 24-25, 27, 35-37, 39-40, 42, 55-56, 58, 61, 75-76, 78, 81, 95, 110	1	LR+ 6.7 [95% CI 0.8-55.9]	1	
111	1-2, 4-6, 8, 10-12, 14, 20-21, 23-24, 26, 28, 35-36, 38-39, 41, 43, 55, 57, 59, 62, 75, 77, 79, 82, 96, 111	23	LR+ 32.7 [95% CI 6.9-154.7]	1	
112	1-2, 4-5, 7-8, 10-11, 13-14, 20, 22-23, 25-26, 29, 35, 37-38, 40-41, 44, 56-57, 60, 63, 76-77, 80, 83, 97, 112	23		44	
113	1-2, 4, 6-8, 10, 12-14, 21-23, 27-29, 36-38, 42-44, 58-60, 64, 78-80, 84, 98, 113	23		44	
114	1-2, 5-8, 11-14, 24-29, 39-44, 61-64, 79, 81-84, 99, 111, 114, 126	29	LR+ 26.2 [95% CI 7.4-92]	44	

115	1, 3-7, 15-18, 20-22, 24-25, 27, 45-47, 49-50, 52, 55-56, 58, 61, 85-86, 88, 91, 95, 115	45	LR+ 12.3 [95% CI 5.5-27.5]	18	
116	1, 3-6, 8, 15-17, 19-21, 23-24, 26, 28, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 85, 87, 89, 92, 96, 116	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
117	1, 3-5, 7-8, 15-16, 18-20, 22-23, 25-26, 29, 45, 47-48, 50-51, 54, 56-57, 60, 63, 86-87, 90, 93, 97, 117	23	LR+ 32.7 [95% CI 6.9-154.7]	90	
118	1, 3-4, 6-8, 15, 17-19, 21-23, 27-29, 46-48, 52-54, 58-60, 64, 88-90, 94, 98, 118	23	LR+ 32.7 [95% CI 6.9-154.7]	90	
119	1, 3, 5-8, 16-19, 24-29, 49-54, 61-64, 85, 91-94, 99, 115, 119, 127	29	LR+ 26.2 [95% CI 7.4-92]	18	
120	1, 4-8, 20-29, 55-64, 95-99, 120	23	LR+ 32.7 [95% CI 6.9-154.7]	60	
121	1-7, 9-13, 15-18, 20-22, 24-25, 27, 30-33, 35-37, 39-40, 42, 45-47, 49-50, 52, 55-56, 58, 61, 65-67, 69-70, 72, 75-76, 78, 81, 85-86, 88, 91, 95, 100-101, 103, 106, 110, 115, 121	103	LR+ 22.3 [95% CI 3.8-129.5]	18	
122	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	23	LR+ 32.7 [95% CI 6.9-154.7]	48	
123	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	23	LR+ 32.7 [95% CI 6.9-154.7]	90	
124	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	23	LR+ 32.7 [95% CI 6.9-154.7]	90	
125	1-3, 5-9, 11-14, 16-19, 24-29, 31-34, 39-44, 49-54, 61-64, 66, 69-74, 79, 81-85, 91-94, 99, 104, 106-109, 111, 114-115, 119, 122, 124-127	29	LR+ 26.2 [95% CI 7.4-92]	18	

126	1-2, 4-8, 10-14, 20-29, 35-44, 55-64, 75-84, 95-99, 110-114, 120, 126	23	LR+ 32.7 [95% CI 6.9-154.7]	44	
127	1, 3-8, 15-29, 45-64, 85-99, 115-120, 127	23	LR+ 32.7 [95% CI 6.9-154.7]	90	
128	1-128	23	LR+ 32.7 [95% CI 6.9-154.7]	90	

Table 0.10 Random forest ruling in and ruling out performance

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
1															
2									✓						
3	45	1 (0.1)	0 (0)	0.0						✓					
4											✓				
5	39	3 (0.2)	0 (0)	0.0	3	3 (0.2)	0 (0)	0.0				✓			
6													✓		
7														✓	
8															✓
9									✓	✓					
10									✓		✓				
11					7	1 (0.1)	0 (0)	0.0	✓			✓			
12									✓				✓		
13									✓					✓	
14									✓						✓
15										✓	✓				
16	46	1 (0.1)	0 (0)	0.0	4	2 (0.2)	0 (0)	0.0		✓		✓			
17										✓			✓		
18										✓				✓	
19										✓					✓
20	36	4 (0.3)	0 (0)	0.0	8	1 (0.1)	0 (0)	0.0			✓	✓			
21	4	8 (0.6)	2 (25)	11.2							✓		✓		
22											✓			✓	
23											✓				✓
24												✓	✓		

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
25					11	329 (25.8)	3 (0.9)	0.3				✓		✓	
26	40	3 (0.2)	0 (0)	0.0	5	2 (0.2)	0 (0)	0.0				✓			✓
27													✓	✓	
28													✓		✓
29														✓	✓
30									✓	✓	✓				
31					6	2 (0.2)	0 (0)	0.0	✓	✓		✓			
32	2	7 (0.5)	2 (28.6)	13.4					✓	✓			✓		
33									✓	✓				✓	
34	35	5 (0.4)	0 (0)	0.0					✓	✓					✓
35	37	4 (0.3)	0 (0)	0.0					✓		✓	✓			
36	30	15 (1.2)	2 (13.3)	5.1					✓		✓		✓		
37									✓		✓			✓	
38									✓		✓				✓
39									✓			✓	✓		
40					13	271 (21.3)	3 (1.1)	0.4	✓			✓		✓	
41	41	3 (0.2)	0 (0)	0.0					✓			✓			✓
42									✓				✓	✓	
43									✓				✓		✓
44									✓					✓	✓
45	5	4 (0.3)	1 (25)	11.2	1	10 (0.8)	0 (0)	0.0		✓	✓	✓			
46	22	12 (0.9)	2 (16.7)	6.7						✓	✓		✓		
47										✓	✓			✓	
48	34	5 (0.5)	0 (0)	0.0						✓	✓				✓
49										✓		✓	✓		
50										✓		✓		✓	

