



BMJ Open Validating machine learning models for the prediction of labour induction intervention using routine data: a registry-based retrospective cohort study at a tertiary hospital in northern Tanzania

Clifford Silver Tarimo ^{1,2}, Soumitra S Bhuyan,³ Quanman Li,¹ Michael Johnson J Mahande ⁴, Jian Wu,¹ Xiaoli Fu¹

To cite: Tarimo CS, Bhuyan SS, Li Q, *et al*. Validating machine learning models for the prediction of labour induction intervention using routine data: a registry-based retrospective cohort study at a tertiary hospital in northern Tanzania. *BMJ Open* 2021;**11**:e051925. doi:10.1136/bmjopen-2021-051925

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2021-051925>).

Received 31 March 2021
Accepted 11 October 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Xiaoli Fu; xlfu66@126.com

ABSTRACT

Objectives We aimed at identifying the important variables for labour induction intervention and assessing the predictive performance of machine learning algorithms.

Setting We analysed the birth registry data from a referral hospital in northern Tanzania. Since July 2000, every birth at this facility has been recorded in a specific database.

Participants 21 578 deliveries between 2000 and 2015 were included. Deliveries that lacked information regarding the labour induction status were excluded.

Primary outcome Deliveries involving labour induction intervention.

Results Parity, maternal age, body mass index, gestational age and birth weight were all found to be important predictors of labour induction. Boosting method demonstrated the best discriminative performance (area under curve, AUC=0.75; 95% CI (0.73 to 0.76)) while logistic regression presented the least (AUC=0.71; 95% CI (0.70 to 0.73)). Random forest and boosting algorithms showed the highest net-benefits as per the decision curve analysis.

Conclusion All of the machine learning algorithms performed well in predicting the likelihood of labour induction intervention. Further optimisation of these classifiers through hyperparameter tuning may result in an improved performance. Extensive research into the performance of other classifier algorithms is warranted.

BACKGROUND

Induction of labour (IOL) is one of the most famous obstetric procedures involving artificially stimulating uterine contractions before they begin spontaneously.^{1–3} Mechanical means or commercially available pharmaceuticals may be used to carry out the procedure.^{4–6} The key indicators for IOL intervention may be grouped as maternal or fetal or both.⁷ Rate of IOL use has increased gradually over the last few decades, owing

Strengths and limitations of this study

- In this modelling, we enrolled a number of deliveries from an extended period (15 years), a sample that have accommodated a diversified group of study participants with contrasting characteristics.
- This is the first study that applied the most popular machine learning algorithms to predict the use of induction of labour intervention in Tanzania.
- The study involved only the deliveries attended at the Kilimanjaro Christian Medical Centre hospital, hence the research output may not be applicable to other hospital setting in Tanzania.

to a greater emphasis on improving pregnancy outcomes.⁸ The prevalence of IOL varies greatly between countries and regions across the world, but higher rates have been recorded in developed countries.^{9–10} Induced deliveries accounts for 20% of all deliveries in the UK and the USA, while in African regions it accounts for only 4.4%.^{11–14}

Major indications for IOL in Sub-Saharan Africa include postdates, intrauterine growth restrictions, fetal macrosomia, oligohydramnios, gestational diabetes, chorioamnionitis, prelabour rupture of membranes (PROM) and hypertensive disorders of pregnancy.^{15–17} However, WHO recommends IOL as a therapeutic option only when the benefits of pregnancy termination surpass the risks of its continuation.¹⁸ Following the growing interest in developing and integrating the AI-based clinical decision support systems in SSA, we validated and assessed the predictive performance of machine learning (ML) models for predicting IOL intervention in obstetrics department. ML approach in healthcare data

is important due to its robustness and incredible ability to learn and classify input data.^{19–25} This prior intelligence may aid in effective resource allocation and mobilisation which eventually plays a significant role in improving pregnancy outcomes. The validation of ML algorithms will as well benefit many other domains including risk management, tailored health communications, decision support system and personalised medicine.

METHODS

Study setting, data source and study design

This study was carried out at Kilimanjaro Christian Medical Centre (KCMC), one of Tanzania's four tertiary hospitals. The facility is located in Moshi urban district, northern Tanzania, serving Kilimanjaro residents as well as nearby regions. Since 2000, the hospital has been recording pregnancy, delivery and newborn information in a database system. In the current study, we analysed deliveries registered from the year 2000 to 2015. Skilled nurses conduct daily interview sessions after each delivery. Interviews are done using structured questionnaire. Records from the hospital birth registry database cover sociodemographic information as well as predelivery and postdelivery mother's health status. More details on KCMC medical birth registry procedures were described elsewhere.²⁶ In our analyses, we excluded observations with missing information on IOL status, deliveries with missing values on covariates as well as those presented non-vertex alignment (figure 1). We remained with 21 578 deliveries that constituted to our final sample size.

Description of the study variables

The main outcome of interest was 'IOL' whereby induced delivery was coded '1' while spontaneous delivery was coded '0' during the analysis. The facility implements IOL in various ways as per the WHO recommendation

for IOL. These include administration of oxytocin infusion, prostaglandins, prostaglandin analogues and the use of mechanical methods such as digital stretching of the cervix and membrane sweeping, hygroscopic cervical dilators, extra-amniotic balloon catheters and artificial rupture of the membranes. Predictors for IOL included maternal characteristics such as maternal age, parity status (nulliparous vs multiparous), gestational age (preterm (<37 weeks of pregnancy), term (37–41 weeks) and post-term (>41 weeks), multiple gestation status (yes, no), maternal body mass index (BMI) (underweight (<18.5 kg/m²), normal weight (18.5–25 kg/m²), overweight (25–30 kg/m²) and obese (≥30 kg/m²)), birth weight (low (<2.5 kg), normal (2.5–3.5 kg), high (>3.5 kg)), referral status (whether referred for delivery or not), number of antenatal care visits (<4 and 4 visits), PROM (categorised as binary, yes, no), alcohol consumption during pregnancy (categorised as binary, yes/no) plus many others as indicated in table 1.

