Application of Data Mining in Healthcare

Application of Data Mining in Healthcare

ALWIS NAZIR

Medical Information Science

United Graduate School of Drug Discovery and

Medical Information Science

Gifu University

JAPAN

September 2014

Contents

Summary

In Japan, medical expenditure is one of the biggest expenses in national financial expenditure. One way to reduce this expenditure is to improve the health condition of community. The purpose of this study was to detect the symptoms of a disease that can prevent the disease becoming more severe. The expected output of this research was to establish the methods for an early detection of severe events for preventive medical treatment. Better health conditions provided by preventive health care can increase work productivity and reduce medical expenditure. In this research we applied data mining technique to process the big data from healthcare.

Data mining is the process to mine a series of valuable information from a database. These information are obtained by extracting and recognizing the patterns from data contained in the database. Data mining is very useful to get information from huge datasets. It seems natural that we can get more good information as the size of dataset increases. However, large dataset often include strong noise, unexpected errors, and complex structure, which make the analysis difficult. Data mining has been developed to overcome these difficulties.

One place that data mining will be useful is healthcare. The data such as the reports of adverse effects caused by drugs, and records of annual health checkups are

very large, but very rarely processed to obtain useful information. The adverse event report, if we processed properly, will give us very useful information such as the relationship between an adverse events with a symptom of a disease. Likewise, the annual health checkups of data, when processed properly, will be very helpful for the prediction of a person's health in the future. In my research, I focused on two data, the adverse event reports and annual health checkup data, and applied two methods of data mining, the network analysis and hidden Markov models. The purpose of this research is

The aim of the first research was to identify the symptoms that would suggest a high suicide risk of depression. To achieve this task, we performed the network analysis of the data obtained from the US Food and Drug Administration Adverse Event Reporting System (FAERS) of selective-serotonin reuptake inhibitors (SSRI). Using FAERS reports from 1997 to the second quarter of 2012, we constructed the co-occurrence network of adverse events. From this network, we extracted the events that were strongly connected to suicidal events (suicidal attempts, suicidal ideation, suicidal behavior, and complete suicide) by means of the community detection method. Using this method, we succeeded in obtaining a list of suicide-related adverse events. Owing to the randomness inherent in the algorithms of community detection, we found that the obtained list differed according to each trial of analysis. However, the lists we derived show considerable efficiency in identifying suicidal events. The network analysis appears to be a promising method for

identifying signals of suicide.

The aim of our second research is to find out the probability of change in the health risks based on annual time series data of health checkup and to determine the level of risk and the progression in health conditions especially for persons with hypertension. For this purpose, we made hidden Markov model analysis of the health checkup data between 2002 and 2007 which include 912,765 records from 279,904 participants, provided by the medical center in Gifu prefecture, Japan. From this dataset, we extracted the data of people with hypertension, i.e. systolic blood pressure (SBP) above 140 mmHg or diastolic blood pressure (DBP) values above 90 mmHg. For the person with hypertension who have a 4-6 year time series of data, we carried out the hidden Markov model analysis using the following test values: the body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), hematocrit (HCT), platelets (PLt), glutamic oxaloacetic transaminase (GOT), glutamic pyruvic transaminase (GPT), total cholesterol (T.Chol), neutral fat (NF) and blood sugar (BS). Because the health condition is strongly dependent on age, we divide the data in several age groups, namely 30's, 40's, 50's and 60's, and carried out the analysis for each group. We evaluated the obtained Markov model by comparison with the classification by professional medical staff. We succeeded to cluster the data in 6 groups. Almost in all states in each age group, the average value of BMI, T.Chol and NF are out of the normal

range. In age group 30's, the model has 4 different levels of risk. In age group 40's, 50's and 60's, the model has 3, 2, and 2 different levels of risk, respectively. Taking the transition probability into account, we found the risk in the future may differs even if the current risk is same. Using hidden Markov model we succeeded to find out the probability of change in the health risks based on annual time series data of health checkup.

From these results, we conclude that data mining methods such as the network analysis and hidden Markov model are useful to process the data contained in the healthcare centers and to derive the valuable information. We can develop an early detection of symptom strongly related to suicide event and we also can develop an early detection symptom for becoming hypertension. Data mining is promising to develop the method of detection of severe events in their early stage, which leads to the reduction of medical expenditure.

Chapter 1

Introduction

Every year, medical expenditure has increased. There are many research about factors causing this increase. For example, Julie Appleby said one of the factors that lead to increased medical expenditure is people growing older, sicker and fatter [1]. Therefore it is very important to maintain the health condition especially for the elder people. The purpose of this study was to detect the symptoms of a disease that can prevent the disease becoming more severe. The expected output of this research is an early detection of severe events. Better health conditions provided by preventive health care can increase work productivity and reduce medical expenditure. The data of healthcare, which has a symptoms of disease data, is so huge that it needs to be processed using techniques from data mining

1.1. Principle of Data Mining

Data mining is the process to mine a series of valuable information from a database. These information are obtained by extracting and recognizing the patterns from data contained in the database. Data mining is very useful to get information from huge datasets. It seems natural that we can get more good information as the size of dataset increases. However, large dataset often include strong noise, unexpected errors, and

complex structure, which make the analysis difficult. Data mining has been developed to overcome these difficulties.

Many terms are used to demonstrate the process of data mining (eq. knowledge discovery, knowledge extraction, data / pattern analysis, data archeology, the data dredging, information harvesting, business intelligence, etc.). In order to provide an understanding of data mining, here are some facts [2]:

- Many organizations, the business and government need to deal with a number of resources and also with the data base management, and it is inevitable to develop the large-scale data warehouse.
- And sometimes the data cannot be directly analyzed using the standard statistical methods, because there are some missing records or the data in a qualitative measure, not in a quantitative measure.

In the book "Advances in Knowledge Discovery and Data Mining" data mining is defined as follows: “Knowledge discovery (data mining) in databases (KDD) is a non-trivial entire process to search and identify a pattern in the data, where patterns are found to be valid, novel, potentially useful, ultimately understandable” [3].

The term data mining and knowledge discovery in databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. Actually, the two words have a different concept but relate to each another. And

one of the stages in the overall KDD process is data mining. KDD process in general can be described as follows [3]:

1. Data Selection

   Selection of data from the data set needs to be done before the excavation phase begins. The results are stored in a file, separate from the operational database.

2. Pre-processing/ Cleaning

   Cleaning processes include discard of duplicate data, check of the data that is inconsistent, and correction of errors in the data, such as printing errors (typography). Also they include enrichment process, i. e. the process of "enriching" the existing data with the data or other information which is relevant and necessary to KDD.

3. Transformation

   Coding is the process of transformation for the data that has been selected, so that the data are suitable for data mining process. Coding in the KDD process is a creative process and is highly dependent on the type or pattern of information to be searched in the database

4. Data Mining

   Data mining is the process of searching for a pattern or interesting information in the selected data using some techniques or methods. Techniques, methods, or algorithms in data mining are highly variable. Selection of appropriate methods or algorithms are

strongly dependent on the destination and the overall KDD process.

5. Interpretation/ Evaluation

   Pattern of information generated from data mining process needs to be displayed in a form that is easily understood by interested parties. This phase includes checking whether a pattern or information found contrary to facts or hypotheses that existed before.

KDD process basically have 5 stages as described previously. However, in the real case, iteration or repetition sometimes occurs at some stage. At each stage, the analyst can be returned to the previous stage. For example, at the time of coding or data mining, the analyst realizes the cleaning process is not done perfectly, or maybe the analyst finds new data or new information to "enrich" the data that already exists. KDD covers the entire process of finding patterns or information in the database, starting from the selection and preparation of the data until the representation of the pattern found in a form that is easily understood by interested parties. Meanwhile data mining is one component in the KDD focused on extracting hidden patterns in the data base.

Here I represent three of most popular data mining analysis technique:

1. Association Rule Mining

   Association rule mining is a mining technique to find an associative rule between item combinations. Examples: the analysis of purchases in a supermarket gives us

information how likely a customer buys a bread along with milk. With this knowledge, the supermarket owner can adjust the placement of the goods or designing a marketing campaign using a combination of discount coupons for certain goods. The most popular algorithm known as a priori generates and tests paradigm, namely it manufactures candidates of the combination of items that might be based on certain rules and then tests whether the combination of items will meet the minimum support requirement. This combination is called frequent item set, which will be used to create rules which qualify the minimum confidence [4]. More efficient new algorithm such as FP-Tree has been developed [5].

2. Classification

Classification is the process of finding a model or a function that describe or distinguish the concept or class of data, with the aim to be able to predict the class from object with an unknown label. The model itself can be the rule "if-then" in the form of a decision tree, neural network, or mathematical formulas. Decision tree is one of the most popular classification methods because it is easy to be interpreted by humans. Every branching is represented as the conditions that must be met and every end of the tree is represented as data class. The most well-known decision tree algorithm is C4.5 [6], but lately has been developed an algorithm that is capable of handling large-scale data that cannot fit in main memory as Rainforest [7]. Other method is Bayesian,

neural networks, genetic algorithms, fuzzy, case-based reasoning, and k-nearest neighbor.

Classification process is usually divided into two phases: learning and test. In the learning phase, most of the data that has been known is fed to create a model of estimation. Then, in the test phase of the model that has been formed, the model is tested with most other data to determine the accuracy of the model page. When the accuracy is enough, this model can be used to predict unknown data class.

3. Clustering

Different from classification which the class has been defined previously, clustering is grouping data without base on specific data class. Even clustering can be used to give label at data class which is not yet known. That why the clustering is often classified as unsupervised learning method.

The principle of clustering is to maximize the similarity between members of the class and minimizing the similarity inter-class / cluster. Clustering can be performed on the data that has some attributes which are mapped as multidimensional space. One of famous method in clustering section is K-Mean method. K-Means is a method that make the modeling process without supervision (unsupervised) and perform grouping of data with the system partition. This method tried to clustering the data into several groups, where the data in the group have a similarity characteristics and have different

characteristics from the data that is in the other group. It means, this method seeks to minimize variations among the data is in same cluster and maximize the variation with the data in other clusters.

Though these methods can be applied in general datasets, there are some other mining methods when dataset has some special structure. One example of method is network community detection, applicable when dataset is represented as "network". Network analysis has been extensively adopted in research into social systems, biological processes, and computer systems [8]. In network analysis, a complex system is modeled using a network—the set of nodes and the set of edges that connect those nodes. For example, in modeling a social network, such as sexual contact, each individual is modeled as a node and edges are made among the people who have sexual contacts [9].

Hidden Markov Model (HMM) is another algorithm used for both classification and clustering, available when we analyze time-series dataset. HMM is a non-supervised learning method, especially popular for voice/speech recognition. Hidden Markov Models (HMM) can be used as a base foundation to create the probabilistic models of linear sequence [10, 11]. HMM is a statistical model of a system that is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters (state) from the observed data. HMM is said to be “a *full probabilistic model*—the model parameters and the overall sequence 'scores' are all probabilities"[12].

1.2. Data mining for Healthcare

One place that data mining will be useful is healthcare. The meaning of healthcare is the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by the medical and allied health professions. Various kinds of data can be found in healthcare such as the number of visits of the patient, the type of disease experienced by patients, and so on. The data such as the reports of adverse effects caused by drugs, and records of annual health checkups are very large, but very rarely processed to obtain useful information. In this section, I briefly introduce medical datasets that is involved with my research.

1.2.1. Adverse Event Report

Adverse events are unwanted and usually harmful outcomes. The event may or may not be related to the treatment, and is not the same as a side effect or an adverse reaction because it is not always clear whether the drug has caused the event [13].

