Increasing Alertness while Coding Secondary Diagnostics in the Medical Record

Ghazar Chahbandarian[1], Nathalie Souf[1], Rémi Bastide[1] and Jean-Christoph Steinbach[2]
[1]*University of Toulouse, IRIT/ISIS, F-81100 Castres, France*
[2]*Department of Medical Information, Pays d'Autan Hospital, Castres, France*

Keywords: Electronic Medical Records, EMR, Context, Data Mining, Decision Tree, ICD, PMSI.

Abstract: In order to measure the medical activity, hospitals are required to manually encode information concerning a patient's stay 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 diagnostics that are not well described in the medical resources such as discharge letter and medical records. Our approach consists of building a decision tree out of big variety of inpatient stay information in particular, contextual information such as age, sex, diagnostic count and other related information, next figure out missing secondary diagnostics. The results are still preliminary, we identify some important information variables that can be interesting to verify while coding certain secondary diagnostics.

1 INTRODUCTION

In medical environment, and especially in hospitals, accurate and fast documenting of diagnosis and of medical procedures becomes a necessity. The estimation of monthly hospital costs will be more accurate, thus leading to better funding in the future. In the US, in 1983 the health system started using a system called Prospective Payment System (PPS) to store all the diagnostics under standardized codes International Classification of Disease (ICD-10) (WHO,), which in turn helped to save more than $50 billion since 1990 (Fetter, 1991). Similarly in France, in 1996 the ministry of health introduced a computerized medical data recording system PMSI '*Programme de médicalisation des systèmes d'information*'- Medical Program Information System (Dubois-Lefrère and Coca, 1992). In the PMSI, each inpatient stay is classified into groups called GHS '*Groupe Homogène de Séjour*' equivalent to Diagnosis Related Group (DRG) in the US. The GHS groups together all the similar inpatient stays in order to evaluate the cost of the hospital activity and finally to fairly refund the hospital.

Each inpatient stay leads to the recording of information such as diagnostics, medical procedures, prescriptions, complications and comorbidities, the diagnostics are coded according to ICD-10.

Each hospital tries to encode all the diagnostics and the medical procedures as accurately as possible to maximize its funding efficiency. However, coding process is a difficult task and it depends on the knowledge of the coders in addition to the interpretation of the coding rules. The process involves reading and understanding diagnostic medical resources such as discharge letters written by the doctors or the physicians, firstly to find the right principal diagnostic that motivates the inpatient stay, secondly to find other secondary diagnostics related to the stay and finally all found diagnostics are encoded into ICD-10 codes.

Coding all the diagnostics accurately is not an easy task. Some hospitals hire specialized people with coding experience to translate diagnostics faster and more accurately. In reality it's not difficult to code the principal diagnostic and medical procedures because they are clearly mentioned in the medical letter most of the time, whereas certain secondary diagnostics are not well described, such as obesity, denutrition 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 (Potignon et al., 2010).

In this paper. we focus on helping the coders by increasing their alertness level to detect secondary di-

Chahbandarian, G., Souf, N., Bastide, R. and Steinbach, J-C.
Increasing Alertness while Coding Secondary Diagnostics in the Medical Record.
DOI: 10.5220/0005824004900495
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 5: HEALTHINF*, pages 490-495
ISBN: 978-989-758-170-0

agnostics even if they are not well described in the medical resources. In order to achieve this goal we will use different kind of information available in the inpatient stay to build a decision tree and finally highlight the important variables used for each diagnostic.

2 RELATED WORK

There are few research works focusing on the use of data mining to predict a diagnostic code and we can report some of the following used methods:

- Sequential patterns. (Djennaoui et al., 2014) used sequential patterns to detect similar medical procedures patterns between different inpatient stays and to extract rules of these patterns to predict missing diagnostics. They studied three diagnostics and they were able to extract three rules, two of them are predicting the same diagnostic. Another similar work is done by (Pinaire et al., 2015).
- Text Mining. Other few works tried to extract the diagnostic codes directly from the medical letter using thesaurus such as MeSH (Medical Subject Heading) in (Pereira et al., 2006) or using probabilistic methods as in (Lecornu et al., 2009).
- Clustering. (Erraguntla et al., 2012) used K-nearest method to cluster all the similar inpatient stays and predict a missing diagnostic.

