Increasing Alertness while Coding Secondary Diagnostics in the Medical Record

Ghazar Chahbandarian¹, Nathalie Souf¹, Rémi Bastide¹ and Jean-Christoph Steinbach²

¹University of Toulouse, IRIT/ISIS, F-81100 Castres, France
²Department of Medical Information, Pays d’Auten 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-
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.

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

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 cate-
gories, there are 2,049 categories and they are usually used for code prediction. 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 (Tuffery, 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.

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

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 preferred 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. (Tuffery, 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 algorithm
        Export the tree
        Cumulate the measurement of ‘goodness of split’ for each variable in a table.
        
    } 
    Export the important variables table
    } 
End.
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 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
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


