ISSN 0975-3303
Mapana J Sci, 9, 1 (2010), 18-30.
https://doi.org/10.12725/mjs.16.3

DATA MINING APPROACH IN PRETERM BIRTH PREDICTION

Jyothi Thomas and G. Kulanthaivel***

ABSTRACT

Data mining refers to the process of discovering patterns in data, typically with the aid of powerful algorithms to automate part of the search. These methods come from the disciplines such as statistics, machine learning, pattern recognition, neural networks and database. In particular this paper reveals out how the problem of preterm birth prediction is approached by a data mining analyst with a background in machine learning. In the health field, data mining applications have been growing considerably as it can be used to directly derive patterns, which are relevant to forecast different risk groups among the patients. Data mining technique such as clustering has not been used to predict preterm birth. Hence this paper made an attempt to identify patterns from the database of the preterm birth patients using clustering.

Key words: *Data mining, preterm birth, patterns, knowledge, clustering*

1. Introduction

Pre Term Birth (PTB), defined as birth before 37 completed weeks gestation, is the leading cause of mortality occurring before 28 days of age, accounting for 85% of

* *Research scholar, Sathyabama University, Chennai.*
j.thomas@christuniversity.in (Corresponding author)
** *NITTTR, Chennai*

all neonatal deaths not due to lethal congenital malformations [1,2]. Due to its direct correlation with infant mortality, reducing the burden of this problem has become the number one neonatal health priority. Many studies have attempted to predict women at risk of PTB; so far, no scoring system has proven itself superior to clinical judgment. One of the major obstacles is that most women who deliver prematurely have no obvious risk factors and over half of all PTBs occur in low-risk pregnancies [3]. In the obstetrical community the most commonly applied predictive model for PTB is based on a combination of obstetric history, fetal fibronectin testing and cervical factors. In a study by the National Institute of Child Health and Human Development (NICHD) Maternal-Fetal Network, a woman with a history of prior preterm delivery at 26 weeks but with a normal cervical length and negative fetal fibronectin status, had an 8% risk of PTB prior to 35 weeks gestation in her next pregnancy [4]. Ability to predict PTB increased from 28% to 30% if either test was positive, and to a maximum sensitivity of 66% when both tests were positive. Fetal fibronectin testing is most commonly used for: (1) symptomatic patients (women who report contractions between 24 and 36 weeks gestation) who have a cervix that is less than 3 cm dilated and are being evaluated for risk of preterm delivery; and (2) at-risk women for whom a negative test result could eliminate the need for intervention [5]. However, current evidence does not support the use of cervical length or fetal fibronectin status in screening for risk of preterm delivery in a low-risk population [5].

2.Need of Focusing on PTB

While mortality rates are improving in many countries worldwide, neonatal mortality rates (deaths in the first 28 days of life) have shown much less progress [20]. Neonatal deaths now account for more than 42% of under-five deaths (Figure 1), up from 37% in the year 2000 when the Millennium Development Goals (MDGs) were set [22]. MDG 4 targets a two-thirds reduction of under-five deaths between 1990 and 2015.

Complications of preterm birth are the leading direct cause of neonatal mortality, accounting for an estimated 27% of the almost four million neonatal deaths every year, and act as a risk factor for many neonatal deaths due to other causes, particularly infections [21]. Hence, achievement of MDG 4 is strongly influenced by progress in reducing neonatal deaths; and since preterm birth is the leading cause of these deaths, progress is dependent on achieving high coverage of evidence-based interventions to prevent preterm delivery and to improve survival for preterm newborns [5]. In some high-income countries, preterm birth has been high on the maternal, newborn and child health (MNCH) agenda for two decades,

