INTERNATIONAL JOURNAL OF BIOLOGY AND BIOMEDICAL ENGINEERING
DOI: 10.46300/91011.2021.15.17 Volume 15, 2021

Nomogram to Early Screen Multiparous Women for Preterm Birth in a Cohort Study

Mayssa A. Traboulsi, Zainab. El Alaoui Talibi, Abdellatif Boussaid

Laboratory of Biotechnology and Molecular Engineering, Faculty of Sciences and Techniques, Cadi Ayyad University, city zip: 40000, Marrakech.
Morocco.

Received: October 24, 2020. Revised: April 20, 2021. Accepted: May 12, 2021. Published: May 17, 2021.

Abstract—Preterm Birth (PTB) can negatively affect the health of mothers as well as infants. Prediction of this gynecological complication remains difficult especially in Middle and Low-Income countries because of limited access to specific tests and data collection scarcity. Machine learning methods have been used to predict PTB but the low prevalence of this pregnancy complication led to rather low prediction values. The objective of this study was to produce a nomogram based on improved prediction for low prevalence PTB using up sampling and lasso penalized regression. We used data from a cohort study in Northern Lebanon of 922 multiparous presenting a PTB prevalence of 8%. We analyzed the personal, demographic, and health indicators available for this group of women. The improved Positive Predictive Value for PTB reached around 88%. The regression coefficients of the 6 selected variables (Pre-hemorrhage, Social status, Residence, Age, BMI, and Weight gain) were used to create a nomogram to screen multiparous women for PTB risk. The nomogram based on readily available indicators for multiparous women reasonably predicted most of the at PTB risk women. The physicians can use this tool to screen for women at high risk for spontaneous preterm birth to improve medical surveillance that can reduce PTB incidence.

Keywords—Preterm Birth, Nomogram, Multiparous, Logistic regression, up-sampling.

I. INTRODUCTION

Although preterm birth (PTB) prevalence varies widely among countries, it is generally estimated to be between 3 and 13% of total pregnancies [1, 2]. PTB is also among the leading causes of morbidity and mortality for under 5-year-old infants, particularly, in Asian and African countries with an important number of low to middle-income households [3].

Therapies such as corticosteroid administration, cervical cerclage, treatment with vaginal progesterone have been applied effectively for women with high risk of PTB [4]. However, screening for PTB remains difficult in the absence of specific tests that would identify potential mothers at high risk of preterm birth. Although, the cervical length and cervico-vaginal fetal fibronectin measurements have been used with some success [5]. In addition, recent meta-analyses show there is no effective risk scoring system for prediction of PTB [6]. Hence, most of the prediction studies have used maternal factors that were associated with PTB.

These maternal factors include non-modifiable parameters such as the history of PTB, extremes in maternal age (<19 and >35 years) [1], multiple pregnancies, short cervical length, uterine abnormalities, and genetic factors [7] along with modifiable parameters. Modifiable factors can be related to nutrition, socioeconomic status, low body mass index (BMI), obesity, poor pregnancy weight gain, smoking, substance abuse, short inter-pregnancy interval, periodontal disease, bacterial vaginosis, late or no prenatal care, untreated antenatal depression, and the use of assisted reproductive technologies [3]. Cohort studies based on these criteria were used to develop models that predict preterm birth [8].

The models range from traditional logistic regression to identify the risk factors and estimate odds ratios to more recent machine learning algorithms including neural networks [9]. Although neural networks algorithms have been shown to lead to very high preterm prediction results, it is difficult to develop a simple version that can be used by physicians. In contrast, the logistic regression model linear coefficients have been used in nomograms and spreadsheets to deliver prediction tools that can be used by all physicians [10]. However, the prediction for PTB were generally low not exceeding 51.5 % [11].

In this work, we report an improvement of the PTB prediction for multiparous women reaching up to 88% using logistic regression models trained on resampled datasets (Up Sampling) to mitigate the problem of the low prevalence of preterm birth. We also used logistic regression regularized models with LASSO (Least Absolute Shrinkage and Selection Operator) to help analyze and select the different covariates for the best possible preterm risk evaluation. These methods have been proven successful in financial studies [12] and [13].

The main objective of this project was to develop a valid and easy to use, tool for physicians to screen among non-nulliparous pregnant women for preterm birth risk based on the data routinely collected such as medical history, demographic, and weight parameters.

