ATHABASCA UNIVERSITY

Healthcare data mining from clinical and administrative systems

By

Siu Hung Keith, LO

An essay submitted in partial fulfillment

Of the requirements for the degree of

MASTER OF SCIENCE in INFORMATION SYSTEMS

Athabasca, Alberta

March, 2012

DEDICATION

To my family members for their love, inspiration, spiritual support and encouragement

ABSTRACT

This research describes a framework to study the data collected in various patient care systems including patient demographic, patient location tracking, and external resource ordering systems. The focused data in this research consists of all the sitter usages within the hospital network, consisting of four adult sites and 1 child & adolescent site, for the entire years of 2008, 2009 and 2010. Due to confidentiality reasons, patient personal data will be removed such as name, address, phone number, etc. Only non-personal data will be used. Collected data will be analyzed using data mining approach to find out any association rules, clusters and patterns. The “discovered” data may uncover any hidden relationships between elements such as date, work shift, gender, age, reason, etc.

Results from mined data may provide hospitals hints to adjust their strategies on resource assignments, in order to better handle patient needs. At the same time, they may be able to provide the government a better picture of current problems in the society. For example, if a lot of agitated patients need to be monitored constantly in a specific month of the year, can there be anything special happening in that month such as temperature, humidity, economy, government policies, etc.? Many things may seem to be unrelated at a first glance. However, with the analyzed data, those “unrelated” things can then become “related” each other.

In addition, a novel method of "predicting" sitter case attribute value will be described. We propose a recommender system that can automatically classify an outcome of sitter case record with initial pre case user inputs, using the vector space model. The prediction accuracy of the system has been verified by comparing the predicted class label with the known class label in the dataset. In most cases, we were able to achieve a fair precision rate (~60%). Initial results showed that the system had been capable of correlating pre case attribute values with post case attribute values, with comparable performance than the Apriori association rule and Naïve Bayes classification engines. However, it could not provide trustable outputs for clustering results, as the recommender system was not originally intended to perform data clustering.

Data mined with proven algorithms revealed several interesting facts. From the results of association rule mining, female patients were more prone to commit suicide, whereas male patients were more likely to have physically related problems such as agitation and violence. Length of stay for agitated or suicidal patients was longer than disorientated patients. In classification, C4.5 offered higher prediction precision than Naïve Bayes in general. However, such precision gap became close when target attributes had fewer values (i.e.: fewer class labels). Naïve Bayes did very badly to classify tuples into correct class label, whereas C4.5 classifier was not much affected by increased number of class labels. In clustering, similar results were found, as the ones by the association rule mining. Teenage patients seemed to have "Suicidal" (mostly female patients)

and "Agitation/Violent" (mostly male patients) as the most popular reasons to require sitter supervision. Such finding appeared to be similar to the results from the association rule finding. For adult sitter cases, the clustering showed that sitter reasons, mission, site and age were the dominant factors, as those factors were strongly related to each other to form centroids.

ACKNOWLEDGMENTS

I would like to express my grateful thanks to all those who have worked or are working in the field of data mining research. I wish to specifically thank Dr. Maiga Chang for his expert knowledge and support during my experiments and writing of this project. I would also like to express my sincere thanks to the McGill University Health Center, Department of Nursing led by Ms. Patricia O'Connor, who allows me to use their internal clinical and administrative data to do my research and analysis. Special thanks to Dr. Judith Ritchie, who provided me advices and support on clinical aspects throughout this project. My sincere thanks also goes to the external examiner, Mr. Mehadi Sayed, for his insightful comments and valuable advises, which helped me improve the quality of this paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 MOTIVATION

Due to shortage of in-hospital health practitioners, sitter is hired as external resources to free up their time, letting them focus on jobs that require more skills. Sitter requests are captured by an in-house tracking system to generate electronic orders to the external agency. Although the reason of sitter use is documented in the tracking system, it is still unclear what really cause some patients constantly in need of sitters. It can be related to certain diseases, post-surgery effects, use of certain medications, etc. Real causes can be so complex that not even experienced health practitioners can easily identify.

With immense number of patients in public hospitals, reviewing long histories of patient charts and doing manual data analysis can be tedious work. Although statistical reports are being generated to report usages, not much is being done to turn data into real knowledge (i.e.: find out any interesting patterns and relationships between different clinical and non-clinical elements). Currently, the only feedbacks on sitter usage are only record counts and dollars spent. A lot of valuable information may still be buried. Such information may be important indications to correlate different clinical and non-clinical factors, which may be critical to provide both management and health practitioners important

information to improve healthcare. Otherwise, some clinical services such as sitters can only be provided reactively to treat symptoms.

There is a wealth of data available in healthcare institutions [1]. However, there is a lack of researches done to discover hidden and meaningful patterns and data trends. The nature of healthcare services is essentially information-based and can be greatly improved with effective information support, including data modeling, archive, retrieval and analysis [2]. So, from the abundantly collected data, valuable information can be uncovered with data analysis. With tighter budgets and dynamic evolution of health service needs, there is an ever-increasing need for efficiently leveraging resources at the existing health institutions [3]. This includes understanding the demand patterns, optimizing resource and facility utilizations.

Data mining methods offer the potential to decipher the service demand and patterns to effect strategic long term and pragmatic short term planning. A method for understanding external clinical resource data will be described by processing and analyzing the collected data using data mining. Data mining has attracted a great deal of attentions in the information industry in recent years, due to wide availability of huge amounts of data and imminent need for turning such data into useful information [4]. It can be used to uncover patterns in collected sample data [5]. Years of effort in data mining have produced variety of

techniques and applications. However, not all of them can be used to produce meaningful outcomes.

This research will investigate mining methods that are more effective and appropriate for the given dataset. Given knowledge found from mined clinical data, many studies have reported that computer-assisted expert systems, such as recommender system, can be built to help health practitioners and patients make reliable diagnoses and management decisions [6, 7]. Clinical recommender systems are active knowledge management systems that bundle basic clinical experience and knowledge with specific information about cases and patients. The main part of a clinical recommender system consists of a knowledge-based rule engine, along with supplemented knowledge of health practitioners and their experiences to validate any generated recommendations. Recommendations generated by the system are being used to support health practitioners in making decisions on short and long-term therapies. Also, those recommendations can assist them in making adjustments to healthcare strategies.

1.2 PURPOSE

The purpose of this research project is to provide a framework to analyze sitter usage data in relation to non-nominative patient information by making use of data mining techniques to uncover association rules, classifications and clusters. Due to different data suitability of data mining techniques, collected data

will first be analyzed with rationale, in order to determine the appropriate techniques to be used. The analysis will be used as a reference to provide healthcare administrators to fine tune staff proportion to better respond patient needs, and adjust certain procedures to carry out treatments more effectively.

From collected data in various patient care systems, data mining techniques and regular expressions will be applied to find out and represent specific patterns about the usage of external resources to watch patients who are “at risk”. Regular expression like technique will be used in our recommender system to perform class label prediction, which can be served as an advisor to help health practitioners find out factors why some patients require higher cost of care and lengthier stay than the others.

1.3 DELIVERABLES AND CONTRIBUTIONS

Works done in this essay can be summarized in two major parts. First, it examines a new way of "predicting" sitter case attribute value using sequence matching, regular expression and vector space model techniques. To better demonstrate how the concept works, a recommender system has been developed that can automatically classify an outcome of sitter case record with initial pre case user inputs.

To find out whether the recommender system is effective in outcome prediction, data mining with proven algorithms has been applied to current dataset. This serves two purposes. It provides results that can be used as

references to compare with the ones generated by the recommender system. It is being used as a base reference to validate how far the prediction from our recommender system is to the ones by proven algorithms. Also, results generated from proven algorithms are trustable to provide health practitioners information about relationships between different clinical and administrative parameters, as well as their characteristics.

1.4 ORGANIZATION OF SECTIONS

The research will be organized in following different sections.

- Chapter I: Motivation, Contributions and Goals
- Chapter II: Background information and mode of inquiry
- Chapter III: Preparation work
- Chapter IV: Recommender system using sequence similarity measure
- Chapter V: Experiment and results
- Chapter VI: Evaluations and discussions
- Chapter VII: Conclusion and future works

Chapter I describes research goal and purpose of this project. It talks about why this project can be useful in providing health administrators useful information about values of data mining on collected data. Chapter II discusses theories and relevant works that researchers have done in the domain of data

mining and regular expression on collected healthcare data mining. They are being used as basis and references to perform various data mining activities on our collected data and to build our recommender system. Also, this chapter talks about the approach and methods being used to perform experiments and data analysis on collected data, as well as to build the recommender system. Chapter III describes how data gets collected from various systems and the processes needed to get it ready to be used. Processed data is being used as reference data (as known as knowledge/experience) for our recommender system. Also, it is being used in data mining with existing algorithms, in order to generate comparison information with our recommender system. Chapter IV focuses on the recommender system design concept and how the experience data can be used to perform unknown outcome prediction. A full example of the entire prediction process is being described as well. Chapter V discusses how the recommender system is being evaluated in a quantitative manner and its performance comparing to the historical data. It also talks about the effectiveness of the recommender system and discusses important findings throughout the experiment with different input parameters. Chapter VI focuses on generating results from existing data mining algorithms then compare them with the ones predicted by our recommender system. Chapter VII summarizes this essay and talks about possible future works.

CHAPTER II

REVIEW OF RELATED LITERATURE

2.1 HEALTH DATA MINING

In healthcare setting, a lot of information about patient episodes is being recorded in various systems. Reports are being generated but they mostly contain only counts, sums and groupings of collected data. Although some manipulations are being done to those reports to facilitate data representation, they are mostly visual appeals or at the most, pivot tables, which do not necessarily provide more knowledge or discovery of new information. Clinical administrative data such as sitter data is not being used sufficiently. Without knowledge discovery using data mining techniques, a lot of information may still be hidden. With only numbers showing on statistical reports, more complex questions about patient care cannot be answered. For example, there may be relationships between gender, culture, length of stay, diagnosis, etc. that can affect patient's need for sitters.

A lot of new and useful information can be discovered with the use of data mining framework. According to Fayyad, Piatetsky-Shapiro and Smyth [8], the tremendous amount of data collected and stored in large and numerous data repositories has far exceeded our human ability for comprehension without machine aided analysis. It is almost unavoidable to have data entry errors or inconsistencies in huge datasets. With data mining techniques, outliers can be

spotted out and further analysis can be done to determine if those are erratic entries. Analyzed data can be served as feedbacks to how system is being used. It may indicate areas of improvements (problematic areas to be addressed).

Data mining performs data analysis that may uncover important data patterns, contributing greatly to business strategies, knowledge bases and scientific and clinical researches. Depending on the nature of clinical administrative data, some different mining techniques will be evaluated. Mined results will retrieve knowledge from the data. However, not all data can equally provide same amount of knowledge. Having the mining exercise done to the dataset can allow us to identify useful data vs. less useful one. It can also serve as a reference to fine tune the system to capture more useful data in the future.

Many meaningful patterns can be analyzed and extracted then be compiled into regular expressions [9, 40]. Regular expression is a metalanguage[1] that describes finite-state automata used to recognize string patterns [10]. It is a context-independent syntax and compact way of describing complex patterns in texts. The purpose of its use is to represent a wide variety of character sets, character set orderings and patterns of characters, in a systematically manner that allows easy identification of patterns inside a sequence. It is generally associated with text processing on text strings. In our recommender system, we

[1] Metalanguage is a language or set of vocabularies used to describe another language.

make use of it to identify repetitive patterns in text sequences generated from a column value. For example, consider the following sequence

AAABBBCCCABBCCCC

Such sequence with repeated patterns can be expressed in the following regular expression

(A+)(B+)(C+)

By looking at the above regular expression, one can identify the patterns (i.e.: repetition of A's, followed by repetition of B's, followed by repetition of C's) more clearly.

Regular expression has been employed in various information extractions and provides an approach to interpret patterns systematically. Researches have been done [50, 51, 52, 53] on transcribing then analyzing physicians' notes, based on regular expressions. Regular expression techniques were used to classify then express data in string search patterns. Statistical methods could then be applied to the collected dataset to find out number of occurrence of particular string patterns. There were also works done using regular expression on prediction analysis [54, 55]. A model based on regular expression and sequence analysis has been proposed to predict outpatient paths and patient flows [54]. Particularly, a diagnosis system [55] has been developed combining

the use of regular expression and simple vector space model[2], which is being used in our recommender system as well. The extraction of knowledge from textbooks was used as the knowledge base in Chinese medicine to provide readers hints about symptoms and possible treatments. The usefulness of such data analysis does not limit to only clinical fields. With text sequence analysis, a research [56] has been done to make use of existing helicopter maintenance records to predict future maintenance needs. Although the researches could not always achieve very accurate prediction results, they can be used as the guiding principle in our recommender system.

Although health authorities and investigators always have interests in extraction of findings and problems from clinical systems, most existing manual and software provider-initiated reporting systems generally only output summaries of numbers, without useful description about patterns and relationships between different administrative and clinical data. With the use of regular expressions, patterns and relationships can be further represented and analyzed more easily.

A recommender system can be built based on analyzed patterns and regular expressions [11]. It can assist in making recommendations from all kinds

---

[2] A simple vector space model is an algebraic model to represent text documents as vectors of identifiers. In our case, term frequencies in sequence strings are used to build vectors.

of collected user sources. For instance, sophisticated page-ranking algorithms employed by web search engines such as Google have greatly improved the relevance of documents retrieved. Similar concept can also be applied to healthcare records. A recommender system can potentially be used to facilitate and improve the healthcare process, particularly when coupled with patient care and clinical information systems. Interesting patterns are being discovered and proposed by the recommender system as hints, which may assist healthcare institutions to alert health practitioners about some higher risk patients and find out factors why some patients require higher cost of care. For example, some patients with aggressive behaviors can be related to side effects of certain treatments and medications. A recommender system can make use of knowledge learned from mined data and express it in a semantic way, which can be more easily understood by most people.

2.2 MODE OF INQUIRY

The primary outcome of this research is to discover new knowledge and provide framework to make use of data mining on various data across systems. Most healthcare institutions collect huge amount of data in different systems. However, systems may use heterogeneous data sources. Each system contains valuable information but within a limited scope. It will be powerful to combine data from different data sources to perform data analysis, since combined data contains information from different aspects and perspectives. Therefore, data

integration is required to relate different types of databases, before data analysis can be performed effectively to bring more values.

Many previous works in knowledge discovery and data mining [12, 13, 14, 15] have been done in healthcare domain with human subjects. The experiment in this project will be carried out with only denominalized data, specifically hospital mission, site, unit, patient's gender, age, marital status, reason for sitters, date of incident, shift, city, length of stay and discharge location. Experiences and challenges by other researchers can be served as guidance and research basis in this project. To ensure gathered information is as accurate as possible, only academically recognized sources will be used such as ACM, IEEE, Ovid, PubMed, etc.

Integrated data sources will first be cleaned up to filter out any outliers and erroneous entries. Depending on the nature of data, multiple mining techniques will be applied against same datasets to find out association rules, classifications and clusters. The time spent in each data mining process will be recorded. It will be used to compare time spent across different automatic knowledge discovery algorithms. Results from mined data (knowledge) will further be analyzed and interpreted using various clinical and non-clinical explanations. A recommender system will be developed as an electronic “advisor” to healthcare administrators, based on regular expression like technique using n-length character finding.

Reference characteristics collected from historically and future sitter cases

will be used as a basis to perform class label prediction of unknown post case attribute values. More sitter cases are collected in the system, more statistically relevant the prediction will become. So, the system will be the electronic learner as well as the advisor. The use of it can be expanded to other systems as well.

2.3 DATA MINING METHODS

2.3.1 Association rules

Association rules mining is one of the most popular techniques in data mining [6, 7]. It is an unsupervised machine learning, in which class label from the training dataset is not known or defined. An association rule implies certain association relationships among a set of objects [7]. The general aim of association rule mining is to find frequent patterns, associations, correlations and structured patterns among sets of items in datasets. It has attracted a lot of attention in current data mining research due to its capability of discovering useful patterns for applications such as decision support, forecast and clinical diagnosis. In general, association rules mining looks for rules in a database that satisfy the predefined minimum support and minimum confidence.

The goal of association rule mining is to detect associations between specific values of categorical variables in large datasets. The use of machine learning techniques in healthcare is typically to find out relationships between health parameters and outcomes from a combination of symptoms, using records of already observed patients. Mined data is being used to build a domain model

applicable for identification of future observations. Clinically related databases often contain hidden frequent patterns. Some items can appear frequently together as frequent itemsets. Some subsequent itemsets that appear as a result from frequent itemsets are sequential patterns. In such case, association rules mining comes in handy.

Association rules have proven successful records with market basket analysis on marketing census and financial data [4]. Association rule discovery in clinical records can improve diagnosis, when multiple target attributes are being used. It deeply searches for hidden patterns, making them suitable for discovering predictive rules involving subsets of the clinical dataset attributes [17, 18]. Also, it can help find and confirm the relationship between different clinical and administrative parameters. However, the following issues must be considered when using association rules mining

1) Clinical dataset may contain a lot of irrelevant association rules. The meaningfulness of rules must be judged by experienced health practitioners.

2) Some relevant rules with high significance may show up with very low support.

3) Number of association rules found can be extremely huge.

Because of the above, search constraints need to be carefully defined to restrict the number of association rules. At the same time, it can also accelerate the rule mining time.

Finding association rules can be very useful to relate what relations data has. Some relations can be very helpful to identify the issues that health practitioners should pay more attention to. For example, does a specific age group have more tendencies to commit suicide? Some rules can help unit managers better plan their staffing level of different specialties at different period of the year. For example, do we have more suicidal patients between the age of 30 and 39 during the fall season? Having this interesting information can also help the healthcare research.

2.3.1.1 Apriori algorithm

Apriori is an algorithm to discover association rules based on exhaustive frequent itemset searches. It is proven to be one of the popular data mining techniques used to extract association rules [24] from transactional databases. With sound data structures and careful implementation, it has been proven to be a competitive algorithm in the contest of Frequent Itemset Mining Implementations (FIMI) [25]. Although it is not the most performing association rule mining algorithm [24], it serves as the basis of many other variants, which offer improved performance and scalability. With support and confidence as inputs, Apriori discovers association rules based on those inputs as filtering

criteria. An itemset means a set of items that occur together. The k in a k-itemset means number of items in an itemset. For instance, if item A and B appear to be together for 5 times in 100 records, the 2-itemset (i.e.: k-itemset where k=2) has support value of 5/100 = 5%, denoted as follows

X = {A, B}

supp(X) = 5/100 = 5%

The support of an itemset X, supp(X), is the proportion of transactions in the dataset, which contain the itemset. The confidence is the probability if itemset X exists, itemset Y also exists. It is denoted as conf(X => Y). It is calculated by dividing the combined support value of both itemsets X and Y over the individual support value of itemset X, as follows.

conf(X => Y) = supp(X U Y) / supp(X)

The confidence value is the probability to represent how likely the itemset Y exists, if the itemset X exists.

In simpler words, the support value represents the counts of the itemset occurrence and the confidence value is the likelihood that both itemset X and Y appear together. Apriori uses those two criteria as the minimum lower bound to filter out any itemsets, which do not satisfy either support or confidence values.

The principle of Apriori is "if an itemset is not frequent, any superset of it cannot be frequent either." [4] The reason why the algorithm is named "Apriori" is

because it uses "prior" knowledge of frequent itemset properties found in the dataset. The algorithm employs breadth-first tree search (BFS) concept by finding 1-itemsets first. Then, 1-itemset will be used to find 2-itemsets and so on, until no more frequent k-itemsets can be found. Candidates of k-itemsets are generated from cross join of (k-1)-itemsets. Only itemsets that can satisfy predefined minimum support will be kept, otherwise, will be pruned during each candidate itemsets generation. To generate k-itemsets, the entire dataset needs to be traversed k+1 times. Rules are being generated based on frequent itemsets found. Only rules that can satisfy the predefined minimum confidence threshold will be kept.

The Apriori algorithm is widely used in the field of data mining. Hospital research has been done using the association rule method, specifically Apriori algorithm, for automatically identifying new, unexpected, and potentially interesting patterns in infection control [26]. In addition, a study [27] has been performed that used the Apriori algorithm to generate the frequent itemsets and designed the model for economic forecasting.

2.3.1.2 Predictive Apriori algorithm

Although the Apriori algorithm offers good results on association rule generation, it heavily relies on user's inputs. The algorithm employs support and confidence thresholds to return only rules, which must be above these lower bounds. It does not make prediction for all database records. For instance, it

does not predict that truly correlated items will more likely correlate in future data. In Apriori, having support or confidence too high will sacrifice a lot of interesting rules. However, some meaningful rules may have either only high support or high confidence but not both. Predictive Apriori algorithm is a modified version of Apriori, which is based on probabilistic model [28]. Without user's input on support and confidence, the algorithm can generate interesting rules based on expected predicted accuracy. It searches association rules with an increasing support threshold for the best n rules concerning a support-based corrected confidence value. This resolves the issue of defining a good balance between support and confidence values, while maximizing the probability of making an accurate prediction for the dataset. The algorithm discovers the rules with corresponding expected predictive accuracies as the output using the Bayesian method. According to Scheffer [28], the definition of predictive accuracy is as follows.

Let D be a data file with r number of records. If [x → y] is an association rule which is generated by a static process P, the predictive accuracy of [x → y] is

c([x → y]) = Pr[r satisfies y|r satisfies x]

where distribution of r is governed by the static process P. The predictive accuracy is the conditional probability of x → r and y → r. The rule is based on a concept of “larger support has to trade against a higher confidence”. The rule

engine tries to maximize the accuracy instead of the confidence in the Apriori algorithm.

2.3.2 Classification

The discovery of knowledge from clinical databases is important to provide references to health practitioners to help them in more effective decision making. Health practitioners must use reasoning and judgment to make decisions [19]. Some decisions must be made without exactly knowing whether the outcome will be positive or not. Health practitioners are instructed to handle such uncertainties based on best available evidence of practice.

However, not every problem has absolute solutions. When there is no formal model or exact solution to a problem, "learning from examples" is a way to provide health practitioners guidance, based on already observed data.

The aim of data mining is to extract knowledge from data and generate clear and understandable description of patterns. Data mining can play a significant role, because of its ability of uncovering clinical evidence from large volumes of clinical data in a visual way [20]. Classification is another powerful technique, which produces interpretable result and thus widely used in clinical purpose. It is known as supervised learning, in which class label from the training dataset is known during data analysis.

A major advantage of classification on clinical datasets is that a decision tree can be generated. It consists of splitting the training dataset by the defined

class label recursively, in order to find out possible sub-populations (i.e.: classes). Each subset contains more or less homogeneous[3] states of predictable attribute. At each split in the tree, all input attributes are evaluated for their impact on the attribute. The recursive splitting process completes when all splitting attributes are used up to split the population into subsets. Then, a decision tree is generated. Since the decision tree is generated based on evaluation of historical data, it can be served as a visual and analytical decision support tool. Based on the generated decision tree, health practitioners may be able to see a clearer picture of what will most likely to happen with the series of decisions or interventions being made.

2.3.2.1 Naïve Bayes Classifier

The Naïve Bayes Classifier is an algorithm from the Naïve Bayes theorem, based on Bayesian statistics[4]. The idea was developed from the conditional probability of a particular event. This method makes use of the probabilistic model to the supervised learning system via training, in order to employ the maximum likelihood pattern. Although the algorithm is relatively simple, it is very capable to take on real life situation or problems. However, due to its intensively

---

[3] A subset is pure or homogenous, if it contains only a single class.

[4] Bayesian statistics is a subset of statistics for describing uncertainty using the mathematical language of probability.

iterative nature, if the dataset is large and each tuple[5] has many attributes, the algorithm will take very long to find the solution. The algorithm is suitable to deal with relatively smaller amount of training data, such as the currently collected datasets.

The largest dataset in our experiment only contains 22229 tuples, with 14 attributes. With the assumption that each attribute is not related to each other, the Naïve Bayes algorithm tries to maximize the likelihood of the probability that the tuple should belong to, by evaluating each attribute independently. Although the name contains the word "Naïve", according to past experiences from other researches [29], it can usually give relatively good classification results.

According to the Bayes rule, $p(C|F_1, ..., F_n)$ is the posterior probability of class membership, where the probability that F belongs to C. In other words, given known pre case attribute values (F_1, ..., F_n), the probability of having a class label C is being optimized. Naïve Bayes assumes that the conditional probabilities of the independent variables are statistically independent. Each attribute is being evaluated independently to get the likelihood of the class label. The algorithm can be summarized with the following equation.

$$classify(f_1,...,f_n) = \operatorname*{argmax}_c p(C=c)\prod_{i=1}^{n} p(F_i = f_i \mid C = c).$$

[5] A tuple refers to one row of record, containing multiple attribute values, in the sitter database.

where p(C) is the known class label and $p(F_1, ..., F_n|C)$ is the pre case sitter information, given known class label C.

Simply speaking, Naïve Bayes Classifier does the prediction by evaluating the conditional probabilities of different class labels. The one having the highest conditional probability is considered to be the predicted class label. The algorithm considers attribute value of each tuple independent of each other. Each attribute value in each tuple is being evaluated with the class label value to calculate its conditional probability. To put the algorithm in practice, let us do the following example, assuming we have the following tuple X.

X = (Reason=Suicidal, Age group=30-39, Gender=Male, Marital status=Single Adult)

Let "gender" be the chosen class label C having possible values,

C_1 = Male

C_2 = Female

The following probabilities will need to be evaluated.

$P(C_1) \rightarrow$ P(Gender = Male)

$P(C_2) \rightarrow$ P(Gender = Female)

Conditional probabilities $P(X|C_i)$ where i = 1 or 2, due to only two class label values present. $P(X|C_1)$ means the probability that X is single adult suicidal patient between age of 30 and 39, given that the patient is "Male".

$P(X|C_1)$ = P(Reason = Suicidal | Gender = Male) x P(Age group = 30-39 | Gender = Male) x P(Marital status = Single Adult | Gender = Male)

$P(X|C_2)$ = P(Reason = Suicidal | Gender = Female) x P(Age group = 30-39 | Gender = Female) x P(Marital status = Single Adult | Gender = Female)

To determine the class of this tuple, we need to find the class Ci, which maximizes $P(X|C_i)P(C_i)$

Comparing the results of

P(X|Gender = Male) x P(Gender = Male) and

P(X|Gender = Female) x P(Gender = Female),

the gender (C_i) in the one that has the highest probability becomes the predicted class value.

2.3.2.2 Decision Tree (C4.5)

C4.5 is a decision tree algorithm, which became a benchmark [48] of newer supervised learning algorithms. The model is an improved model of ID3 classification algorithm [47], which provides higher accuracy of prediction [46]. It is a greedy approach so backtracking is not possible. Each attribute is being

tested and calculated on the information gain[6], which the algorithm tries to maximize. Information gain is the difference between the information before and after splitting. Information is calculated based on entropy, which is a measure of data disorder. Given a probability distribution P_x where x represents each element's index, the information required to predict an event is the distribution's entropy. A uniform distribution results in a splitting criteria with high entropy, whereas a varied (non-uniform, with peaks and valleys) distribution results in low entropy. An information gain is the difference in entropy that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to be the splitting criteria. Through a top-down approach and evaluations on information gain, the algorithm chooses the best test to split the data to create a branch. The process runs recursively until all attributes are tested and split with maximized information gain.

2.3.3 Clustering

Clustering is a process of partitioning a dataset into meaningful sub-classes. Sub-classes are meaningful if they can help users understand the natural grouping in a dataset. It is an unsupervised approach, in which class label from the training dataset is not known or defined. In other words, clustering is the

[6] Information gain is the change in information from one state to another. More difference between two states results in higher information gain.

unsupervised classification of patterns into groups, as known as clusters. Within a cluster, it contains a collection of data objects that resemble to each other. The purpose of clustering is to group the objects based on the principle of maximizing the intra-class similarity, while minimizing the inter-class similarity.

Clustering analysis has proven records and plays a long standing important role in various fields, including its use on healthcare data [21, 22]. In clinical environment, patient information consists of many attributes. With abundantly new data collected in hospitals, similar cases may not be obvious to be noticed. There may often be hidden correlations between different attributes. In order to identify specific groups of patient population, all attributes must be assessed as potential targets from the full dataset. Grouping patients of similar type together can help health practitioners address treatment and take precautions more accurately. The application of knowledge discovery and data mining techniques potentially help health practitioners improve their daily tasks by finding out something special about a particular patient population [23].

2.3.3.1 K-means

K-means clustering is one of the most well known and commonly used partitioning clustering methods [32]. K-means is an algorithm to group the objects together, based on attribute features into K number of groups (clusters). The

main idea is to first randomly define K centroids[7]. The algorithm aims to minimize the sum of squared distance between an object to the centroid, which is called the sum of squared error. According to the Euclidean distance between each point and centroid, each point is associated to the cluster by the closest distance. The centroid is then recalculated based on the mean of all the points within a cluster. The algorithm keeps iterating until there is no change or very little change of the centroid position [33].

2.3.3.2 Expectation-Maximization (EM)

The EM algorithm is used to approximate a probability function, which is related to K-means algorithm. It is explained in detail in Witten and Frank (2001). Unlike K-means, instead of basing on cluster mean to assign object to appropriate clusters, the cluster assignment is based on the probability of cluster membership of the object. It consists of a two folded process – expectation and maximization of it. The goal of EM clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data. It attempts to approximate the observed distributions of values, based on mixtures of different distributions in different clusters. Each object is given a probability to belong to a cluster. Then, cluster centers are recomputed

[7] A centroid is the mean value of all the objects in the cluster.

based on the average of all objects weighted by their probability of belonging to the cluster.

As in K-means, the algorithm starts by first randomly assign objects to represent centroids. Then, the expectation step computes the probability that each data belongs to each cluster. The maximization step computes the distribution parameters and their likelihood to belong to a cluster. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function. At each iteration, the likelihood of data that belongs to specific clusters is being optimized - increased quality of cluster. The algorithm stops when the iteration can no longer provide increased quality of clustering.

CHAPTER III

PREPARATION

3.1 DATA SOURCE

The proposed recommender system and data analysis require data, which consists of all the sitter usages within the hospital network across five hospitals (4 adult sites and 1 child & adolescent site), for the entire years of 2008, 2009 and 2010. Sitter is an external on-call resource hired to "watch" patients who are at risk and need constant supervision. In case something happens to the patient, the sitter informs nurses for intervention. It is more cost effective to use sitters to watch patients because

- clinical knowledge is often not required
- lower wage than health practitioners
- on-call basis; hospitals call the sitters in and only pay for the shifts used

According to the hospital's guideline, no sitter orders can last for more than one entire shift. For patients who need sitter supervision for more than one shift, additional orders must be placed. In every order, hospital wards must provide a primary reason (chosen from a predefined list) to indicate why the patient requires sitter supervision. The reasons can be one of the followings.

- Agitation

- Avoid use of 4-point restraints
- Avoid use of Other type of restraint
- Avoid use of poset vest/jacket
- Away without leave
- Behavior problem
- Constant observation in 4-point restraints
- Delirium
- Dementia
- Disorientation
- Eating disorder
- Psychosis
- Risk of falls
- Suicidal
- Trauma
- Violent
- Youth Protection

- Other
- The original data consists of
- Date of order
- Department's mission (aka sub-division)
- Order shift (Day, Evening, Night)
- Primary reason (Patient's problem)
- Units that placed the order
- Health practitioner who placed the order
- Health practitioner (supervisor level) who authorized the order
- Patient's medical record number
- Patient's family and first name
- Patient's gender
- Patient's primary spoken language
- Patient's bed number and location

Data is gathered from different administrative and clinical systems. The sitter system depends on the hospital's ADT system[8] to get more detailed patient information. Other than the medical record card number and basic information about the patients, the sitter system does not store any other patient specific information but information about the sitter case.

With data consolidation between the sitter system and hospital's ADT system, it is able to provide us the following non-nominative data about the patient.

- Date of birth → Age
- Gender
- Marital status
- Preferred language
- Municipality
- Diagnosis
- Type of admission

[8] ADT system refers to admission, discharge and transfer system. The system contains patient's identification and general information about each hospital visit, including the date/time stamp, location, diagnosis, etc.

- Date of admission → Length of stay
- Date of discharge → Length of stay
- Discharge location

With some calculations on the above data, age and length of stay can be obtained easily.

3.2 DATA COLLECTION

Data will be first collected from sitter administration and hospital patient tracking systems.

Currently, because of heterogeneous information systems running across healthcare institutions, clinical and administrative data is stored in all kinds of proprietary formats, in a multitude of clinical information systems available on the market. They include relational database tables, structured document-based storage in various formats, and unstructured document storage in proprietary document formats. This results in a severe interoperability problem in the healthcare informatics domain. Since the experiment requires data from different systems, data must first be related across them to form a consolidated data source, before data analysis can be performed. Fortunately, most SQL-based

database management systems offer a common ground to communicate to each other via ODBC[9].

In our experiment, the sitter system is based on MSSQL and ADT system is based on MySQL. ODBC data gateways can be used to join data from both systems into MSSQL format. Data transformation with specific queries is needed to convert then port some incompatible data types from MySQL into MSSQL. It is an offline operation which will not be needed in the future. The anticipated future sitter system will make use of message queues, interface engines, data adapters and Web services to communicate with any other corporate systems.

Then, data from different systems will be merged logically according to the patient's medical record numbers and hospital site. Once merged, data columns that contain any patient specific or any information that can lead to find anyone will be removed (denominalized). This is in agreement with the hospital's research ethics board (REB) to protect patient confidentiality. After all, nominal information is not necessary for this research.

Data cleanup will first be done to identify possible erratic entries and eliminate non-useful data. Some data will be discretized into uniform ranges (concept hierarchy generalization) to facilitate data analysis. Data mining will

[9] ODBC refers to Open Database Connectivity. It is a middleware application programming interface for accessing different database management systems.

then be performed to the merged datasets. Data mining techniques will be selected to best suit the nature of data. Depending on the outcomes, repeated analysis may be needed to find out more knowledge from the datasets. Results from mined data will be analyzed and represented. A framework for building a recommender system will be described.

3.3 DATA PREPROCESSING

Most of the software systems nowadays use databases as the backend to store data. However, depending on the software and database design, data quality issues can originally exist in applications and databases. Once different sources are being combined, the issue can be amplified. In real world situations, raw data collected from different systems is often dirty. It can be

- incomplete

  e.g.: missing attribute values, values missing enough level of detail

- noisy

  e.g.: out of range / outlier values, exaggerated values which do not make sense

- inconsistent

  e.g.: free text user inputs, use of complete terms vs. abbreviation vs. non-standard user-defined terms, discrepancy between different records

- duplicated

Dirty data leads to poor data quality and cannot be mined effectively. Data quality is an important contributor in the overall success of any data mining analysis [4]. Inconsistent data leads to wrong assumptions and analysis. Duplicated data can cause issues such as slowdowns and incorrect record counts, which can lead to incorrect data analysis output. These errors in the data arise due to mistyping of the word, or when the data is collected from different sources. This happens especially to clinical databases. Large quantities of information are being collected about patients and their clinical conditions. But, they may be collected as free text or non-standardized data formats. This causes erratic entries and inconsistencies within the database. Many different records may contain contradictory information about the same entity. Before any data mining can take place, data preprocessing is almost a must to do. Data preprocessing contains the following tasks [4, 16].