In our work we want to explore the use of decision tree method. Decision trees are useful in a context when clear results are needed, visually understandable, specially when they need to be validated by non specialist. We made the hypothesis that if we are able to determine which variables may help to predict a secondary diagnosis, we could help the coders to pay attention to these variables while coding.

3 METHODS

3.1 Used Data

We used an anonymous sample data extracted from the PMSI database of "Pays d'Autan" hospital, it contains around 75,000 inpatient stays between 2011 and 2014. We decided to use the information recorded in the PMSI database which are often well encoded as they are easy to detect (primary diagnoses, sex, age, stay duration...) to build decision tree. We also used two levels of diagnostics grouping, the first level groups the diagnostics into 19 general categories depending on thier similarities, the second level groups the diagnostics into 126 more specific categories. After fixing the primary and the secondary diagnostic of the inpatient stay, we retained the following information to include in the construction of the decision tree:

Table 1: Used variables in building the decision tree.

Sex	Male or Female
Mode of Entry (ME)	Patient acceptance mode in the inpatient stay. (GUIDE, 2006)
Mode of Sortie/Exit (MS)	Patient leaving mode of the inpatient stay. (GUIDE, 2006)
Age	Patient's age when accepted in the inpatient stay.
Duration	The duration of the inpatient stay in days.
Season	The season of the inpatient stay when the patient is accepted.
Frequency	Patient's inpatient stay count in the hospital
Gap	The gap in days between the entry date and the first medical procedure.
Passage count	The movements count between different sections during the inpatient stay.
Medical procedures count	Medical procedure count while the inpatient stay.
ICR	The quota cost of medical procedures in teh inpatient stay.
Classified	Whether the inpatient stay contains a classified/important medical procedure.
Emergency	Whether the inpatient stay contains an emergency case.
Example/Label	Positive if the inpatient stay has both the principal and the secondary diagnostics. Negative if it has only the principal diagnostic
Medical procedure chapters	A set of 19 variables each variable indicates if the inpatient stay contains a corresponding medical procedure.
Urgent medical procedure chapters	A set of 5 variables each variable indicates if the inpatient stay contains a corresponding urgent medical procedure category.
First level diagnostic grouping	A set of 19 variables each variable indicates if the inpatient stay contains a corresponding diagnostic grouping.
Second level diagnostic grouping	A set of 126 variables each variable indicates if the inpatient stay contains a corresponding diagnostic grouping.

In total, we have 181 information variables we can use to learn our model. The diagnostics were encoded according to the 10th revision of the International Classification of Diseases (ICD-10) (WHO,). The French version of it 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 predication. In our work we used the categories instead of the full code in order to increase the learning set as much as possible when studying a certain diagnostic.

We decided to focus on interesting and frequent secondary diagnostics but difficult to detect as they are usually not well described in the medical letters. For this reason, the responsible doctor of the Medical Information Department (DIM) in the 'Pays d'Autan' hospital helped us to choose some secondary diagnostics that fulfil the criteria. Table 2.

As for the machine learning method, we used Classification and Regression Tree (CART) algorithms to build a decision tree (Tufféry, 2010), we have chosen decision tree because it generates simple rules, easy to interpret and can be validated by doctors who are not necessarily specialist in the domain.

Table 2: Summary of the chosen secondary diagnostics.

