Increasing secondary diagnosis encoding quality using data mining techniques

Ghazar Chahbandarian, Nathalie Bricon-Souf, Rémi Bastide, Jean-Christophe Steinbach

▶ **To cite this version:**

Ghazar Chahbandarian, Nathalie Bricon-Souf, Rémi Bastide, Jean-Christophe Steinbach. Increasing secondary diagnosis encoding quality using data mining techniques. 10th IEEE International Conference on Research Challenges in Information Science (RCIS 2016), Jun 2016, Grenoble, France. pp. 1-10. hal-01809380

HAL Id: hal-01809380
https://hal.science/hal-01809380

Submitted on 6 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 19022

The contribution was presented at RCIS 2016 :
http://www.rcis-conf.com/rcis2016/
To link to this article URL :
https://doi.org/10.1109/RCIS.2016.7549339

To cite this version : Chahbandarian, Ghazar and Bricon-Souf, Nathalie and Bastide, Rémi and Steinbach, Jean-Christophe *Increasing secondary diagnosis encoding quality using data mining techniques.* (2016) In: 10th IEEE International Conference on Research Challenges in Information Science (RCIS 2016), 1 June 2016 - 3 June 2016 (Grenoble, France).

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

Increasing Secondary Diagnosis Encoding Quality Using Data Mining Techniques

Ghazar Chahbandarian
University of Toulouse, IRIT/ISIS,
F-81100 Castres, France
Email: ghazar.chahbandarian@irit.fr

Nathalie Bricon-Souf
University of Toulouse, IRIT/ISIS,
F-81100 Castres, France
Email: nathalie.souf@irit.fr

Rémi Bastide
University of Toulouse, IRIT/ISIS,
F-81100 Castres, France
Email: remi.bastide@irit.fr

Jean-Christoph Steinbach
Department of Medical Information,
Centre Hospitalier Intercommunal de Castres Mazamet
F-81100 Castres, France
Email: jean-christophe.steinbach@chic-cm.fr

***Abstract*—In order to measure the medical activity, hospitals are required to manually encode information concerning an inpatient episode using International Classification of Disease (ICD-10). This task is time consuming and requires substantial training for the staff. We propose to help by speeding up and facilitating the tedious task of coding patient information, specially while coding some secondary diagnoses that are not well described in the medical resources such as discharge letter and medical records. Our approach leverages data mining techniques in order to explore medical databases of previously encoded secondary diagnoses and use the stored structured information (age, gender, diagnoses count, medical procedures...) to build a decision tree that assigns the proper secondary diagnosis code into the corresponding inpatient episode or indicates the impatient episodes that contains implausible secondary diagnoses. The results suggest that better performance could be achieved by using low level of diagnoses granularity along with adding some filters to balance the repartition of the negative and positive examples in the training set. The obtained results show that there is big variation in the evaluation scores of the studied diagnoses, the highest score is 75% using F1 measurement and the lowest 25% using F1 measurement which indicates further enhancements are needed to achieve better performance regardless of the encoded diagnosis. However, the average accuracy of all the studied secondary diagnoses is around 80% which indicates better negative predictions therefore it could be useful in the prevention or the detection of wrong coding assignments of secondary diagnoses in the inpatient stay.**

I. Introduction

In France, since 1991, by recommendation of the ministry of health, all the public healthcare facilities are mandated to record patient diagnosis and medical procedures in a national database called PMSI (*Programme de Médicalisation des Systèmes d'Information*) equivalent to the PPS (*Prospective Payment System*) used in the USA [1]. The system was initially used for the purpose of reporting hospital activity and comparing the productivity between different facilities. In 1998, PMSI was used by all public and private hospitals for the main purpose of hospital fair funding. Since its creation, millions of records have been registered in PMSI database, which makes it an attractive target for data analysis, in order to solve different problems using data mining techniques.[2].

Each inpatient episode in France consists of one or several standard patient discharge reports called RUM (*Résumé Unité Médicale*). The RUM contains administrative information such as gender, age and length of stay. The RUM also contains medical information such as diagnoses and medical procedures performed during the stay in the medical unit. At the end of the inpatient episode, all the reports are combined into one report called RSS (*Résumé de Sortie Standardisé*). Then, an anonymisation process is applied, thus producing a so-called anonymised episode summary RSA (*Résumé de Sortie Anonymisé*). Finally, the RSA reports are sent to the Regional Health Agencies ARS (*Agences Régionales de Santé*) where they are stored in the national PMSI database. Each hospital is eventually refunded according to the activity described in the RSA reports. Hospitals try to document their activities as accurately as possible to get fair payment. Inaccurate encodings of inpatient episode information could cause diminished refundings, or penalties up to 5 per cent of their annual budget [2]. Consequently, a lot of effort is made by hospitals to increase encoding accuracy of the diagnoses and medical procedures. The Medical Information Unit (*Département d'Information Médicale, DIM*) is responsible for the encoding process which involves within each hospital. One of the encoding challenges is encoding secondary diagnoses. Unlike for main diagnosis, which is not too difficult to detect, some secondary diagnoses require an extra effort in order to identify them, because sometimes they are not clearly mentioned in the medical reports and cannot be directly implied. Another challenge is to find out if there is a way other than medical encoding rules to detect miscoded secondary diagnoses such as checking the plausibility of existence of certain diagnoses together under certain contexts like length of stay, type of admission or medical procedures performed during the stay.