II. MATERIALS AND METHODS

Pre-Term Birth (PTB) was defined as babies born alive before 37 weeks of pregnancy are completed. Only spontaneous preterm was considered in this study.

Source of data: Data were obtained from the medical records in five hospitals in North-Lebanon (private and public Islamic hospitals, Sayyidet Zgharta hospital, governmental hospital of Akkar, and governmental hospital of Tripoli). In addition to the aforementioned collection of data from medical records, we also collected data directly from 688 women under the supervision of local gynecologists.

Outcome: The objective was to develop a model that can be used to predict spontaneous preterm risk for multiparous women but also be able to be expressed in the form of a nomogram easy to use for physicians.

Predictors: The cohort study included binary responses to 14 variables. The positive class for each predictor corresponded to: 25-35 years (vs lower than 25) for Age, Obese BMI (vs Normal), University degree (vs lower degree) for Education-husband, University degree (vs lower degree) for Education-mom, presence in last pregnancy (vs absence in last pregnancy) for Pre-Cesarean, presence in last pregnancy (vs absence in last pregnancy) for Pre-Diabetes presence in last pregnancy (vs absence in last pregnancy) for Pre-Hemorrhage presence in last pregnancy (vs absence in last pregnancy) for Pre-Induction, city (vs village) Residence, presence in last pregnancy (vs absence in last pregnancy) of spontaneous preterm, smoker (vs nonsmoker) for smoking, high Social-status (vs low: income lower than 1500$), excess (vs normal) Weight-gain, external job (vs no job) for Work-husband, and similarly for Work-mom.

The Body Mass Index (BMI) of each woman was calculated using the formula: Weight (kg)/Height (m^2). Women were divided into obese and non-obese weight groups based on WHO guidelines [14] (BMI below or above 30). The underweight group was discarded due to a negligible number of representatives. Excess weight was based on the Institute of Preventive Medicine (IOM) guidelines as follows: normal weight women (BMI: 18.5-24.9) are recommended to gain between 11.4 and 15.9 kg during pregnancy, overweight women (BMI: 25.0-29.9) between 6.8 and 11.4 kg and obese women (BMI: $\geq$ 30) between 5.0 and 9.0 kg.

Missing data: There were no missing data because samples with incomplete data, women aged under 17 or above 35 or suspected to have fetuses with congenital malformation were discarded from the study. Women under 20 and those above 35 are at a greater risk for pregnancy complications [15, 16].

Sample size: The data used in this work were part of a program to evaluate pregnancy fetal complications in Northern Lebanon. The number of multiparous women were 922 among 1996 that gave birth between January 2014 and January 2016. We divided the multiparous data into two files. The first called test data corresponding to 65% of the data (600 profiles) randomly withdrawn from the total multiparous women. The remaining 35% of the data (322 profiles) constituted the second file called validation set.

Statistical analysis methods: All the predictors were coded as binary variables. The first model (glm) used was a logistic regression using the test data file. The second model (glmup) was also a logistic regression model using a new file generated from the test data file using Up-sampling. This file called up sampled data included 1108 profiles representing 554 profiles of non-preterm women and 554 randomly generated profiles, by the up-sampling algorithm, for women with a preterm. The third model (glmnetup) was a logistic LASSO penalized regression. The final model (glmglmnetup) was a logistic regression using only the predictors selected by the LASSO penalized regression trained using the up sampled data.

All models were validated first using the test data and then using the validation data set (validation set). The models were compared in terms of statistically significant predictors along with the percentage of true positives and false positives. True positives were identified for a risk (probability) higher than 50%. We also compared the risk distribution profiles given by each model.

Chi-square test, Fisher test, and Principal Component Analysis for categorical variables were performed using SPSS. The logistic regression modeling, up-sampling, and LASSO penalization were carried out using R version 3.6.1. The Nomogram was created using the lrm package in R version 3.6.1.

III. RESULTS

The multiparous women seem to form a distinct group with a higher PTB incidence. Indeed, the projection of all the 1996 women profiles on the Principal Components Analysis (Fig. 1) revealed that the multiparous women form a separate group characterized by a relatively lower social status and a higher incidence of gynecological complications. Multiparous women also showed a higher PTB incidence (8%) that was more than double that of nulliparous women (3%).

