SIEMENS

Pursuing clinical data quality improvements through the use of electronic health records for coding support

José C. Ferrão[1,2], Mónica D. Oliveira[2], Filipe Janela[1], Henrique M. G. Martins[3]

www.siemens.pt/healthcare

Impresso em Portugal | 05.2014 © Siemens, S.A. | HSOOOS16200109670

Introduction

- Electronic health record (EHR) systems produce large volumes of data, with great (and unexplored) potential for decision support and clinical research
- Clinical coding is the groundwork for provider financing and developing efficiency, quality of care and patient safety studies
- Clinical data quality has impact in the quality of coding and, therefore, on the development of studies based on episode data reported with ICD codes
- Clinical coding is a complex process – its partial automation using EHR data may help mitigate workload and improve quality of ICD databases
- Structured data promotes data standardization and allows circumventing most issues of unlocking information stored in free-text format, although it also entails challenges of data quality, data extraction/processing and model development

Objectives

- Develop a framework to automatically extract and process data from a structured EHR system in order to allow the development of prediction models
- Employ strategies to reduce data dimensionality and handle the existence of multiple codes per episode
- Build prediction models based on structured EHR data to support the assignment of ICD-9-CM codes
- Evaluate and compare performance of 4 prediction models: decision trees, naïve Bayes, logistic regression and support vector machines in assisting clinical coding
- Investigate codes exhibiting lower predictive power in order to identify critical aspects influencing clinical data quality and, consequently, model predictive power

Materials and Methods

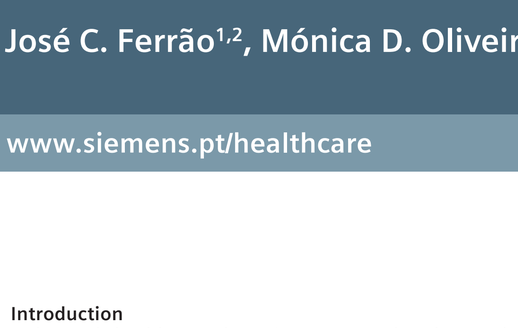

Figure 1 – Data contents of the EHR system Soarian® used in this study

Dataset

- *5089 inpatient episodes from medical wards (HFF)*
- *4820 features (variables) with non-missing values*
- *2272 unique, highly imbalanced ICD-9-CM codes*
- *Analysis focused on the top 50 ICD codes (around 50% of total code occurrences)*

Code order	Description	Code	Frequency
1	*Hypertension NOS*	*401.9*	*40,15%*
2	*Hyperlipidemia NEC/NOS*	*272.4*	*17,72%*
3	*Atrial fibrillation*	*427.31*	*17,55%*
...	...	...	...
50	*Obstr chronic bronchitis w/o exac.*	*491.20*	*3,14%*

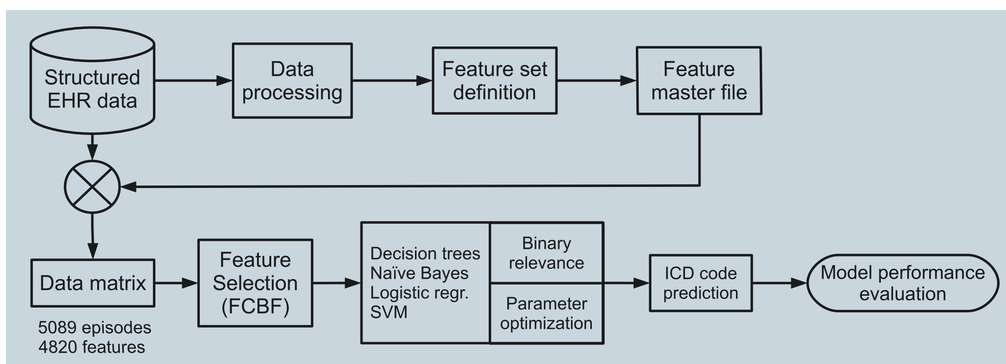

Figure 2 – Methodological workflow developed in this study, including the stages of data extraction and processing, feature definition and selection, testing prediction models and comparing results

