

RESEARCH

Open Access



Analysis of risk factors progression of preterm delivery using electronic health records

Zeineb Safi^{1†}, Neethu Venugopal¹, Haytham Ali², Michel Makhlouf³, Faisal Farooq⁴ and Sabri Boughorbel^{4*} 

[†]Zeineb Safi and Sabri Boughorbel contributed equally to this work.

*Correspondence:

sboughorbel@hbku.edu.qa

²Qatar Computing Research Institute, HBKU, Doha, Qatar
Full list of author information is available at the end of the article

Abstract

Background: Preterm deliveries have many negative health implications on both mother and child. Identifying the population level factors that increase the risk of preterm deliveries is an important step in the direction of mitigating the impact and reducing the frequency of occurrence of preterm deliveries. The purpose of this work is to identify preterm delivery risk factors and their progression throughout the pregnancy from a large collection of Electronic Health Records (EHR).

Results: The study cohort includes about 60,000 deliveries in the USA with the complete medical history from EHR for diagnoses, medications and procedures. We propose a temporal analysis of risk factors by estimating and comparing risk ratios and variable importance at different time points prior to the delivery event. We selected the following time points before delivery: 0, 12 and 24 week(s) of gestation. We did so by conducting a retrospective cohort study of patient history for a selected set of mothers who delivered preterm and a control group of mothers that delivered full-term. We analyzed the extracted data using logistic regression and random forests models. The results of our analyses showed that the highest risk ratio and variable importance corresponds to history of previous preterm delivery. Other risk factors were identified, some of which are consistent with those that are reported in the literature, others need further investigation.

Conclusions: The comparative analysis of the risk factors at different time points showed that risk factors in the early pregnancy related to patient history and chronic condition, while the risk factors in late pregnancy are specific to the current pregnancy. Our analysis unifies several previously reported studies on preterm risk factors. It also gives important insights on the changes of risk factors in the course of pregnancy. The code used for data analysis will be made available on github.

Keywords: Preterm, Pregnancy, EHR, Epidemiology, Risk factors, Progression, Temporal analysis, Precision medicine, Predictive models

Introduction

The WHO defines prematurity as births before 37 weeks of completed gestation [1]. Over the recent years there is a global increase in the rate of prematurity, ranging from 9-12% of all birth. In 2016, prematurity affected around 15 million live-born babies worldwide. [2] The reporting of the short and long-term outcomes of prematurity attracted the interest of scientists, clinicians, as well as policymakers. Of particular concern is the high rate of prematurity within the low socioeconomic class [3]. Extreme prematurity, defined as birth before 28 weeks of gestation, poses life-long consequences on the health, education, and the social life of children [4]. From the societal perspective, almost one-third of extremely premature infants required support in special education systems. The majority had poor educational attainment at school age [5]. It was found that 10% of the deaths were due to preterm birth among the six major causes that were attributed to 73% of yearly deaths of children under the age of 5 years [6]. Children that are born preterm are at a higher risk of a plethora of psychological and physiological health implications, especially those with a low birth weight. A quarter of the very low birth weight children develop severe or multiple psychological problems, and another quarter develop moderate to mild problems. The psychological development is measured in different domains, namely cognitive development, behavioral and emotional status, social functioning and school adaptation [7]. Children that are very preterm have abnormal brain morphology when compared to those born full-term at seven years of age [8]. A significant proportion of the children that are born very premature, i.e., before 27 week gestation, are found to have difficulties in motor and academic skills in early school years [9]. They are also more likely to develop respiratory problems [10]. Preterm deliveries have many negative implications on the mother's and child's health alike. Mothers that deliver preterm babies, are at a greater risk to suffer from complications, particularly those who undergo a Cesarean delivery. The complications include hemorrhage, infection, ICU admission and death [11].

Identifying mothers that are at a higher risk through quantifying risk factors of preterm delivery at a population level helps clinicians to take preventive measures and mitigate the risks [12, 13]. Traditionally, the identification of such risk factors is done through prospective studies. This method presents challenges, some of which are difficulty in recruiting trial participants and the tediousness of the data collection process. With the increased adoption of Electronic Health Records (EHR) in hospitals and health care facilities around the world, some of these challenges have been alleviated. EHR data is being extensively used in clinical research, even though it presents some challenges of its own, it provides fast and easy access to large amounts of data, that is more representative of the general population than data collected from clinical trials. Some of the research application of EHR data are epidemiology and observational research, safety surveillance and regulatory use, and prospective clinical research [14].

Most studies are also concerned with extracting risk factors from a limited set of suspected attributes, while in our analyses we extract risk factors from all possible attributes that are available to us in Electronic Health Records. We included diagnosis information (ICD-9 and ICD-10), medication information (NDC codes), laboratory orders (ICD-9 and ICD-10).

The purpose of this work is to perform a retrospective cohort study using a large EHR dataset to identify risk factors of preterm delivery. We hypothesize that risk factors are dynamics and can change in the course of pregnancy. In the early phase of pregnancy,

genetic predisposition might dominate the risk factors while in advanced stages of the pregnancy, other factors could play more important roles [12]. The analysis of the variation of risk factors as a function of time could help in building models for the early prediction of preterm delivery.

In this work, we identify the risk factors that are associated with preterm delivery in the literature, and present a few examples of the use of EHR data to perform epidemiology studies in “[Related work](#)” section. We present the methodology adopted for extracting relevant preterm and full-term deliveries from the EHR database and statistical analysis of the data in “[Methodology](#)” section. We present our results and discuss our findings in “[Results and discussion](#)” section. Risk factors obtained from association studies should not be understood as causal factors. Among the risk factors that are statistically significant, we curated the ones related to spontaneous preterm deliveries from the medically induced ones. “[Conclusion](#)” section concludes our findings and presents prospects for future work.

Related work

Preterm delivery risk factors

The identification of risk factors of preterm deliveries is a well researched problem. The history of previous preterm deliveries is one of the main risk factors that have been identified by different research studies [15–18]. Maternal age is also a risk factor, adolescent pregnancies and advanced maternal age are associated with increased risk of preterm deliveries [15, 16]. Other factors are low maternal BMI [15, 18], obesity [16], women of African race [16–18], and short inter-pregnancy intervals [16, 17]. The mother anatomy, such as short cervical length [16–18], and uterine anomalies [17] are also well known risk factors. Existing conditions in the mother whether they are infectious conditions (HIV, chlamydia, and urinary tract infection) or other conditions (pre-eclampsia, low maternal vitamin D, pregestational and gestational diabetes [16], and anemia [19]) also increase the risk of preterm delivery. Another risk factor is multiple gestation [18] or singleton and twin pregnancies formed through in vitro fertilization [16], in addition to some of the mother’s behaviors such as smoking and drug use during pregnancy [16]. Some work also investigated the role of environmental factors on preterm birth [20], where an increase in NO_2 concentration was shown to have an association with preterm birth.

