Predicting the primary medical procedure through personalization

Mamoun Almardini[1], Ayman Hajja[1], Zbigniew W. Raś[1,2], Lina Clover[3], and David Olaleye[3]

[1] Univ. of North Carolina, College of Comp. and Informatics, Charlotte, NC 28223, USA
[2] Warsaw Univ. of Technology, Inst. of Computer Science, 00-665 Warsaw, Poland
[3] SAS Institute Inc, Cary, NC 27513
{malmardi,ahajja,Ras}@uncc.edu
{Lina.Clover,David.Olaleye}@sas.com

Abstract. *Healthcare spending has been increasing in the last few decades. This increase can be attributed to hospital readmissions; which is defined as a re-hospitalization of a patient after being discharged from a hospital within a short period of time. The correct prediction of the primary medical procedure is the first step in the treatment process and considered as one of the main reasons for hospital readmission. In this paper, we propose a recommender system that can accurately predict the primary medical procedure for a new admitted patient, given his or her set of diagnoses. The core of the recommender system relies on identifying other existing patients that are considered similar to the new patient.*

Keywords: hospital readmission, main procedure prediction, clustering, personalization.

1 Introduction

Recently, expenditure on healthcare has risen rapidly in the United States. According [1], healthcare spending has been rising at twice the rate of growth of our income, for the past 40 years; the projection of the growth rate in healthcare spending is 5.8 percent during the period 2014-2024, which means that the spending will rise to 5.4 trillion by 2024. That said, the gross domestic product (GDP) growth rate is 4.7 percent (as of 2014) [2]. This increase can be attributed to several factors as listed by Price Waterhouse Coopers (PWC) research institute: over-testing, processing claims, ignoring doctors orders, ineffective use of technology, hospital readmissions, medical errors, unnecessary ER visits, and hospital acquired infections [3]. Figure 1 shows that 25 billion are spent annually on readmissions. Hospital readmissions and surgery outcomes prediction has taken a great interest recently [4–7]. Analyzing the reasons behind readmissions and reducing them can save a great amount of money. A hospital readmission is defined as a hospitalization of the patient after being discharged from the hospital. The period in average is 30 days [7].

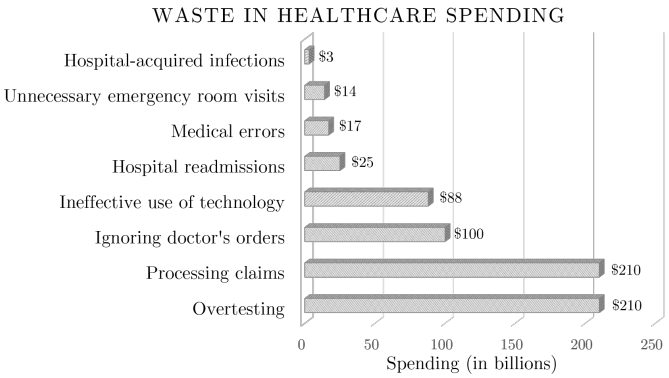

Fig. 1. Waste in healthcare spending as listed by Price Waterhouse Coopers (PWC) Research Institute [3]

One of the reasons for hospital readmissions is the wrong diagnosis of the patients. It is very important to provide the patients with the proper diagnosis in order to avoid any future readmissions and reduce the healthcare spending. In this paper, we extend our work [8] were we introduced a system for physicians that recommends diagnoses transitions which would, as a result, yield to reduction in the number of anticipated hospital readmissions. The input for our system were the set of diagnoses of a new admitted patient, and the primary medical procedure assigned for that patient. In the proposed recommender system however, we mine the medical dataset to predict the primary medical procedure for the patient by clustering the patients according to their set of diagnoses. We propose two approaches to identify the patients, from the dataset, that are similar to the newly admitted patient.

2 HCUP Dataset Description

In this paper, we used the Florida State Inpatient Databases (SID) that is part of the Healthcare Cost and Utilization Project (HCUP) [10]. The Florida SID dataset contains records from several hospitals in the Florida State. It contains over 7.8 million visit discharges from over 3.6 million patients. The dataset is composed of five tables, namely: AHAL, CHGH, GRPS, SEVERITY, and CORE. The main table used in this work is the *Core* table. The *Core* table contains over 280 features; however, many of those features are repeated with different codification schemes. In the following experiments, we used the The Clinical Classifications Software (CCS) [9] that consists of 285 diagnosis categories, and 231 procedure categories. In our experiments, we only used the features, listed in Table 1, that are relevant to the problem. Visit linkage feature *VisitLink* is an encrypted identifier of the patient. It can be used with the *DaysToEvent* feature to keep track of the patient's multiple visits. Each record

