Identifying the Early Signs of a Preterm Birth: A Large Cohort Study

Alireza Ebrahimvandi [1,2*], Niyousha Hosseinichimeh[3], Zhenyu James Kong[1]

[1]Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, United States of America

[2]University of Maryland Medical System, Linthicum Heights, Maryland, United States of America

[3]Industrial and Systems Engineering, Virginia Tech, Falls Church, Virginia, United States of America

* Corresponding author

Alireza Ebrahimvandi: alvandi@vt.edu; alireza.ebrahimvandi@umm.edu
Niyousha Hosseinichimeh: niyousha@vt.edu
Zhenyu James Kong: zkong@vt.edu

Corresponding authors:

Alireza Ebrahimvandi, Ph.D.
Email: alvandi@vt.edu; alireza.ebrahimvandi@umm.edu

Abstract:

Background and Purpose— Preterm birth (PTB) is the leading cause of infant mortality in the U.S. and globally. The goal of this study is to increase understanding of PTB risk factors that are present early in pregnancy by leveraging statistical and machine learning techniques on big data.

Methods—The 2016 U.S. birth records is obtained and combined with two other area-level datasets, Area Health Resources File and County Health Ranking. Then, we applied multiple machine learning techniques to study a cohort of 3.6 million singleton deliveries to identify generalizable preterm risk factors.

Results—The most important predictors of preterm birth are gestational and chronic hypertension, interval since last live birth, and history of a previous preterm birth that can respectively explain 14.91%, 6.92%, and 6.50% of the AUC. Parents education is one of the influential variables in prediction of PTB explaining 10.5% of the AUC. The relative importance of race declines when parents are more educated or have received adequate prenatal care. The gradient boosting machines outperformed other machine learning techniques with an AUC of 0.75 (recall: 0.64, specificity: 0.73) for the validation dataset.

Conclusions—Application of ML techniques improved the performance measures in prediction of preterm birth. The results emphasize the importance of socioeconomic factors such as parental education as one of the most important indicators of a preterm birth. More research is needed on the mechanisms through which the socioeconomic factors affect the biological responses.

Keywords: Racial disparities, education, statistical analysis, neural networks, socioeconomic factors

1 Introduction

Preterm birth (PTB), which is defined as a birth before 37 weeks of pregnancy, is the leading cause of infant mortality in the U.S. and in the world (1). In 2013, PTB accounted for 36% of U.S. infant deaths in their first year of life (2). In addition to the monetary cost of PTB, which exceeds 25 billion dollars annually, these babies may suffer from life-long deficiencies (3, 4). Many of the current interventions for reducing the likelihood of a preterm delivery like progesterone therapy are effective only if administered early—between 16 and 24 weeks of gestation—in the pregnancy (5). In prenatal care settings, patients can be enrolled in helpful interventions for reducing the behavioral risks without significant disruption of services (6). Therefore, it is critical to study risk factors of a preterm delivery that are present early or even before pregnancy. In addition, identifying the risk factors might help define a population useful for studying specific interventions. The identification of risk factors might also provide insight into the mechanisms of preterm birth which is still largely unknown (7, 8).

A large and growing body of literature has focused on finding the individual risk factors of preterm birth (7, 9, 10). The most important individual risk factor for predicting preterm delivery is a history of a previous PTB (both indicated and spontaneous) (11-13). Race is another major predictor for a PTB. The preterm birth rate (PBR) among non-Hispanic (NH) Black is 52% more than NH White—13.77 vs. 9.04 respectively (14). Other significant risk factors of preterm birth include age (15), short cervix between 16 to 28 weeks of pregnancy (16), and chronic medical disorders like hypertension (17) or diabetes (18). Some studies attempted to increase the generalizability of the risk factors by including large cohorts in their studies (19). Machine learning techniques are extensively used in advancing the understanding of spontaneous PTB risk factors (20-24).