Risk factors from electronic health records

With the increasing worldwide adoption of Electronic Health Records, many research groups made use of the availability of such large data volumes for different medical research purposes. Epidemiology studies are use cases that are becoming increasingly popular. An example application is finding associations between late-preterm birth and persistent asthma in young children, inhaled corticosteroid use and more acute respiratory visits, by conducting a retrospective cohort analyses [10]. EHR data was also used to examine the association between the exposure to psychotherapy during pregnancy and the risk reduction of adverse pregnancy outcomes among women with history of major depressive disorder and obstetric outcomes [21]. Other examples include associating glucocorticoid use to predicting fracture risk [22], using EHR data to predict the epidemiology of disease biomarkers [23], and identifying the risk factors of Angiotensin

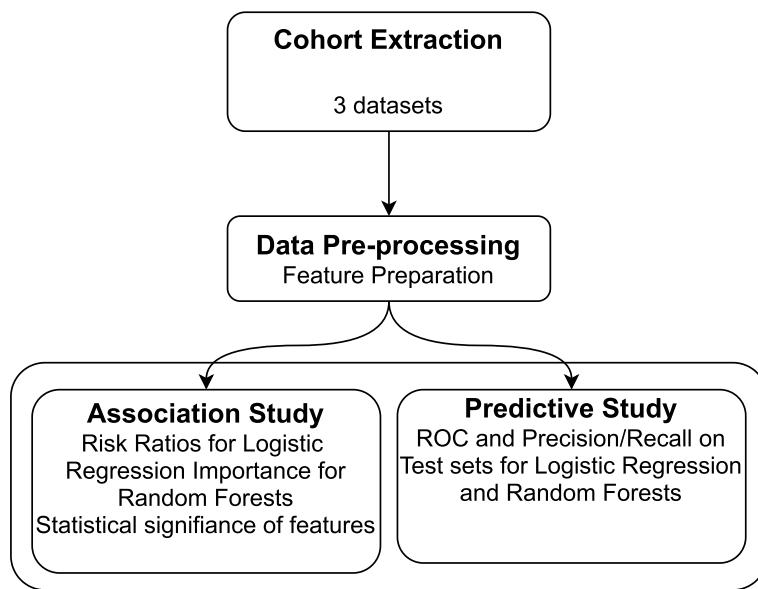


Fig. 1 Workflow of the data analysis

-converting enzyme inhibitors (ACEIs) by conducting a retrospective cohort study from EHR data [24].

Methodology

The analysis plan is illustrated in Fig. 1. We considered association analysis for risk factor identification and predictive analysis for preterm-birth prediction. In order to identify the variation of risk factors in the course of pregnancy, we performed association studies at three time points during pregnancy. These time points are 0, 12 and 24 gestational weeks. At each time point, only the past medical history with respect to the time point is included for each subject. For example, if the time point of analysis is 12 gestational weeks, all the data between 12 gestational weeks and the delivery event are discarded. This mimics a realistic scenario where future data is not available at each time point of analysis. Figures 2 and 5 illustrate the data timeline used in the analysis. The gestational age are extracted based on ICD-10 codes as described in the following section. The inclusion of the time dimension in the analysis gives us a predictive tool. The risk factors at different time points can be interpreted as risk predictors for preterm delivery. The association study is based on multivariate logistic regression model.

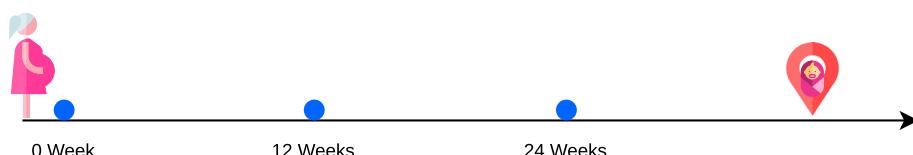


Fig. 2 Timeline of the clinical history for different pregnancies included in the cohort. Each line in the figure depicts a mother's history timeline. The dots on the lines represent delivery events. Green dots represent full-term deliveries, while red dots represent preterm deliveries. The post delivery visits were excluded from the data in the analyses. The dark gray portion before the delivery event is the time gap between the analysis time point and delivery. The data in this time interval were also omitted from the mother's history

Data extraction

We used Health Facts[®] Cerner EHR database to extract the dataset that is used in our analyses. The dataset consists of a retrospective and longitudinal cohort of full-term and preterm deliveries in the period 2001-2017. The dataset of our cohort is gathered from 120 hospitals in the USA. Gestational ages of the pregnancies were extracted using Z3A.xx codes which indicate weeks of gestation. For example Z3A.21 specifies 21 weeks gestation of pregnancy. We selected pregnancies with at least two ICD-10 codes of gestational age. We validated these codes based on the visit time stamps. We checked if the duration using EHR time stamps between the visits having gestational codes is consistent with the duration based on gestational ages extracted from the ICD-10 codes. For example Let's assume a pregnancy had two visits with ICD-10 codes of Z3A.25 (March 1) and the Z3A.38 (June 20). The difference based on gestational age is 13 weeks (91 days) while the difference between the visit time stamps is 110 days. This difference could be due to errors or delay in reporting the information in the EHR system. We discarded pregnancies if the error is higher than one week. Figure 3 shows a validation of the extracted gestational ages. Our assumption is that if the gestational ages extracted from the diagnosis codes are accurate then the time elapsed between two visits should be the same if calculated as the difference of gestational ages or as the difference of time stamps of the visits from EHR. For each subject we took the first visit as a reference point and subtracted the time stamps (x-axis) from the reference as well as the gestational ages (y-axis). Each point in the figure corresponds to information about one visit. In order to clean the data from erroneous gestational age we removed subjects having a large deviation in terms of time difference based on gestational age and timestamps. We discarded samples with a Root Mean Square Error (RMSE) larger than one week. Figure 4 shows an example of samples after removing the potential outliers.

