SYSTEMATIC REVIEW

A Comparative Review of Data Mining Techniques for Prediction of Risk Factors of Low Birth Weight

TAHIRA ASHRAF[1,2], ASIF HANIF[1,3], NYI NYI NAING[4], NADIAH WAN-ARFAH[5]
[1]Ph.D. Scholar, Biostatistics, Faculty of Medicine, Universiti Sultan Zainal Abidin, Medical Campus, Kuala Terengganu, Malaysia
[2]Assistant Professor: University Institute of Radiological Sciences & Medical Imaging Technology, Faculty of Allied Health Sciences, The University of Lahore
[3]Associate Prof. Biostatistics: University Institute of Public Health, Faculty of Allied Health Sciences, The university of Lahore
[4]Professor, Faculty of Medicine, Universiti Sultan Zainal Abidin, Medical Campus, Kuala Terengganu, Malaysia
[5]Faculty of Health Sciences, Universiti Sultan Zainal Abidin, Gong Badak Campus, Kuala Terengganu, Malaysia
Correspondence to Prof. Nyi Nyi Naing, Email: syedhatim@unisza.edu.my, Phone: +609-6688760, Fax: +609-6275771

ABSTRACT

Background: Low Birth Weight is a serious public health issue and has major contribution in neonatal morbidity and mortality worldwide. Logistic regression (LR) has been conventionally used to predict low birth weight and identify its risk factors. However, latest data mining techniques like Artificial Neural Network (ANN) have not been used much for this purpose.
Aim: To review the predictive ability of two data mining techniques (Artificial Neural Network and Logistic Regression) for prediction of risk factors of Low Birth Weight.
Methods: All studies that compared predictive ability of ANN and LR for risk factors of LBW were searched on Google scholar, PubMed, Cochran library and web of science using BOOLEAN search strategy and 6 studies following PRISMA guidelines were included. Studies were stored on ENDNOTE version 7 and were critically analyzed. Any disagreements were handled with consensus.
Results: Studies ranged from 1999 to 2019 and all the studies were retrospective cohort. Total of 3,293 subjects were included in all 6 studies. Commonly compared statistical tests were AUC, sensitivity, specificity, negative predictive value, positive predictive value, concordance index, F-statistics, precision and recall. Almost all studies reported that ANN performed better against all these statistical tests or atleast equal in prediction of risk factors of low birth weight.
Conclusion: ANN is a reliable, powerful, and sophisticated tool for handling complex data with high accuracy. ANN can be advantageous over LR specially if considerable inter and intra-relationships of outcome with risk factors and complicated non-linear relationships exist in data.
Keywords: Data mining, Artificial Neural Network, Logistic Regression, Fetal Weight, Low Birth Weight, Pregnancy

INTRODUCTION

Birth weight in normal range is crucial for ensuring healthy delivery and lesser chances of complications after birth[1]. Low Birth Weight (LBW) is a major public health issue that increases the chances of many physical as well as neurodevelopmental disorders for newborns such as mental retardation, hypothermia and hypoglycemia[2]. According to World Health Organization (WHO), the global prevalence of LBW is 15.5% whereas, almost 96.5% LBW births occur in developing countries[3]. Moreover, LBW is responsible for 60% infant mortality in first year of life and LBW infants have 40% increased risk of death in first few months of their lives compared to Normal Weighted Births (NWBs)[4].

With advancements in technology and science, the statistical tools for predicting low birth weight and its risk factors have also become more powerful and sensitive. Hospitals and healthcare centers are focusing on adding large amount of clinical data for beneficial analysis that can lead to huge contribution in health sector[5]. Recently, data mining approaches have become quite prevalent for managing the enormous amount of data and extract valuable patterns, knowledge, and predict the status of a particular disease or outcome in patients[6]. Moreover, the data mining techniques have an important role in treating complex interactions of patients with their disease, treatment options and other conditions[7].

There are two main objectives fulfilled by data mining, one presentation and the other prediction. Different techniques of data mining constitute one or both parts of these depending upon the situation and spectrum of data.[8] The major tasks involved in this process include summarization, association, stratification or classification, clustering, and trend analysis. A number of techniques serve this purpose in healthcare such as regression analysis, decision tree, Artificial Neural Network (ANNs), and Support Vector Machine (SVM)[5]. Regression analysis is considered as one of the very first techniques being used for prediction of desired outcomes for many years. Now, even with advent of new applications, regression analysis is still used mostly as a gold standard to compare its effectiveness and predictive accuracy with these relatively newer data mining techniques[9,10].