Despite the vast body of literature on the risk factors of PTB, very few interventions have been proven to effectively prolong gestational age in at-risk women (13, 25). This is partly because two-thirds of preterm deliveries happen to women with no risk factors (26). The current risk assessment in the obstetrical population shows limitation because of the low prevalence of individual risk factors in the general obstetric population (27). For example, the most important risk factor for preterm birth in singleton pregnancies is the history of a previous PTB (14, 27). However, the history of a previous PTB is not applicable to the women without a prior birth (nulliparous) which includes more than a third of the total births. Many of the proposed studies consider only the main effect of the individual risk factor of PTB while controlling for a limited number of confounding variables and interactions that were selected manually (10, 20, 26, 28). In sum, previous studies have not examined PTB risk factors that are present in early pregnancy on a dataset that is representative of the whole population while controlling for diverse confounding factors and their interactions. To address this issue, we use proper machine learning (ML) methods that have the capability of checking high-order interactions with minimal supervision. The importance of considering interactions is that it enhances the capability of the model to capture complex relationships. We also use a comprehensive dataset that can increase generalizability and our understanding of preterm birth risk factors, as it enables us to check interactions of risk factors in the general obstetrical population. In this study, we focus on identifying the risk factors of preterm birth (for both indicated and spontaneous), which are present early in pregnancy.

2 Methods

2.1 Study Population

We obtained the 2016 birth records that are collected by the CDC (29). We then combined the birth records with other data sources including the County Health Rankings, and the Area Health Resources File (see Appendix A). All datasets were linked using a common geographical identifier, the FIPS county codes. This allows us to integrate and examine multiple influences on preterm birth. We performed the data cleaning and preparation in *STATA 14.0* and the processing has been coded in *R 4.0.3.* Data preprocessing

2.2 Data Elements

The merged dataset includes 3,664,509 observations with 77 variables. The CDC dataset contains variables that are collected through a self-reported survey at the time of birth from both practitioners and parents. We trained an unsupervised autoencoder deep neural network (DNN) to detect any possible anomalies in the data. This step removes 5.07% of the records and at the same time, it keeps the proportion of singleton preterm birth at 7.73%, which is close to the initial distribution at 8.02% (see Appendix A for more details). The final dataset includes 3,610,827 observations with 77 variables (see Appendix B for a complete list of variables).

Data visualization is a challenging but insightful task in this study due to a large number of observations. We used *Violin* graphs from the *ggplot2* package in *R* to plot the data and gain more information about the features and their relationship with preterm birth. Appendix C shows the visualization of each variable.

2.3 Model Development

Our dataset has five characteristics that guide us in the selection of the methods. First, the distribution of the response variable is imbalanced. Preterm birth in singleton pregnancies occurs only in eight percent of the deliveries and the remaining are full-term. Second, many of the features such as age and education have collinearity (Pearson's correlation coefficient= 0.41). This will limit the use of methods like logistic regression which has the assumption of little or no multicollinearity between independent features. Third, we are interested in finding significant interactions among the variables. One of the best methods for learning the interactions with minimal supervision is decision trees (30). Fourth, our dataset has 3.6 million records with 77 variables, which limits the use of methods that are memory intensive like support vector machines. Fifth, the dataset has 20 categorical variables. This will limit the application of distance-based methods like K-Nearest Neighbor. Based on these five characteristics, we apply regularized logistic regression, random forest, gradient boosting machines (GBM), and LightGBM on our dataset (see Appendix D for more details).

We used a grid search to find the best hyperparameters of logistic regression and random forest. However, we coupled Bayesian optimization (BO) with the ML performance measures to reduce training time for the GBM and lightGBM. The BO reduces the training time by sequentially solving an optimization problem that tries to find the best set of hyperparameters that have the potential to improve the outcomes in fewer iterations compared to an exhaustive grid search (31-33). To prevent overfitting and reducing run-time, we also use early stopping methods (1e-4 after 5 rounds). We used a system equipped with a Core i7 2.50 GHz processor, and a 32.0 GB memory, with an *Ubuntu 18.04.3* operating system.

2.4 Handling Missing Values and Model Assessment

To handle missing observations and categorical variables, we use a method in which strings are internally mapped to integers, and splits are done over these integers. The performance metrics that we use in this study focuses on the true positive rate (Sensitivity or Recall) because it is more important to correctly identify a preterm birth rather than mislabeling a full-term as otherwise.

2.5 Interpretation Techniques

To get the 'effect size' of each variable on the response, we use partial dependence plots (PDP). This is a useful tool for our study, particularly because we consider high-order interactions between our independent variables. Partial dependence plot returns the marginal 'effect size' of each variable on the response after accounting for the effect (average) of other responses:

$$\overline{f_s}\left(X_s\right)=\frac{1}{N}\sum_{i=1}^{N} f\left(X_s, x_{ic}\right).$$

Where X_c and X_s complement the set of X, and $\left\{x_{1c}, x_{2c}, \ldots, x_{Nc}\right\}$ are the values of X_c occurring in the training dataset of X.