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
51	42	3 (0.2)	0 (0)	0.0						✓		✓			✓
52										✓			✓	✓	
53										✓			✓		✓
54										✓				✓	✓
55	20	11 (0.9)	2 (18.2)	7.4							✓	✓	✓		
56											✓	✓		✓	
57	38	4 (0.3)	0 (0)	0.0	2	5 (0.4)	0 (0)	0.0			✓	✓			✓
58											✓		✓	✓	
59											✓		✓		✓
60											✓			✓	✓
61												✓	✓	✓	
62												✓	✓		✓
63					14	320 (25.1)	6 (1.9)	0.6				✓		✓	✓
64													✓	✓	✓
65	12	5 (0.4)	1 (20)	8.4					✓	✓	✓	✓			
66	23	12 (0.9)	2 (16.7)	6.7					✓	✓	✓		✓		
67									✓	✓	✓			✓	
68	6	12 (1.3)	3 (25)	10.9					✓	✓	✓				✓
69	24	12 (0.9)	2 (16.7)	6.7					✓	✓		✓	✓		
70									✓	✓		✓		✓	
71	43	3 (0.2)	0 (0)	0.0	9	1 (0.1)	0 (0)	0.0	✓	✓		✓			✓
72									✓	✓			✓	✓	
73	1	6 (0.5)	2 (33.3)	16.7					✓	✓			✓		✓
74									✓	✓				✓	✓
75	13	5 (0.4)	1 (20)	8.4					✓		✓	✓	✓		
76									✓		✓	✓		✓	

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
77	14	5 (0.4)	1 (20)	8.4					✓		✓	✓			✓
78	15	5 (0.4)	1 (20)	8.4					✓		✓		✓	✓	
79	3	7 (0.5)	2 (28.6)	13.4					✓		✓		✓		✓
80									✓		✓			✓	✓
81									✓			✓	✓	✓	
82									✓			✓	✓		✓
83									✓			✓		✓	✓
84									✓				✓	✓	✓
85	11	10 (0.8)	2 (20)	8.4						✓	✓	✓	✓		
86										✓	✓	✓		✓	
87	44	3 (0.2)	0 (0)	0.0						✓	✓	✓			✓
88	16	5 (0.4)	1 (20)	8.4						✓	✓		✓	✓	
89	25	12 (0.9)	2 (16.7)	6.7						✓	✓		✓		✓
90										✓	✓			✓	✓
91										✓		✓	✓	✓	
92										✓		✓	✓		✓
93										✓		✓		✓	✓
94										✓			✓	✓	✓
95											✓	✓	✓	✓	
96	8	9 (0.7)	2 (22.2)	9.6							✓	✓	✓		✓
97											✓	✓		✓	✓
98	17	5 (0.4)	1 (20)	8.4							✓		✓	✓	✓
99												✓	✓	✓	✓
100	9	9 (0.7)	2 (22.2)	9.6					✓	✓	✓	✓	✓		
101									✓	✓	✓	✓		✓	
102	7	4 (0.4)	1 (25)	10.9					✓	✓	✓	✓			✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
103	33	17 (1.3)	2 (11.8)	4.5					✓	✓	✓		✓	✓	
104	27	13 (1)	2 (15.4)	6.1					✓	✓	✓		✓		✓
105					12	412 (32.3)	4 (1)	0.3	✓	✓	✓			✓	✓
106									✓	✓		✓	✓	✓	
107	26	12 (0.9)	2 (16.7)	6.7					✓	✓		✓	✓		✓
108									✓	✓		✓		✓	✓
109									✓	✓			✓	✓	✓
110					10	490 (38.4)	3 (0.6)	0.2	✓		✓	✓	✓	✓	
111									✓		✓	✓	✓		✓
112									✓		✓	✓		✓	✓
113	18	5 (0.4)	1 (20)	8.4					✓		✓		✓	✓	✓
114									✓			✓	✓	✓	✓
115	28	13 (1)	2 (15.4)	6.1						✓	✓	✓	✓	✓	
116	21	11 (0.9)	2 (18.2)	7.4						✓	✓	✓	✓		✓
117										✓	✓	✓		✓	✓
118	19	5 (0.4)	1 (20)	8.4						✓	✓		✓	✓	✓
119										✓		✓	✓	✓	✓
120											✓	✓	✓	✓	✓
121	32	8 (0.6)	1 (12.5)	4.8					✓	✓	✓	✓	✓	✓	
122	29	7 (0.5)	1 (14.3)	5.6					✓	✓	✓	✓	✓		✓
123									✓	✓	✓	✓		✓	✓
124	31	16 (1.3)	2 (12.5)	4.8					✓	✓	✓		✓	✓	✓
125									✓	✓		✓	✓	✓	✓
126									✓		✓	✓	✓	✓	✓
127	10	15 (1.2)	3 (20)	8.4						✓	✓	✓	✓	✓	✓
128									✓	✓	✓	✓	✓	✓	✓