- Data cleaning

  fill in missing values, remove outliers and correct inconsistencies

- Data integration

  merges data from multiple data sources

- Data transformation

  normalization, noisy data smoothing, aggregation/generalization (roll-up)

- Data reduction

  discretization, selective attributes removal, replacement of values into smaller data representations

In data preprocessing, data cleaning is a vital part to start with. It tackles problems and cleans the database from duplicated data, outlier values, mistaken entries and incorrect information. "Cleaned" data cannot be used right away for analysis. It may not always be in a form that is appropriate for systematic analysis. Data transformation in data preprocessing standardizes the data for further computation and improve the quality of the data for mining. The process helps remove noisy data, transform the data into roll-up classes (also reduces data) and scale data within defined ranges (normalization).

Normalization is especially useful in clustering. Clustering classifies data into different group by observing difference between data. However, some attributes tend to have greater difference than the others. For example, length of stay vs. age. Length of stay between patients can be different as many as hundreds of days, but this kind of difference can never happen to ages of patients. As a result, attributes with usually greater differences dominate the clustering result, while the others tend to be "ignored".

Before doing any clustering to the dataset, every attribute in the dataset must first be normalized. This is because some attributes can have very large range of numeric values and some others may have much narrower range. For

instance, in the pediatric sitter case dataset, age could vary from 0 to 18 years old and length of stay could vary from 0 to over 100 days.

Without normalization

Tuple 1: Age = 0, Length of stay = 10

Tuple 2: Age = 1, Length of stay = 30

Tuple 3: Age = 17, Length of stay = 15

Euclidean distance between tuples 1 and 2: $\text{sqrt}((0-1)^2 + (10-30)^2) = 20.025$

Euclidean distance between tuples 1 and 3: $\text{sqrt}((0-17)^2 + (10-15)^2) = 17.720$

If the age and length of stay remain not normalized, the distance calculation will consider the distance between tuples 1 and 2 greater than the distance between tuples 1 and 3. Such calculation put an emphasis on the length of stay. Due to potentially much larger numeric data range of "length of stay", if left not normalized, its impact on distance calculation will be much greater than the age. Although tuples 1 and 3 have two extreme ages (0 vs. 17 years old) in pediatric population, its impact on distance calculation cannot give as much as slight variation in length of stay.

To address such unfairness, normalization plays an important role to scale numeric values within same data range. It is done by having each value divided by the difference between the highest value and the lowest value in the dataset

for that attribute. For age, the difference between the highest and lowest value is 17 - 0 = 17. For length of stay, it is 100 - 0 = 100.

With normalization

Tuple 1: Age = 0, Length of stay = 0.1

Tuple 2: Age = 0.0588, Length of stay = 0.3

Tuple 3: Age = 1, Length of stay = 0.15

Euclidean distance between tuples 1 and 2: $\text{sqrt}((0-0.0588)^2 + (0.1-0.3)^2) = 0.2085$

Euclidean distance between tuples 1 and 3: $\text{sqrt}((0-1)^2 + (0.1-0.15)^2) = 1.0012$

In fact, the distance between tuples 1 and 3 turns out to be much greater, since all data ranges have been normalized to have the same scale. There is no longer bias in distance calculation towards any specific attribute. Normalization makes attributes to have equal or more "controlled" influence on analysis results in clustering.

Data reduction tries to reduce the data volume by not having major impacts to the analytical results. In addition to reduce data representation using data aggregation and rollup, data reduction also includes removing some attributes that may not be relevant or have too little impact to the analysis.

For instance, in the original dataset, date of birth has been retrieved for each patient record. However, having many different dates to perform analysis not only slow down the data mining process, it also makes mined results to be very scattered, since the mining algorithms treat each date as different value. Without any data reduction, many different rules are being associated with specific dates, even some dates are just one day or two apart. This does not bring more significance to the mined results. The goal of data mining is to have a general picture of how data is related to each other. Similar values should be grouped together to simplify the process as well as the mined outputs. Using date of birth as the example, in data reduction, it is being turned into age. Then, in order to reduce number of different ages (i.e.: patients have varied ages between 0 and 100), age group of every 10 years is used. Instead of having potentially 101 different ages, only 10 age groups are being used. Data reduction can significantly enhance data mining performance, which makes data mining analysis more practical on huge datasets. Also, it makes cleaner and more understandable mined result output.

3.4 DATA CLEAN UP CHALLENGES

Although the hospital's admission, discharge and transfer system offers some useful information about patients, due to the fact that it is mostly based on user free text inputs, data gathered from that system is not standardized and seriously unorganized. For instance, there are 2574 distinct clinical diagnosis related to sitter cases. A lot of them contain spelling mistakes, unofficial user

defined terms and abbreviations, missing words/characters, improper mixed use of languages, unexpected spacing and punctuations between words. Actual distinct number of clinical diagnosis related to sitter cases is expected to be significantly less. Although some of the tuples can be "fixed" by using database queries and partial string matching, they cannot be cleaned without health practitioners' heavy involvement and further standardization in clinical terminologies, which will take considerably longer than the timeframe of this paper. Such issue is beyond the scope of this paper. Unfortunately, the column "diagnosis" needs to be dropped from the data analysis.

CHAPTER IV

REGULAR EXPRESSION BASED DATA MINING

4.1 PROBLEM ANALYSIS

Even with long enough periods of sitter case data collection, many investigations are still currently carried out using manual review of data to correlate attributes to each other to discover hidden patterns, which typically requires expensive labor of health practitioners over long periods of time. Although long histories of clinical data are present as references, one cannot easily tell what post sitter case information will likely happen, given pre sitter case information as inputs. The motivation of this research is to find a novel method of "predicting" post sitter case attribute value. We propose a system that can automatically classify an outcome of sitter case record with initial pre case user inputs, using the vector space model. Then, in later chapters, we will compare how well the system performs, comparing to existing data mining algorithms.

4.1.1 Sitter orders

The objective of our recommender system is to try to predict a post sitter case attribute value, based on pre case attribute values as inputs. Sitter cases are recorded by data and shift. No sitter orders can last for more than one entire shift, according to the hospital's guideline. For patients who need sitter supervision for more than one shift, additional orders must be placed. Every time

when a sitter is ordered, all the following pre case information about the case must be specified in the ordering system.

- Date of order
- Department's mission (aka sub-division)
- Order shift (Day, Evening, Night)
- Primary reason (Patient's problem)
- Units that placed the order
- Health practitioner who placed the order
- Health practitioner (supervisor level) who authorized the order
- Patient's medical record number
- Patient's family and first name
- Patient's gender
- Patient's primary spoken language

Some data is only for administrative purpose and is not useful in our recommender system design. The sitter ordering system in use only gathers information about the case to generate statistical reports from the entered data. It

does not tell the user what will likely happen, given those pre case information entered.

4.1.2 Problem formulation

Data mining can discover patterns and relationships. However, it does not tell the user the significance or relationship between data records [34]. Common ways of data mining evaluate records independently of each other. Although sequence mining methods can be used to take adjacent sequential data records into consideration, they are mostly based on support count to identify frequent sub sequences. Sub sequences with low support counts are often undetected and under looked by common data mining engines.

Our proposed system uses sequence similarity to predict most probable post case attribute values. It is developed to predict the value of an outcome (post case) of a related attribute, based on information before the sitter case happens (pre case). For example, before a sitter case is being created, health practitioners must know the following generic pre case information about the patient, such as

- Mission and hospital site where the patient is hospitalized
- Shift requiring the sitter service
- Gender
- Type of admission

- Marital status

And, the tool will try to predict any of the post case information such as

- Length of stay
- Discharge location

The discovery of information from sequential data consists of

1) data representation in sequential form
2) similarity measure between sequences

We need to first define how we can get the sequences from sitter orders. Due to the fact that all sitter orders contain date and shift stamp, they can be seen as sequential records. Because of the sequential nature of collected sitter cases, data operations to extract sequences from the dataset will be performed with the order of date and shift stamps preserved. Since all the orders are from the past, post case information (e.g.: length of stay and discharge location) is already known and can be related to the orders.

Table 1 - Sitter orders with post case information consist of multiple table columns

Mission	Site	Shift	Reason	Age group	Gender	Length of stay	Discharge location
Surgery	RVH	Night	Away without leave	70-79	M	20-29	Home

ER	RVH	Day	Disorientation	60-69	F	0-9	Hospital
ER	RVH	Evening	Agitation	70-79	F	0-9	Hospital
ER	RVH	Night	Disorientation	50-59	F	0-9	Hospital
Medicine	MGH	Night	Suicidal	80-89	M	0-9	Home

Filtering criteria using pre case attribute values is first applied to the dataset to get a smaller resulting dataset. Sequence can only be generated with one table column at a time. The chosen table column to generate sequences is considered as the “seed”. Users define a “seed” (e.g.: sitter reason), which will be used as the element in sequence generation. To facilitate the representation of the sequence element, instead of using the full table column value, a single alphabet index is being used to represent each attribute value.

Based on the above example data, the following sequence is produced.

AwayWithoutLeave Disorientation Agitation Disorientation Suicidal

(E) (J) (A) (J) (O)

EJAJO

From a filtered dataset by pre case attribute values, a sequence of the chosen table column is being generated. The generated sequence may resemble to some other sequences generated by different filtering values. The resemblance between two sequences may also indicate similar post case

attribute values. Our proposed recommender system makes use of the sequence similarity to discover relationship between pre case and post case attribute values. In other words, it is assumed that the symbolic sequence of an attribute may contain hints to reveal other attribute values. For example, a series of sitter reasons (pre case attribute) can be used as a predictor to predict length of stays (post case attribute).

The recommender system requires the following steps to predict post case attribute values about sitter cases.

1) Users let the system know which attribute to be used to generate sequences. This attribute will be used to predict other attribute values. For example, if users want to predict "length of stay" by using "sitter reason", "sitter reason" is chosen.

2) Users let the system know what to predict (e.g.: as in the above example, users want to predict "length of stay").

3) Users provide the system about the pre case information with multiple attribute value selectors.

4) The system will generate a sequence to represent the case, based on the selected attribute in 1) and criteria defined in 3). The generated sequence will be used a reference.

5) Sequences with different filtering values (i.e.: different lengths of stay) will be generated with the attribute selected in 2).

6) Words of different lengths will be found from all the generated sequences in 4) and 5).

7) Similarities between the reference sequence and sequences generated in 5) will be measured. The one with the highest similarity will be chosen and its filtering value will be considered as the predicted value.

With the above steps, we need to find out

1) How can we measure similarity between different sequences?

2) How can the system generate recommended result?

3) How much can we trust the prediction?

Since all the sitter cases were with date and shift, such date and shift stamps could be made use to perform further analysis. With proper filtering and sorting, sitter cases could be organized in ascending order by date and shift. Each case parameter could then be seen as a sequence.

In order to measure the correctness of the novel approach, the predicted value will be compared with actual historical data, which has known class label value. Prediction accuracy will be measured. The approach will be explained in more details in the next section.

4.2 CONCEPT

With data organized in sequences, data analysis in regular expression can be used to express patterns found. In a sequence, some cases always show up together in a certain way. For example, there may be a sequence of sitter cases with X number of agitated cases followed by Y number of suicidal cases. If pre case attribute values of a sitter case lead to similar sequence, such case may share similar post case attribute values as well. All the existing data mining methods used in later chapter were not able to spot out any similarities between sitter case sequences. Most of the existing algorithms evaluate database records one-by-one, without looking at a group of records together at the same time. Therefore, a lot of potentially similar sitter cases might have been missed.

The unknown attribute value prediction is being carried out by making use of regular expression like technique and sequential nature of other known attribute values. Regular expression technique has been used in information extraction both inside and outside of medicine, and could provide an alternative approach to more complex semantic parsers [41, 42, 43]. It offers the advantage to allow shorter and simpler representation of long sequences, which often contains repeated patterns. At the same time, regular expression syntax is mostly standard across all implementations and regular expressions developed for one application can usually be transferred to any of the others with minimal modification.

The proposed recommender system uses word matching technique to determine whether sequences are similar. A word is a series of items in a sequence that is repeated. A word finding engine with regular expression like approach has been developed to find out possible sub-strings of different lengths. Then, those sub-strings are being stored in a dictionary object with only distinct words. By calculating the term frequencies of each sequence and convert them into vector space, Cosine similarity can then be applied to measure the similarities between the reference sequence and the other generated sequences.

Cosine similarity is one of the most popular similarity measures in text mining, especially in information retrieval applications [35]. With normalized vectors of term frequencies, it can be applied to see how close each sequence is to the reference string. Cosine similarity measure has been widely used in clinical analysis to compare sequences generated by data collection tools with timestamps [36, 37, 38]. It has also been proven to be a robust metric for scoring the similarity between two strings, and it is increasingly being used in text mining related queries [39].

The objective of the novel prediction algorithm is to find out whether sequence similarity of some pre sitter case attributes can be used to predict post case attribute values. In a way, we would like to find out if the sequence of a specific attribute contains any hints to reveal any other attribute values. Since sequence strings are being dealt with, the developed system is designed to be

based on regular expressions to extract different words from sequences of a selected attribute. The information extracted by the system is being compared to the data stored in the database.

The idea of regular expression that we use is to identify repeated terms of any length within the training dataset. For instance, let us consider the following example generated sequence by the pre case attribute values. Each alphabet represents a sitter reason for one sitter case, in ascending order by date and shift.

Table 2 - Sequence generated using the "sitter reason" based on pre case attribute values

Pre case attribute values	Sequence generated using the "sitter reason" column value
Mission = Medicine Site = MGH Shift = Day Gender = Male Admission type = Clinic Marital status = Single	JJJJJJJAAAAJAAAAAALAAAAA

Table 3 - Sequence generated using the "sitter reason" based on different filtering values

Column to be predicted	Sequence label	Filtering Value	Sequence generated using "sitter reason" column value
Length of stay group	S1	A	JJJ
	S2	B	JJJJJJJJAAAAAALAAAAA
	S3	C	AAAA

Table 1 is a simple example showing pre case parameters specified by the user. Assuming the user selected "sitter reason" to predict the attribute value "Length of stay group", the system generates a sequence of sitter reasons based on the selected criteria. Such sequence is being considered as a reference sequence and is being compared with other generated sequences later.

Table 2 shows the generated sequences, based on all the possible values of "Length of stay group" as the filtering values. All the letters in the sequence are the column values of "sitter reason".

Having the sequences generated, the word finding engine will try to find all the possible words. A word is defined as a term with any length that can repeat itself within the sequence. For example, let us take the reference sequence and progressively represent it in regular expression form.

JJJJJJJAAAAJAAAAAALAAAAA

→ (JJJJJJ)(JAAAA)(JAAAA)(AA)(L)(AAAAA)

→ (J+)(JAAAA)+(A+)L(A+)

From the above example, 4 distinct words can be found.

Table 4 - Words found within the sequence

Word	Can mean
J+	J, JJ, JJJ, JJJJ, JJJJ, ...
(JAAAA)+	JAAAA, JAAAAJAAAA, ...
A+	A, AA, AAA, ...
L+	L, LL, LLL, ...

All of them are repeated at least once. Any repeating patterns are being spotted by the system, as words to be searched in other sequences.

4.3 FINDING WORD PATTERNS

Algorithm for finding repeated words within sequences

```
// lengthOfItemToSearch = character length to be used for word search
// curPos = cursor position
// wordList = global collection object to hold all the words found
Begin
For each sequence generated from the selected column value
    For lengthOfItemToSearch = Floor(sequence length / 2) to 1
        For each curPos in a sequence
            Let substring1 = subsequence with lengthOfItemToSearch characters,
from curPos
```

```
            Let substring2 = subsequence with lengthOfItemToSearch characters,
from curPos + sLengthNext
            If substring1 = substring2 Then
             If wordlist does not contain substring1 Then
                 Add substring1 to the wordList collection object
        Next
    Next
Next
End
```

In order to spot the words from sequences, an algorithm has been implemented in the recommender system. Taking the reference sequence as an example, the algorithm tries to find words with maximum length of the current sequence.

JJJJJJJAAAAJAAAAAALAAAAA

First, the algorithm tries to find the repeated words with maximum lengths. In our example, there are 24 characters in the sequence. The repeated word with longest length is

Floor(Length of the sequence / 2) = 24 / 2 = 12 characters

JJJJJJJAAAAJAAAAAALAAAAA

Table 5 - Word finding iterative process

Number of characters to look for	Cursor position	Word candidate	Sub string to be verified against	Result
12	1	JJJJJJJAAAAJ	AAAAAALAAAAA	No match
	2	JJJJJJAAAAJA	AAAAALAAAAA	Sub string does not have enough characters
11	1	JJJJJJAAAA	AAAAAALAAAA	No match
	2	JJJJJJAAAAJ	AAAAAALAAAA	No match
...				
5	1	JJJJJ	JJAAA	No match
	2	JJJJJ	JAAAA	
	3	JJJJJ	AAAAJ	
	...			
	7	JAAAA	JAAAA	Match!

Starting from the left most position of the sequence (cursor position = 1), the algorithm takes 12 characters (highlighted in yellow) then save the string into memory. Then, the cursor shifts 12 characters to the right to see if the next 12 characters (highlighted in green) after the string matches the previous one. Since there is no match, JJJJJJJAAAAJ is not being considered as a word.

Now, the cursor starts at position 2 and also looks for words with same number of characters (12 characters).

JJJJJJJAAAAJAAAAAALAAAAA

The algorithm takes 12 characters from position 2 instead. So, the partial string JJJJJJAAAAJA is being used to compare with the next 12 characters to see if there is any match. However, in this case, only 11 characters can be found after the string JJJJJJAAAAJA, which is AAAAALAAAAA. So, JJJJJJAAAAJA is not being considered as a word.

The algorithm cannot find any words with 12 characters with further starting cursor positions, due to not enough characters (i.e.: fewer than 12 characters) after the first 12-character string.

The algorithm basically works by first looking for repeated words of maximum characters (maximum possible length of repeated word = half of length of the sequence). The search is being done by looking for n-length sub strings at each cursor position, where n starts from the maximum length of potentially

repeated words down to 1-character sub string. In our example, the algorithm starts by looking for 12-character word at each cursor position, followed by 11-character word, then followed by 10-character word and so on, until 1-character word at each cursor position then stops.

Every time a "word" is discovered, it is being stored into a "dictionary". The dictionary holds all the distinct discovered words at the end of the search. In our example, the following words have been found.

Table 6 - Words identified by the word finding process

W0	JAAAA	W1	JJJ	W2	JJ
W3	J	W4	A	W5	L
W6	JJJJ	W7	AA		

Then, the algorithm counts the number of occurrence of each dictionary word in each sequence.

Table 7 - Identified word count of each sequence

Seq. #	Filtering value*	W0	W1	W2	W3	W4	W5	W6	W7
Ref. string	-	2	2	3	8	15	1	1	7
S1	A	0	0	0	1	0	0	0	0
S2	B	1	2	4	8	11	1	2	5
S3	C	0	0	0	0	4	0	0	2

The next section will talk about how the term frequencies are being represented into vector space to do the analysis.

4.4 TURNING TERM FREQUENCIES INTO VECTOR SPACE

Algorithm for populating vector space with term frequencies

```
// To be able to do comparisons between vectors, all vectors must have equal length, which is
// number of words in wordList collection object
Begin
Let totalNumOfWords = number of words stored in the wordList collection object
For each sequence Si
    Prepare an empty vector Vi with number of elements = totalNumOfWords
    For each word in wordList collection object
        Let n = number of selected word occurrence in the sequence (term frequency of that word)
        Store n as an element into the vector
    Next
Next

// This part is to normalize the term frequencies within the vector spaces
For each element in vector with term frequencies
    Divide the element by total number of words found for the sequence (combined count of all
    words within the sequence)
Next
End
```

With all the word counts summarized in the above table, vector space model can be used to treat sequences as a vector of keywords. There are 8 words found by the algorithm, across all the sequences. A vector of 8 elements is being created to represent each sequence, where each element represents number of dictionary word occurrences, as known as term frequency (tf).

Reference string = [2 2 3 8 15 1 1 7]

S1 = [0 0 0 1 0 0 0 0]

S2 = [1 2 4 8 11 1 2 5]

S3 = [0 0 0 0 4 0 0 2]

Since each vector always contains exactly the same number of elements, comparisons can be made between them to find out which sequence resembles the most to the reference string. However, term frequencies of dictionary words can have bias towards longer sequences. With a very long sequence, it is more likely to have higher term frequencies of dictionary words, regardless the significance of the term. A term is more significant if it has a larger proportion of occurrences, with respect to total number of words within the sequence. For example, let us consider the following sequences.

S1 = AAB

S2 = AAAAABBCCCCCCCCCC

The sequence S1 has 3 items, with two distinct words called A and B. In this case, A has a term frequency of 2 and B has 1. Out of 3 items, A has 2/3 = 0.66667 = 66.667% of occurrence and B has 1/3 = 0.33333 = 33.333%. In the case of S2, we have three distinct words, A, B and C. A has a term frequency of 5, B has 2 and C has 10. If we only compare term frequencies of words between the two sequences, S2 has more A's and B's. However, A only occupies 5/17 = 29.4118% and B only occupies 2/17 = 11.7647%, with respect to the total

number of words within the sequence. The 2 A's and 1 B in S1 have more significance (more important) than S2, since within the sequence, they have more occupancies.

So, in order to reduce the impact of potential term frequency bias towards longer sequences, the term frequency is being normalized by dividing the term frequency by total number of words found within the sequence. In our example, the vector elements are being normalized with the total number of words within the sequence.

Table 8 - Normalized vector spaces representing word count for each sequence

Sequence name	Vector of term frequencies	Number of words found within the sequence	Vector of normalized term frequencies
Reference string	[2 2 3 8 15 1 1 7]	39	[0.05128 0.05128 0.07692 0.20513 0.38462 0.02564 0.02564 0.17949]
S1	[0 0 0 1 0 0 0 0]	1	[0 0 0 1 0 0 0 0]
S2	[1 2 4 8 11 1 2 5]	34	[0.02941 0.05882 0.11765 0.23529

			0.32353 0.02941 0.05882 0.14706]
S3	[0 0 0 0 4 0 0 2]	6	[0 0 0 0 0.66667 0 0 0.33333]

According to our algorithm, whichever the sequence (S1, S2 or S3) has the highest similarity with the reference sequence, its post case attribute value is considered to be the prediction result.

4.5 IDENTIFYING THE VECTOR WITH THE HIGHEST SIMILARITY

Algorithm for identifying the sequence with the highest Cosine similarity

```
// This part is to compare the reference sequence S0 with all the other sequences, in order to determine which sequence most resembles S0
Begin
For each vector Vi other than V0
    Calculate the Cosine similarity between V0 and Vi
    If CosineSimilarity(Vi) > CosineSimilarity(previous Vi) Then
        Record it as the output
Next

Output the Vi that has the highest Cosine similarity with V0.  Since Vi comes from Si , filtering value of Si becomes the predicted result.
End
```

Cosine similarity is a vector-based measure of the similarity between two strings. To do so, strings must be represented in vectors with same dimension. In information retrieval, vectors are usually composed of term frequencies of discovered words, as in our example. To calculate how close two strings are, two

vectors can be compared by dividing the dot product of the two vectors by the product of each vector's magnitude.

$$\text{similarity} - \cos(\theta) - \frac{A \cdot B}{\|A\|\|B\|} - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

The cosine of the angle between two vectors is a measure of how similar the two strings are. Cosine of an angle can range from 0 to 1, where 0 means no similarity and 1 means total similarity between the two strings. The algorithm identifies the sequence with the highest cosine similarity with respect to the reference string. The filtering value of the selected sequence is the predicted class label, since each sequence (S1, S2, S3) is generated based on a filtering value. In our example, we get the following cosine similarity values

Table 9 - Cosine similarity of each sequence against the reference sequence

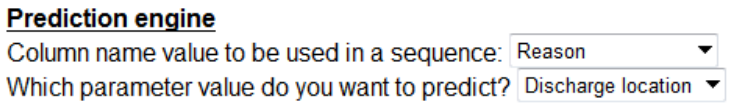
Seq. #	Filtering value*	Cosine similarity
S1	A	0.4234049
S2	B	0.9818717
S3	C	0.875755

The sequence S2 has the highest cosine similarity value. Its filtering value "B" is the predicted "Length of stay group", which translates into "10 to 19" days.

In order to demonstrate the effectiveness of this novel approach of classification, it is necessary to show that such technique employed has relatively better precision than any manual guesswork. We therefore evaluated the precision by comparing the predicted result against records with existing class label values.

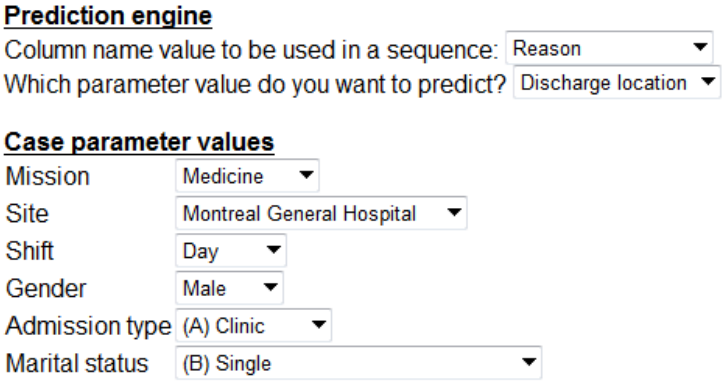
4.6 PROTOTYPE SYSTEM AND ITS USAGE

The recommender system has two major parts, which require user inputs. All those required inputs can be supplied before the sitter case actually happens.



In the prediction engine section, since the prediction engine is based on a single parameter to generate sequences, user must tell the system which parameter is to be used. The system also requires the user to specify the outcome parameter to be predicted (i.e.: post case parameter prediction). The recommender system uses the above inputs to generate sequences and perform similarity calculation accordingly, in order to find out the most probable outcome parameter value.

In the "Case parameter values" section, there are multiple drop down lists that allow users to specify the pre case attribute values. They will be used as filtering criteria to first narrow down the dataset and then generate sequences. The sequence generated with the pre case attribute values is considered as the reference sequence. It will be compared with sequences that are generated with different post case attribute values. The similarity between the reference and each generated sequence will be calculated. The sequence with the highest similarity will be considered as the most probable outcome value. To put the system in practice, let us run a real example with the following input parameter values.



The prediction result comes up as follows.

Prediction engine

Column name value to be used in a sequence: Reason

Which parameter value do you want to predict? Discharge location

Case parameter values

Mission: Medicine

Site: Montreal General Hospital

Shift: Day

Gender: Male

Admission type: (A) Clinic

Marital status: (B) Single

Do it for all the case parameter values and display me the precision in a summary table

Go

Words found from records with different values of Discharge location

W0	JAAAA	W1	JJJ	W2	JJ
W3	J	W4	A	W5	L
W6	AAA	W7	AA	W8	JJJJ

Number of words found in each sequence string (S)

Seq. #	Filtering value*	W0	W1	W2	W3	W4	W5	W6	W7	W8
Ref. string	-	2	2	3	8	15	1	4	7	1
S1	L	1	0	0	1	15	1	4	7	0
S2	M	0	2	4	8	0	0	0	0	2

Cosine similarities between sequence strings (Ref. string vs. different S)

Seq. #	Filtering value*	Cosine similarity
S1	L	0.9104965
S2	M	0.4526037

Discharge location is predicted to be L

** Filtering value is the column filtering value used to obtain the sequence string. For example, if user wants to predict the column 'Shift', all possible values of 'Shift' will be used as the filtering value (i.e.: Day, Evening, Night).*

Seq. to be compared (Ref. string):	JJJJJJJAAAAAJAAAAAALAAAAA
Seq. with highest Cosine similarity:	AAAAJAAAAAALAAAAA
Predicted result:	Discharge location = L
Seq. to be validated* :	MMMMMMMLLLLLLLLLLLLLLL
Precision:	17 / 24 = 0.7083333

** This sequence contains values of the predicted column. It will be used as a reference to evaluate the predicted result accuracy.*

The system identifies the number of words (as known as term frequencies) found from all the sequences. Those term frequencies are being used to construct vectors, which will later be normalized. Cosine similarity measure is being applied to the normalized vectors, in order to calculate the similarity quantitatively. The highest similarity between the reference sequence and generated sequence with post case parameter value is identified as the prediction result (i.e.: Discharge location = L).

To evaluate how accurate the predicted result is, it is being compared against existing class label values in the dataset. In our example, the column "Discharge location" was chosen to be predicted. Based on the input parameter values as filtering criteria, "discharge location" values could be retrieved to produce a sequence. Values in such sequence are being compared with the

predicted result (i.e.: Discharge location = L). Precision is calculated by dividing the number of correctly matched value (i.e.: L in the sequence) by the total number of items in the sequence. This outputs the hit rate of the prediction. It gives user an idea how reliable is the prediction, comparing to historical records.

CHAPTER V

EXPERIMENT AND RESULTS

5.1 EXPERIMENT

To put the algorithm in practice, the system has a feature to perform the class label prediction with combinations of attribute values. A summary of results is shown below with the precision.

Table 10 - Classification result precision

Attribute chosen to be used in sequences	Post case attribute value to be predicted	Average precision	Median precision	# predictions over 70% precision
Reason	Discharge location	0.666874	0.666667	220
Reason	Length of stay	0.567721	0.537037	186
Admission type	Discharge location	0.687383	0.664286	217
Admission type	Length of stay	0.602015	0.529412	181

Marital status	Discharge location	0.688287	0.664286	217
Marital status	Length of stay	0.602015	0.529412	181
Age group	Discharge location	0.679217	0.666667	221
Age group	Length of stay	0.577544	0.537037	186
Discharge location	Length of stay	0.577544	0.537037	186
Length of stay	Discharge location	0.682886	0.666667	221

There are 471 combinations of pre case parameter values that have returned results. Although the sequence pattern of only one attribute is used to predict the post case attribute values, in general, results are quite promising with fair accuracies. It is interesting to notice that the algorithm can retrieve higher average precision with "Discharge location" as the attribute value to-be-predicted. The prediction precision seems to be driven by the selected attribute value to be predicted.

By examining the data, “Length of stay” and “Discharge location” have 11 and 22 distinct values respectively. With such number of distinct values in those two columns, it is supposed to provide enough variations to values to generate sequences. However, it is not quite the case. By looking at the data distribution of each of those columns, here is the graphical observation.

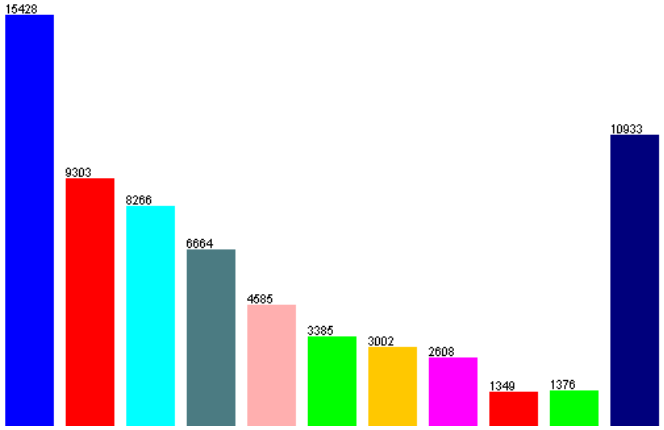


Figure 1 - Distribution of Length of stay

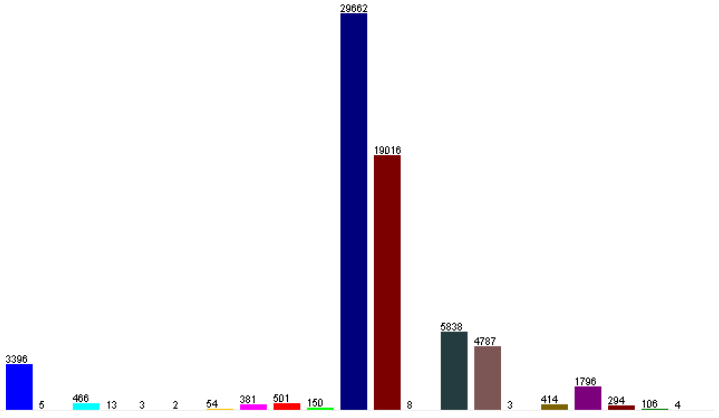


Figure 2 - Distribution of Discharge location

The data distribution of “Discharge location” had strong bias towards one specific value - Home, as shown in the highest bar. The variation of data frequencies between different column values is very extreme – from 2 to 29662. Most patients were discharged “Home”, after their sitter episodes, which occupied 44% of sitter cases. Although "Length of stay" also had strong bias towards the left most bar - "0-9 days", its distribution was more even than "Discharge location". Each value of "Length of stay" still had significant amount of sitter cases. The most popular "Length of stay" occupied 23% of sitter cases, which was just around half of the most popular value in "Discharge location = Home". Other than the two middle highest bars for the "Discharge location" column, other values did not seem to occupy many cases. Since the proposed prediction engine generates sequences based on different filtering values of the column-to-be-predicted, filtering value with "Discharge location = Home" is more likely to result in closer match with the reference sequence. The reference sequence likely contains a lot of sitter cases with "Discharge location = Home", due to very high frequency of such value. Even one makes a wild guess on "Discharge location" = Home (29662 cases, 44%), a correct prediction can still be achieved.

5.2 MEASURING PRECISION

Larger sample sizes generally lead to increased precision when estimating unknown parameters. In order to make use of enough historical records (experience) to estimate to-be-predicted attribute values, sitter records of three

entire years are being used. To evaluate the proposed approach, we use different combinations of sitter pre case parameters to predict different post case results. The accuracy is being measured by comparing the predicted results with the known class label values in the dataset. A commonly used measure for evaluating the performance of text mining approach is precision.

precision = (|known class labels| ∩ |approach class label|) / |known class labels|

The above gives how well the proposed approach (hit rate) does to predict the class label value using sequence matching algorithm. Based on user's pre sitter case inputs, many returned results from historical records with known class label values can be retrieved. However, the algorithm only outputs one predicted class label value. As a result, this only value is being compared with the known class label values to see how many of them are matched, which means, successful predictions. For example, we have a sequence of known class labels as follows.

AABBBBBCCBBD

If the predicted class label value turns out to be "B", 7 of 12 items in the above sequence match the predicted result "B". So, the precision is 7/12 = 0.58333 = 58.333%. Such precision measure is only based on historical sitter cases of three entire years. It is very possible that the precision can change significantly in the future years, due to various clinical and administrative factors

in hospitals. But, as a minimum, assuming future sitter cases have similar trends, the precision measure can indicate how well the algorithm can perform.