The use of these data mining techniques is relatively commoner in some healthcare problems in general such as cancer and very few maternal and child health issues in particular such as preterm birth and neonatal mortality but for LBW, the studies using these data mining techniques are very limited[11]. Although few comparisons of ANN with logistic regression have reported ANN to better or at-least not worse than logistic, the consensus on the better technique for predictive accuracy of risk factors has not been established so far. Therefore, this methodological synthesis compares and reviews the predictive accuracy of logistic regression and ANN for determination of risk factors

of LBW. This review also compares the common statistical tests of the two techniques that have been compared in published literature.

The objective of the study was to review the predictive ability of two data mining techniques (Artificial Neural Network and Logistic Regression) for prediction of risk factors of Low Birth Weight.

MATERIALS AND METHODS

Reporting: Studies published in local and international journals freely available on internet were selected for this review. The results of this review were reported using Preferred Reporting Items for Systematic Review and Meta-Analysis statement (PRISMA) guidelines[12].

Inclusion and Exclusion Criteria: All type of analytical observational studies, including analytical cross sectional, case control, and retrospective cohort were included. No restriction on time duration or study period was applied. Studies that compared the predictive ability of LR and ANN for risk factors or outcomes related to low birth weight were included. Studies that compared at least one statistical test of LR with ANN were included. Studies with irrelevant title or statistical approaches were excluded. Also, incomplete, ambiguous or anonymous studies and those with only abstracts were excluded. Similarly, case studies, editorials, letters to editor, reviews and qualitative studies were excluded.

Before selection of studies a brief checklist was made to assess the title, quality, sample size, sampling methods, time and place of study. After initial selection, the studies were appraised critically on this checklist for final selection.

Search Strategy and Information Sources: Google scholar, PubMed, Web of Science and Cochrane library were accessed for searching the articles. The BOOLEAN search strategy was used to find the related studies. The terms used in phrases and/or keywords included “birth weight”, “low birth weight”, “abnormal birth weight”, “neonates”, “birth outcome”, “preterm birth”, “risk factors”, “causes”, “factors”, “Regression”, “Logistic Regression”, “Artificial neural network”, and “data mining”. Additionally, to fit advanced PubMed search, MeSH terms such as “Newborn OR neonate OR infant AND birth weight OR low birth weight OR abnormal birth weight OR underweight AND risk factors OR factors OR causes AND Logistic regression AND Artificial neural network OR data mining” and synonyms were also used.

Study Selection: In the first step, studies were retrieved in a references management software named ENDNOTE version 7 for storage and avoiding duplication. The retrieved studies were then assessed through the abovementioned checklist. Irrelevant or ambiguous studies were excluded. In the second step, two authors (AH and FZ) critically analyzed the contents of articles. Those articles that were not in line with the title, had irrelevant variables or inappropriate analysis, had statistical and methodological errors and other issues were excluded from the study. Any disagreement of the individuals was resolved by consensus.

Data Extraction: A structured data extraction form was made for the purpose of extracting information from selected studies. First author, Year, Study Design, sample size, comparative tests and values against these comparative tests for LR and ANN were components of the form. The two reviewers independently extracted the data from the articles. Any discrepancy in reported data was rechecked and corrected by a third reviewer.

RESULTS

Initially, 84 studies were found, however, after first screening, only 9 studies were found appropriate. Out of these 9 studies, 2 had irrelevant statistical tests and one study focused on methodological description of the techniques rather than comparison on data and hence these were also excluded. Therefore, for final synthesis and reporting, 6 studies were shortlisted. A Total of 3,293 subjects were included in all 6 studies.

One recent publication in April 2019 compared the predictive accuracy of LR and ANN for determining the risk factors of Low Birth Weight (LBW). There were 223 new-borns included in this study which was conducted in Istanbul, Turkey. The records about the risk factors listed were analyzed using LR and ANN. They reported that the values for AUC (SD) for LR were 0.909(0.019) and for ANN were 0.941(0.0012). Their conclusion was that despite of slightly higher values of ANN compared to LR, the difference was not considerable enough, which means that both LR and ANN have equal potential for helping clinicians understand the risk factors of LBW and make clinical decisions accordingly[13].