Table 0.11: Selected random forest models for ruling in and ruling out for each scenario

Scenario	Usable models	Ruling in		Ruling out	
		Model	LR+	Model	LR-
1	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34	LR+ 0 [95% CI 0-NaN]		
2	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34			
3	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34			
4	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34			
5	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
6	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
7	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
8	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
9	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34	LR+ 0 [95% CI 0-NaN]		
10	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34			
11	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
12	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
13	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
14	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
15	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34	LR+ 0 [95% CI 0-NaN]		
16	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]

17	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
18	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
19	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
20	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
21	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
22	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
23	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
24	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
25	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
26	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
27	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
28	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			

29	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
30	1-4, 8-10, 14-15, 19, 23, 30, 34, 38	34	LR+ 0 [95% CI 0-NaN]		
31	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
32	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
33	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
34	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
35	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
36	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
37	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
38	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
39	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
40	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
41	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
42	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67,	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]

	72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124				
43	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
44	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
45	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
46	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
47	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
48	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
49	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
50	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
51	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
52	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]

53	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
54	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
55	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
56	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
57	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
58	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
59	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
60	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
61	1-47, 49-67, 69-101, 103-128	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
62	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	

63	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
64	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
65	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 51, 57, 65, 71, 77, 87	45	LR+ 11.2 [95% CI 1.2-104.7]	45	LR- 0 [95% CI 0-NaN]
66	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 53, 59, 66, 73, 79, 89, 104	73	LR+ 16.7 [95% CI 3.2-88.5]		
67	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47, 54, 60, 67, 74, 80, 90, 105	34	LR+ 0 [95% CI 0-NaN]	105	LR- 0.3 [95% CI 0.1-0.8]
68	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	68	LR+ 10.9 [95% CI 3.1-38]		
69	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
70	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
71	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
72	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]

73	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
74	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	
75	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
76	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
77	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
78	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
79	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
80	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
81	1-47, 49-67, 69-101, 103-128	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
82	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	

83	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
84	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
85	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73		45	LR- 0 [95% CI 0-NaN]
86	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
87	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
88	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
89	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
90	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
91	1-47, 49-67, 69-101, 103-128	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
92	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55,	73		45	

	57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122				
93	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
94	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
95	1-47, 49-67, 69-101, 103-128	73		45	LR- 0 [95% CI 0-NaN]
96	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	
97	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
98	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
99	1-128	73		45	LR- 0 [95% CI 0-NaN]
100	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 49, 51, 53, 55, 57, 59, 62, 65-66, 69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 104, 107, 111, 116, 122	73		45	

101	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47, 50-51, 54, 56-57, 60, 63, 65, 67, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
102	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	45		45	
103	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-47, 52-54, 58-60, 64, 66-67, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
104	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	73			
105	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	68	LR+ 10.9 [95% CI 3.1-38]	105	LR- 0.3 [95% CI 0.1-0.8]
106	1-47, 49-67, 69-101, 103-128	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
107	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	
108	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
109	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
110	1-47, 49-67, 69-101, 103-128	73		45	LR- 0 [95% CI 0-NaN]

111	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73	LR+ 16.7 [95% CI 3.2-88.5]	45	
112	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
113	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
114	1-128	73		45	LR- 0 [95% CI 0-NaN]
115	1-47, 49-67, 69-101, 103-128	73		45	
116	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	
117	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
118	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73	LR+ 16.7 [95% CI 3.2-88.5]	105	LR- 0.3 [95% CI 0.1-0.8]
119	1-128	73		45	LR- 0 [95% CI 0-NaN]

120	1-128	73		45	
121	1-47, 49-67, 69-101, 103-128	73		45	
122	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	73		45	
123	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	45	LR+ 11.2 [95% CI 1.2-104.7]	45	
124	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	73		105	LR- 0.3 [95% CI 0.1-0.8]
125	1-128	73	LR+ 16.7 [95% CI 3.2-88.5]	45	LR- 0 [95% CI 0-NaN]
126	1-128	73		45	
127	1-128	73		45	
128	1-128	73		45	