ICD-10 codes	Lables	Count in DB
J96	Respiratory failure	2381
L97	Ulcer of lower limb	332
B96	Other specified bacterial agents as the cause of diseases classified to other chapters	4008
T81	Complications of procedures	590
R29	Other symptoms and signs involving the nervous and musculoskeletal systems	744
R26	Abnormalities of gait and mobility	1456

3.2 Implementation

For the implementation phase we used R language in R-studio environment, R is famous for its algorithms and statistics libraries that are easy to implement and test, such as rpart. The first step is to choose the right configuration such as:

- The secondary diagnostic list to study.
- Manuel or automatic discretization of continues variables.
- the granularity level of variables.
- The weight of positive and negative examples.

Then, for each secondary diagnostic we query the most ten frequent principal diagnostics. Afterwards, for each principal and secondary diagnostic we query the positive and negative examples. The positive examples are all the inpatient stays that contain both the principal and the secondary diagnostic at the same time. Whereas, the negative examples are all the inpatient stay that contain only the principal diagnostic without the secondary one. We used the example variable as an output variable to teach our model on the positive examples.

We discretized the continuous variables into three ranges (below average - average - over average), knowing that rpart library in R give us the possibility to discretize the variables into ranges, we prefered to do it ourselves to avoid non meaningful ranges and to avoid to cut them into to many ranges. For future tests, the automatic discretization option is always available.

Then, we use all the processed data to build the decision tree using rpart library (Therneau and Atkinson, 2015). Rpart is using CART decision tree with Gini impurity measure to choose the right variables to split and it performs pruning on the trees using cross-validation. (Tufféry, 2010)

Finally to calculate the important variables we used a built-in function in rpart library that is defined by "A variable may appear in the tree many times, either as a primary or a surrogate variable. An overall measure of variable importance is the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness * (adjusted agreement) for all splits in which it was a surrogate." (Therneau and Atkinson, 2015), consequently we cumulated all the measurement of the important variables so we have a final measurement at the end of the loop. We export the important variables table for further analysis.

```
Begin
 Choose the right configuration (granularity
 level of diagnostics chapter  variables to
 include  secondary diagnostics)
 For each secondary diagnostic do
 {
  Choose the most 10 frequent principal
    diagnostics
  For each principal diagnostic do
  {
   query the positive and negative examples
     Process continues variables to discrete
     ones (age-duration-frequency - medical
     procedures count-ICR-MS-ME)
   Build the decision tree using CART algori-
     thm
   Export the tree
   Cumulate the measurement of 'goodness of
     split' for each variable in a table.
  }
  Export the important variables table
}
End.
```

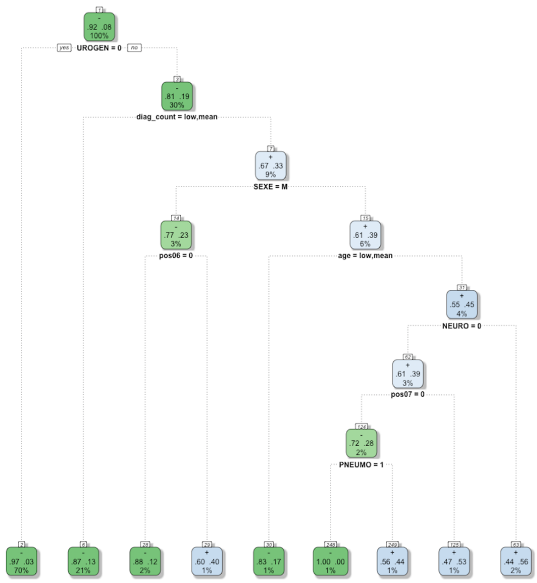


Figure 1: Decision tree of B96 ”Other specified bacterial agents” with R10 ”Acute abdomen” with error rate of 5.91% blue nodes represent positive nodes and green nodes represents negative nodes; each node has three numbers top left is the percentage of the negative examples, the number top right is the percentage of the positive examples; the number at bottom is the percentage of overall examples.