One study in 2015 compared the predictive ability of LR with five other techniques of data mining including neural network in order to find the most impactful risk factors of low birth weight (LBW). The statistical measures of comparison used in this study were on specificity,sensitivity, F-statistics, accuracy, AUC, recall and the precision. For logistic regression, these values were 0.3390, 0.9231, 0.8304, 0.7407, 0.7724, 0.9231 and 0.7547 and for ANN these values were 0.3729, 0.9385, 0.8443, 0.7619, 0.7804, 0.9385 and 0.7673. The most impactful variables found in this study for prediction of the LBW included maternal weight before conceiving in pounds, maternal age, history and frequency of previous premature births and frequency of antenatal visits in the first three months of pregnancy. They concluded that almost all statistical tools of ANN gave better results in prediction of risk factors of LBW compared to LR.[14]

Another study aimed to compare ANNs and LR for prediction of clinical outcome among extremely low birth weight neonates. In this study, set of 23 variables were selected and studied among 810 extremely low birth weighted born babies. They were later divided in three sets of training (502), test set (249) and validation (59) in a random order. LR was applied in forward step-wise direction on the entire set in order to find any variables that were significant. They reported significant risk factors as baby’s weight at birth, ethnicity, age at gestation, the Apgar score taken at 5min, intake of any steroids, multiple babies, and any respiratory disorders. Both ANN and LR models were subsequently implemented using training, validation and then test sets on the significant variables first and afterwards, excluding one variable at a time. They reported that values for AUC were similar for both ANN and LR

using this data i.e., p<0.3. For both the LR and ANN, the result for AUC was found to be better for only significant risk factors compared to the AUC calculated for complete dataset (p < 0.005). Most significant among all the studied risk factors were weight at birth, age of gestation and Apgar score taken at 5min. Moreover, the values for PPV, NPV and specificity were same for both LR and ANN at sensitivity of 80% reported as 72%, 90% and 85% respectively[15].

Another study predicted developmental diseases and complications including low birth weight at infancy using an ANN model. In order to develop this model, hundreds of variables lying under the domains of socio-demographic, maternal, clinical, and newborn related risk factors were studied among 1,232 newborns. The outcomes studied under the domain of newborn related factors also included the LBW. The statistical tool compared in this study was concordance index and was reported as 83.1% for ANN compared to 79.5% for LR. Also, the values for ROC were reported as 0.79 for ANN and 0.68 for LR, whereas sensitivity and specificity of ANN compared to LR were 93.2% versus 92.7% and 39.1% versus 21.7% respectively. They also concluded that the predictive ability of ANN was relatively much better compared to LR for identification of risk factors of LBW[16].

Another study also compared logistic regression and ANN for predicting outcome of ELBW neonates. The data was collected from total of 810 ELBW neonates by first developing the model through the training set, than validation was done and then tested using test set for predicting the outcomes. There was no statistical difference in values calculated for AUC i.e. 0.87±0.03; *p*= 0.31 among the both LR and ANN. Apgar score and gestational ages were significant risk factors for redicting the LBW. The conclusion was that both LR and ANN have excellent ability to predict the risk factors of LBW, and moreover, the ability of ANN is no more superior specially in case of non-linear relationships[17].

One retrospective cohort study, on the other hand, aimed to compare the predictive accuracy of ANN and LR in identifying the neurodevelopmental diseases among ELBW babies. There were 21 variables in this study that were segregated in the training(144) and the test-sets(74). First the neural network was trained as well as LR model was formed using the training set first and than their respective outcomes were compared using the test sets. In this study, although both models were equally suitable for predictive purposes, the sensitivity as well as correlation with worse clinical outcome were not significant. They also reported that although the prediction of these disorders was accurate enough by both models, the variance of these was not explained much by either of these models[18].