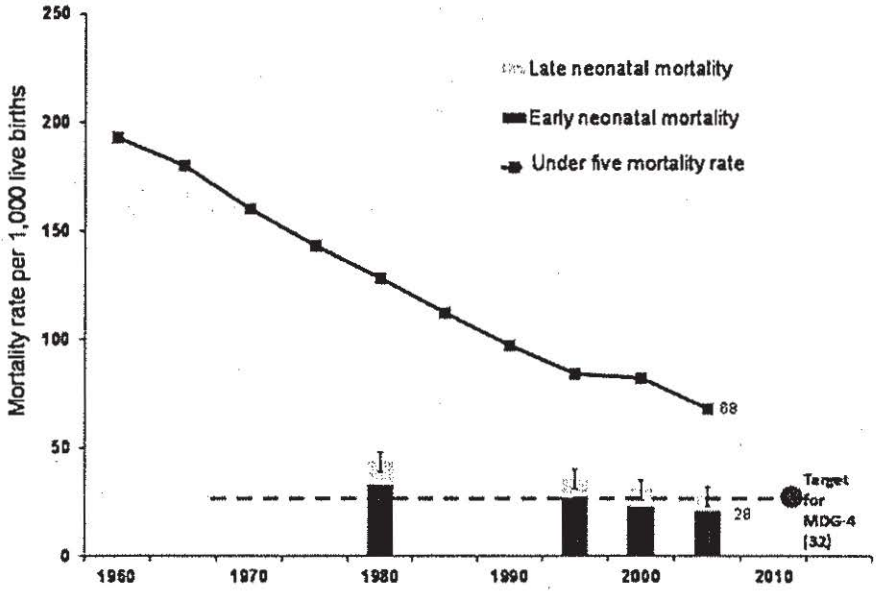

Figure 1. Mortality rate

but is now starting to receive wider public health attention because of increasing preterm birth rates. However, only recently has this issue started to reach the attention of higher-level policy makers in low- and middle-income countries. Many countries have recognized the importance of preterm birth and are looking for solutions in prevention as well as improved care. Understanding and improving the current data are critical to setting priorities for action and for tracking progress.

Increasing attention for preterm birth and stillbirth interventions, alongside increasing investment for mothers, will accelerate progress for these inextricable maternal, fetal, newborn and child health outcomes. Improved data on these pregnancy outcomes are crucial to guiding investment and tracking progress.

Prediction

Most preterm deliveries follow spontaneous onset of preterm labour or preterm prelabour rupture of the amniotic membranes (pPROM). Much work has been done (with limited success) to find diagnostic tests that predict accurately if a woman who is at risk of preterm delivery will go on to deliver preterm. For these women, who may have a history of preterm birth or clinical signs of preterm labour, such tests would allow early and targeted use of antenatal interventions. These interventions,

especially antenatal corticosteroids, improve neonatal and long term outcomes for preterm infants. Length of the endocervix can be measured using the transvaginal sonography (Figure 2).

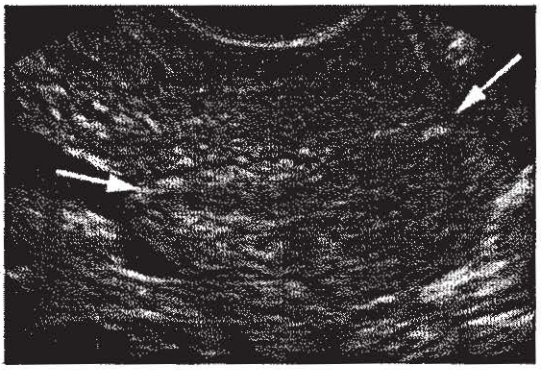

Figure 2. Sonography image

The most common clinical tests used to determine the risk of preterm labour are transvaginal sonography (to measure the length of the endocervix) and the cervicovaginal fetal fibronectin test. These tests have high negative predictive values - that is, if results are negative then the women probably will not progress to preterm delivery. Although there does not seem to be a role for routine use of the fibronectin test or transvaginal sonography to screen women for preterm birth, women thought to be at high risk can be reassured by negative results. This may help women to avoid unnecessary interventions such as antenatal transfer to a distant perinatal unit.

Maternal and Fetal Indications

About 15% to 25% of preterm births are caused by obstetric or medical complications of pregnancy. Obstetric complications such as pre-eclampsia may result in maternal morbidity or mortality and perinatal death if the infant is not delivered. Maternal risks of pre-eclampsia include eclamptic seizures, cerebral haemorrhage, HELLP (haemolysis, elevated liver enzymes, low platelets) syndrome, and maternal death.

Women with diabetes, renal disease, autoimmune disease, and congenital heart disease need intensive surveillance. Preterm delivery may be indicated by deterioration of maternal or fetal health, and obstetric complications may occur. When planning the timing and mode of delivery of preterm infants in these circumstances, it is necessary to weigh the risks to the mother and fetus of continuing the pregnancy against the risks of preterm birth and delivery. With the potentially

compromised very preterm fetus, the aim is to allow the pregnancy to continue to a point before damage occurs without taking unnecessary risks that may harm the mother.

