31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

REFEROCOD: A probabilistic method to medical coding support

L. Lecornu, G. Thillay, C. Le Guillou, P. J. Garreau, P. Saliou, H. Jantzem, J. Puentes, and J. M. Cauvin

Abstract— **Choosing diagnosis codes is a non-intuitive operation for the practitioner. Mistakes are frequent with severe consequences on healthcare evaluation and funding. French physicians have to assign a code for everything they do and they are not spared with these kinds of errors. We propose a tool named REFEROCOD to support the medical coding task in order to minimize errors without losing time, by suggesting a list of codes in accordance with the physician activities and of the patient medical context. The proposed method uses probabilistic knowledge and indicates the probability to have a proper diagnosis code considering the realized procedure, age, sex and other information available in the discharge abstract.**

I. Introduction

In the French medical information system, each procedure and diagnosis must be coded in the discharge abstract. Since 2004, the hospital funding has been directly calculated according to this information. However, coding is complicated and usually considered by physicians as a boring task. Since coding mistakes are common, each hospital may suffer the consequences of these errors, which can lead to wrong statistics and insufficient funding. Thus, the proposed method is directly related to coding support tools for medical procedures and diseases. The aim of this tool is to indicate to the physician which are the closest possible diagnosis codes and their relevance according to patient history in terms of procedures, sex, and other specifications. The proposed method is based on large amounts of anonymized discharge abstracts, so called RSA in French for anonymised and standardized discharge abstracts, which have been gathered previously and usually contain numerous similar histories to the current case. A data mining method was applied on this base to identify the searched codes. First, links between information in the RSA database are modeled to discover knowledge. Second, using this knowledge a list of codes is suggested to the physician. After presenting the study context and the state of art (section 2), the developed method (section 3) that permits to predict diagnosis codes will be described. Some results will be presented (section 4) before the conclusion (section 5).

This work was supported by a TECSAN/ANR project under the name Med1dex associating Brest CHU, Telecom Bretagne, INSERM Unit U650 LaTIM and PRISMEDICA company.

L. Lecornu, G. Thillay and J. Puentes are with INSTITUT Telecom; Telecom Bretagne UEB; Dpt ITI, Brest, France. (email Laurent.Lecornu@telecom-bretagne.eu)

C. Le Guillou, P. J. Garreau, P. Saliou, H. Jantzem, and J. M. Cauvin are with CHU Brest, Medical Information Departement, Brest, France.

L. Lecornu, C. Le Guillou, P. J. Garreau, J. Puentes and J. M. Cauvin are with Inserm, U650, IFR 148 ScInBioS, Brest, France

II. Context

A brief description of the healthcare information system and main database elements is proposed here.

A. Healthcare information system

As for many western countries, the French healthcare information system measures and funds hospitals on the basis of standardized information [1]. This information is gathered in the discharge abstract, and all these data are transmitted after anonymisation, to government health services.

Only hospitalizations are concerned with the discharge abstract, as described in the table 1. The hospital discharge abstract is a set of elementary discharge abstracts which gather information from all medical units that provided healthcare during the patient stay. Each abstract contains patient demographics (age, sex), and a main diagnosis in accordance with resources consumption, as well as associated diagnoses for related diseases and adverse events.

TABLE 1 MAIN ELEMENTS OF THE UNIT DISCHARGE ABSTRACT

Primary keys	
Birth date	
Sex	
Medical unit	
Admission date and modalities	Discharge date and modalities
Primary Diagnosis	Secondary Diagnosis
Severity Simplified Index	
List: Associated Diagnosis Codes, 1…N	
List: Diagnostic and therapeutic procedure codes (1,…N)	

The International Statistical Classification of Diseases and Related Health Problems – 10th revision (ICD-10), is used for coding the diagnosis part: diseases, traumatisms, symptoms, and other reasons for using health service [3]. ICD-10 is published and maintained by the World Health Organization and is used in many countries, mainly for registering morbidity, mortality causes, and for organizing healthcare services. The ICD-10 gathers nearly 15000 entries classified in 22 chapters and has been updated by the French Health Ministry. Diagnostic codes are composed of a letter followed by 2 to 4 digits.

The Common Classification of Medical Procedures (CCMP) [2] is a French nomenclature, which describes

medical procedures. The CCMP code is composed by four letters followed by three digits. Each letter has a signification, which allows knowing a part of the technical procedure context.

CCMP and ICD-10 propose a hierarchical organization from chapter to subchapter and from paragraph to subparagraph, permitting to code for both CCMP and ICD-10, and from chapter to category. This tree organization is consistent with medical knowledge and allows handling information at different aggregate levels.