Statistical and ML analyses

Data analysis was performed using R package V.4.0.3. Pearson χ^2 test was used to determine association between a set of fetomaternal variables and IOL status. The primary outcome that was assessed in the current study was either the delivery was induced using any method (mechanically or pharmaceutical) or it was achieved spontaneously.

Important features, model validation and decision curve analysis

While it has been shown that variable importance is an essential aspect in maintaining the accuracy of predictive models, we used an inbuilt function in random forest (RF) algorithm in R known as 'VarImp' for identifying and displaying top five covariates that are highly predictive of IOL intervention in the database.^{27 28}

In predictive modelling, we compared the performance of logistic regression (Lreg), RF, naïve Bayes (NB), artificial neural networks (ANN), boosting and bagging algorithms in predicting the use of IOL. RF consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the RF spits out a class prediction and the class with the most votes become the model's prediction.^{29 30} In this model, we estimated Out-of-Bag error (tested against training data subsets that are not included in subtree construction) and validation error (tested against the test data) to come up with the best possible predictive model.³¹ We used the 'Random-Forest' package to build this model in R. NB is a robust classifier algorithm belonging to a family of simple probabilistic classifiers based on Bayes' theorem with strong independence and equal-importance assumptions among the features under observation.^{32 33} The calculation of Bayes theorem can be simplified by making some assumptions, such as each input variable is independent of all other input variables.^{34 35} We used 'naiveBayes' function in R-package to fit NB models. The ANN is an inspired

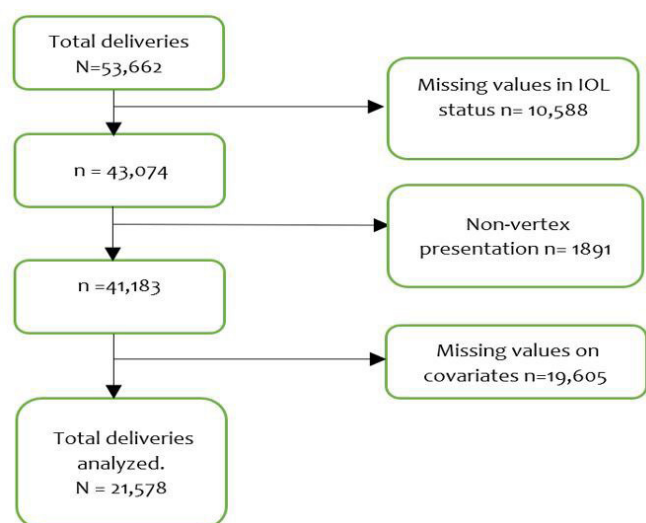


Figure 1 CONSORT diagram for sample size determination. CONSORT, Consolidated Standards of Reporting Trials; IOL, induction of labour.

Table 1 Sociodemographic characteristics of study participant (N=21 578)

Characteristics	Induced delivery n (%)	Spontaneous delivery n (%)	χ^2 p value
Maternal age			
<25	3728 (42.3)	4304 (33.72)	
25–35	4162 (47.22)	6122 (47.96)	
>35	924 (10.48)	2338 (18.32)	<0.001
Maternal religion			
Muslim	3464 (39.3)	4830 (37.84)	
Christian	5350 (60.7)	7934 (62.16)	0.03
Gestational age			
Preterm	1018 (11.55)	1802 (14.12)	
Term	7029 (79.75)	9906 (77.61)	
Post-term	767 (8.70)	1056 (8.27)	<0.001
Maternal residence			
Rural	2904 (32.95)	4795 (37.57)	
Urban	5910 (67.05)	7969 (62.43)	<0.001
Maternal BMI			
Underweight	57 (0.65)	72 (0.56)	
Normal weight	2807 (31.85)	3813 (29.87)	
Overweight	3667 (41.60)	5022 (39.35)	
Obese	2283 (25.90)	3857 (30.22)	<0.001
Birth weight			
Low	975 (11.06)	1481 (11.60)	
Normal	6245 (70.85)	8786 (68.83)	
High	1594 (18.08)	2497 (19.56)	0.006
Parity status			
Nulliparous	6036 (68.48)	3964 (31.06)	
Multiparous	2778 (31.52)	8800 (68.94)	<0.001
Multiple gestation			
No	8466 (96.05)	12 171 (95.35)	
Yes	348 (3.95)	593 (4.65)	0.014
PROM			
No	8623 (97.83)	12 567 (98.46)	
Yes	191 (2.17)	197 (1.54)	0.001
Child sex			
Male	4275 (48.5)	6243 (48.91)	
Female	4539 (51.50)	6521 (51.09)	0.555
Circumcision			
No	7810 (88.61)	10 785 (84.50)	
Yes	1004 (11.39)	1979 (15.50)	<0.001
Referred for delivery			
No	7418 (84.16)	10 517 (82.4)	
Yes	1396 (15.84)	2247 (17.60)	0.001
Alcohol use during pregnancy			

Continued

Table 1 Continued

Characteristics	Induced delivery n (%)	Spontaneous delivery n (%)	χ^2 p value
No	6999 (79.41)	9578 (75.04)	
Yes	1815 (20.59)	3186 (24.96)	<0.001
Maternal occupation			
Employed	6570 (55.64)	4210 (43.09)	
Unemployed	5237 (44.36)	5561 (56.91)	<0.001
Maternal education			
None	1378 (18.97)	3547 (24.78)	
Primary	3231 (44.48)	5879 (41.07)	
Secondary	1756 (24.17)	2877 (20.10)	
Higher	899 (12.38)	2011 (14.05)	<0.001
No of ANC visits			
≥4	4874 (62.72)	9765 (70.72)	
<4	2897 (37.28)	4042 (29.28)	<0.001
Marital status			
Married	7479 (84.85)	11 459 (89.78)	
Not married	1335 (15.15)	1305 (10.22)	<0.001