Attributes with more distinct values can produce more complex sequences with potentially more patterns found. It is been noted that attribute with very few distinct values can only produce mostly sequences with one value, which do not contain any interesting patterns. Thus, such sequences with monotonic value are not being used in our experiment to do the class label value prediction. Only attributes with more than 5 distinct values are being used to generate sequences. For example, let us consider the attributes "reason" vs. "gender". The attribute "reason" has 18 distinct values, whereas "gender" only has 2 distinct values. Sequences represented with the attribute "reason" selected (18 distinct values) can potentially become more complex with many more patterns found. Sequences represented with the attribute value "gender" (2 distinct values) only contain a lot of repeating "Male" or "Female". In many cases, sequences only contain either "Male" or "Female".

5.3 OBSERVATION

The proposed prediction approach gives users a convenient way to predict a post case class label value, with only generic pre case inputs. The algorithm is based on matching the sequence arrangements to "guess" the probable class label value. Such trend matching totally relies on data of the past. The question is, will the future data have more or less the same trends? Will there be significant

changes to any clinical guidelines or/and administrative procedures to order sitters? Any changes to ways of doing things will break the approach. However, this is also true to all other data mining approaches, since they work with current and historical data.

Our approach assumes that pre case attribute values are somehow related to post case attribute ones. Although it can bring in some promising results, it still needs to be validated and subject to major changes. The way data gets collected in the application can cause major impact to the sequence generation. For example, if a clerk orders multiple sitter cases for exactly the same date and time, the order of case entries potentially causes very different item ordering in sequences.

Prediction precisions of different combinations of pre sitter case inputs result an almost perfect normal distribution, with precision averages almost equal to medians. This can be explained by the central limit theorem. The theorem states that the mean of a sufficiently large number of independent cases, each with finite mean and variance, will be approximately normally distributed [44]. Due to sufficiently large number of sitter cases in the dataset, it is expected to have variability of post case attribute values, even with same pre case attribute values.

There are also clinical factors that can cause impacts to the precision of our approach. In the meantime, 18 sitter reasons are defined. The reason why a

sitter is being hired relies on the judgment of health practitioners. There is no clear line drawn between some of the sitter reasons. Standardization and clear definition of sitter reasons are required to ensure that everyone understands the same thing, in order to avoid any bias towards certain sitter reasons over the others.

With more attributes captured in the sitter ordering system and integration with future clinical information systems in the hospital, our recommender system can be potentially more powerful to predict more post case class label values, with more pre case attribute input values. Also, with more data collected from more years, our recommender system can have more experience data to base on, in order to determine the class label value.

While the system brings in some good results, it can only be served as a reference to health practitioners. In some cases, the recommender system predicts a class label value which does not match any of the known class label values. Professional judgment and clinical experience of health practitioners remain the key determination factor to evaluate the outcome of any sitter cases.

In order to validate the effectiveness of the proposed approach, the system needs to be compared with well-known data mining approaches. Different data mining algorithms will be used to perform prediction of sitter post case parameter value. Their accuracies will be compared with the ones from our proposed system. Although some data mining algorithms are not specifically designed to

perform class label prediction, their results will be interpreted as if prediction takes place. This will be discussed in the next chapter.

CHAPTER VI

EVALUATIONS AND DISCUSSIONS

Since some data is considered to be confidential information, only a subset of data has been chosen in this data mining activity. To bring the number of tuples to a manageable range, only the most recent completed years, 2008, 2009 and 2010 have been selected.

Data has been consolidated from sitter and hospital's admission, discharge and transfer (ADT) systems by data transformation and massive cleanup has been performed to correct mistaken entries. Some data attributes such as "Patient's age", "Length of stay" and "Discharge location" have been "rolled up" and discretized for easier and more readable classification.

Since there are 4 adult hospitals and only 1 pediatric hospital, it is expected to have more sitter orders from the adult hospitals. If association rule finding is being run on the combined dataset, a lot of interesting association rules from the pediatric hospital may be filtered out. Due to much larger adult population than pediatric population, insufficient support in pediatric population (i.e.: less than minimum support threshold) can occur, even there are frequent itemsets with high count. For example, in pediatric population, 20% of sitter orders have primary reason = agitation, gender = male, shift = night. However, when we look

at the data with all the adult hospitals together, only 5% of orders have the above parameters.

In order to maximize the meaningfulness of association rule findings, the association rule mining will be run separately, focusing on dataset with

- Pediatric population (Age < 18 years old)
- Adult population (Age >= 18 years old)

According to the statistics, each of the following fields only has one value.

- Mission

  The entire pediatric population belongs to the pediatrics mission.

- Site

  Only the Montreal Children's hospital (MCH) handles pediatric patients.

- Marital status

  The entire pediatric population was considered as "Single Adult", according what has been entered in the system.

- Discharge location

  There is no discharge location indicated for pediatric patients.

In that case, there will be no significance to use the above data for pediatric population. They will be ignored just for this population.

6.1 COMPARE WITH ASSOCIATION RULE EXTRACTION

To be able to compare the effectiveness between our system and proven association rule algorithms, results from our system must be represented as an implication of the form X ==> Y, where X represents the given pre case attribute values and Y represents post case attribute value. From the previous section, a list of pre case attributes and post case attributes has been identified. Since our system is currently designed to have fixed number of pre case attribute values as inputs and one post case attribute value as the output each time, the following association rule can be used to represent the prediction.

Mission Λ Site Λ Shift Λ Gender Λ Type of admission Λ Marital status ==> Length of stay

or

Mission Λ Site Λ Shift Λ Gender Λ Type of admission Λ Marital status ==> Discharge location

Not every association rules found may contain all the attributes like the above rules. Therefore, we will also consider rules with fewer pre case attributes. In addition to observe any hidden knowledge found by the rule engines, we will also pay attention to discovered association rules that match or at least mostly match the above two rules. Comparisons will be made between results from our system versus proven algorithms.

In association rules, two frequent terms must be used.

- Support

  Percentage of the population, which satisfies the discovered rule.

- Confidence

  If the above support percentage is satisfied, confidence is the percentage of how likely the consequence is also satisfied.

These two metrics are being used to judge if discovered rules by proven algorithms are trustable.

Only child population (2008 = 1207 instances, 2009 = 1143 instances, 2010 = 988 instances)

Since the analysis is being done on fiscal years (April 1 of the selected year to March 31 of the year after), it is expected to have more sitter cases in the selected year as it contains 9 months, from April to December. The year after only contains 3 months, from January to March. As a result, it is biased to perform data analysis using the data field "year". The field "year" should not be taken in consideration during data mining.

All patients in the pediatric population belong to the "pediatric" mission and "MCH" hospital site. These fields have no use to data mining and should be taken away. All child patients are "single", as indicated in the marital status. This field should be taken away.

Discharge location does not apply to pediatric patients since all pediatric patient records show no values in that field. It should be taken away from the mining activity too.

Before starting the rule mining, we need to know what we actually look for. A few things can be interesting.

- Do any combinations of attributes have something to do with month(s) and date(s)?
- Are there are any relationships between the primary reason (problem) and other attributes? Can some combinations of attributes more likely to cause patients to have specific problems?
- Are there are any relationships between the patient's gender and other attributes? Can some problems be more gender-specific? Can a specific gender more likely to be affected by other attributes?
- Are there are any relationships between the order shift and other attributes?
- Do certain age groups have more specific reasons that require sitter supervision?

With these questions in mind, the association rule mining can be targeted to specific classes. Class attributes are set in our experiments to guide the system

to find out rules we are looking for. Class association rule mining is being used to help focus our thought. It is used as a guide to mine the association rules.

We may first identify attributes with highest frequencies. This gives us an idea how frequent those attribute appear in discovered rules. Supposedly, attributes with higher frequencies are more likely to be picked by the association rule engine.

Table 11 - Number of distinct values and frequencies of each attribute

Attribute	Number of distinct values in the grouping	Value with highest frequency (number of occurrence and corresponding %)
Month	12	9 (229 → 18.97%)
Day	31	29 (63 → 5.22%)
Cost center	6 (only pediatric cost centers)	32806 (609 → 50.46%)
Shift	3	Day (419 → 34.71%)
Reason	17	Other (507 → 42.00%)
Age group	2 (only pediatric population)	10-19 (886 → 73.41%)
Gender	2	Female (716 → 59.32%)

Language	3	English (685 → 56.75%)
Municipality	278	Montreal (432 → 35.79%)
Admission type	9	Clinic (1194 → 98.92%)
Length of stay group	11	0-9 (493 → 40.85%)

With class attribute = “Month” (using Apriori algorithm)

minimum support threshold = 5% and minimum confidence threshold = 80%

Year 2008

The association rule engine found many multidimensional association rules. Interestingly, the generated association rules showed the following results

1) None of the rules found had attributes "Day" or "Shift". By looking at those attributes, they both had pretty evenly distributed values within their groups in the training dataset.

2) Attribute "Month" always indicated the month of September (Month = 9) as the resulting attribute.

   According to the statistics of the training dataset, September had the most cases throughout the entire fiscal year. It had 229 sitter cases (43% more

cases than the month of December, second highest number of sitter case month, 160 cases).

3) Attribute "Cost center" existed in all the rules and is always 32801, which means, Pediatric Surgery unit. However, Pediatric Surgery unit did not occupy most of the sitter cases in this training dataset.

4) Although the age group "0-9" did not represent most of the pediatric population in this training dataset (321 cases for age group "0-9" vs. 886 cases for age group "10-19"), it was the only age group that showed in the discovered rules.

5) Attribute "Language" existed in most of the rules and was always English. This is understandable as Montreal Children's hospital is an English hospital.

6) Attribute "Gender" existed in most of the rules and was always female. Statistics showed that number of female sitter cases (716 cases) was significantly more than male ones (491 cases).

7) Attribute "Municipality" existed in most of the rules and had always a value of "Montreal". Montreal was the city where most patients resided.

8) Attribute "Length of stay" existed in many rules with value of "80-89". However, this value only appeared in ~10% (124 out of 1207) of sitter cases in this training dataset.

Out of the rules found, a lot of them did not seem to be meaningful. Many of them were often spurious and obvious information. Insignificant rules were not useful and could be eliminated without loss of generality.

With the following rules with top confidence and support,

1. Cost center=32801 Gender=F Municipality=MONTREAL 175 ==> Month=9 151 conf:(0.86)

2. Cost center=32801 Reason=Other Gender=F Municipality=MONTREAL 175 ==> Month=9 151 conf:(0.86)

3. Cost center=32801 Gender=F Language=English Municipality=MONTREAL 175 ==> Month=9 151 conf:(0.86)

4. Cost center=32801 Gender=F Municipality=MONTREAL Admission type=Clinic 175 ==> Month=9 151 conf:(0.86)

5. Cost center=32801 Reason=Other Gender=F Language=English Municipality=MONTREAL 175 ==> Month=9 151 conf:(0.86)

6. Cost center=32801 Reason=Other Gender=F Municipality=MONTREAL Admission type=Clinic 175 ==> Month=9 151 conf:(0.86)

7. Cost center=32801 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 175 ==> Month=9 151 conf:(0.86)

8. Cost center=32801 Reason=Other Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 175 ==> Month=9 151 conf:(0.86)

Among the above rules, none of them had attributes "Day", "Shift", "Age group" or "Length of stay group". The attribute "Length of stay group" and "Age group" only started to exist in rule 11 and 13 respectively, with lower confidence and support than the above.

11. Cost center=32801 Length of stay group=80-89 107 ==> Month=9 88 conf:(0.82)

13. Cost center=32801 Age Group=0-9 Language=English 107 ==> Month=9 88 conf:(0.82)

All the attributes except "Day" and "Shift" showed up in the following rule. It had relatively high confidence and support values.

98. Cost center=32801 Reason=Other Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=80-89 107 ==> Month=9 88 conf:(0.82)

This rule contained exactly the attribute values that appeared in all the previous rules. Even with all those attribute values together, the confidence and support values were not much lower than the previous rules with fewer attributes.

Year 2009

96. Reason=Other Age Group=10-19 Gender=M Language=English Municipality=ROSEMERE Admission type=Clinic Length of stay group=60-69 93 ==> Month=3 84 conf:(0.9)

All the rules were different combination of the attribute values in the above rule 96. The rule gave very different result than the one in 2008.

Year 2010

16. Cost center=32806 Reason=Other Age Group=10-19 Gender=M Municipality=LASALLE Admission type=Clinic 51 ==> Month=3 51 conf:(1)

97. Cost center=32801 Reason=Agitation Language=French Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY Admission type=Elective Length of stay group=20-29 54 ==> Month=11 50 conf:(0.93)

All the generated rules only contained two resulting attributes values, March (Month=3) and November (Month=11). The remaining rules were just combination of attributes of the above rules 16 and 97.

Using Predictive Apriori algorithm

Since there were many rules generated, only the meaningful ones with not too low support count were selected, as follows.

Year 2008

1. Reason=Suicidal Municipality=SAINT-LAURENT 54 ==> Month=12 54 acc:(0.98891)

2. Municipality=SAINT-LAURENT Length of stay group=20-29 53 ==> Month=12 53 acc:(0.98871)

3. Gender=F Language=French Length of stay group=20-29 53 ==> Month=12 53 acc:(0.98871)

7. Cost center=32801 Municipality=MONTREAL Length of stay group=20-29 39 ==> Month=9 39 acc:(0.98475)

26. Cost center=32806 Shift=Day Reason=Suicidal Age Group=10-19 Gender=F Admission type=Clinic Length of stay group=20-29 15 ==> Month=12 15 acc:(0.95639)

46. Reason=Suicidal Gender=F Length of stay group=20-29 46 ==> Month=9 44 acc:(0.93952)

67. Cost center=32806 Gender=F Municipality=SAINT-LAURENT 56 ==> Month=12 53 acc:(0.93196)

69. Cost center=32806 Reason=Suicidal Gender=F Length of stay group=20-29 56 ==> Month=12 53 acc:(0.93196)

Rules discovered by Predictive Apriori algorithm were very different than the ones found by Apriori. The Predictive Apriori algorithm showed a very important observation that had not been caught by the Apriori. Most of the rules discovered with top accuracies resulted in September and December. A number of rules showed that girls from St-Laurent city between the ages of 10 and 19 seemed to be more likely to commit suicide. September and December seemed to be the trouble months for them.

Year 2009

1. Cost center=32207 Length of stay group=70-79 39 ==> Month=8 39 acc:(0.99112)

2. Cost center=32207 Reason=Agitation Age Group=0-9 39 ==> Month=8 39 acc:(0.99112)

3. Cost center=32207 Reason=Agitation Gender=F 39 ==> Month=8 39 acc:(0.99112)

4. Cost center=32207 Reason=Agitation Municipality=MONTREAL 39 ==> Month=8 39 acc:(0.99112)

5. Gender=M Municipality=LONGUEUIL 33 ==> Month=5 33 acc:(0.98978)

6. Reason=Away without leave Length of stay group=20-29 32 ==> Month=8 32 acc:(0.98949)

7. Municipality=PIERREFONDS Length of stay group=20-29 32 ==> Month=8 32 acc:(0.98949)

8. Cost center=32806 Language=English Length of stay group=20-29 32 ==> Month=8 32 acc:(0.98949)

9. Reason=Agitation Municipality=LONGUEUIL 27 ==> Month=5 27 acc:(0.98761)

10. Reason=Agitation Length of stay group=10-19 27 ==> Month=5 27 acc:(0.98761)

52. Cost center=32806 Reason=Away without leave Age Group=10-19 Gender=F Admission type=Clinic Length of stay group=0-9 11 ==> Month=10 11 acc:(0.96487)

65. Cost center=32207 Shift=Evening Reason=Other Gender=M Language=English Admission type=Clinic Length of stay group=60-69 10 ==> Month=3 10 acc:(0.96056)

Rules discovered had relatively low support in general. Most rules had very high accuracies (i.e.: > 0.9).

Year 2010

3. Age Group=0-9 Length of stay group=20-29 44 ==> Month=8 44 acc:(0.98844)

5. Age Group=0-9 Gender=F Language=French 44 ==> Month=8 44 acc:(0.98844)

8. Cost center=32806 Municipality=ROSEMERE 43 ==> Month=4 43 acc:(0.98819)

11. Reason=Suicidal Municipality=NOTRE-DAME-DE-L ILE-PERROT 39 ==> Month=12 39 acc:(0.98705)

Very different results were produced by the predictive Apriori algorithm. Those rules could not be found by the original Apriori, due to low support values.

Only adult population (2008 = 21765 instances, 2009 = 22229 instances, 2010 = 19567 instances)

Similar rule mining is being applied, as in the above pediatric population. But this time, “Mission”, “Age group”, "Marital status" and "Discharge location" are included since these attribute values can be very different between tuples, unlike those in the pediatric population. The attribute "Cost center" is removed from the analysis, since it is a related attribute to "Mission". Every "Cost center" must belong to a specific mission and such relationship never changes. If both attributes are being kept, many obvious rules will be generated thus reducing the meaningfulness of the analysis.

With class attribute = “Reason”

Year 2008

1. Mission=Surgery Site=MGH Gender=F Marital status=SINGLE_ADULT Admission type=ER 1271 ==> Reason=Suicidal 1139 conf:(0.9)

2. Mission=Surgery Site=MGH Marital status=SINGLE_ADULT Admission type=ER Length of stay group=>=100 1252 ==> Reason=Suicidal 1114 conf:(0.89)

3. Mission=Surgery Marital status=SINGLE_ADULT Admission type=ER Length of stay group=>=100 1254 ==> Reason=Suicidal 1114 conf:(0.89)

4. Mission=Surgery Site=MGH Marital status=SINGLE_ADULT Length of stay group=>=100 1258 ==> Reason=Suicidal 1114 conf:(0.89)

5. Mission=Surgery Marital status=SINGLE_ADULT Length of stay group=>=100 1267 ==> Reason=Suicidal 1114 conf:(0.88)

6. Mission=Surgery Gender=F Marital status=SINGLE_ADULT Admission type=ER 1355 ==> Reason=Suicidal 1177 conf:(0.87)

7. Mission=Surgery Site=MGH Admission type=ER Length of stay group=>=100 1348 ==> Reason=Suicidal 1114 conf:(0.83)

8. Mission=Surgery Site=MGH Gender=F Marital status=SINGLE_ADULT 1411 ==> Reason=Suicidal 1147 conf:(0.81)

The Apriori algorithm could only find 8 rules. All resulting attributes were "Suicidal", as the reason for hiring sitters. Many attribute values were the same across rules, such as mission = "Surgery", Marital status = "Single Adult", Gender = "F", Site = "MGH" and Length of stay >= 100.

Year 2009

1. Marital status=MARRIED_ADULT Language=English Length of stay group=>=100 Discharge location=Home 1210 ==> Reason=Agitation 1209 conf:(1)

16. Age Group=80-89 Gender=M Marital status=MARRIED_ADULT Language=English Length of stay group=>=100 Discharge location=Home 1131 ==> Reason=Agitation 1130 conf:(1)

35. Site=RVH Gender=M Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Admission type=Elective Length of stay group=>=100 1113 ==> Reason=Agitation 1112 conf:(1)

All the rules found had the same resulting attribute, that was, "Agitation". This value was the most frequently appeared value among reasons (9132/22229 = 41.08%). Even the attribute value "Disorientation" was the second most frequent in the dataset (8579/22229 = 38.59%), it did not show up in any rules. Most rules were very similar to each other, with same attribute values but with different combination of attributes.

Year 2010

1. Site=RVH Age Group=70-79 Language=Other 1051 ==> Reason=Disorientation 1011 conf:(0.96)

2. Site=RVH Language=Other Municipality=MONTREAL 1143 ==> Reason=Disorientation 1046 conf:(0.92)

3. Site=RVH Gender=M Language=Other 1282 ==> Reason=Disorientation 1150 conf:(0.9)

4. Site=RVH Gender=M Marital status=MARRIED_ADULT Language=Other 1159 ==> Reason=Disorientation 1036 conf:(0.89)

5. Site=RVH Language=Other 1505 ==> Reason=Disorientation 1313 conf:(0.87)

6. Site=RVH Marital status=MARRIED_ADULT Language=Other 1228 ==> Reason=Disorientation 1071 conf:(0.87)

7. Site=RVH Age Group=70-79 Gender=M Municipality=MONTREAL 1338 ==> Reason=Disorientation 1121 conf:(0.84)

8. Site=RVH Age Group=70-79 Municipality=MONTREAL 1583 ==> Reason=Disorientation 1314 conf:(0.83)

9. Age Group=70-79 Gender=M Length of stay group=>=100 1260 ==> Reason=Disorientation 1044 conf:(0.83)

10. Age Group=70-79 Language=Other 1284 ==> Reason=Disorientation 1060 conf:(0.83)

11. Site=RVH Age Group=70-79 Gender=M Discharge location=Home 1238 ==> Reason=Disorientation 992 conf:(0.8)

All the rules found had the same resulting attribute value - "Disorientation". All attribute values were the same with only different combination of attributes across rules. They probably pointed to specific cases of same patient.

Using Predictive Apriori algorithm

Year 2008

1. Age Group=20-29 Gender=M Length of stay group=>=100 Discharge location=Home 298 ==> Reason=Suicidal 298 acc:(0.99485)

15. Age Group=70-79 Discharge location=Long term care 288 ==> Reason=Disorientation 288 acc:(0.99483)

21. Age Group=40-49 Gender=F Language=English Length of stay group=>=100 282 ==> Reason=Suicidal 282 acc:(0.99482)

22. Age Group=40-49 Gender=F Admission type=ER Length of stay group=>=100 282 ==> Reason=Suicidal 282 acc:(0.99482)

28. Mission=Surgery Site=MGH Age Group=40-49 Gender=F Admission type=ER Discharge location=Home 282 ==> Reason=Suicidal 282 acc:(0.99482)

29. Gender=M Marital status=MARRIED_ADULT Length of stay group=60-69 Discharge location=Hospital 278 ==> Reason=Disorientation 278 acc:(0.99481)

30. Age Group=70-79 Language=English Length of stay group=>=100 Discharge location=Long term care 273 ==> Reason=Disorientation 273 acc:(0.9948)

31. Site=MGH Age Group=50-59 Gender=M Language=English Admission type=Elective 273 ==> Reason=Agitation 273 acc:(0.9948)

The Predictive Apriori algorithm could find out more resulting attribute values in the association rules. Each reason for hiring sitter was found to be associated to specific attribute values. For instance, "Disorientation" happened more likely to elderly patients. "Suicidal" patients were mostly middle aged female. Agitated patients were mostly middle aged male.

Year 2009

1. Age Group=80-89 Length of stay group=90-99 Discharge location=Long term care 298 ==> Reason=Disorientation 298 acc:(0.99483)

2. Admission type=ER Length of stay group=90-99 Discharge location=Long term care 298 ==> Reason=Disorientation 298 acc:(0.99483)

3. Mission=Medicine Municipality=DOLLARD-DES-ORMEAUX Admission type=Elective 287 ==> Reason=Suicidal 287 acc:(0.9948)

4. Mission=Medicine Municipality=DOLLARD-DES-ORMEAUX Length of stay group=>=100 287 ==> Reason=Suicidal 287 acc:(0.9948)

5. Age Group=70-79 Municipality=DOLLARD-DES-ORMEAUX Admission type=Elective 287 ==> Reason=Suicidal 287 acc:(0.9948)

45. Language=English Length of stay group=90-99 Discharge location=Long term care 271 ==> Reason=Disorientation 271 acc:(0.99476)

48. Mission=Surgery Marital status=MARRIED_ADULT Municipality=LACHINE 269 ==> Reason=Agitation 269 acc:(0.99476)

55. Mission=Medicine Age Group=80-89 Marital status=MARRIED_ADULT Length of stay group=90-99 269 ==> Reason=Disorientation 269 acc:(0.99476)

Although "Agitation" was the most frequent sitter reason in the dataset, it did not appear as top rules found with highest accuracy.

Year 2010

1. Site=RVH Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Language=French Admission type=ER Length of stay group=60-69 295 ==> Reason=Disorientation 295 acc:(0.99496)

5. Mission=Medicine Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Language=French

Admission type=ER Length of stay group=60-69 291 ==> Reason=Disorientation 291 acc:(0.99496)

25. Age Group=50-59 Admission type=Clinic Length of stay group=>=100 221 ==> Reason=Agitation 221 acc:(0.99492)

76. Mission=Medicine Site=RVH Marital status=MARRIED_ADULT Language=French Admission type=Clinic Length of stay group=>=100 176 ==> Reason=Away without leave 176 acc:(0.99485)

94. Mission=Neuro Language=English Length of stay group=60-69 147 ==> Reason=Constant observation in 4-point restraints 147 acc:(0.99477)

Two additional reasons were found in rules in 2010 but not in previous years. They were "Away without leave" and "Constant observation in 4-point restraints". However, not many supports were found in those rules. Other rules found gave more or less similar information than the ones from previous years.

6.2 COMPARE WITH CLASSIFICATION

From the association rules mining analysis, it has been noted that dividing datasets into different years had given no significant benefits to the results. In order to simplify the process of classification, the analysis is being carried out on combined datasets of all three years, from 2008 to 2010. Not all the attributes should be classified since classification done on some of them can return meaningless results. Naïve Bayes and C4.5 classifiers are chosen for

comparison because they cover a variety of techniques with different representational models. Naïve Bayes is based on probabilistic models and C4.5 algorithm is based on decision tree models.

Only child population (2008 = 1207 instances, 2009 = 1143 instances, 2010 = 988 instances)

Since the datasets are not very large, the experiment will be performed to the combined datasets. With datasets of all 3 years together, there are only 3338 instances for the child population.

Same as in association rule analysis, some of the attributes must be removed since they would provide biases to the outcomes, as indicated in the above section "Approach".

Naïve Bayes Classifier vs. C4.5

With attribute = “Month”

Correctly Classified Instances = 1652 (49.4907 %)

Incorrectly Classified Instances = 1686 (50.5093 %)

Kappa statistic = 0.4354

Contingency table (Row = Real class value, Column = Class label to be classified as)

Table 12 - Contingency table of the Month attribute by Naïve Bayes classification

Class	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
Jan	84	15	5	1	20	5	0	18	7	11	18	3
Feb	19	30	71	2	3	0	0	19	10	22	13	12
Mar	0	11	281	6	4	10	0	23	23	39	22	45
Apr	1	3	77	36	22	1	0	7	9	50	2	10
May	0	0	29	2	75	27	0	30	4	37	11	20
June	12	7	28	12	10	55	0	7	10	8	6	4
July	0	0	8	1	0	1	43	19	4	3	3	4
Aug	2	0	11	0	0	4	25	272	37	17	4	7
Sep	0	1	8	0	1	6	2	58	259	66	1	14
Oct	1	4	33	3	2	6	4	12	70	232	22	18
Nov	10	10	27	4	1	0	0	5	23	42	139	39
Dec	0	9	27	0	20	3	0	0	29	9	43	146

Assuming the attribute "Month" was unknown, the Naïve Bayes algorithm was not able to predict most of the "Month" class labels correctly. For instance, whatever showed in red in the above contingency matrix represented correct prediction of class labels. The sum of all the diagonal numbers in red was the total number of correctly classified instances. Anything outside of the diagonal line (non-red numbers) were number of wrong predictions. The sum of all other numbers in the matrix represented the total number of incorrectly classified instances. Kappa statistic measures the agreement of prediction with the true class. A perfect prediction should show a value of 1. Kappa is widely used to measure inter-observer variability (i.e.: How often two or more observers agree in their interpretations?) The proportion of agreements between yes and no is not a good measure of agreement because it does not correct for chance and take it into consideration. Kappa is the preferred statistic because it accounts for chance. A classification can have a high "hit" rate (high percentage of correctly classified instances) but low Kappa statistic or vice versa. According to Landis and Koch [31], the result from the Kappa statistic can be interpreted as follows.

Table 13 - Kappa statistic general interpretation

Value	Strength of agreement
Less than 0	Poor
0 to 0.2	Slight

0.21 to 0.4	Fair
0.41 to 0.6	Moderate
0.61 to 0.8	Substantial
0.81 to 1	Almost perfect

However, in this experiment, it was only 0.4354. So, it means that the prediction of "Month" was not reliable.

Although the Naïve Bayes could not produce good classification result, the C4.5 classifier could assign most tuples correctly to the right class label. The Kappa statistic was also very high. Comparing to the other classifier Naïve Bayes, C4.5 did a much better job with much higher accuracy in this case.

Similar experiments were being done with other attributes. To summarize, here is a table with classification results by Naïve Bayes classification algorithm and C4.5, with different attributes.

Pediatric sitter cases

Table 14 - Results of performance indicators of pediatric sitter case classification

	Naïve Bayes					C4.5				
	Correct	%	Incorrect	%	Kappa	Correct	%	Incorrect	%	Kappa
Month	1652	49.49	1686	50.51	0.4354	2708	81.13	630	18.87	0.7904
Day	174	5.21	3164	94.79	0.0163	356	10.67	2982	89.33	0.0746
Cost center	2434	72.92	904	27.08	0.5248	2929	87.75	409	12.25	0.7832

Shift	1059	31.73	2279	68.27	-0.029	1148	34.39	2190	65.61	0.0007
Reason	2655	79.54	683	20.46	0.7164	3183	95.36	155	4.64	0.9365
Age group	3201	95.90	137	4.10	0.8909	3303	98.95	35	1.05	0.9719
Gender	2835	84.93	503	15.07	0.6987	3218	96.41	120	3.59	0.9281
Language	2720	81.49	618	18.51	0.6761	3242	97.12	96	2.88	0.9491
Municipality	2282	68.36	1056	31.64	0.6461		0.00		0.00	
Admission type	3211	96.20	127	3.80	0.6484	3310	99.16	28	0.84	0.9087
Length of stay group	2330	69.80	1008	30.20	0.6167	3144	94.19	194	5.81	0.9264

Only adult population (2008 = 21765 instances, 2009 = 22229 instances, 2010 = 19567 instances)

Adult sitter cases

Table 15 - Results of performance indicators of adult sitter case classification

	Naïve Bayes					C4.5				
	Correct	%	Incorrect	%	Kappa	Correct	%	Incorrect	%	Kappa
Month	11823	28.60	29509	71.40	0.2217	32456	78.53	8876	21.47	0.7655
Day	1261	3.05	40071	96.95	-0.0025	2531	6.12	38801	93.88	0.0293
Mission	30522	73.85	10810	26.15	0.6037	39826	96.36	1506	3.64	0.945
Site	33337	80.66	7995	19.34	0.6133	40661	98.38	671	1.62	0.9675
Shift	14870	35.98	26462	64.02	0.0319	15075	36.47	26257	63.53	0.0395
Reason	25834	62.50	15498	37.50	0.4434	40267	97.42	1065	2.58	0.9626
Age group	22006	53.24	19326	46.76	0.4369	40238	97.35	1094	2.65	0.9683
Gender	31018	75.05	10314	24.95	0.4436	40559	98.13	773	1.87	0.9593
Marital status	28034	67.83	13298	32.17	0.5141	40479	97.94	853	2.06	0.9688
Language	27424	66.35	13908	33.65	0.4105	40516	98.03	816	1.97	0.9662
Municipality	23713	57.37	7619	18.43	0.452	40150	97.14	1182	2.86	0.9644
Admission type	32611	78.90	8721	21.10	0.6219	40465	97.90	867	2.10	0.9645

Length of stay group	22156	53.60	19176	46.40	0.4631	40716	98.51	616	1.49	0.9828
Discharge location	24690	59.74	16642	40.26	0.4083	40384	97.71	948	2.29	0.9666

6.3 COMPARE WITH CLUSTERING

The same dataset as above, sitter usage, is being used in this experiment. This time, the emphasis is focused on the pediatric population. Problems with teenagers always cause concern to the society. By focusing on a specific subset of population, some interesting facts about teenagers may be discovered. Some teenage patients may belong to specific "category", as known as cluster. Health practitioner can put more focus and adjust their practice to target those patients. Also, potential social problems can be identified in the cluster. For example, why are there so many female suicidal patients around the age of 14? Is it because of the pressure caused by the transition from primary to high school? The government may be able to use the clustered data to see what the society needs. Additional services may be able to provide to reduce such incidents.

Since the sitter usage dataset contains a lot of non-numerical values, some clustering algorithms cannot be used. To do the clustering for such data, algorithms that support similarities are being used.

In order to be able to make comparisons, two iterative approach algorithms are chosen, as follows.

- Expectation-Maximization (EM)

- K-means

The main reason why the above algorithms are chosen is that they are all related to K-means. K-means is the original algorithm. EM is the variant. Attributes of clinical data like the sitter usage data can be both continuous and categorical. The K-means algorithm can handle both continuous and discrete data and perform clustering based on anticipated likelihood attributes with core attributes of patient characteristics in data point.

K-means clustering algorithm needs to have number of clusters predefined, unlike the expectation-maximization (EM) algorithm. EM can determine the number of clusters automatically by cross validation to maximize the probability. In this experiment, expectation-maximization is being used first to determine the number of clusters. Then, the subsequent algorithm, K-means, will use the same number of clusters than what is determined by the EM. Comparisons will be done based on the characteristics and frequently occurred tuples between the generated clusters, from both text and graphic results.

In general, clustering algorithms do their job by evaluating Euclidean distances between attribute values. The algorithms try to group tuples with closest distances together while maximizing the distances between different clusters. The Euclidean distance is being calculated as follows, given p and q are attribute values from two different tuples.

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

In order to have fair distance measure between attributes, attributes to be clustered must first have the values normalized, as discussed in chapter 3 earlier. Dataset used in the following experiments, specifically the attributes age and length of stay, had first been normalized before clustering took place.

Only child population (2008 = 1207 instances, 2009 = 1143 instances, 2010 = 988 instances)

Expectation-Maximization

Number of clusters is automatically set by cross validation.

Clustered Instances

0 279 (8%)

1 396 (12%)

2 41 (1%)

3 127 (4%)

4 832 (25%)

5 524 (16%)

6 128 (4%)

7 354 (11%)

8 297 (9%)

9 78 (2%)

10 282 (8%)

A total of 11 clusters could be found by the EM. However, some of them were so small and only occupied tiny percentages (< 10%). Those clusters could be thrown away because they were not statistically relevant. No good conclusions could be drawn from those clusters. So, we ended up with 4 clusters (in bold and italic) selected from the cluster analysis.

With 4 clusters chosen, the EM clustering was re-run. Attribute values having means that stood out from others were chosen to represent the characteristics of the cluster.

Table 16 - Characteristics of clusters found from pediatric sitter cases (EM)

Cluster	Characteristics of centroids	%
0	Month=August, CC=32806, Reason=Other, Age=0.3181, Gender=Female, Language=English, Municipality=Montreal, Admission type=Clinic, Length of stay=0.1238	16
1	Month=October, CC=32806, Reason=Suicidal, Age=0.8075, Gender=Female, Language=English, Municipality=Montreal,	47

	Admission type=Clinic, Length of stay=0.0103	
2	Month=March, CC=32806, Reason=Other, Age=0.651, Gender=Male, Language=English, Municipality= SAINT-LAZARE, Admission type=Clinic, Length of stay=0.0275	24
3	Month=October, CC=32801, Reason=Agitation, Age=0.7705, Gender=Male, Language=English, Municipality= SAINTE-CLOTILDE-DE-CHATEAUGUAY, Admission type=Clinic, Length of stay=0.0226	14

The graph of "Reason vs. Cluster number" is chosen because certain areas clearly appear to be denser.