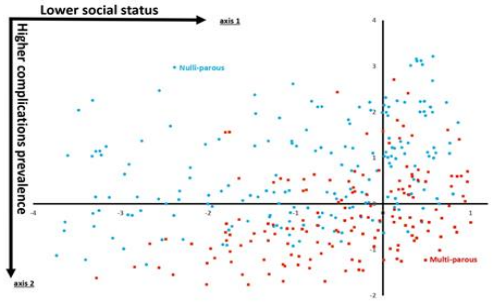

Figure 1. Projection on the first and second axes (34% of total variance) of a Principal Component Analysis for all the retrospective data showing the separation between nulliparous (blue) and multiparous women (red).

Table I gives the number and percentage of women presenting PTB. The multiparous women of the sample reached 922 among the 1996 total representing 46 % of the total retrospective data. This group of women were in majority urban, rather older working women with high education in a good income household (Table I). They have dominantly university-level education (79 %) along with their husbands (81 %). About 65 % reported having a job. Almost all the

husbands reported having a job (96 %) with a good social level (high income by 71 %). They were also dominantly in the age bracket of 25 to 35 (64%), residing in the city (76 %). About 33 % of the women had rather an obese BMI with 47 % presenting an excessive weight gain during the pregnancy. The percentage of mothers who smoked during pregnancy was 13%.

The dominant gynecological complications during past pregnancies were diabetes (5%) and Pre-eclampsia (4%). Approximately 31 % of them have had induction and 29 % hemorrhage.

Among all these factors, the covariates that presented the highest difference of percentage between the PTB positive and the negative classes were Pre-hemorrhage, Weight gain, Age, BMI, and Social status (Table I). The Chi-square test confirmed that most of these variables were statistically significant at least at the 5% level (Table I). Smaller, non-statistically significant, differences were observed for pre-diabetes, work husband, and pre-eclampsia. Pre-eclampsia and Pre-diabetes were discarded from further modeling analysis because they gave a low prevalence reaching even 0 for the positive class. Physicians watch very closely women with these complications for PTB risk. This may explain the low number of PTB incidence observed for women with these complications.

Table I. Characteristics for all multiparous women for the term and preterm classes.

Characteristic	Term (n=847)	Preterm (n=75) (Spontaneous <37 weeks)	P-value	Total (n=922) (Positive /total)
Age (25-35 years)	536(63)	58(77)	0.016	594(64)
BMI (obese)	254(30)	50(67)	0.000	304(33)
Education_husband(high)	691(81.6)	62(82.7)	0.816	753(81)
Education_mom (high)	667(78.7)	61(81.3)	0.660	728(79)
Pre_cesarean (presence)	297(35.1)	29(38.7)	0.532	326(35)
Pre_diabetes (presence)	40(4.7)	6(8)	0.438	46(5)
Pre_eclampsia (presence)	34(4)	0(0)	0.077	34(4)
Pre_hemmorrhage(presence)	215(25.4)	54(72)	0.000	269(29)
Pre_induction (presence)	263(31.1)	24(32)	0.865	287(31)
Residence(city)	644(76)	60(80)	0.438	704(76)
Smoking (smoker)	109(12.9)	11(14.7)	0.657	120(13)
Social_status (high)	614(72.5)	41(54.7)	0.010	655(71)
Weight_gain (excess)	378(44.6)	54(72)	0.000	432(47)
Work_husband (external job)	816(96.3)	74(98.7)	0.291	890(96)
Work_mom (external job)	560(66.1)	44(58.7)	0.206	604(65)

a: probability value for the Chi-square test.
(): percentage

Based on the above results we focused the remaining of this work on the multiparous women. We defined the original dataset including all the initial 922 women described by all the variables except Pre-eclampsia and Pre-diabetes.

The logistic regression analysis of the original dataset (glm) led to almost the same significant variables as the Chi-square test, except that Age was not significant while Residence was added to the list of significant co-factors (Table II).

Table II. Linear coefficients of each logistic regression model (significant at the level 5% *, 1%** and 1‰ ***)

Factors	Models[a]			
	glm	**Glmup**	**glmnetup**	**glmglmnetup**
Intercept	-4.56**	-1.39**	-3.72	-1.97***
Age1	.54	0.86***	0.33	0.68***
BMI1	1.07**	0.75***	0.35	0.70***
Education_hus1	-0.52	-0.02	.	
Education_mom1	-0.01	0.12	.	
Pre_cesarean1	-0.29	-0.52**	.	
Pre_hemmorrhage1	1.98***	2.11***	1.62	1.93***
Pre_induction1	-0.12	0.12	.	
Residence1	1.27*	1.30***	0.47	1.11***
Smoking1	0.12	0.24	.	
Social_status1	-1.42**	-1.82***	-1.04	-1.79***
Weight_gain1	1.03*	1.06***	0.76	1.07***
Work_hus1	-0.28	-0.64	.	
Work_mom1	-0.14	0.09	.	