In this paper we address two main challenges:

- Encoding secondary diagnosis support.
- Avoid encoding implausible secondary diagnoses.

PMSI national database is the richest and the most valuable source of documented standard diagnoses and medical procedures in France. It contains millions of records collected over years which make it fall under big data definition which requires certain type of tools to be explored. Recently, it has been made accessible for research purposes. Among the available methods in data mining, decision trees method is interesting because of its result that can be exploited by a non specialist in the domain.

II. Related work

The investigated problem of supporting the encoding of secondary diagnoses using structured data stored in medical documentation databases falls under data prediction category, which is a common phenomenon in most databases, and researchers address this problem in a variety of applications such as marketing, e-business and other industrial sectors. However, data prediction in the healthcare domain has its particular constraints since it is dealing with medical data which is considered unique in terms of heterogeneity, privacy-sensitive, ethical, legal, and social issues [3]. Therefore, previous researches used various data analysis methods to overcome the difficulties and to solve the encoding diagnosis problem. Medical data is heterogeneous in its nature, it is collected from different sources such as laboratory data, interviews with the patient, radiology images, observation and interpretation of the physician etc...In order to identify all the diagnoses and to assign codes to them, coders need to look at many sources and to interpret information to find out the right code. Automatic code assignment or the support of code assignment simulate the coders by looking at these heterogeneous information and interpret them. Medical photos are rich of information, and it is often used by the coders to identify the diagnosis. Using image processing is one way to extract information from medical photos, it is used to support the radiologists to identify the diagnoses in the image [4] but it is still not used to support the coders assigning codes to the diagnoses. One attempt is proposed by [5] to assist the coders with the assignment of medical codes using image processing by proposing a list of diagnoses codes corresponding to the viewed medical image during the coding, this work is in progress. Image processing to assign medical codes is still new research area that is not matured yet additional techniques should be invested in this domain to investigate the benefits using such techniques. In this paper, we mainly focus on processing conventional data, rather than images. Other main sources of diagnoses are the clinical reports and physician's interpretations, discharge letters and other medical documents that are usually written in free text and that are frequently used by the coders to determine the medical code. According to the reviewed papers, the best suitable technique to extract information from free text is Natural Language Processing (NLP) methods. The idea of NLP is to translate free text into formal representation or features so that computers can understand the text and manipulate it [6]. NLP has good results in predicting diagnosis and medical procedures codes. One way to extract medical codes using NLP is using expert rules and applying it directly to the medical reports, it can reached up to 88% F1 measure score [7][8]. To achieve a high accuracy, experts' knowledge on how they code from radiology reports are translated into hand crafted rules. The rules aim at extracting lexical elements from radiology reports written in free text, lexical elements can be generated using semantic features to include negations, synonyms and uncertainty. The problem of these methods is that they are in most of the cases language dependant and it is difficult to be generalised. One of the objectives of our work is to find a general method that uses structured input and that avoids the ambiguities raised by any language. Another way of extracting medical codes is using NLP machine learning techniques by analysing medical database of previously coded patient episodes and extracting the feature matrix of medical reports of each corresponding patient episode. Finally, machine learning methods are applied on these matrices to generate models that can predict a diagnosis code. Different algorithms are used to tackle the problem such as decision trees [7], K-Nearest Neighbor (KNN) [9] [10] citeErraguntla2012 naïve Bayes classifier [11],[12] regression [13] [14], Support Vector Machine (SVM) [15], Medical Subject Heading MeSH [16]. The problem of these methods is still the same. They are applicable only in certain conditions and they can not be generalised. However, the machine learning methods used in these methods are useful in our work since we are planning to use machine learning methods on standard structured data. We can use some of the experience used in dealing with feature matrices, such as dealing with highly biased negative examples in the training set [13]by using automatic weights scheme and dealing with multi label code assignment cases [7].

Few works used structured patient data other than images and free text. The data are mostly extracted from medical records, such as patient information i.e. (age, sex, length of stay etc...) clinical information i.e. (prescription, medications) and other related medical data such as medical procedures and diagnoses. An interesting study in the reviewed papers is using statistic method and probabilities [17]. Three different input were tested to estimate a diagnosis code probability, the first input was patient information (age, sex, length of stay) , the second input was medical unit information and the last input was medical procedures. A diagnosis prediction is considered valid if it was within the first 10 diagnoses ordered by probability score. The results showed that medical procedures were the most informative input whereas the patient information was the least informative input and better results could be achieved using all the inputs together by defining the right coefficient for each input [17]. The limitation of probabilistic/statistical approaches is that imperfect results are obtained when used with imperfect data, missing data or erroneous codes. Data mining approaches are good alternative, since data preprocessing techniques can help reducing the effect of imperfect data[18]. Two studies in France tackled the problem of assigning medical codes to inpatient episodes [19] and [20]. They used other diagnoses occurred in previous inpatient episodes and constructed sequential patterns rules

to predict a diagnosis code in the current patient episode. Two out of three diagnoses were successfully predicted using sequential patterns in [19]. In fact, sequential patterns work well using one input variable, in our work we are going to use all the available structured variables in the medical files to enhance the results. The last reviewed work was done by Ferrao, he used well structured data extracted from electronic medical records and converted them to around 5000 features. He used different data mining algorithms in several steps, naïve bayes and decision trees algorithms in [21], SVM in [22] and finally regression algorithms in [23] trying to assign codes during different periods of the patient episode. All algorithms gave about similar evaluation of F1-measure but they still didn't reach NLP techniques accuracy on radiology report. Finally, our method is inspired from previous studies to tackle a problem not addressed so far, assigning or denying secondary diagnoses codes to patient episodes. We are going to use adapted structured data as input so it can be generalised on any language. As for the data mining algorithm we are going to use decision trees because it showed good results with this kind of problem compared to other methods in addition to the interpretability of its model i.e. the extracted tree can be verified easily from non specialist in informatics such as a physician. The scalability of the decision trees is another reason to use decision trees since we are going to use first local data set and then move to national dataset.