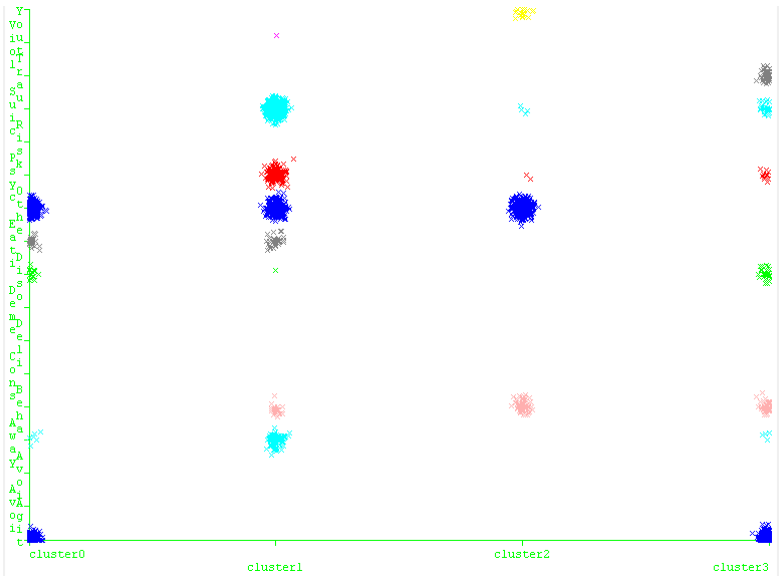


Figure 3 - Graph of Reason vs. Cluster (EM). Colors denote Reason.

By observing the denser areas of different graphs with different attribute combinations, characteristics of the centroids could be discovered. In this chart, upper blue and cyan mass appeared to be obviously larger and denser than the others. The bottom blue mass denoted agitated cases and the upper ones denoted cases with "Other/Unspecified" reasons. The bottom cyan mass denoted "Away without leave" cases, whereas the upper ones denoted suicidal cases. The above chart basically showed the same observations as in the previous table. Clusters were not divided mainly on sitter reasons, since each cluster consisted of very mixed reasons.

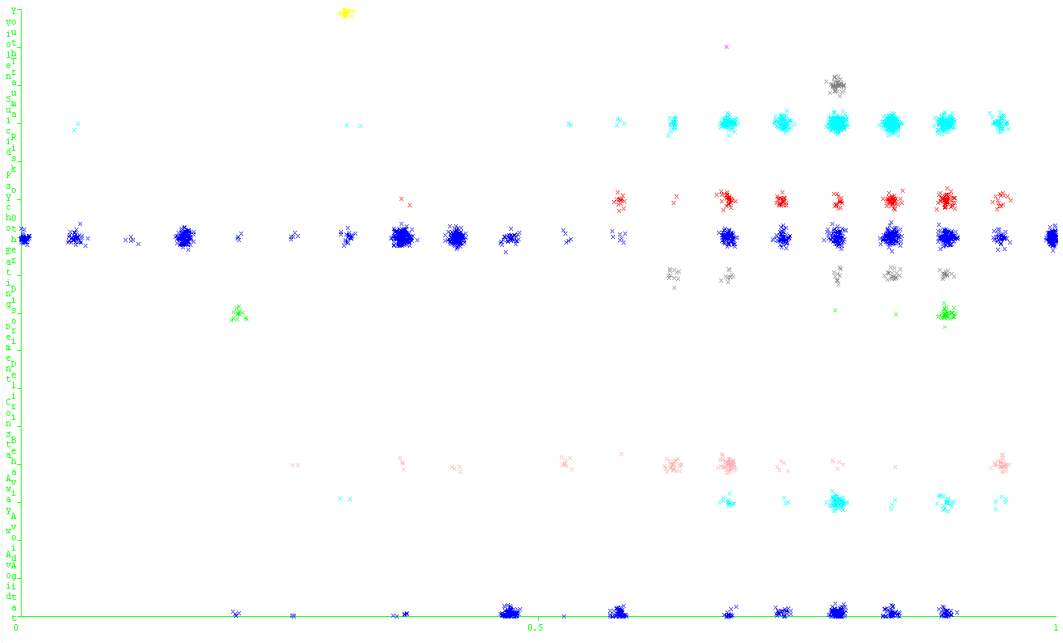


Figure 4 - Graph of Reason vs. Age (EM)

As the chart clearly showed, most denser areas appeared after the 0.5 scale of age, which translated into 9 years and older. Most sitter case reasons

had bias towards older teenagers. This especially happened to suicidal patients (top cyan). They tended to be more concentrated in the right side of the chart. Same happened to "Away with leave" (lower cyan), "Behavior problem" (pink) and "Trauma" (top gray) patients. This might due to the fact patients needed to reach certain ages to develop those problems, thus requiring sitter supervision.

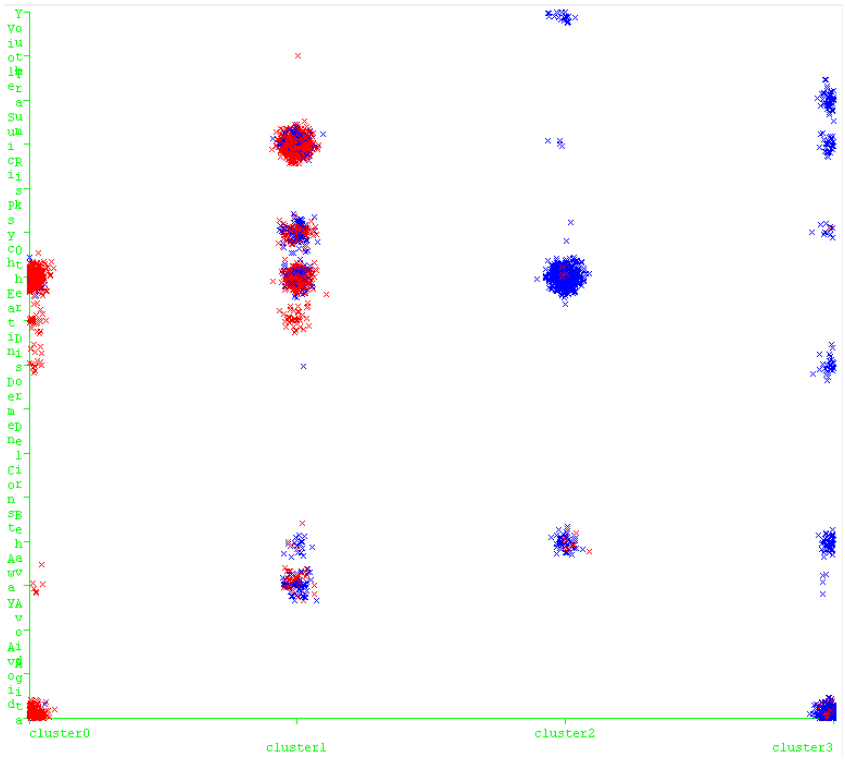


Figure 5 - Graph of Reason vs. Cluster (EM). Colors denote Gender.

This is almost the same graph as the previous one, except the masses were color coded to represent gender (female = red, male = blue). It seemed like the clustering algorithm had used gender as the primary attribute to divide the datasets. Cluster 0 and 1 had mostly female patients and the others had mostly

male patients. The top red mass in cluster 1 showed that a lot of suicidal cases were with female patients. The top blue mass in cluster 2 showed that many cases with "Other/Unspecified" reason were with male patients. Both cluster 0 and 3 had significant density of agitated cases, representing female and male patients respectively. "Psychosis", "Violent" and "Youth protection" as sitter reasons seemed to be specific to cases with male patients, as they only appeared in cluster 2 and 3 in blue.

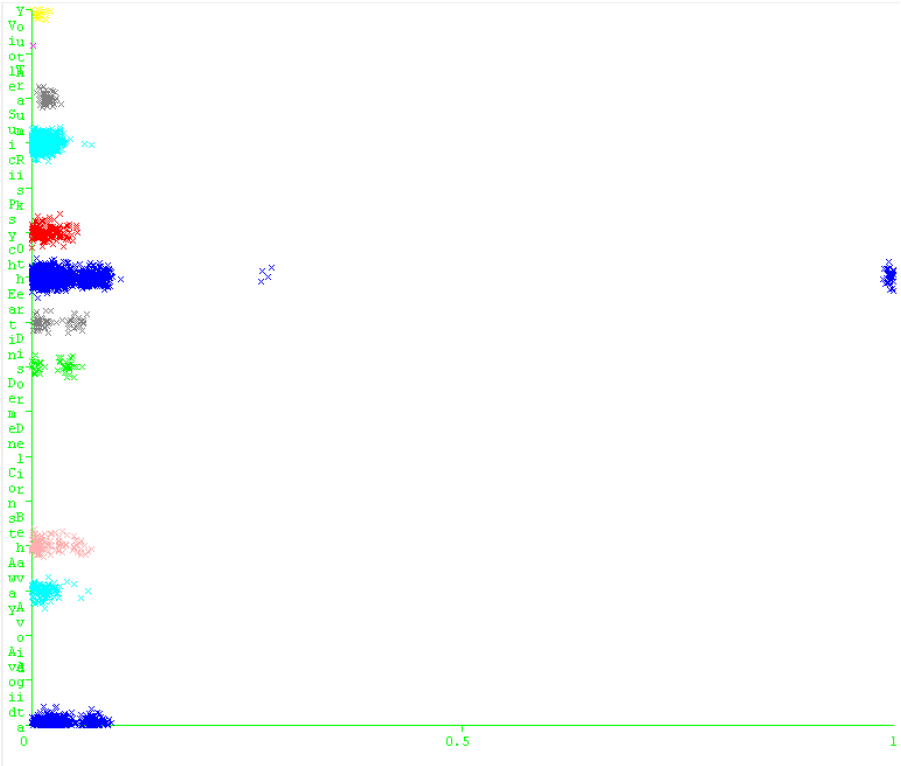


Figure 6 - Graph of Reason vs. Length of stay (EM)

As most of the masses appeared to the left side of the chart, those cases had mostly relatively short length of stay. The only cases with longer length of

stay were with the reason "Other/Unspecified". Other combinations of attributes resulted in charts with very scattered bubbles without any concentrated denser areas. Some of them resulted in only one color, due to strong bias in some attribute value distribution. For instance, admission type. As almost 95% of admission types were "Clinic", there was nothing interesting to see with such attribute.

From the observed characteristics, cluster #1 had the largest patient population, with 47% of pediatric sitter cases. It was also interesting to find out that male patients tended to be related to physical problems such as "Agitation", "Violent" and "Youth protection", like what were discovered in the association rules. The mean age of patients requiring sitters was around 12 with standard deviation of 0.2. This might indicate that teenage patients around that age tended to have more problems and they needed to be constantly supervised. However, most pediatric sitter cases did not require long lengths of stay. Most cases seemed to happen around the school period, from August to October. However, those cases were mostly with the sitter reason "Other/Unspecified".

K-means

To make comparisons, number of clusters was set to 4 to match the number of clusters and result from the EM algorithm.

Table 17 - Characteristics of clusters found from adult sitter cases (K-means)

Cluster	Characteristics of centroids	%
0	Month=December, CC=32806, Reason=Suicidal, Age=0.8139, Gender=Female, Language=French, Municipality=Montreal, Admission type=Clinic, Length of stay=0.0107	28
1	Month=March, CC=32806, Reason=Other, Age=0.6906, Gender=Male, Language=English, Municipality=SAINT-LAZARE, Admission type=Clinic, Length of stay=0.0358	30
2	Month=October, CC=32801, Reason=Agitation, Age=0.7514, Gender=Male, Language=French, Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY, Admission type=Clinic, Length of stay=0.0193	16
3	Month=August, CC=32806, Reason=Other, Age=0.5198, Gender=Female, Language=English, Municipality=Montreal, Admission type=Clinic, Length of stay=0.064	27

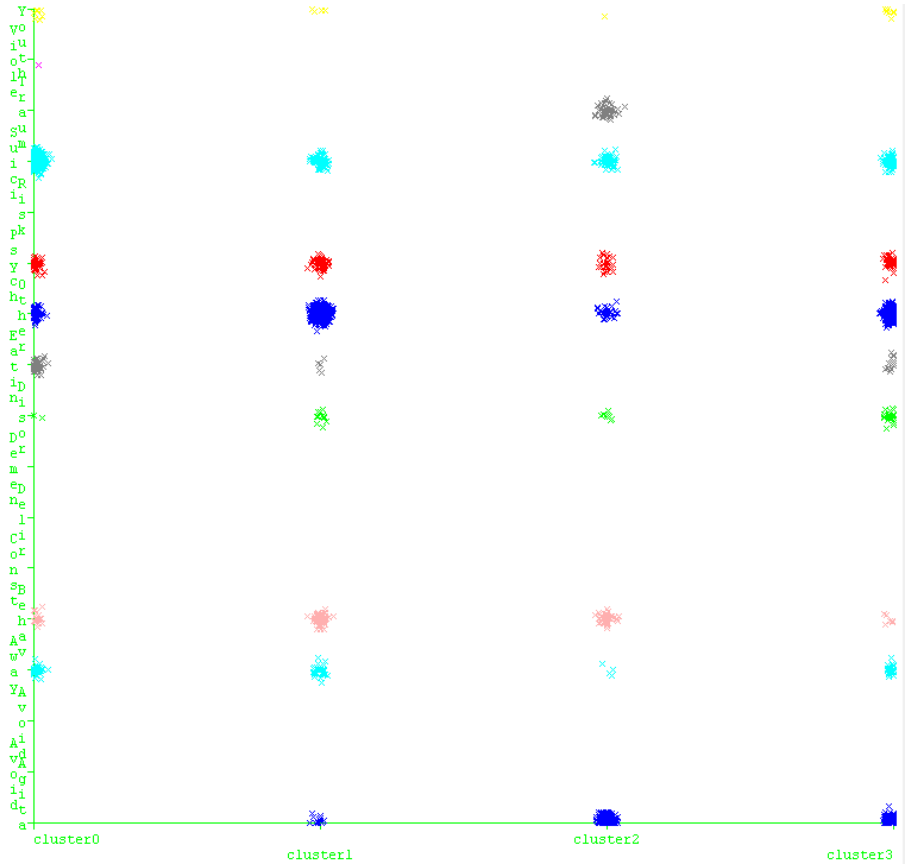


Figure 7 - Graph of Reason vs. Cluster (K-means). Colors denote Reason.

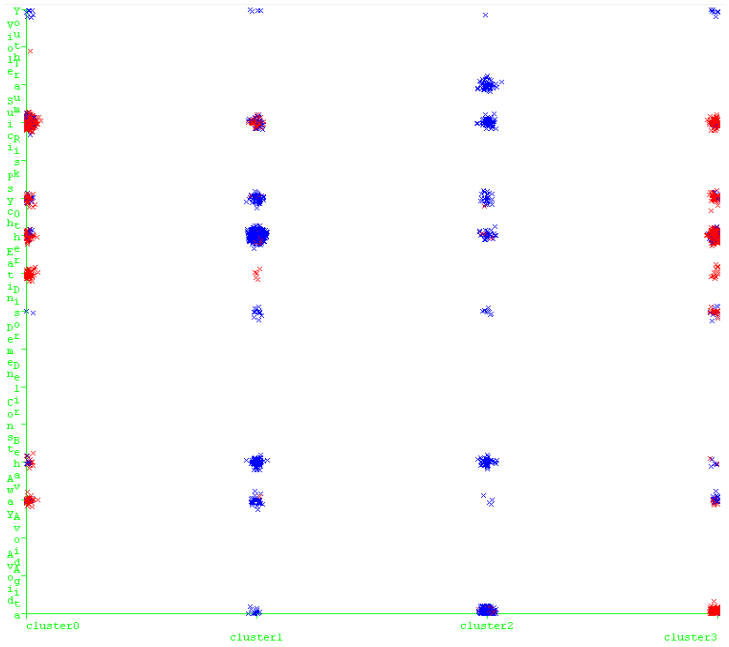


Figure 8 - Graph of Reason vs. Cluster (K-means). Colors denote Gender.

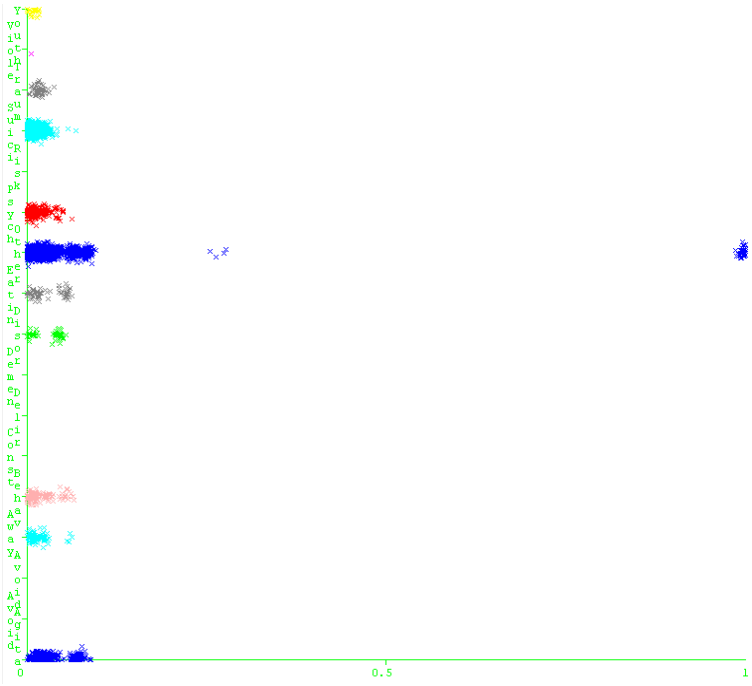


Figure 9 - Graph of Reason vs. Length of stay (K-means)

Although the generated clusters had different percentages, the characteristics of centroids were quite similar to the ones from EM. Significant portions of male agitated patients and female suicidal patient could be found.

Only adult population (2008 = 21765 instances, 2009 = 22229 instances, 2010 = 19567 instances)

Before clustering took place, here was some discovered info about relationships between parameters. Relationship with interesting observations were captured and shown in the following charts. They might offer some important hints about what clusters the algorithms would come up.

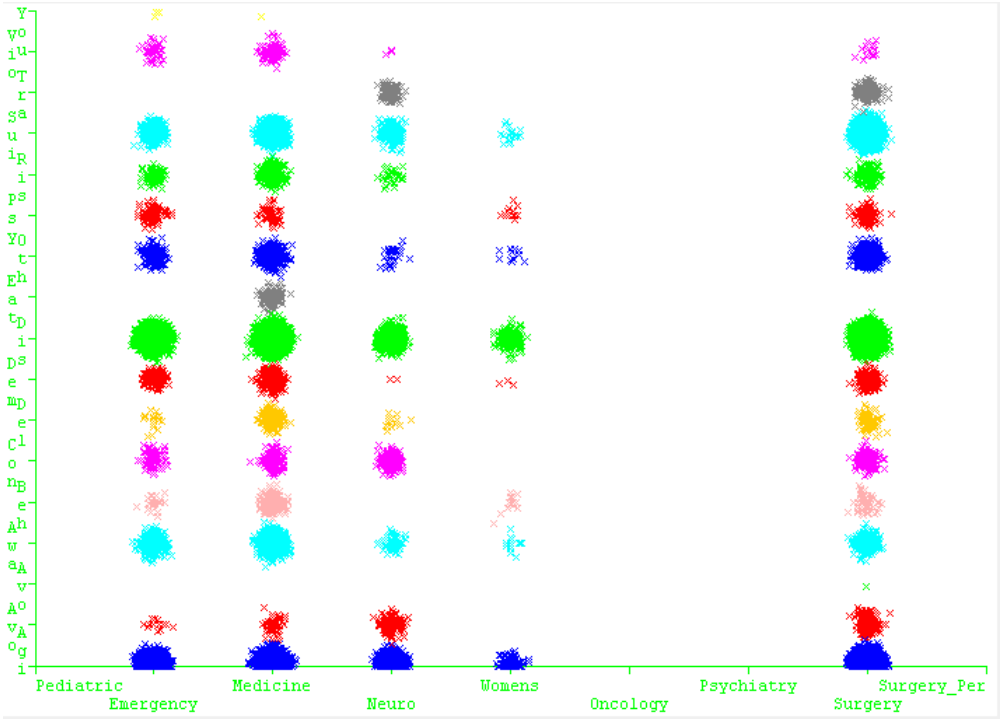


Figure 10 - Graph of Reason vs. Hospital mission

Mission seemed to play a role in the clustering. Surgical mission had most varieties of reasons for sitter cases. In Womens' health mission, most cases were with disoriented and agitated patients, shown in blue and green masses. Sitter cases with "eating disorder" patients only happened in Medical mission, whereas trauma cases only happened in Neuroscience mission, shown in lower gray and upper gray bubbles respectively. In all cases, most sitter cases were with agitated, disoriented and suicidal patients.

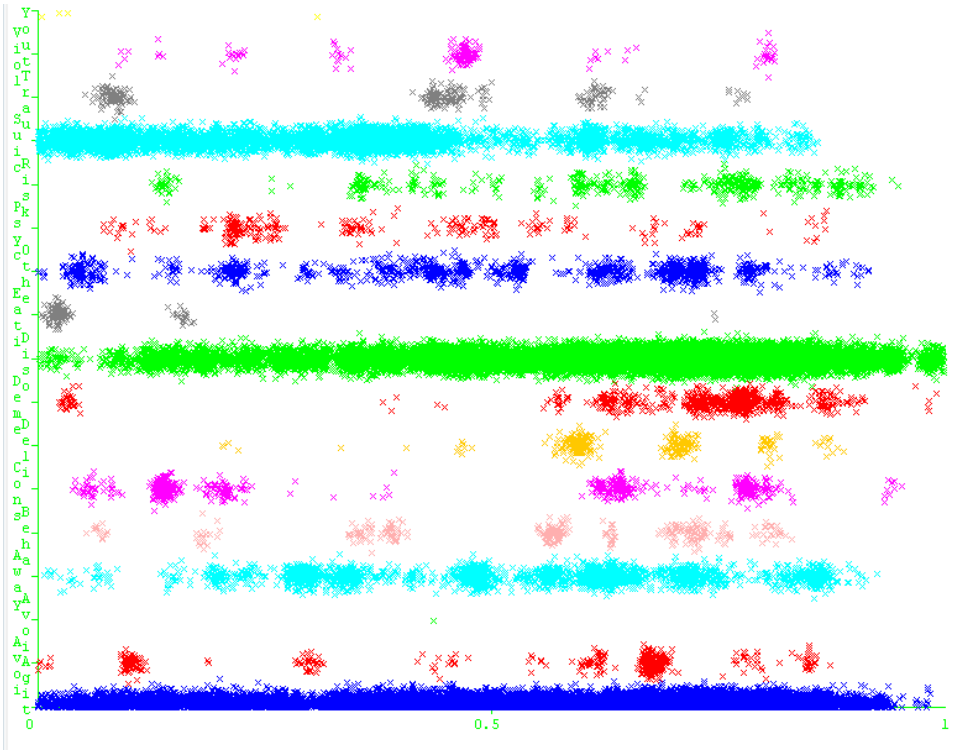


Figure 11 - Graph of Reason vs. Age

The chart clearly showed the age distribution for different sitter reasons. "Disoriented" sitter cases became more (more dense) with older age patients,

whereas suicidal cases (top cyan mass) seemed to have younger patients. Agitated cases happened to patients across all ages quite evenly. "Demantia" cases seemed to happen to elderly patients, shown in red (2nd red mass from the bottom).

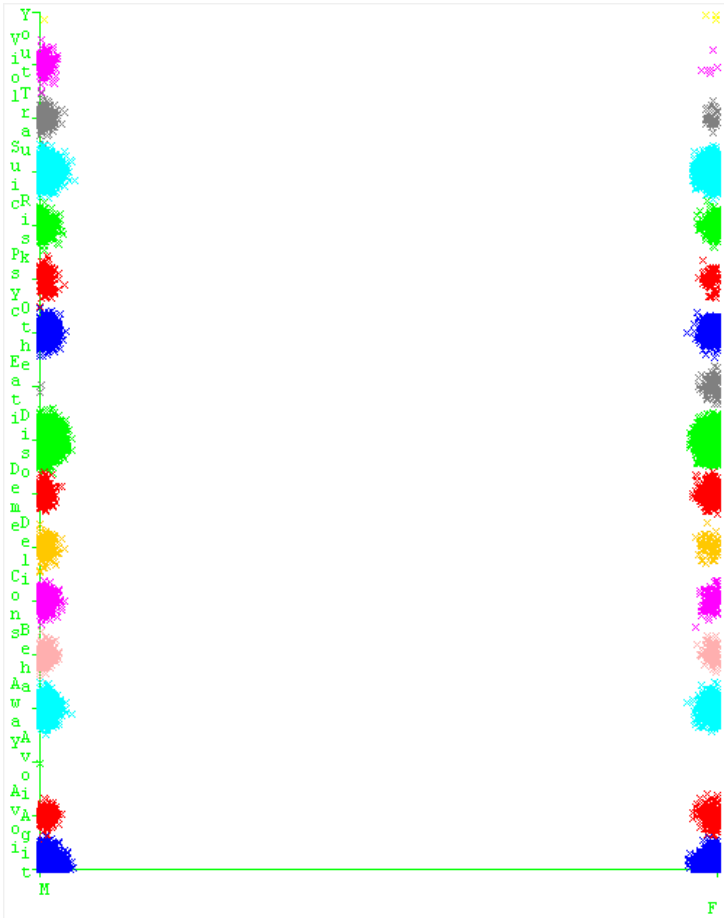


Figure 12 - Graph of Reason vs. Gender

Some sitter reasons were quite related to patients' genders. For instance, "Eating disorder" seemed to happen only to female patients, shown in the first gray circle from the bottom. Violent and trauma sitter cases were mostly with

male patients, shown in first pink and gray masses from the top. Other reasons did not seem to be too gender specific.

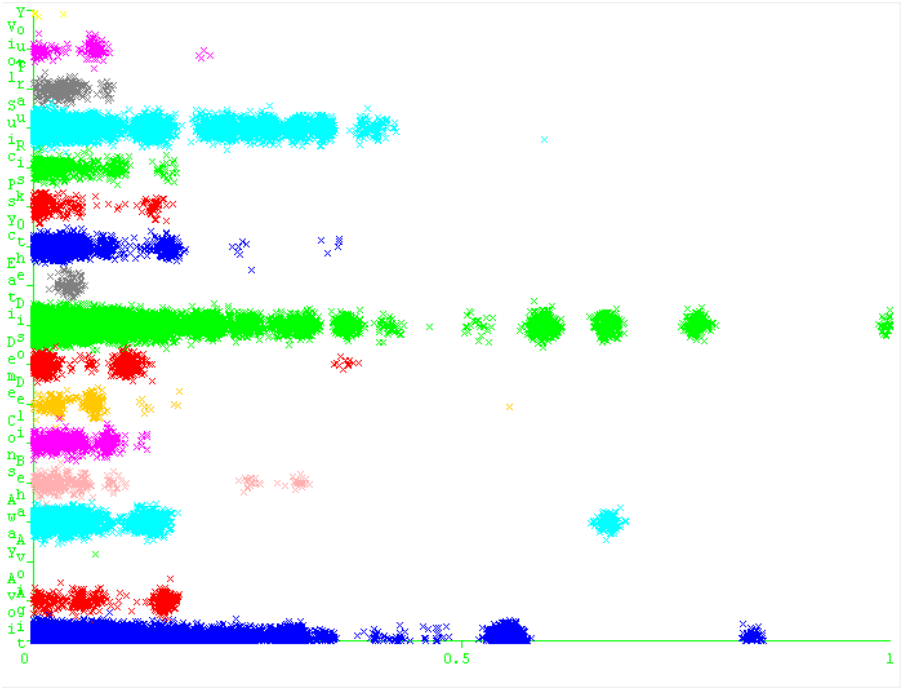


Figure 13 - Graph of Reason vs. Length of stay

The charts showed the relationship between sitter reason and length of stay. The longest middle green mass represented disoriented cases. Those cases caused longest lengths of stay. The bottom cyan mass showed that "Away without leave" cases had resulted mostly short lengths of stay. However, some much lengthier cases did occur. This was the same for agitated cases, represented in the bottom blue. The top cyan mass represented suicidal cases. Almost none of those cases resulted in lengthy stays (less than half way mark 0.5 → 50 days).

Expectation-Maximization

Number of clusters was automatically set by cross validation.

Clustered Instances

0 959 (2%)

1 878 (1%)

2 1438 (2%)

3 962 (2%)

4 1151 (2%)

5 1898 (3%)

6 4862 (8%)

7 298 (0%)

8 1391 (2%)

9 6385 (10%)

10 2169 (3%)

11 2866 (5%)

12 1635 (3%)

13 615 (1%)

14 2149 (3%)

15 6364 (10%)

16 598 (1%)

17 2779 (4%)

18 1987 (3%)

19 2053 (3%)

20 1308 (2%)

21 1779 (3%)

22 728 (1%)

23 2236 (4%)

24 4980 (8%)

25 3356 (5%)

26 5737 (9%)

Although 26 clusters were found, most of them were statistically irrelevant due to very low support value resulting very small clusters. After throwing away all the clusters with less than 10% of count, only two clusters 9 and 15 could

remain. To be able to make comparisons with the pediatric sitter case clustering, same number of clusters (i.e.: 4 clusters), was chosen instead.

With 4 clusters chosen (in bold and italic), the EM clustering was re-run. Attribute values having means that stood out from others were chosen to represent the characteristics of the cluster.

Table 18 - Characteristics of clusters found from adult sitter cases (EM)

Cluster	Characteristics of centroids	%
0	Month=April, Mission=Medicine, Site=RVH, Cost center=52603, Reason=Disorientation, Age=0.6502, Gender=Male, Marital status=Married, Language=English, Municipality=Montreal, Admission type=ER, Length of stay=0.1525, Discharge location=Home	34
1	Month=October, Mission=Surgery, Site=MGH, Cost center=22101, Reason=Suicidal, Age=0.2997, Gender=Male, Marital status=Single, Language=French, Municipality=Montreal, Admission type=ER, Length of stay=0.0829, Discharge location=Home	24
2	Month=August, Mission=Emergency, Site=MGH, Cost center=22101, Reason=Disorientation, Age=0.6728, Gender=Male, Marital status=Married, Language=English, Municipality=Montreal,	28

	Admission type=ER, Length of stay=0.0685, Discharge location=Home	
3	Month=March, Mission=Emergency, Site=RVH, Cost center=53201, Reason=Disorientation, Age=0.6595, Gender=Male, Marital status=Married, Language=English, Municipality=Montreal, Admission type=ER, Length of stay=0.047, Discharge location=Hospital	15

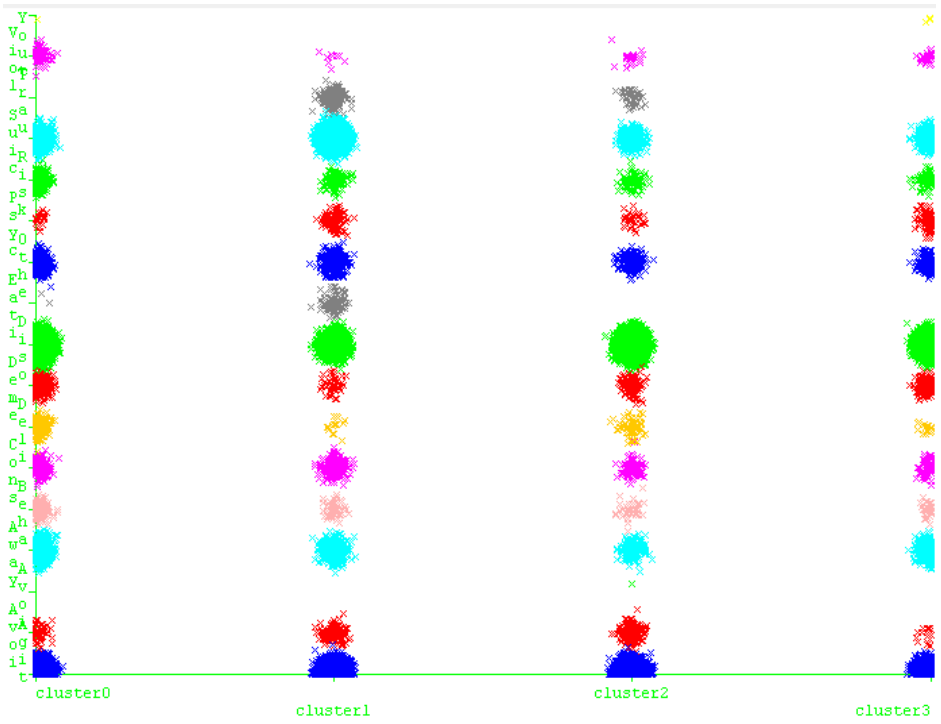


Figure 14 - Graph of Reason vs. Cluster (EM). Colors denote Reason.

Clustering was done not only based on reasons, since varieties of reasons existed in each cluster. The three masses from the bottom, blue, green and cyan represented agitated, disoriented and suicidal patients respectively. Those three masses were dominant in all the clusters. Overall, the distribution of reasons was quite even between clusters. "Reason" was probably not the only parameter which had the most major influence to divide the clusters.

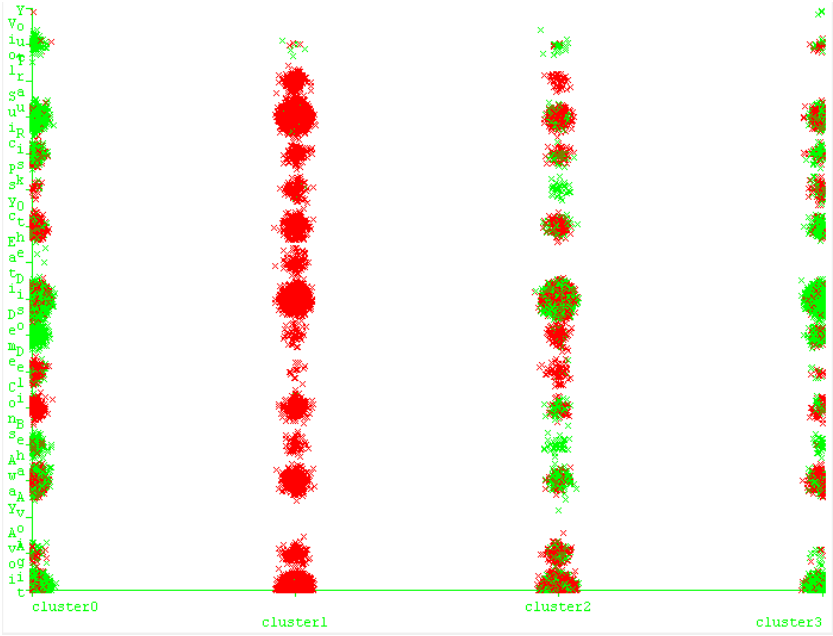


Figure 15 - Graph of Reason vs. Cluster (EM). Colors denote Hospital site.