As the premarketing survey of adverse events, clinical trials usually record all adverse events that occur during the trial, to help determine which ones might be associated with the medicine. On the other hand, spontaneous adverse event report play an important role for the post-marketing surveillance of adverse events. One of the biggest institutions that handle adverse event reporting is the U.S. Food and Drug Administration (FDA). FDA

makes a system called the FDA Adverse Event Reporting System (FAERS). Content of this database is information on adverse event and medication error report which has been submitted to FDA spontaneously. FDA designing this database to support post-marketing safety surveillance program for drug and therapeutic biologic products. Adverse events and medication errors are coded to terms in the Medical Dictionary for Regulatory Activities (MedDRA) terminology [14]. Several research using FAERS have been performed by Eriksson *et al* [15], try to analyze techniques for temporal data mining of electric patient records and then using the techniques to detect adverse drug reactions in a patient- and dose-specific manner. In other research, DuMouchel *et al* [16] try to use a disproportionality design as method to analyze an adverse drug reaction risk identification system using adverse event reports. There are other researches in pharmacovigilance analysis, on which we only show references [17-19].

1.2.2. Annual Health Checkup

An annual evaluation of a person's health status, which includes a physical exam and routine screening tests, help us to insure continued health or to identify early and often treatable stages of a disease [20]. According to the American Medical Association (AMA), there are five important reasons to have an annual exam [21]:

1. Screenings

   If we do screening test, we can detect early for disease as heart disease or cancer. For women, the most popular screening test is Pap smear test and breast exam (mammogram).

2. Health Measurements

   We must check out blood pressure and heart rate periodically to anticipating if we have abnormality in our body. If blood pressure this year has increase compared on last year, we should consult to physician to make sure why this happen. If our weight increase rapidly, we must be aware about obesity. Periodical checking of the health measure can reduce the morbidity and mortality risk.

3. Counseling

   By doing annual health checkup, we have a chance to discuss about our lifestyle with physician. We can improve health condition by suggestion from physician.

4. Medical Records Update

   Of course, by doing annual health checkup, all data in medical record will be updated with new data from annual health checkup result. It will help us if we have an accident or have a serious illness. The paramedic or physician can refer from the last status at medical record.

5. Alleviate Worries

If we know your health condition is good, we don't worry about medical expenses. We just need to keep the good lifestyle and our stress level will be reduced.

Likewise other datasets, the data of annual health checkups, when processed properly, will be very helpful for the prediction of a person's health in the future. Research about the health check-up has been started long ago. Martha Crumpton Hardy conducted research in 1948 on health checkup at a group of children in Chicago [22]. Another research was performed by Krogsbøll to quantify the benefit and harms of general checkup. He made emphasis on the value of morbidity and mortality [23]. Also several researcher from Japan conducted researches using result of annual health checkup [24-25].

1.3. Summary of My Thesis

In my research, I focused on two data, the adverse event reports and annual health checkup data, and applied two methods of data mining, the network analysis and hidden Markov models.

1.3.1. Network Analysis of Adverse Event

The aim of the first research was to identify the symptoms that would suggest a high suicide risk of depression. To achieve this task, we applied the network analysis of

the data obtained from FAERS of selective-serotonin reuptake inhibitors. Using FAERS reports from 1997 to the second quarter of 2012, we constructed the co-occurrence network of adverse events. From this network, we extracted the events that were strongly connected to suicidal events (suicidal attempts, suicidal ideation, suicidal behavior, and complete suicide) by means of the community detection method. Using this method, we succeeded in obtaining a list of suicide-related adverse events. Owing to the randomness inherent in the algorithms of community detection, we found that the obtained list differed according to each trial of analysis. However, the lists we derived show considerable efficiency in identifying suicidal events. The network analysis appears to be a promising method for identifying signals of suicide.

1.3.2. HMM analysis of Health Checkup

The aim of our second research is to find out the probability of change in the health risks based on annual time series data of health checkup and to determine the level of risk and the progression in health conditions especially for persons with hypertension. For this purpose, we made hidden Markov model analysis of the health checkup data between 2002 and 2007 which include 912,765 records from 279,904 participants, provided by the medical center in Gifu prefecture, Japan. From this dataset, we extracted the data of people with hypertension, i.e. systolic blood pressure (SBP) above 140 mmHg or diastolic blood pressure (DBP) values above 90 mmHg. For the person with hypertension who

have a 4-6 year time series of data, we carried out the hidden Markov model analysis using the following test values: the body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP), hematocrit (HCT), platelets (PLt), glutamic oxaloacetic transaminase (GOT), glutamic pyruvic transaminase (GPT), total cholesterol (T.Chol), neutral fat (NF) and blood sugar (BS). Because the health condition is strongly dependent on age, we divide the data in several age groups, namely 30's, 40's, 50's and 60's, and carried out the analysis for each group. We evaluated the obtained Markov model by comparison with the classification by a professional medical staff. We succeeded to cluster the data in 6 groups. Almost in all states in each age group, the average value of BMI, T.Chol and NF are out of the normal range. In age group 30's, the model has 4 different levels of risk. In age group 40's, 50’s and 60’s, the model has 3, 2, and 2 different levels of risk, respectively. Taking the transition probability into account, we found the risk in the future may differs even if the current risk is same. Using hidden Markov model we succeeded to find out the probability of change in the health risks based on annual time series data of health checkup.

From these results, we conclude that data mining methods such as the network analysis and hidden Markov model is useful to process the data contained in the healthcare centers and to derive the valuable information.

Chapter 2

Identification of suicide-related events through network analysis of adverse event reports

2.1. Introduction

Suicide is a critical symptom of depression [26]. To reduce suicide deaths, it is essential to estimate accurately the suicide risk of patients and to give proper treatment. Researchers have extensively investigated suicide risk factors, such as acute mood episodes, personal history, and family history [27, 28]. However, more reliable medical signs that indicate suicidal risk are required.

The aim of this study was to identify the symptoms strongly related to suicide. To this end, we analyzed adverse event reports on selective serotonin reuptake inhibitors (SSRIs). SSRIs are the most widely used antidepressant worldwide. They are believed to decrease the risk of suicide [29]; however, even if depressive patients take SSRIs properly, their suicide risk is still greater than that of individuals without depression. Among the reported adverse events of SSRIs are suicidal ideation, suicidal attempt, suicidal behavior, and complete suicide. These adverse events related to suicide provide useful information toward estimating suicide risk.

In this study, we attempted to compile a list of the adverse events that are strongly related to suicide. For this purpose, we analyzed the reports in the US Food and Drug

Association (FDA) Adverse Event Reporting System (FAERS). FAERS is the spontaneous reporting system of adverse events and is widely used in pharmacovigilance analysis [17–19]. To analyze FAERS data, we employed the method of community detection in networks.

Network analysis has been extensively adopted in research into social systems, biological processes, and computer systems [30]. In network analysis, a complex system is modeled using a network—the set of nodes and the set of edges that connect those nodes. For example, in modeling a social network, such as sexual contact, each individual is modeled as a node and edges are made among the people who have sexual contacts [9]. In network analysis of a gene regulatory network, each gene is taken as a node and an edge is placed between two genes if one gene regulates the other.

The detailed analysis of such networks provides considerable interesting information. For example, it has been demonstrated that social networks usually have a small number of important nodes called hubs [9]. This property plays an important role in strategic vaccination because the spreading of a virus may be effectively prevented by focusing vaccination efforts on such hubs [31]. From an analysis of the gene regulatory network of *Escherichia coli* and a yeast species, it was shown that gene regulatory networks have several patterns, called network motifs [32]. These motifs display some particular functions, such as creating pulse when the environment changes. Network

analysis also has promising applications in medicine. Barabási proposed a network called the diseasome, in which the nodes are diseases and the edges are the connections between two diseases [33]. He believed that this network was useful in predicting the progress of diseases and in identifying the common causes of different diseases.

In the present study, we constructed the network of symptoms reported as adverse events in FAERS and extracted the symptoms that are strongly correlated with suicidal symptoms. Compared with other data mining methods, such as association analysis and Bayesian analysis, network analysis presents several advantages and disadvantages. The major advantage with network analysis is that it can analyze indirect connections. Because it addresses the whole network structure, it can capture hidden relations that would fail to be recognized by association analysis or Bayesian analysis, which deal with the correlation among a restricted number of objects. However, the method of network analysis is still under development. There is no established method for analyzing a network, and many new methods are proposed every year. With FAERS data, several studies have applied association analysis or Bayesian analysis [34, 35], but none have been based on network analysis.

In the present study, we constructed a symptom network of adverse events. We conducted modularity-based community detection [36] to extract the list of suicide-related adverse events that were believed to have a strong connection with suicidal events:

suicidal behavior, suicidal ideation, suicidal attempts, and complete suicide. We investigated the relationship between suicide risk and those events.

2.2. Methods

2.2.1. Data Preparation

We obtained the original dataset of FAERS containing the reports received by the FDA from 1997 to the second quarter of 2012. From this dataset, we extracted reports of adverse events in which SSRI involvement was suspected. Each report in FAERS is composed of a unique identification number, a list of drugs that the patient took, and a list of patient reactions. The names of reactions are coded by the preferred term (PT) defined by the Medical Dictionary for Regulatory Activities (MedDRA®). Though the terms defined by MedDRA® depend on the version employed, we neglected such differences in our analysis: since most PTs did not change between 1997 and 2012, we believed that these differences would not lead to severe errors in our study. However, the names of drugs are not standardized in FAERS: drugs may appear under their generic or proprietary names. For the standardization of drug names, we used American Hospital Formulary Service (AHFS) “Drug Information 2010” [37] and DrugBank [38]. DrugBank was employed to link the generic and proprietary names of each drug; AHFS drug information was used to combine the generic drug name with its function, such as “SSRI.”

Each FAERS report includes such details as patient demographic information,

patient outcome, and prime suspect for the adverse event. From those data, we extracted reports related to SSRIs as follows. First, we selected reports in which only a single drug was used and the drug was citalopram, escitalopram, fluoxetine, paroxetine, or sertraline. From this list, we selected those reports in which SSRIs were regarded as the prime suspect. In this way, we produced a list that contained the report identification number and the patient reactions. A schema of these procedures is presented in Fig. 2.1.

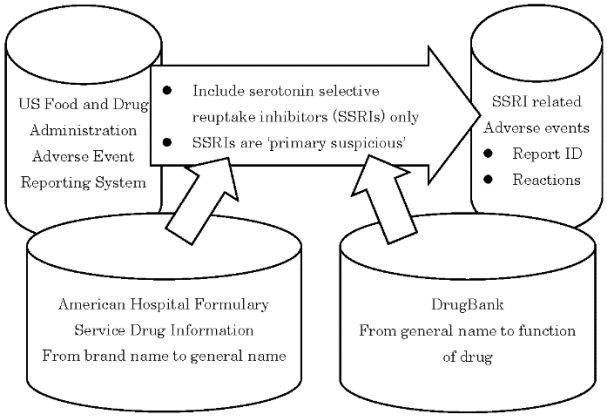

Fig. 2.1. Construction of databases of adverse events of selective serotonin reuptake inhibitors

The resulting reports were randomly divided into two groups. Group 1 was used to construct the adverse event networks and compile a list of suicide-related adverse events. Group 2 was used for quantitative evaluation of this list. The number of reports appears in Table 2.1.

Group ID	Number of reports	Reports including suicidal events
#1	21497	2697
#2	21497	2652

Table 2.1. Summary of the reports used in the analysis

2.2.2. Network Analysis

From the data obtained with the procedures described in section 2.1, we constructed the network of adverse events in the same manner as Barabási constructed the diseasome network [33]. With our network, the reactions were represented as nodes; the edge between nodes represented the co-occurrence of adverse events, whose weight was given by Pearson's correlation $\sigma_{\mathrm{AB},}$, defined by

$$\sigma_{\mathrm{AB}} = (N\, n_{AB} - n_A n_B)/\sqrt{n_A n_B (N - n_A)(N - n_B)}\,,$$

where n_A, n_B, n_{AB}, and N represent the numbers of reports that indicate adverse event A, that indicate adverse event B, that report the adverse events of both A and B, and the total number of reports, respectively. σ_{AB} can take both positive and negative values, but we made an edge only when σ_{AB} was positive; this was because community detection fails if the edge weight is negative. This process is illustrated in Fig. 2.2. At this stage, we removed isolated nodes, i. e. adverse events that have no positive correlation with others.

