Association Rules Mining for Diseases of Preschool Children in Thailand Using R Programming

Parkpoom Chaisiriprasert
Department of Information Technology, College of Information and Communication Technology,
Rangsit University, Thailand
Corresponding author, e-mail: parkpoom.c@rsu.ac.th

Abstract

This paper presents a process to discover groups of diseases frequently diagnosed in preschool children using association rules mining approach based on the simple Apriori algorithm. The researcher applied such approach on the real dataset using R programming. Our dataset is obtained from Office of Policy and Strategic, Ministry of Public Health. The analyzed results show interesting information, e.g., the male children's disease rate is higher than the female's children disease rate, and the children under the age of 1 year have the highest disease rate. In addition, the results obtained from association rules mining indicate that the groups of highest frequent diseases are (1) Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism, (2) Diseases of the respiratory system, and (3) Certain conditions originating in the perinatal period.

Keywords**:** *data mining, association rules mining, data analytics, R programming, preschool*

1. Introduction

Preschool children or children aged between 0-3 years are having rapid growth on neurons. As a result, learning behavior and developing skills in various areas has a direct impact on the development of children in the long run. When children are unhealthy, they cannot effectively learn or develop themselves. Another problem we cannot overlook is the illness or disease in children. It can prevent the normal development of any part of the body. There are several groups of diseases that are commonly found in preschool children, for example, diseases caused by a virus, endocrine system disease, respiratory diseases, diseases caused by infection, and diseases with congenital disorders.

One might see that the aforementioned diseases are serious and can cause severe pain. Understanding the patterns and factors associated with the disease will make us understand that some diseases are very serious and may cause death for children (Mungkornnintra, 2009). Some diseases may relate to other diseases which may result the complications. By knowing the relationships between diseases, we would be able to prevent such complications. Furthermore, studying the rate of diseases commonly found in preschool children is crucial in order to better handle such diseases. As a result, we are interested to study disease rates in preschool children in Thailand. To do so, we apply the simple Apriori algorithm to find patterns and relationships between groups of diseases. We select such algorithm because it is suitable for discovering sets of frequent (i.e., likely to occurred) items in a dataset. In addition, using the simple Apriori algorithm is a good starting point for data analytics in healthcare in Thailand which is a relatively new research area. The results can be used to plan for prevention of some diseases, or plan for policies to improve the quality of life in early childhood for children to grow bright and having good quality of life.

Nowadays, data science is emerging as a new interdisciplinary field of scientific methods for extracting knowledge from data, similar to data mining. Many tools have been proposed to facilitate the process of data science. Such tools are divided into 2 types: (1) Data science platforms are the tools that assist the process of data science and data mining with no coding skills required. Some examples of the data science platforms are KNIME, RapidMiner and Pentaho. (2) Programming languages for data science, e.g., R, Julia and Python, are suitable for handling more complex data, however with coding skills required. In this work, we demonstrate the process of discovering groups of the diseases frequently found in preschool children using a data mining technique, association rules mining.

2. Related Research

Data mining can be used to discover knowledge from database in a variety of areas, such as finance, social media, and healthcare. In this work, we focus on the area of data mining in health information. In fact, there are many works that introduced and applied data mining techniques on healthcare information. The example of such works are as follows:

Takeuchi et el. (2006) have applied a data mining technique to develop an automated response analysis system based on the foundation of personal health records. Their system extracts useful information such as rules and regulations about lifestyle and health conditions. By storing chronological information on an individual mobile devices and web applications, this system will allow users to save their health data daily via smart phone and visualize summary of their heath information through the web application. The system analyzes the data obtained, extracts the results, and provides useful health care analysis and guidance back to users.

Several works focus on applying data mining techniques in diagnosis and treatment data of different diseases. For example, Gosain and Kumar (2009) have studied on the analysis of health information using data mining techniques. They have extracted the results as a strategy for treatment. In particular, the disease of H.I.V. (Human Immunodeficiency Virus) infection or AIDS, and the application of this guideline can have a great impact on management and a strategy that will help prepare users to deal with H.I.V in the future. Shouman et al. (2012) have applied varieties of data mining techniques in a treatment data of heart disease. The study shows that by applying appropriate techniques helps improve the treatment of such disease.

Nuaimi (2014) has studied different data mining techniques for constructing predictive models in healthcare. Concretely, the predictive models are used for predicting demand of healthcare services in Abu Dhabi. Different techniques have been tested and compared to find the most suitable technique with the most optimal model.