It is important to note that the PDP does not ignore the effect X_c. The latter case can be estimated by $\tilde{f}_s\left(X_s\right)=\frac{1}{N}\sum_{i=1}^{N} f\left(X_s, x_{ic} \middle| X_s\right)$. The quantities $\overline{f_s}$ and $\tilde{f}_s$ will be the same only if the two events of c and s are independent, which is an unlikely situation.

3 Results

We randomly separated 75% of the data for the training set and the remaining 25% for validation purposes. The performance metrics are reported for the test set that is not part of the training process. The number of cross-validations for the methods is five-fold.

3.1 Study Design

The parameters for Logistic Regression with Elastic Net regularization (LR-EN) are set as $\alpha = 0.25$, and $\lambda = 2.125E-4$ after performing a grid search. The results of Bayesian optimization for tuning the parameters of Gradient Boosting Machines return 480 decision trees (*ntrees*) with a learning rate of $\eta = 0.04$ and an annealing rate of 0.99. The maximum depth is 13 for each tree. This means that each tree checks up to 13 interactions among variables. Each tree is trained on a random sample of observations, n = 0.55* N, and each split of the tree is performed on a random sample of features, p = 0.80*M. The optimization result for LightGBM returns *ntrees* =280, $\eta = 0.008$, and maximum depth of 14. We also used "Lossguide" for the "grow" policy, "dart" for booster type, and "histogram" for tree method in the LightGBM method. For a detailed list of the hyperparameters, see Appendix E.

3.2 Results of the machine learning algorithms

Table 1 provides the performance metrics for each method. Recall, specificity, and accuracy are a function of the cut-off threshold. Therefore, we report these metrics corresponding to the threshold that returns the highest mean per-class accuracy for all of the methods. Logistic regression with elastic net regularization (LR-EN) and random forest return very close testing and training AUC, which shows that they do not overfit to the noise. However, their AUC metrics are less than the gradient boosting machines (GBM) and the LightGBM on both testing and training datasets. The LightGBM returns the highest testing AUC at 75.91%. We pick the GBM as the best model for the prediction of preterm birth because it returns a slightly higher recall (TPR) at 64.82% while maintaining the specificity at a comparable rate (73.01%) with LightGBM (73.93).

Table 1 Performance metrics for machine learning models

Method	Train AUC (%)	Test AUC (%)	Recall (%)	Specificity (%)	Accuracy (%)
LR-EN	66.59	66.61	51.98	71.68	70.22
RF	70.24	70.78	57.36	73.01	71.78
GBM	**77.94**	**75.58**	**64.82**	**73.01**	**72.37**
LightGBM	78.34	75.91	62.24	73.93	72.99

3.3 Comparison with other studies

There are few similar studies that used high-dimensional dataset in their studies. Weber, Darmstadt (20) developed their model on a high dimensional dataset with 1000 initial features and 2.7 million observations. However, they developed their predictive model for the early spontaneous preterm birth, which happens at a much lower rate of 1.02% compared to the singleton preterm deliveries at 7.63% in our study. Another study by Alleman, Smith (19) has the closest setup in terms of developing the predictive model for singleton pregnancies but has a smaller dataset compared to our study.

Table 2 shows the comparison between the performance of our best GBM with the most relevant preterm birth studies. The criteria for inclusion of a paper is that it has to either use data with a large sample size that includes demographical information as predictors or it has used machine learning techniques for building a predictive model for preterm birth. We report the sample size, prevalence of the positive class, test AUC, recall, and specificity for each study. As can be seen in Table 2, our best GBM model outperforms the frameworks in these studies by improving the AUC by more than 5%, 9%, and 13% compared to the work of Goodwin, Iannacchione (34), Alleman, Smith (19) and Weber, Darmstadt (20), respectively. The improvement in the combined AUC, recall, and accuracy stems from pre-processing steps that remove anomaly and noise removal, regularization methods, an optimized set of hyperparameters, and the superior ability of the GBM algorithms in the extraction of high-level features in the data.

Table 2 Performance comparison with the most related studies.