Table 0.12 LASSO ruling in and ruling out performance

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
1	29	9 (0.7)	1 (11.1)	4.2	11	42 (3.3)	0 (0)	0.0							
2	30	6 (0.5)	0 (0)	0.0	13	39 (3.1)	0 (0)	0.0	✓						
3	15	20 (1.6)	4 (20)	8.4	9	58 (4.5)	0 (0)	0.0		✓					
4	34	3 (0.2)	0 (0)	0.0	34	13 (1)	0 (0)	0.0			✓				
5					35	13 (1)	0 (0)	0.0				✓			
6	35	2 (0.2)	0 (0)	0.0									✓		
7	23	6 (0.5)	1 (16.7)	6.7	20	30 (2.4)	0 (0)	0.0						✓	
8															✓
9	19	22 (1.7)	4 (18.2)	7.4	94	116 (9.1)	2 (1.7)	0.6	✓	✓					
10	36	3 (0.2)	0 (0)	0.0	104	31 (2.4)	1 (3.2)	1.1	✓		✓				
11					17	34 (2.7)	0 (0)	0.0	✓			✓			
12	45	1 (0.1)	0 (0)	0.0	88	65 (5.1)	1 (1.5)	0.5	✓				✓		
13	31	6 (0.5)	0 (0)	0.0	18	33 (2.6)	0 (0)	0.0	✓					✓	
14					47	1 (0.1)	0 (0)	0.0	✓						✓
15	18	21 (1.6)	4 (19)	7.9	109	44 (3.5)	2 (4.5)	1.6		✓	✓				
16					105	85 (6.7)	3 (3.5)	1.2		✓		✓			
17	1	3 (0.2)	1 (33.3)	16.7	6	96 (7.5)	0 (0)	0.0		✓			✓		
18	26	13 (1)	2 (15.4)	6.1	44	4 (0.3)	0 (0)	0.0		✓				✓	
19	25	36 (4)	6 (16.7)	6.5	32	10 (1.1)	0 (0)	0.0		✓					✓
20					45	3 (0.2)	0 (0)	0.0			✓	✓			
21	37	2 (0.2)	0 (0)	0.0	77	77 (6)	1 (1.3)	0.4			✓		✓		
22	46	1 (0.1)	0 (0)	0.0							✓			✓	
23											✓				✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
24	38	3 (0.2)	0 (0)	0.0	83	70 (5.5)	1 (1.4)	0.5				✓	✓		
25					36	11 (0.9)	0 (0)	0.0				✓		✓	
26					73	463 (50.9)	6 (1.3)	0.4				✓			✓
27	47	1 (0.1)	0 (0)	0.0	96	56 (4.4)	1 (1.8)	0.6					✓	✓	
28					55	113 (12.4)	1 (0.9)	0.3					✓		✓
29	17	10 (1.1)	2 (20)	8.2	39	5 (0.5)	0 (0)	0.0						✓	✓
30	21	24 (1.9)	4 (16.7)	6.7	80	141 (11.1)	2 (1.4)	0.5	✓	✓	✓				
31					90	129 (10.1)	2 (1.6)	0.5	✓	✓		✓			
32	2	3 (0.2)	1 (33.3)	16.7	51	182 (14.3)	1 (0.5)	0.2	✓	✓			✓		
33	28	15 (1.2)	2 (13.3)	5.1	42	5 (0.4)	0 (0)	0.0	✓	✓				✓	
34	11	15 (1.6)	4 (26.7)	11.9	27	16 (1.8)	0 (0)	0.0	✓	✓					✓
35					29	18 (1.4)	0 (0)	0.0	✓		✓	✓			
36	48	1 (0.1)	0 (0)	0.0	87	68 (5.3)	1 (1.5)	0.5	✓		✓		✓		
37	27	14 (1.1)	2 (14.3)	5.6	108	48 (3.8)	2 (4.2)	1.5	✓		✓			✓	
38					43	4 (0.4)	0 (0)	0.0	✓		✓				✓
39	32	4 (0.3)	0 (0)	0.0					✓			✓	✓		
40					10	45 (3.5)	0 (0)	0.0	✓			✓		✓	
41					7	49 (5.4)	0 (0)	0.0	✓			✓			✓
42	39	2 (0.2)	0 (0)	0.0	92	61 (4.8)	1 (1.6)	0.6	✓				✓	✓	
43					69	86 (9.5)	1 (1.2)	0.4	✓				✓		✓
44					46	2 (0.2)	0 (0)	0.0	✓					✓	✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
45					26	26 (2)	0 (0)	0.0		✓	✓	✓			
46	3	3 (0.2)	1 (33.3)	16.7						✓	✓		✓		
47	22	12 (0.9)	2 (16.7)	6.7	40	6 (0.5)	0 (0)	0.0		✓	✓			✓	
48	16	25 (2.7)	5 (20)	8.2						✓	✓				✓
49					59	192 (15.1)	2 (1)	0.4		✓		✓	✓		
50					21	30 (2.4)	0 (0)	0.0		✓		✓		✓	
51					8	45 (4.9)	0 (0)	0.0		✓		✓			✓
52					82	72 (5.6)	1 (1.4)	0.5		✓			✓	✓	
53					65	91 (10)	1 (1.1)	0.4		✓			✓		✓
54	14	29 (3.2)	6 (20.7)	8.5	38	6 (0.7)	0 (0)	0.0		✓				✓	✓
55	40	3 (0.2)	0 (0)	0.0	101	80 (6.3)	2 (2.5)	0.9			✓	✓	✓		
56					37	11 (0.9)	0 (0)	0.0			✓	✓		✓	
57					74	305 (33.5)	4 (1.3)	0.4			✓	✓			✓
58	49	1 (0.1)	0 (0)	0.0	110	18 (1.4)	1 (5.6)	2.0			✓		✓	✓	
59					103	32 (3.5)	1 (3.1)	1.1			✓		✓		✓
60					48	1 (0.1)	0 (0)	0.0			✓			✓	✓
61	41	3 (0.2)	0 (0)	0.0								✓	✓	✓	
62					23	21 (2.3)	0 (0)	0.0				✓	✓		✓
63					14	28 (3.1)	0 (0)	0.0				✓		✓	✓
64													✓	✓	✓
65					95	58 (4.5)	1 (1.7)	0.6	✓	✓	✓	✓			
66	4	3 (0.2)	1 (33.3)	16.7					✓	✓	✓		✓		
67	20	11 (0.9)	2 (18.2)	7.4	41	7 (0.5)	0 (0)	0.0	✓	✓	✓			✓	
68	24	42 (4.6)	7 (16.7)	6.5	49	1 (0.1)	0 (0)	0.0	✓	✓	✓				✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
69					4	107 (8.4)	0 (0)	0.0	✓	✓		✓	✓		
70					62	105 (8.2)	1 (1)	0.3	✓	✓		✓		✓	
71	7	10 (1.1)	3 (30)	14.0	86	65 (7.1)	1 (1.5)	0.5	✓	✓		✓			✓
72	5	3 (0.2)	1 (33.3)	16.7	78	77 (6)	1 (1.3)	0.4	✓	✓			✓	✓	
73					2	89 (9.8)	0 (0)	0.0	✓	✓			✓		✓
74	10	11 (1.2)	3 (27.3)	12.3	15	27 (3)	0 (0)	0.0	✓	✓				✓	✓
75	33	4 (0.3)	0 (0)	0.0	53	149 (11.7)	1 (0.7)	0.2	✓		✓	✓	✓		
76					28	19 (1.5)	0 (0)	0.0	✓		✓	✓		✓	
77					54	442 (48.6)	4 (0.9)	0.3	✓		✓	✓			✓
78	50	1 (0.1)	0 (0)	0.0	97	53 (4.2)	1 (1.9)	0.6	✓		✓		✓	✓	
79					12	29 (3.2)	0 (0)	0.0	✓		✓		✓		✓
80					50	1 (0.1)	0 (0)	0.0	✓		✓			✓	✓
81	42	3 (0.2)	0 (0)	0.0					✓			✓	✓	✓	
82					52	135 (14.8)	1 (0.7)	0.2	✓			✓	✓		✓
83					22	22 (2.4)	0 (0)	0.0	✓			✓		✓	✓
84					81	74 (8.1)	1 (1.4)	0.4	✓				✓	✓	✓
85					57	114 (8.9)	1 (0.9)	0.3		✓	✓	✓	✓		
86					102	35 (2.7)	1 (2.9)	1.0		✓	✓	✓		✓	
87					31	11 (1.2)	0 (0)	0.0		✓	✓	✓			✓
88	6	3 (0.2)	1 (33.3)	16.7	72	83 (6.5)	1 (1.2)	0.4		✓	✓		✓	✓	