Several tests of fetal wellbeing are available. In high risk pregnancies, fetal growth is usually monitored using serial ultrasonography to measure circumference of the head and abdominal girth. A fall in the growth velocity of the abdominal circumference indicates intrauterine growth restriction. Many factors must be taken into account when deciding the timing and type of delivery.

Cardiotocography and fetal biophysical profiling are two tools often used to determine the physiological status of the potentially compromised fetus. Unfortunately these tools have no benefit in predicting and preventing poor outcomes in high risk pregnancies. Some evidence shows, however, that computerized cardiotocography is more accurate in predicting poor outcome than subjective clinical assessment alone. The biophysical profile takes into account the tone, movement, breathing, heart rate pattern of the fetus, and liquor volume.

Doppler

Doppler measurement of fetoplacental blood velocity may be more a useful test of fetal wellbeing than cardiocotography or biophysical profiling. Umbilical arterial blood flow becomes abnormal when there is placental insufficiency. A recent systematic review of randomized controlled trials did not indicate that Doppler measurement of fetoplacental blood velocity is associated with a substantial reduction in prenatal mortality. Additionally, there is uncertainty over the ideal frequency of examination and the optimum threshold for intervention. Umbilical artery Doppler ultrasonography to detect fetal compromise is part of routine obstetric practice for high risk pregnancies in many countries, so there will probably be further randomized controlled trials in high risk populations. Growth charts are used to plot the circumference of the head and abdomen over time (menstrual weeks) (Figure 3). This chart shows the progress of a fetus with intrauterine growth restriction.

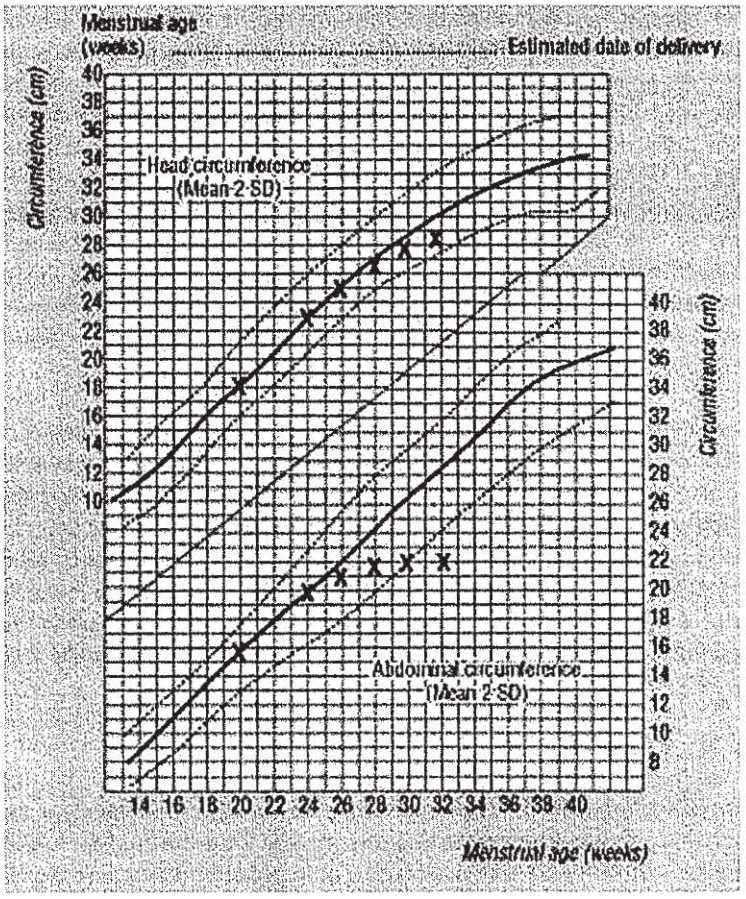

Figure 3. Growth chart (menstrual weeks)

Recent studies have investigated the use of middle cerebral artery and ductus venosus Doppler waveforms in evaluating cardiovascular adaptations to placental insufficiency. Results are promising, although the effect on important outcomes when used as part of clinical practice has yet to be evaluated.

