MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Data Mining as a Tool of Quality in Obstetrics

Diana Lungeanu[a], Stefan Holban[b], Daniela Zaharie[c], Dan B. Navolan[d], Raul F. Horhat[a], Calin Muntean[a], Anca Kigyosi[a, d]

[a] *Department of Medical Informatics, University of Medicine and Pharmacy of Timisoara, Romania*
[b] *Department of Computer Science, Polytechnic University of Timisoara, Romania*
[c] *School of Mathematics and Informatics, West University of Timisoara, Romania*
[d] *University Hospital of Obstetrics-Gynaecology „Dr. Dumitru Popescu" from Timisoara, Romania*

Abstract and objective

There are large amounts of data about patients and their medical condition and evolution, which are stored in electronic databases or simply in archives. These deposits of information contain important nugets which have not been explored. We present our approach in using data mining techniques to investigate the factors associated with care quality and obstetrical risk factors. We also aim at developing suitable models and prediction systems for the two related aspects in obstetrical care.

Keywords:

medical data mining, quality models, obstetrics

Introduction

Following the international trend in health care, the long term strategy of the two Romanian official authorities in the field (the Ministry of Health and the National House for Health Insurance) states that by 2008 there should exist a reliable system for quality assessment in health care and it should serve as a basis for competition on the market. Taking this into consideration, we started a project focused on obstetrical care aiming at identifying the characteristics of our population, so being able to develop a model and a prediction system for this concrete population and environmental context.

Methods

Defining quality measures is not a trivial task and the usual approach is to use a set of measures focused on medical care processes and outcomes [1]. In Romania, we use mainly those available from the DRG system, focused on the outcomes, very limited and difficult to link to the process as they are based on a limited number of 'rate-based' indicators (i.e. occurrence rates of desired or undesired events) [1, 2]. Our approach is to combine pieces of data known as being directly related to quality of obstetrical services (based on criteria like importance as indicators, scientific soundness, and feasibility) [2], and use data mining tools to relate them to other assessments and plans of intervention [3].

Preliminary results and discussion

The work presented in this poster is part of a project aiming at identifying the factors associated with obstetrical outcomes and pre-term risk factors. We began by statistically analyzing reported outcome data from two university hospitals in distant geographic locations of each other, trying to identify patterns of diagnoses (based on the ICD-10 system compulsory in our country). Table 1 presents some raw statistics we obtained, which represented the starting point for further investigations.

Table 1 - Comparison between university Hospital B and university Hospital T (including the limited EHR data set for which we analyzed the pre-term birth risk factors). The ICD-10 codes specified represent the primary DRG reported codes

Year 2006	Hospital B DRG report	Hospital T DRG report	Hospital T EHR data set
No of new born babies	413	2864	211
No of pre-term babies	47 (11.4%)		14 (6.6%)
No of pregnant women (not necessarily distinct cases)	587	4560	211
ICD-10 O80 and O83 vaginal delivery	250 (42.5% of total ob-gyn cases)	2232 (49% of total ob-gyn cases)	157 (74% of deliveries)
ICD-10 - O82.0 elective caesarian delivery	4 (0.7%)	6 (0.1%)	54 (26%)
ICD-10 - O82.1 emergency caesarian section	74 (12.6%)	903 (19.8%)	0
ICD-10 - P03.4 newborn affected by caesarian delivery	40 (10.9% of full-term babies)	263 (9.3%)	22 (11.2% of full-term babies)

The discrepancies between the two hospitals in percent of caesarian section deliveries, as well as between the DRG reported data and the actual data existing in the small internal EHR data set could be related to both the insurance considerations (medical doctors' financial incentives to assign codes for reimbursement purposes combined with little experience in appropriately using the coding) and the patients' preferences or demands (convenience reasons or care for the well-being of the baby). We also tried to calculate some of the proposed performance indices for measuring performance [4, 5].