Table 1: Comparison of Statistical Tests used by Artificial Neural Network (ANN) and Logistic Regression (LR)

Author	year	Study design	Sample Size	Comparison	ANN value	Logistic Regression value
Kirişci, (2019)	2019	Retrospective cohort	223	AUC	0.941	0.909
Senthilkumar and Paulraj, (2015)	2015	Not given	--	sensitivity, specificity, accuracy, AUC, F-statistics, precision and recall	0.9385, 0.3729, 0.7619, 0.7804, 0.8443, 0.7673 and 0.9385	0.9231, 0.3390, 0.7407, 0.7724, 0.8304, 0.7547 and 0.9231
Soleimani et al., (2013)	2013	Retrospective	1232	Concordance Index, AUC, sensitivity, specificity	0.83, 0.79, 0.93, 0.39	0.79, 0.68, 0.92, 0.22
Ambalavanan and Carlo, 2001)	2001	Retrospective Cohort	810	AUC±SE	0.83±0.03	0.82±0.03
Ambalavanan et al., 2000	2000	Retrospective Cohort	218	AUC for major handicap, MDI, PDI	0.62,0.66,0.75	0.68, 0.75,0.69
Ambalavanan and Carlo, (1999)	1999	Retrospective Cohort	810	Full data AUC, significant variables AUC	0.83, 0.87	0.82, 0.87

DISCUSSION

The modern healthcare systems are rapidly adapting data mining techniques to utilize huge amounts of clinical data for better decision making regarding disease management and health issues[19]. Use of techniques such as ANN, SVM, classification trees and others have enabled medical professionals and researchers to early screen vulnerable cases and manage high risk patients timely to reduce fatal health consequences and health costs[20]. In recent years, data mining techniques are getting adequate attention for identification of risk factors of LBW[21]. Traditional logistic method is considered the gold standard statistical technique with which other data mining techniques are compared. These techniques include ANN, SVM, Random Forest, and others[22]. Literature suggests that use of these data mining techniques for LBW may potentially screen high risk mothers in advance and reduce the incidence of LBW in our local setting.

The published literature, both local and global, has mostly focused on identification of risk factors of LBW using any one of the mentioned statistical techniques. However, studies that compare the accuracy and ability of these statistical modeling techniques to identify risk factors of LBW are lacking[23]. Establishing the better predictive technique is important as it may serve as cornerstone in prediction and prevention of Low Birth Weight[21]. Therefore, such study has been conducted to compare the predictive ability of ANN and LR and identify the common statistical tests used for this comparison.

In this review, 6 studies were shortlisted that compared the diagnostic accuracy and predictive ability for identification of risk factors of LBW. Among these the latest one was published in 2019 whereas the oldest study was of 1999. The study designs of all these studies were

retrospective and a total of 3,293 subjects were included in all 6 studies.

The commonest statistical test compared among the two methods was Area Under Curve (AUC) reported in all 6 studies. The AUC for neural network performed slightly better compared to logistic in all reported studies.[13-18] Sensitivity and specificity were compared in three studies. Senthilkumar and Paulraj where sensitivity (0.9385 vs 0.9231) and specificity (0.3729 vs 0.3390) both were better for neural network compared to logistic regression. [14] In study by Soleimani et al., the sensitivity (0.93 vs 0.92) was slightly better for ANN and specificity (0.39 vs 0.22) was considerably better for ANN compared to logistic regression as well[16]. Ambalavanan et al., also reported better sensitivity and specificity for ANN compared to logistic regression.[18] Other statistical tests compared included PPV, NPV, Accuracy, F-statistics, precision, recall and concordance index by Senthilkumar et al., and Soleimani et al. [14] Other studies also suggest that use of ANN can help identify pregnancy related issues, estimate fetal birth weight and improve feto maternal outcomes[24,25]. This study, hence, presents ANN as a reliable, effective and accurate statistical tool that can be used independently or complementary with logistic regression for determining prevalence and risk factors of LBW.

Lack of comparative literature a major limitation in this study. There are only few studies that have compared the predictive accuracy of these two statistical techniques, especially for risk factors of low birth weight. So only few studies were freely available for comparison of these statistical tests. Also, these studies used different study designs and tests that were compared. Therefore establishing consensus in deciding the best statistical tests for comparison is difficult.

Therefore it is recommended that more studies should be done for comparing the predictive ability of the two data mining techniques so that the definitive superiority of the better procedure could be determined.

CONCLUSION

Data mining techniques are reliable, powerful, and sophisticated tools for handling complex data with high accuracy. ANN can be advantageous over LR specially if considerable inter and intra-relationships of outcome with risk factors and complicated non-linear relationships exist in data. In our study, ANN performed better though some numeric results showed very close to LR. We conclude that both models are beneficial and if used to complement each other, can be quite helpful for physicians in decision making.