Report ID	Adverse events
1	A, B
2	A,B,C
3	C,D
4	A,C
…	…

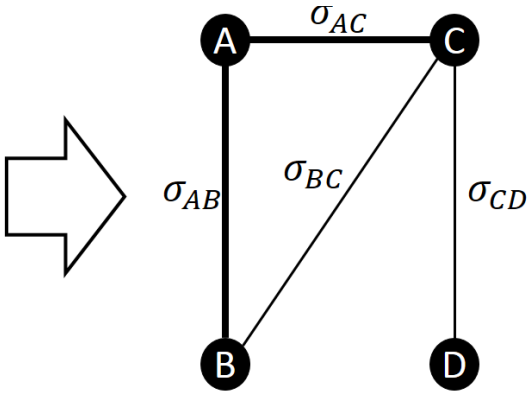

Fig. 2.2. Construction of adverse-event network

Using modularity-based community detection, the obtained network was divided into small groups called communities [36]. We derived the list of suicide-related adverse events, which were those events that were in the same community as suicidal events.

In modularity-based community detection, the best partitioning of the network is that with maximal modularity. Modularity is defined using the following equation:

$$\mathrm{Q} = \frac{1}{2m}\sum_{ij}[\sigma_{ij} - \frac{k_i k_j}{2m}]\delta(c_i, c_j),$$

where σ_{ij} is the weight of the edge between node i and j, $k_i = \sum_l \sigma_{il}$,$k_{\mathrm{j}} = \sum_l \sigma_{jl}$, $m = \frac{1}{2}\sum_{ij}\sigma_{ij}$, and $\delta(c_i, c_j)$ is 1 if node i and j belong to the same community and 0 otherwise. From an intuitive perspective, modularity is the characteristic that indicates the average density of the internal edges in each community. To understand the meaning of modularity more clearly, consider two nodes, i and *j*, in the same community. If there is an edge between these nodes, $\sigma_{ij} - \frac{k_i k_j}{2m}$ is positive provided that σ_{ij} is sufficiently large. Conversely, if there is no edge between them, $\sigma_{ij} - \frac{k_i k_j}{2m} = -\frac{k_i k_j}{2m}$ is always negative. Consider, for example, the network represented in Fig. 2.3(a), where for simplicity the setting is $\sigma_{ij} = 1$. In the partitioning depicted in Fig. 2.3(b), nodes B, C, and D are members of the same community, whereas node A is not. In this case, we do not sum the pairs of nodes (A, B), (A, C), and (A, D) in calculating modularity because A belongs to a different community from the others. However, if we divide the network as in Fig. 2.3(c), these three pairs provide contributions to the modularity. The existence of edges (A, B)

and (A, D) increases the modularity, though the absence of edge (A, C) decreases it. In Fig. 2.3, the positive contribution from the existence of the edges exceeds the negative one from the absence of one edge. This implies that nodes B and C intermediate the connection between A and C so strongly that they compensate for the absence of a direct connection. Therefore, we would conclude that nodes A and C are in the same cluster.

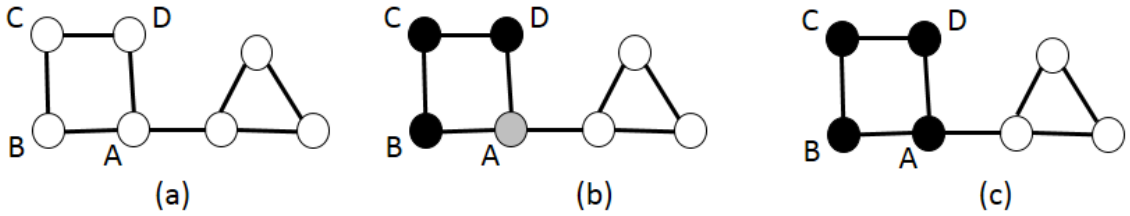

Fig. 2.3. Modularity-based partitioning of the network (a). Partitioning (b; Q=13/32) gives lower modularity (c; Q=33/64).

Two points should be noted here. First, high computational power is necessary to obtain the best partitioning with a large network. To overcome this difficulty, several methods have been proposed [39–42]. In the present study, we applied the algorithm developed by Blondel et al. [42]. Our approach consisted of two phases. In the first phase, we decomposed the network into many small communities. Then, we aggregated the nodes within the same community and constructed a new network, in which the nodes were the small communities obtained in the first phase. Repeating these two processes iteratively, we obtained the community structure of networks. This algorithm is sufficiently fast for application, though it is stochastic: we obtained a slightly different result with each trial of community detection. In section 2.3, we present the results of 20 community

detection trials.

Second, in modularity-based community detection, we often encounter the problem of limit of resolution: we fail to identify small communities [43]. This problem may be partially solved by slightly modifying the definition of modularity [44]. In this approach, we employ the tunable parameter of resolution, which controls the accuracy of community detection. The exact definition of the resolution is provided by Lambiotte et al. [45]. In our analysis, we set the resolution parameter as 1.0. These analyses were conducted using Gephi [46]. The process is depicted in Fig. 2.4.

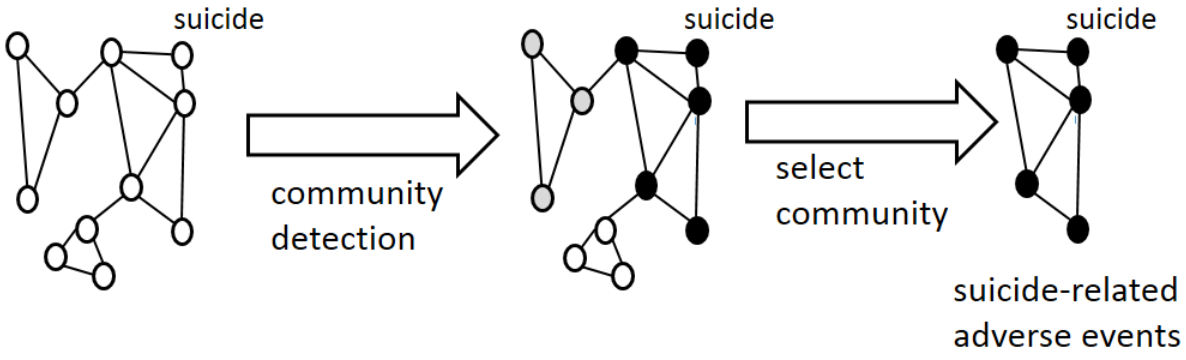

Fig. 2.4. Process of extracting suicide-related adverse events

2.2.3. Evaluation of Suicide-Related Adverse Events

There may be more than one suicide-related adverse event, and it is natural to assume that a patient has a higher suicide risk if they have more suicide-related adverse events. Therefore k, the number of suicide-related adverse events of a patient, can be taken as the control parameter in risk evaluation. We estimated the risk as follows. First, we determined the proportion of patients who had suicidal events and k (suicide-related

adverse events). Second, we plotted the receiver-operating characteristic (ROC) when k varied between 1 and 50. The ROC reflects the effectiveness of suicide-related adverse events as indicators.

These evaluations were conducted with the data for Group 2. Those data were not used for the first network analysis.

2.3. Results

The network of adverse events of SSRIs had 3795 nodes. Using the community detection method described in section 2.2, we obtained lists of suicide-related adverse events. As noted above, the algorithm for community detection is stochastic, and we obtained a different list for each run of the analysis. A typical list appears in Table 2.2, which includes 141 events. We conducted network analysis 20 times, and the number of items on the list varied from 141 to 290 events. The lists included both well-known risk factors, such as social problems and stress, and other adverse events, such as cardiorespiratory arrest and gastric ulcer.

Abnormal behavior	Activation syndrome	adverse drug reaction	Adverse event
Affect lability	Affective disorder	Agoraphobia	Akathisia
Alcohol poisoning	Alcohol use	Anger	Anhedonia
Antisocial behavior	Apathy	Asphyxia	Balance disorder
Bipolar disorder	Bipolar i disorder	Breast cancer	Cardio-respiratory arrest
Cognitive disorder	Coma	Communication disorder	Condition aggravated

Crying	Decreased appetite	Decreased interest	Delusion
Dementia	Depersonalization	Depressed mood	Depression
Derealization	Disease recurrence	Disinhibition	Dissociation
Drug abuse	Drug abuser	Drug administration error	Drug dose omission
Drug effect decreased	Drug ineffective	Drug interaction	Drug intolerance
Drug screen false positive	Drug screen positive	Economic problem	Educational problem
Electric shock	Emotional disorder	Emotional distress	Fear
Fear, focus	Feeling abnormal	Feeling of despair	Flashback
Flat affect	Formication	Gastric disorder	Gastric ulcer
General physical health deterioration	Gun shot wound	Hallucination, auditory	Hallucination, visual
Hallucinations, mixed	Homicidal ideation	Homicide	Hostility
Hypersomnia	Ill-defined disorder	Immobile	Impaired work ability
Imprisonment	Impulsive behavior	Incorrect dose administered	Increased appetite
Indifference	Injury	Injury asphyxiation	Intentional overdose
Intentional self-injury	Judgment impaired	Laceration	Legal problem
Libido increased	Logorrhea	Loss of employment	Loss of libido
Major depression	Mania	Marital problem	Mental disorder
Mental impairment	Mental status changes	Mood altered	Mood swings
Multiple drug overdose	Murder	Muscular weakness	Musculoskeletal pain

Negative thoughts	Nervous system disorder	No adverse drug effect	Obsessive thoughts
Obsessive-compulsive disorder	Overdose	Panic attack	Panic disorder
Paranoia	Personality change	Personality disorder	Pharmaceutical product complaint
Physical assault	Poor quality sleep	Post-traumatic stress disorder	Psoriasis
Psychiatric symptom	Psychomotor hyperactivity	Psychotic disorder	Refusal of treatment by patient
Relationship breakdown	Respiratory arrest	Restlessness	Self esteem decreased
Self-injurious behavior	Self-injurious ideation	Shock	Skin laceration
Social avoidant behavior	Social problem	Stress	Tearfulness
Tension	Theft	Therapeutic response decreased	Therapeutic response unexpected with drug substitution
Thinking abnormal	Tic	Treatment noncompliance	Violence-related symptom
Withdrawal syndrome			

Table 2.2. List of suicide-related adverse events obtained by analysis

It should be noted that suicide-related adverse events do not always have a high correlation with suicidal events. Table 2.3 presents the top 10 adverse events that had a high Pearson's correlation with suicidal events. Interestingly, agitation, which showed the highest such correlation, is not in Table 2.2. This is because agitation also has a high correlation with other events. For example, agitation had a correlation of 0.462 with a confusional state, 0.349 with a disturbance in attention, and 0.328 with tremors. As a result of these high correlations, agitation was clustered into a different community from

suicidal events.

By contrast, cardiorespiratory arrest appeared among the suicide-related adverse events, though it had a low Pearson's correlation with suicidal events (0.03 for complete suicide, 0.008 for suicide ideation). However, this event is listed as suicide-related because it has a relatively high correlation with other suicide-related events. Cardiorespiratory arrest has a low correlation with other events. The greatest correlation was 0.096, which was the correlation for cardiac arrest. Cardiorespiratory arrest has a fairly high correlation with suicide-related events, such as flat affect (0.075), social avoidance behavior (0.06), and restlessness (0.038). It is natural to assume that cardiorespiratory arrest is related to suicidal behavior because suicidal attempts often result in cardiorespiratory arrest. Network analysis succeeded in revealing this relationship, which could not be identified using standard statistical analysis.

