Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-419

Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes

Mehdi DJENNAOUI[a,1], Grégoire FICHEUR[a], Régis BEUSCART[a] and Emmanuel CHAZARD[a]
[a] *Department of Public Health, EA 2694, University of Lille, F-59000 Lille, France.*

Abstract.
Introduction. Diagnoses and medical procedures collected under the French system of information are recorded in a nationwide database, the "PMSI national database", which is accessible for exploitation. Quality of the data in this database is directly related to the quality of coding, which can be of poor quality. Among the proposed methods for the exploitation of health databases, data mining techniques are particularly interesting. Our objective is to build sequential rules for missing diagnoses prediction by data mining of the PMSI national database.
Method. Our working sample was constructed from the national database for years 2007 to 2010. The information retained for rules construction were medical diagnoses and medical procedures. The rules were selected using a statistical filter, and selected rules were validated by case review based on medical letters, which enabled to estimate the improvement of diagnoses recoding.
Results. The work sample was made of 59,170 inpatient stays. The predicted ICD codes were E11 (non-insulin-dependent diabetes mellitus), I48 (atrial fibrillation and flutter) and I50 (heart failure).We validated three sequential rules with a substantial improvement of positive predictive value:
{E11,I10,DZQM006}=>{E11}
{E11,I10,I48}=>{E11}
{I48,I69}=>{I48}
Discussion. We were able to extract by data mining three simple, reliable and effective sequential rules, with a substantial improvement in diagnoses recoding. The results of our study indicate the opportunity to improve the data quality of the national database by data mining methods.

Keywords. Electronic Health Records, Decision Support Techniques, Data mining, Nationwide Database.

Introduction

The main purpose of electronic health records (EHRs) is to support health care of the patient. However, they are increasingly used by medical research as part of data reuse [1-3]. Parts of EHR data are generated automatically or in the course of medical management. Other data, such as medical diagnoses, are collected and not used for

[1] Corresponding Author.

patient care, but for payment or epidemiological purposes. The collection of such data often depends on the knowledge and interpretation of coding rules by coders, and on the time they accept to dedicate to this task. Health data can be of poor quality [4,5], especially in case the encoded data are not useful for patient care. To our knowledge, this problem is more important regarding medical diagnoses.

Diagnoses and medical procedures collected under the French hospital payment system are recorded in a nationwide database called the "PMSI national database" [6]. This database is accessible for research exploitation, with the possibility to link information from a single patient, through a process of anonymous chaining. The quality of the data in this database is directly related to the quality of coding. Many control procedures of this coding process exist; most of them rely heavily on identifying diagnoses in medical letters. With millions of past inpatient stays, the national database falls within the definition of big data, which requires appropriate operating tools [7]. Among the proposed methods, data mining techniques are particularly interesting; they are increasingly used for the exploitation of health databases [8,9]. The objective of this work is to build diagnoses prediction rules from the national database by data mining. These rules would use information from the patient's previous stays to automatically predict diagnoses of the studied stay.

1. Methods

Our working sample was constructed from the PMSI national database for years 2007 to 2010 (n=94,862,015 inpatient stays). The information retained for rules construction were medical diagnoses encoded according to the 10th revision of the International Classification of Diseases (ICD10) [10], and medical procedures encoded according to the Common Classification of Medical Procedures (the French CCAM). The working sample was then filtered so that all the inpatient stays could be related to patients who had at least one stay in a community hospital from the North of France. In this hospital, we could access the complete inpatient stay and free-text anonymous medical letters in the frame of the permissions granted to the PSIP (Patient Safety through Intelligent Procedures in medication) project [11]. This was necessary for the validation step as described below. The sample was split in two samples: a learning sample (70%) for the creation of rules and a test sample (30%) for validation of the rules.

The French version of the ICD10 is made of a set of codes and labels (n=33,816). The first 3 characters of the codes stand for code categories (n=2,049) and are usually used for code prediction. From this point, "diagnostic codes" will denote the 3-digits truncated ICD10 codes. Diagnostic codes to predict should be frequent, characterize chronic diseases and have a strong impact on hospital payment, assuming that these codes are most often coded. To select these diagnostic codes, we based on the results of the Valodiag study [12] that ranked diagnostic codes according to their average impact on payment and their frequency.

