Early Prediction of Severe Maternal Morbidity Using Machine Learning Techniques

Conference Paper · November 2016

DOI: 10.1007/978-3-319-47955-2_22

Eugenia LUZ Arrieta
Universidad del Sinú

William Caicedo
Universidad Tecnológica de Bolívar

Juan Carlos Martinez Santos
Universidad Tecnológica de Bolívar

Early Prediction of Severe Maternal Morbidity Using Machine Learning Techniques

Eugenia Arrieta Rodríguez[1], Francisco Edna Estrada[1], William Caicedo Torres[2], and Juan Carlos Martínez Santos[2](✉)

[1] E.S.E Clínica de Maternidad Rafael Calvo, Cartagena, Colombia
sios@maternidadrafaelcalvo.gov.co
[2] Universidad Tecnológica de Bolívar, Cartagena, Colombia
jcmartinezs@unitecnologica.edu.co
http://www.maternidadrafaelcalvo.gov.co/,
http://www.unitecnologica.edu.co/

Abstract. Severe Maternal Morbidity is a public health issue. It may occur during pregnancy, delivery, or puerperium due to conditions (hypertensive disorders, hemorrhages, infections and others) that put in risk the women's or baby's life. These conditions are really difficult to detect at an early stage. In response to the above, this work proposes using several machine learning techniques, which are considered most relevant in a bio-medical setting, in order to predict the risk level for Severe Maternal Morbidity in patients during pregnancy. The population studied correspond to pregnant women receiving prenatal care and final attention at E.S.E Clínica de Maternidad Rafael Calvo in Cartagena, Colombia. This paper presents the preliminary results of an ongoing project, as well as methods and materials considered for the construction of the learning models. AQ1 AQ2

Keywords: Severe maternal morbidity · Machine learning · Logistic regression

1 Introduction

The term Severe Maternal Morbidity (SMM) includes a set of complications that can have a severe adverse effect on women and baby health, and happen during pregnancy, delivery, or puerperium. When any of these appear, it is necessary to provide the patient with immediate attention, in order to avoid death [10]. Although maternal health outcomes have shown positive variation, complications of pregnancy still are an important public health issue. Each year around 585.000 women die during pregnancy, delivery or puerperium worldwide [5], and annually close to 50 million complications in maternal health are registered, and approximately 300 million women suffer from short and long-term illnesses and injuries related to pregnancy, childbirth and postpartum [10]. Currently, there is an epidemiological surveillance strategy which consists in identifying SMM cases, reporting them to the public surveillance system (SIVIGILA) [16], and following them

M. Montes-y-Gómez et al. (Eds.): IBERAMIA 2016, LNAI 10022, pp. 1–12, 2016.
DOI: 10.1007/978-3-319-47955-2_22

up. This allows to characterize SMM and have a better understanding about the main factors of risk in the population and devise policies to help lower incidence. However, the number of SMM cases continue to be very high.

Studies conducted to identify causes of SMM show that this condition is related with hypertensive disorder, hemorrhage, and infections. The main risk factors associated with occurrence of SMM are black race, obesity, multi parity and backgrounds of previous cesarean sections and presence of co-morbidities [3,6,7,14].

The development of adequately sensitive and specific predictive tests for these outcomes has received significant focus in perinatal research. According to the literature, machine learning approaches are used frequently to identify patterns and make predictions. Specially in medicine, logistic regression [11,13], support vector machines [1], neural networks [9] have been used successfully.

The World Health Organization and PanAmerican Health Organization during the last decades have tried to reduce mortality and Severe Maternal Morbidity. For this, the action plan of 2012–2017 was proposed [3]. It consists in strengthening information systems and monitoring of maternal health in the countries of the region. The reduction of maternal mortality is a millennium goal and a national purpose. Actions, as epidemiological surveillance, the availability of statistical data, and the identification of risk factors related to these events, have contributed to its decrease.

Institutions and doctors strive to avoid SMM, because it is not easy to detect and prevent such situations. Especially, when the volume of pregnant women is quite high in a day, or when novice doctors do not have enough experience. Even with the implementation of the above actions, failure to meet the stated goal persist. Because of that it is necessary to implement new mechanisms for early warning and monitoring of SMM cases.

This paper proposes the use of machine learning techniques to build a risk classifier for SMM. With this, we expect to have early detection of morbidity cases, providing support for medical staff in decision-making, to enable a timely intervention of patients. This would help reduce the risk the mother and baby may have during this stage, and in turn to reduce social and economic repercussions.