Suicidal event	Related event	Pearson's correlation
Suicidal ideation	Agitation	0.21
Suicidal ideation	Fatigue	0.151
Suicidal ideation	Crying	0.138
Suicidal ideation	Nervousness	0.132
Suicidal ideation	Nausea	0.123
Suicidal ideation	Drug withdrawal syndrome	0.122
Suicidal attempt	Non-accidental overdose	0.104
Suicidal behavior	Alcohol poisoning	0.075
Suicidal ideation	Restlessness	0.068
Suicidal attempt	Agitation	0.061

Table 2.3. Top 10 adverse events with a high Pearson's correlation with suicidal events

In the following section, we evaluate the effectiveness of the lists obtained in evaluating the risk of suicide. Unless otherwise indicated, the dataset for Group 2 was used for the evaluation.

First, we plotted the population of suicidal events as the function, defined in section 2.3 (Fig. 2.5). In that figure, the results of 20 trials of community detection are indicated by open triangles. It is evident in Fig. 2.5 that the results allow a division into groups. In 12 times among 20 trials of community detection, we obtained lists of adverse events that were strongly related to suicide: in those cases, if a patient had one suicide-related adverse event, the probability that they would also have a suicidal event was about 25%. That was about twice the figure for a subject with no suicide-related adverse events. The probability of suicidal events increased with an increase in k and became about 60% at $k = 5$. These results indicate that the obtained adverse events displayed a strong relationship with suicidal events. Other lists, such as those obtained eight times among 20 trials of community detection, were more weakly related with suicidal action. In those cases, the proportion of suicidal events was less than 20% at $k = 1$; the local maximum was approximately 40% at k. of about 10. This was due to the list of suicide-related adverse events in such cases being excessively long. Each of those lists included over 200 adverse events, whereas the length of “high-performance” lists had 140–200 items. Therefore, we conclude that those “low-performance” lists included events that

held only a weak connection with suicidal events.

Figure 2.5 also presents a plot of the proportion with the following: when the dataset for Group 1 was used instead of that for Group 2 (open circles); when all adverse events were employed rather than suicide-related adverse events (closed circles); and when 200 adverse events were used with a high Pearson's correlation with suicidal events (closed triangles). Clearly, there is no quantitative difference between the results obtained with the dataset for Group 1 and that for Group 2. In both cases, there were low- and high-performance lists. Compared with the results obtained with the suicide-related adverse event lists, the performance of the list for all adverse events was low. For example, the proportion of suicidal events at $k = 5$ was less than 0.2, whereas the high-performance list gave k of about 0.6. Interestingly, when we used the list of 200 adverse events with a high Pearson's correlation, the performance was better than the low-performance list; it was, however, worse than the high-performance list. Particularly, if k was under 15, the performance of the list with high Pearson's correlation was almost the same as with the low-performance list. These results suggest that community detection may capture the hidden relationship better than the simple Pearson's correlation.

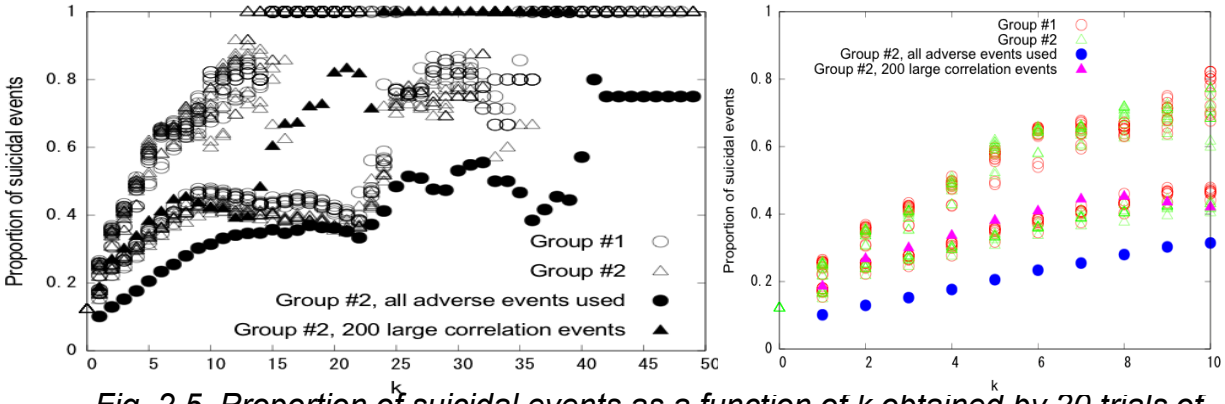

Fig. 2.5. Proportion of suicidal events as a function of k obtained by 20 trials of community detection.

We plotted the ROC to investigate the false-positive and true-positive rates. For each list obtained with 20 trials of modularity analysis, we calculated the true-positive and false-positive rates when we varied k from 0 (no accompanying adverse events) to 50. The result is indicated by the triangles in Fig. 2.6. The results obtained with each list were consistent: we obtained a true-positive rate of 60% and a false-positive rate of 20%. Using the trapezoidal rule for calculation, we obtained an area under curve (AUC) of 0.648–0.697 (0.681 on average). To demonstrate that our method effectively extracted suicide-related adverse events, we plotted the ROC when all adverse events were taken to be suicide-related (the circles in Fig. 2.6). The ROC obtained with our network analysis clearly showed a better performance than the one without network analysis, which resulted in an AUC of 0.529. This result suggests that our method was successful in effectively extracting suicide-related adverse events.

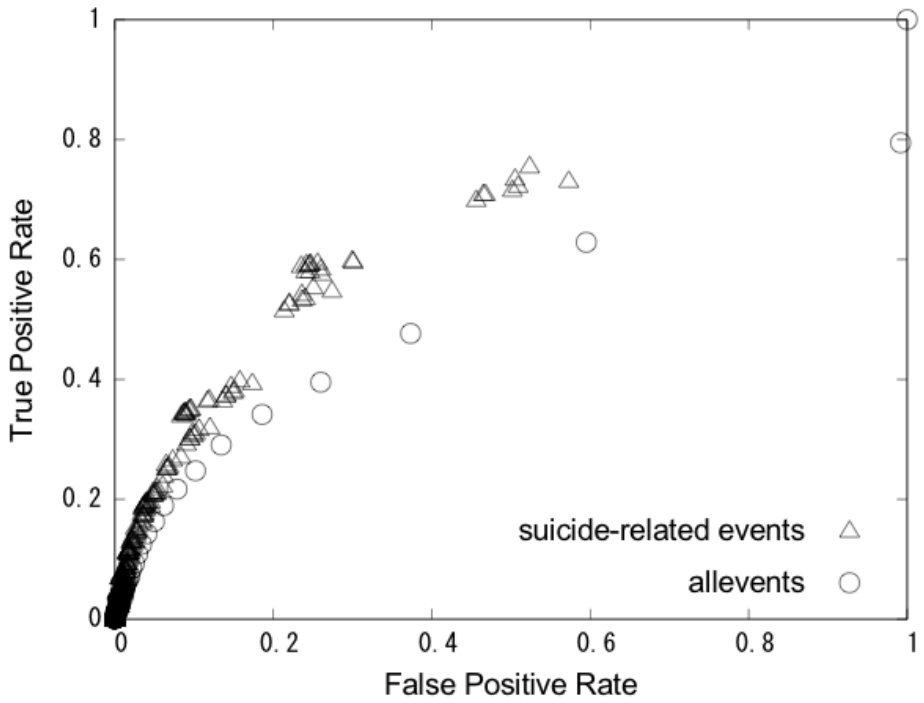

Fig. 2.6. Receiver operating characteristic curve obtained with the dataset in Group 2 for suicide-related adverse events (triangles) and for all adverse events (circles)

2.4. Discussion

The method adopted in the present study will be helpful in analyzing other diseases or adverse events. Compared with other analysis techniques, such as association analysis and Bayesian analysis, an advantage of our method lies in its ability to capture the indirect connection between symptoms, for example, cardiorespiratory arrest and suicide. Another advantage is that we were able to control the sensitivity and specificity. With association analysis and Bayesian analysis, it is possible to investigate the relationship among one or a few symptoms. In contrast, our method identifies the group of symptoms that belong to the same community with target symptoms. With this approach, sensitivity or specificity can be modified by stipulating the number of symptoms,

k, and we were able to adjust for maximal expected patient benefit.

However, our method presents several problems, and clinical application will demand further improvement. First, the list of symptoms we obtained was too long for clinical use. The smallest set of symptoms we obtained included over 140 symptoms. For practical application, that should be reduced to fewer than 20. This can be achieved in several ways. Refinement of the community detection algorithm may help in reducing that list. The modularity-based method is standard for community detection, but it is not the only approach. Different methods [39–41] may offer better performance for extracting suicide-related adverse events. Adjusting the resolution will also help control the size of the adverse event list. Such approaches will result in a smaller set of suicide-related events. It should be noted that our method for network construction is not unique. For example, Barabási constructed a two-diseasome network—one based on Pearson's correlation and the other on relative risk [33]. The relative risk-based network may provide a better indicator of suicide. Further study is required on the way to construct a network of symptoms and to detect communities.

The second problem with our study is the high false-positive detection rate. To surmount this problem, it would be useful to apply different signal detection techniques in combination with the community detection method. For example, a method based on the likelihood ratio test has been demonstrated to control false-positive rates [46]. Many other

algorithms of signal detection exist, such as proportional reporting ratio, reporting odds ratio, simplified Bayes, multi-item gamma Poisson shrinker, and Bayesian confidence propagation neural networks [47]. Applying these methods in combination with community detection will reduce the false-positive rate and increase the AUC.

Finally, we note that our result shows only correlation, not causation. For example, such items as cardiorespiratory arrest and gunshot wounds in our lists should not be taken as the cause but the result of suicidal behavior. We cannot determine whether or not overdose is the result of a suicidal attempt. Some of the adverse events, such as logorrhea and tics, seem to have no apparent relation to suicide. This problem is intrinsic in the analysis of spontaneous adverse reports. Investigating the cause of suicidal events demands another approach, such as a prospective cohort study.

2.5. Conclusion

In this study, we constructed lists of suicide-related adverse events from the FAERS using network analysis. We found that the more listed adverse events a person possessed, the greater the risk of suicide. This result suggests that such lists allow a qualitative estimation of the risk of suicide. Though our method presents problems, such as the high level of false-positive results and excessively long symptom lists, such drawbacks may be countered by refining the modularity detection algorithm, modifying the means of network construction, and applying different signal detection methods.

Chapter 3

Health Risk Estimation Using Time-series Analysis of Health Checkup Data

3.1. Introduction

General health checkup has played the essential role in the health care in several countries. In the UK, the publicly funded NHS Health Check Program was introduced in 2009, and in Denmark an organized health check program for the general public has been suggested. Health checkup are also performed by some primary care medical staff outside organized programs and by commercial clinics [23].

Japan has one of the most advanced health checkup systems in the world. In Japan, the Industrial Safety and Health Law obliges all workers to undergo annual health checkup in their workplaces. Workers who have one or more abnormal findings pointed out in their annual health checkup are summoned by occupational health staff and subsequently attend health consultations conducted by occupational health nurses each year [48]. This annual health checkup may contribute to reduction of medical expenditures in Japanese middle-aged workers [49].

