**Individual Assignment 5: Towards Quantitative Medicine**

The application of data mining techniques for knowledge discovery holds big promise. Data mining algorithms have been successfully applied in many different application areas, including but not limited to, retail, telecommunications, and more. However, applying these methods in the medical domain has its challenges because the data sets are often very large and complex, with numerous rare variables such as diagnosis, procedures and drugs. These variables are rare as most people are healthy (and therefore have few diagnosis, procedures and drugs), and those who are not, suffer from a wide variety of conditions (and therefore relatively few people have the same set of conditions).

The data used in this assignment is based on claims data. Claims data is generated when hospitals, pharmacies and other health care providers send claims to third party payers to receive reimbursement for their services. The claims data include information on diagnoses, procedures and drugs prescriptions as well as place of service, and patient's age, gender and ZIP-code. A collection of members' claims has the benefit of giving a "bird's eye view'' of patients’ health care.

Diabetes is a disease in which the body does not produce or properly use insulin. Diabetes is a chronic disease, and it is believed that over 18 million Americans suffer from it. Effective disease management of this group, and advancing understanding of the disease, is therefore of great importance.

The data (members.csv) used in this study has information on over 17,000 diabetic patients. We have data on their diagnosis, procedures and drugs over a 12 month observation period (as well as some other details from the observation period) together with their overall health care costs in the 12 months following the observation period, called the result period. We will apply Association Rules to gain insights into diabetes, risky health care patterns and exploratory data-mining in general.

**Associations Rule Analysis**

1. For each of the diagnosis columns, count the number of members with the diagnosis.
   1. Provide a readable (translate the diagnosis codes into actual conditions) table as an exhibit of the top 10 diagnosis and their counts.

**Table

Description automatically generated**

1. Delete Diag\_DD0046 from the data (or otherwise exclude it from the analysis), as R will not allow you to run the analysis with a column that has only one value (all of our patients are diabetic).  
     
   Reduce the dataset, use all members with a number of diagnosis above some threshold that you define (suggested value is 3). Start with a higher threshold, and then reduce it as your computer allows.
   1. What was your threshold?

*3*

* 1. How many members does that correspond to?

13482 from original 17433 total members.

1. Run the association rule algorithm, with the default settings. How many rules did the algorithm find?

Support = 0.001, conf = 0.001, 59446 rules

1. One of the difficulties with creating good rules with medical data is that most diseases are rare; therefore, setting the support too high will create uninteresting rules, if any at all. Run the association rule algorithm again with all default settings except changing the minimum support settings:
2. How many rules were created when the minimum support is 10%?
   * 13
3. How many rules were created when the minimum support is 1%?
   * 1038
4. How many rules were created when the minimum support is 0.1%?
   * 59446
5. Select one of your runs that resulted in at least one rule! Analyze the three rules that have the maximum lift. Provide an exhibit that summarizes the rules and a brief interpretation of these three rules (translate codes to descriptions).

Graphical user interface, application, table, Excel

Description automatically generated

According to the result, it seems that the rules with maximum lift have some of the lowest support and confidence, and minimum lift tend to associate with high support and confidence.

Created rules with minimum support at 10%, 1% and 0.1% and minimum confidence at 0.1%. These rule parameters allow algorithmic interpretation on the rate (%) which RHS item also appear when LHS items are added (support), and the likelihood (%) of this combination appearing (confidence).

1. Rerun the Association Rules algorithm with minimum support at 1% and confidence at 10%.
2. How many rules were created?

1038

1. Select the rule with the highest lift ratio.
   1. Explain the rule in words.

Rule with support at 1% and confidence of 1%, the 7.25 lift ratio indicates that the antecedent set and the consequent set appear more often together than expected.

* 1. Hypothetically, if you were a doctor and a diabetic patient walks in already diagnosed with {the rules antecedent diagnosis}, how could the rule potentially guide your work (with all the simplifying assumptions needed to answer this question!).

You can see what are the consequent diagnosis that tend to also appear, and assess the likeliness of this combination appearing. You then can use lift ratio to see the probability of you correctly predicting this combination.

1. One of the main reasons behind the interest in data mining in health care is the hope that through intervention and prevention, one can help reduce health care costs by identifying patients early who are at risk of high health care cost. To use association rules to contribute to this goal, it is not enough to run association rules on the diagnosis data, as there is nothing that distinguishes between costly patients and not costly patients

In order to use association rules to distinguish high risk patients from low risk patients, we need to identify rules that have good support and high confidence on high cost population, but low support for the not-high-cost population. In particular, we are interested in identifying rules of the type:

{group of diagnosis codes} -> {High cost in a future period}

The variable TA2 contains the overall cost in the year following the observation period. Define a new variable that equals one if the overall cost is ≥ $30,000, and zero otherwise.

Run the Association Rule algorithm again using all diagnosis variables (continue to exclude the diabetes diagnosis) and the new high-cost variable.  Set the support as low as 0.1% and the keep the confidence at 10%. Identify the rules with high-cost in the consequent set (**HINT:** here the appearance parameter we pointed out in class may be helpful, alternatively run as you did before and sort the rules afterwards in R or Excel (per our Sushi exercise)).

* 1. How many rules of the form {group of diagnosis codes} -> {High cost in a future period} did you identify ?

4386

* 1. Briefly summarize the top 5 rules with the highest confidence (an Exhibit that summarizes the rules translating the diagnosis codes into actual conditions, their support, confidence and lift may be helpful).

Graphical user interface

Description automatically generated with low confidence

According to the screenshot above, the diagnose codes (in combination) shown in the top 5 rules have the given rate % (support) of which that lead to high costs, and the given likeliness % (confidence) of this happening. These 5 rules have the given probability (lift) of correctly predicting this combination.

* 1. Briefly summarize the top 5 rules with the highest support (again an Exhibit may be helpful).



Same comment from above. The difference here is that the case count is much larger than previous sorting and lower lift. This indicate that the prediction is more likely to sustain given larger sample.

* 1. Comment on the potential role of these rules in a prediction setting.

The different sorting view of the rules can help us better examine the importance of the relationships between antecedent and consequent items, and the rule’s performance depending on the size of sample data.