4 RESULTS

Since we have two levels of granularity of diagnostic grouping we run the program two times. the first run with 56 variables in total by including the first level of diagnostic grouping. For the second run we used 163 variables in total by including the second level of diagnostic groupings. In the following, we show the results of the two runs of B96 which is ”Other specified bacterial agents” as secondary diagnostic.

4.1 Decision Tree

We built a decision tree for each secondary diagnostic mention in table 2 with the ten principal diagnostics. For example figure 1 is showing the decision tree of the first run of B96 ”Other specified bacterial agents” as secondary diagnostic with R10 ”Acute abdomen” as principal diagnostic, the error rate is 5.91% using 10 fold validation. We notice for instance that if the inpatient stay has urogenital diseases, the diagnostic count is over average and the patient is over aged male and has neurological diseases then probably the inpatient stay has B96 ”Other specified bacterial agents” as secondary diagnostic.

In the second run Figure 2 we had more details about the medical procedure used in the inpatient stay

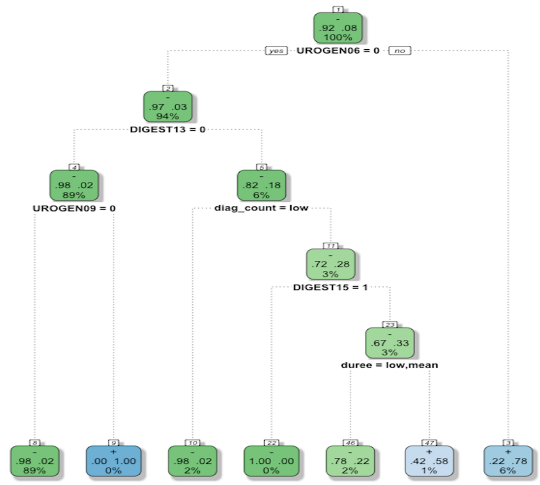


Figure 2: Decision tree of B96 ”Other specified bacterial agents” with R10 ”Acute abdomen” with error rate of 3.21% blue nodes represent positive nodes and green nodes represents negative nodes; each node has three numbers top left is the percentage of the negative examples, the number top right is the percentage of the positive examples; the number at bottom is the percentage of overall examples.

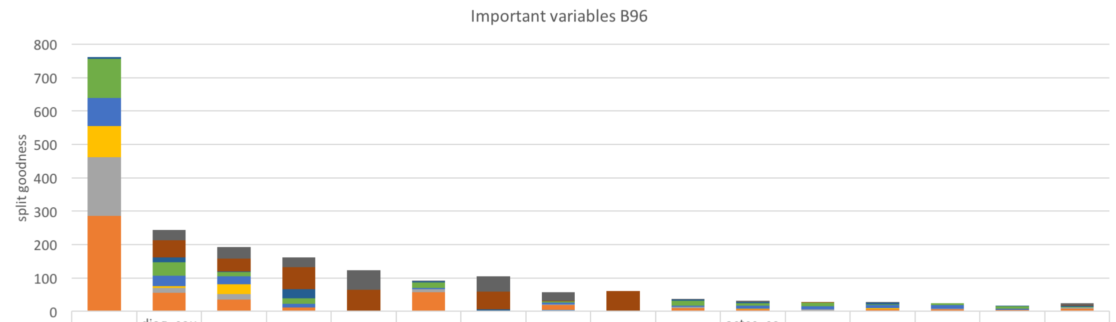
such as UROGEN06 ”Urinary tract infection” instead of UROGEN ”Pelvic pain, urogenital diseases”. In addition to less error rate 3.21

4.2 Variable Selection

To select the important variables for each studied secondary diagnostic we compared the important variables list from the first and the second run of our program and then select the important variables with the highest scores for each diagnostic, that help us to distinguish the interesting variables to verify while coding. For instance, in the first run figure 3 represents the important variables of B96 ”Other specified bacterial agents” as secondary diagnostic studied over the most ten frequent principal diagnostic, the variables are in order (Pelvic pain, urogenital diseases - diagnostic count- 8th chapter of medical procedure - duration - emergency - sex - 4th chapter of medical procedure - age...) In the second run figure 4 represents the important variables of B96 ”Other specified bacterial agents” as secondary diagnostic with the most ten frequent principal diagnostic, the variables are in order (UROGEN06- UROGEN09 - emergency - 4th chapter of medical procedure- diagnostic count- duration- 8th chapter of medical procedure ...)