Application of data mining techniques to the health checkup data will enable us to find the hidden factors that have relation with particular disease and make the further reduction of expenditures possible. For example, Chang et al. applied the data mining

techniques to identify the risk factors for hypertension and hyperlipidemia [50]. In this research, they used Multivariate Adaptive Regression Splines method to construct a multiple predictive model. In other research, Akdag et al. tried to find risk factors for hypertension using the classification tree method [51]. Here we notice that the data obtained by annual health checkups should be analyzed as time-series data. Though there were several studies applying data-mining technique for investigating the risk factors of hypertension or hyperlipidemia, there had been no study in which health checkup data are analyzed as time-series data. In the last year, our group carried out the time-series analysis of health checkup data using hidden Markov model (HMM) [52]. Using this model, we classified the people into 6 groups and showed that the persons in different group have different health risks.

In our previous study, we used the data that include both healthy and unhealthy people so that the results obtained indicates a health risk of both. However, we found the difficulty in determining health risk clearly, because the data include many kind of diseases, such as hypertension, diabetes, and hyperlipidemia. Moreover, it is difficult to advice healthy people to prevent life-style related diseases for the reduction of medical expenditures. They are healthy and do not conduct with medical staff. To overcome these difficulties, we analyzed the health checkup data of hypertension, i. e. Systolic Blood Pressure (SBP) > 140 mmHg or Diastolic Blood Pressure (DBP) > 90mmHg, in this

work. By using the data from one disease, the risk of changes can be seen more clearly. Hypertension is one of the primary risk factors for heart disease and stroke, the leading causes of death worldwide [53]. It is also the most common disease among life-style related diseases, and we have enough data for analysis. Moreover, hypertension is disorder and people with hypertension make periodic visit to medical staff. Therefore we can advise them to prevent the progress of diseases.

In this study, we made a time-series analysis of yearly health checkup data using HMM. HMM is usually used as a tool to analyze genome and voice recognition, and also has medical application such as health monitoring engine in order to avoid unexpected failures, and the study of infectious diseases when individuals can be in a number of imperfectly observed states [54-56]. This is the main reason to use the HMM as a method to analyze the health checkup data. By using the HHM, we can get the probability of a transition between states and can determine the level of health risk.

3.2. Material and Method

3.2.1. Data preparation

We made the datasets of health checkups for the time-series analysis. From the data provided by the medical center of Gifu, which has 912,765 records from 279,904 people, we selected the data of male that satisfy all the following conditions.

1. SBP values must be above 140 mmHg.

2. DBP values must be above 90 mmHg.

3. Have a 4-6 year time series of data.

Here we removed the data from females, because the number of data are too small for analysis. After extracting the data, we classified them using the age by decade. In Table 3.1, we summarize the number of records used in the following analysis.

Number of	Age Groups				
	20's	**30's**	**40's**	**50's**	**60's**
People	19	207	681	1,266	427
Records	85	940	3,127	5,734	1,902

Table 3.1. Number of records used in the analysis.

3.2.2. Analysis by HMM

HMM is a non-supervised learning method, especially popular for voice/speech recognition. HMM is a statistical model of a system that is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters (state) from the observed data. In a Hidden Markov model there is an underlying unobserved state of the system that changes in time according to a Markov process. The distribution of observations at a given time is determined by the system's state at that time [57]. The specified Markov process can then be used for further analysis, for example for the application of pattern recognition. In the following, we shortly explain HMM.

Suppose that at time t=1, 2, 3, … the person's health state is C_t, and his health

checkup data is X_t. In HMM, we assume that

$$\Pr(C_t \mid \mathbf{C}^{(t-1)}) = \Pr(C_t \mid C_{t-1}),\ t = 2, 3, \ldots \quad (1)$$

$$\Pr(X_t \mid \mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = \Pr(X_t \mid C_t),\ t \in N \quad (2),$$

where $\mathbf{C}^{(t-1)}=(C_1, C_2, \ldots, C_{t-1})$ and $\mathbf{X}^{(t-1)}=(X_1, X_2, \ldots X_{t-1})$, and Pr(A|B) indicates the probability to get A under the condition B.

The model consist of two parts: firstly, an unobserved state $\{C_t: t = 1, 2, \ldots\}$ satisfying the Markov property, i. e. that the state at time t is determined only by the state at time t-1. Secondly, it includes the state-dependent observation $\{X_t: t = 1, 2, \ldots\}$ such that, when C_t is known, the distribution of X_t depends only on the current state C_t and not on previous states or observations [57]. In HMM we construct the Markov model described by Eqs. (1) and (2) from observations. In this paper, we used the following 10 test values as the observations: Body Mass Index (BMI), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Hematocrit (HCT), Platelets (PLt), Glutamic Oxaloacetic Transaminase (GOT), Glutamic Pyruvic Transaminase (GPT), Total Cholesterol (T.Chol), Neutral Fat (NF) and Blood Sugar (BS). These 10 test values are the most commonly measured parameters during regular health checkups.

One of the problems of HMM is that we need to determine the topology of Markov model before carrying out the analysis. In this study, we tried the HMM analysis for 4 Markov model with different topologies depicted by Fig. 3.1, and selected the best model

among them using Bayes Information Criterion (BIC), explained in the next subsection.

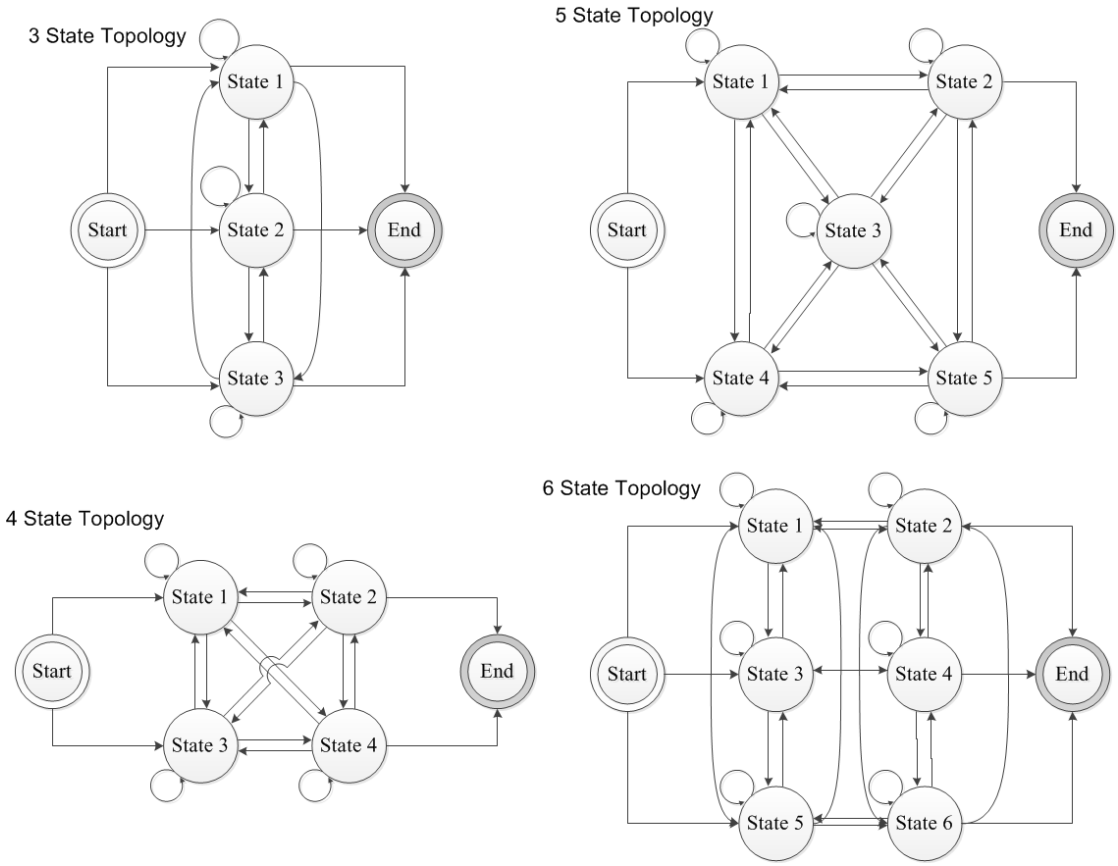

Figure 3.1.Topologies used in the hidden Markov model analysis of health checkup data.

In this study, we used Hidden Markov Model Toolkit (HTK), which is a toolkit of HMM developed by Cambridge University, for analysis [58].

3.2.3. Evaluation of the result

After performing data processing with HTK, we selected the appropriate model using BIC. BIC was introduced by Schwarz (1978) to select a best model from a set of candidate models by maximizing the posterior probability [59]. The definition of BIC is given by $\sum_i \log P_i - \frac{K_M}{2} \log N$, where P_i, K_M, N represent the likelihood that *i*-th data

takes the state determined by HMM, the number of free parameters included in HMM, and total number of data used for analysis, respectively. In this approach, the topology that has the smallest BIC is the best model. BIC has been widely used to estimate the topology of HMM in Speech Activity Detection and Speaker Diarization tasks [60], and handwriting recognition [61].

To estimate the health risk of each state, we used the decision by professional medical staffs. In Japan, the results of the health checkup are usually classified into five categories by professional medical staffs: A, B, C, D1 and D2. A means "within normal values", B is "slightly above the normal value", C is "retesting required". Usually someone who is getting the health check-up with C value will get advice and suggestions from doctor or health consultant to change lifestyle or diet. Results D2 means "Advanced examination required" and D1 means "Medical treatment should be required". For the case of D2 and D1, it will usually be followed by a more detailed examination in accordance with the results of health check-up. The guideline of this classification is shown in the table 3.2, though not rigorous one [62, 63].

Parameters	**A**	**B**	**C**	**D**
	Normal	**Mild abnormality**	**Need to take a wait-and see approach**	**Need medical treatment**
BMI (kg/m^2)	18.5~24.9		~18.4 or 25.0~	
SBP (mmHg)	~129	130~139	140~159	160~
DBP (mmHg)	~84	85~89	90~99	100~
Hematocrit(%)	38.5~48.9	49.0~50.9	35.4~38.4	~35.3, 51,0~

Platelet(×10^4/ml)	13.0~34.9	35.0~39.9	10.0~12.9	~9.9, 40.0~
GOT(IU/l)	0~30	31~35	36~50	51~
GPT(IU/l)	0~30	31~40	41~50	51~
T. Cholesterol(mg/dl)	140~199	200~219	220~259	~139, 260~
Neutral Fat(mg/dl)	30~149	150~199	200~399	~29, 400~
Blood Sugar(mg/dl)		~139	140~199	200~

Table 3.2. Range of normal values in health checkup data.

Using this professional decision, we evaluated the health risk of each state as show in Table 3.3. For example, if the category A + B is more dominant than the other categories where the other categories not exceed 30% in people classified into a state, we call this state Low Risk. Similarly, if the category C is dominant and the other categories do not exist in excess of 30% in a state, we call it Medium Risk.

Risk Name Categories	**A+B**	**C**	**D**
Very Low	>70%		
Low	Dominant		
Low (Me)	Dominant	>30%	
Low (Hi)	Dominant		>30%
Medium (Lo)	>30%	Dominant	
Medium		Dominant	
Medium (Hi)		Dominant	>30%
High (Lo)	>30%		Dominant
High (Me)		>30%	Dominant
High)			Dominant
Very High			>70%

Table 3.3. Risk Categories

3.3. RESULT

3.3.1. BIC Result

After the calculation for BIC, we got the result in table 3.4.