REFERENCES

1. Mahumud RA, Sultana M, Sarker AR. Distribution and Determinants of Low Birth Weight in Developing Countries. J Prev Med Public Health. 2017;50(1):18-28.
2. Rezende Chrisman J, Mattos IE, Koifman RJ, Koifman S, Moraes Mello Boccolini P, Meyer A. Prevalence of very low birthweight, malformation, and low Apgar score among newborns in Brazil according to maternal urban or rural residence at birth. J Obstet Gynaecol Res. 2016;42(5):496-504.
3. Sachdev HPS. Low birth weight in South Asia. Int J Diab Dev Countries. 2001;21(1):13-33.
4. Wardlaw TM. Low birthweight: country, regional and global estimates: Unicef; 2004.
5. Pandey SC, editor. Data mining techniques for medical data: a review. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES); 2016: IEEE.
6. Al-Shawwa MO, Abu-Naser SS. Predicting Birth Weight Using Artificial Neural Network. Int J Acad Health Med Res. 2019;3: 9-14.
7. Krishnaiah V, Narsimha G, Chandra NS. A study on clinical prediction using Data Mining techniques. Int J Comput Sci Engin Infor Technol Res. 2013;1(3):239-48.
8. Durairaj M, Ranjani V. Data mining applications in healthcare sector: a study. Int J Sci Technol Res. 2013;2(10):29-35.
9. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol. 2004;25(8):690-5.
10. Padmavathi J. Logistic regression in feature selection in data mining. Int J Sci Engin Res. 2012;3(8):1-4.
11. Mohammad S, Suheil M. ANN for Predicting Birth Weight. Int J Acad Health Med Re. 2020;3(1):9-12.
12. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162(11):777-84.
13. Kirişci M. Comparison of artificial neural network and logistic regression model for factors affecting birth weight. SN Appli Sci. 2019;1(4):1-9.
14. Senthilkumar D, Paulraj S, editors. Prediction of low birth weight infants and its risk factors using data mining techniques. Proceedings of the 2015 international conference on industrial engineering and operations management; 2015.
15. Ambalavanan N, Carlo WA. Prediction of extremely low birth weight (ELBW) neonatal mortality by neural networks and logistic regression. Pediatr Res. 1999;45(7):236-.
16. Soleimani F, Teymouri R, Biglarian A. Predicting developmental disorder in infants using an artificial neural network. Acta Med Iran. 2013;51(6):347-52.
17. Ambalavanan N, Carlo WA. Comparison of the prediction of extremely low birth weight neonatal mortality by regression analysis and by neural networks. Early Hum Dev. 2001;65(2):123-37.
18. Ambalavanan N, Nelson KG, Alexander G, Johnson SE, Biasini F, Carlo WA. Prediction of neurologic morbidity in extremely low birth weight infants. J Perinatol. 2000;20(8 Pt 1):496-503.
19. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform. 2008;77(2):81-97.
20. Milovic B, Milovic M. Prediction and decision making in health care using data mining. Kuwait Chapt Arab J Busi Manag Rev. 2012;1(12):1-11.
21. Hange U, Selvaraj R, Galani M, Letsholo K, editors. A Data-Mining Model for Predicting Low Birth Weight with a High AUC. International Conference on Computer and Information Science; 2017: Springer.
22. Ahmadi P, Alavimajd H, Khodakarim S, Tapak L, Kariman N, Amini P, et al. Prediction of low birth weight using Random Forest: A comparison with Logistic Regression. Arch Advan Biosci. 2017;8(3):36-43.
23. Kitsantas P, Hollander M, Li L. Using classification trees to assess low birth weight outcomes. Artif Intell Med. 2006;38(3):275-89.
24. Tayal DK, Meena K, Kumar S, editors. Analysis of various Data Mining Techniques Techniques for Pregnancy related issues and Postnatal health of infant using Machine Learning and Fuzzy Logic. 2018 3rd International Conference on Communication and Electronics Systems (ICCES); 2018: IEEE.
25. Moreira MW, Rodrigues JJ, Furtado V, Mavromoustakis CX, Kumar N, Woungang I, editors. Fetal Birth Weight Estimation in High-Risk Pregnancies Through Machine Learning Techniques. ICC 2019-2019 IEEE International Conference on Communications (ICC); 2019: IEEE.