in the *Core* table represents a visit discharge. A patient may have several visits in the table. This table reports up to 31 diagnoses and up to 31 procedures per discharge as it has 31 diagnosis columns and 31 procedure columns. It is worth mentioning that it is often the case that patients examination returns less than 31 diagnoses. Furthermore, even though a patient might have gone through several procedures during a given visit, the primary procedure that occurred at the visit discharge is assumed to be the first procedure column. The Present on Admission Indicator *DXPOAn* identifies the diagnoses that were present when the patient was admitted. In addition to the features explained above, there are several demographic data that are reported in this table as well, such as race, age range, sex, living area, etc. Table 1 maps the features from the *Core* table to the concepts and notations used in this paper.

Table 1. Description of the used core table features.

Features	**Concepts**
VisitLink	Patient Identifier
DaysToEvent	Temporal visit ordering
DXn	n^{th} Diagnosis, flexible feature
PRn	n^{th} Procedure, meta-action
DXPOAn	Present on Admission Indicator

3 Predicting the Primary Procedure

In the previous section, we provided a concise description of our information system (HCUP), in which each instance (or visit) consists of one primary procedure and a set of diagnoses; when a new patient is admitted to the hospital, the physicians examine his or her set of diagnoses and assign a primary procedure accordingly. In [8], Almardini et al. introduced a system for physicians that recommends diagnoses transitions which would, as a result, yield to reduction in the number of anticipated hospital readmissions. The input for their system were the set of diagnoses of a new admitted patient, and the primary procedure assigned for that patient. The recommender system presented in [8] however, was not built to provide any recommendations on what the primary procedure should be. In this paper, we examine few approaches that address the challenge of predicting the primary procedure for a patient, given his or her set of diagnoses.

The goal of our system, which is to accurately predict the primary procedure for a newly admitted patient, is almost wholly determined by its ability to identify other existing patients that are considered similar to our admitted patient. The basis for determining similarities between different patients however, which we will explore next, is an intricate endeavor, given that the input of our patients is a set of diagnoses that differ greatly in the level of significance.

3.1 Minimum Similarity Match

The first approach that we propose to predict the primary procedure, is to have our similarity function be defined in a way that marks a newly admitted patient (p_n) similar to an existing patient (p_e) if and only if the existing patient exhibits every single diagnoses present in the admitted patient:

$$\text{similarity}(p_n, p_e) = \begin{cases} 1, & \text{if diag}(p_n) \subset \text{diag}(p_e) \\ 0, & \text{otherwise} \end{cases}$$

where 1 indicates that the new patient (p_n) is similar to the existing patient (p_e), and 0 otherwise. Consequently, we can define A_n as the set of all existing patients that the new patient (p_n) is similar to:

$$A_n = \bigcup_{i=1}^{m} \{p_i : \text{similarity}(p_n, p_i) = 1\}$$

where m is the number of existing patients in our dataset, and p_i is the i^{th} existing patient in the dataset. The final output for our recommender system is a probability distribution for the primary procedures obtained by the set of all similar existing patients (A_n). To demonstrate with an example, say we have a newly admitted patient (p_n) with the following set of diagnoses: diag(p_n) = {d1, d3, d5}, let us also assume that our dataset consists of the set of seven patients shown in Table 2.

Looking at our dataset of patients in Table 2, we can conclude the following:

- similarity(p_n, p_i) = 0, for i = 1 and 4
- similarity(p_n, p_i) = 1, for i = 2, 3, 5, 6, and 7

Table 2. Dataset S, containing all existing patients

	Diagnoses	Primary Procedure
p_1	{d1, d2, d5, d8}	Procedure 6
p_2	{d1, d2, d3, d5}	Procedure 3
p_3	{d1, d3, d4, d5, d9}	Procedure 6
p_4	{d1, d2, d3, d6}	Procedure 2
p_5	{d1, d3, d4, d5, d8}	Procedure 3
p_6	{d1, d2, d3, d4, d5}	Procedure 2
p_7	{d1, d3, d5, d6, d7}	Procedure 3

According to our previous definitions, A_n will contain the set of elements: p_2, p_3, p_5, p_6, and p_7; and the output to our recommender system will therefore be 60% Procedure 3, 20% Procedure 6, and 20% Procedure 2, which is the probability distribution of the primary procedures of A_n.