ANC, Antenatal care; BMI, body mass index; PROM, prelabour rupture of membrane.

computational model of biological neural networks meant to imitate the human brain.^{36 37} The algorithm can learn from intertwined inputs, hidden and output layers to achieve the desired outputs while guided by set of rules in Backpropagation process. The input units are informed based on the internal weighting system, and the neural network tries to learn more and ultimately produce the desired results. Furthermore, ANN can accomplish multitasking without affecting system performance.^{38 39} We used ‘nnet’ package to fit the ANN model in R. Boosting is an ensemble meta-algorithm integrating weak learners to form a firm classification rule by executing several iterations that enhance the accuracy of the prediction.^{40 41} These algorithms seek to improve prediction power by forming a number of weak models that are converted into strong ones. Bagging or bootstrap aggregation method also uses ensemble learning to evolve ML models.⁴² This algorithm is based on the hypothesis that combining multiple models together can often produce a much more powerful model.⁴³ Lreg is one of the simplest and most common ML algorithms that has been used often in low dimension data for binary classification problems. This algorithm uses the sigmoid function to perform prediction task.^{44 45} We used the generalised linear model function found in ‘glm’ package in ‘R-software’ to execute Lreg algorithm. After training the selected models, we computed each model’s prediction performance using a testing dataset (which is the 30% proportion of the primary data set reserved and

unseen by the model) for the validation task. The primary purpose of using the testing data set (also known as holdout method) is to test the generalisation ability of the trained models as well as to avoid model overfitting. To evaluate the models' validity and performance, we used 'area under the receiver operating characteristic curve' (AUROC). The AUROC is a performance measurement used in ML for classification problems that uses the true positive and false positive rate that represents the degree or measure of separability and describes how much the model can distinguish between classes.^{46 47} In these ML models, we uniformly applied hold-out method, 10-fold cross-validation and Synthetic Minority Oversampling Technique techniques to minimise potential overfitting as well as handling class imbalances. Cross-validation in ML involves randomly dividing the set observations into groups (folds) of equal sizes. We estimated the sensitivity, specificity, AUROC, positive predictive values and negative predictive values based on 30% samples left out of the classifier training procedure, during a 10-fold cross validation process. Since the AUROC method assumes uniform distribution of threshold probabilities, a scenario which may not be always true, we ran the decision curve analysis (DCA). The approach estimates the distribution of threshold probabilities without the need of additional data. Using the estimated distribution of threshold probabilities, the weighted area under the net benefit curve serves as the summary measure to compare risk prediction models in a range of interest.⁴⁸ In brief, DCA calculates a clinical 'net benefit' for one or more prediction models or diagnostic tests in comparison to default strategies of treating all or no patients. In this framework, a clinical judgement of the relative value of benefits (treating a true positive case) and harms (treating a false positive case) associated with prediction models is made. As such, the preferences of patients or policymakers are accounted for by using threshold probability metric.⁴⁸ The net benefit is calculated for each possible threshold probability, which puts benefits and harms on the same scale.

Patient and public involvement

No patient involved.

RESULTS

Characteristics of study participants

The current study analysed 8814 induced and 12 764 spontaneous deliveries. The mean maternal age of study participants was 28 (SD=6) years. About half of deliveries were from mothers aged between 25 and 35 years. Our study population had a good balance between nulliparous (46%) and multiparous (54%) mothers. The majority of deliveries (78%) were at term while post-term deliveries constituted of about 8% and preterm accounted for 14% of all deliveries. Sociodemographic and clinical characteristics of study participants are clearly displayed in [table 1](#).

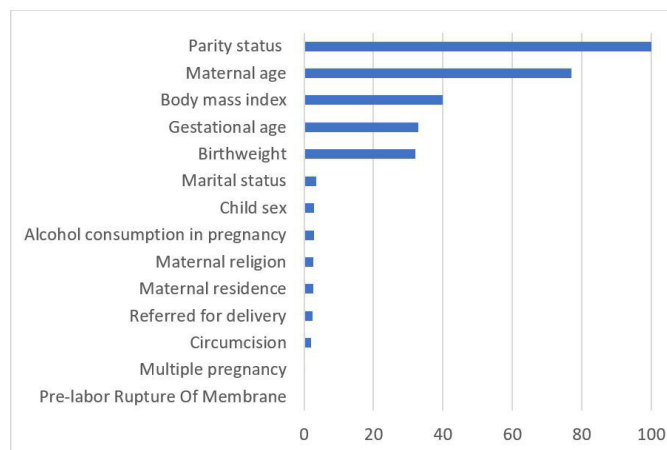


Figure 2 Variable importance measures for prediction of labour induction intervention.

Variable importance measures for IOL intervention

Body mass index, maternal age, gestational age, parity and birth weight to be the essential features to consider when predicting IOL intervention [figure 2](#).

Predictive models for IOL intervention

The overall performance of the selected ML models can be visualised in [figure 3](#) and [table 2](#). All models showed a somewhat similar performance in terms of AUROC. However, Delong's test for similarity of AUROC indicated that the Lreg model was outperformed by all other models except for Bagging algorithm. Simply put, the performance of bagged trees was not significantly different from that of Lreg at 5% significance level. In addition, we performed subgroup analyses for predictive performance by parity and maternal age presented in [tables 3 and 4](#). The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement for the transparent reporting of multivariable prediction models has been included (online supplemental file 1) as

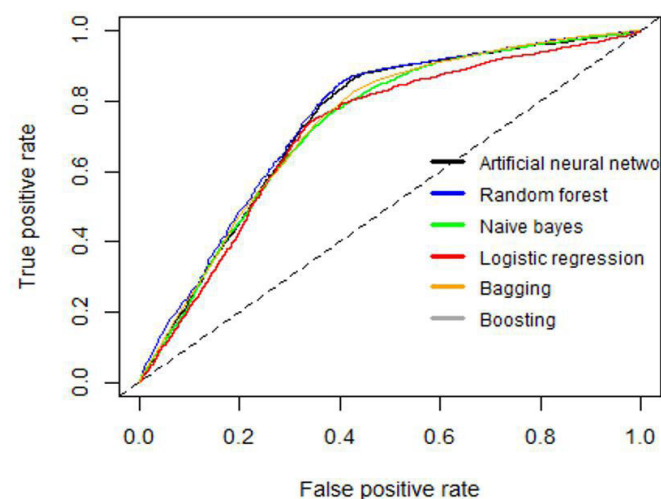


Figure 3 ROC curve for comparing the performance of ML algorithms. ML, machine learning; ROC, receiver operating characteristic.