The paper is organized as follows. Section 2 presents related work. Section 3 shows the methods and materials used in our approach. Section 4 shows the preliminary results of our on-going research. Finally, Sect. 5 states the conclusions of this paper.

2 Related Work

Every time, it is more frequent to find the use of machine learning in medicine, especially for classification problems. Some studies of prediction of at least one of the major diseases associated with SMM are mentioned below.

In Poon et al. [13], authors show work about early prediction of hypertensive disorders during first-trimester pregnancy, in the population of London, UK.

Logistic regression was used and a detection rate around 90 % for early preeclampsia was obtained, and a false-positive rate of 5 %.

In Park et al. [11], authors present an algorithm based on multiple logistic regression to predict the risk of preeclampsia in an Australian population. The algorithm correctly predicted preeclampsia in 95 % of women with a 10 % false positive rate.

In Nanda et al. [8], authors present a model for prediction of gestational diabetes mellitus in the first-trimester of pregnancy, based on bio markers and some maternal features. The use of logistic regression gave them a 74.1 % of correct predictions, with 20 % false positive rate.

In Farran et al. [4], authors implemented logistic regression, k-nearest neighbours (k-NN), multifactor dimensionality reduction and support vector machines for predicting diabetes, hypertension and comorbidity. The techniques were satisfactory implemented and similar results were obtained.

According to the reviewed literature there is evidence that the results obtained from the implementation of machine learning for classification problems in medicine are quite satisfactory. The authors of this paper have not been able to find similar studies or proposals, using machine learning as a tool to help to avoid or mitigate the risk of SSM.

3 Methods and Materials

3.1 Participants

A retrospective cohort study was done through clinical histories of prenatal controls obtained between 2014 and 2015. The population selected for this study include patients with ages between 12 and 45 years who had at least one control at E.S.E Clínica de Maternidad Rafael Calvo and whose labor was cared for in this institution.

Cohort patients were classified according SMM outcome, in two groups: patients who did not present SMM, and patients which presented SMM (which also were reported to the public healthcare surveillance system). For the first group, we used random sampling, and for the second group, we used convenience sampling. This method is known as mixed sampling [2].

3.2 Data Set

The construction of the machine learning model was based on features or risk factors. These factors were selected according to the risk factor characterization described by Latin American Center of Perinatology (Centro Latino Americano de Perinatología, CLAP), compared to the 2015 SMM protocol from the Colombia Ministry of Health and Social Protection (Ministerio de Salud y de Protección Social) [3]. The selected predictor set was then supplemented with socio demographic data and the gynecological and obstetrical history for each patient, Tables 1 and 2.

Table 1. Socio demographic characteristics

Feature	Options
Age in years	Younger than 14, between 14 and 19, between 19 and 34, older than 35
Ethnicity	Native Colombians, Gypsy, Raizal, Palenquero, Black, Mulatto, Afro-descendants, Other
Scholarship	Basic Primary, Secondary Basic, Technical, University, None
Socio economic strata	Strata 1, Strata 2, Strata 4, Strata 5, Strata 6, Unknown
Health care regulation	Contributory Regime (CR), Subsidized Regime (SR)
Origin	Capital city, Village, Rural zone
Marital status	Single, Married, Domestic partnership, Separated, Widowed

The selected population included treated patients in gynecological antenatal consultations at the ESE clinic during pregnancy, and whose labor was cared for, between years 2014 and 2015.

For the data set construction we used Google Forms. Two forms were designed, the first one to record Obstetrics Gynecology (OBGYN) and socio-demographic background data, and the second one to record diagnoses for each gestational week, using the International Classification of Diseases (ICD-10) codification [15].

The data set is being built with help of sixth year medicine students. They were trained on SMM, review of medical records and filling out Google forms. The manual review of prenatal medical records of each patient is necessary, because the information is sometimes scattered and not totally centralized on the hospital information system. To perform preliminar training and validation, two patient groups were generated according to outcomes (SMM and not SMM). First group was sub-sampled in order to reduce class imbalance, given that the number of patients that exhibited SMM is lower than the non SMM group. The collection process is still underway, and subsequent validation will be carried to monitor progress and performance of the predictive models.