III. Materials and Methods

Medical databases are rich of data but poor of knowledge. Data mining is a way to extract previously unknown hidden data that could be useful. Machine Learning (ML) provides the necessary tools and techniques to data mining in order to discover knowledge from raw data. The idea is to identify strong patterns in a database and generate a model that can predict similar cases in the future. The choice of ML technique and the data used to build its model play a big role in the results. Each ML technique has inputs and an output, the data used to build the ML model is called input or features, the prediction made by the model is called output or label. In order to produce better output/prediction, the input should be chosen carefully i.e. handling the missing data, discretizing the continuous numeric values and discarding non informative features. In this section, we are going to describe the data structure used to build the ML model, feature selection in addition to data preprocessing and finally ML technique used to tackle the problem.

A. Data structure

We used the PMSI database of "*Centre Hospitalier Intercommunal de Castres Mazamet*" hospital in France. The united structure of PMSI permits us to evaluate our model in the future on different scales regional and national. The PMSI contains anonymous discharge summaries (*Résumé de Sortie Anonymisé, RSA*). Each summary consists of a set of elements that characterise an inpatient episode.

- Administrative information: Admission date, Discharge date, Admission mode, Discharge mode (transfer, death), Length of stay, Gender, Age…
- Clinical information: the main diagnosis that motivates the inpatient episode, secondary diagnoses and related diagnoses. It also contains all the medical procedures performed during the inpatient episode.

The (*International Classification of Disease, ICD-10*) [24] is used to encode all diagnoses. The French version of ICD-10 contains 33,816 codes, the first three characters of the codes stand for code categories, there are 2,049 categories and they are usually used for code prediction. The Common Classification of Medical Procedures CCAM (*Classification Commune des Actes Médicaux*) is used to encode the medical procedures. It consists of four characters and three numbers. There are around 1,700 standard medical procedure codes classified under 19 chapters depending on their category [25].

B. Feature selection

Our problem is the assignment of secondary diagnoses to the inpatient episodes or denying them. In order to increase ML model efficiency, it is necessary to choose the most relevant features to the problem and discard the non relevant ones. The first set of relevant features is composed of personal information which includes the patient's gender and his/her age at admission to the inpatient episode. We discarded the postal code, as all the patients come from the same area where the hospital is located, but this information would be interesting to investigate in case of using the national version of PMSI database. The second set of retained features concerned inpatient episode including the length of stay, the patient admission type, the patient discharge status, the time interval between the admission date and the first medical procedure performed , the transfer count between medical units in the inpatient stay, the medical procedures count, the season of the admission and the previous inpatient episode count calculated thanks to a process of anonymous chaining available in the PMSI databases which permits to link information from a single patient. We chose to use medical procedure chapters instead of using each medical procedure as a separate feature, therefore we have 19 features for the medical procedures each feature indicate if one or many medical procedures in the corresponding chapter are occurred during. Similarly, we used diagnoses chapters as features instead of using each diagnosis as a separate feature. Two levels of diagnosis granularity are available, the high level granularity contains 19 general chapters, the low level of diagnoses granularity contains 126 specific categories. The chapters group the diagnoses based on their similarity. Each chapter represents a feature indicating if a diagnosis occurred during an inpatient episode or not. In addition, all the medical procedures and all diagnoses other than the predicted one are considered as input features to the ML model. The choice of using chapters, instead of the code itself is to limit the extra large number of features which does not yield necessarily to good results for ML learning algorithm, specially for decision trees case [26]. Finally, the

output or the label of the ML model is a ICD-10 code for a secondary diagnosis, positive if the code exists in the inpatient episode and negative if it doesn't. In total, we have 181 features used to build our ML model. A detailed description can be found in the table I.

C. Data preprocessing

Data preprocessing is a very important step in data mining process. It consists in dealing with noisy and inconsistent data because of their huge size. Low quality data will lead to low quality mining results [18]. In our work we processed the database to deal with continuous numeric data. There are two kinds of data: numerical and categorical. Numerical data can be in two forms: continuous or discrete. Data mining algorithms prefer categorical or discrete numerical values therefore we discretized the continuous variables into discrete values. Binning, entropy-based or interval merging are common ways to discretise the data. To avoid arbitrary discretization of data and add meaning to the values so it can be interpretable by physicians, we have studied our database to extract statistical information such the mean of each diagnosis and the standard deviation. From these statistical information we have chosen to discretise the continuous features into three ranges (below - mean - over) where 'below' refers to values smaller than the mean minus one standard deviation, 'mean' refers to data between the mean plus minus one standard deviation and 'over' refers to data above mean plus one standard deviation. The following features have been discretized (frequency, transfer count, medical procedure count, diagnoses count, age, length of stay, Delay).