Model	Method	Sample size (n)	Prevalence of Positive Class (%)	Test AUC (%)	Recall (%)	Specificity (%)

Goodwin et al., 2002	Neural nets, Stepwise LR	19970	22.20	72.00	NR	NR
Vovsha et al., 2014	SVM with Radial Basis kernel	3002	NR	NR	57.60	62.10
Alleman et al., 2014	LR	2509	7.50	69.50**	31.20	90.60
Weber et al., 2018*	Super learner (Combination of RF, lasso, ridge)	336,214	1.02*	67.00	62.00	65.00
Best model in this study	GBM	3,610,827	7.73	75.58	64.82	73.01

*Early (before 32 weeks) spontaneous preterm

** Training AUC

NR= Not reported, LR= Logistic regression, RF= Random forest, SVM= Support vector machine

3.4 Interpretations

Figure 1 shows the scaled importance of the top 15 variables in the prediction of preterm birth in the obstetric population (See Appendix F for more details). The absolute percentage of AUC attributed to each variable is also shown in front of each variable. Hypertension ("hyper"), interval since last live birth ("interval"), and history of PTB ("Previous_preterm") are the most important predictors of preterm birth that can respectively explain 14.91, 6.92, and 6.5% of the AUC. Mothers' pre-pregnancy BMI is also an important predictor of preterm birth. Figure 1 shows this interesting result that race has less relative importance when we consider factors like parent's education, age, and adequacy of care during pregnancy.

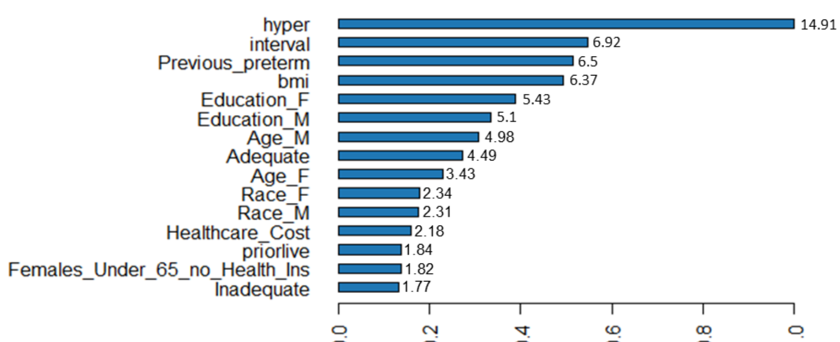


Figure 1 Variable importance plot

Building the model on a high dimensional dataset that is representative of almost all the deliveries in the U.S. indicates that the level of parent education is a more important predictor than demographic characteristics like race. If considered as the single explanatory variable, race is a significant predictor of both preterm birth and infant mortality where African American mothers have consistently been at a higher risk of preterm delivery (14, 35). In 2016, 10.88% of Black singleton pregnancies resulted in a preterm baby versus 7.11% for White mothers. Our results in Figure 2 show that this likelihood is 7.02% (P-Value<0.001) for Black versus 6.32% (P-Value<0.001) for White mothers when we account for the (average) effect of all factors such as education and age of parents, and adequacy of care during pregnancy in each class.

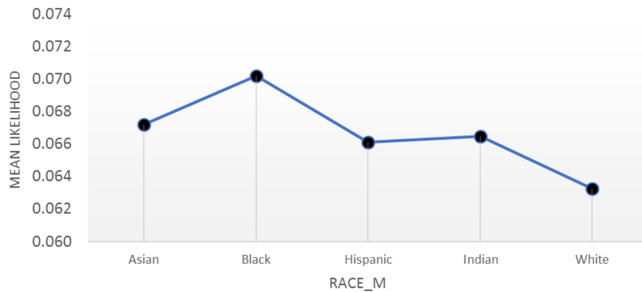
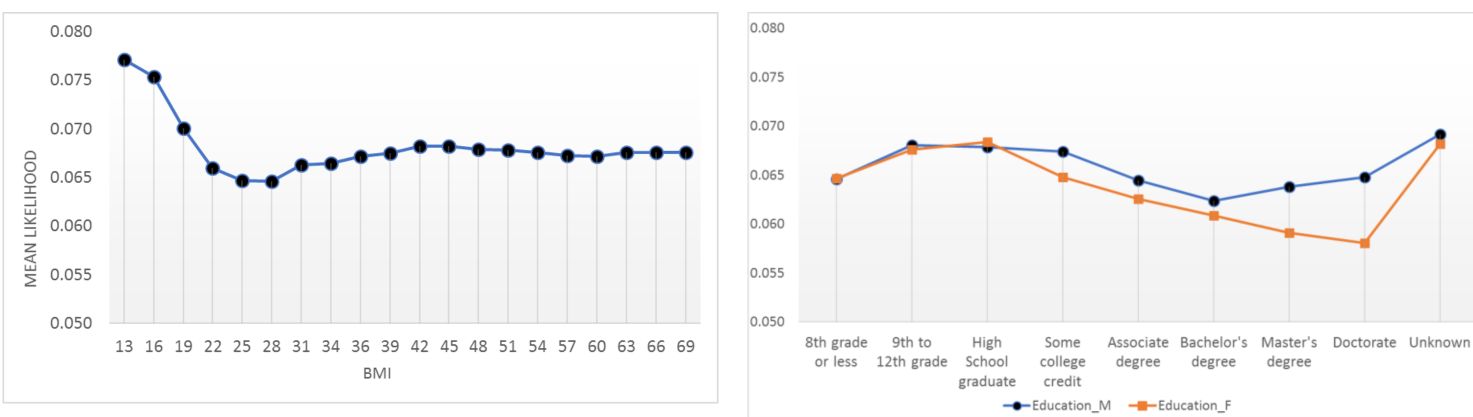