3.3 Statistical Data Analysis

The filtering features allow to select the set of variables that represent variability in the occurrence of SMM. Once the variables are defined, and the database constructed, we proceeds to do an analysis statistical to obtain a database only with the information of the variables that we considered predictors for the model. We use descriptive statistics to identify the frequency with which diagnoses are presented, and multi-factor analysis of variance (ANOVA) to determine which variables are more likely to be considered predictive for the model that we want to implement. It was tested with different levels of confidence for the group of variables most likely had to influence the behavior of the response variable.

Table 2. Gynecological and obstetrical history

Feature	Options
Maternal parity	Nulliparous, Multiparous
Pregnancy spacing less than two years	Yes, No
Multiple birth	Simple, Twins, Triplets or more
Prenatal care	0, More than 0 less than 4, More than 4
Gestational age in first-prenatal care	First-trimester, Second-trimester,Third-trimester
Micronutrient intake	Yes, No
Personal history of preeclampsy	Yes, No
Pregnancy induced hypertension PIH	Yes, No
Chronic hypertension	Yes, No
Superimposed preeclampsy	Yes, No
Diabetes	Yes, No
Autoimmune disorders	Yes, No
Human immunodeficiency virus HIV	Yes, No
Congenital syphilis	Yes, No
Hepatitis B	Yes, No
Previous perinatal mortality	Yes, No
Incompatible with life VIP	Yes, No
Maternal causes VIP	Yes, No
Sexual abuse VIP	Yes, No
Urinary tract infection, UTI	Yes, No
Drinking/Smoking	Yes, No
Illicit and nonillicit drug use	Yes, No
Anemia in pregnancy	Yes, No
TORCH infections	Yes, No
Obesity in pregnancy	Yes, No
Under-nutrition during pregnancy	Yes, No

3.4 Learning Models

This section lists some of the machine learning techniques most commonly used for classification or pattern recognition.

Logistic Regression. The logistic regression has been historical an important tool for data analysis in medical investigation and epidemiology. This allows differentiating between some classes, in terms of a set of numerical variables, as a predictor. The basic goals of a logistic regression model are:

– Get an unbiased or adjusted estimate of the relationship between variable dependent (or result) and an independent variable.
– Simultaneously evaluate several factors that are allegedly related somehow (or not) with the dependent variable.
– Build a model and get a hypothesis prediction purpose or calculating risk.

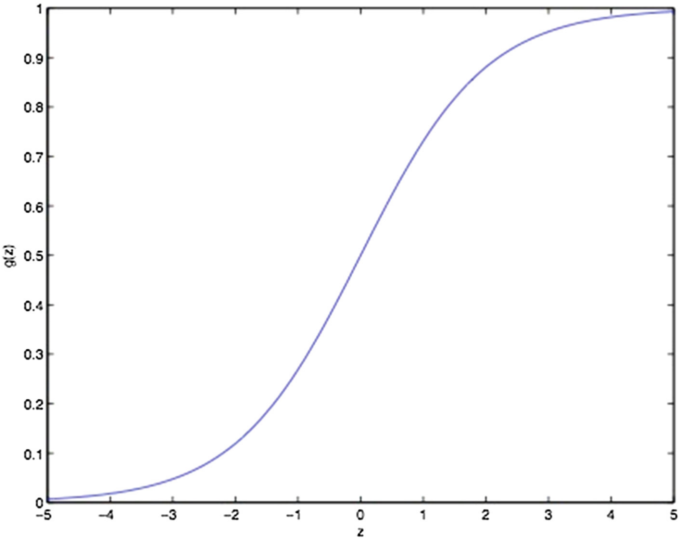

Fig. 1. Sigmoid function

Logistic Regression is a linear classification algorithm widely used in medicine where a logistic sigmoid function is coupled to a linear regression model. The Sigmoid function used to make the prediction algorithm is shown in Fig. 1.

Results can be interpreted as the probability for the input to belong to the positive class, $p(Y = 1|x; \theta)$. In this study, Logistic Regression was trained with Cross-Entropy loss. A L2 regularization penalty was introduced to account for model complexity and avoid over-fitting:

$$min \ -\frac{1}{N}\sum_{i=1}^{N}[y_i log h_\theta(x_i) + (1-y_i) log(1-h_\theta(x_i))] + \lambda||\theta||^2 \tag{1}$$

where

$$h_\theta(x) = \frac{1}{1+e^{-\theta^t x}} \tag{2}$$

It is possible to increase its capabilities by applying a polynomial transform to the input, in which case the decision boundary can be non-linear and handle more difficult problems.