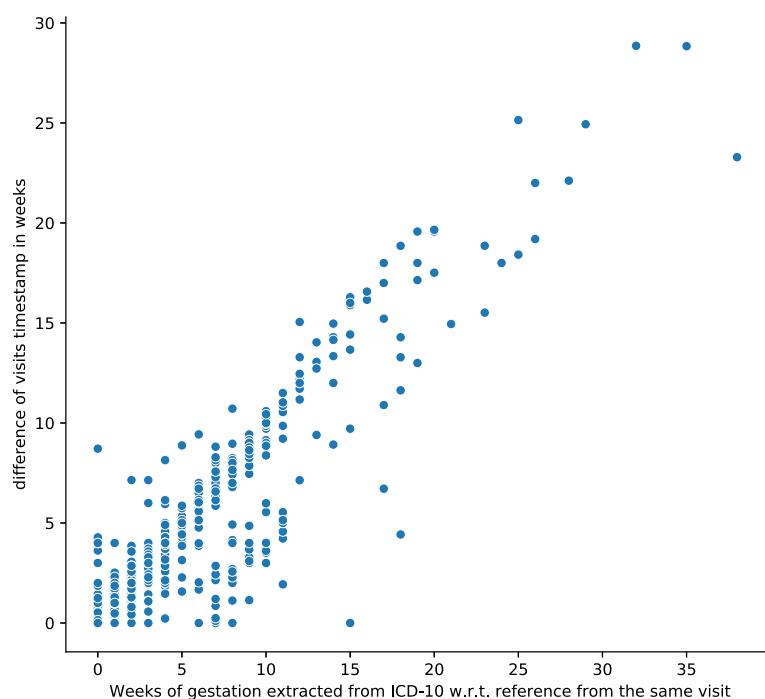


Fig. 3 Time difference of EHR visits based on extracted gestational age from ICD-10 visit codes

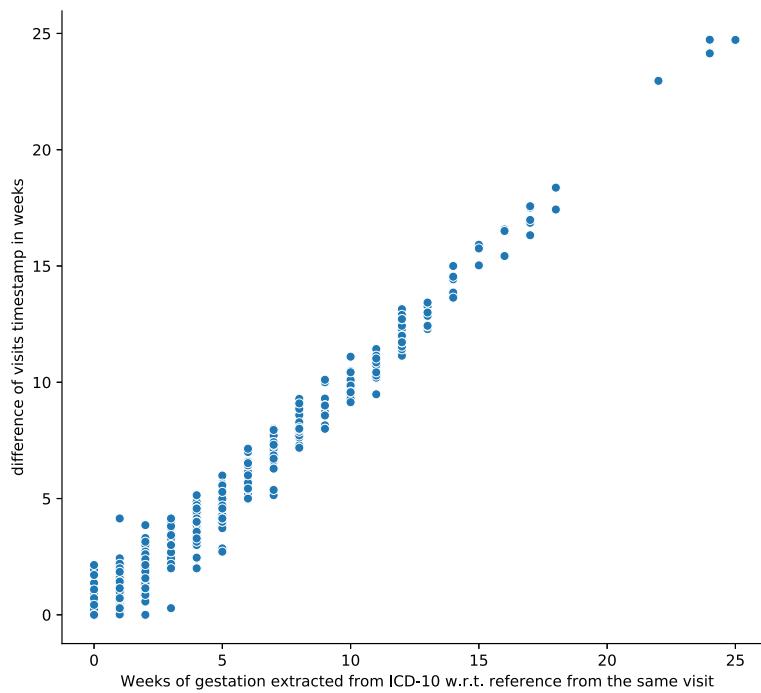


Fig. 4 Similar plot as in Fig. 3 with a filtered visits based on RMSE ≤ 1 week between the time difference from extracted gestational ages and visit timestamps

We relied on ICD-10 diagnosis codes to identify preterm and full-term pregnancies. The ICD-10 codes that was used to identify preterm and full deliveries are provided in the supplementary material. In order to validate the data extraction, we randomly selected 100 samples and presented them to a clinician for a manual check. He concluded that for 11 records among 100 it was not possible to assert whether the delivery was full or preterm. For the remaining records, the associated labels (full term delivery vs. preterm delivery) were accurate. We note the clinician decision was based on the information available from EHR which did not include clinical notes.

To evaluate the risk factors at different time intervals prior to delivery event, we excluded patient history from the delivery event until a defined time point. Figure 2 illustrates the data selection process. We first used a time gap of 0 week of gestation, and selected mothers that had at least two hospital visits in their history prior to the time gap. Mother E in Fig. 2 would not be selected in this case, as not enough history data is available for the analyses. These time points are illustrated in Fig. 5.

The medical history of each patient encompasses the following information: drug orders (using NDC encoding), procedures (using ICD-9 ICD-10, CPT4, and HCPCS encoding), diagnosis codes (using ICD-9 and ICD-10 encoding) and lab orders (using LOINC encoding). In addition to these, the dataset includes demographic information of the patients. The latter included are race, marital status, age, medical specialty and hospital ID. The demographic information are included to reduce the effects of confounding factors.

Risk factors identification

We used logistic regression as shown in Eq. (1) [25] for risk factor identification using four datasets, each containing the same set of patients, with a portion of the data omitted according to the previously mentioned time gaps.

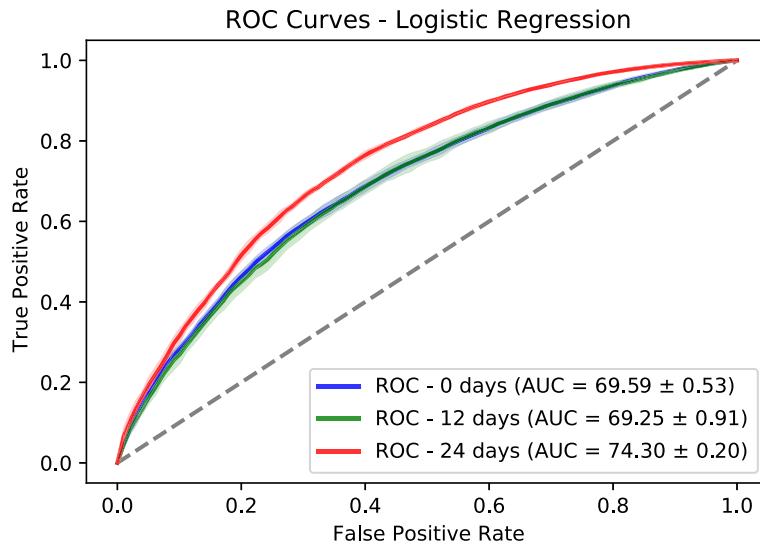


Fig. 5 Time points used to predict preterm-birth

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} \quad (1)$$

In Eq. (1), β_i are the coefficients to be estimated and x_i values are the corresponding independent variables that represent medications, diagnosis, lab orders, procedures and demographics. The value of i ranges from 1 to p , where p is the total number of dependent variables.

The dependent variable in the analysis y indicates whether the delivery is preterm $y_i = 1$ or full-term $y_i = 0$. We used Python's scikit-learn¹ implementation of logistic regression. We used an SGD training with L_2 regularization ($\alpha=0.01$) and tolerance=0.001, batch size of 5000 and 30 epochs.

After estimating the multiple logistic regression coefficients, we extracted, from the multivariate model, the p -values associated with each predictor. This allowed us to evaluate the statistically significant variables in the analysis. A Bonferroni correction was applied to these p -values. The covariates with the significant adjusted p -values were selected as risk factors in the corresponding time gap. The p -value cut-off was chosen to be 0.01 to retain the statistical significant variables. We report the risk ratios as well as their 95% confidence intervals corresponding to each identified statistically significant risk factor.