The discharge abstract is analyzed first by an algorithm applied at national level and then classified in diseases related groups, based on which, decisions about hospitalization fees are taken. This classification procedure according to the prevalence of diagnosis codes groups has a noticeable effect on healthcare fares: statistics covering the whole country are continuously estimated to determine those fees. As a consequence, physicians should attach particular attention to notifications concerning classifying procedures changes.

B. Problem definition

The introduction of abounding codes through the CCMP and the ICD-10 unavoidably induces coding errors. Indeed, a missing or a wrong code may alter a patient fee by thousands Euros. For this reason potential sources of errors at each step of the described inpatient International Classification of Diseases coding process have been examined [4]. Moreover, these errors are common and concern about 20% of discharge abstracts [5]. As a consequence, it is becoming critical to limit these errors and to optimize the discharge abstract from a financial evaluation point of view. Given the fact that coding diagnosis is more difficult than coding procedures, the aim of our work is to help physicians to code diagnosis, by predicting the most appropriate ICD-10 codes through the REFERECOD tool.

C. State of the art

The investigated problem belongs to the prediction domain. It is consequently important to examine the two categories of prediction methods [6]: qualitative and quantitative. Qualitative methods are rather instinctive and based on a subjective hypothesis, which can depend on past data, or not. On the other hand, quantitative methods are based instead on statistical and/or mathematical models. Once the underlying model has been chosen, predictions are automatically calculated and can be reproduced by anyone. For these reasons we decided to choose a combination of quantitative methods. Among existing approaches – probabilistic, Bayesian network, and neural network – the probabilistic method was found to be more adapted to the large size of the learning database.

Having access to a considerable amount of data, it was thus possible to identify and count on the given learning database, available diagnosis-procedures pairs, necessary to calculate thereafter the required frequency of occurrence probabilities, according to a certain number of entries.

III. CODING SUPPORT METHODS

A. Principles

Our main objective is to display a proposed list of codes depending on the related evidence, from the most to the less pertinent, in order to decrease the amount of potential coding errors, giving the user an easy choice. According to this probabilistic approach, results are classified according to their frequency of occurrence probability.

The proposed method is based on conditional probabilities: considering an event E and events R_i, the probability of E given R_i is defined by:

$$P\left(E_{|R_1,\ldots,R_n}\right)=\frac{P\left(R_1\cap\ldots\cap R_n\cap E\right)}{P\left(R_1\cap\ldots\cap R_n\right)} \quad (1)$$

Obtained results become more relevant when the database size increases, given that the estimator uses computed values from a sample of the whole population, or from random results of the experiment.

Estimation by frequency of occurrence is suitable to determine an event probability, when there is a finite number of possible events and the experiment can be independently reproduced many times.

As our database has a sufficient size (more than 2 millions clinical stays), it is thus possible to approach this conditional probability by means of equation (1). Counting every occurrence of each event E and each event R_i, we can find an estimation of the probability $\hat{P}\left(E|R_1,\ldots,R_n\right)$ with n entries R_i that gives the following equation:

$$\hat{P}\left(E|R_1,\ldots,R_n\right)=\frac{occ\left(E,R_1,\ldots,R_n\right)}{occ\left(R_1,\ldots,R_n\right)} \quad (2)$$

where *occ(x)* defines the number of times that the variable x is present in the database.

B. Probability fusion

Let two information sources (or sensors) (respectively C_1 and C_2) giving each probabilistic information of a D_1 diagnosis appearance (respectively $P_{C_1}(D_1)$ and $P_{C_2}(D_1)$). In order to take a decision, it is necessary to merge these two pieces of information. Grandin [7] indicates several methods to carry out this fusion, consisting mainly in combining each sensor result according to its intrinsic performance, i.e. $P(D_1)=\alpha_1\cdot P_{C_1}(D_1)+\alpha_2\cdot P_{C_2}(D_1)$. We applied this approach to merge the diagnosis probability of each information source.

C. Proposed method

In order to use the described probabilistic approach to implement REFEROCOD, a database containing hospital data with more than one million of anonymous discharge abstracts was provided, providing the required data volume for relevant probabilities computation.

Such a database may seem quite adequate. However,

knowing that the CCMP describes more than 7600 procedures and the ICD-10 over 18000 diagnoses, combined with some additional criteria like sex or age, more than 10^{12} different combinations become possible.

Even if the procedure is limited to CCMP subparagraphs and the first three letters of ICD-10 diagnoses, the number of theoretically possible combinations is still over 5.10^8.

If all those combinations were taken into account the database would become insufficient. Nevertheless most of them, besides being theoretical, are impossible at a practical level. This fact highly reduces the number of real possible combinations and makes possible to apply the law of large numbers.