Table 2 Overall prediction performance of the machine learning models

Model	Logistic regression	Artificial neural network	Random forest	Naïve Bayes	Bagging	Boosting
ACC	0.69 (0.68–0.70)	0.74 (0.73–0.75)	0.73 (0.72–0.74)	0.72 (0.71–0.73)	0.71 (0.70–0.72)	0.74 (0.73–0.75)
AUROC	0.71 (0.70–0.73)	0.73 (0.72–0.75)	0.74 (0.72–0.75)	0.73 (0.72–0.75)	0.72 (0.71–0.73)	0.75 (0.73–0.76)
P value*	Reference	<0.001	<0.001	<0.001	0.3326	<0.001
Sensitivity	0.70 (0.68–0.71)	0.86 (0.85–0.87)	0.84 (0.83–0.85)	0.82 (0.80–0.83)	0.80 (0.79–0.82)	0.85 (0.83–0.86)
Specificity	0.68 (0.66–0.70)	0.58 (0.56–0.60)	0.57 (0.56–0.59)	0.57 (0.56–0.59)	0.57 (0.55–0.59)	0.59 (0.57–0.61)
PPV	0.76 (0.74–0.77)	0.75 (0.73–0.76)	0.74 (0.73–0.75)	0.74 (0.72–0.75)	0.73 (0.72–0.74)	0.75 (0.74–0.76)
NPV	0.61 (0.59–0.62)	0.74 (0.72–0.75)	0.71 (0.69–0.73)	0.68 (0.66–0.70)	0.67 (0.65–0.69)	0.73 (0.71–0.75)

N=21 578.

ACC, Accuracy; AUROC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

well as the R-syntax for modelling and validation (online supplemental file 2).

Results from the DCA (figure 4) have shown that RF and boosting methods demonstrated the best net benefit over the range of threshold probabilities compared with other models.

DISCUSSION

Parity, maternal age, BMI, gestational age and birth weight were identified as important predictive features for IOL intervention. Previous studies have shown the existing link between these features and likelihood of pregnancy

interventions including IOL. It is shown that pregnancies beyond 41 weeks of gestation are prone to IOL due to increased risk of uterine rupture after 40 weeks of gestation.^{49–51} The birth rate for women aged 35 years and older has increased more than 30% since 1990, and they have been shown to be at an elevated risk of adverse pregnancy outcome.⁵² Literature shows that obese women require more prolonged IOL involving more extensive and frequent applications of both cervical ripening methods and synthetic oxytocin.⁵³ The concept of a risk threshold for the relationship between parity and pregnancy outcomes has been of concern for decades. In some studies, associations have been

Table 3 Prediction performance of the ML algorithm by maternal age

Model	Logistic regression	Artificial neural network	Random forest	Naïve Bayes	Bagging	Boosting	
Maternal age <25 (n=8032)	ACC	0.74 (0.72–0.75)	0.74 (0.72–0.76)	0.74 (0.72–0.76)	0.74 (0.72–0.76)	0.71 (0.69–0.73)	0.76 (0.74–0.77)
	AUROC	0.75 (0.73–0.77)	0.74 (0.71–0.75)	0.76 (0.72–0.76)	0.76 (0.74–0.78)	0.74 (0.72–0.76)	0.77 (0.75–0.78)
	Sensitivity	0.78 (0.76–0.80)	0.85 (0.83–0.87)	0.82 (0.80–0.84)	0.79 (0.77–0.81)	0.76 (0.74–0.78)	0.85 (0.83–0.87)
	Specificity	0.68 (0.66–0.71)	0.62 (0.59–0.64)	0.65 (0.63–0.68)	0.68 (0.65–0.70)	0.65 (0.63–0.68)	0.64 (0.62–0.67)
	PPV	0.74 (0.72–0.76)	0.72 (0.69–0.74)	0.73 (0.71–0.75)	0.74 (0.72–0.76)	0.72 (0.69–0.74)	0.73 (0.71–0.76)
	NPV	0.73 (0.70–0.76)	0.78 (0.75–0.80)	0.76 (0.73–0.78)	0.74 (0.71–0.77)	0.70 (0.67–0.73)	0.79 (0.76–0.82)
Maternal age 25–35 (n=10284)	ACC	0.74 (0.73–0.76)	0.74 (0.73–0.76)	0.73 (0.73–0.76)	0.75 (0.73–0.76)	0.71 (0.70–0.73)	0.75 (0.73–0.76)
	AUROC	0.74 (0.72–0.76)	0.74 (0.72–0.76)	0.74 (0.72–0.76)	0.74 (0.72–0.75)	0.73 (0.71–0.74)	0.73 (0.73–0.76)
	Sensitivity	0.83 (0.82–0.85)	0.84 (0.82–0.85)	0.84 (0.82–0.86)	0.84 (0.82–0.85)	0.81 (0.79–0.83)	0.84 (0.82–0.85)
	Specificity	0.62 (0.59–0.64)	0.61 (0.58–0.64)	0.57 (0.54–0.60)	0.61 (0.58–0.64)	0.57 (0.54–0.60)	0.61 (0.58–0.64)
	PPV	0.76 (0.74–0.78)	0.76 (0.74–0.78)	0.74 (0.72–0.76)	0.76 (0.74–0.78)	0.73 (0.72–0.75)	0.76 (0.74–0.78)
	NPV	0.72 (0.69–0.74)	0.72 (0.69–0.74)	0.71 (0.68–0.74)	0.72 (0.69–0.74)	0.67 (0.64–0.70)	0.72 (0.69–0.74)
Maternal age >35 (n=3262)	ACC	0.73 (0.71–0.74)	0.75 (0.74–0.76)	0.74 (0.73–0.76)	0.74 (0.72–0.75)	0.72 (0.71–0.74)	0.75 (0.74–0.77)
	AUROC	0.75 (0.74–0.78)	0.76 (0.75–0.78)	0.77 (0.75–0.79)	0.77 (0.76–0.79)	0.76 (0.74–0.77)	0.77 (0.76–0.79)
	Sensitivity	0.69 (0.67–0.71)	0.80 (0.78–0.82)	0.77 (0.75–0.79)	0.72 (0.70–0.74)	0.74 (0.72–0.76)	0.78 (0.76–0.80)
	Specificity	0.76 (0.74–0.78)	0.70 (0.68–0.72)	0.72 (0.69–0.74)	0.76 (0.74–0.78)	0.70 (0.68–0.72)	0.73 (0.71–0.75)
	PPV	0.75 (0.73–0.77)	0.73 (0.71–0.75)	0.73 (0.71–0.75)	0.75 (0.73–0.77)	0.71 (0.69–0.73)	0.74 (0.72–0.76)
	NPV	0.71 (0.69–0.73)	0.78 (0.76–0.80)	0.75 (0.73–0.77)	0.73 (0.71–0.75)	0.73 (0.71–0.75)	0.77 (0.75–0.79)