D. Building the decision tree

The machine learning method we have chosen is *decision trees* from class-labeled training tuples. We chose decision tree because it generates simple models, it is easy to interpret and can be validated by physicians who are not necessarily specialists in the domain. Decision trees are scalable and can produce efficient models even when using large amounts of data. We decided to use *Classification and Regression Tree* (CART) algorithms as induction algorithm as it allows to build a binary decision tree with Gini impurity index to select the features representing the nodes of the tree. We applied postpruning the decision tree to avoid the overfitting problem, it occurs when the model is more accurate on the training set than new unseen data. We pruned the subtree that generates the less error rate.[18]

Our objective is then to detect secondary diagnoses, from the PMSI database information, using decision trees. Two main questions are highlighted in order to produce better performance model.

- Administrative information: Admission date, Discharge date, Admission mode, Discharge mode (transfer, death), Length of stay, Gender, Age…
- Clinical information: the main diagnosis that motivates the inpatient episode, secondary diagnoses and related diagnoses. It also contains all the medical procedures performed during the inpatient episode.

How to limit the effect of the large number of negative examples compared to the positive ones. Which granularity level of information representation to use for building the decision tree.

To deal with the first issue : the PMSI database contains by nature more negative examples than positive ones, we make the hypothesis that the we can build a better performance decision tree by balancing the number of positive and negative examples. To achieve this balance we tried to solve it using two methods:

- The first method gives the positive examples more weights than the negative ones . To answer this question we compared the performances of two decision trees, the first decision tree is built by giving equal weights to the negative and positive examples. The second decision tree is built by giving the positive examples double weight.
- The second method extracts a specific training set using certain contextual filters to focus on a sub-database. In our work we used the most frequent primary diagnoses occurred along with the secondary diagnosis as filter : the sub-database will then concern all the cases dealing with a specific primary diagnose, and will have cases with or without the chosen secondary diagnoses. We hope it could improve the significance of the results. To answer this question we compared the performances of 11 decision trees, the first decision tree is built using all the database. The other decision trees are built by filtering the data using 10 sub-databases concerned with the 10 most frequent primary diagnoses.

The second issue concerns the choice of which diagnoses granularity level leads to a better decision tree performance. We can propose two levels of diagnoses granularity, either high level with 19 features (general chapters) or low level of diagnoses granularity with 126 features (more specific chapters). To select the most efficient choice, we compared the performances of two decision trees, each one is built using different level of diagnoses granularity.

The steps followed to build and to evaluate the decision tree are described in Algorithm 1. The first step allows to choose the right configuration by fixing the 3 parameters we have just mentioned :

- The weight of positive and negative examples (for instance, we decide to weight a positive example twice in order to highlight its importance) ;
- The use of complete or specific database (if it is off all the database will be considered, if it is on it will be split into sub-databases concerned with most frequent primary database) ;

The granularity level of diagnosis (for instance, we choose a decision tree based on the 19 features issued from general chapters). If the primary filter option is chosen, we query the most frequent primary diagnoses occurred with the studied secondary diagnosis. (for example, in case of "B96" bacterial

TABLE I: Used features in the decision trees.

	Variable Name	Description	Valid values
Personal information	Gender	Patient's gender	F=Female M=Male
	Age	Patient's age at admission	Below=the age is less than the average minus the standard deviation Mean= the age is inside the average $\pm$ the standard deviation Over= the age is more than the average plus the standard deviation
Inpatient variables	Length of stay	Time interval between admission date and discharge date	Below=the interval is less than the average minus the standard deviation Mean= the interval is inside the average $\pm$ the standard deviation Over= the interval is more than the average plus the standard deviation
	Admission type	Patient's admission type	1= Emergency 2=Urgent 3=Elective 4=Newborn 5=Trauma 9=Information not available
	Disposition	Patient's discharge status	1=Discharge to home 2=Transferred to short-term facility 3=Transferred to skilled nursing facility 4=Transferred to intermediate care facility 5=Transferred to other healthcare facility 6=Transferred to home health care 7=Left AMA(Against Medical Advice) 20=Expired/Mortality
	Season	The season at the admission	Summer Winter Fall Spring
	Frequency	The count of the inpatient episodes of the patient during his life.	Below=the count count is less than the average minus the standard deviation Mean= the count is inside the average $\pm$ the standard deviation Over= the count is more than the average plus the standard deviation
	Delay	Time interval between admission date and first medical procedure	Below=the interval is less than the average minus the standard deviation Mean= the interval is inside the average $\pm$ the standard deviation Over= the interval is more than the average plus the standard deviation
	Inpatient transfer count	The count of the transfers between medical units in the inpatient episode	Below=the count is less than the average minus the standard deviation Mean= the count is inside the average $\pm$ the standard deviation Over= the count is more than the average plus the standard deviation
	Medical procedures count	The count of the medical procedures during the inpatient episode	Below=the count is less than the average minus the standard deviation Mean= the count is inside the average $\pm$ the standard deviation Over= the count is more than the average plus the standard deviation
Derived flags	Classified	A flag indicating whether the inpatient stay has a classified/important medical procedure or not.	0=No 1=Yes
	Emergency	A flag indicating whether the inpatient stay has an emergency case or not.	0=No 1=Yes
	Medical procedure groupings	19 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding medical procedure category.	0=No 1=Yes
	Urgent medical procedure grouping	5 flags, each flag indicates whether the inpatient stay has a medical procedure within the corresponding urgent medical procedure category.	0=No 1=Yes
	First level diagnoses granularity	19 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding diagnosis granularity.	0=No 1=Yes
	Second level diagnoses granularity	126 flags, each flag indicates whether the inpatient stay has a diagnosis within the corresponding diagnosis granularity.	0=No 1=Yes
Output	Label	A flag indicating whether the inpatient stay has the studied secondary diagnosis or not.	0=Negative 1=Positive