In order to predict possible diagnoses, it was initially decided to use simultaneously all information sources. The ideal method would be to estimate the conditional probability to have a diagnosis knowing the age, sex, stay length, medical unit, and the already coded procedure and diagnoses. However, this probability cannot be estimated directly. This is why it was decided afterwards to distinguish four medical information sources: (1) age, sex, and stay length; (2) medical unit; (3) procedures; and (4) diagnoses already coded. As a consequence, for each of these information sources, we separately estimated the respective diagnosis probability. Then, we merged them according to each source performance.

Moreover, some technological constraints were imposed to handle data: the probability estimation must be saved on a Microsoft Access database and the time to find an estimation had to be very short (less than 3 seconds).

- Diagnosis probability according to three inputs (age, sex, stay length) with customizable ranking

Using equation (5), the probability estimation corresponding to $\hat{P}\left(D_j \middle| age, sex, stay\ length\right)$is computed. This assessment was done for every diagnosis, age segments, sex, and stay length of the anonymized discharge abstract database.

- Diagnosis probability according to the medical unit

In a second step, the diagnosis probability is estimated according to the medical unit, reflecting each specific hospital organization and the usual activity of medical units. For evaluation tests, the diagnostic probabilities were calculated according to the coding profile of the medical units of Brest university hospital.

- Diagnosis probability according to procedures

In a third step, every coded procedure is used to estimate a diagnosis probability. The number of coded procedure varies from 0 to *N*. Because of combination limits, it is not possible to determine the independence between procedures (from a statistical point of view, it is very likely that some procedures could be mutually dependant, while others would not exhibit such feature).

The diagnosis probability according to a unique procedure $\hat{P}\left(D_j \middle| proc_i\right)$ has been estimated applying (2). To merge the different pieces of information coming from each procedure, it is assumed that every procedure has the same importance. Then, procedures are distinguished according to their classifying effect on the disease related group:

$$\hat{P}\left(D_j \middle| proc_1,...,proc_N\right) = \alpha_1 \cdot \frac{1}{L} \sum_{k=1,\ proc_k \text{ is classified}}^{L} \hat{P}\left(D_j \middle| proc_k\right) + \alpha_2 \cdot \frac{1}{M} \sum_{k=1,\ proc_k \text{ is unclassifiedt}}^{M} \hat{P}\left(D_j \middle| proc_k\right) \quad (3)$$

with $L + M = N$ and α_1 and α_2 depending on the classifying procedure performance compared with non-classifying one. Each procedure added value is tested afterwards.

- Diagnosis probability according to diagnoses

During the fourth step, every coded diagnosis is used to estimate diagnosis probability. Coded diagnoses vary from 0 to *M*. As in the procedures case, it is not possible to determine the statistical independence between diagnoses, since some would be mutually dependant and others would not.

The diagnosis probability according to a unique diagnosis $\hat{P}\left(D_j \middle| D_i\right)$ was estimated in the same way as in (2). Again, in order to merge each piece of information coming from each diagnosis, it was assumed that each diagnosis has the same importance, before fusing them by means of:

$$\hat{P}\left(D_j \middle| D_1,...D_M\right) = \frac{1}{M}\sum_{k=1}^{M} \hat{P}\left(D_j \middle| D_k\right) \quad (4)$$

- Diagnosis probability according to the four inputs

$$\begin{aligned}\hat{P}\left(D_j\right) = {} & \beta_1.\hat{P}(D_j / age, sex, length\ stay) \\ & + \beta_2.\hat{P}(D_j / MU) \\ & + \beta_3.\hat{P}(D_j / proc_1,...,proc_N) \\ & + \beta_4.\hat{P}(D_j / D_1,...,D_M)\end{aligned} \quad (5)$$

Where β_1, β_2, β_3 and β_4 depend on their respective sensor performance.

Algorithm

Based on coding expertise, it was decided to limit the prediction to the diagnosis category. The final code choice is determined by the physician. As a result, the iterative algorithm for proposing diagnosis codes is as follows:

1. Display the 10 most probable diagnostic categories.
2. The coder chooses diagnoses among these categories.
3. Display the 10 consecutive most probable diagnosis families, taking into account already chosen and not chosen diagnoses.
4. The coder chooses diagnoses among these new displayed categories.

IV. Results - Discussion

The method was applied on a 1000 discharge abstracts sample for evaluation, each of them with classifying and non-classifying procedures. Every discharge abstract was randomly extracted from the hospital database. All contained

patient demographic data, length of stay, severity index, procedures list, and the diagnoses chosen by physicians. The aim of the test was to find the diagnosis list according to other abstract information. A diagnosis code was considered to be found by the method if its category was ranked within the 10 more likely responses, for each algorithm iteration. The validation criteria were the rate of discharge abstracts for which all diagnoses were found, and the rate for which at least one diagnosis was found.