In this work, we used L2 Regularized Logistic Regression with polynomial transform of 2nd degree. The model was implemented using the Scikit-Learn Python Machine Learning Library, version 2.7.11. [12]

The model was trained with Stratified K-Folds cross validation iterator. The algorithm was probe in a server in a High Performance Computing Laboratory through a batch system to run tasks (HTCondor).

4 Results

4.1 Statistical Analysis

This paper presents preliminary results of the project that is still ongoing. Achieving determine the variables set related to the occurrence of extreme maternal morbidity. The entire population is 1838 patients. 72 belong to the first group, and the remain 1766 belong to the second group. For a 95 % confidence level with a confidence interval of 4.5, we obtained a total sample of 377 patients. Once the system is trained, we will validate the obtained data in 145 patients. Initially, we performed a descriptive data frequency analysis of the patients who developed SMM.

We perform an analysis of frequency of the diagnosis. As we can see in Fig. 2, the amount of diagnostics is very large and the graph of frequency showed a data distribution dispersed and difficult to interpret. It was decided to group them according to the ICD-10 , leaving a moderate amount of variables. The new results are shown in the Table 3.

Followed by this, we did the frequency analysis by trimester. In the first trimester only the diagnostic group Z30-Z39 had high frequency. For the second trimester, the results show that the most frequent group are Z30-Z39, O30-O48, O20-O29, E65-E68, and N70-N77. In the third trimester is obtained that the diagnostic groups most frequently are Z30-Z39, O30-O48, O10-O16, N70-N77, O95-O99, D50-D64, and O60-O75.

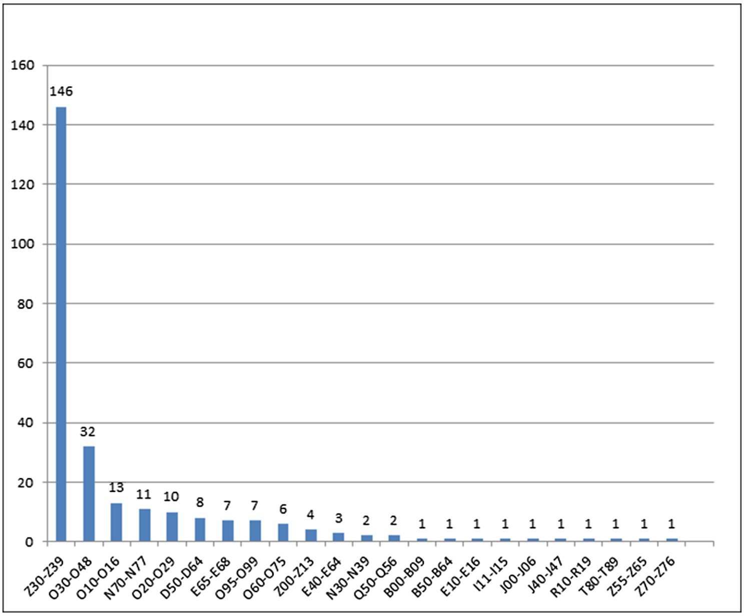

Fig. 2. Diagnostics group frequency

Table 3. Diagnostics frequency analysis

CODE (ICD-10)	Diagnostic group
Z30-Z39	Persons encountering health Services
O30-O48	Maternal care related to the fetus and amniotic cavity and possible delivery problems
O10-O16	Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
N70-N77	Inflammatory diseases of female pelvic organs
O20-O29	Other maternal disorders predom1inantly related to pregnancy
D50-D64	Nutritional anaemias
E65-E68	Obesity and other hyperalimentation
O95-O99	Other obstetric conditions, not elsewhere classified
O60-O75	Complications of labour and delivery

After having a notion of the data trend, we decided to carry out an analysis of variance ANOVA to the data of the patients with SMM and not SMM. It was organized per trimester. Similarly, the diagnostics were grouped according to SMM. The analysis of the first trimester was done initially with levels of confidence of 95 %, 90 % and finally 85 %. The results indicate that none of the variables have a high probability of being connected with the SMM. It notes that the diagnosis Z30-Z39 shows more likely to have any connection with the response variable, but it is the default diagnostic.

The analysis of the data obtained for the second trimester was tested with the 90 %, 80 % and 75 % confidence levels, and the results are shown in Table 4.

Table 4. Anova second trimester 75 % Confidence

Description	Feature	P-value
Obesity and other hyperalimentation	E65-E68	0.0661
Persons encountering health services for examination and investigation	Z00-Z13	0.2411
Persons with potential health hazards related to socioeconomic and psychosocial circumstances	Z55-Z65	0.2196

The analysis of variance of the third trimester was tested with the 95 %, 85 % and 80 % confidence levels. Results with 80 % confidence level are shown in Table 5.