ACC, Accuracy; AUROC, area under the receiver operating characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

Table 4 Prediction performance of the ML algorithm by parity status

	Model	Logistic regression	Artificial neural network	Random forest	Naïve Bayes	Bagging	Boosting
Nulliparous women (n=10 000)	ACC	0.66 (0.64–0.69)	0.65 (0.64–0.67)	0.73 (0.72–0.75)	0.72 (0.70–0.73)	0.71 (0.70–0.73)	0.74 (0.72–0.75)
	AUROC	0.64 (0.61–0.66)	0.67 (0.65–0.69)	0.74 (0.72–0.76)	0.72 (0.70–0.74)	0.73 (0.71–0.75)	0.75 (0.73–0.77)
	Sensitivity	0.31 (0.28–0.33)	0.44 (0.42–0.47)	0.55 (0.52–0.58)	0.53 (0.50–0.56)	0.56 (0.53–0.59)	0.54 (0.51–0.57)
	Specificity	0.89 (0.88–0.91)	0.80 (0.78–0.81)	0.85 (0.83–0.87)	0.84 (0.82–0.86)	0.82 (0.80–0.83)	0.86 (0.85–0.88)
	PPV	0.66 (0.62–0.70)	0.59 (0.56–0.62)	0.71 (0.68–0.74)	0.69 (0.65–0.72)	0.67 (0.64–0.70)	0.72 (0.69–0.75)
	NPV	0.66 (0.64–0.68)	0.69 (0.67–0.71)	0.74 (0.72–0.76)	0.73 (0.71–0.75)	0.74 (0.72–0.76)	0.74 (0.72–0.76)
Multiparous women (n=11 578)	ACC	0.84 (0.83–0.85)	0.84 (0.82–0.85)	0.83 (0.82–0.85)	0.80 (0.79–0.81)	0.82 (0.81–0.83)	0.84 (0.83–0.85)
	AUROC	0.85 (0.84–0.86)	0.84 (0.82–0.86)	0.84 (0.83–0.86)	0.84 (0.82–0.85)	0.84 (0.82–0.85)	0.85 (0.84–0.86)
	Sensitivity	0.99 (0.98–1.00)	0.98 (0.97–0.99)	0.96 (0.95–0.97)	0.89 (0.87–0.90)	0.93 (0.92–0.94)	0.99 (0.98–1.00)
	Specificity	0.67 (0.64–0.69)	0.67 (0.64–0.69)	0.69 (0.67–0.71)	0.71 (0.68–0.73)	0.70 (0.67–0.72)	0.67 (0.65–0.69)
	PPV	0.76 (0.75–0.78)	0.77 (0.75–0.78)	0.77 (0.76–0.79)	0.77 (0.75–0.78)	0.77 (0.75–0.79)	0.77 (0.75–0.78)
	NPV	0.98 (0.97–1.00)	1.00 (0.99–1.00)	0.95 (0.93–0.96)	0.85 (0.83–0.87)	0.91 (0.89–0.92)	0.99 (0.99–1.00)

ACC, Accuracy; AUROC, area under the receiver operating characteristic curve; ML, machine learning; NPV, negative predictive value; PPV, positive predictive value.

found between parity and adverse pregnancy outcomes, while others concluded that multiparity was not a risk for any pregnancy intervention.^{54–59} The current study used ML methods to predict the likelihood of IOL intervention at the tertiary hospital using the maternal birth registry database. With the AUROC value of 0.71 (95% CI 0.70 to 0.73), Lreg model was outperformed by all other models, while Boosting algorithm showing the best performance (AUROC=0.75 (95% CI 0.73 to 0.76)). Boosting method is likely to achieve the best overall performance as it takes care of the weightage of the higher accuracy sample and lower accuracy sample and eventually gives the combined results. In addition, boosting evaluates the net error in each step prior to model building and can work better with interactions compared with other ML models. When we disaggregated our analyses by parity and maternal age, the ML models' performance in terms of