Figure 2 Partial dependence plot for Mother's Race

A partial dependence plot shows the 'effect' of a variable on the response—the likelihood of preterm birth—while accounting for the effect of other variables. Figure 3 shows two examples of partial dependence plots (PDP). Figure 3.a shows the relationship between a mother's BMI and the likelihood of preterm delivery. The PDP shows that mothers with very low BMI—less than 22—are at higher risk of delivering a preterm baby.

Figure 3.b shows the relationship between the parent's education and the likelihood of preterm birth. The likelihood of having a preterm infant for fathers decreases as their level of education increases. However, mothers with a Bachelor's degree are the least likely group to have a preterm baby (6.24% with P-Value<0.001), and the likelihood increases for any degree more or less than that. The graph also shows an important insight about the interpretation of missing values. A missing value in the education of a father or mother carries an important information showing that the likelihood of a preterm delivery for these types of observations is the highest (6.92% with P-Value<0.001) compared to other groups. Appendix G shows the PDP of other major risk factors.



a. Preterm delivery likelihood for different values of BMI

b. Preterm delivery likelihood for different levels of education

Figure 3 Partial dependence plot for BMI and parent's education

4 Discussion

In this study, we deployed statistical and machine learning techniques to first build a predictive model and then extract the risk factors of preterm birth (PTB) that are present during the early stages of pregnancy. This study is novel in that the application of ML techniques to a large cohort increases the generalizability of the risk factors. We included both nulliparous and multiparous mothers, spontaneous and indicated preterm birth, but excluded multifetal pregnancies that also increase the generalizability of our PTB prediction model. We reported the variable importance and partial dependence plots for the first time in the study of PTB.

The reported metrics indicate that our best GBM model improves the performance of preterm prediction compared to the similar works that combined maternal characteristics with important biological markers like serum analytes (19, 34). One of the major findings of this study is that the importance of race in predicting preterm birth can be explained when both individual risk factors such as interval since live birth, education of parents, and whether the person received adequate care during pregnancy, and their interactions are added to the model. This analytical finding is consistent with the theory of Lifecourse for addressing the racial disparities in the preterm birth outcomes (36-38). The theory of Lifecourse emphasizes the socioeconomic factors as the main determinants of health that can result in a positive shift in the long-term individual's health trajectory.

Hypertension is the most important predictor of preterm birth in a large cohort study, where 14.91 percent of the AUC improvement is attributed to this variable. The relative importance of hypertension is partly because of the deliveries that are scheduled preterm to prevent further complications in the pregnancy, especially when the placenta is not providing enough nutrients and oxygen to the baby (39). The other important finding of this paper is that history of a previous PTB is not the most important variable in the prediction of PTB and it can only explain 5.63 percent of the AUC. This finding can be explained in two ways. First, a history of preterm is useful only when the mother had a previous pregnancy. Second, the frequency of hypertension among the preterm population is almost two times the population of those with a history of a PTB in singleton pregnancies. In 2016, the number of singleton pregnancies that resulted in preterm birth was 290,584. Among this population, 56,768 were hypertension positive, while a much smaller group—28,501—had a history of PTB.