The rate of preterm birth is reported to be 1:10 in the US according to CDC². In our dataset the preterm birth rate is about 9.5%, which is very close to the reported risk ratios. In order to account for the potential mismatch between the observed and the real prevalence of preterm birth we use Eq. (2) to calculate the corrected risk ratio, where OR is the odds ratio and P_0 is the incidence of the covariate in the full-term group [26].

$$RR = \frac{OR}{(1 - P_0) + (P_0 \times OR)} \quad (2)$$

In addition to Logistic Regression model, we trained a Random Forests model based on a fast implementation in Ranger package [27]. The number of trees is set to 500. The

¹<https://scikit-learn.org>

²<https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm>

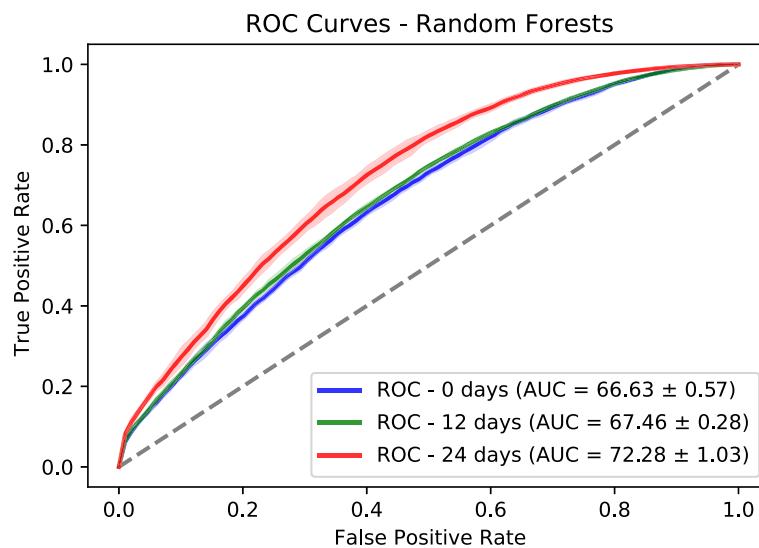


Fig. 6 ROCs on test sets based on Logistic Regression

number to possibly split in each node is 185. The regularization factor (gain penalization) is 1 [28]. These are the default values in Ranger package. We extracted the variable importance of each feature and the associated *p*-value based on an approximation of a permutation test [29]. We performed a Bonferroni *p*-value correction and we used the value of 0.01 to identify statistically significant features. We used a threshold of 3 on the variable importance to retain the most relevant features. We evaluated the predictor power of both logistic regression and random forests using a held-out test sets corresponding to 30% of the dataset size. The splitting and evaluation was repeated 10 times for each model. We used area under ROC and Precision-Recall curves as performance metrics. We provide mean and standard deviation for each metric. The results are reported in Figs. 6, 7, 8 and 9. Logistic regression performed better than random forests on the prediction of preterm-birth. Therefore we present the results of logistic regression in the main

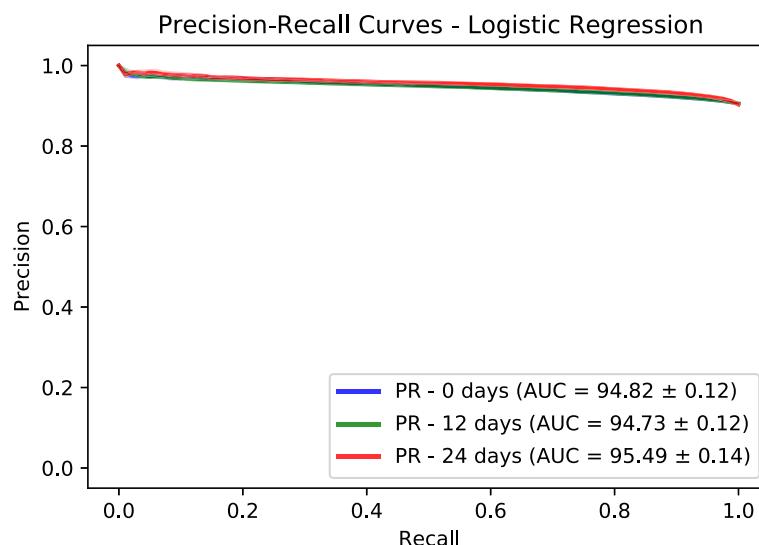


Fig. 7 ROCs on test sets based on Random Forests

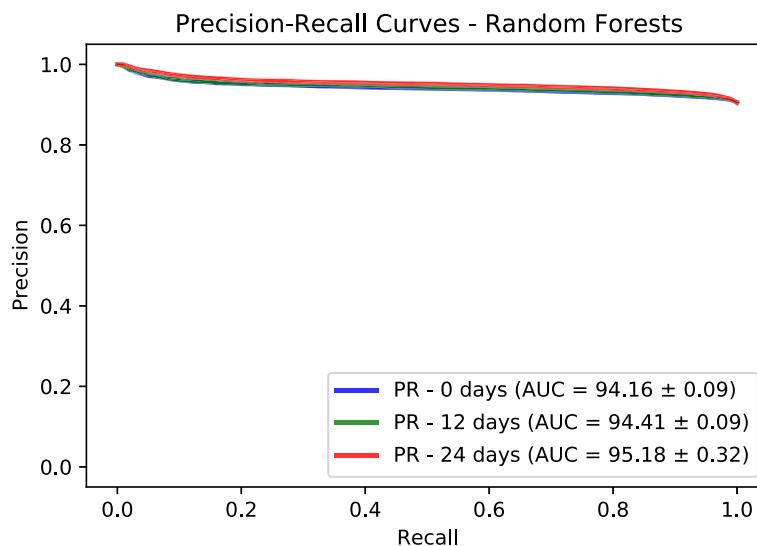


Fig. 8 Precision-Recall curves on test sets based on Logistic Regression

text and provide Random forests variable importance in the supplementary material. The code used for data analysis will be made available on github³.

Results and discussion

Datasets

The datasets that were used in the experiment contained about of 60,000 delivery events with a prevalence of about 9.5% of preterm deliveries. The size of the datasets are reported in Table 1. The covariates in the experiments are the different medications, procedures, lab orders, and diagnosis that occurred more than 100 times across all patient visits at each time gap. The number of covariates changes across the datasets, as some of them that appeared in a specific time gap from the delivery would disappear as the time gap between the delivery event and the start of considered history is made bigger. A summary description of the covariates for the different clinical modalities is given in Table 1.