	Average of log P_i	**Number of Samples**	**$\sum$ log P_i**	**K_M**	**N**	**BIC**
20 - 3 state	-161.58235	19	-3070.06	71	85	(3,227.78)
20 - 4 state	-158.51743	19	-3011.83	95	85	**(3,222.86)**
20 - 5 state	-176.28613	19	-3349.44	119	85	(3,613.77)
20 - 6 state	**Not Enough Data**					
30 - 3 state	-168.70935	207	-34922.8	71	940	(35,165.86)
30 - 4 state	-168.59285	207	-34898.7	95	940	(35,223.90)
30 - 5 state	-168.01443	207	-34779	119	940	(35,186.32)
30 - 6 state	-167.30862	207	-34632.9	140	940	**(35,112.10)**
40 - 3 state	-171.56807	681	-116838	71	3127	(117,123.55)
40 - 4 state	-171.92404	681	-117080	95	3127	(117,462.54)
40 - 5 state	-171.40434	681	-116726	119	3127	(117,205.20)
40 - 6 state	-170.66441	681	-116222	140	3127	**(116,785.81)**
50 - 3 state	-168.60489	1266	-213454	71	5734	(213,761.01)
50 - 4 state	-169.1748	1266	-214175	95	5734	(214,586.37)
50 - 5 state	-168.67477	1266	-213542	119	5734	(214,057.18)
50 - 6 state	-167.82216	1266	-212463	140	5734	**(213,068.65)**
60 - 3 state	-164.15295	427	-70093.3	71	1902	(70,361.36)
60 - 4 state	-164.47632	427	-70231.4	95	1902	(70,590.05)
60 - 5 state	-164.08369	427	-70063.7	119	1902	(70,513.00)
60 - 6 state	-163.43199	427	-69785.5	140	1902	**(70,314.01)**

Table 3.4. Bayesian Information Criterion (BIC) for High Blood Pressure Data

The smallest BIC is shown in bold text in Table 3.4. Based on the result, we find that the 6 state model is the most appropriate because it has the smallest BIC in all age group except at age group 20's. In the case of age group 20's, we cannot obtain the BIC data due to the smallness of datasets. In the following, we show the result of the analysis by 6-state model over 30's.

3.3.2. Average parameters for each state

In table 3.5, we show the average of test values in each state

Parameter	30's						40's					
	State						State					
	1	2	3	4	5	6	1	2	3	4	5	6
BMI	23.54	24.24	24.11	24.17	**27.10**	**27.55**	22.37	22.71	**25.78**	**27.17**	**25.96**	**25.34**
SBP	**145.30**	**150.66**	**159.18**	**157.07**	**150.45**	**158.26**	**151.61**	**155.65**	**154.38**	**153.50**	**151.59**	**157.29**
DBP	**90.33**	**93.21**	**97.53**	**95.76**	**94.34**	**98.16**	**94.14**	**95.41**	**94.33**	**94.76**	**94.72**	**97.44**
HCT	45.90	46.05	46.00	45.80	47.13	46.78	44.94	44.61	45.72	45.50	46.04	45.24
PLt	25.34	26.59	23.51	23.84	25.19	25.46	25.67	25.47	25.89	25.70	25.02	25.00
GOT	20.70	22.51	24.79	21.16	29.68	27.73	23.70	20.89	22.34	21.05	32.01	32.96
GPT	22.53	25.32	29.34	23.58	**49.21**	**42.74**	24.22	19.01	29.11	25.18	**46.63**	**42.37**
T.Chol	**208.17**	**214.91**	189.30	198.03	**215.12**	**222.63**	**203.94**	**206.66**	**213.32**	**209.13**	**218.15**	**210.39**
NF	**172.96**	**169.23**	125.82	134.85	**204.62**	**250.17**	**154.37**	131.64	**195.67**	**191.65**	**217.53**	**209.46**
BS	101.82	92.35	103.11	121.05	105.18	111.23	101.67	106.42	121.61	103.22	108.37	123.89
Parameter	**50's**						**60's**					
	State						State					
	1	2	3	4	5	6	1	2	3	4	5	6
BMI	21.49	21.69	**25.38**	**25.53**	24.99	**26.11**	23.23	23.36	**27.03**	**27.67**	23.19	23.35
SBP	**154.43**	**158.15**	**154.43**	**155.63**	**156.18**	**156.50**	**156.70**	**154.19**	**155.69**	**155.04**	**157.60**	**157.63**
DBP	**93.00**	**93.97**	**94.23**	**94.05**	**94.10**	**92.68**	**92.90**	**90.51**	**92.45**	**92.16**	**92.59**	**92.36**
HCT	44.22	43.89	44.96	44.64	45.43	45.07	44.32	43.64	44.58	44.35	44.67	44.02
PLt	23.92	24.41	24.80	24.67	24.29	23.22	23.29	23.31	23.69	22.30	23.44	22.43
GOT	23.68	22.76	22.71	20.51	**32.42**	28.49	23.57	21.47	27.03	**30.27**	29.37	25.73
GPT	20.88	19.41	25.39	20.42	40.39	33.77	19.83	16.71	31.65	34.87	29.97	23.21
T.Chol	**204.24**	**209.60**	**211.60**	**207.47**	**212.61**	**206.45**	**207.44**	**202.71**	**216.28**	**204.51**	**207.88**	198.26
NF	117.15	146.55	**184.33**	**166.10**	**197.70**	**173.72**	129.31	138.44	**196.69**	**180.30**	**178.47**	**178.57**
BS	106.77	119.66	110.05	103.99	118.72	137.03	101.73	105.98	117.75	130.09	124.06	135.77

Table 3.5. Mean test values in each state. The values over the normal range are shown in bold.

In all age groups, in addition to parameter SBP and DBP that from beginning exceed the normal ranges, there are three parameters whose values exceed normal

value in the worst state; BMI, T.Chol and NF. These three values have close connection with hypertension and cardiovascular diseases. If both T.Chol and blood pressure are high, it can lead to heart disease. A BMI that exceeds the normal range means that the person is overweight and could be an obesity. Obesity can cause abnormalities in blood pressure. Obesity also give effect to increasing risk of health problems such as coronary heart disease and hypertension, together with type 2 diabetes, arthritis and cancer [64]. Excess of neutral fat can cause buildup on walls of blood vessels which can lead to hypertension [65]. Evident from the results shown in Table 3.5, the average value of BMI, T.Chol and NF are over the normal range in almost every state in every age group. We also found that when the BMI is over the normal range, then the T.Chol and NF also are.

3.3.3. Transition Probability

In table 3.6, we show the transition probability between each state. Self-looping usually have the biggest probabilities with the lowest value 38.18% in the age group 40's, state 3. There are a two exception in the 40's; the transition probability from state 1 to state 2 is higher than the that of self- looping in state 1, and the transition probability from state 3 to state 4 is higher than that of self- looping in state 3. In the age group 50's, transition probability from state 3 to state 4 is higher than that of self- looping in state 3.

	State /State	1	2	3	4	5	6
30's	**1**	**47.69%**	44.27%	4.17%	0.00%	3.87%	0.00%
	2	0.00%	**44.60%**	0.00%	9.46%	0.00%	0.00%
	3	0.00%	0.00%	**54.07%**	45.93%	0.00%	0.00%
	4	0.00%	8.96%	0.00%	**42.17%**	0.00%	0.00%
	5	6.39%	0.00%	5.30%	0.00%	**45.37%**	42.94%
	6	0.00%	0.00%	0.00%	5.36%	0.00%	**55.94%**
40's		**1**	**2**	**3**	**4**	**5**	**6**
	1	47.96%	52.04%	0.00%	0.00%	0.00%	0.00%
	2	0.00%	**56.97%**	0.00%	0.00%	0.00%	5.83%
	3	3.43%	0.00%	38.18%	45.99%	12.40%	0.00%
	4	0.00%	0.70%	0.00%	**58.82%**	0.00%	0.00%
	5	2.83%	0.00%	13.31%	0.00%	**48.04%**	35.83%
	6	0.00%	0.00%	0.00%	1.19%	0.00%	**45.15%**
50's		**1**	**2**	**3**	**4**	**5**	**6**
	1	**58.01%**	41.99%	0.00%	0.00%	0.00%	0.00%
	2	0.00%	**48.00%**	0.00%	0.00%	0.00%	0.00%
	3	0.00%	0.00%	48.69%	51.31%	0.00%	0.00%
	4	0.00%	0.00%	0.00%	**50.35%**	0.00%	10.94%
	5	5.46%	0.00%	4.52%	0.00%	**53.38%**	36.64%
	6	0.00%	0.03%	0.00%	4.27%	0.00%	**49.18%**
60's		**1**	**2**	**3**	**4**	**5**	**6**
	1	**50.50%**	48.60%	0.89%	0.00%	0.00%	0.00%
	2	0.00%	**51.03%**	0.00%	0.00%	0.00%	6.64%
	3	0.03%	0.00%	**56.04%**	43.93%	0.00%	0.00%
	4	0.00%	0.00%	0.00%	**51.29%**	0.00%	0.00%
	5	7.89%	0.00%	0.00%	0.00%	**53.63%**	38.48%
	6	0.00%	0.00%	0.00%	0.00%	0.00%	**51.78%**

Table 3.6. Transition Probability. Self-looping is shown in bold.

Self-looping (shown in bold text at table 3.6) implies that the health condition of a person does not changed from that in the previous year. In the 30's in state 1, a self-looping have a transition probability 47.69%, which means that 47.69% of the people who

had previously been in state 1, are still in state 1 in the next year. The health conditions of these people do not change. On the other hand, there is a transition probability 44.27% from state 1 to state 2. This means that 44.27% of people who had previously been in state 1 are changed into state 2. The health condition of these people may be much better or worse, which depend of every state's health risk. Next, we evaluate the health risk of each state.

3.3.4. Comparing with Health Checkup Examination

In this subsection, we evaluate the risk of each state using the table 3.3 defined in the previous section.

The result for each age group is shown in Fig. 3.2 – 3.5. In case of the 30's in Fig. 3.2, there are four kinds of risk. State 1 has Medium risk, where the dominant category is C with more than 70%. State 2 has Medium (Hi) risk, where the dominant category is C followed by the category D with more than 30%. Third risk level is High (Me) risk at state 3, 4 and 5, where the category D is dominant followed by the category C with more than 30%. Finally, the state 6 has the highest risk: Very High, where more than 70% of people are classified into the category D. We note that state 3, state 4 and state 5 are not combined into one state. Among these 3 states with the High (Me) risk, state 5 have the highest risk in the future. The state 5 has a fairly large transition probability (42.94%) to state 6 which have a Very High risk. In addition to the fact that the mean

values of characteristics are not the same, these states have different transition probabilities. Even if two state has same health risk, the risk in the future also depends on the transition probability. If one state has the larger transition probability to very high risk state than another, the risk in the future will be increased. Therefore state 5 has high health risk in the future, while current risk is same as other two states. This conclusion is consistent with the averages of test values in table 3.5. From 10 test values, 6 values have averages that exceed the normal ranges in state 5; BMI, SBP, DBP, GPT, T.Chol and NF. On the other hand, other two states have 2 values (SBP and DBP) that exceed the normal ranges.