Knowledge Discovery

Human analysts with no special tools can no longer make sense of enormous volumes of data that require processing in order to make informed business decisions. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support

system or assessed by a human analyst. The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various businesses, scientific and government organizations around the world is daunting. Hospital Scientific and business organizations store each day about 1 TB (terabyte) of new information. It becomes impossible for human analysts to cope with such overwhelming amounts of data. Two other problems that surface when human analysts process data are the inadequacy of the human brain when searching for complex multifactor dependencies in data and the lack of objectiveness in such an analysis. A human expert is always a hostage of the previous experience of investigating other systems. Sometimes this helps, sometimes this hurts, but it is almost impossible to get rid of this fact.

One benefit of using automated data mining systems is that this process has a much lower cost than hiring highly trained (and paid) professional statisticians. While data mining does not eliminate human participation in solving the task completely, it significantly simplifies the job and allows an analyst who is not a professional in statistics and programming to manage the process of extracting knowledge from data. The process of storing knowledge discovery in data bases consists of sequence of steps including problem understanding, data understanding and preparation, data mining result interpretation and evaluation finally the use of discovered knowledge.

Clustering Revisited

The need to analyze data for decision making is growing exponentially, since data collection through electronic version grows rapidly. Thus the field of data mining has emerged at the intersection of statistics, data bases and machine learning for development of the techniques to obtain information and the knowledge from vast amounts of micro data, which are of numerical and categorical in nature. The development of hardware and software and the rapid computerization of business have made capturing the data easy and digitized information, this makes the collection and storing the data to grow at a phenomenal rate. As a result, traditional adhoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing such data.

Raw data is rarely of direct use. Its true value is predicted on the ability to extract information useful for decision support or exploration and understanding the phenomena governing the data source. One or more analyst may be intimating familiar with the data and with the help statistical techniques provide summaries

and generate reports. Hence the analysts are acting as a sophisticated query processor. However such manual query processing has its own limitations as the size of data grows and the number of dimension increases. Since the scale of data manipulation, exploration and inference go beyond human capacities, computing technologies become inevitable. Partitioning a set of objects into homogeneous clusters is fundamental operation in Data mining and the operation is needed in a number of Data Mining tasks such as unsupervised classification and Data Summation. This operation is also used in segmentation of large heterogeneous Data sets into smaller homogeneous subsets that can be easily managed, separately modeled and analyzed. Clustering is a popular approach used to implement this operation. Clustering methods partition a set of objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. In statistical clustering methods [8, 9], we use similarity measure to partition objects, whereas in conceptual clustering methods [10], we cluster the objects according to the concept of objects.

The Data mining community has recently put a lot of efforts on developing fast algorithms for clustering large Data sets. Some popular algorithms include CLARA program [11], CLARNS [11], DBSCAN [12], K-modes algorithm [12], K-prototypes [13], and PCBClu [14]. These algorithms are often revisions of some existing clustering methods by using some carefully designed search methods (e.g. combination of sampling procedure and the clustering program PAM in CLARA program, randomized search in CLARANS), organizing structures (e.g. CF-Tree in BIRCH and PC-tree in PCBClu). Indices (e.g., R*-Tree in DBSCAN) and statistical methods (frequency and dissimilarity measure in K-prototypes and K-modes). These algorithms have shown some significant performance still based on complex schemes and procedures. They cannot be used to solve massive categorical data clustering problems as simple as K-means clustering methods in numerical domain.

The K-means based methods [15] are efficient for processing the large data sets, thus very attractive for Data mining. The major handicap for them is that they are often limited to numeric data. The reason is these algorithms optimize a cost function defined on the Euclidean distance measure between the data points and means of cluster. Minimizing the cost function by calculating means limits they used numerical data.

The K-means Algorithm

The K-means algorithm [9,14] is build upon the following operations.

Step 1: Choose initial cluster Centers Z1, Z2, ...Zk randomly from the n points w1, w2, wN, wl € Rm

Step 2: Assign point Wi, I = 1, 2 , N
to Cluster Cj = 1, 2,.... K
If and only if ||Wq - Zq,
P = 1,2, K and J != P.
Ties are resolved arbitrarily

Step 3: Compute the new cluster centers $Z_1^*, Z_2^*, \ldots Z_k^*$
as follows $Zi^* = (1/n)\,\Sigma Wj$
I= 1,2,....k Wj € CJ

Step4: If Zi * = Zi, I = 1, 2,...K
then terminate.
Otherwise Zi – Zi* and go to step 2.