- Obtained results from $\hat{P}(D_j|age,sex,length\ stay)$, $\hat{P}(D_j|proc_1,...,proc_N)$ and $\hat{P}(D_j|MU)$

The first step was to search using one after another the 3 information sources. Figure 1 shows obtained results. Each vertical bar is composed of 3 parts, representing from top to bottom the rates of discharge abstracts for which no diagnosis was found, at least one diagnosis was found, and all diagnoses were retrieved, respectively. The first four columns present the classification rates after the first iteration and after the third one, using in each case three inputs, the procedure, the medical unit, and all inputs.

These results show that the most relevant piece of information is the medical unit where the patient was, then the given procedures, and finally demographic data.

- Obtained results by adding $\hat{P}(D_j|D_1,...,D_M)$

The main problem was to define coefficients β_i values. Based on test results, the β_i coefficients for demographic data, procedures, and medical units were chosen proportionally to the rates of discharge abstracts for which all diagnoses were found. The fourth one, for diagnoses was approached by experiments looking for the β_i value that optimized results.

- Obtained results by combining every information source

Again, the main inconvenient was to define coefficients β_i values. For the first three coefficients, each value is proportional to the percentages of RUM (standardized discharge summary) for which all diagnoses were found. As for the fourth one, experimental results led to identify the optimal value for this criterion.

During the first iteration, results were only obtained from the first three medical information sources, whereas during the second and third iterations, all information sources were exploited.

V. Conclusion

Our study suggests that the REFEROCOD coding-aid method, based on a probabilistic approach obtained encouraging results even if perfectible. Particularly, the learning database is imperfect with missing or erroneous codes, resulting in imperfect probability estimation and imperfect results. This is perhaps the major limitation of this approach.

Perspective developments of the proposed work will be to validate this approach in real conditions. The experiment has begun by developing a human interface, which suggests to the physician the diagnosis categories according to other information sources. A comparison between this methodology and the usual research method by keywords is currently realized by 3 physicians, studying 30 randomized discharge abstracts. Preliminary results indicate that the two methods have a similar usability time. Nevertheless, our tool seems to be more efficient, that means faster, for finding the first diagnostic code.

References

[1] Noury JF. La gestion médicalisée des établissements de santé, Le PMSI et l'information médicale. CNRS Editions, 2000.

[2] Maravic M, Le Bihan C and Landais P. La classification commune des actes médicaux (CCAM) : de la description à la tarification, *Revue du Rhumatisme* 2003, 70 9 : 785-9.

[3] *International Statistical Classification of Diseases and Health Related Problems (The) ICD-10 Second Edition Tenth Revision. Volume 1,2,and3*, World Health Organization (ed), 2004.

[4] Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton Measuring Diagnoses: ICD Code Accuracy, Health Serv Res. 2005 October; 40(5 Pt 2): 1620–1639.

[5] Cauvin JM, Hardy B, Jehan P, Josso F, Collet JY, Gourlaouen A, Gicquel B, Le Beux P.Qualité du codage et conséquences en ISA et journées théoriques : à partir des fichiers de 100 dossiers recodés par établissement. *Journal d'Economie Médicale* 1997, 15 : 195-206.

[6] Abraham B et Ledolter J. *Statistical methods for forecasting*. In: Wiley-interscience (ed), 1983.

[7] Grandin JF. *Fusion de données : Théorie et méthode*. In. Technique de l'ingénieur (ed) 2006.

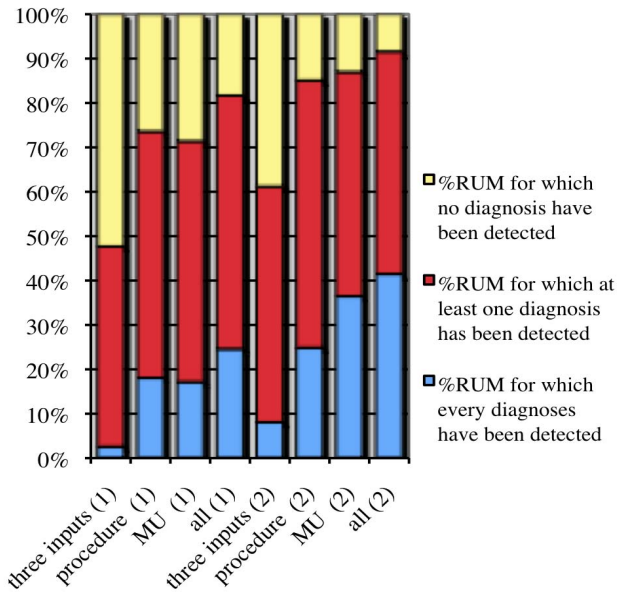

Figure 1 : Obtained results after the first two steps of the algorithm (1) and at the end of the algorithm (2)