AUROC was still almost similar but significantly improved despite the reduced sample sizes. Studies indicated that the subgrouped data may be problematic for pattern recognition, and only a limited number of papers have systematically investigated how the ML validation process should be designed to help avoid potential optimistic performance estimates due to reduced sample size.⁶⁰ However, it is not yet clear how sample size affects the accuracy of the learning algorithms.⁶¹ In addition, as we subgrouped the dataset, the proportion of induced delivery relatively changed, a scenario that may have created the so-called class-imbalance problem (the change in proportions between the induced and spontaneous deliveries). The literature indicates that the extent of class-imbalance has important role on the learning process of ML algorithms.⁶⁰ We used DCA to portray the impact of false-negative and false-positive misclassification errors. From the DCA, we estimated the 'net benefit' metric which is calculated across a range of threshold probabilities. We observed that the net benefit for RF and Boosting models surpassed all other models under investigation. The net benefit metric provides information about the consequences of using the model in question unlike AUROC that solely reports the model's accuracy. Taking the case where falsely predicting a case as 'not induced' (false-negative) is much more harmful than a false-positive result, a model that has a much greater specificity but slightly lower sensitivity than another would have a higher AUROC but would be a poorer choice for clinical use.^{62–64} Simply put, applying Lreg method for predicting the utilisation of IOL intervention using this registry database may be more clinically consequential than using any ML algorithm tested in the current study.^{65–68} To our knowledge, the current study is the first study that applied the most popular ML algorithms to predict the use of IOL intervention in Tanzania. We enrolled a number of deliveries from an extended period (15 years), a sample that may have accommodated a diversified group of study

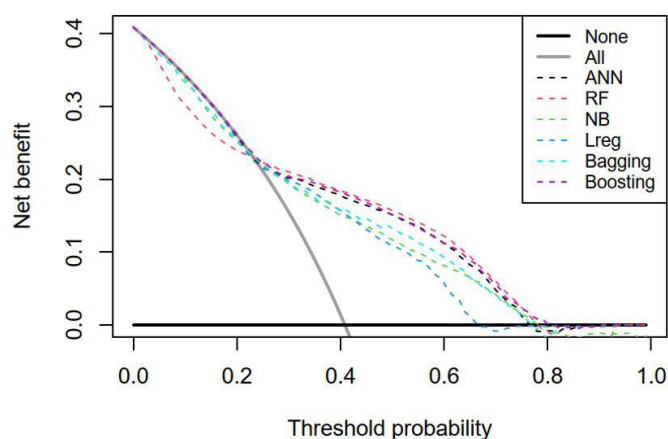


Figure 4 The decision curve analysis showing the net-benefit of machine learning models for predicting likelihood of labour induction over the range of threshold probabilities. ANN, artificial neural networks; Lreg, logistic regression; NB, naïve Bayes; RF, random forest.

participants with contrasting characteristics. We also derived the preliminary predictors of IOL intervention from diverse maternal features that are routinely recorded in hospital setting. Forasmuch as the validity of predictive models and their application in the general population highly depends on their goodness of fit, our modelling procedures based on bootstrap re-sampling and repeated random split (training and validation sets) techniques to ensure model generalisability.

However, our study had some limitations that should be taken into consideration during interpretations of the results. Observations with missing values in both the outcome and predictors were excluded from the analyses. We argue that this might not necessarily be an optimal approach for dealing with missing data because important information could be lost when incomplete rows of data are discarded. However, learning algorithms are significantly affected missing values more than other statistical models, including Lreg, as they rely heavily on data to learn the underlying input/output relationships of the attributes being modelled. Studies have explored the extent of damage to the performance of learning algorithms due to missing data in a field-scale application. In this regard, we call on a prospectively designed study that will consider improvement in data entry prior to assessment of predictive models. In addition, our study did not perform feature engineering or variable selection prior to model building, a scenario which may have an impact to classifier performance as well as the possibility of model overfitting. Furthermore, we think that it would be interesting if future studies will identify and consider medically important subgroups as far as IOL intervention is concerned and conduct the comparative performance of the ML algorithms. Lastly, the study involved only the deliveries attended at the KCMC, hence a potential for selection bias, indicating that the output may not be applicable to other hospital setting.

CONCLUSION

Parity, maternal age, gestational age and body mass index have been shown to be stable and relatively important variables in the preliminary prediction of IOL intervention. Boosting algorithms shows the promising performance in predicting IOL intervention. However, extensive studies may be required to assess the performance of additional ML methods, particularly the ones applies ensemble learning methods such as Adaptive boosting, extreme boosting and gradient boosting.

Author affiliations

¹College of Public Health, Zhengzhou University, Zhengzhou, China

²Science and Laboratory Technology, Dar es Salaam Institute of Technology, Dar es Salaam, Tanzania, United Republic of

³School of Planning and Public Policy, Rutgers University-New Brunswick, New York, New York, USA

⁴Institute of Public Health, Kilimanjaro Christian Medical University College, Moshi, Tanzania, United Republic of

Twitter Clifford Silver Tarimo @clifford_silver

Acknowledgements We would like to thank the staff of the Birth Registry, Department of Obstetrics & Gynaecology of the Kilimanjaro Christian Medical Centre and the Department of Epidemiology and Applied Biostatistics of the Kilimanjaro Christian Medical University College for their substantial support during this study. Special thanks to women who participated in the KCMC birth registry study and the Norwegian birth registry for partnering with us in providing the limited dataset used for this study. Finally, we would like to acknowledge Mr. Innocent B. Mboya (Lecturer at Kilimanjaro Christian Medical University College (KCMUCo)-Tanzania) for the motivation he provided on machine learning techniques particularly on R-syntax.

Contributors All authors made a substantial contribution to this study. CST conceived the idea, designed the study, analysed and drafted the manuscript. SSB cosupervised the project and provided technical assistance on global health perspective. QL reviewed the final manuscript. MJJM data management and retrieval, read and approved the final manuscript. JW supervised the project and reviewed the final manuscript. All authors read and approved the final version of the manuscript. CST acts as a guarantor for this publication.

Funding This study was funded by the 'Research on CDC-Hospital-Community Trinity Coordinated Prevention and Control System for Major Infectious Diseases, Zhengzhou University 2020 Key Project of Discipline Construction' with grant number XKZDQY202007, 2021 Postgraduate Education Reform and Quality Improvement Project of Henan Province with grant number (YJS2021KC07) and National Key R&D Programme of China with grant number (2018YFC0114501).