The results of our GBM model agree with the findings of previous studies. The variables like hypertension ("hyper"), interval since last live birth ("interval"), and history of PTB ("Previous_preterm") are among the most important predictors of a preterm birth, which is consistent with past studies (7, 27). The variable importance plot (VIP) reveals a novel and insightful finding compared to the previous studies. While the plot shows that the variables like a previous preterm are important predictors for preterm birth, it attributes larger relative importance to factors like hypertension or interval since last live birth in the prediction of preterm in the general obstetric population. This new finding can be explained by the limitation of traditional studies. Logistic regression models have no direct way to provide variable importance plots. This capability of the DTs provides insights about the variables that can explain a larger portion of the AUC. The new hierarchy of the important variables in the prediction of PTB can address a gap in literature where already known risk factors cannot predict many actual preterm deliveries.

This study contributes to the literature in several ways. First, the results are generalizable to the US population. Past studies lacked generalizability for different reasons (19, 20). For example, some studies used a majority White population or their sample was from one geographical location to assess the PTB risk factors (19, 21). A major strength of this study was the application of data science on a population-based linked singleton births in the U.S. to address this gap. However, using the U.S. birth dataset had its own challenges like the existence of anomalous observations and random errors. To mitigate this problem, we applied one of the advanced machine learning techniques, auto-encoders with deep neural nets, to perform data cleaning and preparation. This study also contributes to the literature of preterm birth study by providing important insights by using advanced visualization techniques. The initial visualization of variables like mother's age versus gestational age (see Appendix C) shows a clear relationship between these two variables in which the risk of a preterm delivery is the highest at the extremes of maternal age. These findings match the results of multiple other in-depth analyses (15, 40). Partial dependence plots (PDP) are the other insightful tool that we used in this analysis. The PDPs like mother's BMI in Figure 3 shows that the extremes of pre-pregnancy BMI is associated with increased rates of PTB, which is compatible with the finding of other studies (27, 41). The PDP provides a better estimation of this association compared to previous studies (42), because it takes the (average) interdependent effect of other variables into account.

There is still significant room for improving the precision of preterm birth in large cohort studies. Positive predictive value (precision) of the past studies varied between 17 to 30 percent depending on the sample used in the analysis (26, 43). Our model shows a maximum precision of 28.13% in a national-level dataset, which approaches the best practices of similar studies. However, this metric is still relatively low. This low precision is due to the lack of knowledge regarding the cause(s) of PTB and the absence of important predictors of preterm birth (e.g., cervical length) in the CDC dataset (26). Our study was subject to other limitations. Despite using the obstetric estimation for categorization of the PTB, there remains potential for errors (44). However, we used large samples and multifold cross-validations that minimize the effect of the incorrect categorization. Also, some of the biomarkers like cervical length or fetal fibronectin that are routinely measured in the obstetrical screenings were unavailable in the U.S. linked birth datasets. The association of these biomarkers and their interactions on the likelihood of a PTB can be assessed in future research.

List of Abbreviations

BO = Bayesian Optimization
CDC = Center for Disease Control and Prevention
GBM = Gradient Boosting Machines
IMR = Infant Mortality Rate
LightGBM = Light Gradient Boosting Machines
ML = Machine Learning
PTB = Preterm Birth
PBR = Preterm Birth Rate

5 References

1. Blencowe H, Cousens S, Chou D, Oestergaard M, Say L, Moller A-B, et al. Born too soon: the global epidemiology of 15 million preterm births. 2013;10(1):S2.

2. Mathews TJ, MacDorman MF, ME T. Infant mortality statistics from the 2013 period linked birth/infant death data set. National vital statistics reports. Hyattsville, MD: National Center for Health Statistics2015.

3. Butler AS, Behrman RE. Preterm birth: causes, consequences, and prevention: National Academies Press; 2007.

4. Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. The Lancet. 2008;371(9608):261-9.

5. Iams JD. Identification of candidates for progesterone: why, who, how, and when? Obstetrics and gynecology. 2014;123(6):1317-26.

6. Katz KS, Blake SM, Milligan RA, Sharps PW, White DB, Rodan MF, et al. The design, implementation and acceptability of an integrated intervention to address multiple behavioral and psychosocial risk factors among pregnant African American women. BMC pregnancy and childbirth. 2008;8(1):1-22.

7. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. The lancet. 2008;371(9606):75-84.

8. Singh N, Bonney E, McElrath T, Lamont RF, Shennan A, Gibbons D, et al. Prevention of preterm birth: Proactive and reactive clinical practice-are we on the right track? Placenta. 2020.