Finally to make the analysis of ANOVA without taking into account the trimester was tested with confidence levels of 95 % and 80 %. Results are shown in Table 6.

Table 5. ANOVA third trimester with 80 % confidence

Description	Feature	P-value
Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium	O10-O16	0.0383
Complications of labour and delivery	O60-O75	0.0408
Obesity and other hyperalimentation	E65-E68	0.109
Acute upper respiratory infections	J00-J06	0.1232
Symptoms and signs involving the digestive system and abdomen	R10-R19	0.1769

Table 6. ANOVA with the whole pregnancy period with 80 % confidence

Description	Feature	P-value
Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium	O10-O16	0.0277
Complications of labour and delivery	O60-O75	0.0408
Symptoms and signs involving the digestive system and abdomen	R10-R19	0.0725
Acute upper respiratory infections	J00-J06	0.1391
Disorders of other endocrine glands	E20-E35	0.1505

Table 7. Anova personal history with 90 % confidence level

Feature	P-value
VIP Maternal causes	0.0009
Preeclampsy	0.0013
Scholarship	0.0227
Diabetes	0.0318
Origin	0.0422
Socio economic strata	0.0697
Perinatal mortality	0.0821
Htacronica	0.0937

In the Table 7 is observed the analysis of ANOVA for the personal history of the patients performed with a confidence level of 90 %. It identifies the factors that have the most significant statistical effect on the occurrence of the SMM.

4.2 Training

The data set was divided into training set and test set, the first corresponding to 80 % (178 instances) and the other 20 % (44 instances). The learning model was trained using 5-fold Stratified Cross-Validation. We Used L2 Regularized

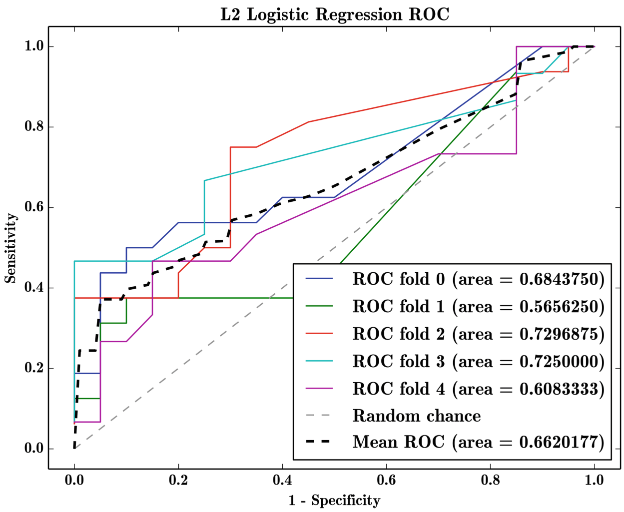

Fig. 3. SMM prediction model ROC

Logistic Regression to control the model complexity. To attack the problem of the disparity of the classes is it penalizes classifier 10 times for each error in the positive kind. The process of training takes 20 min in a server using one processor core. Figure 3 show the result for a Logistic Regression classifier with the selected predictors. The subsets are represented by a Receiver Operation Characteristic Score Area Under the Curve (ROC AUC). It is a graphical representation of the sensitivity vs. (1 - specificity).

To interpret the curve ROC is recommended to analyze the following intervals: $[0.5 - 0.6]$ Bad test, $[0.6 - 0.75]$ Regular test $[0.75 - 0.9]$ Good test, $[0.9 - 0.97]$ Very Good test, $[0.97 - 1]$ Excellent test.

In the Fig. 3, the ROC shows that the average area under the curve is 0.66. Then, the results are considerate Regular. To improve it, we suggest to try the following options:

1. Take the original data set and to use Recursive Feature Elimination for logistic regression, in this way obtain the best predictors to the model.
2. To Prove with polynomial transform higher than 2nd degree.
3. If the above options fail, we need to try a vector support machine.
4. After, we need to compare the results of ROC graphics for polynomial transform and support vector machine and select the model that shows the best especificity and sensibility.