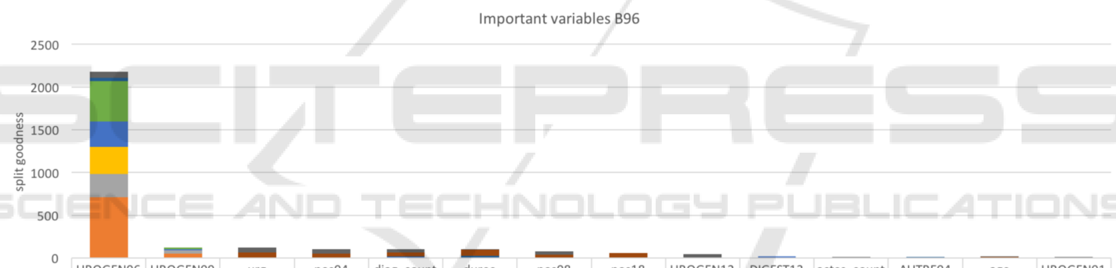
5 CONCLUSIONS

In this paper we presented our approach to help the



	UROGEN	diag_count	pos08	duree	urg	SEXE	pos04	age	pos18	MS	actes_count	DIGESTIF	icr	NB_PAS	pos05	pos01
N39	0	31,02692	35,82	29,39253	58,86579	1,851964	46,102	26,26209	0	1,999167	3,134428	0	2,481904	0	0	5,177273
N41	0	51,00139	38,84115	66,28404	63,68758	0	52,26029	1,766736	59,98065	0,829797	0	0,96456	0	0	0	1,44684
J44	4,911174	14,42603	1,351414	27,22381	0	3,961551	3,979553	0	0	3,853842	3,07085	0	5,507525	0	0,782518	4,659002
I50	116,9356	40,39039	12,1165	16,63673	0	16,16296	0,173608	3,142797	0	14,80802	8,894788	11,53877	1,739628	5,449817	8,651742	0,972785
R10	84,26798	31,19038	23,26416	10,63536	0	4,165069	0	6,724103	0,053398	2,587737	8,677127	9,115782	9,722279	10,63536	3,427133	0,123792
R06	94,25973	4,795383	30,55083	0	0	0	0,705996	0	0	0	2,189915	0	3,200645	0	0	0
R50	175,0033	15,46834	16,03299	0	0	8,706085	0	0,890454	0,313543	3,511244	0	3,151549	0	1,533714	0	3,52522
R53	285,8101	54,03896	33,77079	9,021096	0,09168	56,46745	0	15,24569	0,1146	7,098619	3,636049	2,38367	6,036743	4,829128	2,979775	7,995392
N10	0	1,397872	1,67433	2,477037	0,130862	0,485426	1,788628	3,976163	0	3,403773	1,871363	0,922392	0,036398	1,438931	0,824193	0

Figure 3: First run important variables chart for the B96 "Other specified bacterial agents" as secondary diagnostic: the table at the bottom of the figure represents the important variable table ordered descending each line represents the values obtained with the appropriate principal diagnostic.