Balasubramanian and Umarani (2012) have studied on the analysis of the health impacts on fluoride in water using data clustering techniques on data collected from Krishnagiri, India. In this research, they use a clustering technique to group the data of water based on the level of fluoride and the area of the city of Krishnagiri because there are many people who consume too much fluoride. Their work also describes the risk factors for fluoridated water for humans which can be applied to health to improve healthcare policy for the Indian health ministry.

Batra et el. (2014) used several data mining techniques to define a set of standard for recording health information in medical electronic devices. Thereafter, such data can be used for supporting the decision making to improve healthcare policies. In this work, they also describe the process for medical information analysis of large database in order to discover knowledge and decision models.

3. Material and Methods

In this work, the researchers aimed to discover association rules of diseases frequently appearing in Thai preschool children. We adopt the Apriori algorithm to mine association rules on a dataset that represents the diagnostic data from public hospitals in Thailand. To do so, the researchers defined a process as depicted in Figure 1. This process consisted of 4 steps: (1) data collection, (2) data preparation, (3) data grouping, and (4) association rules mining.

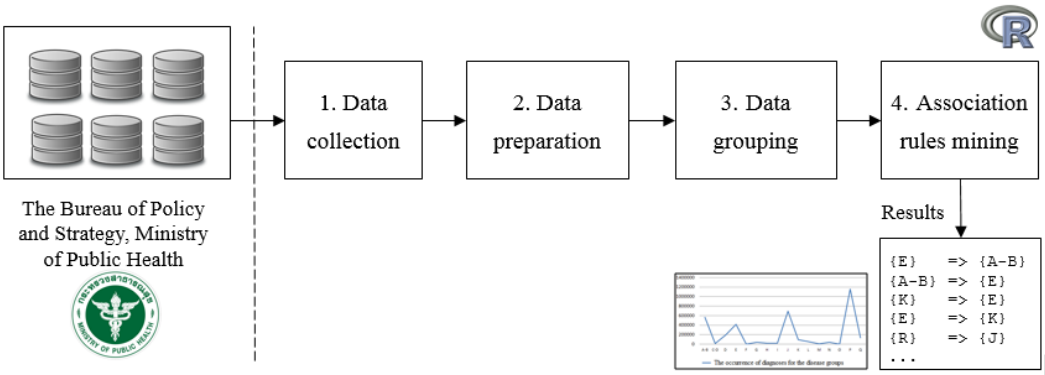

Figure 1 The process of discover association rules using R programming

3.1 Data collection

The researchers appreciated the support of the Bureau of Policy and Strategy, Ministry of Public Health, for providing the dataset. Such data is the diagnosis of various diseases for preschool children collected from public hospitals in Thailand. Basically, the dataset is represented as electronic health records modeled with Thai healthcare data model standard (Ministry of Public Health, 2017). Table 1 depicts an excerpt from the dataset. Each record details a hospital admission of a patient diagnosed with multiple diseases. For example, in the first record, a patient is admitted to the hospital with *hospcode = 10660*. This patient is diagnosed with 4 diseases described by the attributes *diag1*, *diag2*, *diag3* and *diag4*. These attributes are coded using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) defined by World Health Organization: WHO (2016). For example, in the first record, *diag1 = P369* means that this patient is diagnosed with bacterial sepsis of newborn and *diag3 = Z380* is hemolytic disease of newborn. Table 2 provides more details about description of attributes for electronic health records in our dataset.

Table 1 Excerpt of our dataset

hospcode	changwat	diag1	diag2	diag3	diag4	admitdate	dischargedate
10660	Phra Nakhon Si Ayutthaya	P369	P240	P599	P012	24/12/2009	1/1/2010
10660	Phra Nakhon Si Ayutthaya	P369	P240	P143	P599	25/12/2009	1/1/2010
10691	Lopburi	P002	P700	Z380		5/3/2010	7/3/2010
10691	Lopburi	P704	P081	Z380		5/3/2010	8/3/2010
10933	Sisaket	P599	P071	Z380		12/12/2010	17/12/2010
10933	Sisaket	Z380				14/12/2010	17/12/2010
11415	Phatthalung	J00				16/8/2010	17/8/2010
11415	Phatthalung	A90				14/8/2010	18/8/2010