Table 3 shows a list of the accuracies for our system when tested on 815 randomly selected instances, each being compared to roughly 4 millions existing patients using our definition of similarity presented earlier. As can be seen in Table 3, the procedure with the highest probability in the existing matches distribution was predicted correctly 18.5% of the time; 23.6% of the time, the correct primary procedure was one of the two procedures with the highest probabilities; and 26.5% of the time, the correct primary procedure was one of the three highest probabilities procedures, so on and so forth. The frequency is the number of instances, out of the 815, for which the primary procedure was predicted correctly.

Table 3. Prediction accuracy of the minimum similarity match using the N most probable primary procedures

N	**Frequency**	**Accuracy**
1	152	18.5%
2	192	23.6%
3	216	26.5%
4	233	28.6%
5	243	29.8%
6	250	30.7%
7	260	31.9%
8	265	32.5%
9	268	32.9%
10	270	33.1%

Although the approach presented in this section is showing reasonably good results, the fact that our definition of similarities requires an existing patient (p_e) to exhibit all diagnoses of the new patient (p_n) makes this system rather limited. The first limiting aspect is the system's inability to find enough similar existing patients in the case of a new patient exhibiting many diagnoses; the second, and perhaps more important reason, is that the level of importance for each diagnoses (with respect to their abilities to predict the primary procedure) differ substantially, and that there are typically only a small number of subsets that are capable of determining what the primary procedure is.

3.2 Selective Similarity Match

In this subsection, we introduce an enhanced system for predicting the primary procedure for new patients. Our approach presented here is based on the fact that there is only a selected number of combinations for diagnoses subsets that are capable of predicting primary procedures. This means that for a new patient exhibiting x number of diagnoses, it would be more likely the case that matching our dataset for patients that exhibit only a subset of the x diagnoses will yield better result; by doing so, our system will not only avoid overfitting, but it will

also result in many more matches in our existing dataset, which will provide a higher level of prediction accuracy. The level of predictability for a subset of diagnoses s, can be determined based on the distribution of the primary procedures for existing patients that exhibit s; by calculating the entropy of main procedures for each subset of diagnoses, we can identify the subsets that are capable of most accurately predicting the primary procedure, which are essentially the subsets that have the least entropy values.

Our system starts by generating all possible combinations of k-diagnosis sets, starting with k=1 and ending with k=3, then calculating the entropy of the primary procedures for each combination. For each combination of diagnoses s, we identify all the existing patients that belong to s, then we calculate the entropy of s according to the distribution of its primary procedures:

$$H(s) = -\sum_{i=1}^{m} p_i \log(p_i)$$

where p_i is the probability of the i^{th} primary procedure, and m is the number of primary procedures in s.

The reason for why we stop at the number 3 is because the number of distinct subsets that can be generated from the set of all 285 diagnoses grows exponentially large as k increases. For example, the number of unique 3-diagnoses subsets that can be chosen from 285 diagnoses is roughly 4 millions; the number of unique 4-diagnoses subsets however, exceeds 250 millions. Table 4 shows few examples for some of low-entropy subsets extracted from a sample of 10,000 instances (existing instances), tested on a sample of 1,000 instances. The clusters in the table are sorted, from largest to lowest, by the number of patients who belonged to that cluster.

For a new admitted patient with x number of diagnoses, we generate all subsets of k-diagnoses for $k = 1, 2,$ and 3; then, using our previously calculated entropies for all possible diagnoses, we identify the subset of the patient diagnoses with the lowest entropy (highest level of predictability), and use its most frequent procedure as the anticipated primary procedures. Next, we provide a real example from our dataset to demonstrate the algorithm.

Let us first assume that the first step of the algorithm, which is to generate all possible combinations of k-diagnosis sets, starting with k=1 and ending with k=3 has been performed. Now, say that a new patient (p_n) has been admitted to the hospital with the following set of diagnoses {53, 98, 101, 164}:

- 53: Disorders of lipid metabolism
- 98: Essential hypertension
- 101: Coronary atherosclerosis
- 164: Hyperplasia of prostate

The next step would be to generate all 1-diagnosis, 2-diagnoses, and 3-diagnoses subsets of (p_n), which is shown in the first column of Table 5.

