Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 565–570

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems (HiPIS 2017)

Step Towards Prediction of Perineal Tear

Francisca Fonseca[a], Hugo Peixoto[b], Filipe Miranda[b], José Machado[b]* and António Abelha[b]

[a]*University of Minho, Campus Gualtar, Braga 4710, Portugal*
[b]*Algoritmi Research Center, University of Minho, Campus Gualtar, Braga 4710, Portugal*

Abstract

The aim of this study is to predict, through data mining tools, the incidence of perineal tear. This kind of laceration developed during child delivery might imply surgery and entails a set of several consequences. Clinical Decision Support Systems, with the information collected from patients' electronic health records combined with the data mining techniques, may decrease the incidence of perineal tears during labour.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Data Mining; Obstetrics; Perineal Tear; Decision Support Systems.

1. Introduction

It is estimated that 85% of women had perineal tear during child birth[1]. These lacerations, which are resultant from episiotomies or spontaneous obstetrics tears that may happen during vaginal deliveries, have severe consequences including chronic perineal tear, dyspareunia, urinary incontinence and fecal incontinence[1,2]. Predicting these situations, based on the information collected during pregnancy, would allow obstetricians along with pregnant women to take some early measures in order to reduce the risk of developing perineal tears.
Decision Support Systems (DSS) can be defined as an interactive computer based systems that supports several phases of the decision making process[3,4]. Applying DSS in healthcare will provide knowledge and information on a

* Corresponding author. Tel.: +351 253 604 430; fax: +351 253 604 471.
E-mail address: jmac@di.uminho.pt

1877-0509
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.284

specific person, which was intelligently filtered and presented at appropriate times enhancing health and healthcare[5,6]. Knowledge based systems are the most popular type of Clinical Decision Support Systems (CDSS). Being also known as expert systems, these systems may perform different types of clinical tasks from alerts and reminders on a patient's condition to diagnostic assistance or recognition and interpretation of clinical images. Allying CDSS with electronic health records will provide the best practice and high quality care to the patient minimizing the errors that may occur[7,8]. An example of a CDSS based on Knowledge Discovery from Databases (KDD) and Agent-Based Systems paradigms, whose target area is intensive care medicine, is INTCare[4,9]. This multi-agent approach collects and processes data in real time providing new knowledge. Besides, it allows the prediction of clinical events or diseases with high sensitivities rates, which might be, for example, organ failure and patient outcome, using data mining techniques and adapting data mining models[4,9].

Data mining, which is a set of approaches that allows the extraction of information from data through its analysis in an automatic/semi-automatic way, is a descriptive or a predictive technique. Since the data referred to is often presented in large datasets, the descriptive processes, which might be, for example, automatic clustering, bring out knowledge present among all the data. On the other hand, the predictive (or explanatory) processes, as classification or scoring for qualitative data and as regression for quantitative, anticipate new information based on the present facts[10,11,12]. Moreover, data mining refers to a precise step of the overall process of discovering useful knowledge from data known as KDD. Unlike data mining, which is only the application of particular algorithms to remove patterns from data, KDD process includes other steps like data preparation, data selection, data cleaning, integration of prior knowledge and also analysis of the results in order to provide valuable knowledge[13].

This article includes six sections. After the Introduction, the second section, Background and Related Work, presents a brief description of the perineal tear role in obstetrics and its consequences followed by some studies similarly to this one. On Methodology, the third section, the tools and methods used are mentioned and described. Section four, Knowledge Discovering Process, all the process of data mining, where Cross Industry Standard Process for Data Mining was adopted, is presented as well as the results. These results are discussed on the fifth section being the conclusion and future work the last section.

2. Background and Related Work

2.1. Perineal Tears in Obstetrics

Not only is obstetrics the surgery sub-specialty responsible for providing care to women throughout pregnancy but also for surgeries related to child delivery[14]. Obstetricians follow future mothers during appointments along their pregnancy. Thus, potential problems might be detected and the patients advised on their future steps to avoid any pregnancy complications. When a woman feels signs of labour, her obstetrician and other medical staff implement the required delivery procedures[15]. During these delivery procedures women might develop perineal tears mostly because of the soft tissue of the birth canal overstretching. The severity of the tear, depending on their depth, varies from first degree, when the tear is small and it only involves skin and will heal naturally – only if it had affected the perineal muscles, as it happens in second degree tears, the mother would have needed stitches –, to fourth degree, where the tear affects not only the vaginal wall but also the area since the perineum to the anal sphincter (this same thing happens on third degrees tears) with an additional damage to the lining of the bowel. With the exception of the first degree tear, the other of the three degrees are only treated by surgery, and the reasons behind the occurrence of a perineal tear aren't yet clear[1]. In contrast, the consequences, as the ones referred to in the introduction, might have quite an impact in day-to-day life mainly due to the poor functional outcomes, after traditional surgery, to repair lacerations involving the anal sphincter complex[2].