[a]glm: logisitc regression on original data,
glmup: logisitc regression up-sample data,
glmnetup: LASSO regression on up-sample data,
glmglmnetup: Logistic regression with selected LASSO variables on up-sample data

In contrast, after creating a balanced sample using the up-sampling algorithm and running the logistic model (glmup) on these datasets, the results were notably improved for the PTB prediction as shown in Table III. Indeed, PTB prediction (PPV) increased from 12% for unbalanced to 92% for the up sampled data. Additionally, despite a small decrease the negative predictive value remained high around 75%. However, the number of misclassified non-PTB women (False Positives) significantly increased from 1 % in the unbalanced sample model (glm) to about 25%. The LASSO regularized model (glmnetup) gave comparable results. Nevertheless, the logistic regression with the selected variables by the LASSO regularization (glmglmnetup) of up-sampled data gave the best compromise between a low number of false positives (lower than 21%) and a high PTB prediction of 88% (PPV) along with a NPV of 75% (Table III).

Table III. Results of the of preterm and non-preterm (false positives) prediction (percentage) and the values of Area Under the Curve for the different models.

Models*	Value Positive Predictive	Negative Predictive Value	False Positives	Area Under the Curve
glm	12	98	1	0.841
glmup	92	71	25	0.846
glmnetup	92	72	25	0.837
glmglmnetup	88	75	21	0.840

*glm: logistic regression on original data, glmup: logistic regression up-sample data,
glmnetup: LASSO regression on up-sample data,
glmglmnetup: Logistic regression with selected LASSO variables on up-sample data.

The comparison of the distribution of the PTB risk estimated by each model to the original data (Fig. 2), showed that logistic regression before up-sampling (glm) and the Lasso model (glmnet) generally underestimate the probabilities in comparison to the other models. Even the last logistic model using the lasso selected variables slightly under-estimated those probabilities. However, both logistic regression with up-sampling before or after lasso regularization gave a closer risk or probability distribution to the original data than the other models (Fig. 2).

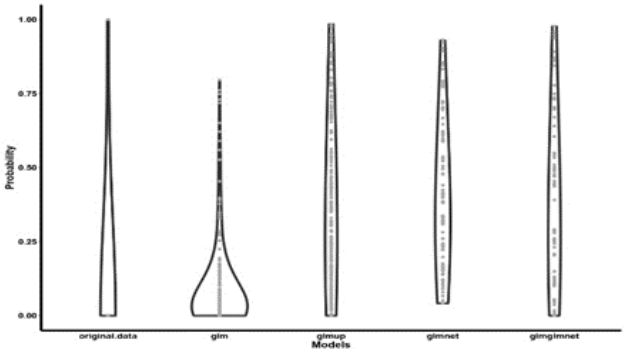

Figure2. Predicted probability distribution for each model (glm: logistic regression on original data, glmup: logistic regression up-sample data,glmnetup: LASSO regression on up-sample data and glmglmnetup: Logistic regression with selected LASSO variables on up-sample data)

.

Along with the improvement of preterm prediction the number of statistically significant covariates (at least at the level 5%) also increased from 5 for glm, to 10 in glmup but the glmnetup reduced this number to 6 (Table II). The regression model using the selected Lasso variables (glmglmnetup) was used to develop a nomogram (Fig. 3). The validation of this nomogram using the data of this study showed the possibility of having a reasonably accurate risk of PTB given the levels of Social status, Residence, Pre-hemorrhage, Age, BMI, and Weight gain for a multiparous woman.

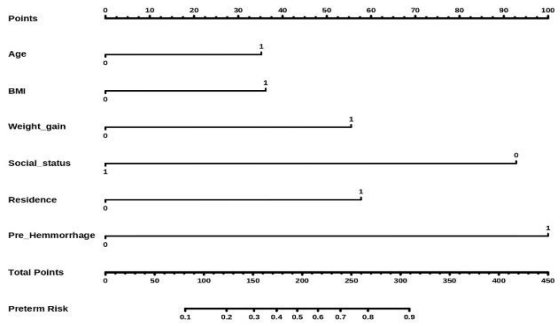

Figure 3. Nomogram for the screening of nulliparous women at risk of preterm birth.

