Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Expert system supporting an early prediction of the bronchopulmonary dysplasia

Marcin Ochab*, Wiesław Wajs

AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland

ARTICLE INFO

Article history:
Received 30 October 2014
Accepted 22 August 2015

Keywords:
Machine learning
Feature selection
Chronic lung disease
Prediction
Bronchopulmonary dysplasia
Expert system
Support vector machine
Logit regression
Prematurity
Low-birth-weight infant

ABSTRACT

This work presents a decision support system which uses machine learning to support early prediction of bronchopulmonary dysplasia (BPD) for extremely premature infants after their first week of life. For that purpose a knowledge database was created based on the historical data gathered including data on 109 patients with birth weight less than or equal to 1500 g. The core of the database consists of support vector machine and logit regression classification results calculated specifically for that system, and obtained by considering 2^{14} different combinations of 14 risk factors. Based on the results obtained and user demands, the system recommends the best methods and the most suitable parameter subset among those currently available to the user. The program is also able to estimate the accuracy, sensitivity and specificity together with their standard deviations. The user is also given information on which additional parameter it is worth adding to his measurement system most and what an increase in prediction efficiency it is expected to trigger. The BPD can be predicted by the system with the accuracy reaching up to 83.25% in the best-case scenario, i.e. higher than for most of the models presented in the literature. This work presents a set of examples illustrating the difficulties in obtaining one single model that can be widely used, and thus explaining why an expert system approach is much more useful in day-to-day clinical practice. In addition, the work discusses the significance of the parameters used and the impact of a chosen method on the sensitivity and specificity.

1. Introduction

Bronchopulmonary dysplasia (BPD) is a chronic pulmonary disorder affecting premature infants [1], which results in significant morbidity and mortality: almost a third of infants with birth weight lower than 1000 g [2] are affected. This chronic lung disease is most common among children with low birth weight and those who received prolonged mechanical ventilation to treat the respiratory distress syndrome. Due to the fact that the disease is poorly understood, many research projects are focused on identifying its risk factors. It is known that steroids applied before the eighth day of life can prevent BPD development; however, the risks of such treatment may outweigh the benefits [3]. Since illness cannot be diagnosed until the 28th day of life [4], it is very important to predict such a result by the end of the first week, which would enable early intervention and an increased likelihood of preventing the disease [3]. Therefore, an intensive work has been done to define a classifier, based on static parameters (gathered just after birth) and dynamic ones (collected during the first week of life), which would be able to predict the diagnosis. The literature [5–15] reports several prediction models of BPD used in the research. However, none of them could be used in common clinical practice due to a variety of reasons. For instance, when for some reason there was no technical capability to measure one of the parameters just after birth or the mensuration was delayed, no models based on that variable can be used. Similarly, when using roentgenographic scoring systems – an image of sufficient quality which could be compared with a database, especially by a machine, is extremely hard to obtain, because of ceaseless child movements. Therefore, it is essential to propose an expert system which could advise on which model and method to choose in a given situation.

2. Background

As already mentioned, there are numerous papers devoted to BPD, its risk factors and prediction [16–21]. The most common factors are derived from the analysis of static data among which gestational age and birth weight are mainly considered [9,12,13,15,18,21]. The other factors covered by the literature are administration of surfactant

* Corresponding author at: ul. Jastrzębia 4, 35-207 Rzeszów, Poland. Tel.: +48 504024195 fax: +48 178630252.
E-mail addresses: mj.ochab@labor.rzeszow.pl (M. Ochab), wwa@agh.edu.pl (W. Wajs).

http://dx.doi.org/10.1016/j.compbiomed.2015.08.016
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

[3,17–19,22], presence of patent ductus arteriosus (*PDA*) [3,9,14,16–18,20,22–24], or respiratory support [14,21]. Some of the papers introduced sex [3,18,19,21] or even race and ethnicity [21] as factors which seem to be promising. Unfortunately, due to the Polish social structure proposing the second as a parameter in our system would require a very big set of data which we do not have. An analysis of dynamical data can be found in the literature much less frequently, because it requires a constant acquisition of data during the first week of life. Most common parameters acquired that way are arterial blood gas variables like fraction of inspired oxygen (FiO_2) [6,7,9,12,13,20,22] or alveolar–arterial ratio (*AA*) [25,26] (which is respiratory distress degree measure Eq. (B.1); blood gas levels like oxygen saturation of arterial hemoglobin (SpO_2) [27] and its standard deviation, mean value, etc. [26], or even time series analysis [28]; heartbeat and its derivatives [26]. One can find several papers regarding BPD prediction with analysis of radiological images [8,10,11]. Unfortunately, during our investigation we found that those to which we had access and which were taken before fourth day of life give ambiguous results. Most likely, the infants' lungs were not developed enough in that stage of life. Therefore, we were not able to use lung images in presented system.