Models evaluation

To evaluate the goodness-of-fit of the logistic regression and random forest models we used the Area Under the receiver operating characteristic Curve (AUC) and the Area under the Precision and Recall Curve (PRC) of the different datasets. We tuned the hyper-parameters of the Logistic Regression (LR) and Random Forests (RF) on the training dataset using a grid search approach. For LR, we tuned the type of regularization [L_1 , L_2] and the regularization factor with possible values as [0.001, 0.01, 0.1, 1]. The choice of L_2 and factor of 0.01 gave the best performance across the three prediction time points. The different AUC values are shown in Table 2. For RF, we tuned two hyper-parameters which are the number of trees (ntree) and the number of variables to possibly split at in each node (mtry). The possible values are [10, 100, 500] for ntree and [10, 50, 250] for mtry. The choice of ntree=10 and mtry=250 was the best from the grid search.

³https://github.com/bsabri/ptb_ehr

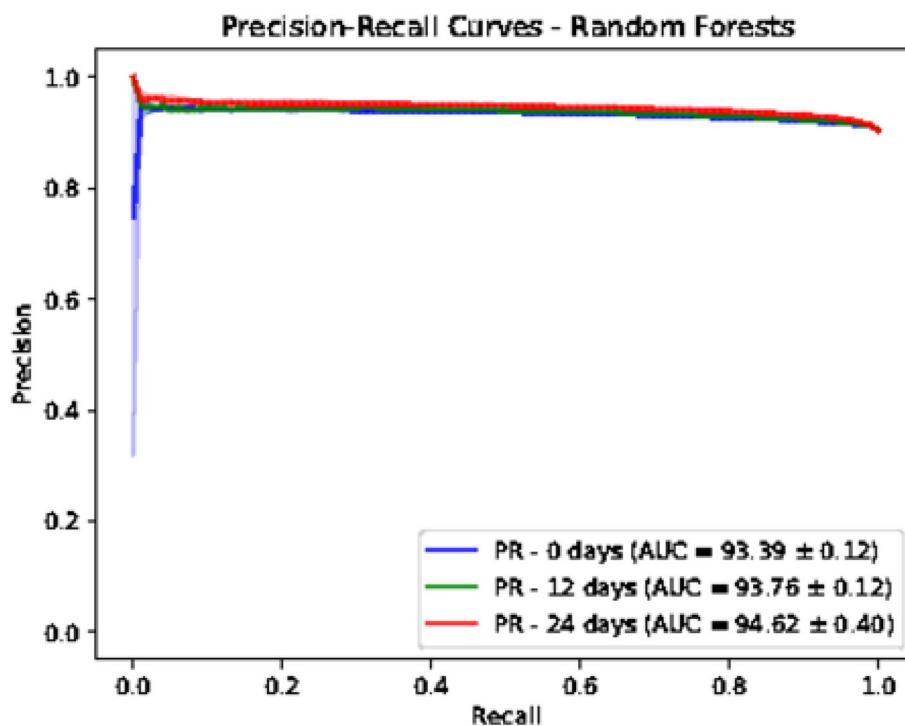


Fig. 9 Precision-Recall curves on test sets based on Random Forests

Identified risk factors based on logistic regression

Tables 3 and 4 summarize the significant risk factors identified. The risk factors that are identified in each time gap are those with a p -value ≤ 0.01 . To follow the progression of risk factors before and during the pregnancy at different stages, we group the risk factors that are common to all three time gaps and two time gaps (Table 3), as well as those that are specific to the 24 gestation-weeks time gap, 12 gestation-weeks time gap and 0 gestation-week time gap (Table 4). The tables show the medication, diagnosis, medical procedure and lab orders codes proceeded by M_, D_, P_ and L_ respectively in the first columns of the table. The corresponding descriptions of these codes are also included in the table. The description of medication codes consist of brand names and generic names of the medication separated by "/". Risk factors with a risk ratio between 0.9 and 1.1 are excluded from the analysis, as the coefficients of these risk ratios are very close to 1, and are found to not have a significant impact.

Table 1 Covariate Counts by Category

Category	0 Week	12 Weeks	24 Weeks
Diagnosis	17250	17169	17266
Medication	12892	12506	12287
Procedure	6282	6110	4596
Lab orders	2878	2811	2770
Total	39302	38596	36919
# samples	63814	60431	58696
Preterm %	9.5	9.4	9.6

Table 2 Prediction performance for Logistic Regression (LR) and Random Forests (RF) models

Number of weeks	0	12	24
AUC - LR	69.59±0.53%	69.25±0.91%	74.30±0.20%
AUC - RF	66.63±0.57%	67.46±0.28%	72.28±1.03%
PRC - LR	94.82±0.12%	94.73±0.12%	95.49±0.14%
PRC - RF	94.16±0.09%	94.41±0.09%	95.18±0.32%

Table 3 Common risk factor from Logistic Regression across all time gaps

	Code	Description	Risk Ratios(95% CI)		
			0 Week	12 Weeks	24 Weeks
All time gaps	D_644.21	Early Onset of Delivery, Delivered, with or without Mention of Antepartum Condition	2.03(1.96-2.1)	1.93(1.82-2.05)	1.78(1.6-1.95)
	M_85056605.0	Celestone Soluspan	1.76(1.66-1.85)	2.05(1.93-2.16)	1.81(1.71-1.92)
	M_517072001.0	Betamethasone Acetate-Betamethasone Sodium Phosphate	1.74(1.66-1.83)	1.92(1.82-2.03)	1.76(1.7-1.83)
2 time gaps	M_409672909.0	Magnesium Sulfate, Injectable/magnesium sulfate	1.68(1.49-1.87)		1.78(1.58-1.98)

Table 4 Risk factors from Logistic Regression unique to 0, 12 and 24 week(s)

	Code	Description	Risk Ratios(95% CI)
0 week	M_409672909.0	Magnesium Sulfate, Injectable/magnesium sulfate	1.68(1.49-1.87)
	L_26464-8	White Blood Cell Count, Blood	1.75(1.54-1.96)
	D_654.20	Previous Cesarean Section, Unspecified as to Episode of Care or Not Applicable	1.59(1.46-1.72)
	L_32766-8	Trichomonas vaginalis	1.57(1.43-1.71)
	M_409672903.0	Magnesium Sulfate, Injectable/magnesium sulfate	1.57(1.42-1.71)
	D_V12.09	Other Personal History of Infectious and Parasitic Disease	1.54(1.37-1.7)
	M_55390012110.0	Ondansetron Hydrochloride/ondansetron	1.51(1.34-1.68)
	D_V15.81	Personal History of Noncompliance with Medical Treatment, Presenting Hazards to Health	1.47(1.32-1.62)
	D_644.20	Early Onset of Delivery, Unspecified as to Episode of Care or Not Applicable	1.46(1.31-1.61)
	L_741-9	Differential Microcytes	1.41(1.28-1.53)
12 weeks	L_10378-8	Polychromasia Qualitative Blood Light Microscopy	1.41(1.27-1.54)
	D_645.11	Post Term Pregnancy, Delivered, with or without Mention of Antepartum Condition	0.49(0.35-0.63)
	D_O30.041	Twin pregnancy, dichorionic/diamniotic, first trimester	1.65(1.62-1.69)
24 weeks	D_O34.219	Maternal care for unspecified type scar from previous cesarean delivery	1.57(1.39-1.75)
	D_O09.91	Supervision of high risk pregnancy, unspecified, first trimester	1.7(1.53-1.86)
	M_51079092920.0	Labetolol Hydrochloride/labetalol	1.73(1.58-1.87)
	D_O30.009	Twin pregnancy unspecified number of placenta and unspecified number of amniotic sacs unspecified trimester	1.45(1.39-1.51)