IV. Discussion

The results of this work improved detection of high risk PTB among multiparous women using routinely collected social, demographic, and health indicators. The model that led to the best result of 88% PTB correctly predicted along with the lowest number of false positives, was used to draw a graphical nomogram that could be easily used by physicians to screen for high-risk PTB. Nevertheless, the physicians will need to put on stricter medical surveillance about 21% (at risk of PTB + false positives) of the total number of multiparous women.

In comparison Mehta-Lee et al. (2016) [10] have reported a significantly lower prediction of 51.5 % for PPV vs 88% in our study and 76.7 % for NPV versus 75% in this work. To achieve this improved level of PTB prediction, data augmentation of the initial sample through up-sampling algorithms was used. Hence, it is probable that the low PTB prediction of the logistic regression model based on the original data was at least partially due to the low prevalence of preterm birth. Furthermore, using logistic regression to predict low prevalence events may lead to meaningless outcomes [17]. Data augmentation by up-sampling randomly increases the number of positive preterm birth profiles in the newly generated dataset without changing the other class comprising women not presenting PTB [18]. This technique has been successfully used in investigations with low or very low prevalence, including some machine learning techniques such as convolutional neural networks [9].

The logistic regression model on low prevalence data clearly under-estimates the general probability [19]. A similar phenomenon was also observed for the Lasso based model, albeit with significantly smaller under-estimation. Furthermore, the regressions on up-sampled data included a higher number of significant variables to explain the model.

The number of significant variables by logistic regression almost exactly corresponded to the variables selected by Lasso regularization. However, the final model using the 6 selected variables from Lasso regularization decreased the number of false positives and hence gave the best results for PTB prediction.

The most statistically significant covariates that seem to affect PTB in this study were Social status, Pre-hemorrhage, Residence, Weight gain, BMI, and Age. Hence, it seems that the possibility of access to adequate medical care through a high income, residing in the city, and avoiding weight problems are key factors to decrease PTB incidence for this group of multiparous women. Nevertheless, urban women presented a slightly higher PTB risk. In China, a similar result has also been reported with higher PTB risk in urban areas [20] related to a higher stress and anxiety. Indicators of excess weight in terms of BMI or during pregnancy weight gain especially coupled to older pregnancy age correlated well with higher preterm risk. These last factors have been identified in other investigations [21, 22] as risk factors for PTB. It is noteworthy that besides the social status, the high incidence of hemorrhage in this group of women was relatively high. Indeed, 29 % of the multiparous women presented hemorrhage during their pregnancy. This incidence is higher even in comparison to some countries of lower national income [23] led to the highest adjusted odds ratio for PTB of 6.88 to 10.24 (95% interval).

Despite showing promising results of PTB prediction in multiparous women in Northern Lebanon, this study presents many limitations. It would be improved with a higher number of women in the sample. On top of the low number of cases, the sample was fairly homogeneous because data are better kept in hospitals treating a bigger number of high social status patients. We are hoping that this type of work will encourage health authorities to establish public databases on births in this type of low to middle-income countries. Pre-eclampsia and Diabetes were not used in the models because of the very low prevalence affecting the interpretation of the models. More variables could be added such as the number of children, stressful work, anxiety, and planned pregnancies among others. Measurements such as cervical length and cervicovaginal fetal fibronectin should be added in the screening model or at least carried out on the group of screened women by the nomogram.

V. CONCLUSION

Using readily available information from past pregnancy along with social and weight indicators, we developed a nomogram that can be used to screen for PTB risk in multiparous women. The nomogram uses the binary response to six covariates including Social status, Pre-hemorrhage, Residence, Weight gain, BMI, and Age. The nomogram could identify about 88% of the high PTB risk women.

In order to achieve a reasonably high prediction for PTB, the logistic regression was trained on a data augmented sample using up sampling. LASSO penalization helped select the final covariates in the model. These methods could improve analysis and prediction of diseases or health complications that present low or very low prevalence.

ACKNOWLEDGMENTS

The authors thank the Hospitals that helped us collect the data for this work.

References

[1] World Health Organization. “Born too soon: the global action report on preterm birth”, 2012.