Cluster 1 seemed to contain only patients of one site - Montreal General Hospital (MGH). Cluster 0 contained mostly patients from the hospital site, Royal Victoria Hospital (RVH).

K-means

To make comparisons, number of clusters was set to 4 to match the number of clusters and result from the EM algorithm.

Table 19 - Characteristics of clusters found from adult sitter cases (K-means)

Cluster	Characteristics of centroids	%
0	Month=August, Mission=Medicine, Site=RVH, Cost center=52051, Reason=Agitation, Age=0.6234, Gender=Male, Marital status=Married, Language=French, Municipality=Montreal, Admission type=ER, Length of stay=0.1728, Discharge location=Home	21
1	Month=September, Mission=Surgery, Site=MGH, Cost center=22101, Reason=Agitation, Age=0.4275, Gender=Male, Marital status=Single, Language=English, Municipality=Montreal, Admission type=ER, Length of stay=0.0794, Discharge location=Home	38
2	Month=July, Mission=Medicine, Site=RVH, Cost center=52603, Reason=Disorientation, Age=0.7068, Gender=Female, Marital status=Separated, Language=English, Municipality=Montreal, Admission type=ER, Length of stay=0.0603, Discharge location=Hospital	25

3	Month=May, Mission=Surgery, Site=RVH, Cost center=52156, Reason=Disorientation, Age=0.6419, Gender=Male, Marital status=Married, Language=French, Municipality=Montreal, Admission type=Urgent, Length of stay=0.0597, Discharge location=Home	16

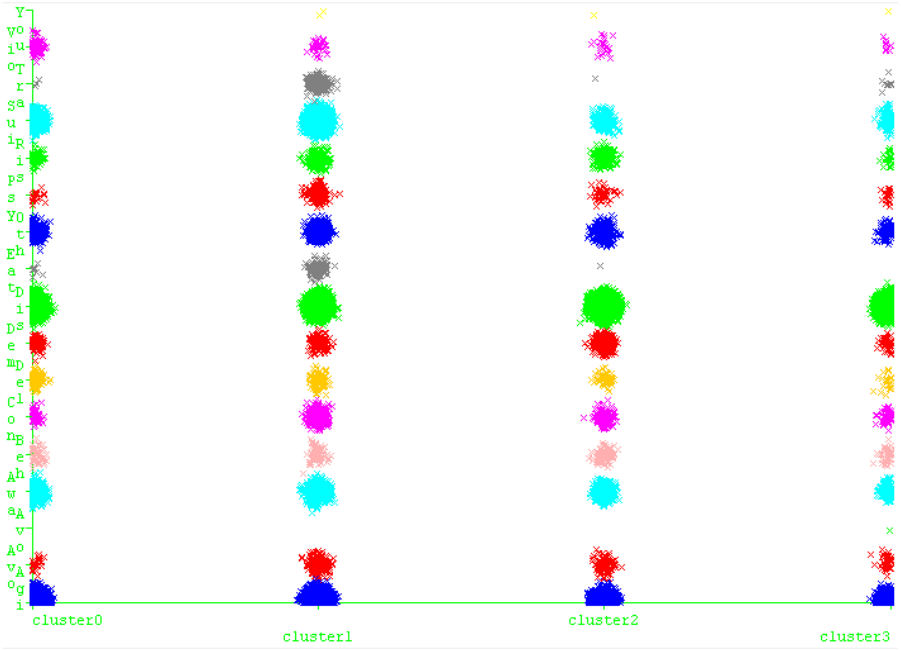


Figure 16 - Graph of Reason vs. Cluster (K-means). Colors denote Reason.

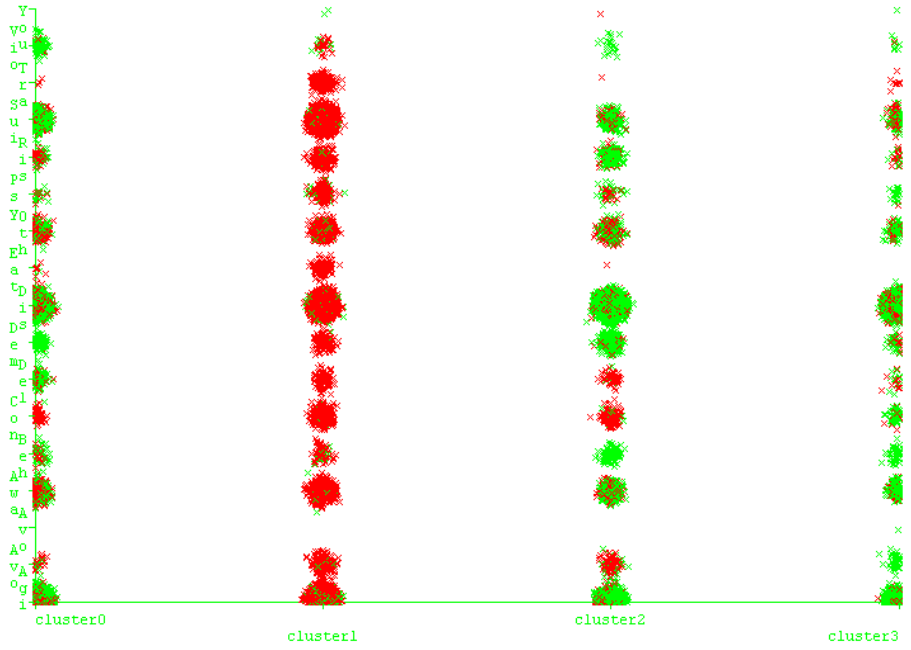


Figure 17 - Graph of Reason vs. Cluster (K-means). Colors denote Hospital site.

The clustering results contained quite a bit of similarities but also with some differences. For instance, both clustering algorithms divided the clusters in medicine and surgery missions, except the EM also had ER in cluster 2 & 3. Both algorithms had cluster centroids with disorientation as the mostly used reason. In EM, 3 of the 4 centroids had disorientation as the reason, whereas K-means had only 2. Both algorithms showed that cases with reason "Disorientation" had all been associated with middle-age patients on average, with age > 0.5, which translated into 50-year-old. All centroids had relatively short lengths of stay.

6.4 OBSERVATION

6.4.1 Association rule extraction

Based on the above experiments on each type of patient populations (pediatric vs. adult), different findings have been observed. For the pediatric patient population, the following facts have been noticed.

- Female pediatric patients were more prone to commit suicide;
- March, September and December seemed to be problematic months for both genders;
- Male pediatric patients were more likely to have physically related problems such as agitation;

For the adult patient population, the following facts have been observed.

- Month and Day did not seem to have any relationships with other attributes in the dataset;
- Municipality did not seem to have any ties with any parameters in the dataset;
- Cases with admission type "ER" seemed to be related to longer length of stay than the ones with "Stretcher".
- The "Medicine", "Neuroscience" and "Surgery" divisions were popular divisions in many association rules;

- Female adult patients were more likely to have "Suicidal" or "Eating disorder" problems. Same as in pediatric male patient population, male adult patients were more likely to have physically related problems such as agitation and constant observation in 4-point restraints;
- Length of stay for agitated or suicidal patients was longer than disorientated patients;
- Some problems seemed to be related to certain divisions. "Suicidal" adult patients were likely to be in the "Surgery" division. "Eating disorder" and "Behavior problem" patients seemed to be likely in the "Medicine" division.
- "Suicidal" adult patients tended to be female for year 2008 and 2009. However, in year 2010, rules were found with male suicidal adult patients too. All suicidal patients had length of stay more than 100 days. Also, they were very likely to be discharged home.
- If the patient requiring sitter supervision had either "Agitation" or "Suicidal" problem in the "Neuroscience" division, the patient had a relatively high chance to be male;
- Patients from the "Neurosciences" division had relatively high chances to have "Agitation" problem; Patients from the "Medicine" division had relatively high chances to have "Disorientation" problem;

- Disoriented patients were associated with older age between 70 and 79, and had "Married adult" as the marital status.
- Patients with more than 70-year-old were more likely to have "Dementia" and "Agitation" problem.

Although datasets were divided into different year and analyses were being done to each of them separately, results came out to be more or less similar year to year. Rules discovered did not show major differences. Indeed, results were expected to be very similar if analysis had been applied to the entire dataset of three years (2008 to 2010). The above observations can serve as a summarized picture of important relationships between attributes. Although the data was from the sitter ordering system, the rule mining could discover information more than just the sitter usage. The rules can provide healthcare workers ideas to pinpoint more problematic areas thus to improve the efficiency of the overall healthcare team. For example, "suicidal" patients were likely to be in the "Surgery" division. Was it mostly due to pain after surgery, insufficient post-op support, short staffing, inflexible visiting hours or anything else?

With a large amount of data, a lot of association rules were generated. A lot of them were not meaningful. Although the class attribute was used as a ”guidance" in rule mining, a careful inspection of each rule was still needed to determine which ones could meet the context. Most of the time, in this experiment, the standard Apriori algorithm could not discover enough meaningful

information from the training data. Predictive Apriori algorithm was needed as a complement to do more thorough rule mining. However, although its accuracy reflected the confidence, it was not proportional to the support count. A lot of rules found (with very high accuracy index) by the predictive Apriori algorithm only had very few support count. Thus, they could not be used to represent something significant. In general, the predictive Apriori algorithm could find more interesting rules than the regular Apriori algorithm.

It has been noted that predictive Apriori algorithm had taken much more time than the regular Apriori algorithm to find the association rules. Some of the above association rule mining on the adult patient population took several hours, on a P4 3.2GHz machine running Windows XP.

Each algorithm has its pros and cons. Regular Apriori needs to have very well defined support and confidence in advance. Predictive Apriori trades support against confidence to find out rules with high probabilities. In this experiment, it offered more in-depth search of association rules than the regular Apriori. In general, it found better results, at the expense of much longer computation time.

Overall, this experiment made use of the power of association rule mining algorithms. Information from a computerized statistical system may not always look interesting before data mining is applied. Data mining allows the finding of often "unseen" messages, hidden inside the data. Although association rule finding can be done "automatically" by computers, a lot of rules found can be

meaningless. Verification by clinical experts is still needed to judge whether the rules discovered are usable.

6.4.1.1 Association rules vs. proposed approach

A list of meaningful association rules discovered by the Apriori approach was extracted. These rules had similar pre case attributes as required by our system. They were being used to validate the prediction results by our proposed recommender system. The recommender system produced results similar to the ones by the classification. Direct comparison between the recommender system and the association rule engines cannot be done in a straight way. Although association rules were done in our experiment with predefined target class attribute, rules can contain any number of predicates, resulting multidimensional rules. However, in the proposed recommender system, a fixed number of predicates must be used to perform the post case attribute value prediction. To be able to do the comparison of results between the recommender system and association rules, predicates of association rules must have very close match with the required input attributes in the recommender system. A list of association rules have been chosen as follows.

Association rules found in the adult dataset

- Mission=Medicine Gender=M Municipality=MONTREAL Admission type=Elective 1242 ==> Discharge location=Home 1082 conf:(0.87)

- Mission=Surgery Site=MGH Gender=M Marital status=SINGLE_ADULT Admission type=ER 1511 ==> Discharge location=Home 1305 conf:(0.86)
- Mission=Surgery Site=MGH Gender=M Marital status=SINGLE_ADULT 1798 ==> Discharge location=Home 1508 conf:(0.84)
- Mission=Surgery Age Group=40-49 Marital status=SINGLE_ADULT Language=French Admission type=ER 263 ==> Discharge location=Home 263 acc:(0.99495)
- Mission=Emergency Gender=F Admission type=Stretcher 1186 ==> Length of stay group=0-9 1148 conf:(0.97)
- Mission=Medicine Site=RVH Admission type=Elective Discharge location=Home 1399 ==> Length of stay group=>=100 1384 conf:(0.99)
- Mission=Medicine Gender=F Marital status=SINGLE_ADULT Admission type=Elective 287 ==> Discharge location=Home 287 acc:(0.99478)
- Mission=Emergency Site=RVH Admission type=Stretcher 2010 ==> Length of stay group=0-9 1879 conf:(0.93)
- Mission=Medicine Gender=M Language=English Admission type=Elective 1253 ==> Length of stay group=>=100 1122 conf:(0.9)
- Mission=Emergency Site=RVH Gender=M 1410 ==> Length of stay group=0-9 1231 conf:(0.87)

Each system (Apriori, Predictive Apriori and recommender system) returns different measurement units to represent how trustful the results can be. Apriori returns support and confidence values for each rule found. Given number of records as "support ", confidence value is the likelihood that such rule can happen. Confidence returned by Apriori can be interpreted as the precision of the prediction. For predictive Apriori, only the accuracy value is returned by the mining engine. The accuracy value is the conditional probability of $x \rightarrow r$ and $y \rightarrow r$, where r is number of records to satisfy the above conditional probabilities. Simply speaking, the confidence must trade against support. A higher confidence value must be supported by higher support value as well. Although definition wise, it is different than the prediction precision from the recommender system, it can provide a feedback to see how accurate the rule is, from the predictive Apriori algorithm. We use the confidence from Apriori, accuracy from predictive Apriori and prediction precision from the recommender system, to compare how well the prediction turned out.

In information retrieval, precision and recall are being used to measure the correctness of the predicted results. F-measure has been used to evaluate various types of data mining activities [57, 58, 59, 60]. According to its definition, it is

F = 2 x precision x recall / (precision + recall)

To be able to get F-measure, precision and recall values are required. In our experiment, they are defined as follows.

precision = |correctly predicted results| ∩ |retrieved results| / |retrieved results|

recall = |correctly predicted results| ∩ |retrieved results| / |relevant results|

|relevant results| means the number of correctly predicted results and |retrieved results| is the number of records filtered by pre case parameter values. Since |relevant results| = |correctly predicted results|, recall value always becomes 1. With recall having the same value all the time, it becomes not meaningful in the F-measure calculation. As a result, only precision value can be used to compare effectiveness of predictions.

Table 20 – Prediction result vs. Apriori rules (Adult dataset)

Rule	Prediction result from our system	Reported precision from our system	Precision/Accuracy
1	Discharge location = Home	0.6843427	0.87
2	Discharge location = Hospital	0.3862645	0.86
3	Discharge location = Home	0.5789563	0.84

4	Discharge location = Home	0.5301262	0.99495
5	Length of stay = 0-9	0.9395185	0.97
6	Length of stay = >=100	0.759461	0.99
7	Discharge location = Home	0.5252708	0.99478
8	Length of stay = 0-9	0.9381583	0.93
9	Length of stay = 0-9	0.01702484	0.9
10	Length of stay = 0-9	0.8345715	0.87

Association rules found in the pediatric dataset

- Gender=M Admission type=Clinic 65 ==> Length of stay group=0-9 65 conf:(1)
- Gender=M Admission type=Elective 66 ==> Length of stay group=20-29 66 acc:(0.99399)
- Gender=F Admission type=Clinic 44 ==> Length of stay group=20-29 44 acc:(0.9928)

Table 21 – Prediction result vs. Apriori rules (Pediatric dataset)

Rule	Prediction result from our system	Reported precision from our system	Precision/Accuracy
1	Length of stay = 0-9	0.2417355	1
2	Length of stay = 20-29	0.5384616	0.99399
3	Length of stay = 20-29	0.2403259	0.9928

With the adult dataset, out of 10 predictions by our system, 2 of them (in italic) got poor results. Predictions done on pediatric dataset all returned very disappointing results. From the reported low precision values, they may already indicate that such predictions are not trustable. It is encouraging to see that most results from our proposed approach match the association rules found by the Apriori, when used with the adult dataset. Since our prediction engine requires a fixed number of pre case user inputs and the rules found by the Apriori can contain any number of attribute values, most of the rules still do not exactly match. After all, the Apriori algorithm finds the rules by using support and confidence, which is a totally different approach to relate attributes than ours. One thing in common between our approach and the Apriori is that post case attributes are somehow related to the pre case attributes. In a number of cases, as results shown in table 20 and 21, pre case attributes can give a rough guesswork on the post case attribute value.

6.4.2 Classification

In general, C4.5 could produce better results than the ones from Naïve Bayes. However, the percentage of correctly classified instances between the two algorithms became close when target attributes had fewer values (i.e.: fewer class labels). As long as there were more class labels for the target attribute, Naïve Bayes did very badly to classify tuples into correct class label. For instance, let us look at the "Age group". In the pediatric population, there were only two age groups, namely, 0-9 and 10-18 years old, resulting only 2 class labels. The Naïve Bayes algorithm could achieve 95.9% accuracy whereas the C4.5 could only achieve slightly better result. The same observation applied to the target attribute "Admission type". When number of class labels within the target attribute increased, the effectiveness of Naïve Bayes dropped significantly. For instance, such phenomenon happened to the target attribute "length of stay group". There were 11 length of stay groups, with each group representing a date range of 10 days. Naïve Bayes could only be able to classify 69.8% instances correctly whereas C4.5 could almost classify all the instances perfectly, with an impressive 94.19% hit rate.

By simply looking at the correctly classified vs. incorrectly classified instances, chance and class distribution were never taken into consideration. To ensure a classification was corrected to chance, Kappa statistic had been used as a chance-corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance

away from the observed agreement and dividing by the maximum possible agreement. If the value is greater than 0, it means that the classifier can do a better job than one does wild guessing. This is important as some target attributes did not have many class labels. Depending on the distribution of class labels, one could always make wild guesses to get some tuples correctly classified. If the chance of a wild guess can achieve more hit rate than the classification algorithm, there will be no meaning to compute classification systematically for such attribute. As a result, it is important to also consider the Kappa statistic to judge whether the classification should be used. Having a high hit rate (i.e.: correctly classified instances) did not always turn into high Kappa statistic. This was demonstrated in "Admission type" classification done by Naïve Bayes. Even it had an impressive 96.2% of accuracy, the Kappa statistic only showed 0.6484. This could be explained by the data distribution. Out of 9 admission types (9 class labels), only 2 of them (Clinic and Elective) were applicable to pediatric patient population. Most sitter cases (3166/3338 = 94.85%) had the admission type "Clinic" and the rest had "Elective". Such serious bias made classification less meaningful, since one could get most tuples classified correctly by putting them all to the admission type "Clinic".

C4.5 seemed to be less sensitive to class label distribution. Although some target attribute values had very bias distribution, both correctly classified instances and Kappa statistic were not hugely affected, unlike the cases in Naïve Bayes. Kappa statistic played some roles, however. For instance, although

"Admission type" had higher hit rate 99.16% than "Length of stay group" 94.19%, its Kappa statistic appeared to be lower than "Length of stay group", 0.9087 vs. 0.9264.

Overall, the C4.5 classification algorithm was able to provide better classification accuracy. Many "unreliable" classifications by Naïve Bayes became "reliable" by using the C4.5. C4.5 could always produce higher number of correctly classified instances than the Naïve Bayes. None of the results by C4.5 appeared to be worse than the Naïve Bayes. The effectiveness of C4.5 was especially prominent with increased number of class labels in target attributes.

In general, adult sitter cases classification results shared similar observations than the pediatric ones. C4.5 could achieve better classification results than the Naïve Bayes, especially with target attributes with more class labels. For instance, "Age group" in the adult population consisted of 9 groups, instead of only 2 groups in the pediatric population. This made Naïve Bayes performed much more poorly, with only 53.24% of correctly classified instances and Kappa statistic of 0.4369. However, C4.5 could still achieve very good hit rate of 97.35% and Kappa statistic of 0.9683, despite increased number of class labels. Due to more variety of class labels with adult population, Naïve Bayes did not perform as well as with pediatric population. All correctly classified instances and Kappa statistic appeared to be lower than the pediatric counterparts. If we compare the classification results with the association rules found, something

can be related. Here are the observations from the association rules discovered in section 6.4.1.

Table 22 - Observation of relationships between classification and association rules mining results

Rule	Related attributes	Naïve Bayes (Accuracy > 70% and Kappa > 0.6)	C4.5 (Accuracy > 70% and Kappa > 0.6)
1	Month, Day	No, No	Yes, No
2	Municipality	No	Yes
3	Admission type, Length of stay	Yes, No	Yes, Yes
4	Mission	Yes	Yes
5	Gender, Reason	No, No	Yes, Yes
6	Reason, Length of stay	No, No	Yes, Yes
7	Mission, Reason	Yes, No	Yes, Yes
8	Gender, Reason	No, No	Yes, Yes

9	Gender, Reason	No, No	Yes
10	Mission, Reason	Yes, No	Yes
11	Reason, Age group, Marital status	No, No, No	Yes, Yes, Yes
12	Age group, Reason	No, No	Yes, Yes

Although classification did completely different thing than the association rules mining, both of them turned out to be somehow related to each other. Since the association rules mining by Apriori algorithm was done with identified class attribute, it could partially tell whether classification would lead to any results. If association rules could not be applied with certain class attributes (i.e.: no rules found), classification on those attributes would not turn out to reliable class label prediction. For instance, both classification algorithms could not reliably classify the attribute "Day". As the association rule mining engine had difficulties to find rules with such class attribute, same phenomenon also applied to the classification rule engine.

Classification done by C4.5 could be interpreted as decision trees. However, due to numerous attributes and class labels, it became impractical to visualize extremely huge trees. It became very hard to traverse through the tree nodes. Although it is believed that the decision tree learning algorithm C4.5 and Naïve

Bayes can yield similar predictive accuracies [49], such statement did not seem to apply to our experiments. As observed in the adult sitter cases classification, Naïve Bayes could not render good class label prediction, once target attributes have more class labels. C4.5 could however achieve more consistent classification performance and less sensitive to number of class labels in attributes than the Naïve Bayes. In general, one common thing we found out from the experiments was much better effective classification from C4.5. C4.5 almost outperformed Naïve Bayes in every experiment, some even with significant difference. Many classifications that could not be done reliably using Naïve Bayes turned out to be good using C4.5.

Classification with any target attributes cannot always provide meaningful information. A good classification can lead to a meaningful predictive model that can potentially help health practitioners address their treatments to patients. For instance, if certain combinations of attribute values lead to violent patient behavior, health practitioners can take extra precautions about such case. Although classification can never replace health practitioners' professional decisions, it can at least give some extra info and hints about some cases, which need to pay special attention on.

6.4.2.1 Classification vs. proposed approach

From table 10 in the previous section, we have a summary of classification performance from our recommender system.

Table 23 – Comparison table of classification results between our approach and proven algorithms (Adult dataset)

	Recommender		Precision from proven classification algorithms	
Post case attribute value to be predicted	Attribute chosen to be used in sequences	Avg. precision	Naïve Bayes	C4.5
Length of stay	Reason	0.5677	0.5360	0.9851
	Admission type	0.6020		
	Marital status	0.6020		
	Age group	0.5775		
	Discharge location	0.5775		
Discharge location	Reason	0.6669	0.5974	0.9771
	Admission type	0.6873		

	Age group	0.6792		
	Length of stay	0.6829		
	Marital status	0.6883		

Since our system focuses on predicting post case attribute values, we are only comparing the prediction performance of the attributes “Length of stay group” and “Discharge location”. Pediatric dataset did not contain any discharge locations. Therefore, only "length of stay" can be used to do the prediction.

Table 24 - Comparison table of classification results between our approach and proven algorithms (Pediatric dataset)

	Recommender		Proven classification algorithms	
Post case attribute value to be predicted	Attribute chosen to be used in sequences	Avg precision	Naïve Bayes	C4.5
Length of stay	Reason	0.373209	0.6980	0.9419

	Admission type	0.438630		
	Marital status	0.438630		
	Age group	0.393841		

As we can clearly see, our proposed system cannot match the performance of the proven algorithm C4.5. The highest precision never exceeds 70%, whereas C4.5 can achieve over 97%, for length of stay group and discharge location attribute value predictions. However, the average precision of our system is comparable to the one by Naïve Bayes, when used on the adult dataset but not on the pediatric dataset. Same as in the association rules section, in general, use of the recommender system on pediatric dataset yields very bad prediction results. This may be explained by much smaller amount of data in pediatric cases. Although our system only uses sequence patterns of one single attribute at a time to perform class label prediction, its performance sits between Naïve Bayes and C4.5, which both take much longer time to get the results. Based on these observations, it is noted that single attributes chosen to be used in sequences (i.e.: sequence element) seemed to dominate the classification results. This phenomenon specifically applied to the sitter dataset. Among the attributes "reason", "admission type", "marital status", "age group" or "length of

stay", no matter which one we picked to predict the class label, classification yielded to results with more or less similar prediction accuracy. This means that those attributes are somehow related to each other. Each of them contained similar hints and patterns within the generated sequence.

Another interesting observation is the difference in prediction accuracies between “Discharge location” and “Length of stay group”. Both Naïve Bayes and our proposed approach could not predict the class label “Length of stay group” as good as “Discharge location”, as the prediction precisions showed. However, C4.5 was able to do good jobs (98.51% and 97.71% respectively) with both attributes.

However, the observation heavily depends on the dataset. Since our system only used one attribute at a time to perform class label prediction, if the attribute is independent, does not have much relationship with other attributes or has mostly monotonic value, use of that attribute to perform prediction will probably yield very poor results. Therefore, the choice of attribute can have tremendous effect on the prediction accuracy. Currently, the sitter dataset does not contain too much clinical information about the patient. Our proposed system does its job, based on whatever data is available from the sitter dataset. It provides a general idea about the possibility to predict post case attribute values by observing sequential patterns of a chosen attribute. Does it work? From the results of our experiments, the adequate accuracies seemed to answer “yes”. We

got the message that sequential patterns can be considered to be used in recommender system, given the sitter dataset. With future consolidated data from other clinical systems, it will be interesting to observe sequential patterns of other clinical attributes to perform post case attribute prediction. Can the proposed system be used on other clinical datasets? The answer is still unknown. Repeated tests need to be performed on different datasets to ensure the prediction reliability, before the system can be used. As mentioned, the prediction performance can heavily vary across different datasets, which contain very different types of data.

6.4.3 Clustering

With a massive amount of data presented, it is often hard to group things together in a meaningful way, without applying computational methods. It deals with finding a structure in a collection of unlabeled data. Each cluster consists of objects that are similar between themselves and dissimilar to objects of other clusters. Having said that, however, in the experiments, different clustering algorithms yielded different results. None of the results were considered to be wrong since they were generated from proven algorithms. The different results were just different representation of clusters.

A challenge to do the above clustering was to define the number of clusters in advance. Having fewer clusters might sacrifice certain fine details, but could achieve simplification. Due to the fact that the expectation-maximization

clustering algorithm can determine the number of clusters by cross validation, it has been used to perform the first experiment. Other experiments used the same number of clusters determined by the EM to facilitate the clustering result comparison.

The above experiments found out something in common. Among teenage patients who required sitter supervision, "Suicidal" (mostly female patients) and "Agitation/Violent" (mostly male patients) were the most popular reasons as they occupied significant percentages and represented in the clusters from any of the above selected algorithms. Another interesting thing was the mean ages in most centroids. They were mostly around 12. This discovered information could mean something significant to the healthcare system. Instead of treating the symptoms, should the government focus more on treating the causes of problems? There was a large cluster of female teenagers who had committed suicide. Could that be because they were bullied or forced to do something they did not want?

For adult sitter cases, it seemed like sitter reasons, mission, site and age were the dominant factors in clustering. According to the charts and relationships found between parameters, those factors were strongly related to each other to form centroids. Only one parameter could not give enough "impact" to have the data divided into meaningful clusters. Among adult patients, "disoriented" sitter cases seemed to occupy most of the centroids. From the cluster centroids, some important observations could be noted. For instance, disoriented sitter cases

seemed to happen to mostly middle aged patients whereas suicidal patients seemed to be on the younger side. Interestingly, those observations from clustering matched the ones from association rules mining.

There is no absolute "best" criterion which would be independent of the final aim of the clustering. Same as in doing the association rule analysis, no single method is said to be the "best" one. One must have familiarity with the data in order to do the judgment on the data clustering correctness. Also, one should not rely on a single algorithm to compute results. Different algorithms can yield to different discoveries and potentially cause ambiguities. Due to different evaluation methods used in algorithms, they may have preferences on some sorts of data contents. As a result, relying on only a single mining method may cause mined results to be biased. Although with the continuous research in the data mining field, more efficient and powerful algorithms will be developed to help people "mine" a lot of valuables with higher complexities from the raw data. However, human's intuition and knowledge on data still remain the most vital part to judge whether mined results make sense or not.

6.4.3.1 Clustering vs. proposed approach

Since our recommender system returns one class label at each prediction, centroids of clusters are driven by the predicted class label. The system predicts either discharge location or length of stay group, with a selected attribute as the sequence element. Among attributes to be selected as the sequence element,

the attribute "Reason" resulted several clusters that could be distinguished from the others. Other attributes resulted clusters with mostly more or less same number of records.

To compare the approach from our recommender system with the clustering results by proven algorithms, common measurement must exist. However, clustering results from the data mining tool WEKA do not provide any details about how each record got distributed to which cluster. The tool only provides centroid information of each cluster.

To be able to do the comparative analysis, we must find a way to get centroid information from the results of the recommender system. First, we need to look at the prediction result of each class label.

With “Discharge location” as the attribute to be predicted and “Reason” as the sequence element

Table 25 - Count of records for each discharge location (Adult dataset)

Predicted discharge location	Count
B	12
D	6
J	3
L	174

M	180
O	27
P	49
R	2
S	16
T	2

In order to have consistent number of clusters with the proven clustering records (i.e.: 4 clusters, as determined by EM), we only take the first 4 class labels with largest counts. Since others have relatively small counts, they are not as statistically relevant.

Table 26 - Top 4 counts of discharge location (Adult dataset)

Predicted discharge location	Meaning	Count
L	Home	174
M	Hospital	180
O	Long_term_care	27
P	Morgue	49

Although we “convert” the prediction from our recommender system into clusters based on the predicted class label, each cluster may also contain other characteristics such as high frequency counts of certain attribute values. Those characteristics are also factors to distinguish each centroid from the others.

Among the 4 predicted class labels, let us try to get statistical counts of each attribute value within each centroid.

Table 27 - Count of attribute values for each discharge location (Adult dataset)

		Discharge location as centroid			
Attribute	Value	L (Home)	M (Hospital)	O (Long term)	P (Morgue)
Mission	ER	19	44	5	8
	Medicine	60	42	14	26
	Neuro	20	19	3	6
	Surgery	75	75	5	9
Site	MGH	99	94	12	31
	RVH	75	86	15	18
Shift	Day	58	47	10	17

	Evening	57	63	10	15
	Night	59	70	7	17
Gender	Male	90	89	5	21
	Female	84	91	22	28
Admission type	Clinic	37	19	8	18
	Elective	41	13	9	3
	ER	65	28	5	19
	Stretcher	0	104	0	0
	Urgent	31	16	5	9
Marital status	Separated	52	66	7	18
	Single	66	58	11	10
	Married	56	56	9	21

Table 28 - Characteristics of each centroid determined by discharge location (Adult dataset)

Centroid	Characteristics of centroids	%
L	Mission=Surgery, Site=MGH, Shift=Night, Gender=Male,	40

	Admission type=ER, Marital status=Single	
M	Mission=Surgery, Site=MGH, Shift=Night, Gender=Female, Admission type=Stretcher, Marital status=Separated	42
O	Mission=Medicine, Site=RVH, Shift=Day/Evening, Gender=Female, Admission type=Elective, Marital status=Single	6
P	Mission=Medicine, Site=MGH, Shift=Day/Night, Gender=Female, Admission type=ER, Marital status=Married	11

With “Length of stay group” as the attribute to be predicted and “Reason” as the sequence element

Table 29 - Count of records for each length of stay group

Predicted length of stay group	Adult	Pediatric
A	175	2
B	78	0
C	53	8
D	38	0
E	24	0
F	23	2
G	5	0
H	10	0
I	8	0
J	5	0
K	52	0

Same as before, in order to have consistent number of clusters for comparison with proven clustering algorithms, we select 4 clusters with the most

record counts. Unfortunately, for pediatric cases, due to very little number of cases associated to length of stay groups, they cannot be used to do further analysis. We focus on the adult dataset instead.

Table 30 - Top 4 counts of length of stay group (Adult dataset)

Predicted length of stay group	Meaning	Count
A	0-9 days	175
B	10-19 days	78
C	20-29 days	53
K	>=100 days	52

Table 31 - Count of attribute values for each length of stay group (Adult dataset)

		Length of stay group (in days) as centroid			
Attribute	Value	A (0-9)	B (10-19)	C (20-29)	K (>=100)
Mission	ER	63	8	0	4
	Medicine	44	27	16	32
	Neuro	14	5	18	3

	Surgery	54	38	19	13
Site	MGH	94	38	28	20
	RVH	81	40	25	32
Shift	Day	53	24	17	17
	Evening	59	28	18	18
	Night	63	26	18	17
Gender	Male	91	38	35	28
	Female	84	40	18	24
Admission type	Clinic	6	15	21	14
	Elective	11	13	11	24
	ER	41	34	11	3
	Stretcher	104	0	0	0
	Urgent	13	16	10	11
Marital status	Separated	56	29	17	15
	Single	62	18	26	14
	Married	57	31	10	23

Table 32 - Characteristics of each centroid determined by length of stay group (Adult dataset)

Centroid	Characteristics of centroids	%
A	Mission=ER, Site=MGH, Shift=Night, Gender=Male, Admission type=Stretcher, Marital status=Single	49
B	Mission=Surgery, Site=RVH, Shift=Evening, Gender=Female, Admission type=ER, Marital status=Married	22
C	Mission=Surgery, Site=MGH, Shift=Evening/Night, Gender=Male, Admission type=Clinic, Marital status=Single	15
K	Mission=Medicine, Site=RVH, Shift=Evening, Gender=Male, Admission type=Elective, Marital status=Married	15

At a first glance, characteristics of centroids are very different between our proposed approach and proven clustering algorithms. Centroid size across different algorithms varied significantly. Cluster centroids from both EM and K-means had length of stay groups of 0-9 (A) and 10-19 (B) days. For discharge location, they only had Home (L) and Hospital (M). Our approach could however assign more length of stay and discharge location values to clusters. Since our proposed approach is based on supervised classification, clustering was dominated by predicted class labels, which resulted unfair clustering and bias

towards one attribute value. However, characteristics within each centroid tended to be similar across EM, K-means and our approach.

Although the recommender system recorded pre case attribute values and predicted post case attribute values, each prediction could not be treated to represent any single record. There was no way to judge if any single record had been assigned to the correct cluster or not. Pre case attribute values were used to produce a sequence of selected attribute value. Sequence similarity measure was used to perform classification and we turned each distinct classification result into a cluster. It did not evaluate difference between attributes to judge how similar the records were. This approach forcefully used the predicted results to form clusters. Unlike EM and K-means, it did not necessarily form exhaustive and mutually exclusive clusters that are locally optimal.