Risk factors common to all time gaps

The risk factors that are common to all time gaps, meaning, they indicate a high risk pregnancy before and during the current pregnancy, are the existence of a previous preterm delivery and possible pregnancy complications of previous pregnancy with preterm delivery. The previous preterm delivery is indicated by the diagnosis code D_644.21, and the medication Celestone Soluspan (Betamethasone), which is a prenatal corticosteroid that is administered in high risk pregnancies in order to reduce the risk of preterm infants having lung problem [30]. We clarify that Betamethasone as mentioned in Table 3 should not be interpreted as a risk factor for preterm birth but it is rather a proxy indicating the presence of previous pregnancy complications in the clinical history treated using Betamethasone.

Risk factors common to two time gaps

There is one significant factor common to two time gaps which is presented in Table 3. The risk factor which is a medication (Magnesium Sulfate, Injectable) prescribed for the treatment of pregnancy complication with risk of preterm delivery . This risk factor is also a proxy indicator of a previous preterm.

Risk factors at a single time gap

The results summary can be found in Table 4. The risk factors are most likely not relevant to the current pregnancy. They reflect preexisting chronic conditions in the mother's history and general predisposition to deliver on preterm. The factors can be mainly grouped into three categories: 1) previous delivery complications: (D_654.20, D_644.20, M_55390012110.0, M_409672903). Magnesium Sulfate (M_409672903) and Ondansetron (M_55390012110.0) are used for the management of pregnancy complications [31]. 2) Complications related to blood: (L_26464-, L_741-9, L_10378-8) such as anemia or iron deficiency and 3) infectious diseases (D_041.9, L_32766-8). Anemia has been previously correlated with the risk preterm delivery [32–34]. Our results indicate that Previous post-term deliveries (D_645.11) reduces the risk of preterm delivery.

For 12 and 24 weeks of gestation, the significant ICD-10 codes are related to current high risk pregnancy. Twin pregnancy (D_O30.041) is known to be a risk factor of preterm delivery [18]. We found that previous cesarean delivery D_O34.219 is associated with preterm delivery. The medication (M_51079092920.0/ Labetolol) is used to treat hypertension. This finding is supported by previous research that pre-existing hypertension increases the risk of preterm delivery [35].

Identified risk factors based on random forest

Similarly to Logistic Regression, Random Forest was used to identify the most important predictors for preterm delivery at different time points. We used a threshold of 2 for variable importance and the same *p*-value of 0.01 for identifying the statistical significant variables. The common risk factors for the three and two (0 and 12 weeks) time gaps as depicted in Table 5 is one factor which is a previous preterm delivery. Betamethasone is a proxy for a previous pregnancy with complications treated using this medication. The variables for two-time gaps are related to previous preterm delivery and twin pregnancy. Table 6 lists the risk factors unique to the time gap of 0, 12 and 24 week(s) of gestation. Three factors have been identified: Type 1 diabetes, blood complication identi-

Table 5 Risk factors from Random Forest common to two and three time gaps

	Code	Description	Variable Importance		
			0 Week	12 Weeks	24 Weeks
3 time gaps	M_517072001.0	Betamethasone Acetate-Betamethasone /betamethasone	3.53	3.73	9.98
2 time gaps	D_644.21	Early Onset of Delivery, Delivered, with or without Mention of Antepartum Condition	4.566	4.11	
	D_V23.41	Pregnancy with History of Pre-Term Labor	2.19	2.42	
	D_O30.041	Twin pregnancy, dichorionic/diamniotic, first trimester	7.17	6.03	

fied by a lab order and JC/BK virus. A previous study based on a large cohort has found an association of type 1 diabetes with an increased risk of preterm delivery [36]. Previous preterm delivery and hypertension condition are the main identified factors. As the pregnancy progresses acute conditions dominate the risk factors. For example, Oligohydramnios (D_O41.02X0) is a condition of reduced amniotic fluid around your baby during pregnancy. A previous study has shown the association between Oligohydramnios and preterm risk [37]. Also Coagulation disorder (Fibrinogen, L_3255-7) is a known risk factor for preterm delivery [38].

Table 6 Risk factors from Random Forest unique to 0, 12 and 24 week(s) of gestation

	Code	Description	Variable Importance
0 week	D_250.01	Diabetes Mellitus without Mention of Complication, Type I	2.64
	L_49024-3	Differential Cell Count Method Blood	2.09
	L_47251-4	JC/BK Virus PCR	2.07
12 weeks	D_401.9	Unspecified Essential Hypertension	2.13
	D_644.20	Early Onset of Delivery, Unspecified as to Episode of Care or Not Applicable	2.11
	D_O60.10X0	Preterm labor with preterm delivery, unspecified trimester, not applicable or unspecified	2.05
24 weeks	D_O30.032	Twin pregnancy, monochorionic/diamniotic, second trimester	7.62
	D_O34.32	Maternal care for cervical incompetence, second trimester	4.36
	D_O09.212	Supervision of pregnancy with history of pre-term labor, second trimester	3.95
	D_O10.912	Unspecified pre-existing hypertension complicating pregnancy, second trimester	3.76
	M_409672903.0	Magnesium Sulfate, Injectable/magnesium sulfate	2.80
	L_3255-7	Fibrinogen Automated	2.49
	D_O41.02X0	Oligohydramnios, second trimester, not applicable or unspecified	2.12

Limitations

While our study supports the importance EHR data in the prediction of preterm delivery, it highlighted several limitations with this method. The accuracy of selection of the appropriate diagnosis or procedure codes is questioned. Similarly, in order for the laboratory values to make sense, they need to be stratified into positive or negative values. A more clinically relevant prediction model will be able to combine the risk factors and produce a risk index.

Conclusion

In this work, we used a large dataset from electronic health records to identify risk factors of preterm delivery based on two models namely Logistic Regression and Random Forest. The obtained factors overlap to a large extend between both models. We explored the progression of these risk factors across different stages of the pregnancy. We performed the association analysis at three time gestational ages (0, 12 and 24 weeks). Risk factors before and in early pregnancy related to chronic condition and general predisposition such as hypertension, diabetes and anatomy. Risk factors at advanced stage of pregnancy are mostly related to the current pregnancy and reflect acute condition such as infections, complications in blood. Most of the identified factors could be confirmed from the literature. The results from this experiment are easily reproducible to other diseases. Some risk factors can change over time with the change in the society such as lifestyle, income level etc. Re-evaluating the risk factors will be easy using electronic health records. As future work, we plan to extend this analysis by adding other modalities (microbiology, physiological measurements) and test results of lab order to better refine the obtained results.