2.2. Related Work

Given the constant progression of data mining and DSS, the applicability of these tools on the healthcare sector may minimize errors, where lives and their quality are involved, by supporting decision-making. Pereira et al. (2015) developed a model for the *Centro Hospitalar do Porto* with the aim of helping physicians making quick decisions on the most advisable type of delivery for a certain patient when following the traditional guidelines is

unfeasible. Applying the data mining methodology of Cross Industry Standard Process for Data Mining along with KDD, four data mining models were induced: Decision Trees, Generalized Linear Models, Support Vector Machine and Naive Bayes. This model is able to identify caesarean sections and vaginal deliveries, based on obstetric risk factors, with good results on sensitivity (90,11%) and specificity (80,05%) values[16]. Although not applied on obstetrics, Portela et al. (2015) takes advantage of data mining classification techniques and combines them with patient data gathered in real time in Intensive Care Units. Using Cross Industry Standard Process for Data Mining to complement Design Science Research Methodology, it aims to avoid future complications for the patient as hypotension or hypertension. By inducing the same data mining models as Pereira et al. (2015), the model predicts the probability of a patient having a blood pressure critical event with a promising sensitivity of 95%[9].

3. Methodology

The model of Cross Industry Standard Process for Data Mining (CRISP-DM) was followed. This methodology breaks the process of data mining in six steps: business understanding; data understanding; data preparation; modelling; evaluation; and deployment[17,18,19]. The machine learning software Weka was used in order to analyze the data along with five approaches to induce the data mining models. The mentioned techniques were Logistic Regression (LR), Naive Bayes (NB), k-Nearest Neighbours (kNN), Classification and Regression Trees (CRT) and Support Vector Machine (SVM). NB, kNN, CRT and SVM are all on the top-10 data mining algorithms being SVM acknowledged as one of the most robust and accurate methods[18,19]. SVM combines linear modelling with learning based on instances. This algorithm chooses a limited number of samples from each group and constructs a linear function building separated boundaries between datasets[18,20]. When no linear separation is feasible, the kernel approach will be used to automatically add the training samples into a higher dimensional space and to learn a separator in that zone[18]. The NB algorithm is based on the Bayesian rule. This probabilistic learner, which is very easy to build, is constructed on the hypothesis that features are conditionally independent from the class label[18,19]. Although this might not be the case of all datasets, this is one of the most practical techniques to certain types of learning problems[18]. Like SVM, the kNN classifier is an instance-based learning algorithm. It is also a simple iterative method where each item on a dataset, having a known class label, is classified by a majority vote of its k nearest neighbour. In other words, a given dataset is divided into a user-specified number of clusters, k, which, in its turn, is usually a small and positive integer and is the only adjustable parameter[18,19,20]. This is one of the most simple and commonly-used methods since it applies directly the data for classification without building the model first[18,20]. LR, as its name implies, is a regression analysis method. This algorithm estimates probabilities recurring to a certain logistic function measuring, this way, the relationship between the dependent variable with at least one independent variable[18]. Finally, CRT is a binary recursive technique able to process not only continuous but also nominal attributes being those targets or predictors. The CRT algorithm is intended to produce a sequence of trees, all candidates to optimal tree. This is identified in the pruning sequence when the predictive performance of each tree is assessed[19].

4. Knowledge Discovering Process

4.1. Business Understanding

The aim of this study is to identify the patterns in labour perineal tear. This study is intended to identify the variables related to perineal tear. Although there's not proven efficacy of the techniques used to reduce the tearing risk, in a clinical point of view it is expected to create knowledge which might help preventing this situation during labour.

4.2. Data Understanding

The data collected and presented for this study comes from a Portuguese hospital during the year of 2016. This information that composed the dataset was related to the pregnancy and moment of labour, the baby and the mother.

Table 1 presents de target variable percentage distribution.

Table 1. Target variable Distribution.

ID	*Variable*		Cases	
Perineal tear	No perineal tear	No	85,47%	85,47%
	First-degree tear	Yes	9,98%	14,53%
	Second-degree tear		3,8%	
	Third/Fourth-degree tear		0,75%	

4.3. Data Preparation

After selecting the data exposed before, a pre-processing phase started. In this phase, all the data with null and noise values were removed leaving 1370 records which will be used by the data mining models. Some of the data had inconsistent values – weight, height and cephalic perimeter in different units and BMI wrongly calculated (this last one was easily corrected since the mother's height and weight came along with the extracted values of BMI) – needing therefore treatment. Finally, to guarantee the data mining performance, the data was normalized converting all the values to a number between 0 and 1 and the point was used as decimal separator. In the case of the variables with continuous values, 0 and 1 correspond respectively to the minimum and maximum values.