Our approach should not be used as the only clustering method in data clustering. It is only valuable when a prior knowledge about the data is available and when an emphasis is required to be put on post case attributes in clustering. After all, our recommender system was not originally and primarily designed to perform data clustering. There is no single "best" clustering method for all possible datasets. The appropriateness of a particular algorithm is dependent on the nature of the data. To make our recommender system useful in clustering, attempts must be made to combine the strength of proven algorithms with our approach. For example, EM or K-means can first be used to obtain gross

partitions of data. Then, within the partition, results from our recommender system may be used to obtain more smaller clusters, in which each cluster must be distinguished by discharge location or length of stay group. Results do not always come out to be meaningful. Therefore, it is important to try out different algorithms to explore the data and get meaningful clustering results through comparisons and inputs from health practitioners.

CHAPTER VII

CONCLUSION AND FUTURE WORKS

CONCLUSION

With the development of large electronic clinically related information systems, there has been an increased interest in data mining or knowledge discovery in clinical areas. Collected data usually contains value adding information. Although statistical reports with data counts are somehow useful to identify frequent cases and trends, they do not often provide real knowledge such as data correlation between attribute values. Many scenarios may likely happen with certain combination of attribute values. With only statistical reports, it is often hard to tell whether some attributes are actually correlated. This is where data mining becomes useful to help people explore "real meanings" from data.

Data mining is the process of extracting previously unknown, valid and actionable information from large information sources and databases. It can be described as business intelligence (BI) technology that has various techniques to extract comprehensible, hidden and useful information from a training dataset. By applying data mining techniques, which are combined power of elements of statistics, artificial intelligence and machine learning, they are able to provide health practitioners what those data can mean and how different attributes are related to each other. Health practitioners need the ability to easily obtain knowledge from collected data for decision-making purposes. The use of data

mining techniques in knowledge discovery in clinical databases is likely to be of increasing importance, as they are likely to be able to inform health practitioners important information (also precautions) before the clinical episode actually happens, based on historical data.

From the results of this research, we find out that each data mining method yields to very different results. While the methods are proven technologies, some are able to discover more meaningful information than the others. A thing in common is that patterns with high enough support count tend to be preferred by more algorithms. Most algorithms used are based on support, confidence and distance measures. Repeated or very similar records tend to favor those measures thus creating somehow bias in the analysis results.

While data mining is a very powerful concept, it is still far away from being self-sufficient. Up to now, there is not a data mining model or algorithm which can be applied anywhere [45]. To be successful, data mining results still rely on interpretation of skilled technical and clinical specialists. Therefore, the limitations of data mining are primarily related to input data and personnel.

FUTURE WORKS

Prior to perform most data mining activities, it is required to have data being first extracted from different corporate systems and preprocessed. Although there are middleware to perform data integration and preprocessing, they still require

heavy involvement on personnel's part. Because of that, data mining activities can only be mostly done on past refined records.

Integration of data mining into existing clinical systems still has a long way to go. Most clinical decision support systems with data mining support are standalone products, which lack interoperability with reporting and electronic clinical systems. Unless the clinical system has data mining features built in, it is found to be rather cumbersome to perform data mining activities. A generic data mining system may be built in the future, which can be attached to the database backend of the existing clinical and other corporate systems, to complement the missing data mining piece. Such concept should probably be based on data federation in order to consolidate data in real time, while leaving the actual data in its original place. It is a proposed idea, which is still under investigation.

Recommender system provides a way to support health practitioners. In situations where health practitioners are in dilemma to make certain decisions, the system can suggest them a possible outcome where previously no other help would have been available. The recommender system with our approach depends on the sequence value of one attribute at a time. Using one attribute to predict other outcome values may not always lead to satisfactory output, since there may not be any relationships between some attributes. There leaves us a lot of room for further improvement and refinement of the algorithm.

For future work, we hope to refine our technique to be able to take multiple attributes in consideration to predict outcome values. Another area of improvement can be done in the user inputs. Health practitioners may not always have all the required input information for some patients. The algorithm can be refined to give users options to specify "Don't know" in some of the input fields. Our system can be potentially extended to link with the hospital's clinical information system to relate more clinical, laboratory testing and administrative data. The algorithm can be further developed to allow health practitioners to define any available input data fields and any available outcome data fields to be predicted. This will offer way more flexibility and potentially more accuracy in the outcome prediction.

The tool can be potentially used for future clinical studies on patients who required sitter supervision. It allows users to access large transactional datasets to predict outcomes of sitter cases. Selected cases with similar outcomes can be further examined by health practitioners to discover more interesting knowledge. Although the initial implementation of the tool is focused on its application in sitter cases with the current dataset, this tool can be further expanded to include more clinically related parameters such as diagnosis, medications used, etc. With modification of the tool, it can be adapted and used in broader areas of health data analysis.

REFERENCES

[1] Gosain, A., & Kumar, A. (2009). Analysis of Health Care Data Using Different Data Mining Techniques. In *Proceedings of the International Conference on Intelligent Agent & Multi-Agent Systems, IAMA*, Chennai, India, July 22-24, 2009.

[2] Berndt, D., Hu, J., & Wei, C.P. (2000). Databases, Data Warehousing, and Data Mining in Health Care. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, Maui, Hawaii, January 4-7, 2000.

[3] Ramachandran, S., Erraguntla, M., Mayer, R., & Benjamin, P. (2007). Data Mining in Military Health Systems – Clinical and Administrative Applications. In *Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering*, Scottsdale, AZ, USA, Sept 22-25, 2007.

[4] Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

[5] Adriaans, P., & Zantinge, D. (1996). *Data Mining*. Massachusetts: Addison-Wesley.

[6] Balas, E. A., Krishna, S., Kretschmer, R. A., Cheek, T. R., Lobach, D. F., and Boren, S. A. (2004). Computerize knowledge management in diabetes care. *Journal of Medical Care, American Public Health Association, 42(6)*, 610-621.

[7] Bodenheimer, T., Lorig, K., Holman, H., and Grumbach, K. (2002). Patient self-management of chronic disease in primary care. *Journal of the American Medical Association, 288(19)*, 2469-75.

[8] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Journal of ACM communications, 39(11)*, 27-34.

[9] Lin, C.H., & Hsiao, H.S. (2009). Hierarchical State Machine Architecture for Regular Expression Pattern Matching. In *Proceedings of the 19th ACM Great Lakes symposium on VLSI*, Boston, USA, May 10-12, 2009.

[10] Jurafsky, D., and Martin J.-H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

[11] Wei, Y.Z., Moreau L., & Jennings, N.R. (2004). *Learning users' interests in a market-based recommender system*. Retrieved July 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.9461&rep=rep1&type=pdf

[12] Bjering, H, and McGregor, C. (2010). A Multidimensional Temporal Abstractive Data Mining Framework. In *Proceedings of the 4th Australasian Workshop on Health Informatics and Knowledge Management*, Brisbane, Australia, January, 2011.

[13] McAullay, D., Williams, G., Chen, J., Jin, H., He, H., Sparks, R., and Kelman, C. (2005). A Delivery Framework for Health Data Mining and Analytics. In *Proceedings of the 28th Australasian Computer Science Conference*, Newcastle, Australia, January/February, 2005.

[14] Khosla, A., Cao, Y., Lin, C., Chie, H.-K., Hu, J., and Lee, H. (2010). An Integrated Machine Learning Approach to Stroke Prediction. In *Proceedings of the 16th ACM SIGKDD conference on knowledge discovery and data mining*, Washington, USA, July 25-28, 2010.

[15] Asha, T., Natarajan, S., and Murthy, KNB. (2011). Associative Classification in the Prediction of Tuberculosis. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*, Mumbai, India, February 25–26, 2011.

[16] Witten, I., and Frank, E. (2001). *Data mining practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann.

[17] Ordonez, C., Ezquerra, N., and Santana, C.A. (2006). Constraining and summarizing association rules in medical data. *Journal of Knowledge and Information Systems, 9(3)*, 259-283.

[18] Ordonez, C., Omiecinski, E., De Braal, L., Santana, C., and Ezquerra, N. (2001). Mining constrained association rules to predict heart disease. In

Proceedings of IEEE Industrial Conference on Data Mining Conference, California, USA, November 29-December 2, 2001.

[19] Thompson, C., and Yang, H. (2009). Nurses decisions, irreducible uncertainty and maximizing nurses contribution to patient safety. *Journal of Healthcare Quarterly, 12(1)*, 178-185.

[20] Nguyen, D., Ho, T., and Kawasaki, S. (2006). Knowledge Visualization in Hepatitis Study. In *Proceedings of the Asia Pacific Symposium on Information Visualization*, Tokyo, Japan, February 1-3, 2006.

[21] Pang, N.T., Michael, S, and Vipin, K. (2006). Introduction to data mining. Massachusetts: Addison Wesley.

[22] Lee, C.-S., Wang, M.-H., Li, H.-C., and Chen, W.-H. (2008). Intelligent Ontological Agent for Diabetic Food Recommendation. In *Proceedings of the 16th IEEE International Conference on Fuzzy Systems*, Hong Kong, China, June 1-6, 2008.

[23] Hsu, W., Lee, M.L., Liu, B., and Ling, T.W. (2000). Exploration Mining in Diabetic Patients Databases: Findings and Conclusions. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, USA, August 20-23, 2000.

[24] Agrawal, R., C. Faloutsos, and A. N. Swami (1994). Efficient similarity search in sequence databases. In *Proceedings of the 4th International*

Conference of Foundations of Data Organization and Algorithms, Chicago, Illinois, USA, October 13-15, 1993.

[25] Goethals, B., and Zaki, M.J. (2003). Advances in frequent itemset mining implementations. In *Proceedings of the IEEE Workshop on Frequent Itemset Mining Implementations*, Melbourne, Florida, USA, November 19, 2003.

[26] Ma, Y., Liu, B., Wong, C. K., Yu, S., and Lee, S.M. (2000). Targeting the Right Student Using Data Mining. In *Proceedings of the 6th ACM international conference on Knowledge discovery and data mining*, Boston, Maine, USA, August 20-23, 2000.

[27] Défit, S., and Noor, S. (2001). An Economic Forecasting Based on Association Roles and Neural Network. *Journal of Information Technology, 13(1)*, 42-55.

[28] Scheffer, T. (2001). Finding association rules that trade support optimally against confidence. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, Freiburg, Germany, September 3-5, 2001.

[29] Zhang, H. (2005). The Optimality of Naïve Bayes. *Journal of Pattern* Recognition and Artificial Intelligence, 19(2), 183-198.

[30] Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: Wiley-Interscience.

[31] Landis, J.R., and Koch, G.G. (1977). A One-Way Components of Variance Model for Categorical Data. *Journal of Biometrics*, 33(4), 671-679.

[32] Gao Y., Qi, H., Liu, D.-Y., and Liu, H. (2008). In *Proceedings of the 7th International Conference on Machine Learning and Cybernetics*, Kunming, China, July 12-15, 2008.

[33] MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, USA, 1967.

[34] Seifert, J. (2004). Data Mining: An overview. Retrieved February 26, 2012, from http://www.fas.org/irp/crs/RL31798.pdf

[35] Zhao, Y., and Karypis, G. (2012). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management. CIKM '02*. ACM, New York, 515-524.

[36] Augustyniak, P. (2007). Optimal Coding of Vectorcardiographic Sequences Using Spatial Prediction. *Journal of IEEE Transactions of Information Technology in Biomedicine, 11(3)*, 305-311.

[37] Bratsas, C., Hatzizisis, I., Bamidis, P., Quaresma, P., and Maglaveras, N. (2005). Similarity Estimation among OWL Descriptions of Computational Cardiology Problems in a Knowledge Base. *Journal of IEEE Computers in Cardiology, 32(5)*, 243-246.

[38] Chen, C.-M., Hong, C.-M., Huang, C.-M., and Lee, T.-H. (2008). Web-based Remote Human Pulse Monitoring System with Intelligent Data Analysis for Home Healthcare. In *Proceedings of Cybernetics and Intelligent Systems*. CIS '08. IEEE, 636-641.

[39] Subhashini, R., and Jawahar, Senthil Kumar, V. (2010). Evaluating the Performance of Similarity Measures Used in Document Clustering and Information Retrieval. In *Proceedings of the 1st International Conference on Integrated Intelligent Computing*, Bangalore, India, 27-31.

[40] Garofalakis, M., Rastogi, R., and Shim, K. (2002). Mining Sequential Patterns with Regular Expression Constraints. *Journal of IEEE transactions on knowledge and data engineering, 14(3)*, 530-552.

[41] Grishman, R. (1997). Information Extraction: *Techniques and Challenges*. SCIE: 10-27.

[42] Mutalik, P.-G., Deshpande, A., and Nadkarni, P.-M. (2001). Use of general-purpose negation detection to augment concept indexing of medical

documents. *Journal of the American Medical Informatics Association, 8(6)*, 598-609.

[43] Chapman, W.-W., Bridewell, W., Hanbury, P., Cooper, G.-F., and Buchanan, B.-G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of PubMed Biomed Inform, 34(5)*, 301-310.

[44] Johnson, J. (2003). *Probability and Statistics for Computer Science* (1st ed.). Washington: Wiley-Interscience.

[45] Chaudhuri, S., Dayal, U., and Ganti, V. (2001). Database Technology for Decision Support Systems. Journal of Computer, 34(12), 48-55.

[46] Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Journal of Machine Learning, 40(3), 203-228.

[47] Quinlan, J.R. (1986). Induction of decision trees. *Journal of Machine Learning, 1(1)*, 81-106.

[48] Ruggieri S. (2002). Efficient C4.5. *Journal of IEEE Transactions On Knowledge And Data Engineering, 14(2)*, 438-444.

[49] Huang, J., Lu, J., and Ling, X. (2003). Comparing Naïve Bayes, Decision Trees, and SVM with AUC and Accuracy. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, Florida, USA, 553-556.

[50] Jia, Z., Li, H., Dong, L. and Long, D. (2011). Temporal Expression Recognition and Temporal Relationship Extraction from Chinese Narrative Medical Records. In *Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering*, Wuhan, China, 1-4.

[51] Bhatia, R., Graystone, A., Davies, R., McClinton, S., Morin, J., and Davies, R. (2010). Extracting information for generating a diabetes report card from free text in physicians notes. In *Proceedings of the NAACL HLT 2010 2nd Louhi Workshop on Text and Data Mining of Health Documents*, Los Angeles, USA, 8-14.

[52] Chapman, W., Chu, D., and Dowling, J. (2007). An Algorithm for Identifying Contextual Features from Clinical Text. In *Proceedings of Biological, translational, and clinical language processing*, Prague, Czech Republic, 81–88.

[53] Boudin, F., Nie, J., and Dawes, M. (2010). Positional Language Models for Clinical Information Retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Massachusetts, USA, 108-115.

[54] Konrad, R., and Lawley, M. (2009). Input modeling for hospital simulation models using electronic messages. In *Proceedings of the Winter Simulation Conference*, Maryland, USA, 134-147.

[55] Zhu, W., Fu, L., Xu, L., and Zhang, B. (2011). A TCM Diagnosis System Based on Textbook Information Extraction. In *Proceedings of the 4th IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing*, Dalian, China, 483-487.

[56] McKenzie, A., Matthews, M., Goodman, N., and Bayoumi, A. (2010). Information extraction from helicopter maintenance records as a springboard for the future of maintenance text analysis. In *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems*, Cordoba, Spain, 590-600.

[57] Qu, G., and Wu, H. (2009). Bucket Learning: Improving Model Quality through Enhancing Local Patterns. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, Florida, USA, 539-544.

[58] Hammouda, K., and Kamel, M. (2004). Efficient Phrase-Based Document Indexing for web document clustering. *Journal of IEEE Transactions on knowledge and data engineering, 16(10)*, 1279-1296.

[59] Gao, K., and Khoshgoftaar, T., and Napolitano, A. (2009). Exploring software quality classification with a wrapper-based feature ranking technique. In *Proceedings of the 21st IEEE international conference on tools with artificial intelligence*, New Jersey, USA, 67-74.

[60] Mooney, R., and Nahm, U. (2005). Text Mining with Information Extraction. In *Proceedings of the 4th International MIDP Colloquium*, Bloemfontein, South Africa, 141-160.

APPENDIX A

ASSOCIATION RULE MINING RESULTS USING CHILD POPULATION DATASET

WITH CLASS ATTRIBUTE = “DAY”

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 60%

2008, 2009, 2010

No association rules could be found.

Using Predictive Apriori algorithm

2008, 2009, 2010

All the rules found had extremely low support values. Most rules only had support values of 2 or 3.

WITH CLASS ATTRIBUTE = “COST CENTER”

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 80%

2008

1. Month=12 Reason=Suicidal 105 ==> Cost center=32806 104 conf:(0.99)

2. Month=12 Reason=Suicidal Age Group=10-19 105 ==> Cost center=32806 104 conf:(0.99)

3. Month=12 Reason=Suicidal Admission type=Clinic 105 ==> Cost center=32806 104 conf:(0.99)

4. Month=12 Reason=Suicidal Age Group=10-19 Admission type=Clinic 105 ==> Cost center=32806 104 conf:(0.99)

26. Reason=Suicidal Age Group=10-19 Gender=F Language=French Admission type=Clinic 151 ==> Cost center=32806 137 conf:(0.91)

99. Reason=Other Age Group=0-9 Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=80-89 119 ==> Cost center=32801 107 conf:(0.9)

Most of the rules found had resulting attribute value of 32806 as the cost center (ward). Similar to the mining done earlier, they were all related to French speaking girls in the age group of 10-19 who had committed suicide tended to stay in the cost center 32806, between 80 and 89 days. A lot of rules found were similar to the ones by Apriori with class attribute "Month". For instance, the rule 99 was very similar to the rule 98 with class attribute "Month", except the order of attribute values became different.

2009

7. Age Group=0-9 Length of stay group=30-39 67 ==> Cost center=32806 67 acc:(0.99142)

13. Reason=Away without leave 55 ==> Cost center=32806 55 acc:(0.98994)

24. Month=9 Reason=Agitation Gender=F 38 ==> Cost center=32806 38 acc:(0.98568)

27. Reason=Eating disorder Length of stay group=10-19 33 ==> Cost center=32806 33 acc:(0.98338)

31. Reason=Disorientation Municipality=SALLUIT 29 ==> Cost center=32801 29 acc:(0.98087)

32. Reason=Disorientation Length of stay group=40-49 29 ==> Cost center=32801 29 acc:(0.98087)

40. Month=2 Reason=Suicidal Language=French 25 ==> Cost center=32806 25 acc:(0.97742)

Same as in 2008, most of the rules discovered had "32806" as the resulting attribute value. It was expected to have such result since "32806" was the most frequent cost center (650/988 = 65.79%) in the dataset. Interestingly, patients with "Disorientation" tended to be in the unit 32801 instead, as indicated in rules 31 and 32.

2010

32. Reason=Other Age Group=10-19 Gender=M Language=English Municipality=SAINT-JACQUES-LE-MINEUR Admission type=Clinic 108 ==> Cost center=32806 107 conf:(0.99)

66. Age Group=10-19 Language=French Admission type=Elective 132 ==> Cost center=32801 114 conf:(0.86)

Same as in the other two previous years, most of the rules discovered had "32806" as the resulting attribute value. There were many subsets of rule 32 in the result, with same attribute values but with different combinations. We discovered quite several sitter cases with patients with "Elective" admissions in unit 32801. It never happened in the other two previous years.

Using Predictive Apriori algorithm

2008

1. Month=9 Length of stay group=80-89 88 ==> Cost center=32801 88 acc:(0.9929)

2. Month=9 Reason=Other Age Group=0-9 88 ==> Cost center=32801 88 acc:(0.9929)

3. Month=9 Age Group=0-9 Gender=F 88 ==> Cost center=32801 88 acc:(0.9929)

4. Month=9 Age Group=0-9 Municipality=MONTREAL 88 ==> Cost center=32801 88 acc:(0.9929)

5. Month=12 Reason=Suicidal Language=French 68 ==> Cost center=32806 68 acc:(0.99162)

6. Month=12 Municipality=SAINT-LAURENT 54 ==> Cost center=32806 54 acc:(0.98993)

7. Reason=Suicidal Municipality=SAINT-LAURENT 54 ==> Cost center=32806 54 acc:(0.98993)

8. Month=12 Length of stay group=20-29 53 ==> Cost center=32806 53 acc:(0.98976)

9. Municipality=SAINT-LAURENT Length of stay group=20-29 53 ==> Cost center=32806 53 acc:(0.98976)

10. Gender=F Language=French Length of stay group=20-29 53 ==> Cost center=32806 53 acc:(0.98976)

The above rules gave more or less the same message than the ones with class attribute "Month".

2009

32. Reason=Disorientation Length of stay group=40-49 29 ==> Cost center=32801 29 acc:(0.98087)

43. Age Group=0-9 Language=English Admission type=Clinic Length of stay group=0-9 23 ==> Cost center=32806 23 acc:(0.97519)

2010

4. Reason=Suicidal Length of stay group=20-29 64 ==> Cost center=32806 64 acc:(0.98892)

5. Age Group=0-9 Municipality=SAINT-LAZARE 64 ==> Cost center=32806 64 acc:(0.98892)

17. Reason=Trauma 50 ==> Cost center=32801 50 acc:(0.98596)

25. Month=12 Reason=Suicidal Gender=F 41 ==> Cost center=32806 41 acc:(0.98279)

54. Month=9 Reason=Agitation 24 ==> Cost center=32801 24 acc:(0.9693)

The rules were very different than the ones by the original Apriori algorithm. However, they were all with very low support values and a lot of them did not seem to be very meaningful.

<u>WITH CLASS ATTRIBUTE = “SHIFT”</u>

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 60%

2008, 2009, 2010

No association rules could be found.

Using Predictive Apriori algorithm

2008, 2009, 2010

All the rules found had extremely low support values. Most rules only had support values of 2 or 3. That means, shift might not contribute to be a factor to have sitter cases.

WITH CLASS ATTRIBUTE = “REASON”

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 80%

2008

32. Age Group=0-9 Gender=M Language=Other Municipality=SAINT-LAZARE Admission type=Clinic 123 ==> Reason=Other 123 conf:(1)

66. Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=80-89 119 ==> Reason=Other 119 conf:(1)

All the rules found had "Other" as the resulting attribute value. All the rules had very high confidence of 95% and up. Most of the rules provided similar information than the ones with class attribute "Month". They were mostly subsets of rule 66. The only new information found this time was rule 32, showing that male in the age group of 0-9, with neither English nor French as mother tongue,

living in the city "Saint-Lazare" were likely to have "Other" as the reason to get sitter service.

2009

32. Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=70-79 119 ==> Reason=Agitation 119 conf:(1)

48. Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=40-49 118 ==> Reason=Other 118 conf:(1)

Rules discovered from 2009 data were very different. Rules with top support and confidence values had "Agitation" as the resulting attribute. Such rules were basically variants of the rule 32, with different combination of attributes, with same values. The remaining rules all had "Other" as the resulting attribute.

2010

15. Age Group=10-19 Gender=M Language=French Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 118 ==> Reason=Agitation 118 conf:(1)

19. Cost center=32806 Age Group=10-19 Language=English Admission type=Clinic Length of stay group=20-29 112 ==> Reason=Other 112 conf:(1)

20. Municipality=SAINT-JACQUES-LE-MINEUR 108 ==> Reason=Other 108 conf:(1)

92. Cost center=32801 Age Group=10-19 Language=French Length of stay group=20-29 94 ==> Reason=Agitation 83 conf:(0.88)

As the resulting attribute value "Other" occupied almost half of the reasons (480/988 = 48.58%) for hiring sitters, it was very likely to appear in the discovered rules and actually, it was the case.

Using Predictive Apriori algorithm

2008

2. Month=9 Cost center=32801 Municipality=MONTREAL 151 ==> Reason=Other 151 acc:(0.99453)

4. Age Group=0-9 Municipality=SAINT-LAZARE 123 ==> Reason=Other 123 acc:(0.99426)

13. Language=French Length of stay group=20-29 90 ==> Reason=Suicidal 90 acc:(0.9936)

Same as previous results using different class attributes, this mining activity did not give additional information.

2009

1. Length of stay group=70-79 119 ==> Reason=Agitation 119 acc:(0.99468)

31. Cost center=32806 Gender=F Language=French Length of stay group=10-19 33 ==> Reason=Eating disorder 33 acc:(0.99111)

40. Cost center=32806 Age Group=10-19 Gender=F Municipality=MONTREAL Length of stay group=0-9 29 ==> Reason=Suicidal 29 acc:(0.9901)

50. Month=8 Cost center=32801 Age Group=10-19 Gender=M Language=English Admission type=Clinic 25 ==> Reason=Disorientation 25 acc:(0.98865)

In general, rules discovered had quite low support. The algorithm was able to find more resulting attribute values than the original Apriori.

2010

4. Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 118 ==> Reason=Agitation 118 acc:(0.99406)

5. Municipality=SAINT-JACQUES-LE-MINEUR 108 ==> Reason=Other 108 acc:(0.99387)

22. Month=10 Admission type=Elective Length of stay group=10-19 49 ==> Reason=Trauma 49 acc:(0.98999)

62. Language=French Municipality=BROSSARD 25 ==> Reason=Suicidal 25 acc:(0.97978)

85. Month=7 Length of stay group=10-19 17 ==> Reason=Psychosis 17 acc:(0.96825)

86. Month=10 Municipality=LASALLE 17 ==> Reason=Behavior problem 17 acc:(0.9682)

Rules with more varieties of resulting attribute values could be found. Many rules contained only very few attributes, unlike the ones discovered by the original Apriori. A lot of rules with low support values could be spotted.

WITH CLASS ATTRIBUTE = “AGE GROUP”

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 80%

2008

24. Cost center=32806 Reason=Suicidal Gender=F Admission type=Clinic Length of stay group=0-9 157 ==> Age Group=10-19 157 conf:(1)

47. Reason=Psychosis Admission type=Clinic 127 ==> Age Group=10-19 127 conf:(1)

70. Reason=Other Gender=M Language=Other Municipality=SAINT-LAZARE Admission type=Clinic 123 ==> Age Group=0-9 123 conf:(1)

All rules had very high confidence values. In every suicidal case, age group was always 10-19 in the ward "32806". Similarly, in rule 47, it indicated that psychosis patients from clinics were always in the age group 10-19. These two reasons became problems for teenagers.

2009

6. Reason=Suicidal Gender=F Admission type=Clinic Length of stay group=0-9 162 ==> Age Group=10-19 162 conf:(1)

56. Reason=Agitation Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=70-79 119 ==> Age Group=0-9 119 conf:(1)

72. Reason=Other Gender=F Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=40-49 118 ==> Age Group=0-9 118 conf:(1)

The discovered rules showed some important information. Suicidal patients tended to be female teenagers and had relatively short length of stay, 0 to 9 days. Agitated patients tended to have younger age and had longer length of stay, 70 to 79 days.

2010

1. Length of stay group=10-19 242 ==> Age Group=10-19 242 conf:(1)

3. Reason=Suicidal 194 ==> Age Group=10-19 194 conf:(1)

10. Reason=Agitation Language=French 152 ==> Age Group=10-19 152 conf:(1)

11. Reason=Suicidal Gender=F 149 ==> Age Group=10-19 149 conf:(1)

17. Gender=M Admission type=Elective 132 ==> Age Group=10-19 132 conf:(1)

32. Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 118 ==> Age Group=10-19 118 conf:(1)

All rules had age group equal to 10-19 as the resulting attribute. This could be explained by the serious bias in age group distribution in the dataset. As majority of patients were between 10 and 19 years old (845/988 = 85.53%), most cases fell into that age group.

Using Predictive Apriori algorithm

2008

1. Cost center=32806 Reason=Suicidal 268 ==> Age Group=10-19 268 acc:(0.99489)

2. Reason=Suicidal Length of stay group=0-9 259 ==> Age Group=10-19 259 acc:(0.99487)

3. Reason=Suicidal Language=English 206 ==> Age Group=10-19 206 acc:(0.99475)

4. Cost center=32806 Gender=F Language=French 198 ==> Age Group=10-19 198 acc:(0.99472)

6. Gender=F Length of stay group=20-29 156 ==> Age Group=10-19 156 acc:(0.99449)

8. Reason=Suicidal Gender=F Language=French 151 ==> Age Group=10-19 151 acc:(0.99445)

9. Shift=Evening Reason=Suicidal 142 ==> Age Group=10-19 142 acc:(0.99436)

18. Gender=F Length of stay group=80-89 119 ==> Age Group=0-9 119 acc:(0.99405)

25. Month=12 Reason=Suicidal 105 ==> Age Group=10-19 105 acc:(0.99377)

61. Month=9 Reason=Suicidal 46 ==> Age Group=10-19 46 acc:(0.98954)

The above rules again showed that French speaking teenage patients who had "suicidal" problem tended to be female. September and December seemed to be problematic months for female teenage patients. Could that be pressure from "back to school" and "final exam" periods?

2009

1. Reason=Suicidal Gender=F 222 ==> Age Group=10-19 222 acc:(0.99493)

3. Length of stay group=10-19 130 ==> Age Group=10-19 130 acc:(0.99476)

6. Length of stay group=70-79 119 ==> Age Group=0-9 119 acc:(0.9947)

7. Reason=Agitation Gender=F 119 ==> Age Group=0-9 119 acc:(0.9947)

17. Reason=Agitation Gender=M 101 ==> Age Group=10-19 101 acc:(0.99458)

74. Reason=Behavior problem Language=English 36 ==> Age Group=10-19 36 acc:(0.99185)

Rules were similar to the ones discovered by the original Apriori. In general, younger patients (age between 0 and 9) were more likely to stay in the hospital longer than the teenagers. Also, their reasons for sitter cases were different.

2010

1. Length of stay group=10-19 242 ==> Age Group=10-19 242 acc:(0.99483)

2. Reason=Suicidal 194 ==> Age Group=10-19 194 acc:(0.99469)

10. Reason=Agitation Gender=M 122 ==> Age Group=10-19 122 acc:(0.99411)

17. Reason=Agitation Length of stay group=20-29 96 ==> Age Group=10-19 96 acc:(0.99355)

52. Month=8 Reason=Other Gender=F Admission type=Clinic Length of stay group=20-29 44 ==> Age Group=0-9 44 acc:(0.98892)

Not many rules could be discovered for the age group "0-9". Rules with higher probabilities and support were mostly similar to the ones discovered by the original Apriori.

<u>WITH CLASS ATTRIBUTE = “GENDER”</u>

Using Apriori algorithm

minimum support threshold = 5% and minimum confidence threshold = 80%

2008

34. Reason=Other Age Group=0-9 Language=Other Municipality=SAINT-LAZARE Admission type=Clinic 123 ==> Gender=M 123 conf:(1)

68. Reason=Other Age Group=0-9 Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=80-89 119 ==> Gender=F 119 conf:(1)

Most rules had "Female" as the resulting attribute value. The ones with "Male" were sub-rules of rule 34 above. Interestingly, only female patients had length of stay attribute in the rules and the attribute value was always "80-89".

2009

5. Length of stay group=30-39 146 ==> Gender=M 146 conf:(1)

54. Reason=Agitation Age Group=0-9 Language=English Municipality=MONTREAL Admission type=Clinic Length of stay group=70-79 119 ==> Gender=F 119 conf:(1)

Most rules were discovered with the resulting attribute value "Female". There were a lot of rules that were subsets of the above rule 54, with same attribute values but different combinations.

2010

46. Cost center=32806 Reason=Other Age Group=10-19 Language=English Municipality=SAINT-JACQUES-LE-MINEUR Admission type=Clinic 107 ==> Gender=M 107 conf:(1)

100. Reason=Suicidal Language=French 127 ==> Gender=F 103 conf:(0.81)

Due to most patients were male, most discovered rules had male as the resulting attribute value. In general, female teenagers were more prone to have suicidal problem and male teenagers were more prone to have agitation and trauma problems.

Using Predictive Apriori algorithm

2008

26. Month=10 Reason=Suicidal Language=English 61 ==> Gender=F 61 acc:(0.99246)

36. Month=9 Reason=Suicidal 46 ==> Gender=F 46 acc:(0.99076)

52. Reason=Suicidal Age Group=10-19 Language=English Length of stay group=20-29 35 ==> Gender=F 35 acc:(0.98818)

72. Reason=Agitation Length of stay group=10-19 23 ==> Gender=M 23 acc:(0.98145)

96. Shift=Night Reason=Agitation Age Group=10-19 Admission type=Clinic 17 ==> Gender=M 17 acc:(0.97348)

Female teenagers tended to be more likely to commit suicide in September and October. The reason "Agitation" seemed to only apply to male teenagers.

2010

1. Language=French Admission type=Elective 132 ==> Gender=M 132 acc:(0.99482)

4. Cost center=32806 Language=French Admission type=Clinic Length of stay group=20-29 110 ==> Gender=F 110 acc:(0.99473)

7. Age Group=0-9 Admission type=Clinic Length of stay group=0-9 93 ==> Gender=M 93 acc:(0.99461)

19. Reason=Suicidal Length of stay group=20-29 64 ==> Gender=F 64 acc:(0.99418)

38. Reason=Trauma 50 ==> Gender=M 50 acc:(0.99367)

43. Cost center=32806 Reason=Suicidal Language=French Admission type=Clinic 98 ==> Gender=F 97 acc:(0.99325)

44. Reason=Behavior problem Language=French 41 ==> Gender=M 41 acc:(0.99306)

Not much information could be gained from the above, as they provided more or less similar information than the ones offered by the original Apriori.

WITH CLASS ATTRIBUTE = “LANGUAGE”

Using Apriori algorithm

2008

44. Reason=Other Age Group=0-9 Gender=M Municipality=SAINT-LAZARE Admission type=Clinic 123 ==> Language=Other 123 conf:(1)

Due to the fact that all the hospitals were Anglophone, patient population was seriously biased towards English speaking patients. Rules that had resulting attribute value equal to English were not as interesting as the ones with other languages. All the rules discovered did not have "French" as the resulting attribute value. All the rules with resulting attribute value as "Other" were subsets of the above rule 44.

2009

All rules had "English" as the resulting attributes. As 62.38% (713/1143) patients were English speaking, the discovered rules did not give any interestingness.

2010

13. Reason=Agitation Age Group=10-19 Gender=M Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 118 ==> Language=French 118 conf:(1)

Although English was the language for most patients, rules with top support and confidence values had "French" as the resulting attribute value. There were a lot of variants of the above rule 13, with different combinations of attributes.

Using Predictive Apriori algorithm

2008

7. Age Group=0-9 Municipality=SAINT-LAZARE 123 ==> Language=Other 123 acc:(0.99485)

20. Municipality=SAINT-LAURENT Length of stay group=20-29 53 ==> Language=French 53 acc:(0.99416)

The predictive Apriori algorithm discovered similar rule than the one by the original Apriori algorithm, except it discovered a rule with "French" as the resulting attribute value. Other rules found had very low support values and contained mostly obvious information with only high frequency attribute values.

2009

10. Gender=M Municipality=SAINT-LAZARE Admission type=Clinic 110 ==> Language=Other 110 acc:(0.99401)

35. Reason=Suicidal Length of stay group=20-29 47 ==> Language=French 47 acc:(0.9897)

The predictive Apriori algorithm discovered some rules with "French" and "Other" as the resulting attribute values.