Table 2 Attribute description

Attribute	Description
hospcode	Hospital code
hospname	Hospital name
tambon	District of hospital
amphur	Amphoe of hospital
changwat	Province of hospital
admit_number	Patient's admission number
sex	Gender (1=Male, 2=Female)
age	Age
diag1	Diagnosis code 1 (coding with ICD-10)
diag2	Diagnosis code 2 (coding with ICD-10)
diag3	Diagnosis code 3 (coding with ICD-10)
diag4	Diagnosis code 4 (coding with ICD-10)
admitdate	Date of patient admitted
dischargedate	Date of patient discharge

3.2 Data preparation

In order to discover association rules from the dataset using data mining techniques, firstly the researchers prepared the dataset using R programming as shown in Listing 1. Concretely, the researchers select records and attributes from our dataset that are relevant to our objective, i.e., finding groups of diseases frequently appears in the dataset. To do so, firstly we read our dataset into R (line 1 in Listing 1). Thereafter, line 2 and 3 represent the transformation of attributes *admitdate* and *dischargedate* from string to date format. Thus, the researchers filtered only the records with the value of attributes *admitdate* and *dischargedate* between 01/01/2010 and 31/12/2014 and attribute *age* between 0-3 (line 4) since the researchers focused on the data of preschool children (0-3 years old) between 2010 and 2014. Finally, for each row, the researchers selected only the attributes: *diag1*, *diag2*, *diag3* and *diag4* (line 5) and removed all the records with only one disease diagnosed because the researchers aimed to discover frequent sets of diseases (line 6).

Listing 1 Data preparation using R programming

```
1  data <- read.csv(file="data")
2  data$admitdate <- as.Date(data$admitdate, format= "%d/%m/%Y")
3  data$dischargedate <- as.Date(data$dischargedate, format= "%d/%m/%Y")
4  filterData <- subset(data, admitdate >= "2010-01-01" & admitdate <= "2014-12-31" &
                        dischargedate >= "2010-01-01" & dischargedate <= "2014-12-31" &
                        age == 3)
5  filteredData <- filteredData[c("diag1", "diag2", "diag3", "diag4")]
6  filteredData <- filteredData[which(filteredData$diag2 != ''),]
```

3.3 Data grouping

In the filtered dataset, the values of attributes *diag1*, *diag2*, *diag3* and *diag4*, are coded with ICD-10 standard. In fact, this standard consists of more than 14,400 different codes to describe diseases. Therefore, the variation of diagnostic data is relatively high. The researchers reduced such variation by recoding diseases based on disease groups defined in ICD-10 Alphabetical Index of Diseases and Nature of Injury (Strategy and Planning Division, 2010) as shown in Table 3. Concretely, the researchers used R to recode the filtered data from section 3.2 as shown in Listing 2.

Table 3 Grouping of disease

Code	Diseases Groups
A-B	Certain infectious and parasitic diseases
C-D	Neoplasms
D	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E	Endocrine, nutritional and metabolic diseases
F	Mental and behavioural disorders
G	Diseases of the nervous system
H	Diseases of the eye and adnexa
I	Diseases of the circulatory system
J	Diseases of the respiratory system
K	Diseases of the digestive system
L	Diseases of the skin and subcutaneous tissue
M	Diseases of the musculoskeletal system and connective tissue
N	Diseases of the genitourinary system
O	Pregnancy, childbirth and the puerperium
P	Certain conditions originating in the perinatal period
Q	Congenital malformations, deformations and chromosomal abnormalities
R	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
S-T	Injury, poisoning and certain other consequences of external causes
U	Code for special purposes
V-Y	External causes of morbidity and mortality
Z	Factors influencing health status and contact with health services

Listing 2 Excerpt of data grouping using R programming

```
>filteredData$diag1 <- ifelse(grepl('A|B',filteredData$diag1),'A-B',
            ifelse(grepl('C|D[1-4]',filteredData$diag1),'C-D',
            ifelse(grepl('D[5-8]', filteredData$diag1), 'D',
            ifelse(grepl('E', filteredData$diag1), 'E',
            ifelse(grepl('F', filteredData$diag1), 'F',
            ifelse(grepl('G', filteredData$diag1), 'G',
            ifelse(grepl('H', filteredData$diag1), 'H',
            ...
            ...
            ifelse(grepl('V|Y', filteredData$diag1), 'V-Y',
            ifelse(grepl('Z', filteredData$diag1), 'Z',
            ''))))))))))))))))))))
```