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
89										✓	✓		✓		✓
90	13	18 (2)	4 (22.2)	9.3	30	12 (1.3)	0 (0)	0.0		✓	✓			✓	✓
91					68	161 (12.6)	2 (1.2)	0.4		✓		✓	✓	✓	
92					60	104 (11.4)	1 (1)	0.3		✓		✓	✓		✓
93					33	10 (1.1)	0 (0)	0.0		✓		✓		✓	✓
94										✓			✓	✓	✓
95	43	3 (0.2)	0 (0)	0.0	89	65 (5.1)	1 (1.5)	0.5			✓	✓	✓	✓	
96					70	84 (9.2)	1 (1.2)	0.4			✓	✓	✓		✓
97					79	424 (46.6)	6 (1.4)	0.5			✓	✓		✓	✓
98					85	66 (7.3)	1 (1.5)	0.5			✓		✓	✓	✓
99					25	20 (2.2)	0 (0)	0.0				✓	✓	✓	✓
100					3	114 (8.9)	0 (0)	0.0	✓	✓	✓	✓	✓		
101					99	46 (3.6)	1 (2.2)	0.7	✓	✓	✓	✓		✓	
102					24	21 (2.3)	0 (0)	0.0	✓	✓	✓	✓			✓
103	12	4 (0.3)	1 (25)	11.2	76	78 (6.1)	1 (1.3)	0.4	✓	✓	✓		✓	✓	
104					61	101 (11.1)	1 (1)	0.3	✓	✓	✓		✓		✓
105	8	10 (1.1)	3 (30)	14.0	16	26 (2.9)	0 (0)	0.0	✓	✓	✓			✓	✓
106					5	100 (7.8)	0 (0)	0.0	✓	✓		✓	✓	✓	
107									✓	✓		✓	✓		✓
108	9	10 (1.1)	3 (30)	14.0	91	64 (7)	1 (1.6)	0.5	✓	✓		✓		✓	✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
109					71	81 (8.9)	1 (1.2)	0.4	✓	✓			✓	✓	✓
110	44	3 (0.2)	0 (0)	0.0	93	61 (4.8)	1 (1.6)	0.6	✓		✓	✓	✓	✓	
111					64	93 (10.2)	1 (1.1)	0.4	✓		✓	✓	✓		✓
112					66	434 (47.7)	5 (1.2)	0.4	✓		✓	✓		✓	✓
113					106	27 (3)	1 (3.7)	1.3	✓		✓		✓	✓	✓
114					75	79 (8.7)	1 (1.3)	0.4	✓			✓	✓	✓	✓
115					100	88 (6.9)	2 (2.3)	0.8		✓	✓	✓	✓	✓	
116					56	107 (11.8)	1 (0.9)	0.3		✓	✓	✓	✓		✓
117					67	413 (45.4)	5 (1.2)	0.4		✓	✓	✓		✓	✓
118										✓	✓		✓	✓	✓
119					58	195 (21.4)	2 (1)	0.3		✓		✓	✓	✓	✓
120					19	23 (2.5)	0 (0)	0.0			✓	✓	✓	✓	✓
121					63	104 (8.2)	1 (1)	0.3	✓	✓	✓	✓	✓	✓	
122					1	110 (12.1)	0 (0)	0.0	✓	✓	✓	✓	✓		✓
123					84	67 (7.4)	1 (1.5)	0.5	✓	✓	✓	✓		✓	✓
124									✓	✓	✓		✓	✓	✓
125									✓	✓		✓	✓	✓	✓
126					107	25 (2.7)	1 (4)	1.4	✓		✓	✓	✓	✓	✓