Competing interests None declared.

Patient consent for publication Consent obtained directly from patient(s)

Ethics approval This study sought and was granted an ethical approval from Kilimanjaro Christian Medical University College Research and Ethics and Review Committee (KCMU-CRERC) with approval number 985. The registry project obtained informed verbal consents from the study subjects during development of the medical registry database. The midwife nurse gave every woman oral information about the birth registry, the data needed to be collected from them and the use of the data for research purposes. Following the consent, the woman could still opt not to reply to individual questions. All consent procedures were approved by the Kilimanjaro Christian Medical Centre ethical committee and the administrative permission to access the data was provided by the KCMC hospital. Furthermore, confidentiality and privacy were assured as per the protocol of the birth registry. Patients' names were coded by the unique hospital registration numbers to ensure anonymity. We declare that all methods adopted in this research were carried out in accordance with the guideline and regulations for involving human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The data used/or analyzed during the current study is available from the first author on a reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Clifford Silver Tarimo <http://orcid.org/0000-0002-0672-9471>

Michael Johnson J Mahande <http://orcid.org/0000-0002-7750-7657>

REFERENCES

- 1 American College of Obstetricians and Gynecologists (ACOG practice bulletin). ACOG practice Bulletin No. 107: induction of labor. *Obstet Gynecol* 2009;114:386–97.

- 2 Mackenzie IZ. Induction of labour at the start of the new millennium. *Reproduction* 2006;131:989–98.
- 3 Sanchez-Ramos L. Induction of labor. *Obstet Gynecol Clin North Am* 2005;32:181–200.
- 4 Mozurkewich EL, Chilimigras JL, Berman DR, et al. Methods of induction of labour: a systematic review. *BMC Pregnancy Childbirth* 2011;11:84.
- 5 Heinemann J, Gillen G, Sanchez-Ramos L, et al. Do mechanical methods of cervical ripening increase infectious morbidity? A systematic review. *Am J Obstet Gynecol* 2008;199:177–88.
- 6 Vaknin Z, Kurzweil Y, Sherman D. Foley catheter balloon vs locally applied prostaglandins for cervical ripening and labor induction: a systematic review and metaanalysis. *Am J Obstet Gynecol* 2010;203:418–29.
- 7 Lueth GD, Kebede A, Medhanyie AA. Prevalence, outcomes and associated factors of labor induction among women delivered at public hospitals of MEKELLE town-(a hospital based cross sectional study). *BMC Pregnancy Childbirth* 2020;20:203.
- 8 Robson S, Pridmore B, Dodd J. Outcomes of induced labour. *Aust N Z J Obstet Gynaecol* 1997;37:16–19.
- 9 Lydon-Rochelle MT, Cárdenas V, Nelson JC, et al. Induction of labor in the absence of standard medical indications: incidence and correlates. *Med Care* 2007;45:505–12. 9.
- 10 Coonrod DV, Bay RC, Kishi GY. The epidemiology of labor induction: Arizona, 1997. *Am J Obstet Gynecol* 2000;182:1355–62.
- 11 Zhang J, Yancey MK, Henderson GE. U.S. national trends in labor induction, 1989–1998. *J Reprod Med* 2002;47:498–9.
- 12 Laughon SK, Zhang J, Grewal J, et al. Induction of labor in a contemporary obstetric cohort. *Am J Obstet Gynecol* 2012;206:486.e1–486.e9.
- 13 Bukola F, Idi N, M'Mimunya M, et al. Unmet need for induction of labor in Africa: secondary analysis from the 2004 - 2005 WHO Global Maternal and Perinatal Health Survey (A cross-sectional survey). *BMC Public Health* 2012;12:722.
- 14 World Health Organization. Neonatal and perinatal mortality: country, regional and global estimates. Available: <https://apps.who.int/iris/handle/10665/43800> [Accessed 30 Dec 2020].
- 15 Thomas J, Kavanagh J, Anthony K. *RCOG evidence-based clinical guidelines induction of labour*, 2001.
- 16 Yeast JD, Jones A, Poskin M. Induction of labor and the relationship to cesarean delivery: a review of 7001 consecutive inductions. *Am J Obstet Gynecol* 1999;180:628–33.
- 17 Sanchez-Ramos L. Induction of labor. *Obstet Gynecol Clin North Am* 2005;32:181–200.
- 18 The World Health Organization. Recommendations for induction of labor. Available: https://www.who.int/reproductivehealth/publications/maternal_perinatal_health/9789241501156/en/ [Accessed 30 Apr 2020].
- 19 Lassi ZS, Mansoor T, Salam RA, et al. Essential pre-pregnancy and pregnancy interventions for improved maternal, newborn and child health. *Reprod Health* 2014;11 Suppl 1:S2.
- 20 Iftikhar P, Kuijpers MV, Khayyat A, et al. Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. *Cureus* 2020;12:e7124.
- 21 McCoy LG, Banja JD, Ghassemi M, et al. Ensuring machine learning for healthcare works for all. *BMJ Health Care Inform* 2020;27:e100237.
- 22 Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019;6:94–8.
- 23 Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.
- 24 Lynam AL, Dennis JM, Owen KR, et al. Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. *Diagn Progn Res* 2020;4:6.
- 25 Shillan D, Sterne JAC, Champneys A, et al. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit Care* 2019;23:284.
- 26 Mahande MJ, Daltveit AK, Mmbaga BT, et al. Recurrence of perinatal death in northern Tanzania: a Registry based cohort study. *BMC Pregnancy Childbirth* 2013;13:166.
- 27 Strobl C, Boulesteix A-L, Kneib T, et al. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:307.
- 28 Strobl C, Zeileis A. Danger: high power! – exploring the statistical properties of a test for random forest variable importance. *Proceedings of the 18th International Conference on Computational Statistics*, Porto, Portugal, 2008.
- 29 Jiang R, Tang W, Wu X, et al. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* 2009;10 Suppl 1:S65.
- 30 Strobl C, Boulesteix A-L, Zeileis A, et al. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 2007;8:25.
- 31 Song L, Langfelder P, Horvath S. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics* 2013;14:5.
- 32 Langarizadeh M, Moghbeli F. Applying naive Bayesian networks to disease prediction: a systematic review. *Acta Inform Med* 2016;24:364–9.
- 33 Sharma RK, Sugumaran V, Kumar H, et al. A comparative study of naive Bayes classifier and Bayes net classifier for fault diagnosis of roller bearing using sound signal. *International Journal of Decision Support Systems* 2015;1:115–29.
- 34 Wolfson J, Bandyopadhyay S, Elidrisi M, et al. A naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Stat Med* 2015;34:2941–57.
- 35 Kazmierska J, Malicki J. Application of the naïve Bayesian classifier to optimize treatment decisions. *Radiother Oncol* 2008;86:211–6.
- 36 Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. *PLoS One* 2019;14:e0212356.
- 37 Lisboa PJ, Taktak AFG. The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Netw* 2006;19:408–15.
- 38 Kawauchi K, Furuya S, Hirata K, et al. A convolutional neural network-based system to classify patients using FDG PET/CT examinations. *BMC Cancer* 2020;20:227.
- 39 Almeida PP, Cardoso CP, de Freitas LM. PDAC-ANN: an artificial neural network to predict pancreatic ductal adenocarcinoma based on gene expression. *BMC Cancer* 2020;20:82.
- 40 Komori O, Eguchi S. A boosting method for maximizing the partial area under the ROC curve. *BMC Bioinformatics* 2010;11:314.
- 41 Freund Y, Schapire RE. A Decision-Theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997;55:119–39.
- 42 Zhang H, Song Y, Jiang B, et al. Two-Stage Bagging pruning for reducing the ensemble size and improving the classification performance. *Math Probl Eng* 2019;2019:1–17.
- 43 Datta S, Pihur V, Datta S. An adaptive optimal ensemble classifier via bagging and RANK aggregation with applications to high dimensional data. *BMC Bioinformatics* 2010;11:427.
- 44 Choi SH, Labadorf AT, Myers RH, et al. Evaluation of logistic regression models and effect of covariates for case-control study in RNA-seq analysis. *BMC Bioinformatics* 2017;18:91.
- 45 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002;35:352–9.
- 46 Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008;29 Suppl 1:S83–7.
- 47 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;39:561–77.
- 48 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.
- 49 Heaman M, Kingston D, Chalmers B, et al. Risk factors for preterm birth and small-for-gestational-age births among Canadian women. *Paediatr Perinat Epidemiol* 2013;27:54–61.
- 50 Vendittelli F, Rivière O, Neveu B, et al. Does induction of labor for constitutionally large-for-gestational-age fetuses identified in utero reduce maternal morbidity? *BMC Pregnancy Childbirth* 2014;14:156.
- 51 Elden H, Hagberg H, Wessberg A, et al. Study protocol of SWEPIS a Swedish multicentre register based randomised controlled trial to compare induction of labour at 41 completed gestational weeks versus expectant management and induction at 42 completed gestational weeks. *BMC Pregnancy Childbirth* 2016;16:49.
- 52 Hamm RF, Srinivas SK, Levine LD. Risk factors and racial disparities related to low maternal birth satisfaction with labor induction: a prospective, cohort study. *BMC Pregnancy Childbirth* 2019;19:530.
- 53 Feresu SA, Wang Y, Dickinson S. Relationship between maternal obesity and prenatal, metabolic syndrome, obstetrical and perinatal complications of pregnancy in Indiana, 2008–2010. *BMC Pregnancy Childbirth* 2015;15:266.
- 54 Hsieh T'sang-T'ang, Liou J-D, Hsu J-J, et al. Advanced maternal age and adverse perinatal outcomes in an Asian population. *Eur J Obstet Gynecol Reprod Biol* 2010;148:21–6.