2010

5. Cost center=32801 Age Group=10-19 Admission type=Elective 114 ==> Language=French 114 acc:(0.99425)

6. Cost center=32806 Reason=Other Age Group=10-19 Length of stay group=20-29 112 ==> Language=English 112 acc:(0.99422)

20. Age Group=0-9 Municipality=SAINT-LAZARE 64 ==> Language=Other 64 acc:(0.99264)

31. Reason=Trauma 50 ==> Language=French 50 acc:(0.99129)

Rules discovered were not very meaningful. Language might not be a factor that had led to sitter cases.

WITH CLASS ATTRIBUTE = “MUNICIPALITY”

Using Apriori algorithm

2008

24. Reason=Other Age Group=0-9 Gender=M Language=Other Admission type=Clinic 123 ==> Municipality=SAINT-LAZARE 123 conf:(1)

Since the hospital resides in Montreal, it was expected to have mostly patients from Montreal. Having such highly expected resulting attribute value might not be as interesting as the others. In this finding, only two resulting attribute values were presented - Montreal and SAINT-LAZARE. Several rules found with "SAINT-LAZARE" as the resulting attribute value were basically subsets of the above rule 24.

2009

100. Cost center=32806 Age Group=0-9 Gender=M Language=Other Admission type=Clinic 110 ==> Municipality=SAINT-LAZARE 110 conf:(1)

As in the previous year 2008, only two resulting attribute values were discovered - Montreal and SAINT-LAZARE. Rules found with "SAINT-LAZARE" as the resulting attribute value were basically subsets of the above rule 100.

2010

35. Reason=Agitation Age Group=10-19 Gender=M Language=French Admission type=Elective Length of stay group=20-29 66 ==> Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 66 conf:(1)

99. Cost center=32806 Reason=Other Age Group=0-9 Gender=M Language=Other Admission type=Clinic Length of stay group=0-9 64 ==> Municipality=SAINT-LAZARE 64 conf:(1)

Although "Montreal" was the municipality where most patients resided, none of the rules had it as the resulting attribute value. There were many rules which related "Agitation" with the municipality "SAINTE-CLOTILDE-DE-CHATEAUGUAY".

Using Predictive Apriori algorithm

2008

15. Cost center=32806 Reason=Other Length of stay group=20-29 73 ==> Municipality=SAINT-LAZARE 73 acc:(0.98613)

18. Gender=F Language=French Length of stay group=20-29 53 ==> Municipality=SAINT-LAURENT 53 acc:(0.98145)

32. Reason=Suicidal Gender=M Length of stay group=20-29 37 ==> Municipality=CARIGNAN 37 acc:(0.97414)

33. Gender=M Language=French Length of stay group=20-29 37 ==> Municipality=CARIGNAN 37 acc:(0.97414)

34. Reason=Other Age Group=0-9 Gender=M Language=English 37 ==> Municipality=IVUJIVIK 37 acc:(0.97414)

78. Cost center=32801 Reason=Agitation Age Group=10-19 Gender=M Language=French 21 ==> Municipality=SAINT-HUBERT 21 acc:(0.95647)

2009

14. Age Group=0-9 Length of stay group=30-39 67 ==> Municipality=SAINT-LAZARE 67 acc:(0.9904)

22. Gender=M Length of stay group=40-49 53 ==> Municipality=SALLUIT 53 acc:(0.98819)

30. Reason=Agitation Length of stay group=30-39 39 ==> Municipality=CANDIAC 39 acc:(0.984)

35. Reason=Away without leave Gender=M Language=English 37 ==> Municipality=PIERREFONDS 37 acc:(0.98309)

54. Reason=Agitation Language=French Admission type=Clinic 27 ==> Municipality=LONGUEUIL 27 acc:(0.97627)

In addition to similar rules found by the original Apriori algorithm, the predictive Apriori algorithm could discover more resulting municipalities in the rules. However, none of them had high support values.

2010

1. Month=11 Reason=Agitation 75 ==> Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 75 acc:(0.98652)

3. Reason=Agitation Age Group=10-19 Admission type=Elective 66 ==> Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY 66 acc:(0.98488)

38. Reason=Agitation Age Group=10-19 Gender=F 30 ==>
Municipality=MANAWAN 30 acc:(0.96869)

86. Reason=Psychosis Gender=M Language=English Length of stay group=10-19 17 ==> Municipality=KIRKLAND 17 acc:(0.94743)

Some reasons for sitter cases seemed to be related to certain municipalities. However, most rules had low support values. Those might be just specific cases.

WITH CLASS ATTRIBUTE = “ADMISSION TYPE”

Using Apriori algorithm

2008, 2009, 2010

Since almost all sitter cases had "Clinic" as the "Admission type" attribute value, it was expected to have "Clinic" in most resulting attribute values. All the rules discovered by the Apriori algorithm only had "Clinic" as the resulting attribute value. No specific rules presented any interesting hidden facts.

Using Predictive Apriori algorithm

2008, 2009, 2010

All the rules discovered had "Clinic" as the resulting attribute value. No interesting rules could be discovered.

WITH CLASS ATTRIBUTE = “LENGTH OF STAY”

Using Apriori algorithm

2008

38. Cost center=32801 Reason=Other Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 107 ==> Length of stay group=80-89 107 conf:(1)

98. Month=9 Cost center=32801 Reason=Other Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 88 ==> Length of stay group=80-89 88 conf:(1)

Interestingly, although the attribute value "80-89" did not appear as frequent as the other values, all the discovered rules had this value. Also, all the rules had very high confidence with a minimum of 99%. Rule 98 contained all the attributes appeared in other rules. All other rules were just subsets of this rule 98.

2009

24. Reason=Agitation Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 119 ==> Length of stay group=70-79 119 conf:(1)

56. Reason=Other Age Group=10-19 Gender=M Language=English Municipality=ROSEMERE Admission type=Clinic 93 ==> Length of stay group=60-69 93 conf:(1)

72. Cost center=32806 Reason=Other Age Group=0-9 Gender=F Language=English Municipality=MONTREAL Admission type=Clinic 103 ==> Length of stay group=40-49 99 conf:(0.96)

This time, 3 different attribute values could be found. Lengths of stay varied from 40 to 79 in the discovered rules. Although "0-9" was the most frequent length of stay attribute value in the dataset, it never appeared in any of the discovered rules. All the remaining discovered rules were subsets of the above rules 24, 56 and 72 with same attribute values but with different combination.

2010

1. Age Group=0-9 Gender=M 93 ==> Length of stay group=0-9 93 conf:(1)

13. Cost center=32801 Reason=Agitation Age Group=10-19 83 ==> Length of stay group=20-29

38. Reason=Agitation Age Group=10-19 Gender=M Language=French Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY Admission type=Elective 66 ==> Length of stay group=20-29 66 conf:(1)

83 conf:(1) 38. Reason=Agitation Age Group=10-19 Gender=M Language=French Municipality=SAINTE-CLOTILDE-DE-CHATEAUGUAY Admission type=Elective 66 ==> Length of stay group=20-29 66 conf:(1)

Using Predictive Apriori algorithm

2008

1. Reason=Other Age Group=0-9 Municipality=MONTREAL 119 ==> Length of stay group=80-89 119 acc:(0.99461)

11. Cost center=32806 Municipality=SAINT-LAZARE 73 ==> Length of stay group=20-29 73 acc:(0.99394)

12. Month=10 Cost center=32806 Gender=F 61 ==> Length of stay group=0-9 61 acc:(0.9935)

20. Cost center=32801 Language=Other 45 ==> Length of stay group=10-19 45 acc:(0.99234)

21. Cost center=32801 Municipality=SAINT-LAZARE 45 ==> Length of stay group=10-19 45 acc:(0.99234)

The algorithm found similar rules than the above ones with different class attributes. Not much information was gained.

2009

1. Reason=Agitation Age Group=0-9 119 ==> Length of stay group=70-79 119 acc:(0.99438)

3. Municipality=ROSEMERE 93 ==> Length of stay group=60-69 93 acc:(0.99397)

11. Month=5 Age Group=0-9 Gender=F 60 ==> Length of stay group=40-49 60 acc:(0.9926)

16. Month=6 Age Group=0-9 Gender=F 58 ==> Length of stay group=40-49 58 acc:(0.99244)

43. Month=5 Reason=Agitation 27 ==> Length of stay group=10-19 27 acc:(0.98528)

66. Month=4 Age Group=0-9 21 ==> Length of stay group=30-39 21 acc:(0.98048)

79. Month=1 Reason=Eating disorder Language=French 20 ==> Length of stay group=10-19 20 acc:(0.97934)

The algorithm returned rules with various lengths of stay. Most of the rules did not provide any meaningful information and had very low support values. Some rules provided important observation about the relationship between "Age group", "Reason" and "Length of stay" as the selected ones above.

2010

1. Age Group=0-9 Gender=M 93 ==> Length of stay group=0-9 93 acc:(0.99449)

2. Municipality=SAINT-LAZARE 91 ==> Length of stay group=0-9 91 acc:(0.99447)

6. Reason=Agitation Age Group=10-19 Admission type=Elective 66 ==> Length of stay group=20-29 66 acc:(0.99399)

40. Reason=Agitation Gender=F Language=French Admission type=Clinic 30 ==> Length of stay group=20-29 30 acc:(0.99057)

77. Reason=Behavior problem Municipality=LASALLE 17 ==> Length of stay group=0-9 17 acc:(0.98327)

78. Reason=Psychosis Municipality=KIRKLAND 17 ==> Length of stay group=10-19 17 acc:(0.98327)

It seemed that age groups and sitter reasons could be somehow related to lengths of stay. Patients in age group 0-9 seemed to have shorter length of stay than the ones in age group 10-19 in general.

ASSOCIATION RULE MINING RESULTS USING ADULT POPULATION DATASET

WITH CLASS ATTRIBUTE = “MONTH”

minimum support threshold = 5% and minimum confidence threshold = 80%

Using Apriori algorithm

2008, 2009, 2010

No association rules could be found.

Using Predictive Apriori algorithm

2008, 2009, 2010

No association rules could be found.

WITH CLASS ATTRIBUTE = “DAY”

Using Apriori algorithm

2008, 2009, 2010

No association rules could be found.

Using Predictive Apriori algorithm

2008, 2009, 2010

No association rules could be found.

WITH CLASS ATTRIBUTE = “MISSION”

Using Apriori algorithm

2008

1. Site=MGH Reason=Suicidal Gender=F Marital status=SINGLE_ADULT Admission type=ER 1191 ==> Mission=Surgery 1139 conf:(0.96)

6. Site=RVH Discharge location=Long term care 1489 ==> Mission=Medicine 1355 conf:(0.91)

17. Reason=Suicidal Marital status=SINGLE_ADULT Length of stay group=>=100 1248 ==> Mission=Surgery 1114 conf:(0.89)

18. Site=MGH Reason=Suicidal Gender=F 1319 ==> Mission=Surgery 1177 conf:(0.89)

24. Site=RVH Reason=Disorientation Admission type=Stretcher 1330 ==> Mission=Emergency 1175 conf:(0.88)

60. Site=MGH Reason=Suicidal 2452 ==> Mission=Surgery 2033 conf:(0.83)

67. Reason=Suicidal Gender=F 1503 ==> Mission=Surgery 1222 conf:(0.81)

Most discovered rules had the mission "Surgery" as the resulting attribute value. A lot of rules with "suicidal" as the sitter reason were found. It seemed like "suicidal" cases happened often at the "MGH" site in units under surgical mission. The reason "Disorientation" only happened in missions "Emergency" and "Medicine".

2009

14. Site=RVH Reason=Agitation Age Group=80-89 Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Discharge location=Home 1142 ==> Mission=Medicine 1133 conf:(0.99)

54. Site=RVH Reason=Agitation Gender=M Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Discharge location=Home 1163 ==> Mission=Medicine 1138 conf:(0.98)

88. Site=RVH Reason=Agitation Age Group=80-89 Gender=M Marital status=MARRIED_ADULT Municipality=MONTREAL Admission type=Elective 1149 ==> Mission=Medicine 1113 conf:(0.97)

Although "Surgery" was really the most frequently appeared attribute value (9073/22229 = 40.82%) in the dataset, all the discovered rules had the mission "Medicine" as the resulting attribute value. Values of other attributes were all the same across rules.

2010

6. Site=MGH Reason=Suicidal Admission type=ER 1365 ==> Mission=Surgery 1270 conf:(0.93)

9. Site=MGH Reason=Suicidal Discharge location=Home 1181 ==> Mission=Surgery 1093 conf:(0.93)

12. Reason=Disorientation Admission type=Elective 1087 ==> Mission=Medicine 992 conf:(0.91)

47. Reason=Suicidal Gender=M 1323 ==> Mission=Surgery 1095 conf:(0.83)

48. Reason=Disorientation Admission type=Urgent 1455 ==> Mission=Surgery 1197 conf:(0.82)

50. Site=MGH Reason=Agitation Gender=F 1195 ==> Mission=Surgery 979 conf:(0.82)

54. Reason=Suicidal 2047 ==> Mission=Surgery 1659 conf:(0.81)

Similar to previous years, disoriented patients were usually in medical mission, where agitated and suicidal patients were in surgical mission. However, unlike previous years, female patients were found to be agitated too, as rule 50 showed.

Using Predictive Apriori algorithm

2008

3. Site=MGH Reason=Agitation Age Group=50-59 Marital status=MARRIED_ADULT Language=English 299 ==> Mission=Medicine 299 acc:(0.99495)

4. Site=MGH Reason=Suicidal Language=French Length of stay group=>=100 298 ==> Mission=Surgery 298 acc:(0.99495)

20. Shift=Day Gender=F Language=English Admission type=Stretcher 297 ==> Mission=Emergency 297 acc:(0.99495)

22. Reason=Suicidal Gender=F Marital status=SINGLE_ADULT Length of stay group=>=100 296 ==> Mission=Surgery 296 acc:(0.99495)

58. Age Group=50-59 Marital status=MARRIED_ADULT Language=English Length of stay group=>=100 251 ==> Mission=Medicine 251 acc:(0.99492)

77. Reason=Suicidal Age Group=20-29 Gender=F Marital status=SINGLE_ADULT Length of stay group=>=100 223 ==> Mission=Surgery 223 acc:(0.99488)

86. Month=12 Reason=Suicidal Length of stay group=>=100 219 ==> Mission=Surgery 219 acc:(0.99488)

89. Site=MGH Shift=Day Reason=Suicidal Age Group=20-29 Length of stay group=>=100 215 ==> Mission=Surgery 215 acc:(0.99487)

96. Reason=Suicidal Language=French Length of stay group=>=100 299 ==> Mission=Surgery 298 acc:(0.99485)

97. Reason=Suicidal Gender=M Marital status=SINGLE_ADULT Length of stay group=>=100 Discharge location=Home 299 ==> Mission=Surgery 298 acc:(0.99485)

98. Reason=Suicidal Gender=M Municipality=MONTREAL Length of stay group=>=100 Discharge location=Home 299 ==> Mission=Surgery 298 acc:(0.99485)

Adult suicidal patients seemed to be singles in their twenties, who needed to be hospitalized longer (more than 100 days) in surgery mission unit. This happened to both genders. Patients in the age group of 50-59 seemed to be more likely to be hospitalized in Medicine mission units.

2009

1. Language=French Length of stay group=50-59 Discharge location=Home 299 ==> Mission=Surgery 299 acc:(0.99498)

35. Age Group=70-79 Marital status=SINGLE_ADULT Municipality=DOLLARD-DES-ORMEAUX Discharge location=Home 287 ==> Mission=Medicine 287 acc:(0.99497)

61. Marital status=MARRIED_ADULT Admission type=ER Length of stay group=90-99 Discharge location=Long term care 269 ==> Mission=Medicine 269 acc:(0.99497)

70. Marital status=SINGLE_ADULT Admission type=Stretcher Discharge location=Home 250 ==> Mission=Emergency 250 acc:(0.99496)

A lot more resulting attribute values could be found, which had not been discovered by the original Apriori algorithm. However, all of them had very low support.

2010

10. Reason=Suicidal Language=French Length of stay group=>=100 280 ==> Mission=Surgery 280 acc:(0.99494)

12. Reason=Suicidal Marital status=SINGLE_ADULT Language=French Municipality=MONTREAL Admission type=ER Discharge location=Home 280 ==> Mission=Surgery 280 acc:(0.99494)

13. Reason=Disorientation Age Group=70-79 Gender=M Language=French Admission type=Urgent 276 ==> Mission=Surgery 276 acc:(0.99494)

16. Reason=Agitation Gender=M Marital status=MARRIED_ADULT Admission type=Urgent 272 ==> Mission=Surgery 272 acc:(0.99493)

87. Reason=Away without leave Age Group=70-79 Municipality=POINTE-CLAIRE 194 ==> Mission=Medicine 194 acc:(0.99483)

95. Site=MGH Reason=Agitation Age Group=70-79 Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=ER Discharge location=Long term care 189 ==> Mission=Surgery 189 acc:(0.99481)

The only new information that was different than previous years was shown in rule 87. The rule engine could find relationship with the reason "Away without leave", age group and municipality. However, due to very low support (194/19567 = 0.99%), it might just represent very rare and specific cases.

WITH CLASS ATTRIBUTE = “SITE”

Using Apriori algorithm

2008

1. Mission=Neuro 1487 ==> Site=MGH 1487 conf:(1)

2. Mission=Neuro Admission type=ER 1362 ==> Site=MGH 1362 conf:(1)

3. Marital status=SINGLE_ADULT Admission type=ER Length of stay group=>=100 1570 ==> Site=MGH 1566 conf:(1)

4. Gender=M Language=French Length of stay group=>=100 1507 ==> Site=MGH 1475 conf:(0.98)

39. Mission=Emergency Reason=Disorientation 1660 ==> Site=RVH 1475 conf:(0.89)

53. Mission=Emergency Language=English 1929 ==> Site=RVH 1666 conf:(0.86)

96. Reason=Agitation Language=French 1785 ==> Site=MGH 1451 conf:(0.81)

97. Reason=Disorientation Age Group=80-89 Language=English 2130 ==> Site=RVH 1717 conf:(0.81)

2009

1. Mission=Surgery Reason=Agitation Language=French Admission type=ER 1798 ==> Site=MGH 1789 conf:(0.99)

4. Mission=Medicine Age Group=80-89 Gender=M Marital status=MARRIED_ADULT Language=English 1939 ==> Site=RVH 1863 conf:(0.96)

45. Reason=Agitation Gender=M Marital status=MARRIED_ADULT Language=English Discharge location=Home 1763 ==> Site=RVH 1567 conf:(0.89)

2010

1. Age Group=20-29 Gender=M 1001 ==> Site=MGH 991 conf:(0.99)

6. Reason=Suicidal Gender=M Admission type=ER 1015 ==> Site=MGH 990 conf:(0.98)

18. Reason=Disorientation Age Group=70-79 Language=Other 1060 ==> Site=RVH 1011 conf:(0.95)

85. Reason=Agitation Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=ER 1333 ==> Site=MGH 1149 conf:(0.86)

86. Reason=Disorientation Gender=M Length of stay group=>=100 1245 ==> Site=RVH 1071 conf:(0.86)

90. Reason=Disorientation Age Group=70-79 Municipality=MONTREAL 1534 ==> Site=RVH 1314 conf:(0.86)

91. Mission=Emergency Gender=F 1354 ==> Site=RVH 1159 conf:(0.86)

The rule engine could find rules associated to both adult hospital sites (MGH and RVH). The discovered rules showed that specific hospital seemed to get patients with certain characteristics. This could be explained because of different specialties within different hospitals.

Using Predictive Apriori algorithm

2008

2. Mission=Neuro Admission type=ER Length of stay group=30-39 299 ==> Site=MGH 299 acc:(0.99498)

3. Mission=Neuro Reason=Disorientation Gender=M Admission type=ER 298 ==> Site=MGH 298 acc:(0.99498)

4. Mission=Neuro Marital status=MARRIED_ADULT Language=French 296 ==> Site=MGH 296 acc:(0.99498)

71. Reason=Suicidal Municipality=POINTE-CLAIRE 234 ==> Site=MGH 234 acc:(0.99495)

Rules with higher accuracies had the hospital site "MGH" as the resulting attribute value. In those discovered association rules, they had neuroscience as the mission. However, none of the discovered rules had high enough support (i.e.: < 5%).

2009

1. Mission=Neuro Marital status=SINGLE_ADULT Admission type=ER Discharge location=Hospital 299 ==> Site=MGH 299 acc:(0.99497)

2. Gender=F Marital status=SINGLE_ADULT Language=French Admission type=Elective Discharge location=Home 298 ==> Site=RVH 298 acc:(0.99497)

10. Mission=Medicine Marital status=SINGLE_ADULT Length of stay group=>=100 Discharge location=Home 290 ==> Site=RVH 290 acc:(0.99497)

11. Mission=Medicine Reason=Suicidal Municipality=DOLLARD-DES-ORMEAUX 287 ==> Site=RVH 287 acc:(0.99497)

12. Mission=Medicine Reason=Suicidal Length of stay group=>=100 287 ==> Site=RVH 287 acc:(0.99497)

2010

2. Mission=Surgery Reason=Agitation Age Group=20-29 Gender=M Admission type=ER 299 ==> Site=MGH 299 acc:(0.99493)

6. Reason=Away without leave Age Group=70-79 Municipality=MONTREAL 297 ==> Site=MGH 297 acc:(0.99493)

24. Mission=Surgery Reason=Suicidal Age Group=20-29 283 ==> Site=MGH 283 acc:(0.99492)

35. Age Group=20-29 Length of stay group=40-49 273 ==> Site=MGH 273 acc:(0.99491) 38. Mission=Neuro Reason=Agitation Gender=M 272 ==> Site=MGH 272 acc:(0.99491)

46. Mission=Medicine Reason=Disorientation Discharge location=Residence 267 ==> Site=RVH 267 acc:(0.9949)

47. Mission=Medicine Reason=Disorientation Gender=F Length of stay group=60-69 267 ==> Site=RVH 267 acc:(0.9949)

A lot of rules were discovered with the resulting both attribute values "MGH and "RVH". Support counts of all the rules were rather low (<2%). Information was rather very scattered and did not bring into any significant conclusion.

<u>WITH CLASS ATTRIBUTE = “SHIFT”</u>

Using Apriori algorithm

2008, 2009, 2010

No association rules could be found.

WITH CLASS ATTRIBUTE = “AGE GROUP”

Using Apriori algorithm

2008, 2010

No association rules could be found.

2009

1. Mission=Medicine Reason=Agitation Marital status=MARRIED_ADULT Municipality=MONTREAL Admission type=Elective 1113 ==> Age Group=80-89 1113 conf:(1)

2. Mission=Medicine Site=RVH Reason=Agitation Gender=M Municipality=MONTREAL Admission type=Elective 1113 ==> Age Group=80-89 1113 conf:(1)

3. Mission=Medicine Site=RVH Reason=Agitation Marital status=MARRIED_ADULT Municipality=MONTREAL Admission type=Elective 1113 ==> Age Group=80-89 1113 conf:(1)

40. Mission=Medicine Site=RVH Reason=Agitation Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Discharge location=Home 1174 ==> Age Group=80-89 1133 conf:(0.97)

There was only one resulting attribute value found, that was, Age group = 80-89. All the rules were just deviations of the above rule 40 with mix and match of same attribute values.

Using Predictive Apriori algorithm

2008

1. Mission=Surgery Reason=Suicidal Language=French Length of stay group=>=100 298 ==> Age Group=20-29 298 acc:(0.99479)

13. Site=RVH Reason=Dementia Gender=F Language=English 297 ==> Age Group=80-89 297 acc:(0.99479)

15. Reason=Disorientation Gender=M Language=English Municipality=COTE-SAINT-LUC Discharge location=Long term care 288 ==> Age Group=70-79 288 acc:(0.99477)

18. Mission=Medicine Reason=Dementia Gender=F 282 ==> Age Group=80-89 282 acc:(0.99476)

19. Mission=Surgery Reason=Suicidal Gender=F Municipality=MONTREAL Length of stay group=>=100 282 ==> Age Group=40-49 282 acc:(0.99476)

The engine could find more or less similar information than the original Apriori algorithm. Not much knowledge could be gained.

2009

4. Reason=Suicidal Admission type=Elective 295 ==> Age Group=70-79 295 acc:(0.99491)

14. Site=RVH Reason=Suicidal Length of stay group=>=100 287 ==> Age Group=70-79 287 acc:(0.9949)

17. Reason=Suicidal Municipality=DOLLARD-DES-ORMEAUX Discharge location=Home 287 ==> Age Group=70-79 287 acc:(0.9949)

18. Reason=Suicidal Length of stay group=>=100 Discharge location=Home 287 ==> Age Group=70-79 287 acc:(0.9949)

42. Mission=Surgery Reason=Agitation Marital status=SINGLE_ADULT Length of stay group=>=100 278 ==> Age Group=20-29 278 acc:(0.99489)

54. Reason=Disorientation Marital status=MARRIED_ADULT Length of stay group=90-99 269 ==> Age Group=80-89 269 acc:(0.99488)

Rules with age group "80-89" were similar to the ones found by the original Apriori. However, the predictive Apriori was able to discover rules with different resulting attribute values (i.e.: other age groups). Different age group seemed to be related to different set of attributes.

2010

10. Mission=Medicine Site=RVH Gender=F Language=French Length of stay group=40-49 220 ==> Age Group=50-59 220 acc:(0.99467)

30. Reason=Agitation 203 ==> Age Group=70-79 203 acc:(0.9946)

38. Reason=Away without leave Gender=F Length of stay group=40-49 198 ==> Age Group=50-59 198 acc:(0.99457)

45. Municipality=OUTREMONT Length of stay group=>=100 189 ==> Age Group=70-79 189 acc:(0.99452)

58. Length of stay group=>=100 Discharge location=ACUTE STATUS 177 ==> Age Group=50-59 177 acc:(0.99444)

Rules with "50-59" and "70-79" were the most popular age groups among the rules found. Some simple rules were found such as rule 30. It indicated that agitated patients were mostly in the age group of 70 to 79. Other rules probably point to very specific cases of individual cases, due to low support values.

WITH CLASS ATTRIBUTE = “GENDER”

Using Apriori algorithm

2008

8. Reason=Disorientation Length of stay group=>=100 1828 ==> Gender=M 1688 conf:(0.92)

14. Admission type=Urgent 1656 ==> Gender=M 1423 conf:(0.86)

18. Mission=Neuro 1487 ==> Gender=M 1250 conf:(0.84)

36. Site=RVH Reason=Agitation Marital status=MARRIED_ADULT 1659 ==> Gender=M 1328 conf:(0.8)

All the discovered rules had "Male" as the resulting attribute value. They pointed out some specifics about sitter cases for male patients.

2009

20. Site=RVH Reason=Agitation Marital status=MARRIED_ADULT Discharge location=Home 1856 ==> Gender=M 1680 conf:(0.91)

21. Mission=Medicine Site=RVH Marital status=MARRIED_ADULT Language=English 2621 ==> Gender=M 2371 conf:(0.9)

22. Reason=Agitation Municipality=MONTREAL Discharge location=Home 2144 ==> Gender=M 1935 conf:(0.9)

23. Site=RVH Reason=Agitation Marital status=MARRIED_ADULT Language=English Discharge location=Home 1739 ==> Gender=M 1567 conf:(0.9)

86. Mission=Medicine Site=RVH Age Group=80-89 2919 ==> Gender=M 2424 conf:(0.83)

98. Reason=Agitation Length of stay group=>=100 2159 ==> Gender=M 1775 conf:(0.82)

All the rules had only one resulting attribute value, "Male". As male patients occupied most sitter cases (15108/22229 = 67.97%), it was not surprised to see "Male" in most of the rules. However, although female patients still occupied significant amount of sitter cases, the attribute value never showed up in any of the rules.

2010

1. Mission=Medicine Municipality=MONTREAL Length of stay group=>=100 1148 ==> Gender=M 1148 conf:(1)

2. Reason=Disorientation Age Group=70-79 Length of stay group=>=100 1047 ==> Gender=M 1044 conf:(1)

17. Reason=Disorientation Admission type=Elective 1087 ==> Gender=M 1038 conf:(0.95)

72. Reason=Disorientation Marital status=SINGLE_ADULT 1930 ==> Gender=M 1606 conf:(0.83)

73. Reason=Agitation Language=English Discharge location=Home 1543 ==> Gender=M 1283 conf:(0.83)

All the rules had "male" as patient's gender. According to the statistics, there were significant higher number of sitter cases for male patients in year 2010 (12725/19567 = 65.03%). It was not a surprise to see rules occupied by male as the patient's gender.

Using Predictive Apriori algorithm

2008

3. Reason=Suicidal Language=French Length of stay group=>=100 299 ==> Gender=M 299 acc:(0.99498)

4. Mission=Surgery Shift=Day Marital status=SINGLE_ADULT Language=English Length of stay group=>=100 299 ==> Gender=F 299 acc:(0.99498)

13. Mission=Surgery Site=MGH Reason=Disorientation Marital status=SINGLE_ADULT Discharge location=Hospital 297 ==> Gender=M 297 acc:(0.99498)

16. Mission=Surgery Shift=Day Reason=Suicidal Language=English Length of stay group=>=100 296 ==> Gender=F 296 acc:(0.99497)

17. Mission=Medicine Reason=Agitation Language=English Admission type=Elective Length of stay group=>=100 292 ==> Gender=M 292 acc:(0.99497)

23. Mission=Surgery Reason=Suicidal Language=English Municipality=MONTREAL Length of stay group=>=100 282 ==> Gender=F 282 acc:(0.99497)

25. Age Group=70-79 Length of stay group=>=100 Discharge location=Long term care 281 ==> Gender=M 281 acc:(0.99497)

2009

3. Reason=Agitation Length of stay group=>=100 Discharge location=Hospital 296 ==> Gender=M 296 acc:(0.99495)

18. Mission=Medicine Reason=Suicidal Municipality=DOLLARD-DES-ORMEAUX 287 ==> Gender=F 287 acc:(0.99494)

19. Mission=Medicine Reason=Suicidal Admission type=Elective 287 ==> Gender=F 287 acc:(0.99494)

20. Mission=Medicine Reason=Suicidal Length of stay group=>=100 287 ==> Gender=F 287 acc:(0.99494)

26. Reason=Suicidal Age Group=70-79 Municipality=DOLLARD-DES-ORMEAUX 287 ==> Gender=F 287 acc:(0.99494)

27. Reason=Suicidal Age Group=70-79 Length of stay group=>=100 287 ==> Gender=F 287 acc:(0.99494)

Most of the rules had "Male" as the resulting attribute. They provided more or less the same information than the ones by the original Apriori algorithm. There were a few rules that had "Female" as the resulting attribute. However, they provided almost the same findings as if with the class attribute "Reason".

2010

1. Age Group=50-59 Language=French Length of stay group=40-49 299 ==> Gender=F 299 acc:(0.99496)

33. Mission=Surgery Site=MGH Reason=Agitation Discharge location=Long term care 249 ==> Gender=F 249 acc:(0.99493)

55. Mission=Womens 223 ==> Gender=F 223 acc:(0.9949)

75. Site=RVH Reason=Away without leave Age Group=70-79 203 ==> Gender=F 203 acc:(0.99487)

100. Reason=Away without leave Length of stay group=>=100 176 ==> Gender=F 176 acc:(0.99481)

The engine was able to discover a number of rules with female patients. Rule 55 showed an obvious fact - Women's health mission had female patients. All the rules had relatively low support, which was less than 2%. The reason "Away without leave" seemed to only happen to female patients and the length of stay of such incidents was always long (>=100 days).

WITH CLASS ATTRIBUTE = “MARITAL STATUS”

Using Apriori algorithm

2008

1. Site=MGH Reason=Suicidal Length of stay group=>=100 1247 ==> Marital status=SINGLE_ADULT 1247 conf:(1)

2. Reason=Suicidal Admission type=ER Length of stay group=>=100 1247 ==> Marital status=SINGLE_ADULT 1247 conf:(1)

3. Site=MGH Reason=Suicidal Admission type=ER Length of stay group=>=100 1247 ==> Marital status=SINGLE_ADULT 1247 conf:(1)

4. Mission=Surgery Reason=Suicidal Length of stay group=>=100 1114 ==> Marital status=SINGLE_ADULT 1114 conf:(1)

5. Mission=Surgery Site=MGH Reason=Suicidal Length of stay group=>=100 1114 ==> Marital status=SINGLE_ADULT 1114 conf:(1)

26. Mission=Surgery Reason=Suicidal Discharge location=Home 1187 ==> Marital status=SINGLE_ADULT 1136 conf:(0.96)

33. Age Group=20-29 1432 ==> Marital status=SINGLE_ADULT 1355 conf:(0.95)

All the rules had "Single Adult" as the resulting attribute, with "Suicidal" as the reason for sitters. Interestingly, patients between age of 20 and 29 were all singles.

2009

1. Site=RVH Reason=Agitation Gender=M Municipality=MONTREAL Admission type=Elective 1196 ==> Marital status=MARRIED_ADULT 1196 conf:(1)

44. Site=RVH Reason=Agitation Gender=M Language=English Municipality=MONTREAL Admission type=Elective Length of stay group=>=100 1112 ==> Marital status=MARRIED_ADULT 1112 conf:(1)

72. Site=RVH Reason=Agitation Age Group=80-89 Gender=M Language=English Municipality=MONTREAL Discharge location=Home 1131 ==> Marital status=MARRIED_ADULT 1120 conf:(0.99)

All the rules had the same attribute values, with mix and match of different attribute combinations. The resulting attribute value was always "Married Adult". This value appeared very frequently in the dataset, 9677/22229 = 43.53%.

2010

1. Age Group=20-29 1217 ==> Marital status=SINGLE_ADULT 1126 conf:(0.93)

2. Site=MGH Age Group=20-29 1106 ==> Marital status=SINGLE_ADULT 1015 conf:(0.92)

3. Site=RVH Gender=M Language=Other 1282 ==> Marital status=MARRIED_ADULT 1159 conf:(0.9)

4. Site=RVH Reason=Disorientation Gender=M Language=Other 1150 ==> Marital status=MARRIED_ADULT 1036 conf:(0.9)

5. Mission=Medicine Site=RVH Length of stay group=>=100 1273 ==> Marital status=MARRIED_ADULT 1114 conf:(0.88)

6. Age Group=70-79 Gender=M Language=Other 1135 ==> Marital status=MARRIED_ADULT 985 conf:(0.87)

7. Age Group=70-79 Language=Other 1284 ==> Marital status=MARRIED_ADULT 1080 conf:(0.84)

8. Reason=Disorientation Age Group=70-79 Gender=M Discharge location=Home 1264 ==> Marital status=MARRIED_ADULT 1040 conf:(0.82)

9. Site=RVH Language=Other 1505 ==> Marital status=MARRIED_ADULT 1228 conf:(0.82)

10. Site=RVH Reason=Disorientation Language=Other 1313 ==> Marital status=MARRIED_ADULT 1071 conf:(0.82)

11. Site=RVH Age Group=70-79 Length of stay group=>=100 1246 ==> Marital status=MARRIED_ADULT 1016 conf:(0.82)

12. Reason=Disorientation Gender=M Language=Other 1289 ==> Marital status=MARRIED_ADULT 1050 conf:(0.81)

Two rules were discovered with "Single Adult" as the marital status. All the remaining ones were with "Married Adult". The number of married adults was very close to single adults (6850 vs. 6670). However, even "separated, divorced or widowed" patients occupied quite a portion in the dataset (6047 counts), no discovered rules had such value. From the above rules, patients in their twenties were "Single adults" and patients in their senior ages were "Married adults". Also, senior patients seemed to have longer length of stay (>=100 days).