As a data mining method, we built sequential rules. Sequential rules are expressed using the syntax {A}=>{B}, {A} constituting the predictive pattern, which stands for a set of one or several items, and {B} the predicted item, the predictive pattern preceding in time the predicted item. The rules generation is performed by algorithms based on the notions of support and confidence; support in this context is the proportion of stays with the rule on the set of stays, while confidence is the proportion of stays containing the predicted item on the set of stays expressing the predictive pattern. The user

determines thresholds in advance for these two parameters. For our study, we set the threshold of support to 0.00075 and the threshold of confidence to 0.5. Sequential rules were constructed using R software [13], with a function implementing the reference algorithm SPADE [14,15]. The selection of rules was made by mean of a statistical filter, based on the value of the product (support*confidence) to obtain the best trade-off between these two values. Evaluation of the rules was made by a review of cases using medical letters. So that this assessment is the closest real working conditions, it was bound to be based on medical mails. For this, we extracted for each rule from the test sample all the stays that met the following condition: for a patient stays sequence, the predictive pattern had to be present at a given stay *n* and the predicted code had to be missing at the next stay *n+1*, i.e. $\{A\} \not\Rightarrow \{B\}$. The careful reading of medical letters enabled to estimate the proportion of stays among which the diagnostic code was really missing, defining the positive predictive value (PPV) of the predictive pattern. The same method was applied for evaluating the sole predicted code as a predictive pattern of its own presence in the next stay, i.e. $\{B\} \not\Rightarrow \{B\}$. We then calculated the lift of each rule, according to the following formula (1).

$$lift = PPV \text{ of the predictive pattern} / PPV \text{ of the sole diagnostic code} \quad (1)$$

2. Results

The work sample was made of 59,170 inpatient stays corresponding to 12,125 different patients. The predicted codes were E11 (non-insulin-dependent diabetes mellitus), I48 (atrial fibrillation and flutter) and I50 (heart failure). Six sequential rules were extracted from the learning sample (40,876 stays) (Table I).

Table I. Sequential rules in the form "Predictive pattern → Predicted code", with Support and Confidence.

Predicted codes	Predictive patterns	Wordings	Support	Confidence
E11	E11	non-insulin-dependent diabetes mellitus	0.0717	0.55
	I10 , DZQM006 , E11	essential (primary) hypertension transthoracic heart Doppler ultrasound non-insulin-dependent diabetes mellitus	0.0069	0.71
	I10 , I48 , E11	essential (primary) hypertension atrial fibrillation and flutter non-insulin-dependent diabetes mellitus	0.0052	0.72
I48	I48	atrial fibrillation and flutter	0.0571	0.51
	I10 , I48 , E78	essential (primary) hypertension atrial fibrillation and flutter disorders of lipoprotein metabolism	0.0035	0.60
	I10 , I48 , Z95	essential (primary) hypertension atrial fibrillation and flutter presence of cardiac and vascular implants	0.0035	0.60
	I48 , I69	atrial fibrillation and flutter sequelae of cerebrovascular disease	0.0037	0.62
I50	I50	heart failure	0.0321	0.37
	I10 , I48 , I50	essential (primary) hypertension atrial fibrillation and flutter heart failure	0.0037	0.50

Rules validation from the test sample (18,294 stays) by estimating the lift enabled to validate three prediction rules (Table II).

Table 2. Lift of the complete rule pattern: PPV of the predictive pattern divided by PPV of the sole past code.