	Ruling in				Ruling out				Variable groups						
model	rank	Very high risk N (%)	N outcomes in very high risk group (%)	LR+	rank	Very low risk N (%)	N outcomes in very low risk group (%)	LR-	Patient information	Symptoms	Medical history	Signs	Bloods part 1	Bloods part 2	Urine
127					98	101 (11.1)	2 (2)	0.7		✓	✓	✓	✓	✓	✓
128									✓	✓	✓	✓	✓	✓	✓

Table 0.13: Selected LASSO models for ruling in and ruling out for each scenario

Scenario	Usable models	Ruling in		Ruling out	
		Model	LR+	Model	LR-
1	1	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
2	1-2	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
3	1, 3	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
4	1, 4	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
5	1, 5	1		1	
6	1, 6	1	LR+ 4.8 [95% CI 0.6-37.9]	1	
7	1, 7	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
8	1, 8	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
9	1-3, 9	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
10	1-2, 4, 10	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
11	1-2, 5, 11	1		1	
12	1-2, 6, 12	1		1	
13	1-2, 7, 13	7	LR+ 5.6 [95% CI 0.7-45.2]	1	
14	1-2, 8, 14	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
15	1, 3-4, 15	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
16	1, 3, 5, 16	3		3	
17	1, 3, 6, 17	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
18	1, 3, 7, 18	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
19	1, 3, 8, 19	3		3	
20	1, 4-5, 20	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
21	1, 4, 6, 21	1		1	LR- 0 [95% CI 0-NaN]

22	1, 4, 7, 22	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
23	1, 4, 8, 23	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
24	1, 5-6, 24	1		1	
25	1, 5, 7, 25	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
26	1, 5, 8, 26	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
27	1, 6-7, 27	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
28	1, 6, 8, 28	1	LR+ 4.8 [95% CI 0.6-37.9]	1	
29	1, 7-8, 29	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
30	1-4, 9-10, 15, 30	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
31	1-3, 5, 9, 11, 16, 31	3		3	
32	1-3, 6, 9, 12, 17, 32	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
33	1-3, 7, 9, 13, 18, 33	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
34	1-3, 8-9, 14, 19, 34	34	LR+ 11.9 [95% CI 4-35]	3	
35	1-2, 4-5, 10-11, 20, 35	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
36	1-2, 4, 6, 10, 12, 21, 36	1		1	
37	1-2, 4, 7, 10, 13, 22, 37	7	LR+ 5.6 [95% CI 0.7-45.2]	1	
38	1-2, 4, 8, 10, 14, 23, 38	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
39	1-2, 5-6, 11-12, 24, 39	1		1	
40	1-2, 5, 7, 11, 13, 25, 40	7	LR+ 5.6 [95% CI 0.7-45.2]	40	
41	1-2, 5, 8, 11, 14, 26, 41	1	LR+ 3.7 [95% CI 0.5-28.6]	41	
42	1-2, 6-7, 12-13, 27, 42	7	LR+ 5.6 [95% CI 0.7-45.2]	1	
43	1-2, 6, 8, 12, 14, 28, 43	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
44	1-2, 7-8, 13-14, 29, 44	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
45	1, 3-5, 15-16, 20, 45	3	LR+ 8.4 [95% CI 2.9-23.8]	3	

46	1, 3-4, 6, 15, 17, 21, 46	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
47	1, 3-4, 7, 15, 18, 22, 47	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
48	1, 3-4, 8, 15, 19, 23, 48	3		3	
49	1, 3, 5-6, 16-17, 24, 49	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
50	1, 3, 5, 7, 16, 18, 25, 50	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
51	1, 3, 5, 8, 16, 19, 26, 51	3		51	
52	1, 3, 6-7, 17-18, 27, 52	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
53	1, 3, 6, 8, 17, 19, 28, 53	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
54	1, 3, 7-8, 18-19, 29, 54	54	LR+ 8.5 [95% CI 3.8-19.2]	3	
55	1, 4-6, 20-21, 24, 55	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
56	1, 4-5, 7, 20, 22, 25, 56	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
57	1, 4-5, 8, 20, 23, 26, 57	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
58	1, 4, 6-7, 21-22, 27, 58	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
59	1, 4, 6, 8, 21, 23, 28, 59	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
60	1, 4, 7-8, 22-23, 29, 60	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
61	1, 5-7, 24-25, 27, 61	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
62	1, 5-6, 8, 24, 26, 28, 62	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
63	1, 5, 7-8, 25-26, 29, 63	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
64	1, 6-8, 27-29, 64	29		1	
65	1-5, 9-11, 15-16, 20, 30-31, 35, 45, 65	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
66	1-4, 6, 9-10, 12, 15, 17, 21, 30, 32, 36, 46, 66	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
67	1-4, 7, 9-10, 13, 15, 18, 22, 30, 33, 37, 47, 67	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
68	1-4, 8-10, 14-15, 19, 23, 30, 34, 38, 48, 68	34	LR+ 11.9 [95% CI 4-35]	3	
69	1-3, 5-6, 9, 11-12, 16-17, 24, 31-32, 39, 49, 69	17	LR+ 16.7 [95% CI 1.6-180.4]	69	