Acknowledgements

We would like to thank Haithum Elhadi for his support on EHR data.

Authors' contributions

ZS implemented the part of the methods, conducted the experiments, and drafted the manuscript. SB prepared the data, implemented part of the methods, supervised the work, and edited the manuscript. HA gave feedback and edited the manuscript. MM supervised the work. FF designed data extraction and experiments and edited the manuscript. All authors read and approved the final manuscript.

Funding

This work is funded by Sidra Internal Research Funds: SDR200033.

Availability of data and materials

The data is part of Cerner Health Facts. It is protected by copyright and obtained from Cerner through a Data Use Agreement.

Declarations

Ethics approval and consent to participate

The project obtained from Sidra IRB office an Exempt IRB protocol # 1807026901.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Research Department, Sidra Medicine, Doha, Qatar. ²Division of Neonatology, Sidra Medicine, Doha, Qatar. ³Department of Maternal-Fetal Medicine, Sidra Medicine, Doha, Qatar. ⁴Qatar Computing Research Institute, HBKU, Doha, Qatar.

Published online: 17 August 2022

References

1. Obstet A. Who: recommended definitions, terminology and format for statistical tables related to the perinatal period and use of a new certificate for cause of perinatal deaths. *Acta Obstet Gynecol Scand.* 1977;56(3):247–53.
2. Chawanpaiboon S, Vogel JP, Moller A-B, Lumbiganon P, Petzold M, Hogan D, Landoulsi S, Jampathong N, Kongwattanakul K, Laopaiboon M, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health.* 2019;7(1):37–46.
3. Chawanpaiboon S, Vogel JP, Moller A-B, Lumbiganon P, Petzold M, Hogan D, Landoulsi S, Jampathong N, Kongwattanakul K, Laopaiboon M, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Health.* 2019;7(1):37–46.
4. Johnson S, Marlow N. Early and long-term outcome of infants born extremely preterm. *Arch Dis Child.* 2017;102(1):97–102.
5. Johnson S, Hennessy E, Smith R, Trikic R, Wolke D, Marlow N. Academic attainment and special educational needs in extremely preterm children at 11 years of age: the epicure study. *Arch Dis Child Fetal Neonatal Ed.* 2009;94(4):283–89.
6. Bryce J, Boschi-Pinto C, Shibuya K, Black RE. WHO estimates of the causes of death in children. *Lancet.* 2005;365(9465):1147–52. [https://doi.org/10.1016/S0140-6736\(05\)71877-8](https://doi.org/10.1016/S0140-6736(05)71877-8).
7. Wolke D. Psychological development of prematurely born children. *Arch Dis Child.* 1998;78(6):567–70. <https://doi.org/10.1136/adc.78.6.567>.
8. Zhang Y, Inder TE, Neil JJ, Dierker DL, Alexopoulos D, Anderson PJ, Van Essen DC. Cortical structural abnormalities in very preterm children at 7years of age. *NeuroImage.* 2015;109:469–79. <https://doi.org/10.1016/j.neuroimage.2015.01.005>.
9. Wocadol C, Rieger I. Motor impairment and low achievement in very preterm children at eight years of age. *Early Hum Dev.* 2008;84(11):769–76. <https://doi.org/10.1016/j.earlhumdev.2008.06.001>.
10. Goyal NK, Fiks AG, Lorch SA. Association of late preterm birth with asthma in young children: Practice based study. *Pediatrics.* 2011;128(4):830–38. <https://doi.org/10.1542/peds.2011-0809>.
11. Reddy UM, Rice MM, Grobman WA, Bailit JL, Wapner RJ, Varner MW, Thorp JM, Leveno KJ, Caritis SN, Prasad M, Tita ATN, Saade GR, Sorokin Y, Rouse DJ, Blackwell SC, Tolosa JE, Spong C, Tolivaisa S, Talucci M, Zylfajaj M, Reid Z, Leed R, Benson J, Forester S, Kitto C, Davis S, Falk M, Perez C, Hill K, Sowles A, Postma J, Alexander S, Andersen G, Scott V, Morby V, Jolley K, Miller J, Berg B, Dorman K, Mitchell J, Kaluta E, Clark K, Spicer K, Timlin S, Wilson K, Moseley L, Santillan M, Price J, Buentipo K, Bludau V, Thomas T, Fay L, Melton C, Kingsbery J, Benezeue R, Simhan H, Bickus M, Fischer D, Kamon T, DeAngelis D, Mercer B, Milluzzi C, Dalton W, Dotson T, McDonald P, Brezine C, McGrail A, Latimer C, Guzzo L, Johnson F, Gerwig L, Fyffe S, Loux D, Frantz S, Cline D, Wylie S, Iams J, Wallace M, Northen A, Grant J, Colquitt C, Rouse D, Andrews W, Mallett G, Ramos-Brinson M, Roy A, Stein L, Campbell P, Collins C, Jackson N, Dinsmoor M, Senka J, Paycheck K, Peaceman A, Moss J, Salazar A, Acosta A, Hankins G, Hauff N, Palmer L, Lockhart P, Driscoll D, Wynn L, Sudz C, Dengate D, Girard C, Field S, Breaule P, Smith F, Annunziata N, Allard D, Silva J, Gamage M, Hunt J, Tillinghast J, Corcoran N, Jimenez M, Ortiz F, Givens P, Rech B, Moran C, Hutchinson M, Spears Z, Carreno C, Heaps B, Zamora G, Seguin J, Rincon M, Snyder J, Farrar C, Lairson E, Bonino C, Smith W, Beach K, Van Dyke S, Butcher S, Thom E, Zhao Y, McGee P, Momirova V, Palugod R, Reamer B, Larsen M, Spangler T, Bhandaru V, VanDorsten JP. Serious maternal complications after early preterm delivery (24–33 weeks' gestation). *Am J Obstet Gynecol.* 2015;213(4):538–15389.
12. Menard MK. Toward achieving risk-appropriate maternity care: Maternal morbidity prediction. *Obstet Gynecol.* 2019;134(2):213–15.
13. AlSaad R, Malluhi Q, Boughorbel S. PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks. *BioData Min.* 2022;15(1):6. <https://doi.org/10.1186/s13040-022-00289-8>.
14. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A, Ong S, Pell JP, Southworth MR, Stough WG, Thoenes M, Zannad F, Zalewski A. Electronic health records to facilitate clinical research. *Clin Res Cardiol.* 2017;106(1):1–9. <https://doi.org/10.1007/s00392-016-1025-6>.
15. Ip M, Peyman E, Lohsoonthorn V, Williams MA. A case-control study of preterm delivery risk factors according to clinical subtypes and severity. *J Obstet Gynaecol Res.* 2010;36(1):34–44. <https://doi.org/10.1111/j.1447-0756.2009.01087.x>.
16. Vogel JP, Chawanpaiboon S, Moller A-B, Watananirun K, Bonet M, Lumbiganon P. The global epidemiology of preterm birth. *Best Pract Res Clin Obstet Gynecol.* 2018;52:3–12. <https://doi.org/10.1016/j.bpobgyn.2018.04.003>.
17. Purisch SE, Gyamfi-Bannerman C. Epidemiology of preterm birth. *Semin Perinatol.* 2017;41(7):387–91. <https://doi.org/10.1053/j.semperi.2017.07.009>.
18. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. *Lancet.* 2008;371(9606):75–84. [https://doi.org/10.1016/S0140-6736\(08\)60074-4](https://doi.org/10.1016/S0140-6736(08)60074-4).
19. Zhang Q, Ananth CV, Li Z, Smulian JC. Maternal anaemia and preterm birth: a prospective cohort study. *Int J Epidemiol.* 2009;38(5):1380–89. <https://doi.org/10.1093/ije/dyp243>.
20. Ji X, Meng X, Liu C, Chen R, Ge Y, Kan L, Fu Q, Li W, Tse LA, Kan H. Nitrogen dioxide air pollution and preterm birth in Shanghai, China. *Environ Res.* 2019;169:79–85. <https://doi.org/10.1016/j.envres.2018.11.007>.
21. Snapper LA, Hart KL, Venkatesh KK, Kaimal AJ, Perlis RH. Cohort study of the relationship between individual psychotherapy and pregnancy outcomes. *J Affect Disord.* 2018;239:253–57. <https://doi.org/10.1016/j.jad.2018.05.083>.
22. Robinson DE, Staa T. P. v., Dennison EM, Cooper C, Dixon WG. The limitations of using simple definitions of glucocorticoid exposure to predict fracture risk: A cohort study. *Bone.* 2018;117:83–90. <https://doi.org/10.1016/j.bone.2018.09.004>.
23. Mosley JD, Feng Q, Wells QS, Driest SLV, Shaffer CM, Edwards TL, Bastarache L, Wei W-Q, Davis LK, McCarty CA, Thompson W, Chute CG, Jarvik GP, Gordon AS, Palmer MR, Crosslin DR, Larson EB, Carroll DS, Kullo JJ, Pacheco JA, Peissig PL, Brilliant MH, Linneman JG, Namjou B, Williams MS, Ritchie MD, Borthwick KM, Verma SS, Karnes JH, Weiss ST, Wang TJ, Stein CM, Denny JC, Roden DM. A study paradigm integrating prospective epidemiologic