agents infection as secondary diagnoses, the most frequent primary diagnoses found in the database are "*Acute tubulointerstitial nephritis*" with the code "*N10*", "*Malaise and fatigue*" with the code "*R53*", "*Fever of other and unknown origin*" with the code "*R50*" , etc...) Afterwards, for each primary diagnosis we query the positive and negative examples. Then, we do all the preprocessing, split the data into training and testing set and use the training set to build the decision tree. Afterwards, We prune the tree in case it produces better performance. Finally, we evaluate the tree using the testing set. The same steps are applied when primary diagnoses filter option is off but without filtering the data.

Algorithm 1 The steps followed to build secondary diagnoses decision tree

```
Set(positive and negative example's weights)
Set(primary diagnoses filter option)
Set(granularity level of diagnoses)
if primary diagnoses filter is off then

    Query the positive and negative examples
    Discretize the continuous features
    Split the data into k folds
    for Each fold do
        Choose the training and testing sets
        Build the decision tree with the training set using
        CART algorithm
        Prune the tree
        Evaluate the tree using testing set
    end for
else

    Query the most frequent principal diagnoses
    for Each principal diagnosis do
        Query the positive and negative examples
        Discretize the continuous features
        Split the data into k folds
        for Each fold do
            Choose the training and testing sets
            Build the decision tree with the training set using
            CART algorithm
            Prune the tree
            Evaluate the tree using testing set
        end for
    end for
end if
```

IV. Results

A. Dataset

Certain secondary diagnoses are not well described, such as obesity, malnutrition and respiratory failure and they are often not coded in PMSI. In France, one hospital reported that more than a third of the patients with denutrition and obesity were not coded in the database [27]. We used an anonymized sample data extracted from the PMSI database of "*Centre Hospitalier Intercommunal de Castres Mazamet*" hospital, it contains around 90,000 inpatient episodes between 2011 and 2014. We decided to focus on interesting and frequent secondary diagnoses which are difficult to detect as they are usually not well described across the medical sources. For this reason, the doctor in charge of the Medical Information Department (DIM) in the '*Centre Hospitalier Intercommunal de Castres Mazamet*' hospital helped us to choose some secondary diagnoses that fulfil the criteria. Table II.

TABLE II: Summary of the studied secondary diagnoses.

ICD-10 codes	Labels	Count in DB
J96	Respiratory failure	4166
B96	Other specified bacterial agents as the cause of diseases classified to other chapters	6514
T81	Complications of procedures	1150
R29	Other symptoms and signs involving the nervous and musculoskeletal systems	1596
R26	Abnormalities of gait and mobility	2378
E66	Overweight and obesity	5453
E44	Malnutrition	2144

B. Evaluation

We implemented the proposed algorithm using rpart library in R language. We evaluated the results using 5-fold cross validation, in each fold we divided the dataset into 80% training set and 20% testing set. Since our model has binary output, positive and negative, we used the standard measurements used in classification Accuracy, Precision, Recall and F1-measure. The measurements are based on the number of instances that are correctly assigned positive examples True Positive (TP), the number of instances that are correctly assigned negative examples True Negative (TN), the number of instances that are incorrectly assigned positive examples False Positive (FP) and the number of instances that are incorrectly assigned negative examples False Negative (FN).[28]

Accuracy is the ratio of correctly assigned negative and positive examples to the total number of examples.

$$A = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

Precision is the ratio of correctly assigned examples to the total number of examples produced by the classifier.

$$P = \frac{TP}{(TP + FP)} \tag{2}$$

Recall is the ratio of correctly assigned examples to the number of target examples in the test set.

$$R = \frac{TP}{(TP + FN)} \tag{3}$$

F1-measure represents the harmonic mean of precision and recall according to the formula in (4):

$$F1 = \frac{2P * R}{(P + R)} \tag{4}$$

TABLE III: Results obtained for B96 (bacterial agents) as secondary diagnosis without fixing any primary diagnosis using high and low levels of diagnoses granularity.

High level of granularity				**Low level of granularity**			
Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	F1
94%	42%	31%	36%	97%	80%	62%	70%

TABLE IV: Results obtained for B96 (bacterial agents) as secondary diagnosis with the most frequent primary diagnoses using high and low levels of diagnoses granularity.

Principal diagnosis	**High level of granularity**				**Low level of granularity**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	F1
N10	75%	82%	89%	85%	77%	83%	90%	86%
R53	83%	45%	52%	48%	95%	82%	85%	84%
R50	81%	51%	69%	59%	92%	83%	77%	80%
R06	88%	40%	21%	27%	94%	87%	53%	66%
R10	90%	33%	28%	31%	95%	70%	66%	68%
I50	89%	25%	20%	22%	97%	78%	82%	80%
J44	67%	35%	44%	39%	77%	52%	52%	52%
N41	68%	88%	72%	79%	71%	83%	82%	82%
N39	63%	76%	72%	74%	64%	78%	71%	75%
J18	83%	28%	29%	29%	95%	77%	79%	78%
Average	79%	50%	50%	49%	86%	77%	74%	75%

TABLE V: Results obtained for J96 (Respiratory failure) as secondary diagnosis without fixing any primary diagnosis using high and low levels of diagnoses granularity.