- 55 Kahveci B, Melekoglu R, Evruke IC, *et al.* The effect of advanced maternal age on perinatal outcomes in nulliparous singleton pregnancies. *BMC Pregnancy Childbirth* 2018;18:343.
- 56 Yogev Y, Melamed N, Bardin R, *et al.* Pregnancy outcome at extremely advanced maternal age. *Am J Obstet Gynecol* 2010;203:558.e1–558.e7.
- 57 Dammer U, Bogner R, Weiss C, *et al.* Influence of body mass index on induction of labor: a historical cohort study. *J Obstet Gynaecol Res* 2018;44:697–707.
- 58 Lewkowitz A, Koser S, Koser S. Relationship Between Maternal BMI and Labor Induction Outcomes [14T]. *Obstetrics Gynecology* 2019;133:216S.
- 59 Athukorala C, Rumbold AR, Willson KJ, *et al.* The risk of adverse pregnancy outcomes in women who are overweight or obese. *BMC Pregnancy Childbirth* 2010;10:56.
- 60 Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013;14:106.
- 61 Figueroa RL, Zeng-Treitler Q, Kandula S, *et al.* Predicting sample size required for classification performance. *BMC Med Inform Decis Mak* 2012;12:8.
- 62 Raita Y, Goto T, Faridi MK, *et al.* Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019;23:64.
- 63 Goto T, Camargo CA, Faridi MK, *et al.* Machine Learning-Based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open* 2019;2:e186937.
- 64 Kuhle S, Maguire B, Zhang H, *et al.* Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. *BMC Pregnancy Childbirth* 2018;18:333.
- 65 Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Med Res Methodol* 2020;20:171.
- 66 Kupek E. Beyond logistic regression: structural equations modelling for binary variables and its application to investigating unobserved confounders. *BMC Med Res Methodol* 2006;6:13.
- 67 Frank E, Hall M, Trigg L, *et al.* Data mining in bioinformatics using Weka. *Bioinformatics* 2004;20:2479–81.
- 68 Gómez D, Rojas A. An empirical overview of the NO free lunch theorem and its effect on real-world machine learning classification. *Neural Comput* 2016;28:216–28.