KNOWLEDGE MANAGEMENT ON PULMONARY COMPLICATION OF HIV PATIENTS WITH DATA MINING APPROACH

Euphrasia Susy Suhendra, PhD, prof.,
University of Gunadarma, Indonezia
Lahcen Boumedjout, Indonezia
Audrey Chandra, Indonezia

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient record, medical devices etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost savings and decision making. Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions.

AIDS, the acquired Immune Deficiency syndrome is a set of symptoms and infections resulting from the damage done to the human immune system caused by the Human Immunodeficiency Virus (HIV). AIDS is one of the most challenging medical issues all over the world. Globally, there are more than 34 million people that are infected in 2011, and this number equals to an increase by 17% from incidence in 2001. Pulmonary complication is one of the most frequent complications that occurs in HIV infection.

The aim of this study is to know the proportion of pulmonary complications which might occur in HIV patients by using data mining technique. This model development, which is a part of knowledge management, is useful to discover the relationship between HIV infections with pulmonary complications which therefore will help clinician to have an early diagnosis and prompt treatment for HIV patients. This will eventually prevent further complications. In addition, this may provide clear and useful results to the data analyst.

Key words: *knowledge management, HIV.*

Introduction

AIDS, the Acquired Immune Deficiency syndrome is a set of symptoms and infections resulting from the damage done to the human immune system caused by the Human Immunodeficiency Virus (HIV) [1]. HIV is transmitted through direct contact of mucous membrane or the blood stream with a bodily fluid containing HIV [2, 3].There is currently no vaccine or cure to both HIV and AIDS. However, treatment such as the antiretroviral therapy slows down the course of the disease thus reduces mortality and morbidity of HIV infection.

The evaluation of respiratory symptoms in HIV-infected patients can be challenging for a number of reasons. Respiratory symptoms are a frequent complaint among HIV-infected individuals and may be caused by a wide spectrum of illnesses. The spectrum of pulmonary illnesses in HIV-infected patients includes both HIV-related and non-HIV-related conditions. The HIV-associated pulmonary conditions include both opportunistic infections (OIs) and neoplasms. The OIs involve bacterial, mycobacterial, fungal, viral, and parasitic pathogens. Each of these OIs and neoplasms has a characteristic clinical and radiographic presentation. However, there can be considerable variation and overlap in these presentations. Therefore, no constellation of symptoms, physical examination findings, laboratory abnormalities, and chest radiographic findings is pathognomonic or specific for a particular disease. As a result, a definitive microbiologic or pathologic diagnosis is preferable to empiric therapy whenever possible. Diagnostic tests include cultures from sputum and blood and from respiratory specimens obtained by invasive procedures such as bronchoscopy, thoracentesis, computed tomography (CT)-guided transthoracic needle aspiration, thoracoscopy, mediastinoscopy, and open-lung biopsy.

The knowledge management system (KMS) is defined as a tool that selectively provides information relevant to the characteristics or circumstances of a clinical situation but which requires human interpretation for direct application to a specific patient. Examples of electronic KMSs include information retrieval tools and knowledge resources that consist of distilled primary literature on evidence-based practices.

The objective of this paper is to highlight the use of a predictive data mining approach Formal Concept Analysis (FCA) in the study of knowledge management in HIV – pulmonary complication patients. This model development, which is a part of knowledge management, is useful to discover the relationship between HIV infections with pulmonary complications which therefore will help clinician to have an early diagnosis and prompt treatment for HIV patients. This will eventually prevent further complications. In addition, this may provide clear and useful results to the data analyst.

Literature Review

Knowledge management and data mining techniques have been widely used in many important applications in both scientific and business domains in recent years. Knowledge management is the system and managerial approach to the gathering, management, use, analysis, sharing, and discovery of knowledge in an organization or a community in order to maximize performance (Chen, 2001). Although there is no universal definition of what constitutes knowledge, it is generally agreed there is a continuum of data, information, and knowledge. Data are mostly structured, factual, and oftentimes numeric, and reside in database management systems. Information is factual, but unstructured, and in many cases textual. Knowledge is inferential, abstract, and is needed to support decision making or hypothesis generation. The concept of knowledge has become prevalent in many disciplines and business practices. For example, information scientists consider taxonomies, subject headings, and classification schemes as representations of knowledge. Consulting firms also have been

actively promoting practices and methodologies to capture corporate knowledge assets and organizational memory. In the biomedical context, knowledge management practices often need to leverage existing clinical decision support, information retrieval, and digital library techniques to capture and deliver tacit and explicit biomedical knowledge.

Data mining is often used during the knowledge discovery process and is one of the most important subfields in knowledge management. Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns (Fayyad et al., 1996). These techniques, such as Bayesian models, decision trees, artificial neural networks, associate rule mining, and genetic algorithms, are often used to discover patterns or knowledge that are previously unknown to the system and the users (Dunham, 2002; Chen and Chau, 2004). Data mining has been used in many applications such as marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, among others.