The first data mining modelling tests didn't present satisfactory results leading to the transformation of the dataset over again. As a result, the target variable, which initially was to determine the type of perineal tear, ended up being to determine whether a mother would develop perineal tear or not during labour. Another change was to apply oversampling to the data. This technique replicates the cases of mothers who developed perineal tear so that the number of their occurrence is similar to the mothers who aren't affected, improving the outcomes of this study.

4.4. Data Preparation

With the data transformed and processed, the data mining models were induced using the methods listed in section 3. Two different data approaches were made, one of them using oversampling and the other without it, both testing on 1/3 of the data (Holdout Sampling) and on all the data (Cross Validation). The different scenarios presented below were also created combining different variables in order to identify which factors have more impact on developing perineal tear.

S1: {All variables}; S2: {BMI, Age, Blood type, Rhesus}; S3: {Sex, Weight, Height, Cephalic perimeter}; S4: {Number of days of pregnancy, Pregnancy programmed, Type of delivery, Analgesia}; S5: {Age, Blood type, Rhesus, Weight, Height, Cephalic perimeter, Number of days of pregnancy, Type of delivery, Analgesia} and S6: {Blood type, Weight, Height, Cephalic perimeter, Type of delivery}

The data mining model can be described through the equation where a data mining model (DMM) can be described by the approach (A), a set of scenarios (S), a sampling method (SM), a data approach (DA), a data mining technique (DMT) and a target (TG):

$$DMM_n = A_f \times S_i \times DMT_Y \times SM_C \times DA_b \times TG_t \quad (1)$$

A_f = {*Classification*}; S_i = {*Scenation1...Scenario6*}; DMT_y = {*LR, NB, kNN, CRT, SVM*}; SM_c = {*Holdout Sampling, Cross Validation*}; DA_b = {*Without Oversampling, With Oversampling*}; TG_t = {*perineal tear versus none perineal tear*}

Therefore, the data mining model will be: DMM = {*Approach, 6 Scenarios, 5 Techniques, 2 Sampling Method, 2 Data Approaches, 1 Target*} with a total of 120 models induced.

4.5. Evaluation

The induced models evaluation was performed by calculating the statistic metrics as it follows based on the results given by the confusion matrix. This matrix provides the number of true positives (TP), false positives (FP),

true negatives (TN) and false negatives (FN). With these in mind it is possible to define Precision, Sensitivity, Specificity and Accuracy.

In order to select the best models, a threshold was used. Ideally, this threshold should have combined three metrics with the goal of finding the most suitable model to predict the development of perineal tear during labour – sensitivity ≥90% – with a satisfactory accuracy and precision to prevent a high number of false positives. However, the results obtained had low values of precision as well as sensitivity. Thus, the best models selected and presented in table 2 were the ones with the highest values of accuracy and precision and sensitivity above 70%, using the sampling method Holdout Sampling. Table 3, likewise Table 2, display the best results, but using the sampling method Cross Validation.

Table 2. Best models achieving the highest values of sensitivity (Holdout Sampling).

Scenario	*Model*	*Approach*	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*
S1	kNN	With oversampling	0,5699	0,7794	0,7990	0,8760
S2	kNN	With oversampling	0,5753	0,7868	0,8015	0,8802
S3	kNN	With oversampling	0,5882	0,7353	0,8241	0,8560
S5	kNN	With oversampling	0,5876	0,8372	0,7990	0,9076
S6	kNN	With oversampling	0,5778	0,7647	0,8090	0,8694

Table 3. Best models achieving the highest values of sensitivity (Cross Validation).

Scenario	*Model*	*Approach*	*Precision*	*Sensitivity*	*Specificity*	*Accuracy*
S1	kNN	With oversampling	0,6484	0,9403	0,8246	0,9655
S2	kNN	With oversampling	0,6487	0,9278	0,8272	0,9585
S3	kNN	With oversampling	0,6654	0,8955	0,8452	0,9413
S5	kNN	With oversampling	0,6786	0,9403	0,8469	0,9661
S6	kNN	With oversampling	0,6891	0,9154	0,8580	0,9527

5. Discussion

From the analysis of the best results stood out that all of them had the model (kNN) and the approach (with oversampling) in common. The metric results were also improved when using the Cross Validation, yet *Scenario5* had the best results despite the sampling methods.

The positive impact of the oversampling on the results proves how unbalanced the dataset is. The reduced number of cases with perineal tear (14,53%), compared to the amount of cases without it (85,47%), is presented as a disadvantage to the algorithm learning process. The sum of different cases should at least be similar, otherwise the predictions of the data mining models, as it has happened, will be compromised.