3.4 Association rules mining

The researchers used the Apriori algorithm to discover association rules from the dataset, i.e., group of diseases frequently appear in preschool students in Thailand. The Apriori algorithm (Srikant and Agrawal, 1995) is an algorithm for mining association rules which present general trends in the dataset. Such an algorithm is designed to perform on transactional database, e.g., collections of products bought by clients and health records of patients diagnosed with multiple diseases in our case. Each transaction represents a set of items, so called item set. Thereafter, the Apriori algorithm constructed such item sets having supported values more than a given minimum support threshold. Listing 3 describes how we apply the Apriori algorithm to the dataset using R programming. Concretely, the researchers loaded the pre-

processed data from section 3.3 into R (line 2). Thus, the researchers used the function *apriori* defined in *arules* library to apply the Apriori algorithm to the dataset. The result returns as item set. Hence, the researchers used the inspect function (line 4) to view the result as item sets (line 5) which will be discussed in the next section.

Listing 2 Excerpt of association rules mining using R programming

```
1  library('arules')
2  TrnsData <- read.transactions("filteredData", sep = ",")
3  itemsets <- apriori(TrnsData, parameter=list(minlen=2,support=0.01,
                                                confidence=0.01))
4  inspect(sort(itemsets, by="lift"))
5  [1]  {E}     => {A-B} 0.11097632 0.68574285 3.1436061
   [2]  {A-B}   => {E}   0.11097632 0.50874146 3.1436061
   [3]  {K}     => {E}   0.01301994 0.37141813 2.2950602
   [4]  {E}     => {K}   0.01301994 0.08045260 2.2950602
   [5]  {R}     => {J}   0.05664044 0.57224075 2.2721177
   [6]  {J}     => {R}   0.05664044 0.22489441 2.2721177
   [7]  {P}     => {Z}   0.34793024 0.85455892 1.8125156
   [8]  {Z}     => {P}   0.34793024 0.73795847 1.8125156
   [9]  {J}     => {D}   0.03265840 0.12967221 1.6694269
   [10] {D}     => {J}   0.03265840 0.42045097 1.6694269
   ...
```

4. Results and Discussion

Based on the dataset obtained from the Bureau of Policy and Strategy, Ministry of Public Health, the researchers can discover the general trends of disease rates and groups of diseases frequently found in preschool children in Thailand. The results can be summarized as follows:

Figure 2 shows the number of patients by age between 2010 and 2014. The highest number of patients aged less than 1 year old followed by 1, 2, and 3 years old respectively.

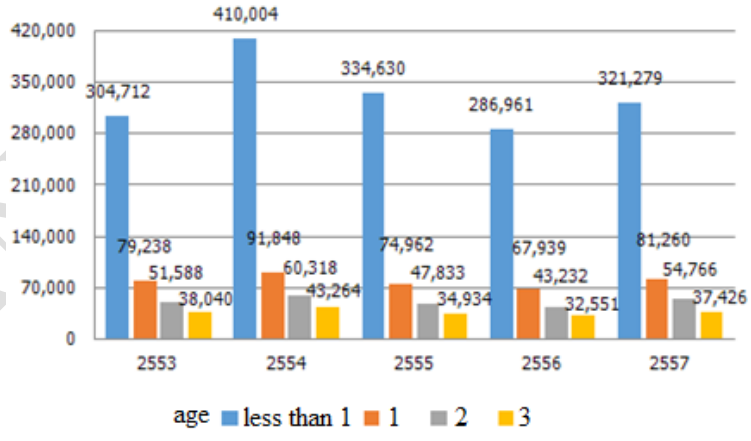

Figure 2 Number of patients by age between 2010 and 2014

Figure 3 depicts the occurrence of disease group found in preschool children between 2010 and 2014. The highest is the group *P* which is certain conditions originating in the perinatal period, followed by group *J* diseases of the respiratory system, and group *E* endocrine, nutritional and metabolic diseases.