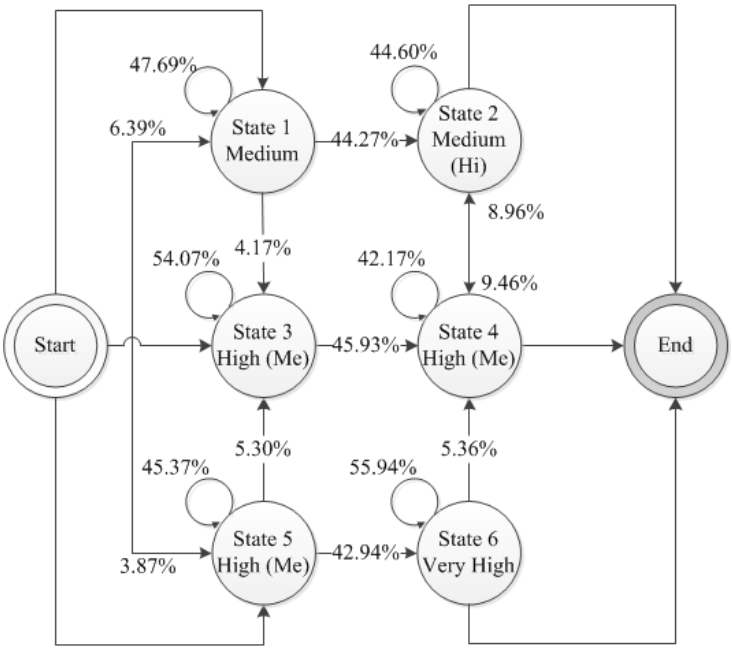

Fig. 3.2. Transition probability with risk level for age group 30's

In case of age 40's shown in Fig. 3.3, there are 3 levels of risk, namely Medium (Hi) in state 1, High (Me) in state 2, 3, 4, and 5, and Very High risk in state 6. Difference

among 4 states with the High (Me) risk is at the transition probability. State 2 only has two transitions; one is the self-looping transition and the other is the transition to state 6 (Very High risk). State 3 has a one-way transitions to state 1 and state 4, and also has a reciprocal transition to state 5. State 4 only has two transitions, a self-looping transition and a transition to state 2. State 5 has a reciprocal transition to state 3, and also has a transition to state 1 (Medium (Hi) risk) and state 6 (Very High risk). Among these four states with the High (Me) risk, state 5 has the highest risk. From 10 test values, 6 values have the averages that exceed the normal range; BMI, SBP, DBP, GPT, T.Chol and NF. On the other hand, state 2 has 3 values that exceed the normal range (SBP, DBP and T.Chol), state 3 and state 4 have 5 test values that exceed the normal range; BMI, SBP, DBP, T.Chol and NF. The state 5 has a transition probability fairly large (35.83%) to state 6 which have a Very High risk. The lowest risk state is state 2, where there are only 3 test values whose averages exceed the normal ranges. Therefore we find the consistency between risks estimated from averages of test values and that from transition probability again.

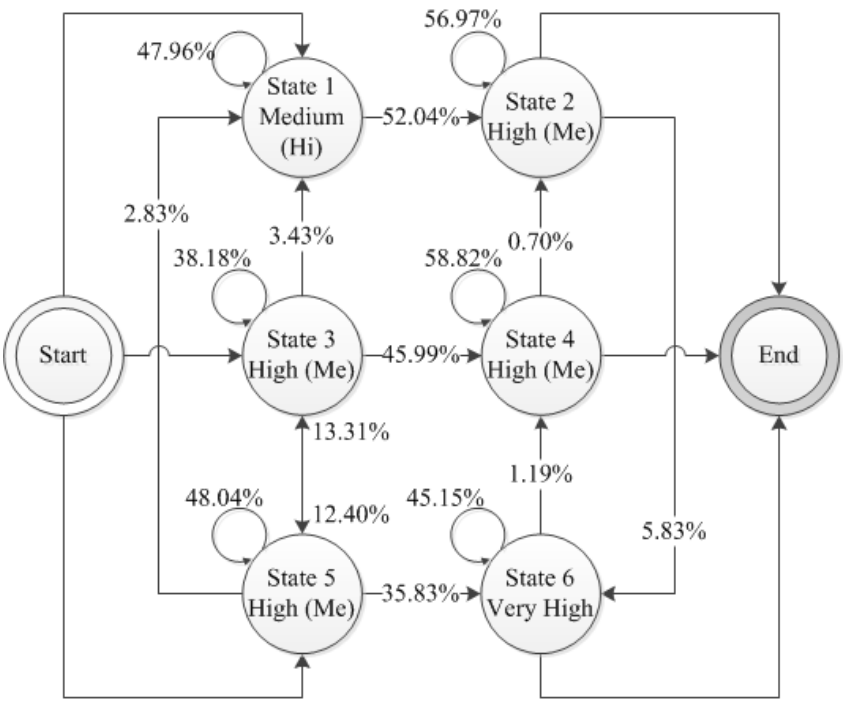

Fig. 3.3. Transition probability with risk level for age group 40's

In age group 50's shown in Fig. 3.4, obtained Markov model has 2 levels of risk; Medium (Hi) risk at state 1 and High (Me) risk at the other states. All states in the High (Me) risk have different transition probabilities. State 2 only has self-looping transition. State 3 has 1 transition to State 4 and received 1 transition from State 5. State 4 has a reciprocal transition to state 6. State 5 does not received transition from the other state, but state 5 has 3 transition to the other states; state 1, 3 and 6. State 6 has a reciprocal transition to state 4, and also has a transition to state 2.

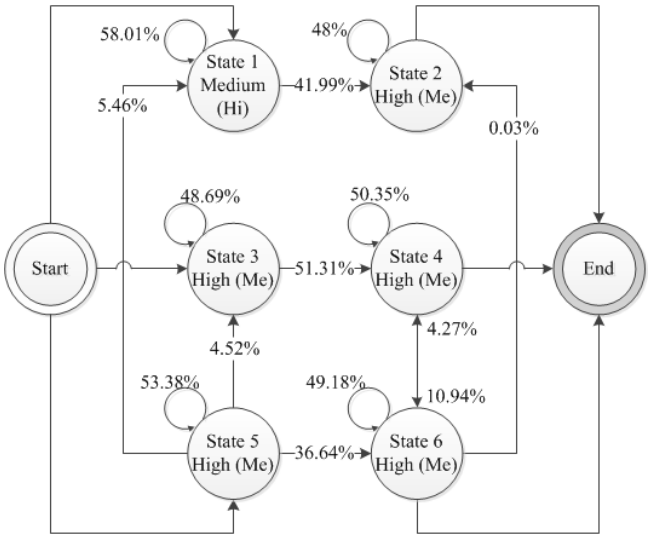

Fig. 3.4. Transition probability with risk level for age group 50's

In the last age group, 60's, the Markov model also has 2 risk level, Medium (Hi) and High (Me). State 1 and 2 have Medium (Hi) risk, while the other have High (Me). In this age group, state 4 has a self-loop only, and there is no transition to another state. This means that the people in this state have a stable health condition, which gets neither better nor worse.

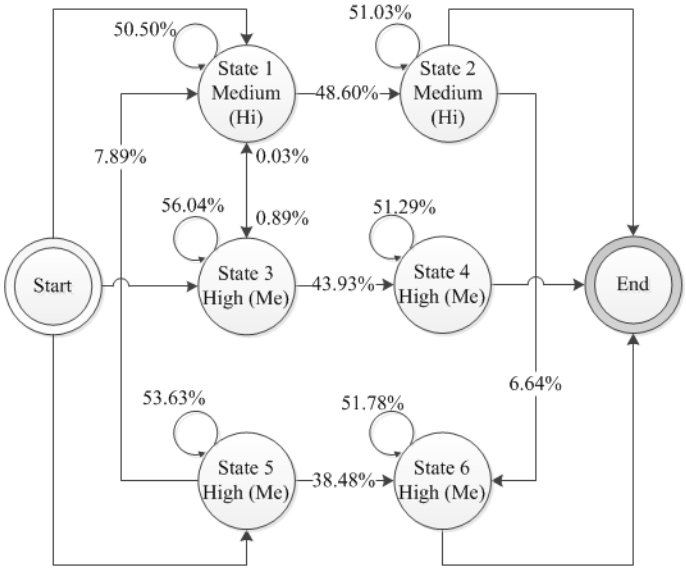

Fig. 3.5. Transition probability with risk level for age group 60's

3.4. Conclusion

Using HMM, we succeeded to cluster the health checkup data into 6 groups with different health risks. Through the calculation of transition probability, we also found out the probability of change in the health risks based on annual health checkup data. Generally, the largest transition probability is that of the self-looping, which means a person's health condition has not changed. Using transition probability, we estimated the health risk of each state in the future. The result is consistent with the health risk estimated by the averages of test values.

Using HMM we succeeded to determine the risk of changes in health conditions of peoples who have hypertension. With the success of this study, we believe that HMM could be used to analyze other life-style related diseases such as diabetes or hyperlipidemia. Unfortunately, we currently have not enough data for these analysis, because the number of patients of these diseases is much smaller than that of hypertension. However, we will be able to make HMM analysis in the future, if we can obtain the health checkup data from multiple medical centers or hospitals. The analysis of these data will enable us to prevent the progress of these diseases, which leads to the reduction of medical expenditure.

Chapter 4

Conclusions

1. We extracted the symptom strongly related to suicide event, and we also developed model to estimate the health condition in the future of hypertension patients. These results are available for the early detection of severe events, such as suicide or aggravation of hypertension.
2. We constructed lists of suicide-related adverse events from the FAERS using network analysis. We found that the more listed adverse events a person possessed, the greater the risk of suicide. This result suggests that such lists allow a qualitative estimation of the risk of suicide. Though our method presents problems, such as the high level of false-positive results and excessively long symptom lists, such drawbacks may be countered by refining the modularity detection algorithm, modifying the means of network construction, and applying different signal detection methods.
3. Using HMM, we succeeded to cluster the health checkup data into 6 groups with different health risks. Through the calculation of transition probability, we also found out the probability of change in the health risks based on annual health checkup data. Generally, the largest transition probability is that of the self-looping, which means a

person's health condition has not changed. Using transition probability, we estimated the health risk of each state in the future. The result is consistent with the health risk estimated by the averages of test values.

4. Using HMM we succeeded to determine the risk of changes in health conditions of peoples who have hypertension. With the success of this study, we believe that HMM could be used to analyze other life-style related diseases such as diabetes or hyperlipidemia. Unfortunately, we currently have not enough data for these analysis, because the number of patients of these diseases is much smaller than that of hypertension. However, we will be able to make HMM analysis in the future, if we can obtain the health checkup data from multiple medical centers or hospitals. The analysis of these data will enable us to prevent the progress of these diseases, which leads to the reduction of medical expenditure.
5. From these results, we conclude that data mining methods such as the network analysis and hidden Markov model is useful to process the data contained in the healthcare centers and to derive the valuable information.
6. From the results, we can said that the amount of data is an important. If number of data too small, the data mining process cannot be run.

References

1. Appleby, Julie. “Seven Factors Driving Up Your Health Care Cost”. Kaiser Health News, 2012. Retrieved September 9, 2014 from http://www.kaiserhealthnews.org/stories/2012/october/25/health-care-costs.aspx
2. Thearling, Kurt. “An Introduction to Data Mining”. Whitepaper. http://www3.shore.net/~kht/dmwhite/dmwhite.htm
3. Fayyad, Usama. “Advances in Knowledge Discovery and Data Mining”. MIT Press. 1996
4. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases”, In Proc. 1994, Int. Conf. Very Large Data Bases (VLDB), 1994.
5. J. Han, J. Pei and Y. Yin. “Mining frequent patterns without candidate generation”, In Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD’00), 2000.
6. J.R. Quinlan. “C4.5: Programs for Machine Learning”, Morgan Kaufmann, 1993.
7. J. Gehrke, R. Ramakrishnan and V. Ganti. “Rainforest: A framework for fast decision tree construction of large datasets”. In Proc. 1998 Int. Conf Very Large Data Bases (VLDB), 1998.
8. Gunguly N., Deutsh A., Mukherjee A., Editors. Dynamics On and Of Complex Networks: Application to Biology, Computer Science, and the Social Sciences. Basel:

Birkhäuser; 2009.