Contrary to what happens with Cross Validation that uses the total information on the dataset, Holdout Sampling only uses 66% of the dataset to train the model. In addition, using Cross Validation, the tests will be performed on already known data. These two factors combined together may explain the lack of quality of the results presented by the sampling method of Holdout. S2, S3 and S4 scenarios were built with the purpose of finding out if perineal tear was dependent on factors directly related with the mother (S2), or with the baby (S3), or even with pregnancy aspects (S4). However, since the interest on this study is the proportion of positives (developing perineal tear) correctly identified – sensitivity –, *Scenario5* demonstrate that perineal tears are a result of a mother, baby and pregnancy attributes combination, putting aside the other hypothesis. The best models show that results are inflicted by some variables, like baby's weight, height and cephalic perimeter or the type of delivery. Nevertheless, S5 doesn't have the best proportion of negatives (not developing perineal tear) correctly identified – specificity. This is not a barrier since it is better to predict that a mother will have a perineal tear during labour and prepare her and try to avoid that than predicting the inverse situation, which means not taking any measures, and the mother ending up developing a laceration during child birth.

6. Conclusions and Future Work

This study aimed to predict, demonstrating the utility of clinical decision support systems, the incidence of perineal tear during child delivery allying real data with data mining models. Acceptable results were obtained when inducing k-Nearest Neighbours algorithm with oversampling and all the data for testing, achieving approximately 94% of sensitivity, 85% of specificity and 97% of accuracy. The inequality on the quantity of cases, existing on the dataset, became the biggest obstacle encountered to accomplish better results. Future work will include more cases of perineal tear in the predictive models and maybe other data mining techniques will be applied.

Acknowledgments

This work has been supported by Compete: POCI-01-0145-FEDER-007043 and FCT within the Project Scope UID/CEC/00319/2013.

References

1. Ward Manager. Perineal tears - information for patients & visitors. Northern Lincolnshire and Goole Hospitals, sep 2016.
2. Lawrence Leeman, Maridee Spearman, and Rebecca Rogers. Repair of obstetric perineal lacerations. *American family physician*, 68(8):1585-1590, 2003.
3. Marek J Druzdzel and Roger R Flynn. Decision support systems. encyclopedia of library and information science. a. kent. *Marcel Dekker, Inc. Last Login*, 10(03):2010, 1999.
4. Pedro Gago, Manuel Filipe Santos, Álvaro Silva, Paulo Cortez, José Neves, and Lopes Gomes. Intcare: a knowledge discovery based intelligent decision support system for intensive care medicine. *Journal of decision systems*, 14(3):241-259, 2005.
5. Eta S Berner. Clinical decision support systems: state of the art. *AHRQ publication*, 90069, 2009.
6. Filipe Portela, Manuel Filipe Santos, Álvaro Silva, Fernando Rua, António Abelha, and José Machado. Preventing patient cardiac arrhythmias by using data mining techniques. In Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on, pages 165-170. IEEE, 2014.
7. Enrico Coiera. Clinical decision support systems. *Guide to health informatics*, 2(1), 2003.
8. Christian Castaneda, Kip Nalley, Ciaran Mannion, Pritish Bhattacharyya, Patrick Blake, Andrew Pecora, Andre Goy, and K Stephen Suh. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics*, 5(1):4, 2015.
9. Filipe Portela, Manuel Filipe Santos, José Machado, António Abelha, Fernando Rua, and Álvaro Silva. Real-time decision support using data mining to predict blood pressure critical events in intensive medicine patients. In *Ambient Intelligence for Health*, pages 77-90. Springer, 2015.
10. Stéphane Tufféry and Rod Riesco. Data mining and statistics for decision making. 2011.
11. Eva Silva, Luciana Cardoso, Filipe Portela, António Abelha, Manuel Filipe Santos, and José Machado. Predicting nosocomial infection by using data mining technologies. In *New Contributions in Information Systems and Technologies*, pages 189-198. Springer, 2015.
12. Shakiba Khademolqorani and Ali Zeinal Hamadani. An adjusted decision support system through data mining and multiple criteria decision making. *Procedia-Social and Behavioral Sciences*, 73:388-395, 2013.
13. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37, 1996.
14. Samuel Valente, Jorge Braga, José Machado, Manuel Santos, and António Abelha. The impact of mobile platforms in obstetrics. *Procedia Technology*, 9:1201-1208, 2013.
15. wiseGEEK. http://www.wisegeek.org/what-is-an-obstetrician.htm. [Online; accessed 31-May-2017].
16. Sónia Pereira, Filipe Portela, Manuel Filipe Santos, José Machado, and António Abelha. Predicting type of delivery by identification of obstetric risk factors through data mining. *Procedia Computer Science*, 64:601-609, 2015.
17. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.
18. Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128-150, 2017.
19. Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1-37, 2008.
20. Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5):352-359, 2002