[2] S. Chawanpaiboon, J.P. Vogel, A.-B. Moller, P. Lumbiganon, M. Petzold, D. Hogan, S. Landoulsi, N. Jampathong, K. Kongwattanakul, M. Laopaiboon, et al. “Global, regional, and national estimates of levels of preterm birth in 2014, a systematic review and modelling analysis”, Lancet Glob. Health 7: e37–e46. 2018. Available: https://pubmed.ncbi.nlm.nih.gov/30389451/.

[3] J. Katz, A.C. Lee, A, N. Kozuki, J.E. Lawn, S. Cousens, H. Blencowe, M. Ezzati, Z. A. Bhutta, T. Marchant, B.A.Willey, L. Adair, F. Barros, A.H. Baqui, P. Christian, W. Fawzi, R. Gonzalez, J. Humphrey, L. Huybregts, P. Kolsteren,, A. Mongkolchati,, CHERG. “Mortality risk in preterm and small-for-gestational-age infants in low-income and middle-income countries: a pooled country analysis”, Lancet (London, England), 382(9890), 417–425, 2013.
Available: https://pubmed.ncbi.nlm.nih.gov/23746775/.

[4] G. U. Eleje, J. I. Ikechebelu, A. C. Eke, P. C. Okam, I. U. Ezebialu, & C. P. Ilika, “Cervical cerclage in combination with other treatments for preventing preterm birth in singleton pregnancies”, The Cochrane Database of Systematic Reviews, (11)2017.
Available:
https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012871.pub2/full.

[5] Z. A. Oskovi Kaplan, & A. S. Ozgu Erdinc, “Prediction of Preterm Birth: Maternal Characteristics, Ultrasound Markers, and Biomarkers: An Updated Overview”, Journal of Pregnancy, 1-8, 2018.
Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6199875/.

[6] L. J. E. Meertens, P. van Montfort, H. C. J. Scheepers, S. M. J. van Kuijk, R. Aardenburg, J. Langenveld, I. M. A. van Dooren, I. M. Zwaan, M. E. A. Spaanderman, L. J. M. Smits. “Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation”, Acta Obstet.Gynecol Scand ;97(8):907-920, Epub 9, PMID: 29663314; PMCID: PMC6099449, 2018.
Available: https://pubmed.ncbi.nlm.nih.gov/29663314/.

[7] R. L. Goldenberg, J. F. Culhane, J. F. Iams, R. Romero, “Epidemiology and Causes of Preterm Birth”, Lancet 371 :75-84, 2018.
Available: https://pubmed.ncbi.nlm.nih.gov/18177778/.

[8] C. E. Kleinrouweler, F. M. Cheong-See, G. S. Collins, A. Kwee, S. Thangaratinam, K. S. Khan, B. W. Mol, E. Pajkrt, K. G. Moons, E. Schuit. “Prognostic models in obstetrics:

available, but far from applicable", Am J Obstet. Gynecol, 214(1):79-90, e36, 2016.
Available: https://pubmed.ncbi.nlm.nih.gov/26070707/
[9] T. Włodarczyk, S. Płotka, P. Rokita, N. Sochacki-Wójcicka, J. Wójcicki, M. Lipa, T. Trzciński. "Spontaneous Preterm Birth Prediction Using Convolutional Neural Networks", In: Hu Y. et al. (eds) "Medical Ultrasound, and Preterm, Perinatal and Pediatric Image Analysis", vol 12437. Springer, Cham. Lecture Notes in Computer Science ASMUS 2020, PIPPI 2020.
Available: https://link.springer.com/chapter/10.1007/978-3-030-60334-2_27.
[10] S. S. Mehta-Lee, A. Palma, P. S. Bernstein *et al.* "A Preconception Nomogram to Predict Preterm Delivery", Matern Child Health *J,* **21,** 118–127, 2017. Available: https://pubmed.ncbi.nlm.nih.gov/27461021/.
[11] B. Koullali, M. D. van Zijl, B. M. Kazemier *et al* "The association between parity and spontaneous preterm birth: a population-based study", BMC Pregnancy Childbirth, 20, 233, 2020. Available: https://pubmed.ncbi.nlm.nih.gov/32316915/.
[12] M. Chabachib, R. H. Kusmaningrum, H. Hersugondo, I. D. Pamungkas, "Financial Distress Prediction in Indonesia, WSEAS Transactions on Business and Economics", ISSN / E-ISSN: 1109-9526 / 2224-2899, Volume 16, Art. #28, pp. 251-260, 2019.
Available:
https://www.wseas.org/multimedia/journals/economics/2019/a505107-730.php.
[13] Y. Alsaawy, A. Alkhodre, M. Benaida, R. A. Khan, "A Comparative Study of Multiple Regression Analysis and Back Propagation Neural Network Approaches on Predicting Financial Strength of Banks: An Indian Perspective, WSEAS Transactions on Business and Economics", ISSN / E-ISSN: 1109-9526 / 2224-2899, Volume 17, Art. #60, pp. 627-637, 2020.
Available:
https://www.wseas.org/multimedia/journals/economics/2020/b225107-978.pdf.
[14] World Health Organization (WHO), "Global Strategy on Diet, Physical Activity and Health", Cited 2020.
[15] D. Koniak-Griffin & C. Turner-Pluta, "Health risks and psychosocial outcomes of early childbearing: a review of the literature", The Journal of perinatal & neonatal nursing, 15(2), 1-17, 2001.
Available: https://pubmed.ncbi.nlm.nih.gov/12095025/.
[16] M. Jolly, N. Sebire, J. Harris, S. Robinson, L. Regan. "The risks associated with pregnancy in women aged 35 years or older", Human Reproduction, Volume 15, Issue 11, Pages 2433–2437, 2000.
Available: https://pubmed.ncbi.nlm.nih.gov/11056148/.
[17] S. Doerken , M. Avalos, E. Lagarde, M. Schumacher, "Penalized logistic regression with low prevalence exposures beyond high dimensional settings", PLOS ONE, 14(5): e0217057,2019.
Available:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0217057.
[18] G. Cheng, S. Osmundson, D. R. Velez Edwards, G. Purcell Jackson, B. A. Malin, Y. Chen, "Deep learning predicts extreme preterm birth from electronic health records", Journal of Biomedical Informatics Volume 100, 103334, ISSN 1532-0464,2019.
Available: https://pubmed.ncbi.nlm.nih.gov/31678588/.
[19] G. Francesco, M. Niglio & M. Restaino. "A new procedure for variable selection in presence of rare events". Journal of the Operational Research Society, 2020. Available: https://www.tandfonline.com/doi/abs/10.1080/01605682.2020.1740620.
[20] L. Li, J. Ma, Y. Cheng, *et al.* "Urban–rural disparity in the relationship between ambient air pollution and preterm birth", Int J Health Geogr 19, 23 2020. Available: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-020-00218-0.
[21] S.W Masho, D.L. Bishop & M. Munn, "Pre-pregnancy BMI and weight gain: where is the tipping point for preterm birth?", BMC Pregnancy Childbirth, 13, 120, 2013. Available: https://pubmed.ncbi.nlm.nih.gov/23706121/.
[22] F. Fuchs, B. Monet, T. Ducruet., N. Chaillet & F. Audibert, "Effect of maternal age on the risk of preterm birth: A large cohort study", *PloS one* 2018, *13*(1), e0191002, 2018. Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5791955/.
[23] B. A Kebede, R. A. Abdo, A. A Anshebo & B. M Gebremariam, "Prevalence and predictors of primary postpartum hemorrhage: An implication for designing effective intervention at selected hospitals, Southern Ethiopia", *PloS one*, *14*(10), e0224579, 2019. Available: https://pubmed.ncbi.nlm.nih.gov/31671143/.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Mrs. Traboulsi Mayssa: is a Ph.D. candidate that collected the data, participated in the design and write up of this work.

Pr. Zainab E. El Alaoui- Talibi: is the Ph.D. main advisor, participated in the design and write up of this work.

Pr. Boussaid Abdellatif: is the Ph.D. co-advisor, participated in the design and write up of this work. Executed and helped in the interpretation of the statistical analyses.

Data from this study will be available on request to the corresponding author:
Mayssa A. Traboulsi, E-mail : Mayssatr@gmail.com

Sources of funding for research presented in a scientific article or scientific article itself

No special funds were used in this study.

Competing interests

The authors declare that they don't have any conflict of interest regarding the data published in this work

List of abbreviations

PTB: Pre-Term Birth
AUC: Area Under the Curve
BMI: Body Mass Index
LASSO: Least Absolute Shrinkage and Selection Operator