70	1-3, 5, 7, 9, 11, 13, 16, 18, 25, 31, 33, 40, 50, 70	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
71	1-3, 5, 8-9, 11, 14, 16, 19, 26, 31, 34, 41, 51, 71	71	LR+ 14 [95% CI 3.8-51.3]	41	
72	1-3, 6-7, 9, 12-13, 17-18, 27, 32-33, 42, 52, 72	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
73	1-3, 6, 8-9, 12, 14, 17, 19, 28, 32, 34, 43, 53, 73	17		73	
74	1-3, 7-9, 13-14, 18-19, 29, 33-34, 44, 54, 74	74	LR+ 12.3 [95% CI 3.4-43.7]	3	
75	1-2, 4-6, 10-12, 20-21, 24, 35-36, 39, 55, 75	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
76	1-2, 4-5, 7, 10-11, 13, 20, 22, 25, 35, 37, 40, 56, 76	7	LR+ 5.6 [95% CI 0.7-45.2]	40	
77	1-2, 4-5, 8, 10-11, 14, 20, 23, 26, 35, 38, 41, 57, 77	1	LR+ 3.7 [95% CI 0.5-28.6]	41	
78	1-2, 4, 6-7, 10, 12-13, 21-22, 27, 36-37, 42, 58, 78	7	LR+ 5.6 [95% CI 0.7-45.2]	1	
79	1-2, 4, 6, 8, 10, 12, 14, 21, 23, 28, 36, 38, 43, 59, 79	1	LR+ 3.7 [95% CI 0.5-28.6]	1	
80	1-2, 4, 7-8, 10, 13-14, 22-23, 29, 37-38, 44, 60, 80	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
81	1-2, 5-7, 11-13, 24-25, 27, 39-40, 42, 61, 81	7	LR+ 5.6 [95% CI 0.7-45.2]	40	
82	1-2, 5-6, 8, 11-12, 14, 24, 26, 28, 39, 41, 43, 62, 82	1	LR+ 3.7 [95% CI 0.5-28.6]	41	
83	1-2, 5, 7-8, 11, 13-14, 25-26, 29, 40-41, 44, 63, 83	29	LR+ 8.2 [95% CI 1.8-36.7]	41	
84	1-2, 6-8, 12-14, 27-29, 42-44, 64, 84	29		1	
85	1, 3-6, 15-17, 20-21, 24, 45-46, 49, 55, 85	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
86	1, 3-5, 7, 15-16, 18, 20, 22, 25, 45, 47, 50, 56, 86	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
87	1, 3-5, 8, 15-16, 19-20, 23, 26, 45, 48, 51, 57, 87	3		51	
88	1, 3-4, 6-7, 15, 17-18, 21-22, 27, 46-47, 52, 58, 88	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
89	1, 3-4, 6, 8, 15, 17, 19, 21, 23, 28, 46, 48, 53, 59, 89	17		17	
90	1, 3-4, 7-8, 15, 18-19, 22-23, 29, 47-48, 54, 60, 90	90	LR+ 9.3 [95% CI 3.3-26.5]	3	
91	1, 3, 5-7, 16-18, 24-25, 27, 49-50, 52, 61, 91	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
92	1, 3, 5-6, 8, 16-17, 19, 24, 26, 28, 49, 51, 53, 62, 92	17		17	
93	1, 3, 5, 7-8, 16, 18-19, 25-26, 29, 50-51, 54, 63, 93	54	LR+ 8.5 [95% CI 3.8-19.2]	51	

94	1, 3, 6-8, 17-19, 27-29, 52-54, 64, 94	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
95	1, 4-7, 20-22, 24-25, 27, 55-56, 58, 61, 95	7	LR+ 6.7 [95% CI 0.8-55.9]	1	
96	1, 4-6, 8, 20-21, 23-24, 26, 28, 55, 57, 59, 62, 96	1	LR+ 4.2 [95% CI 0.5-32.6]	1	
97	1, 4-5, 7-8, 20, 22-23, 25-26, 29, 56-57, 60, 63, 97	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
98	1, 4, 6-8, 21-23, 27-29, 58-60, 64, 98	29		1	
99	1, 5-8, 24-29, 61-64, 99	29		1	
100	1-6, 9-12, 15-17, 20-21, 24, 30-32, 35-36, 39, 45-46, 49, 55, 65-66, 69, 75, 85, 100	17	LR+ 16.7 [95% CI 1.6-180.4]	100	
101	1-5, 7, 9-11, 13, 15-16, 18, 20, 22, 25, 30-31, 33, 35, 37, 40, 45, 47, 50, 56, 65, 67, 70, 76, 86, 101	3	LR+ 8.4 [95% CI 2.9-23.8]	3	
102	1-5, 8-11, 14-16, 19-20, 23, 26, 30-31, 34-35, 38, 41, 45, 48, 51, 57, 65, 68, 71, 77, 87, 102	71	LR+ 14 [95% CI 3.8-51.3]	41	
103	1-4, 6-7, 9-10, 12-13, 15, 17-18, 21-22, 27, 30, 32-33, 36-37, 42, 46-47, 52, 58, 66-67, 72, 78, 88, 103	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
104	1-4, 6, 8-10, 12, 14-15, 17, 19, 21, 23, 28, 30, 32, 34, 36, 38, 43, 46, 48, 53, 59, 66, 68, 73, 79, 89, 104	17		73	
105	1-4, 7-10, 13-15, 18-19, 22-23, 29-30, 33-34, 37-38, 44, 47-48, 54, 60, 67-68, 74, 80, 90, 105	105	LR+ 14 [95% CI 3.8-51.3]	3	
106	1-3, 5-7, 9, 11-13, 16-18, 24-25, 27, 31-33, 39-40, 42, 49-50, 52, 61, 69-70, 72, 81, 91, 106	17	LR+ 16.7 [95% CI 1.6-180.4]	69	
107	1-3, 5-6, 8-9, 11-12, 14, 16-17, 19, 24, 26, 28, 31-32, 34, 39, 41, 43, 49, 51, 53, 62, 69, 71, 73, 82, 92, 107	17		73	
108	1-3, 5, 7-9, 11, 13-14, 16, 18-19, 25-26, 29, 31, 33-34, 40-41, 44, 50-51, 54, 63, 70-71, 74, 83, 93, 108	71	LR+ 14 [95% CI 3.8-51.3]	41	
109	1-3, 6-9, 12-14, 17-19, 27-29, 32-34, 42-44, 52-54, 64, 72-74, 84, 94, 109	17	LR+ 16.7 [95% CI 1.6-180.4]	73	
110	1-2, 4-7, 10-13, 20-22, 24-25, 27, 35-37, 39-40, 42, 55-56, 58, 61, 75-76, 78, 81, 95, 110	7	LR+ 5.6 [95% CI 0.7-45.2]	40	
111	1-2, 4-6, 8, 10-12, 14, 20-21, 23-24, 26, 28, 35-36, 38-39, 41, 43, 55, 57, 59, 62, 75, 77, 79, 82, 96, 111	1	LR+ 3.7 [95% CI 0.5-28.6]	41	