Predictive patterns	Wordings	Nb stays reviewed	PPV	Lift
E11	non-insulin-dependent diabetes mellitus	117	0.53	(reference)
I10 , DZQM006 , E11	essential (primary) hypertension transthoracic heart Doppler ultrasound non-insulin-dependent diabetes mellitus	32	0.69	**1.30**
I10 , I48 , E11	essential (primary) hypertension atrial fibrillation and flutter non-insulin-dependent diabetes mellitus	20	0.75	**1.42**
I48	atrial fibrillation and flutter	92	0.30	(reference)
I10 , I48 , E78	essential (primary) hypertension atrial fibrillation and flutter disorders of lipoprotein metabolism	16	0.25	0.83
I10 , I48 , Z95	essential (primary) hypertension atrial fibrillation and flutter presence of cardiac and vascular implants	25	0.24	0.80
I48 , I69	atrial fibrillation and flutter sequelae of cerebrovascular disease	23	0.39	**1.30**
I50	heart failure	70	0.21	(reference)
I10 , I48 , I50	essential (primary) hypertension atrial fibrillation and flutter heart failure	37	0.21	1.00

3. Discussion

Consistent with the main purpose of our study, we validated three sequential rules for diagnosis prediction. These rules have a higher positive predictive value than a simple renewal of previously encoded diagnoses, with a lift greater than or equal to 1.30. The three rules are:

- {E11,I10,DZQM006}=>{E11}
- {E11,I10,I48}=>{E11}
- {I48,I69}=>{I48}

The recoding improvement with an increase in the PPV is necessarily associated with a decrease in recall, which does not reflect the evaluation by the lift. However, this pitfall is acceptable if we consider the context for the recoding, which is marked by a lack of time; then the benefit provided by the prediction rules used for automatic recoding offsets loss of recall. Other pitfall, we could not retain rules for code I50 (heart failure). This can be explained by the fact that this code was less often coded, hence producing less trustworthy rules.

Despite these limitations, we were able to extract three simple, reliable and effective sequential rules, enabling for a substantial improvement in recoding diagnoses. The originality of our work lays in the use of data mining methods for the development of control rules, which contrasted with conventional approaches of control, mainly based on the identification of diagnoses by reading medical letters. The evaluation of the rules by returning to medical letters allowed an accurate and objective assessment, with testing these rules as a tool for recoding in real situations of monitoring and improvement data quality. The assessment by the lift was well aware of the improvement of recoding.

The results of our study indicate the opportunity to improve the data quality of the national database by data mining methods. Extending prediction to other diagnostic codes would be an interesting perspective.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under Grant Agreement n°216130 – the PSIP project.

References

[1] Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. juin 2012;13(6):395‑405.
[2] Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data : A transnational perspective. Int J Med Inf. 2013;82(1):1‑9.
[3] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144‑51.
[4] Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. juill 2012;50 Suppl:S21‑9.
[5] Richesson RL, Horvath MM, Rusincovitch SA. Clinical research informatics and electronic health record data. Yearb Med Inform. 2014;9(1):215‑23.
[6] Agence Technique de l'Information Hospitalière ATIH. http://www.atih.sante.fr/l-atih/presentation
[7] Bacardit J, Widera P, Lazzarini N, Krasnogor N. Hard Data Analytics Problems Make for Better Data Analysis Algorithms: Bioinformatics as an Example. Big Data. 1 sept 2014;2(3):164‑76.
[8] Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag JHIM. 2005;19(2):64‑72.
[9] Bailey S, Singh A, Azadian R, Huber P, Blum M. Prospective data mining of six products in the US FDA Adverse Event Reporting System: disposition of events identified and impact on product safety profiles. Drug Saf Int J Med Toxicol Drug Exp. 1 févr 2010;33(2):139‑46.
[10] WHO International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/
[11] Beuscart R, McNair P, Brender J, PSIP consortium. Patient safety through intelligent procedures in medication: the PSIP project. Stud Health Technol Inform. 2009;148:6‑13.
[12] Ficheur G, Genty M, Chazard E, Flament C, Beuscart R. Proposition d'une méthode automatisée calculant la valeur moyenne d'un diagnostic. Rev DÉpidémiologie Santé Publique. 2013;61:S18‑9.
[13] R Core Team. R: A language and environment for statistical computing.2013.http://www.R-project.org/.
[14] Zaki MJ. SPADE: An efficient algorithm for mining frequent sequences. Mach Learn. 2001;42:31‑60.
[15] Hahsler CB and M, Diaz with contributions from D. arulesSequences : Mining frequent sequences 2014. http://cran.r-project.org/web/packages/arulesSequences/index.html