9. Boots AB, Sanchez-Ramos L, Bowers DM, Kaunitz AM, Zamora J, Schlattmann P. The short-term prediction of preterm birth: a systematic review and diagnostic metaanalysis. American journal of obstetrics and gynecology. 2014;210(1):54. e1-. e10.

10. Davey MA, Watson L, Rayner JA, Rowlands S. Risk‐scoring systems for predicting preterm birth with the aim of reducing associated adverse outcomes. Cochrane Database of Systematic Reviews. 2015(10).

11. Bhattacharya S, Raja EA, Mirazo ER, Campbell DM, Lee AJ, Norman JE, et al. Inherited predisposition to spontaneous preterm delivery. Obstetrics & Gynecology. 2010;115(6):1125-33.

12. Laughon SK, Albert PS, Leishear K, Mendola P. The NICHD Consecutive Pregnancies Study: recurrent preterm delivery by subtype. American journal of obstetrics and gynecology. 2014;210(2):131. e1-. e8.

13. Webb DA, Mathew L, Culhane JF. Lessons learned from the Philadelphia Collaborative Preterm Prevention Project: the prevalence of risk factors and program participation rates among women in the intervention group. BMC pregnancy and childbirth. 2014;14(1):1-10.

14. Martin JA, Hamilton BE, Osterman MJ, Driscoll AK, Drake P. Births: Final data for 2016. 2018.

15. Fuchs F, Monet B, Ducruet T, Chaillet N, Audibert F. Effect of maternal age on the risk of preterm birth: A large cohort study. PloS one. 2018;13(1):e0191002.

16. Newman R, Goldenberg R, Iams J, Meis P, Mercer B, Moawad A, et al. Preterm prediction study: comparison of the cervical score and Bishop score for prediction of spontaneous preterm delivery. Obstetrics and gynecology. 2008;112(3):508.

17. Magee L, Von Dadelszen P, Chan S, Gafni A, Gruslin A, Helewa M, et al. The control of hypertension in pregnancy study pilot trial. BJOG: An International Journal of Obstetrics & Gynaecology. 2007;114(6):770-e20.

18. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001:1189-232.

19. Alleman BW, Smith AR, Byers HM, Bedell B, Ryckman KK, Murray JC, et al. A proposed method to predict preterm birth using clinical data, standard maternal serum screening, and cholesterol. American Journal of Obstetrics & Gynecology. 2013;208(6):472. e1-. e11.

20. Weber A, Darmstadt GL, Gruber S, Foeller ME, Carmichael SL, Stevenson DK, et al. Application of machine-learning to predict early spontaneous preterm birth among nulliparous non-Hispanic black and white women. Annals of epidemiology. 2018;28(11):783-9. e1.

21. Gao C, Osmundson S, Edwards DRV, Jackson GP, Malin BA, Chen Y. Deep learning predicts extreme preterm birth from electronic health records. Journal of biomedical informatics. 2019;100:103334.

22. Goodwin L, VanDyne M, Lin S, Talbert S. Data mining issues and opportunities for building nursing knowledge. Journal of biomedical informatics. 2003;36(4-5):379-88.

23. Woolery LK, Grzymala-Busse J. Machine learning for an expert system to predict preterm birth risk. Journal of the American Medical Informatics Association. 1994;1(6):439-46.

24. Chen H-Y, Chuang C-H, Yang Y-J, Wu T-P. Exploring the risk factors of preterm birth using data mining. Expert systems with applications. 2011;38(5):5384-7.

25. Nelson DB, McIntire DD, McDonald J, Gard J, Turrichi P, Leveno KJ. 17-alpha Hydroxyprogesterone caproate did not reduce the rate of recurrent preterm birth in a prospective cohort study. American journal of obstetrics and gynecology. 2017;216(6):600. e1-. e9.

26. Robinson JN, Norwitz E. Preterm birth: Risk factors, interventions for risk reduction, and maternal prognosis. https://wwwuptodatecom/contents/preterm-birth-risk-factors-interventions-for-risk-reduction-and-maternal-prognosis. 2019.

27. Iams JD. Prevention of preterm parturition. New England Journal of Medicine. 2014;370(3):254-61.

28. He J-R, Ramakrishnan R, Lai Y-M, Li W-D, Zhao X, Hu Y, et al. Predictions of preterm birth from early pregnancy characteristics: born in guangzhou cohort study. Journal of clinical medicine. 2018;7(8):185.