Except for the first operation, the other three are repeatedly performed in the algorithm until the algorithm converges. Note that in case the process does not terminate normally at Step 4, then it is executed for a maximum fixed number of iterations.

The optimality of this algorithm can be estimated by Inter and Intra clustering metric values which has been calculated by sum of the Euclidean distances.

Mathematically, the clustering intra metric ì for K clusters $C_1, C_2, \ldots, C_k$

$$\mu (C_1, C_2, \ldots C_k) = \sum_{i=1}^{k} \sum_{xj \in Ci} || Zi - Zj ||$$

Where Ci are Clusters and Zj are cluster centers And inter Cluster metric v for k clusters $C_1, C_2, \ldots C_k$

$$V (C_1, C_2, \ldots C_k) = \sum_{i=1}^{k} \sum_{j=k+1} ||Zi - Zj ||$$

The task of the proposed clustering technique is to search for the appropriate cluster centers $Z_1, Z_2, \ldots Z_k$ such that the clustering intra-cluster metric i is minimized and inter cluster metric v is maximized.

There exist a few variants of algorithm which differ in selection of the initial K-means, dissimilarity calculations and strategies to calculate cluster means [8,16]. The sophisticated variants of the K-means algorithm include the well known ISODATA algorithm and the fuzzy K-means algorithms [19].

Most K-means type algorithms have been proved convergent [15,17]. The K-means algorithm has the following important properties.

- It is efficient in processing large data sets. The computational complexity of the algorithm is O (tkmn), where m is the number of attributes, n is the number of objects, k is the number of clusters and t is the number of iterations over the whole data set. Usually, k. m, t * n. In clustering large data sets, the K-means algorithm is much faster than the hierarchical clustering algorithms whose general computational complexity is $O(n^2)$.
- It often terminates at a local optimum [15,17]. To find out the global optimum, techniques such as deterministic annealing [18] and genetic algorithm can be incorporated with the K-means algorithm.
- It works only on numeric values because it minimizes a cost function by calculating the means of clusters.
- The clusters have convex shape [8]. Therefore, it is difficult to use the K-means algorithm to discover clusters with non-convex shapes.

The main difficulty in using the K-means algorithm is to specify the number of clusters. Some variants like ISODATA include procedure to search for the best K at the cost of some performance. But the extensive studies dealing with comparative analysis of different clustering methods suggest that there is no general strategy, which works equally well in different problem domain. However it has been found that it is usually beneficial to run schemes that are simpler and execute them several times like K-means, rather than using schemes that are very complex but need to be run only once.

The K-Means algorithm is best suited for data mining because of its efficiency in processing large data sets. However working only on numeric values limits its use in data mining because data sets in data mining often have categorical values.

Case Study

Full term births are between 37 and 42 gestational weeks long, while those happening before are considered to be preterm. Although infants born after 20 weeks of gestation can survive, they frequently suffer from life long and severely debilitating handicaps. Moreover the care for these preterm neonates costs very high. In conclusion preventing preterm birth and prolonging gestation is not only an important medical issue but also a public health and a health care managing problem. Identifying the risk factors has been concentrating substantial research efforts, as this would help in developing models for risk prediction [1,23].

Data Set

The data for this study were provided by the Department of Gynecology, Fr. Muller Medical College, Mangalore. Data consisted of prenatal information collected from the medical records of patients who received prenatal care in the hospital during Jan – Dec 2009. The analyzed data consist of 150 records. The set of data was obtained by preliminary preprocessing which includes eliminating the records with missing values and attributes which are irrelevant in predicting preterm. The number of attributes obtained after the preliminary preprocessing is 20 which are shown in the table.

Table 1. Selected attributes

Code	Attribute	Code	Attribute
1	Maternal age	10	Weight gain during pregnancy
2	Body Mass	11	Fundus uterus height
3	Hemoglobin level	12	Gestational age
4	Low/high red cell count	13	Type of birth
5	Glucose level	14	Child sex
6	Systolic BP	15	Child head perimeter
7	Diastolic BP	16	Child weight
8	Abdominal perimeter	17	Child height
9	No. of pregnancies	18	Live/still birth

The data are then clustered with K-means algorithm using MATLab.

3. Conclusion

It has been observed that the prediction of preterm birth patient groups from existing or easily measured demographic data using clustering yield better solution for the problem. The result obtains from this study are shown to be consistent with traditional medical diagnosis techniques and are really useful in prediction of non linear groups, which are essentially different risk groups.