High level of granularity				**Low level of granularity**			
Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
96%	15%	12%	13%	97%	34%	15%	20%

TABLE VI: Results obtained for J96 (Respiratory failure) as secondary diagnosis with the most frequent primary diagnoses using high and low levels of diagnoses granularity.

Principal diagnosis	**High level of granularity**				**Low level of granularity**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
I50	71%	36%	31%	34%	82%	31%	9%	14%
R06	56%	51%	76%	61%	73%	43%	51%	47%
J96	85%	34%	11%	16%	58%	53%	80%	63%
J44	58%	50%	67%	57%	67%	59%	83%	69%
J18	75%	22%	18%	20%	72%	41%	54%	47%
R53	75%	25%	27%	26%	93%	24%	14%	18%
J20	92%	27%	11%	16%	76%	32%	35%	33%
J15	76%	27%	22%	25%	74%	31%	36%	33%
Z51	70%	21%	19%	20%	97%	37%	33%	35%
J69	67%	43%	54%	48%	59%	26%	29%	27%
Average	73%	33%	34%	32%	75%	38%	42%	39%

Using these measurements we aimed to evaluate three aspects, firstly and most importantly the possibility to assign codes to secondary diagnoses or denying their existence in the inpatient episode using decision tree method. Secondly, to see the effect of our proposed solution to answer question of limiting the effect of the large number of negative examples compared to the positive ones. Finally to evaluate which diagnoses granularity level produces a better performing decision tree (low level when specific diagnoses groupings are considered, low level when general diagnoses groupings are considered) .

In the tables III V we present the evaluation measurements without using any filter and using equal weights for negative and positive examples for B96 (bacterial agents) and J96(Respiratory failure) as secondary respectively.

In the tables IV VI we present the evaluation measurements

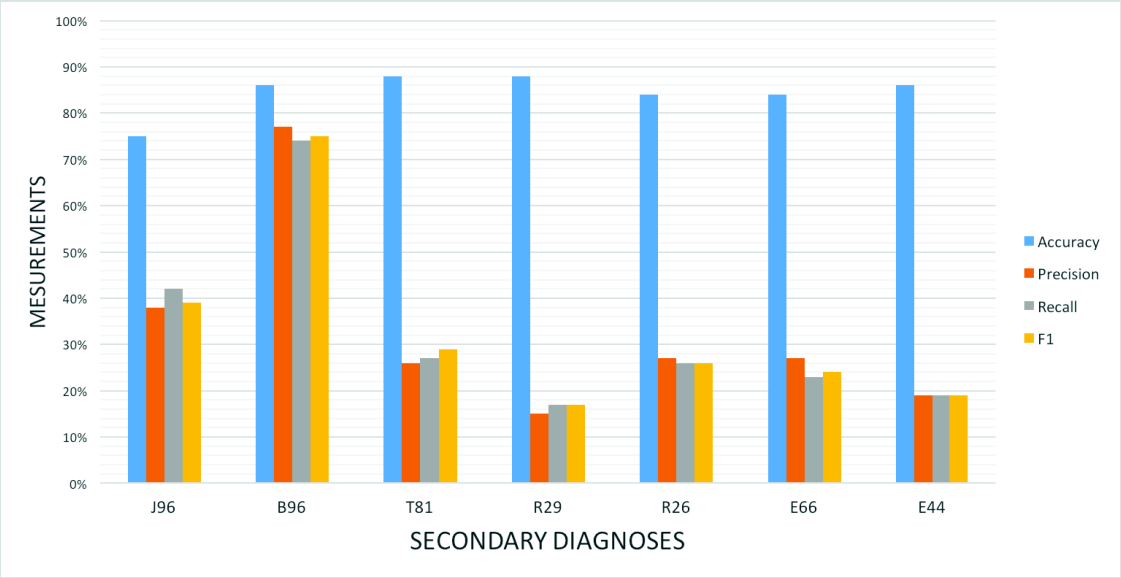

Fig. 1: Summary of the average measurements of the studied secondary diagnoses, using high level of granularity for all the encoded diagnoses.

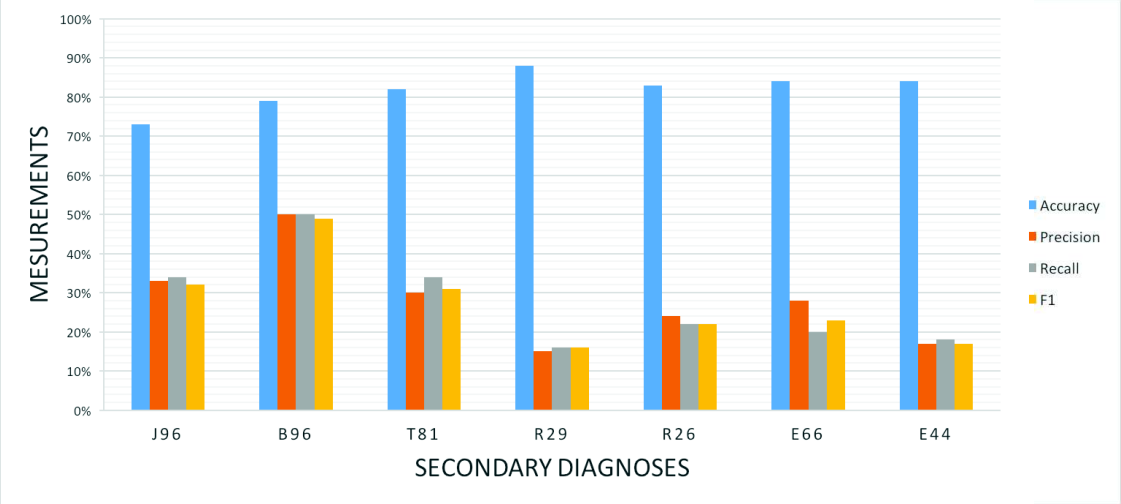

Fig. 2: Summary of the average measurements of the studied secondary diagnoses, using low level of granularity for all the encoded diagnoses.