	UROGEN06	UROGEN09	urg	pos04	diag_count	duree	pos08	pos18	UROGEN12	DIGEST13	actes_count	AUTRE04	age	UROGEN01
N39	68,5708646	0	63,139509	51,8117304	37,3573873	0,44050928	38,6984087	0	47,5243616	0	5,48682063	0	0,29153866	0,44050928
N41	2,4893918	0	63,2783693	52,2602919	49,5837779	66,2840388	38,8411534	60,9744388	0	0	0	0	1,92911947	0
J44	36,7855357	0	0,3823339	0	3,62646133	18,7854882	0	0	0	0	6,82444828	12,7465706	5,35122719	0,09733414
I50	473,048253	15,8923008	0	0	0	0	0	0	0	0	0	0	0	2,12129261
R10	295,03815	13,5474456	0	0,83472269	12,9115461	6,46453631	0,62604202	0	0	20,3728146	0,80806704	0	2,92152942	0
R06	318,552714	0	0	0	0	0	0	0	0	0	0	0	0	4,24736952
R50	272,961356	38,5731982	0	0	0	0	0	0	0	0	0	0	0	0
R53	709,983705	53,6844004	0	0	0	0	0	0	0	0	0	0	0	0
N10	0	0	0	2,1885557	2,05411032	3,25345521	2,02496511	0	0	0	0,79689452	0	2,16123233	3,31007815

Figure 4: Second run important variables chart for the B96 "Other specified bacterial agents" as secondary diagnostic: the table at the bottom of the figure represents the important variable table ordered descending each line represents the values obtained with the appropriate principal diagnostic.

coders encoding secondary diagnostics that are often neglected when coding, because related information are often spread into different information type. For this reason, we used the inpatient stay information available in PMSI database to build a decision tree that detects the missing secondary diagnostics. We studied a group of secondary diagnostic suggested by the responsible doctor of PMSI database. As a result, we built a model to detect the missing secondary diagnostics in addition to identification of the important variables used to determine them. The work is still preliminary we hope to validate the results on unseen data and generalize the result to include more secondary diagnostics and finally to validate the results by using national PMSI database.

ACKNOWLEDGEMENTS

We thank the Midi-Pyrénées region, the metropole Castres-Mazamet, INU Champollion and Paul Sabatier University for funding the research.

REFERENCES

Djennaoui, M., Ficheur, G., Beuscart, R., and Chazard, E. (2014). Improvement of the quality of medical databases: data-mining-based prediction of diagnostic codes from previous patient codes. *Studies in health technology and informatics*, 210:419–423.

Dubois-Lefrère, J. and Coca, E. (1992). Maîtriser lévolution des dépenses hospitalières: le PMSI.

Erraguntla, M., Gopal, B., Ramachandran, S., and Mayer, R. (2012). Inference of missing ICD 9 codes using text mining and nearest neighbor techniques. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1060–1069. IEEE.

Fetter, R. B. (1991). Diagnosis Related Groups: Understanding Hospital Performance. *Interfaces*, 21(1):6–26.

GUIDE (2006). Guide Mthodologique De Production Des Rsums De Sjour Du Pmsi En Mdecine, Chirurgie Et Obsttrique.

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

Pereira, S., Névéol, A., Massari, P., Joubert, M., and Darmoni, S. (2006). Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. In *MIE*, pages 845–850.

Pinaire, J., Rabatel, J., Azé, J., Bringay, S., and Landais, P. (2015). Recherche et visualisation de trajectoires dans les parcours de soins des patients ayant eu un infarctus du myocarde. In *3ème Symposium Ingénierie de lInformation Médicale (SIIM 2015)*, Rennes, France.

Potignon, C., Musat, A., Hillon, P., Rat, P., Osmak, L., Rigaud, D., Vergès, B., and Others (2010). P146-Impact financier pour les établissements hospitaliers du mauvais codage PMSI de la dénutrition et de lobésité. Étude au sein du pôle des pathologies digestives, endocriniennes et métaboliques du CHU de Dijon.

Therneau, T. M. and Atkinson, E. J. (2015). An Introduction to Recursive Partitioning Using the RPART Routines.

Tufféry, S. (2010). *Data mining et statistique décisionnelle: l'intelligence des données*. Editions Technip.

WHO. International Classification of Diseases (ICD)-10.