References

1. L. Goodwin and S Maher, "Data mining for preterm birth prediction," Proceedings of the 2000 ACM Symposium on Applied Computing (SAC'00), 2000, pp. 46-50.

2. Y. Dawood, "The obstetric view of premature labor" chapter in *Historical review and recent advances in neonatal and prenatal medicine* (G.F. Smith, D. Vidyasagar, Eds.), Mead Johnson, 1980.

3. J.D. Iams, R.L Goldenberg, B.M. Merber, A.H. Moawad, P.J. Meis, A.F. Das and S. Caritis, "The preterm prediction study: can low-risk women destined for spontaneous preterm birth be identified?" *General Obstetrics and Gynecology*, vol. 184(4), pp. 652-655, 2001.

4. Iams J.D. The Preterm Prediction Study: A model for estimation of risk of spontaneous preterm birth in parous women. NICHD Network. *American Journal of Obstetric and Gynecology.*, vol.176:S51, 1997.

5. Aerts M., Iams J., "Prevention of spontaneous preterm birth," *Contemporary OB/GYN.* May 1999; pp. 128-136.

6. C. Catley, M. Frize, D.C. Petriu, C.R. Walker and L. Yang, "Towards a Web Services Infrastructure for Perinatal, Obstetrical, and Neonatal Clinical Decision Support", in *Proceedings of IEEE-EMBS/ BMES 2004*, 2004.

7. Meis P.J., Goldenberg R.L., Mercer B.M., Iams J.D., Moawad A.H., Miodovnik M., Menard K., Caritis S.N., Thurnau G.R., Bottoms S.F., Das A., Roberts J.M. McNellis D., "The preterm prediction study: Risk factors for indicated preterm births". *American Journal of Obstetrics and Gynecology*, vol. 178(3), pp. 562-567, 1998.

8. Anderberg, *Cluster Analysis for Applications*, Academic Press, 1973.

9. Jain A.K., R.C. Dubes, 1988, *Algorithm for Clustering Data*, Prentice Hall.

10. Fisher, D.H, *Knowledge acquisition via incremental conceptual clustering, Machine learning*, 1987, 139-172.

11. Kaufman and Rousseeuw, *Finding groups in Data – An introduction to Cluster Analysis*, 1990.

12. Huang, Clustering large data sets with numeric and categorical values, Proceedings of the First Pacific Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific, 1997, pp. 21-34.

13. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of the SIGMOD workshop on research issues on data mining and knowledge discovery.

14. Ananthanarayana, M.N Murthy, Subramanian, *Efficient clustering of large data sets, Pattern recognition*, 2001, vol 34., pp. 2561-2563.

15. MacQueen, Methods of classification and analysis of multivariate observations, proceedings of Fifth Berkely Symposium on Mathematical statistics and Probability, 1967, pp. 281-297.

16. Bobrowski and Bezdeck, c-means clustering with the 11 and 14 norms, *IEEE transactions on System, Man and Cybernetics,* 1991, vol. 21, pp. 545-554.

17. Bezdeck, A Convergence Theorem for fuzzy ISODATA clustering algorithms, *IEEE transactions on Pattern Analysis and Machine Intelligence,* 1980, vol. 2, pp. 153-155.

18. Murtagh, Comments on Parallel Algorithm for hierarchical clustering and cluster validity, *IEEE transactions on Pattern Analysis and Machine Intelligence,* 1992, vol. 4, pp. 1056-1057.

19. Rupsini, *A new approach to clustering, Information control,* vol. 19, 273-284.

20. Lawn J.E., Kerber K., Enweronu-Laryea C., Massee Bateman O, Newborn survival in low resource settings—are we delivering? *BJOG* 2009, 116(Suppl 1):49-59.

21. Lawn J.E., Cousens S., Zupan J., 4 million neonatal deaths: when? Where? Why? *Lancet* 2005, 365(9462):891-900.

22. Lawn JE, Wilczynska-Ketende K., Cousens S.N., *Estimating the causes of 4 million neonatal deaths in the year 2000.*

23. Lovell, Rosario, Niranjan, Design construction and evaluation of systems to predict risk in obstetrics, *International Journal of Medical Informatics,* 1997, vol 46, pp. 159-173.