In parallel, trying to identify the targets for our further efforts in developing models for the obstetrical care outcomes, we focused on the factors associated with spontaneous preterm labour [6, 7, 8]. We have tried a selection of the significant attributes from the EHR database at our university hospital, processing them with Weka 3.4.8, some regressional models with DataFit 7.1.44, and applying a comparative analysis based on the following methods:

M1: a filter type ranking of attributes based on the informational gain (IG) [9];

M2: a filter type ranking of attributes based on the symmetrical uncertainty (SU) [9];

M3: a ranking based on the coefficients obtained by logistic regression (LR); in this case, the higher was the absolute value of a regression coefficient, the higher was the corresponding attribute's relevance;

M4: a ranking method based on the evolutionary multi-objective approach (EMOA) proposed in [10].

Although we have succeeded in identifying only 14 perinatal attributes as contributing to the pre-term delivery risk and the findings mainly confirmed what we already knew from literature or clinical experience [10], the classification results were encouraging (Table 2).

Table 2 - Classification results for attributes subsets selected based on different rankings

Ranks of selected attributes	Correct classification ratio (%)		
	M1(IG), M2(US)	M3 (LR)	M4 (EMOA)
1-14	77.96	77.96	77.96
1-13	78.53	78.53	79.09
1-12	76.27	76.83	77.40
1-11	76.27	77.4	78.53
1-10	78.53	79.09	80.79
1-9	78.53	78.53	80.79
1-8	79.09	80.79	80.22
1-7	79.66	80.22	80.79
1-6	77.96	79.66	81.92
1-5	76.83	81.35	81.92
1-4	78.53	82.48	83.05
1-3	76.83	81.35	82.48
1-2	80.79	81.92	81.92
1	79.09	79.09	79.09

Conclusions

A major problem consisted of the data themselves, which were sparse and incomplete (out of more than 2000 records we started with, we could effectively use only 211 for the pre-term risk model) and in the reported DRG data there were confounding factors in estimating the quality of the perinatal care.

At this stage, we concluded we should reconsider the data collecting methodology as better models require more discriminative clinical information. Despite the difficulties, data mining offers useful tools in this research endeavour.

Acknowledgements

This work is supported by grant 99-II CEEX 03 - INFOSOC 4091/31.07.2006 from the Romanian Ministry of Education and Research. We kindly thank Dr. Maria Bari from the University of Medicine and Pharmacy of Bucharest for her valuable contribution to this project.

References

[1] Committee on Quality of Health Care in America, Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press, 2001.

[2] Marshall M, Leatherman S, Mattke S. OECD working paper no. 16: Selecting indicators for the quality of health promotion, prevention and primary care at the health systems level in OECD countries. 2004: *http://www.oecd.org/dataoecd/27/52/33865865.pdf* (Last access 28 May 2007).

[3] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26: 1-24.

[4] Spangler WE, May JH, Strum DP, Vargas LG. A data mining approach to characterizing medical code usage pattern. Journal of Medical Systems 2002; 26(3): 255-275.

[5] Neumann A, Holstein J, Le Gall JR, Lepage E. Measuring performance in health care: case-mix adjustment by boosted decision trees. Artificial Intelligence in Medicine 2004; 32: 97-113.

[6] Goodwin L, Maher S. Data mining for preterm birth prediction. Proceedings of the 2000 ACM Symposium on Applied Computing (SAC'00): 46-50.

[7] Lovell DR, Rosario B, Niranjan M, Prager RW, Dalton KJ, et al. Design, construction and evaluation of systems to predict risk in obstetrics. International Journal of Medical Informatics 1997; 46: 159-173.

[8] Resnik R. Issues in the management of preterm labor. J. Obstet. Gynaecol. Res. 2005; 31(5): 354-358.

[9] Witten IH, Frank E. Data mining. Practical machine learning tools. Berlin: Elsevier, 2005.

[10] Zaharie D, Holban S, Lungeanu D, Navolan D. A computational intelligence approach for ranking risk factors in preterm birth. 4th International Symposium on Applied Computational Intelligence and Informatics SACI 2007; in press.

Address for correspondence

Diana Lungeanu. Department of Medical Informatics, University of Medicine and Pharmacy, P-ta E. Murgu 2, 300041 Timisoara, Romania. Tel/fax +40-256-490288; *dlungeanu@gmail.com*