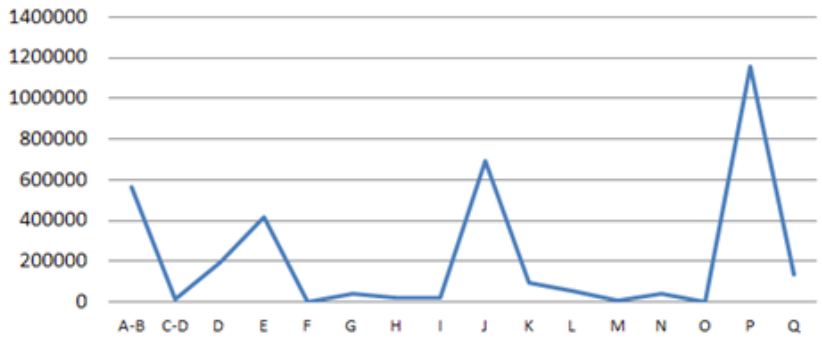

Figure 3 The occurrence of disease group found in preschool children between 2010 and 2014

The researchers adopted the Apriori algorithm to mine association rules between groups of diseases from our data set. Such rules represent disease groups that are likely to occur in preschool children in Thailand during 2010-2014. Table 4 shows the association rules between the diseases groups A-B to Q. We define association rules in the form of $LHS \rightarrow RHS$, where left-hand-side (LHS) and right-hand-side (RHS) are two subsets of disease groups. The researchers measured the significance of each rule using 3 metrics (1) support: the frequency of disease groups LHS and RHS ($LHS \cup RHS$) occur together with respect to all records (R) in our dataset; (2) confidence: the probability that the rule is correct; and (3) lift: the ratio of dependent between LHS and RHS as described below:

$$support(X) = \frac{|\{X \in R\}|}{|R|}$$

$$confidence(LHS \rightarrow RHS) = \frac{support(LHS \cup RHS)}{support(LHS)}$$

$$lift(LHS \rightarrow RHS) = \frac{support(LHS \cup RHS)}{support(LHS) \times support(RHS)}$$

If the lift value is greater than 1, it would imply that the LHS and RHS are dependent of each other. The results show that disease group E and A-B are highly dependent by having the highest lift, followed by K and E, R and J, P and Z, and J and D, respectively. Figures 4 and 5 are alternative representations of the association rules visualized by R programming.

Table 3 Association rules for various disease groups

LHR	RHS	Support	Confidence	Lift
E	A-B	0.11	0.69	3.14
A-B	E	0.11	0.51	3.14
K	E	0.13	0.37	2.30
E	K	0.13	0.08	2.30
R	J	0.57	0.57	2.27
J	R	0.57	0.22	2.27
P	Z	0.35	0.85	1.80
Z	P	0.35	0.74	1.80
J	D	0.03	0.13	1.67
D	J	0.03	0.42	1.67