Data mining is a process of selecting, exploring and modeling a set of data in order to discover unknown patterns or relationships which provides clear and useful results to the data analyst. The goal of predictive data mining is to derive models that can use patient specific information to predict the outcome of interest. Predictive data mining methods can be used for medical diagnosis, prognosis, treatment planning and also for general screening purposes. Criteria for good predictive data mining technique include: Good performance, transparency, ability to deal with missing data and noise (outliers), ability to work with small sample and ability to explain the decisions being made. In this research a predictive data mining technique based on Formal Concept Analysis (FCA) is used in the prediction of relationship between HIV infection with pulmonary complications.

Formal Concept Analysis (FCA: Wille, 1982) is a mathematical technique based on lattice theory. Central to FCA is the notion of a concept. The term concept comes from logic and refers to a category which can be used to classify objects. Within the FCA framework a concept is defined as having two parts: (i) the objects that can be categorized using the concept and (ii) the attributes or properties that are shared by the objects belonging to the concept (Wormuth & Becker, 2004). In addition, certain objects have certain attributes; in other words, objects are related to attributes. Taken together, the set of objects, the set of attributes and the relation defined among the objects and attributes is known as a *formal context* (Wormuth & Becker, 2004).

Methodology research

This research used a retrospective method by opening patients' medical records that are infected by Human Immunodeficiency Virus (HIV) with pulmonary infection in Atma Jaya Hospital, Jakarta, in the year of 2010-2013.

The variables in this research are age, gender, nutritional status, HIV status, and pulmonary
infection status. Pulmonary status is identified according to thorax radiography.

The data is gathered into binary, for age, gender and nutritional status, here are the classifications:

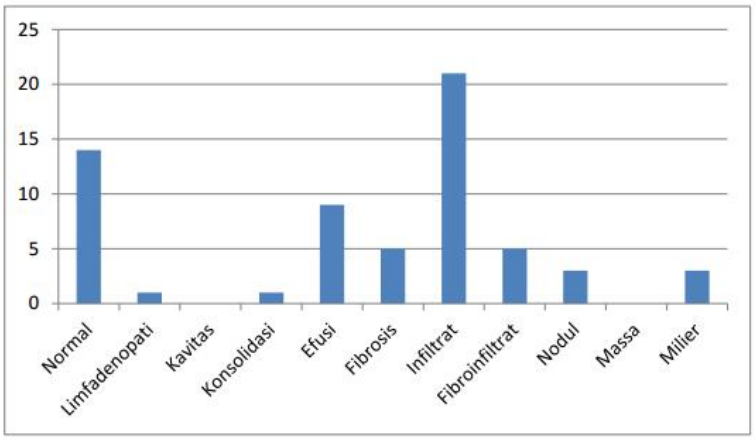
1. For "Age" variable
   - Age_geq_32 : age greater or equal to 32
   - Age_less_32: age less than 32
2. For "Gender" variable
   - Male
   - Female
3. For "Nutritional Status" variable
   - value 1 means normal or good
   - value 0 means low or NA

The next step taken is Extracting Knowledge: the use of FCA will classify data into many clusters, each cluster is called as a concept or formal concept, these concepts are in order and called as form a lattice.

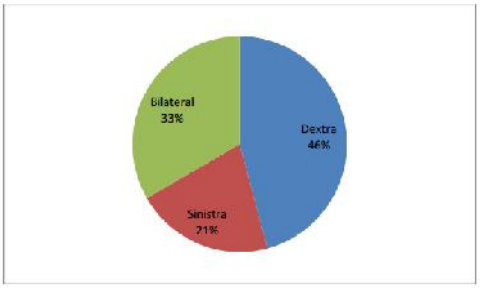
Result and Discussion

Table 1. Characteristic of Research Subject

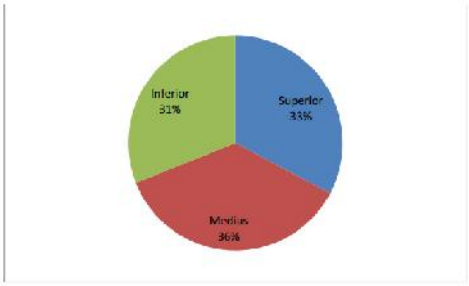
Variable		Percentage
Gender	Male	70.83 %
	Female	29.17 %
Age	< 32 years	41.67 %
	> 32 years	56.25 %
	N/A	2.08 %
Nutritional status	< 18.5	27.08 %
	> 18.5	35.42 %
	N/A	37.5 %



Picture 1. Pulmonary infection morphology according to thorax radiograhy of HIV patients



Picture 2. The area of pulmo infection based on thorax radiography of HIV patients



Picture 3.Pulmo infection location based on thorax radiography of HIV patients

By using FCA, here is the table obtained:

Tabel 2. Data Analysis with FCA

#	Antecedent	=>	Consequence	Support	Confidence
1.	{male}	=>	{Pulmo TB Status}	43.75%	61.76%
2.	{male}	=>	{Age_greq_32}	41.66%	58.82%
3.	{Pulmo TB Status}	=>	{male}	43.75%	84.0%
4.	{Pulmo TB Status}	=>	{nutritional status}	27.08%	52.0%
5.	{Pulmo TB Status}	=>	{Age_greq_32}	31.25%	60.0%
6.	{Age_less_32}	=>	{male}	29.16%	69.99%
7.	{nutritional status}	=>	{Pulmo TB Status}	27.08%	76.47%
8.	{nutritional status}	=>	{male}	29.16%	82.35%
9.	{superior}	=>	{male}	25.0%	85.71%
10.	{Pneumonia Status}	=>	{male}	29.16%	73.68%
11.	{Pneumonia Status}	=>	{Infiltration}	31.25%	78.94%
12.	{Pneumonia Status}	=>	{inferior}	25.0%	63.15%
13.	{Pneumonia Status}	=>	{Age_greq_32}	27.08%	68.42%
14.	{Age_greq_32}	=>	{male}	41.66%	71.42%
15.	{Age_greq_32}	=>	{Pulmo TB Status}	31.25%	53.57%
16.	{Infiltration}	=>	{Pneumonia Status}	31.25%	71.42%
17.	{Infiltration}	=>	{inferior}	27.08%	61.9%
18.	{Infiltration}	=>	{Age_greq_32}	27.08%	61.9%
19.	{Infiltration}	=>	{medius}	25.0%	57.14%
20.	{Infiltration}	=>	{male}	29.16%	66.66%
21.	{inferior}	=>	{Infiltration}	27.08%	86.66%
22.	{inferior}	=>	{Pneumonia Status}	25.0%	80.0%
23.	{medius}	=>	{Infiltration}	25.0%	75.0%
24.	{medius}	=>	{male}	25.0%	75.0%
25.	{Pulmo TB Status, male}	=>	{Age_greq_32}	27.08%	61.9%
26.	{Age_greq_32, male}	=>	{Pulmo TB Status}	27.08%	65.0%
27.	{Age_greq_32, Pulmo TB S...	=>	{male}	27.08%	86.66%

These association rules give some knowledge extracted from data. This knowledge expresses the relationship between different variables. For instance, if we take the rule number 26 from the table above:" (age_greq_32, male) => (Pulmo TB status)" hence we have 2 parts separated by the implication symbol " => "; whereas the left part is called antecedent or left-hand-side; and the right part is called *consequent* or right-hand-side. In this rule number 26, the antecedent is formed by 2 variables "age_greq_32" and " male" and the consequent is formed by only one variable "pulmo

TB status". Thus, the meaning of this rule is: if an HIV patient is a male and his age is greater or equal to 32 than his pulmo TB status is (or maybe will be) positive.

We have two measures to measure the interestingness of an association rule which are support and confidence.

Support is the proportion (among all the 48 patients) of patients that are male with the age greater or equal to 32 and their Pulmo TB status is postive. We can see from the data that we have 13 patients that are male with the age greater or equal to 32 and their Pulmo TB status is positive and since we have 48 patients, hence the support of this rule is:

Support = 13/48 = 0.2708 = 27.08%

In other words the support measures the proportion of patients that verify the rule among all the patients (in all the data).

Confidence is the proportion of patients that are male with the age greater or equal to 32 and their Pulmo TB status is positive among the patients that are male with the age greater or equal to 32 (with Pulmo TB status can be positive or negative). We can see from the data that we have 20 patients that are male with the age greater or equal to 32 and also we have 13 patients that are male with the age greater or equal to 32 and their Pulmo TB status is positive. Hence, the confidence of this rule is:

Confidence = 13/17 = 0.65 =65%

In other words the confidence measures the proportion of patients that verify the rule among the patients that verify only the antecedent of the rule.

The greater the support or confidence (close to 100%) then the rule is more interesting.

Conclusion

The relationship between HIV infection and pulmonary complication can be obtained by using data mining, hence we can use the knowledge management optimally. The use of data mining by FCA method will classify the data into clusters, whereas each cluster is called as concept or formal concept. The relationship analysis between HIV infection and pulmonary complication can be drawn with support and confidence value.

REFERENCE

1. Chen, H. (2001). Knowledge Management Systems: A Text Mining Perspective, Tucson, AZ: The University of Arizona.
2. Divisions of HIV/AIDS Prevention, HIV and Its Transmission, Center for Disease Control & Prevention, 2003.
3. Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From Data Mining to Knowledge Discovery in Databases," AI Magazine, 17(3), 37-54.
4. Dunham, M. H. (2002). Data Mining: Introductory and Advanced Topics, New Jersey, USA: Prentice Hall.
5. Kononenko, I, Machine learning for medical diagnosis: History, state of the art and perspective, *Artificial Intelligence in Medicine*,Vol. 23, No. 1, 2001, pp 89-109.
6. San Francisco AIDS Foundation, How HIV is Spread, 2006
7. Weiss RA, How does HIV cause AIDS, *Science Journal*, Vol. 260, No. 5112, 1993, pp 1273-9. PMID8493571.
   Wille, R (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival (ed.), *Ordered sets*. (pp. 445-470), DordrechtBoston: Reidel
8. Wormuth, B & Becker, P (2004). Introduction to formal concept analysis. Paper presented at the 2nd International Conference of Formal Concept Analysis, 2004, Sydney Australia.