5 Conclusions

Severe maternal morbidity remains a public health-care problem that affects much the pregnant women population, and in many cases, it is possible to avoid this. The problem lies in the early identification of risk patients who have finished in SMM. In response to the above mentioned, this paper presents the usage of the logistic regression for SMM detection. It is a pattern recognition technique commonly used in the medical field to solve problems of classification, prediction and identification of patterns. By the using of this technique, it is expected to build a tool for risk identification or risk classification of a patient having SMM. The goal is to provide a timely and adequate attention to each patient depending on the risk level to be determined.

With the implemented logistic regression model, we obtained regular results, for this, we to continue to proving others techniques of machine learning from to obtain a model with best results.

Acknowledgements. Special thanks for their cooperation to the High-Performance Computing Laboratory (HPCLab) at Universidad Tecnológica de Bolívar and to research group on maternal safety of Center of research for maternal health, Perinatal and women at E.S.E Clińica de Maternidad Rafael Calvo.

References

1. Carty, D.M., Siwy, J., Brennand, J.E., Zürbig, P., Mullen, W., Franke, J., McCulloch, J.W., North, R.A., Chappell, L.C., Mischak, H., et al.: Urinary proteomics for prediction of preeclampsia. Hypertension **57**(3), 561–569 (2011)
2. Casal, J., Mateu, E.: Tipos de muestreo. Rev. Epidem. Med. Prev **1**(1), 3–7 (2003) AQ3
3. Duran, M.E.M., García, O.E.P., CArey, A.C., Bonilla, H.Q., Espitia, N.C.C., Barros, E.C.: Protocolo de vigilancia en salud pública morbilidad materna extrema
4. Farran, B., Channanath, A.M., Behbehani, K., Thanaraj, T.A.: Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from kuwaita cohort study. BMJ Open **3**(5), e002457 (2013)
5. Haaga, J.G., Wasserheit, J.N., Tsui, A.O., et al.: Reproductive Health in Developing Countries: Expanding Dimensions, Building Solutions. National Academies Press, Washington, D.C. (1997)
6. Mariño Martínez, C.A., Fiesco, V., Carolina, D., et al.: Caracterización de la morbilidad materna extrema en el Instituto Materno Infantil-Hospital la Victoria/Characterization of extreme morbidity disease in the Instituto Materno Infantil-Hospital la Victoria. Ph.D. thesis, Universidad Nacional de Colombia
7. Morales-Osorno, B., Martínez, D.M., Cifuentes-Borrero, R.: Extreme maternal morbidity in Clinica Rafael Uribe Uribe, Cali, Colombia, from January 2003 to May 2006. Revista Colombiana de Obstetricia y Ginecología **58**(3), 184–188 (2007)
8. Nanda, S., Savvidou, M., Syngelaki, A., Akolekar, R., Nicolaides, K.H.: Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. Prenat. Diagn. **31**(2), 135–141 (2011)
9. Neocleous, C.K., Anastasopoulos, P., Nikolaides, K.H., Schizas, C.N., Neokleous, K.C.: Neural networks to estimate the risk for preeclampsia occurrence. In: International Joint Conference on Neural Networks, IJCNN 2009, pp. 2221–2225. IEEE (2009)
10. Organization, W.H., UNICEF.: Revised 1990 estimates of maternal mortality: a new approach. World Health Organization (1996)

11. Park, F.J., Leung, C.H., Poon, L.C., Williams, P.F., Rothwell, S.J., Hyett, J.A.: Clinical evaluation of a first trimester algorithm predicting the risk of hypertensive disease of pregnancy. Aust. N. Z. J. Obstet. Gynaecol. **53**(6), 532–539 (2013)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
13. Poon, L.C., Kametas, N.A., Maiz, N., Akolekar, R., Nicolaides, K.H.: First-trimester prediction of hypertensive disorders in pregnancy. Hypertension **53**(5), 812–818 (2009)
14. Rojas, J.A., Cogollo, M., Miranda, J.E., Ramos, E.C., Fernández, J.C., Bello, A.M.: Morbilidad materna extrema en cuidados intensivos obstétricos. Cartagena (Colombia) 2006–2008 maternal near miss in obstetric critical care. Cartagena, Colombia, 2006–2008. Revista Colombiana de Obstetricia y Ginecología **62**(2), 131–140 (2011)
15. de la Salud, O.P.: Clasificación estadística internacional de enfermedades y problemas relacionados con la salud: décima revisión: CIE-10. Pan American Health Org (1995)
16. de Vigilancia, S.: Control en salud pública (sivigila). Informe de Intoxicaciones por plaguicidas. Instituto Nacional de Salud, INS. Bogotá, Colombia (2012)