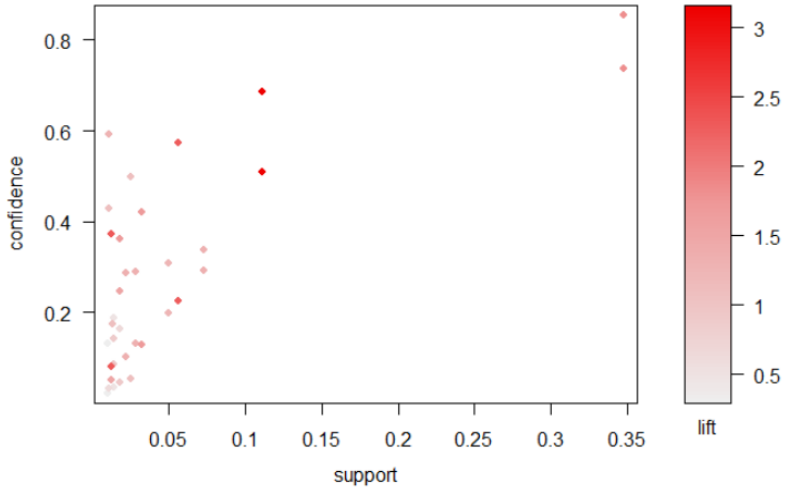

Figure 4 The Association rules visualized as scatter plot

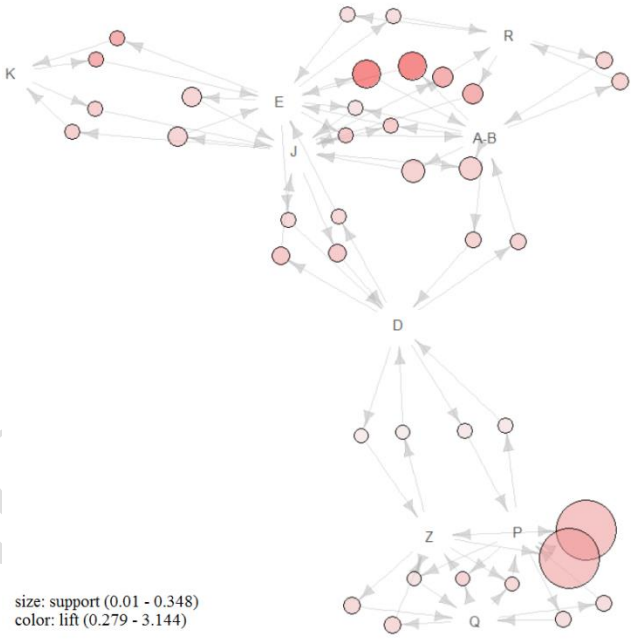

Figure 4 The Association rules visualized as network

6. Conclusion

This work presents the results of a study on the association between disease groups by the occurrence in preschool children in Thailand using R programming. To the best of our knowledge, existing work relating to data analytics in healthcare do not use data mining technique on a national healthcare dataset in order to discover association between diseases grouped by ICD-10 standard. Our results show that the disease group having the highest rate of occurrence in preschool children is group *P* and the certain conditions originating in the perinatal period. Such disease group includes fetus and newborn affected by maternal factors and by complications of pregnancy, labour and delivery, birth trauma, and so on. Furthermore, we discover association rules from our data set showing that the disease group *P*, i.e.,

endocrine, nutritional and metabolic diseases, is likely to occur with the disease group *A-B*, i.e., certain infectious and parasitic diseases. We also found that children under 1 years of age have the highest chance of developing the disease which corresponds to the immune system in the age range of newborn baby. We believe that such knowledge that was discovered from the dataset could be used to improve the healthcare policy in order to have a better quality of life for preschool children.

For future studies, the researchers want to consider other factors for mining association rules in healthcare data, such as residential data, environmental data, temperature, and seasonal data.

7. Acknowledgements

The researchers appreciated the kind support of the Bureau of Policy and Strategy, Ministry of Public Health, for providing the dataset of patients diagnosed with a variety of diseases from public hospitals in Thailand during the year 2010-2014.

8. References

Takeuchi, H., Kodama, N., Hashiguchi, T., & Hayashi, D. (2006). Automated healthcare data mining based on a personal dynamic healthcare system. *Proceedings of the 28th IEEE Annual International Conference*, 3604-3607.

Gosain, A., & Kumar, A. (2009). Analysis of health care data using different data mining techniques. *Proceeds of International Conference on Intelligent Agent & Multi-Agent Systems*, 1-6.

Shouman, M., Turner, T., & Stocker, R. (2012). Using data mining techniques in heart disease diagnosis and treatment. *Proceedings of the Japan-Egypt Conference on the Electronics, Communications and Computers*, 173-177.

Balasubramanian, T., & Umarani, R. (2012). An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. *Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering*, 370-375.

Batra, S., Sachdeva, S., Mehndiratta, P., & Jyotsana H. (2014). Mining standardized semantic interoperable electronic healthcare records, biomedical informatics and technology, *Communications in Computer and Information Science*, *404*, 179-193.

Nuaimi, N. A. (2014). Data mining approaches for predicting demand for healthcare services in Abu Dhabi. *Proceedings of the 10th International Conference on Innovations in Information Technology*, 42-47.

Srikant, R., & Agrawal, R. (1995). Mining generalized association rules. *Proceedings of the 21th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.*, 407-419.

Srikant, R., & Agrawal, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*, 487-499.

Mungkornnitra, K. *(2009). Patterns and factors related to disease*. Retrieved from https://www.gotoknow.org/posts/271990

Ministry of Public Health (2017). *Thai healthcare data model standard references*. Retrieved from https://www.moph.go.th/

Strategy and Planning Division (2012). *Alphabetical Index of Diseases and Nature of Injury, ICD-10-TM for PCU (International Classification for Primary Health Care Unit)*. Retrieved from http://www.hiso.or.th/

World Health Organization: WHO (2016). *International Classification of Diseases and Related Health Problems (ICD)*. Retrieved from http://www.who.int/classifications/icd/en/