The vast majority of studies make use of logit regression (LOGIT) and the best of such models are able to achieve accuracy in the range 73–82%; some authors use neural networks [26] with accuracy over 80%. A few mention that use of support vector machine (SVM) [29] could give interesting results, but nobody has really investigated that method in the context of BPD prediction. That is why we compared SVM with LOGIT classifiers in our previous papers [30,31]. Generally, SVM models have proved to be more unstable than LOGIT and should be used with a particular care. However, we proved that in certain situations choosing a proper SVM model even from a limited group of randomly constructed ones may lead to better results.

In the a lack of a generally accepted model, a multitude of the ones proposed in the literature and considering our previous scientific experience in this respect, we propose to use an expert system. Such a system would assist doctors in deciding which parameters to measure, and which method to use in certain circumstances instead of searching for a single universal method of BPD forecasting. Since we did not find any mention of such a system in the literature we decided to construct one.

3. Materials and methods

3.1. Data collection

Thanks to the Neonatal Intensive Care Unit of The Department of Pediatrics at the Jagiellonian University Medical College, we were able to collect data with our own software. It includes 109 patients born prematurely with birth weight less than or equal to 1500 g, admitted no later than on the second day of life. For 46 of them BPD has been diagnosed after the fourth week of life. Data has been normalized to $[-1,1]$.

In the proposed expert system we consider 14 different features which are used in BPD prediction:

(a) binary parameters such as
- presence of patent ductus arteriosus (*pda*),
- use of a respirator (*respimv*) during the first week of life,
- administration of surfactant (*surfact*) in the same period;

(b) real-valued (range in parentheses) such as
- birth weight (*bweight*) (550–1500 g),
- gestational age (*gage*) (22–34 weeks),
- alveolar–arterial ratio (*aa*) (0.05–1) measured during patient admission, which depends on FiO_2 Eq. (B.1),
- the percentage of time during the first week for which the oxygen saturation of hemoglobin was less than 85% (*low*85) (0.03–12.45%) or higher than 94% (*high*94) (14.56–99.02%) [27],
- average number of heartbeats per minute (*bpmmean*) (124.69–161.42 bpm),
- mean and standard deviation of oxygen saturation (*spo2mean*, *spo2dev*) (89.89–98.99% and 1.19–7.98, respectively) and their trends (first day to first week ratio: *bpmmean_tr*, *spo2mean_tr*, *spo2dev_tr*) (0.8–1.18 , 0.96–1.07 and 0.51–2.36, respectively).

3.2. Prediction methods

As a prediction methods in our system we used SVM and LOGIT. A brief description of these two algorithms, with the equations and an explanation of the parameters, is provided in Appendix A. All computations were performed in the Matlab R2013a environment. To obtain probability of positive diagnosis, in LOGIT calculations we used functions *glmfit* and *glmval*, whereas in SVM we used LIBSVM library (version 3.17) [32] in C-SVC mode with a sigmoid kernel function, Eq. (1). As mentioned in the previous paper [31], the analysis of several arbitrary tested models revealed that for the specified problem C-SVC method is more effective than nu-SVC (which simply means that the acceptable range of penalty parameter *c*, Eq. (A.6), is from zero to infinity, rather than between [0,1]). It has also been investigated that the sigmoid kernel function gives better results and is much faster in finding the separating hyperplane than the radial basis function, Eq. (2):

$$Sigmoid: K(X_i, X_j) = \tanh(\gamma X_i^T X_j + r), \tag{1}$$

$$RBF: K(X_i, X_j) = e^{-\gamma \|X_i - X_j\|^2}, \quad \gamma > 0, \tag{2}$$

where γ, r are kernel parameters.

Accuracy (*ACC*) defined as below was considered as a preliminary result measure. The sensitivity (*TPR*) and specificity (*SPC*) were also obtained the same way:

$$ACC_i = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$TPR_i = \frac{TP}{TP + FN}, \tag{4}$$

$$SPC_i = \frac{TN}{TN + FP}, \tag{5}$$

$$ACC = \frac{1}{n} \sum_{i=1}^{n} ACC_i, \tag{6}$$

$$ACC_{dev} = \sqrt{\frac{\sum_{i=1}^{n} (ACC_i - ACC)^2}{n - 1}}, \tag{7}$$

where *TP* is True Positives, *FP* is False Positives, *FN* is False Negatives, *TN* is True Negatives, *i* is the Jackknife iteration, $n=30$ is the number of iterations.

Please cite this article as: M. Ochab, W. Wajs, Expert system supporting an early prediction of the bronchopulmonary dysplasia, Comput. Biol. Med. (2015), http://dx.doi.org/10.1016/j.compbiomed.2015.08.016