Table 4. Examples of the extracted clusters from our dataset

List of Diagnoses in Cluster	Entropy	Correctly Predicted Procedure	Accuracy
189 (Previous C-Section)	1.3297	134(Cesarean section)	100%
205 (Disc Disorders)	5.4023	158(Spinal fusion)	66.67%
82 (Paralysis) 141 (Stomach and Duodenum) 663 (Screening Mental Health)	0.0	70(Upper gastrointestinal endoscopy biopsy)	16.67%
49 (Diabetes Mellitus) 82 (Paralysis) 141 (Stomach and Duodenum)	0.0	70(Upper gastrointestinal endoscopy biopsy)	20%
181 (Other complications of pregnancy)	2.431	137(Other procedures to assist delivery)	100%
98 (Essential hypertension) 153 (Gastrointestinal hemorrhage) 250(Nausea and vomiting)	0.0	70(Upper gastrointestinal endoscopy biopsy)	75%
184(Early or threatened labor)	1.6627	121(Ligation or occlusion of fallopian tubes)	100%
61(Sickle cell anemia)	3.3063	222(Blood transfusion)	100%

Table 5. An example of one of the tested patients

List of Diagnoses in Cluster	**Entropy**	Primary Procedure
53	5.530	54
98	5.650	54
101	5.339	54
164	5.438	54
53, 98	-	-
53, 101	5.183	47
53, 164	5.19	54
98, 101	5.237	47
98, 164	5.112	54
101, 164	5.039	222
53, 98, 101	-	-
53, 101, 164	4.762	54
98, 101, 164	2.845	47

According to Table 5, the list of diagnoses that has the least entropy is {98, 101, 164}, in which the most probable primary procedure is 47 (Diagnostic cardiac catheterization coronary arteriography), which is indeed the correct primary procedure for our patient (p_n). Following is a description of the procedure codes found in Table 5:

- 54: Other vascular catheterization not heart.
- 47: Diagnostic cardiac catheterization coronary arteriography.
- 222: Blood transfusion.

4 Conclusion

Predicting the primary medical procedure for a new patient is of great help to physicians; since it gives them confidence of their medical decisions and helps to achieve the desired outcomes. In this research, we proposed a recommender system that can accurately predict the primary medical procedure for a newly admitted patient through finding the similarity with the old patients according to their set of diagnoses. We proposed two approaches to address the challenge of predicting the primary procedure for a patient, given his or her set of diagnoses. The two approaches showed a high level of predictability. However, the second approach is more accurate due to its ability of identifying the significant diagnoses that are responsible for the patient's admission. As a future work, we are planning to mine the procedures that are medically associated with the primary procedure and then recommend a set of procedures according to the patient's demographic and medical details.

Acknowledgment

This work was supported by SAS Institute under UNC-Charlotte Internal Grant No. 15-0645

References

1. Goodman, J.C., Priceless: Curing the Healthcare Crisis. Independent Institute, 2012.
2. Keehan, S.P., et al., "National health expenditure projections, 201424: spending growth faster than recent trends," Health Affairs 34.8 (2015): 1407-1417.
3. Coopers, Pricewaterhouse, "The Price of excess. Identifying waste in healthcare spending." (2006).
4. Touati, H., Raś, Z.W., Studnicki, J. and Wieczorkowska, A., "Mining Surgical Meta-Actions Effects with Variable Diagnoses Number," In Roskilde, Denmark, LNAI, Vol. 8502, Springer, 254-263 (2014)
5. Lally, A., Bachi, S., Barborak, M.A., Buchanan, D.W., Chu-Carroll, J., Ferrucci, D.A., Glass, M.R., Kalyanpur, A., Mueller, E.T., Murdock, J.W. and Patwardhan, S., "WatsonPaths: scenario-based question answering and inference over unstructured information," Technical Report Research Report RC25489, IBM Research, 2014.

6. Tremblay, M.C., Berndt, D.J. and Studnicki, J., "Feature selection for predicting surgical outcomes," In System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on, vol. 5, pp. 93a-93a. IEEE, 2006.
7. Silow-Carroll, S., Edwards, J.N. and Lashbrook, A., "Reducing hospital readmissions: lessons from top-performing hospitals." CareManagement 17, no. 5 (2011): 14.
8. Almardini, M., Hajja, A., Raś, Z.W., Clover, L., Olaleye, D., Park, Y., Paulson, J. and Xiao, Y., "Reduction of Readmissions to Hospitals Based on Actionable Knowledge Discovery and Personalization," In Beyond Databases Architectures and Structures - BDAS 2016, Conference Proceedings, Communications in Computer and Information Science, Vol. 613, Springer, 2016, 39-55
9. Healthcare Cost and Utilization Project (HCUP). Clinical classifications software (ccs), http://www.hcup-us.ahrq.gov.
10. Healthcare Cost and Utilization Project (HCUP), `http://www.hcup-us.ahrq.gov`