29. Linked Birth / Infant Death Records 2007-20017 [Internet]. United States Department of Health and Human Services (US DHHS), National Center for Health Statistics (NCHS), Division of Vital Statistics (DVS). [cited March 2, 2019 5:01:19 PM]. Available from: http://wonder.cdc.gov/lbd-current.html

30. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics New York, NY, USA:; 2001.

31. Bergstra J, Yamins D, Cox DD. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. 2013.

32. Mendoza H, Klein A, Feurer M, Springenberg JT, Hutter F, editors. Towards automatically-tuned neural networks. Workshop on Automatic Machine Learning; 2016.

33. Feurer M, Hutter F. Hyperparameter optimization. Automated Machine Learning: Springer; 2019. p. 3-33.

34. Goodwin LK, Iannacchione MA, Hammond WE, Crockett P, Maher S, Schlitz K. Data mining methods find demographic predictors of preterm birth. Nursing research. 2001;50(6):340-5.

35. Meis PJ, Goldenberg RL, Mercer BM, Iams JD, Moawad AH, Miodovnik M, et al. The preterm prediction study: risk factors for indicated preterm births. American journal of obstetrics and gynecology. 1998;178(3):562-7.

36. Manuck TA, editor Racial and ethnic differences in preterm birth: a complex, multifactorial problem. Seminars in perinatology; 2017: Elsevier.

37. Lu MC, Kotelchuck M, Hogan V, Jones L, Wright K, Halfon N. Closing the Black-White gap in birth outcomes: a life-course approach. Ethnicity & disease. 2010;20(1 Suppl 2):S2-62-76.

38. Wadhwa PD, Entringer S, Buss C, Lu MC. The contribution of maternal stress to preterm birth: issues and considerations. Clinics in perinatology. 2011;38(3):351-84.

39. Krishna U, Bhalerao S. Placental insufficiency and fetal growth restriction. J Obstet Gynaecol India. 2011;61(5):505-11.

40. Fraser AM, Brockert JE, Ward RH. Association of young maternal age with adverse reproductive outcomes. New England journal of medicine. 1995;332(17):1113-8.

41. Hendler I, Goldenberg RL, Mercer BM, Iams JD, Meis PJ, Moawad AH, et al. The Preterm Prediction Study: association between maternal body mass index and spontaneous and indicated preterm birth. American journal of obstetrics and gynecology. 2005;192(3):882-6.

42. Honest H, Bachmann LM, Ngai C, Gupta JK, Kleijnen J, Khan KS. The accuracy of maternal anthropometry measurements as predictor for spontaneous preterm birth—a systematic review. European Journal of Obstetrics & Gynecology and reproductive biology. 2005;119(1):11-20.

43. Meertens LJ, van Montfort P, Scheepers HC, van Kuijk SM, Aardenburg R, Langenveld J, et al. Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation. Acta obstetricia et gynecologica Scandinavica. 2018;97(8):907-20.

44. Martin JA, Osterman M, Kirmeyer S, Gregory E. Measuring gestational age in vital statistics data: transitioning to the obstetric estimate. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2015;64(5):1-20.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets analyzed during the current study are available in the three different repositories.

1. The first dataset, 2016 Period Linked Birth-Infant Death Data Files, can be accessed via this link. To obtain the same dataset with geographical identifiers, researchers should submit a formal request to the National Center for Health Statistics. The files are in the plain text format. Due to the large size, data dictionaries should be used to read these files. These dictionaries can be found on National Bureau of Economic Research website.
2. The second dataset, 2016 CHR CSV Analytic Data, County Health Ranking Data, is publicly available via this link.
3. The third dataset, Area Health Resources file, is publicly available to researchers via this link. The historical data can also be accessed by sending an email to arf@qrs-inc.com.

The codes for preparing the data files are uploaded on my personal GitHub. The processed files for the second and third datasets are also uploaded in the same repository.

Competing interests

The authors declare that they have no competing interests

Funding
Support for this study was provided solely via institutional and/or departmental sources.

Authors' contributions

AE: Acquisition of the data, Conception and design of the study, Analysis of the data, Implementation of the code, Interpretation of the findings, Writing the original draft, Participating in discussions of the results

NH: Conception and design of the study, Editing the manuscript, Interpretation of the findings, Participating in discussions of the results

ZJK: Conception and design of the study, Interpretation of the findings, Participating in discussions of the results.

All authors read and approved the final manuscript.

Acknowledgements

Not applicable.