using primary diagnoses as filter and giving double weight to the positive examples compared to the negative ones for B96 (bacterial agents) and J96(Respiratory failure) as secondary respectively.

The first four columns of each table represent the evaluation using high level of granularity the remaining four columns represents the evaluations using low level of granularity.

In figure 1 and figure 2 we show a summary of the average measurements of 10 most frequent primary diagnoses for all secondary diagnoses using high level of granularity in the figure 1 and low level of granularity in the figure 2.

V. Discussion

In the lights of the results and the evaluations obtained in figures 1 and 2, the measurement varied between different diagnoses. On the one hand, B96 (bacterial agents) scored the best F1, precision and recall measurements around 75% which is considered very good compared to similar reviewed studies. On the other hand, other diagnoses scored low percentages using the same measurements. The variation of measurements means that not all diagnoses have the same learning ability and confirms the complexity of the problem [29], where the same methodology applied to different diagnoses produced different results.

Therefore, the first part of our objective needs development in order to reach a level where it can be used to assign medical codes to the secondary diagnoses using structured information extracted from inpatient episodes. However, the second part of our objective which is detecting miscoded diagnoses could be a good potential application of our model because all the diagnoses scored very good accuracy measurement 1 against the other measurements in all the diagnoses. It is explained by the high percentage of True Negative predictions against the low percentage of True Positive predictions in the model.

Concerning the first highlighted issue about the effect of balancing the large number of negative examples against the negative ones, we notice that the tables IV VIhave better measurements compared to the tables III V which means that our proposition of using primary diagnoses as filter in addition to adding some extra weight to the positive examples is useful to produce better performing decision tree.

Finally, concerning the second highlighted issue about the effect of the granularity level we notice that the first four columns of all the tables III IV V VI have better measurements compared to the second four columns which means that the granularity level of the diagnoses at the input plays a big role in getting better scores. The scores are better using the low level of diagnoses granularity which has 126 chapters of diagnoses than using the high level of diagnoses granularities which has 19 chapters. Therefore, decision trees produce better results using the appropriate level of granularity, in order to generate suitable number of features to be used in the input. Further research is needed to explore new methods to enhance the results and in order to apply the model in real world application that support professionals coders by providing accurate codings.

VI. Conclusion

The paper outlined preliminary results of our methodology to develop an automatic model able to assign secondary medical codes or deny their existence. To achieve this objective we have built a model based on decision tree that uses structured data extracted from PMSI database as an input.

We enhanced the results by fixing the most frequent primary diagnoses before building the model. We achieved better results using a low level of diagnoses granularity input which contains 126 diagnoses chapters than using high level of granularity which contains 19 diagnoses chapters.

The results suggest that the performance of the model varies among codes: the best result obtained is assigning B96 code 75% F1 measure, the worst result obtained is around 20% F1 score. The variety of measurements between different diagnoses indicates that further research is needed to apply the model in real world application. On the contrary the high accuracy in the results suggest that true negative predictions are better than positive ones which makes denying secondary diagnoses while wrong coding possible.

For future work, we are planning to extend the research on larger database, such as national PMSI database of France, in addition to exploring new methods in order to balance the positive and negative examples in the training set.

Acknowledgment

The Ph.D. work of one author is funded by a grant from Midi-Pyrénées region, Castres-Mazamet Technopole, INU Champollion and Paul Sabatier University.

References

[1] R. B. Fetter, "Diagnosis Related Groups: Understanding Hospital Performance," *Interfaces*, vol. 21, no. 1, pp. 6–26, feb 1991. [Online]. Available: http://dx.doi.org/10.1287/inte.21.1.6

[2] R. Busse, A. Geissler, and W. Quentin, *Diagnosis-Related Groups in Europe: Moving towards transparency, efficiency and quality in hospitals*. McGraw-Hill Education (UK), 2011.

[3] K. J.Cios and G. Moore, "Uniqueness of Medical Data Mining," *Artificial Intelligence in Medicine Journal*, vol. 26, no. 1, pp. 1–24, 2002.