9. Liljeros F., Edling CR., Amaral LAN. Stanley HE. Aberg Y. The web of human sexual contacts. Nature. 2001; 411: 907-908.
10. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286 (1989).
11. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G.J. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge UK, 1998).
12. Eddy, S.R, What is a hidden Markov model?, *Nature Biotechnology,* 22, 1315 - 1316 (2004).
13. Adverse Events. (2014). Retrieved August 5, 2014, from http://www.nps.org.au/glossary/adverse-events
14. FDA Adverse Event Reporting System (FAERS) (formerly AERS) (2014), Retrieved August 5, 2014, from http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm
15. Eriksson. R., Werge, T., Jensen, LJ. Brunak, S. Dose-Specific Adverse Drug Reaction Identification in Electronic Patient Records: Temporal Data Mining in an Inpatient Psychiatric Population. Drug Saf. 2014; 37(4):237-247.

16. DuMouchel, W., Ryan, PB. Schuemie, MJ. Madigan, D. Evaluation of Disproportionality Safety Signaling Applied to Healthcare Databases. Drug Saf. 2013; 36(1):123-132.
17. Szarfman A, Machado SG, O'Neill, RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Safety, 2002; 25: 381-392.
18. Hauben, M, Madigan, D, Gerrits CM, Walsh L, Van Puijenbroek EP. The role of data mining in pharmacovigilance. Expert Opin Drug Saf. 2005; 4: 929-48
19. Piccinni C, Motola D, Marchesini G, Poluzzi E. Assessing the association of pioglitazone use and bladder cancer through drug adverse event reporting. Diabetes care. 2011; 34: 1369-1371.
20. McGraw-Hill Concise Dictionary of Modern Medicine. © 2002 by the McGraw-Hill Companies, Inc.
21. Important Reasons for an Annual Checkup (2014), Retrieved August 5, 2014, from http://www.livestrong.com/article/169496-important-reasons-for-an-annual-checkup/
22. Hardy, M. FOLLOW-UP OF MEDICAL RECOMMENDATIONS: Results of a Health Checkup of a Group of Well Children in Chicago. JAMA. 1948; 136(1):20-27.
23. Krogsbøll LT, Jørgensen K, Gøtzsche PC. General Health Checks in Adults for Reducing Morbidity and Mortality from Disease. JAMA. 2013; 309(23):2489-2490.

24. Okubo, Y., Sairenchi, T., Irie, F., Yamagishi, K., Association of Alcohol Consumption With Incident Hypertension Among Middle-Aged and Older Japanese Population: The Ibarakai Prefectural Health Study (IPHS). Hypertension. 2014; 63:41-4

25. Ohno, Y., Ishimura, E., Naganuma, T., Kondo, K., et al. Prevalence Of and Factors Associated With Chronic Kidney Disease In Japanese Subjects, Who Have No Known Chronic Disease, Undergoing Anannual Health Checkup. Nephrology Dialysis Transplantation. 2012; 27:395-395

26. Harris EC, Barraclough B. Suicide as an outcome for mental disorders. A meta-analysis. Br J Psychiatry. 1997; 170: 205-28.

27. Mann JJ, Waternaux C, Haas GL Malone KM. Toward a clinical model of suicidal behavior in psychiatric patients. Am J Psychiatry. 1999; 156: 181-189.

28. Reihmer Z. Suicide and bipolar disorder. In: Zarate CA. Manji HK, editors. Bipolar Depression: Molecular Neurobiology, Clinical Diagnosis and Pharmacotherapy. Basel: Birkhäuser; 2009. p. 47-56.

29. Vaswani MK, Linda FK, Ramesh S. Role of selective serotonin reuptake inhibitors in psychiatric disorders: a comprehensive review. Prog Neuropsychopharmacol Biol Psychiatry. 2003; 27: 85-102.

30. Gunguly N, Deutsh A, Mukherjee A, Editors. Dynamics On and Of Complex Networks: Application to Biology, Computer Science, and the Social Sciences. Basel:

Birkhäuser; 2009.

31. Pastor-Satorras R, Vespignani A. Immunization of complex networks. Phys Rev E. 2002; 65: 036104.
32. Alon U. An Introduction to Systems Biology: Design Principles of Biological Circuits. London: Chapman and Hall; 2006.
33. Barabási AL. Network Medicine-From obesity to the “Diseasome”. N Engl Jour Med. 2007; 357: 404-407.
34. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting systems. Am Stat. 1999; 53: 177-190.
35. Harpaz R, Chase HS, Friedman C. Mining multi-item drug adverse effect associations in spontaneous reporting systems. BMC Bioinformatics 2010; 11(Supp 9): S7.
36. Newman MEJ, Modularity and community structure in networks. Proc Natl Acad Sci U S A. 2006; 103: 8577-8582.du
37. American Society of Health-System Pharmacists. AHFS Drug Information 2010. Maryland: American Society of Health-System Pharmacists; 2010.
38. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008; 36: D901-6.
39. Newman MEJ. Finding and evaluating community structure in networks. Phys Rev E. 2004; 69: 026113.

40. Guimerà R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Phys Rev E. 2004; 70: R025101.

41. Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex network. Proc Natl Acad Sci U S A. 2007; 104: 7327-7331.

42. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008; P10008.

43. Fortunato S, Barthélemy M. Resolution limit in community detection. Proc Natl Acad Sci U S A. 2007; 104: 36-41.

44. Lambiotte R, Delvenne J -C, Barahona M. Laplacian dynamics and multiscale modular structure in networks [internet]. 2008; arXiv: 0812.1770.

45. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media; 2009 May 17-20; San Jose, United States. Available at http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154 . Accessed 23 Mar 2014.

46. Huang L, Zalkikar J, Tiwari RC. A Likelihood Ratio Test Based Method for Signal Detection with Application to FDA's Drug Safety Data. J Am Stat Assoc. 2011; 106: 1230-1241.

47. Huang L, Guo T, Zalkikar JN, Tiwari RC. A Review of Statistical Methods for Safety Surveillance. Therapeutic Innovation & Regulatory Science. 2014; 48: 98-108.

48. Kudo Y, Satoh T, Kido S, *et al.* The Degree of Workers' Use of Annual Health Checkup Results among Japanese Workers. *Ind Health* 2008;46:223–32.

49. Suka M, Yoshida K, Matsuda S. Effect of annual health checkups on medical expenditures in Japanese middle-aged workers. *J Occup Environ Med* 2009;51:456–61.

50. Chang C-D, Wang C-C, Jiang BC. Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors. *Expert Syst Appl* 2011;38:5507–13.

51. Akdag B, Fenkci S, Degirmencioglu S, *et al.* Determination of risk factors for hypertension through the classification tree method. *Adv Ther* 2006;23:885–92.

52. Kawamoto R, Nazir A, Kameyama A, *et al.* Hidden Markov Model for Analyzing Time-Series Health Checkup Data. *Stud Health Technol Inform* 2013;192:491 – 495.

53. Chockalingam A, Campbell NR, Fodor JG. Worldwide epidemic of hypertension. *Can J Cardiol* 2006;22:553–5.

54. Yu J. Health Condition Monitoring of Machines Based on Hidden Markov Model and Contribution Analysis. *IEEE Trans Instrum Meas* 2012;61:2200–11.

55. Wall MM, Li R. Multiple indicator hidden Markov model with an application to medical utilization data. *Stat Med* 2009;28:293–310.

56. Satten GA, Longini Jr IM. Markov Chains With Measurement Error: Estimating the `True' Course of a Marker of the Progression of Human Immunodeficiency Virus Disease. *J R Stat Soc Ser C (Applied Stat)* 1996;45:275–309.

57. Zucchini W, MacDonald IL. *Hidden Markov Models for Time Series - An Introduction Using R*. Chapman and Hall/CRC 2009.

58. HTK Speech Recognition Toolkit. http://htk.eng.cam.ac.uk

59. Hirose K, Kawano S, Konishi S, *et al.* Bayesian Information Criterion and Selection of the Number of Factors in Factor Analysis Models. *J Data Sci* 2011;9:243–59.

60. Leeuwen DAv, Huijbregts M. The AMI speaker diarization system for NIST RT06s meeting data. *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation, RT06s*. Washington DC, USA: Springer Verlag, 2007:371-84.

61. Li D, Biem A, Subrahmonia J. HMM topology optimization for handwriting recognition. In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. IEEE 2001. 1521–4.

62. Japan Society of Ningen Dock Diagnosis Guidelines, 2013. http://www.ningen-dock.jp/

63. Treatment Guide for Diabetes 2012-2013, edited by Japan Diabetes Society.

64. Walley AJ, Blakemore AIF, Froguel P. Genetics of obesity and the prediction of risk for health. *Hum Mol Genet* 2006;15 Spec No:R124–30.

65. Murakami NYHKK, Okada YYHKK. Astaxanthin-containing agent for lowering neutral fat concentration in blood: Google Patents, 2006

Figure list

Chapter 2

Chapter 3

Table list

Chapter 2

Chapter 3

List of Publications

1. A. Nazir, T. Ichinomiya, N. Miyamura, Y. Sekiya, Y. Kinosada

   “Identification of suicide-related events through network analysis of adverse event reports”

   Drug Safety, 2014, 37(7).

2. Kawamoto R, Nazir A, Kameyama A, Ichinomiya T, Yamamoto K, Tamura S, Yamamoto M, Hayamizu S, Kinosada Y.

   “Hidden Markov model for analyzing time-series health checkup data.”

   Stud Health Technol Inform. 2013; 192:491-5.

List of Presentations

1. Alwis Nazir, Ryouhei Kawamoto, Keiko Yamamoto, Satoshi Tamura, Takashi Ichinomiya, Satoru Hayamizu, Yasutomi Kinosada

   “Time-series analysis of health checkup data using Hidden-Markov Model”, Annual Symposium American Medical Information Association 2013, 16 - 20 November 2013, Washington DC, USA (Poster Presentation)

Curriculum Vitae

Alwis Nazir

Born on August 7, 1974 in Padang, West Sumatera, Indonesia

Married with Roza Linda

Father of M. Anugrah Syauqi Alzazirka

1989 – 1992 : Senior High School at SMAN 8 Padang, West Sumatera, Indonesia

1995 – 2000 : Bachelor at Department of Informatics Engineering, STMIK Budi Luhur Jakarta, Indonesia

2006 – 2008 : Master Degree at Post Graduate University of Putra Indonesia, Padang, Indonesia

2009 – Now : Lecturer at Department of Informatics Engineering, State Islamic University of Sultan Syarif Kasim, Pekanbaru, Indonesia

2011 – 2014 : Doctoral Degree at Medical Information Science Division, United Graduate School of Drug Discovery and Medical Information Science, Gifu University, Gifu, Japan

Acknowledgement

Thanks and praise to God (Allah Almighty), the merciful and compassionate, for blessing me in all my life and providing me the great opportunity in my education.

I would like to express my sincerest gratitude and appreciation to my advisor Professor Yasutomi Kinosada, for great support to me throughout my study with his wide knowledge, patience and understanding. His precious suggestion, encouragement and effort have provided a good basis for my dissertation.

I would like to express my gratitude to Dr. Takashi Ichinomiya for big support and helped me in all the time of research, writing paper and dissertation. Thank you so much for everything what you have done to me. Also my gratitude to Professor Satoru Hayamizu and Dr. Satoshi Tamura for your guidance. Nobuteru Miyamura, Yasuaki Sekiya and Ryouhei Kawamoto for your help in research. Also thank you to Atsuyuki Kameyama who teach me about HTK.

My gratitude to all person at office of United Graduate School of Drug Discovery and Medical Information Science, Gifu University who always help me. For big family of Indonesian Student Association at Gifu (PPI Gifu), thank you for your kind support and help. Thank you for all friendship and being family during I stayed in Gifu.

Last but not least, I would like to thank my deceased father and my mother at Padang-Indonesia, my father and mother in law, my brothers and my brothers. Special word of gratitude to my wife Roza Linda and my son M. Anugrah Syauqi Alzazirka for their patience, deep understanding, great support, help and their unconditional love. This dissertation is dedicated to booth of them.

Alwis Nazir