Results

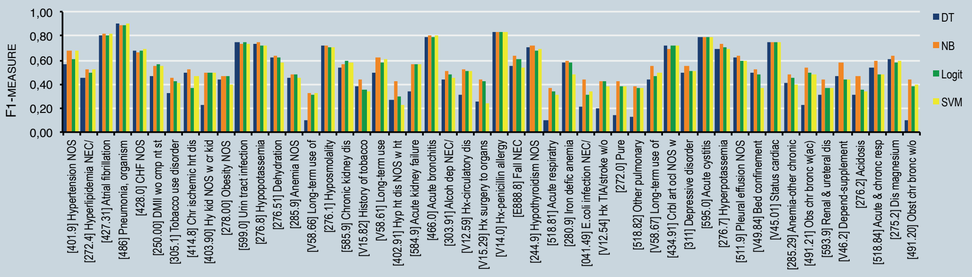

Figure 3 – Precision, recall and F1-score values obtained for the 50 most frequent ICD-9-CM codes with the 4 tested prediction models

Evaluation Metrics

Precision

$$P_i = TP_i / (TP_i + FP_i)$$

Recall

$$R_i = TP_i / (TP_i + FN_i)$$

F_1-score

$$F1_i = 2 P_i R_i / (P_i + R_i)$$

TP – True Positives; FP – False Positives; FN – False Negatives

Key results

- The tested models are able to support clinical coding to different extents, with decision trees exhibiting higher precision, SVM showing lower recall, and Bayesian and logistic models exhibiting higher overall performance
- Variables selected with the FCBF method to reduce dimensionality were mostly clinically meaningful, although some unexpected variables are used for model development, likely due to statistical effects captured by the method
- Mixing different levels of specificity of clinical concepts (e.g. in assigned diagnoses or prescriptions) seems to be closely associated with lower model predictive power
- Ambiguity and heterogeneity in data recording practices of clinicians, namely upon choosing between general and specific conditions and using modifiers (e.g. benign/malignant/unspecified), also seem to be related with lower model results
- Structured EHR systems often capture redundant data, which leaves room for ambiguities and inconsistencies in database entries
- Multiple variables exhibit high missing rates, which precludes their use in coding support in cases where they might be relevant for coding

Conclusions

- Our methodology sheds light onto the use of structured EHR data for coding support
- Data quality issues directly influence model results and likely have negative impact on quality of coding, consequently influencing financing and care quality studies
- Structured EHR systems, although rendering clinical information usable for clinical research and quality studies, demand coherent configuration throughout the entire system and the incorporation of mechanisms to improve coherence in the level of specificity used by clinicians in recording clinical data

Acknowledgments

This work has received financial support from Fundação para a Ciência e a Tecnologia (SFRH/BDE/51605/2011), Siemens SA and the Centre for Management Studies of Instituto Superior Técnico (CEG-IST, University of Lisbon). The authors wish to thank colleagues from Hospital Professor Doutor Fernando Fonseca for close collaboration and availability throughout this research.

Bibliography

I. Stanfill MH et al., "A systematic literature review of automated clinical coding and classification systems" Journal of the American Medical Informatics Association, vol. 17, no. 6, pp. 646-51, 2010
II. Bishop CM., Pattern Recognition and Machine Learning, Singapore: Springer; 2006
III. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 48(1):38–44

[1] Healthcare Sector, Siemens S.A., Rua Irmãos Siemens 1, Amadora, Portugal
[2] CEG-IST, Centro de Estudos de Gestão do Instituto Superior Técnico, University of Lisbon, Av. Rovisco Pais 1, Lisboa, Portugal
[3] Centro de Investigação e Criatividade em Informática, Hospital Prof. Doutor Fernando Fonseca, IC-19 Venteira, Amadora, Portugal

TÉCNICO LISBOA

Answers for life.