[4] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging." *The British journal of radiology*, vol. 78 Spec No, pp. S3–S19, jan 2005. [Online]. Available: http://www.birpublications.org/doi/abs/10.1259/bjr/82933343?journalCode=bjr

[5] V. Jain, "System and method for image processing with assignment of medical codes," 2014. [Online]. Available: https://www.google.com/patents/US8751920

[6] R. Collobert and J. Weston, "A unified architecture for natural language processing," in *Proceedings of the 25th international conference on Machine learning - ICML '08*. New York, New York, USA: ACM Press, jul 2008, pp. 160–167. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390156.1390177

[7] R. Farkas and G. Szarvas, "Automatic construction of rule-based ICD-9-CM coding systems." *BMC bioinformatics*, vol. 9 Suppl 3, p. S10, jan 2008. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2352868{\&}tool=pmcentrez{\&}rendertype=abstract

[8] I. Goldstein, A. Arzrumtsyan, and O. Uzuner, "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports." *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pp. 279–83, jan 2007. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2655861{\&}tool=pmcentrez{\&}rendertype=abstract

[9] A. R. Aronson, O. Bodenreider, D. Demner-Fushman, K. W. Fung, V. K. Lee, J. G. Mork, A. Névéol, L. Peters, and W. J. Rogers, "From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches," pp. 105–112, jun 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1572392.1572412

[10] P. Ruch, J. Gobeill, I. Tbahriti, P. Tahintzi, C. Lovis, A. Geissbühler, and F. Borst, "From clinical narratives to ICD codes: automatic text categorization for medico-economic encoding," *Swiss Medical Informatics*, vol. 23, no. 61, pp. 29–32, 2007.

[11] S. V. S. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." *Journal of the American Medical Informatics Association : JAMIA*, vol. 13, no. 5, pp. 516–25, jan 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1067502706001186

[12] K. Okamoto, T. Uchiyama, T. Takemura, T. Adachi, N. Kume, T. Kuroda, T. Uchiyama, and H. Yoshihara, "Automatic Selection of Diagnosis Procedure Combination Codes Based on Partial Treatment Data Relative to the Number of Hospitalization Days," *Proc. APAMI 2012*, no. 4, p. 1031, 2012.

[13] J. W. Xu, S. Yu, J. Bi, L. V. Lita, R. S. Niculescu, and R. B. Rao, "Automatic medical coding of patient records via weighted ridge regression," in *Proceedings - 6th International Conference on Machine Learning and Applications, ICMLA 2007*. IEEE, 2007, pp. 260–265.

[14] L. V. Lita, S. Yu, S. Niculescu, and J. Bi, "Large Scale Diagnostic Code Classification for Medical Patient Records," in *Proceeding sof the International Joint Conference on Natural Language Processing (IJCNLP'08)*. Citeseer, 2008, pp. 877–882.

[15] Y. Yan, G. Fung, J. G. Dy, and R. Rosales, "Medical Coding Classification by Leveraging Inter-Code Relationships," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2010, pp. 193–201.

[16] S. Pereira, A. Névéol, P. Massari, M. Joubert, and S. Darmoni, "Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding." in *Studies in health technology and informatics*, vol. 124, 2006, pp. 845–50. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17108618

[17] L. Lecornu, G. Thillay, C. Le Guillou, P. J. Garreau, P. Saliou, H. Jantzem, J. Puentes, and J. M. Cauvin, "REFEROCOD: a probabilistic method to medical coding support," in *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*. IEEE, 2009, pp. 3421–3424.

[18] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier, 2012.

[19] M. Djennaoui, G. Ficheur, R. Beuscart, and E. Chazard, "Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes." *Studies in health technology and informatics*, vol. 210, pp. 419–23, jan 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25991178

[20] J. Pinaire, J. Rabatel, J. Azé, S. Bringay, and P. Landais, "Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde," jun 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01180416/

[21] J. C. Ferrao, M. D. Oliveira, F. Janela, and H. M. G. Martins, "Clinical coding support based on structured data stored in electronic health records," in *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. IEEE, oct 2012, pp. 790–797. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6470241

[22] J. Ferrao, F. Janela, M. Oliveira, and H. Martins, "Using Structured EHR Data and SVM to Support ICD-9-CM Coding," in *2013 IEEE International Conference on Healthcare Informatics*. IEEE, sep 2013, pp. 511–516. [Online]. Available: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6680527

[23] J. C. Ferrão, S. Healthcare, and F. Janela, "Predicting length of stay and assignment of diagnosis codes during hospital inpatient episodes," in *Proceedings of the First Karlsruhe Service Summit Workshop-Advances in Service Research, Karlsruhe, Germany, February 2015*, vol. 7692. KIT Scientific Publishing, 2015, p. 65.

[24] WHO, "International Classification of Diseases (ICD)-10." [Online]. Available: http://www.who.int/classifications/icd/

[25] A. T. d. l. s. I. (ATIH). (2015) Version 39 de la CCAM. [Online]. Available: http://www.atih.sante.fr/version-39-de-la-ccam

[26] M. Sebban, R. Nock, J. H. Chauchat, and R. Rakotomalala, "Impact of learning set quality and size on decision tree performances," *International Journal of Computers, Systems and Signals*, vol. 1, no. 1, pp. 85–105, 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.62.6365

[27] C. Potignon, A. Musat, P. Hillon, P. Rat, L. Osmak, D. Rigaud, B. Vergès, and Others, "P146-Impact financier pour les {é}tablissements hospitaliers du mauvais codage PMSI de la d{é}nutrition et de l'ob{é}sit{é}. {É}tude au sein du p{ô}le des pathologies digestives, endocriniennes et m{é}taboliques du CHU de Dijon," 2010.

[28] S. Tuffery, "Data mining et statistique decisionnelle : l'intelligence des donnees," 2007.

[29] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, and W. R. Hersh, "A systematic literature review of automated clinical coding and classification systems." *Journal of the American Medical Informatics Association : JAMIA*, vol. 17, no. 6, pp. 646–51, jan 2010. [Online]. Available: http://jamia.oxfordjournals.org/content/17/6/646.abstract