112	1-2, 4-5, 7-8, 10-11, 13-14, 20, 22-23, 25-26, 29, 35, 37-38, 40-41, 44, 56-57, 60, 63, 76-77, 80, 83, 97, 112	29	LR+ 8.2 [95% CI 1.8-36.7]	41	
113	1-2, 4, 6-8, 10, 12-14, 21-23, 27-29, 36-38, 42-44, 58-60, 64, 78-80, 84, 98, 113	29		1	
114	1-2, 5-8, 11-14, 24-29, 39-44, 61-64, 81-84, 99, 114	29		41	
115	1, 3-7, 15-18, 20-22, 24-25, 27, 45-47, 49-50, 52, 55-56, 58, 61, 85-86, 88, 91, 95, 115	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
116	1, 3-6, 8, 15-17, 19-21, 23-24, 26, 28, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 85, 87, 89, 92, 96, 116	17		17	
117	1, 3-5, 7-8, 15-16, 18-20, 22-23, 25-26, 29, 45, 47-48, 50-51, 54, 56-57, 60, 63, 86-87, 90, 93, 97, 117	90	LR+ 9.3 [95% CI 3.3-26.5]	51	
118	1, 3-4, 6-8, 15, 17-19, 21-23, 27-29, 46-48, 52-54, 58-60, 64, 88-90, 94, 98, 118	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
119	1, 3, 5-8, 16-19, 24-29, 49-54, 61-64, 91-94, 99, 119	17		17	
120	1, 4-8, 20-29, 55-64, 95-99, 120	29	LR+ 8.2 [95% CI 1.8-36.7]	1	
121	1-7, 9-13, 15-18, 20-22, 24-25, 27, 30-33, 35-37, 39-40, 42, 45-47, 49-50, 52, 55-56, 58, 61, 65-67, 69-70, 72, 75-76, 78, 81, 85-86, 88, 91, 95, 100-101, 103, 106, 110, 115, 121	17	LR+ 16.7 [95% CI 1.6-180.4]	100	
122	1-6, 8-12, 14-17, 19-21, 23-24, 26, 28, 30-32, 34-36, 38-39, 41, 43, 45-46, 48-49, 51, 53, 55, 57, 59, 62, 65-66, 68-69, 71, 73, 75, 77, 79, 82, 85, 87, 89, 92, 96, 100, 102, 104, 107, 111, 116, 122	17		122	
123	1-5, 7-11, 13-16, 18-20, 22-23, 25-26, 29-31, 33-35, 37-38, 40-41, 44-45, 47-48, 50-51, 54, 56-57, 60, 63, 65, 67-68, 70-71, 74, 76-77, 80, 83, 86-87, 90, 93, 97, 101-102, 105, 108, 112, 117, 123	71	LR+ 14 [95% CI 3.8-51.3]	41	
124	1-4, 6-10, 12-15, 17-19, 21-23, 27-30, 32-34, 36-38, 42-44, 46-48, 52-54, 58-60, 64, 66-68, 72-74, 78-80, 84, 88-90, 94, 98, 103-105, 109, 113, 118, 124	17	LR+ 16.7 [95% CI 1.6-180.4]	73	
125	1-3, 5-9, 11-14, 16-19, 24-29, 31-34, 39-44, 49-54, 61-64, 69-74, 81-84, 91-94, 99, 106-109, 114, 119, 125	17		73	

126	1-2, 4-8, 10-14, 20-29, 35-44, 55-64, 75-84, 95-99, 110-114, 120, 126	29	LR+ 8.2 [95% CI 1.8-36.7]	41	
127	1, 3-8, 15-29, 45-64, 85-99, 115-120, 127	17	LR+ 16.7 [95% CI 1.6-180.4]	17	
128	1-128	17		122	