Using Predictive Apriori algorithm

2008

2. Mission=Surgery Reason=Suicidal Gender=M Length of stay group=>=100 299 ==> Marital status=SINGLE_ADULT 299 acc:(0.99494)

4. Reason=Suicidal Age Group=20-29 Gender=M Length of stay group=>=100 298 ==> Marital status=SINGLE_ADULT 298 acc:(0.99494)

17. Mission=Medicine Reason=Disorientation Age Group=70-79 Municipality=COTE-SAINT-LUC 292 ==> Marital status=MARRIED_ADULT 292 acc:(0.99494)

18. Mission=Medicine Age Group=70-79 Gender=M Municipality=COTE-SAINT-LUC 291 ==> Marital status=MARRIED_ADULT 291 acc:(0.99494)

62. Mission=Medicine Site=MGH Reason=Away without leave Language=French 240 ==> Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT 240 acc:(0.99489)

Similar than the original Apriori, most of the rules found by the predictive Apriori had "Single Adult" as the resulting attribute. However, it was able to find the others such as "Married Adult" and "SEPARATED_DIVORCED_WIDOWED_ADULT". Marital status seemed to have impact to patient characteristics.

2009

1. Language=French Length of stay group=50-59 Discharge location=Home 299 ==> Marital status=MARRIED_ADULT 299 acc:(0.99493)

2. Site=MGH Reason=Suicidal Age Group=30-39 Admission type=ER 299 ==> Marital status=SINGLE_ADULT 299 acc:(0.99493)

3. Mission=Medicine Municipality=BROSSARD Admission type=ER 298 ==> Marital status=MARRIED_ADULT 298 acc:(0.99493)

4. Reason=Agitation Gender=M Language=French Length of stay group=50-59 297 ==> Marital status=MARRIED_ADULT 297 acc:(0.99493)

5. Month=3 Mission=Surgery Site=RVH Gender=M Language=English 294 ==> Marital status=MARRIED_ADULT 294 acc:(0.99493)

15. Site=RVH Reason=Suicidal Municipality=DOLLARD-DES-ORMEAUX 287 ==> Marital status=SINGLE_ADULT 287 acc:(0.99492)

16. Site=RVH Reason=Suicidal Length of stay group=>=100 287 ==> Marital status=SINGLE_ADULT 287 acc:(0.99492)

17. Reason=Suicidal Age Group=70-79 Municipality=DOLLARD-DES-ORMEAUX 287 ==> Marital status=SINGLE_ADULT 287 acc:(0.99492)

90. Site=RVH Reason=Agitation Age Group=80-89 Admission type=Elective 1197 ==> Marital status=MARRIED_ADULT 1183 conf:(0.99)

Although "Single Adult" was not the most frequently appeared marital status, a lot of rules could be found with such value. Again, rules offered us more or less same findings than with other class attributes, such as with "Reason".

2010

4. Reason=Suicidal Age Group=20-29 Gender=M 289 ==> Marital status=SINGLE_ADULT 289 acc:(0.99493)

5. Reason=Suicidal Language=French Length of stay group=>=100 280 ==> Marital status=SINGLE_ADULT 280 acc:(0.99493)

12. Mission=Neuro Age Group=20-29 263 ==> Marital status=SINGLE_ADULT 263 acc:(0.99491)

15. Site=MGH Reason=Agitation Discharge location=Long term care 256 ==> Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT 256 acc:(0.9949)

23. Age Group=50-59 Gender=F Length of stay group=>=100 240 ==> Marital status=SINGLE_ADULT 240 acc:(0.99488)

61. Reason=Away without leave Age Group=50-59 Length of stay group=40-49 198 ==> Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT 198 acc:(0.9948)

62. Reason=Away without leave Gender=F Length of stay group=40-49 198 ==> Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT 198 acc:(0.9948)

Rules with more resulting attribute values were found. Same as in previous years, "Away without leave" seemed to be specific to separated, divorced or widowed adult patients. For some reasons, female patient with ages between 50 and 59 who had stayed for more than 100 days were single adults. Due to low support value, the rule might only represent specific incidents.

WITH CLASS ATTRIBUTE = “LANGUAGE”

Using Apriori algorithm

2008

1. Site=MGH Gender=M Length of stay group=>=100 Discharge location=Home 1269 ==> Language=French 1190 conf:(0.94)

8. Mission=Surgery Site=MGH Gender=M Marital status=SINGLE_ADULT 1696 ==> Language=French 1399 conf:(0.82)

Since English speaking patients occupied most of the sitter cases (12314/21765 = 56.58%), it was expected to have English as the resulting attribute value in most discovered rules. It was more interesting to see rules with other attribute values. There were no rules with "Other" as the language resulting attribute value. Although some rules with "French" were discovered, it seemed like there was no relationship between language and patient characteristics.

2009

1. Site=RVH Reason=Agitation Marital status=MARRIED_ADULT Length of stay group=>=100 1401 ==> Language=English 1401 conf:(1)

6. Reason=Agitation Gender=M Marital status=MARRIED_ADULT Admission type=Elective Discharge location=Home 1387 ==> Language=English 1342 conf:(0.97)

As in year 2008, all the rules had "English" as the resulting attribute value. All the rules had same attribute values with different combinations of attributes.

2010

No association rules could be found.

Using Predictive Apriori algorithm

2008

4. Reason=Suicidal Age Group=20-29 Municipality=MONTREAL Length of stay group=>=100 298 ==> Language=French 298 acc:(0.99498)

12. Mission=Surgery Shift=Day Reason=Suicidal Gender=F Length of stay group=>=100 296 ==> Language=English 296 acc:(0.99498)

All discovered rules had either "English" or "French" as the resulting attribute value. Same as before, information was very scattered. No clear relationships could be found between language and patient characteristics.

2009

1. Site=RVH Gender=F Marital status=SINGLE_ADULT Admission type=Elective Discharge location=Home 298 ==> Language=French 298 acc:(0.99495)

2. Mission=Medicine Marital status=SINGLE_ADULT Municipality=DOLLARD-DES-ORMEAUX 292 ==> Language=French 292 acc:(0.99495)

3. Age Group=70-79 Marital status=SINGLE_ADULT Length of stay group=>=100 292 ==> Language=French 292 acc:(0.99495)

4. Mission=Medicine Age Group=70-79 Gender=F Municipality=DOLLARD-DES-ORMEAUX 292 ==> Language=French 292 acc:(0.99495)

5. Mission=Medicine Marital status=SINGLE_ADULT Length of stay group=>=100 Discharge location=Home 290 ==> Language=French 290 acc:(0.99494)

6. Site=RVH Marital status=SINGLE_ADULT Length of stay group=>=100 Discharge location=Home 290 ==> Language=French 290 acc:(0.99494)

7. Mission=Medicine Site=RVH Gender=F Admission type=Elective Discharge location=Home 289 ==> Language=French 289 acc:(0.99494)

8. Mission=Medicine Gender=F Marital status=SINGLE_ADULT Length of stay group=>=100 288 ==> Language=French 288 acc:(0.99494)

9. Site=RVH Gender=F Municipality=DOLLARD-DES-ORMEAUX Discharge location=Home 288 ==> Language=French 288 acc:(0.99494)

10. Mission=Medicine Reason=Suicidal Municipality=DOLLARD-DES-ORMEAUX 287 ==> Language=French 287 acc:(0.99494)

47. Site=RVH Admission type=Elective Length of stay group=70-79 282 ==> Language=English 282 acc:(0.99494)

Although "English" was the predominant language of most patients, most of the rules found by the predictive Apriori had "French" as the resulting attribute value. Most rules gave similar findings than ones with other class attributes.

2010

1. Age Group=50-59 Gender=F Length of stay group=40-49 299 ==> Language=French 299 acc:(0.99495)

2. Site=MGH Age Group=50-59 Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=Clinic 290 ==> Language=French 290 acc:(0.99495)

3. Site=RVH Gender=M Marital status=SINGLE_ADULT Length of stay group=70-79 289 ==> Language=English 289 acc:(0.99495)

4. Age Group=50-59 Gender=M Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Municipality=MONTREAL 284 ==> Language=French 284 acc:(0.99494)

5. Site=MGH Reason=Agitation Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Length of stay group=30-39 274 ==> Language=French 274 acc:(0.99494)

6. Mission=Medicine Reason=Disorientation Length of stay group=60-69 Discharge location=Home 262 ==> Language=French 262 acc:(0.99493)

7. Reason=Disorientation Municipality=MONTREAL Length of stay group=60-69 Discharge location=Home 262 ==> Language=French 262 acc:(0.99493)

8. Mission=Medicine Site=RVH Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=ER Length of stay group=40-49 258 ==> Language=French 258 acc:(0.99493)

9. Mission=Medicine Age Group=40-49 Municipality=MONTREAL Admission type=ER 256 ==> Language=French 256 acc:(0.99492)

10. Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=ER Length of stay group=60-69 Discharge location=Home 255 ==> Language=French 255 acc:(0.99492)

All discovered rules had either "English" or "French" as the resulting attribute value. Same as before, information was very scattered. No clear relationships could be found between language and patient characteristics.

WITH CLASS ATTRIBUTE = “MUNICIPALITY”

Using Apriori algorithm

2008

1. Reason=Suicidal Language=French Admission type=ER 1355 ==> Municipality=MONTREAL 1094 conf:(0.81)

The algorithm could only find one rule, as above. It did not tell much as a majority of patient population (9384/21765 = 43.12%) was from the municipality "Montreal".

2009

1. Language=English Admission type=Elective Length of stay group=>=100 Discharge location=Home 1303 ==> Municipality=MONTREAL 1303 conf:(1)

5. Mission=Medicine Site=RVH Reason=Agitation Age Group=80-89 Marital status=MARRIED_ADULT Admission type=Elective 1113 ==> Municipality=MONTREAL 1113 conf:(1)

27. Reason=Agitation Language=English Admission type=Elective Length of stay group=>=100 1150 ==> Municipality=MONTREAL 1137 conf:(0.99)

70. Mission=Medicine Site=RVH Reason=Agitation Age Group=80-89 Marital status=MARRIED_ADULT Language=English Discharge location=Home 1175 ==> Municipality=MONTREAL 1133 conf:(0.96)

The municipality "Montreal" was found to be the only resulting attribute value in all the rules. Almost half of the patients in the dataset resided in "Montreal" (10692/22229 = 48.1%). Rules did not give additional knowledge than the previous findings with other class attributes.

2010

1. Mission=Medicine Gender=M Admission type=Elective Discharge location=Home 1113 ==> Municipality=MONTREAL 1082 conf:(0.97)

2. Mission=Medicine Admission type=Elective Discharge location=Home 1122 ==> Municipality=MONTREAL 1086 conf:(0.97)

3. Gender=M Admission type=Elective Discharge location=Home 1232 ==> Municipality=MONTREAL 1126 conf:(0.91)

4. Admission type=Elective Discharge location=Home 1269 ==> Municipality=MONTREAL 1148 conf:(0.9)

5. Mission=Medicine Age Group=70-79 Discharge location=Home 1157 ==> Municipality=MONTREAL 1032 conf:(0.89)

6. Mission=Medicine Gender=M Admission type=Elective 1399 ==> Municipality=MONTREAL 1242 conf:(0.89)

7. Gender=M Length of stay group=>=100 Discharge location=Home 1113 ==> Municipality=MONTREAL 985 conf:(0.88)

8. Mission=Medicine Site=RVH Reason=Disorientation Discharge location=Home 1176 ==> Municipality=MONTREAL 1028 conf:(0.87)

9. Mission=Medicine Admission type=Elective 1432 ==> Municipality=MONTREAL 1247 conf:(0.87)

10. Mission=Medicine Reason=Disorientation Gender=M Discharge location=Home 1290 ==> Municipality=MONTREAL 1123 conf:(0.87)

11. Length of stay group=>=100 Discharge location=Home 1210 ==> Municipality=MONTREAL 1042 conf:(0.86)

12. Mission=Medicine Reason=Disorientation Discharge location=Home 1525 ==> Municipality=MONTREAL 1298 conf:(0.85)

Same as the previous year, the municipality "Montreal" was found to be the only resulting attribute value in all the rules. Almost half of the patients in the dataset resided in "Montreal" (8331/19567 = 42.58%). Rules did not give additional knowledge than the previous findings with other class attributes.

Using Predictive Apriori algorithm

2008

32. Mission=Medicine Marital status=MARRIED_ADULT Language=English Length of stay group=>=100 Discharge location=Long term care 271 ==> Municipality=COTE-SAINT-LUC 271 acc:(0.99438)

34. Discharge location=ACUTE STATUS 251 ==> Municipality=DORVAL 251 acc:(0.99423)

44. Reason=Agitation Age Group=50-59 Language=English Length of stay group=>=100 251 ==> Municipality=DORVAL 251 acc:(0.99423)

73. Site=MGH Reason=Suicidal Age Group=20-29 Language=English Admission type=ER Length of stay group=>=100 223 ==> Municipality=POINTE-CLAIRE 223 acc:(0.99395)

Since most patients (9384/21765 = 43.12%) lived in Montreal, it was expected to have most rules containing such resulting attribute value. Rules that contained other attribute values more likely showed more useful facts. All the rules found did not have high support. Some patient characteristics might be linked to where they lived.

2009

1. Mission=Surgery Site=MGH Age Group=50-59 Gender=M Language=English Admission type=ER Discharge location=Home 289 ==> Municipality=MONTREAL 289 acc:(0.9946)

5. Reason=Suicidal Age Group=70-79 Length of stay group=>=100 287 ==> Municipality=DOLLARD-DES-ORMEAUX 287 acc:(0.99459)

20. Marital status=SINGLE_ADULT Admission type=Elective Length of stay group=>=100 Discharge location=Home 287 ==> Municipality=DOLLARD-DES-ORMEAUX 287 acc:(0.99459)

43. Reason=Disorientation Marital status=MARRIED_ADULT Length of stay group=90-99 269 ==> Municipality=BROSSARD 269 acc:(0.9945)

55. Age Group=80-89 Marital status=MARRIED_ADULT Length of stay group=90-99 Discharge location=Long term care 269 ==> Municipality=BROSSARD 269 acc:(0.9945)

61. Reason=Suicidal Gender=M Language=English Length of stay group=>=100 260 ==> Municipality=MONTREAL 260 acc:(0.99445)

2010

1. Site=MGH Reason=Away without leave Age Group=70-79 297 ==> Municipality=MONTREAL 297 acc:(0.99497)

16. Age Group=40-49 Discharge location=Long term care 244 ==> Municipality=MONTREAL 244 acc:(0.99494)

24. Age Group=50-59 Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Length of stay group=40-49 220 ==> Municipality=LAVAL 220 acc:(0.99493)

51. Mission=Medicine Reason=Away without leave Age Group=50-59 Gender=F 198 ==> Municipality=LAVAL 198 acc:(0.9949)

66. Site=RVH Reason=Away without leave Discharge location=Long term care 176 ==> Municipality=POINTE-CLAIRE 176 acc:(0.99487)

100. Site=MGH Age Group=70-79 Length of stay group=70-79 142 ==> Municipality=WESTMOUNT 142 acc:(0.99478)

The rule engine was able to find a lot of rules with different municipalities. However, due to similarity (similar attribute values and combination) between different rules, there was no evidence to show relationship between sitter cases and where patients resided. Rules seemed to point to specific sitter cases with same patients.

WITH CLASS ATTRIBUTE = “ADMISSION TYPE”

Using Apriori algorithm

2008

9. Mission=Surgery Site=MGH Marital status=SINGLE_ADULT Length of stay group=>=100 1258 ==> Admission type=ER 1252 conf:(1)

10. Mission=Surgery Site=MGH Reason=Suicidal Gender=F Marital status=SINGLE_ADULT 1147 ==> Admission type=ER 1139 conf:(0.99)

11. Mission=Surgery Reason=Suicidal Gender=F Marital status=SINGLE_ADULT 1186 ==> Admission type=ER 1177 conf:(0.99)

12. Mission=Surgery Site=MGH Reason=Suicidal Discharge location=Home 1099 ==> Admission type=ER 1089 conf:(0.99)

34. Mission=Emergency Length of stay group=0-9 Discharge location=Hospital 1748 ==> Admission type=Stretcher 1692 conf:(0.97)

37. Mission=Emergency Site=RVH Length of stay group=0-9 Discharge location=Hospital 1455 ==> Admission type=Stretcher 1403 conf:(0.96)

60. Reason=Disorientation Length of stay group=0-9 Discharge location=Hospital 1175 ==> Admission type=Stretcher 1104 conf:(0.94)

52. Reason=Suicidal Gender=F Marital status=SINGLE_ADULT 1307 ==> Admission type=ER 1240 conf:(0.95)

93. Mission=Emergency Reason=Disorientation Length of stay group=0-9 1379 ==> Admission type=Stretcher 1246 conf:(0.9)

Most of the rules had "ER" as the resulting attribute value. This was due to the fact that most sitter cases (13972/21765 = 64.19%) had such value. There were a few rules found with the admission type "Stretcher". Patients admitted with admission type "Stretcher" seemed to have shorter length of stay.

2009

4. Mission=Surgery Site=MGH Length of stay group=>=100 1128 ==> Admission type=ER 1128 conf:(1)

14. Site=RVH Reason=Agitation Gender=M Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Length of stay group=>=100 1112 ==> Admission type=Elective 1112 conf:(1)

24. Mission=Emergency Site=RVH Length of stay group=0-9 Discharge location=Hospital 1450 ==> Admission type=Stretcher 1433 conf:(0.99)

42. Mission=Medicine Site=RVH Reason=Agitation Age Group=80-89 Gender=M Marital status=MARRIED_ADULT Municipality=MONTREAL 1145 ==> Admission type=Elective 1113 conf:(0.97)

Most rules found were with the resulting attribute "Elective". However, comparing to the most frequently appeared admission type "ER", it only occupied 3048/22229 = 13.71%, which was way less than 12892/22229 = 58% for the admission type "ER".

2010

3. Site=RVH Reason=Disorientation Length of stay group=0-9 Discharge location=Hospital 1155 ==> Admission type=Stretcher 1077 conf:(0.93)

7. Reason=Disorientation Length of stay group=0-9 Discharge location=Hospital 1275 ==> Admission type=Stretcher 1151 conf:(0.9)

12. Mission=Surgery Site=MGH Gender=M Marital status=SINGLE_ADULT Discharge location=Home 1508 ==> Admission type=ER 1305 conf:(0.87)

21. Language=French Length of stay group=30-39 1265 ==> Admission type=ER 1081 conf:(0.85)

Most rules had "ER" as the admission type. It was the expected outcome since 11172/19567 = 57.1% of sitter cases had such value. An interesting observation was that cases with "Stretcher" as the admission type seemed to have shorter length of stay than the ones with "ER".

Using Predictive Apriori algorithm

2008

22. Mission=Medicine Reason=Agitation Marital status=SINGLE_ADULT Language=English Discharge location=Home 294 ==> Admission type=ER 294 acc:(0.99493)

29. Mission=Emergency Reason=Disorientation Gender=F Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Discharge location=Hospital 288 ==> Admission type=Stretcher 288 acc:(0.99493)

37. Mission=Emergency Reason=Disorientation Gender=F Language=English Discharge location=Hospital 278 ==> Admission type=Stretcher 278 acc:(0.99492)

54. Reason=Agitation Marital status=MARRIED_ADULT Municipality=DORVAL 254 ==> Admission type=Elective 254 acc:(0.99489)

The predictive Apriori algorithm was able to find more admission type values, for instance, "Elective". Even with same resulting attribute values, in addition to rules found by the original Apriori, the predictive Apriori algorithm was able to discover more relationships with other attributes.

2009

5. Reason=Agitation Age Group=70-79 Length of stay group=>=100 Discharge location=Long term care 298 ==> Admission type=ER 298 acc:(0.99496)

64. Month=8 Mission=Emergency Discharge location=Hospital 247 ==> Admission type=Stretcher 247 acc:(0.99493)

32. Mission=Medicine Marital status=SINGLE_ADULT Municipality=DOLLARD-DES-ORMEAUX Discharge location=Home 287 ==> Admission type=Elective 287 acc:(0.99495)

82. Mission=Medicine Reason=Disorientation Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Length of stay group=>=100 221 ==> Admission type=Elective 221 acc:(0.9949)

The rule engine was able to associate different attributes with only 3 of the admission types - ER, Stretcher and Elective. Most of the rules had the value "ER" and only one rule with the value "Stetcher". It was able to associate combination of attributes with the chosen class attribute, although with very low support.

2010

1. Age Group=50-59 Gender=F Length of stay group=40-49 299 ==> Admission type=ER 299 acc:(0.99495)

4. Site=MGH Reason=Suicidal Gender=M Length of stay group=40-49 298 ==> Admission type=ER 298 acc:(0.99495)

13. Mission=Emergency Reason=Disorientation Gender=F Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Length of stay group=0-9 Discharge location=Hospital 280 ==> Admission type=Stretcher 280 acc:(0.99493)

93. Reason=Agitation Discharge location=ACUTE STATUS 177 ==> Admission type=Clinic 177 acc:(0.99477)

96. Length of stay group=>=100 Discharge location=ACUTE STATUS 177 ==> Admission type=Clinic 177 acc:(0.99477)

98. Reason=Away without leave Length of stay group=>=100 176 ==> Admission type=Clinic 176 acc:(0.99476)

The predictive Apriori was able to find the admission type "Clinic" in the rules, although with low support. Rules with "ER" as the admission type had very mixed attributes and values. They seemed to refer to many various cases. Since "ER"

occupied most of the cases, it was not as interesting as other attribute values. Rules with "Clinic" as the admission type seemed to refer to cases of same patient, due to same attribute values.

WITH CLASS ATTRIBUTE = “LENGTH OF STAY”

Using Apriori algorithm

2008

1. Age Group=80-89 Admission type=Stretcher 1158 ==> Length of stay group=0-9 1142 conf:(0.99)

6. Admission type=Stretcher 2796 ==> Length of stay group=0-9 2659 conf:(0.95)

21. Reason=Disorientation Admission type=Stretcher Discharge location=Hospital 1196 ==> Length of stay group=0-9 1104 conf:(0.92)

All the discovered rules had "0-9" as the resulting attribute. It was the most frequent attribute value (4584/21765 = 21.06%) in the dataset. However, the second most frequent attribute value did not appear, even it had 4169/21765 = 19.15%, which was just a little bit behind than the most frequent value.

2009

16. Site=RVH Reason=Agitation Gender=M Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL Admission type=Elective 1142 ==> Length of stay group=>=100 1112 conf:(0.97)

18. Gender=F Admission type=Stretcher 1370 ==> Length of stay group=0-9 1331 conf:(0.97)

35. Mission=Emergency Site=RVH Reason=Disorientation Admission type=Stretcher 1253 ==> Length of stay group=0-9 1190 conf:(0.95)

91. Mission=Emergency Gender=M 1811 ==> Length of stay group=0-9 1584 conf:(0.87)

97. Mission=Emergency Gender=F 1489 ==> Length of stay group=0-9 1290 conf:(0.87)

99. Reason=Agitation Age Group=80-89 Gender=M Discharge location=Home 1359 ==> Length of stay group=>=100 1176 conf:(0.87)

Most of the rules showed length of stay greater or equal to 100 days, specially for agitated patients. Rule 91 and 97 showed an interesting fact. Both gender admitted in an emergency unit had relatively short length of stay, between 0 and 9 days.

2010

1. Gender=M Admission type=Stretcher Discharge location=Hospital 1022 ==> Length of stay group=0-9 1021 conf:(1)

4. Site=RVH Reason=Disorientation Admission type=Stretcher Discharge location=Hospital 1079 ==> Length of stay group=0-9 1077 conf:(1)

9. Reason=Disorientation Admission type=Stretcher 1494 ==> Length of stay group=0-9 1449 conf:(0.97)

19. Admission type=Stretcher 2389 ==> Length of stay group=0-9 2284 conf:(0.96)

All the rules found had 0 to 9 days as the length of stay. They basically showed similar information than the ones with different class attributes. Rule 19 indicated a simple fact that patients had admitted by "Stretcher" tended to have short length of stay.

Using Predictive Apriori algorithm

2008

6. Age Group=80-89 Gender=M Municipality=MONTREAL Admission type=Stretcher 298 ==> Length of stay group=0-9 298 acc:(0.99495)

9. Mission=Surgery Reason=Suicidal Age Group=20-29 Municipality=MONTREAL Discharge location=Home 298 ==> Length of stay group=>=100 298 acc:(0.99495)

25. Site=RVH Reason=Agitation Gender=F Admission type=Stretcher Discharge location=Hospital 286 ==> Length of stay group=0-9 286 acc:(0.99494)

26. Reason=Agitation Age Group=80-89 Language=English Admission type=Stretcher 285 ==> Length of stay group=0-9 285 acc:(0.99494)

31. Mission=Surgery Reason=Suicidal Age Group=40-49 Gender=F Language=English 282 ==> Length of stay group=>=100 282 acc:(0.99494)

32. Mission=Surgery Reason=Suicidal Age Group=40-49 Language=English Municipality=MONTREAL 282 ==> Length of stay group=>=100 282 acc:(0.99494)

80. Reason=Dementia Marital status=SINGLE_ADULT Discharge location=Residence 221 ==> Length of stay group=70-79 221 acc:(0.99488)

100. Reason=Disorientation Age Group=30-39 Admission type=Clinic 201 ==> Length of stay group=>=100 201 acc:(0.99485)

Most rules contained the resulting attribute values "0-9" and ">=100", the two extreme lengths of stay in the dataset. There were a few rules with lengths of stay "70-79". Agitated patients did not seem to have long length of stay. In contrast, reasons like dementia, disorientation and suicidal seemed to cause longer lengths of stay. Those reasons also seemed to be related to patients' ages.

2009

1. Site=RVH Gender=M Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=Stretcher 299 ==> Length of stay group=0-9 299 acc:(0.99493)

9. Mission=Emergency Site=RVH Reason=Disorientation Gender=F Municipality=MONTREAL Admission type=Stretcher Discharge location=Hospital 295 ==> Length of stay group=0-9 295 acc:(0.99493)

20. Reason=Suicidal Age Group=70-79 Municipality=DOLLARD-DES-ORMEAUX 287 ==> Length of stay group=>=100 287 acc:(0.99492)

96. Mission=Surgery Reason=Agitation Age Group=80-89 Municipality=SAINTE-ANNE-DE-BELLEVUE Discharge location=Hospital 206 ==> Length of stay group=80-89 206 acc:(0.99479)

99. Reason=Agitation Age Group=70-79 Admission type=Stretcher 199 ==> Length of stay group=0-9 199 acc:(0.99477)

From the above rules, similar facts were indicated. Like the previous rules, patients who had suicidal problems tended to be either senior or twenties. They needed to stay for more than 100 days. Patients who were agitated usually came from the mission "Emergency" and were senior. In most cases, they did not need to be hospitalized for long.

2010

1. Reason=Disorientation Marital status=MARRIED_ADULT Language=English Admission type=Stretcher 299 ==> Length of stay group=0-9 299 acc:(0.99493)

9. Age Group=80-89 Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=Stretcher 295 ==> Length of stay group=0-9 295 acc:(0.99493)

16. Reason=Agitation Gender=M Admission type=Stretcher 290 ==> Length of stay group=0-9 290 acc:(0.99492)

24. Mission=Emergency Reason=Disorientation Gender=F Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Admission type=Stretcher Discharge location=Hospital 280 ==> Length of stay group=0-9 280 acc:(0.99491)

74. Mission=Emergency Site=RVH Reason=Disorientation Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Municipality=MONTREAL Admission type=Stretcher Discharge location=Hospital 216 ==> Length of stay group=0-9 216 acc:(0.99481)

79. Reason=Disorientation Age Group=40-49 Discharge location=Long term care 207 ==> Length of stay group=70-79 207 acc:(0.99479)

88. Site=RVH Reason=Agitation Gender=F Admission type=Stretcher 197 ==> Length of stay group=0-9 197 acc:(0.99476)

100. Reason=Agitation Discharge location=ACUTE STATUS 177 ==> Length of stay group=>=100 177 acc:(0.99469)

Length of stay and admission type both seemed to have strong relationship with reasons, as the previous rules with other class attributes had showed. Both "agitated" and "disoriented" patients seemed to have "stretcher" as the admission type. "Disoriented" patients had "Hospital" or "Long term care" as the discharge location.

WITH CLASS ATTRIBUTE = “DISCHARGE LOCATION”

Using Apriori algorithm

2008

1. Site=MGH Marital status=MARRIED_ADULT Language=French 1542 ==> Discharge location=Home 1257 conf:(0.82)

2. Mission=Medicine Site=MGH Language=French 1504 ==> Discharge location=Home 1219 conf:(0.81)

3. Gender=M Language=French Length of stay group=>=100 1507 ==> Discharge location=Home 1216 conf:(0.81)

4. Site=MGH Gender=M Language=French Length of stay group=>=100 1475 ==> Discharge location=Home 1190 conf:(0.81)

5. Language=French Length of stay group=>=100 1578 ==> Discharge location=Home 1269 conf:(0.8)

6. Site=MGH Language=French Length of stay group=>=100 1503 ==> Discharge location=Home 1206 conf:(0.8)

The engine could only find 6 rules. All of them were with the resulting attribute value "Home". The value existed very frequently in the dataset (9569 / 21765 = 43.97%). Although the next highest frequent attribute value "Hospital" appeared in 6680 sitter cases, no rules could be found with such value.

2009

43. Mission=Medicine Site=RVH Reason=Agitation Age Group=80-89 Marital status=MARRIED_ADULT Language=English Municipality=MONTREAL 1149 ==> Discharge location=Home 1133 conf:(0.99)

All the rules had "Home" as the discharge location. "Home" as the discharge location was the most frequently appeared attribute value (10766/22229 = 48.43%) in the dataset. Basically, all the discovered rules were variants of the above rule, but with different attribute combinations and same attribute values.

2010

1. Mission=Medicine Gender=M Municipality=MONTREAL Admission type=Elective 1242 ==> Discharge location=Home 1082 conf:(0.87)

2. Mission=Medicine Municipality=MONTREAL Admission type=Elective 1247 ==> Discharge location=Home 1086 conf:(0.87)

3. Mission=Surgery Site=MGH Gender=M Marital status=SINGLE_ADULT Admission type=ER 1511 ==> Discharge location=Home 1305 conf:(0.86)

6. Language=English Admission type=Stretcher Length of stay group=0-9 1274 ==> Discharge location=Hospital 1073 conf:(0.84)

11. Language=English Admission type=Stretcher 1332 ==> Discharge location=Hospital 1088 conf:(0.82)

15. Gender=M Admission type=Stretcher Length of stay group=0-9 1270 ==> Discharge location=Hospital 1021 conf:(0.8)

Only two resulting attribute values could be found - Home and Hospital. As the previous analysis also showed, "Stretcher" as the admission type seemed to be related to shorter length of stay. Also, patients admitted through "Stretcher" tended to be discharged back to the hospital.

Using Predictive Apriori algorithm

2008

28. Mission=Medicine Reason=Disorientation Age Group=70-79 Gender=M Length of stay group=>=100 272 ==> Discharge location=Long term care 272 acc:(0.99477)

29. Mission=Medicine Site=MGH Reason=Away without leave Admission type=ER 263 ==> Discharge location=Home 263 acc:(0.99475)

30. Mission=Surgery Language=French Admission type=ER Length of stay group=60-69 256 ==> Discharge location=Hospital 256 acc:(0.99473)

31. Age Group=50-59 Municipality=DORVAL 251 ==> Discharge location=ACUTE STATUS 251 acc:(0.99471)

32. Municipality=DORVAL Length of stay group=>=100 251 ==> Discharge location=ACUTE STATUS 251 acc:(0.99471)

33. Mission=Medicine Reason=Agitation Municipality=DORVAL 251 ==> Discharge location=ACUTE STATUS 251 acc:(0.99471)

82. Admission type=Elective Length of stay group=70-79 221 ==> Discharge location=Residence 221 acc:(0.99459)

83. Mission=Medicine Reason=Dementia Length of stay group=70-79 221 ==> Discharge location=Residence 221 acc:(0.99459)

The algorithm was able to discover rules with relatively infrequent attribute values, such as "Acute status" (251/21765 = 1.15%) and "Residence" (796/21765 = 3.66%). Since those rules had pretty high accuracy, they might contain important facts about sitter cases.

2009

1. Reason=Disorientation Age Group=50-59 Municipality=MONTREAL Admission type=ER 299 ==> Discharge location=Home 299 acc:(0.99481)

2. Reason=Disorientation Age Group=80-89 Length of stay group=90-99 298 ==> Discharge location=Long term care 298 acc:(0.9948)

4. Marital status=SINGLE_ADULT Language=English Length of stay group=>=100 297 ==> Discharge location=Hospital 297 acc:(0.9948)

5. Reason=Suicidal Admission type=Elective 295 ==> Discharge location=Home 295 acc:(0.9948)

26. Reason=Suicidal Gender=F Language=French Length of stay group=>=100 287 ==> Discharge location=Home 287 acc:(0.99478)

35. Site=RVH Age Group=70-79 Gender=F Language=French Length of stay group=>=100 287 ==> Discharge location=Home 287 acc:(0.99478)

2010

1. Site=MGH Reason=Suicidal Language=French Municipality=MONTREAL Admission type=ER 294 ==> Discharge location=Home 294 acc:(0.99497)

30. Mission=Medicine Site=MGH Reason=Disorientation Length of stay group=70-79 207 ==> Discharge location=Long term care 207 acc:(0.99491)

43. Mission=Neuro Reason=Constant observation in 4-point restraints 184 ==> Discharge location=Home 184 acc:(0.99488)

47. Site=MGH Gender=M Marital status=SEPARATED_DIVORCED_WIDOWED_ADULT Language=French Admission type=Clinic Length of stay group=>=100 177 ==> Discharge location=ACUTE STATUS 177 acc:(0.99487)

62. Reason=Constant observation in 4-point restraints Length of stay group=60-69 164 ==> Discharge location=Home 164 acc:(0.99484)

95. Age Group=80-89 Municipality=BROSSARD 142 ==> Discharge location=Morgue 142 acc:(0.99477)

More discharge locations were picked up by the predictive Apriori algorithm for both years 2009 and 2010. In most cases, patients were discharged "Home". Elderly (at least 70 years old) disoriented patients were discharged to long term care facilities. However, suicidal patients were mostly discharged "Home", after a long enough length of stay (>= 100 days). There were few cases that elderly patients between 80 and 89 had been discharged to "Morgue".