- cohorts and electronic health records to identify disease biomarkers. *Nat Commun.* 2018;9(1):3522. <https://doi.org/10.1038/s41467-018-05624-4>.
- 24. Banerji A, Blumenthal KG, Lai KH, Zhou L. Epidemiology of ACE Inhibitor Angioedema Utilizing a Large Electronic Health Record. *J Allergy Clin Immunol Pract.* 2017;5(3):744–49. <https://doi.org/10.1016/j.jaip.2017.02.018>.
 - 25. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression vol. 398. Hoboken: Wiley; 2013.
 - 26. Zhang J, Yu KF. What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA.* 1998;280(19):1690–91. <https://doi.org/10.1001/jama.280.19.1690>.
 - 27. Wright MN, Ziegler A. ranger: A fast implementation of random forests for high dimensional data in c++ and r. arXiv preprint arXiv:1508.04409. 2015.
 - 28. Deng H, Runger G. Feature selection via regularized trees. In: The 2012 International Joint Conference on Neural Networks (IJCNN). Brisbane: IEEE; 2012. p. 1–8.
 - 29. Janitza S, Celik E, Boulesteix A-L. A computationally fast variable importance test for random forests for high-dimensional data. *ADAC.* 2018;12(4):885–915.
 - 30. Crowther CA, McKinlay CJ, Middleton P, Harding JE. Repeat doses of prenatal corticosteroids for women at risk of preterm birth for improving neonatal health outcomes. *Cochrane Database Syst Rev.* 2011;6:003935. <https://doi.org/10.1002/14651858.CD003935.pub3>.
 - 31. Suarez EA, Boggess K, Engel SM, Stürmer T, Lund JL, Funk MJ. Ondansetron use in early pregnancy and the risk of late pregnancy outcomes. *Pharmacoepidemiol Drug Saf.* 2021;30(2):114–25.
 - 32. Dotters-Katz SK, Grotegut CA, Heine RP. The effects of anemia on pregnancy outcome in patients with pyelonephritis. *Infect Dis Obstet Gynecol.* 2013;2013:780960. <https://europepmc.org/articles/PMC3863467>.
 - 33. Ma M, Zhu M, Zhuo B, Li L, Chen H, Xu L, Wu Z, Cheng F, Xu L, Yan J. Use of complete blood count for predicting preterm birth in asymptomatic pregnant women: A propensity score-matched analysis. *J Clin Lab Anal.* 2020;34(8):23313.
 - 34. Rahmati S, Azami M, Badfar G, Parizad N, Sayehmiri K. The relationship between maternal anemia during pregnancy with preterm birth: a systematic review and meta-analysis. *J Matern Fetal Neonatal Med.* 2020;33(15):2679–89.
 - 35. Catov JM, Nohr EA, Olsen J, Ness RB. Chronic Hypertension Related to Risk for Preterm and Term Small-for-Gestational-Age Births. *Obstet Gynecol.* 2008;112(2 Pt 1):290–96. <https://doi.org/10.1097/AOG.0b013e31817f589b>. Accessed 28 Feb 2019.
 - 36. Ludvigsson JF, Neovius M, Söderling J, Gudbjörnsdóttir S, Svensson A-M, Franzén S, Stephansson O, Pasternak B. Maternal glycemic control in type 1 diabetes and the risk for preterm birth: a population-based cohort study. *Ann Intern Med.* 2019;170(10):691–701.
 - 37. Park JS, Yoon BH, Romero R, Moon JB, Oh S-Y, Kim JC, Jun JK. The relationship between oligohydramnios and the onset of preterm labor in preterm premature rupture of membranes. *Am J Obstet Gynecol.* 2001;184(3):459–62.
 - 38. Teraoka Y, Miyoshi H, Oshima K, Urabe S, Tanaka N, Kudo Y. Prenatal and peripartum management of patients with hypofibrinogenemia resulted in two successful deliveries. *Case Rep Obstet Gynecol.* 2017;2017:9427359. <https://doi.org/10.1155/2017/9427359>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

