Iranian Journal Of Medical Informatics 2016; 5(1): 24-27

www.ijmi.ir

RESEARCH Open Access

Detecting Hidden Patterns from Brucellosis Patients' Data in Khorasan Razavi Province Using Appriori Algorithm

Soheil Hashtarkhani [1], Ali Akbar Heidari[2], Kobra Etminani *[1]

1- Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Science, Mashhad, Iran.
2- Department of Infectious Disease, Faculty of Medicine, Mashhad University of Medical Science, Mashhad, Iran.

ABSTRACT

Introduction: Brucellosis is a transmissible disease between humans and animals through infected animals and their products.The disease exist in most parts of the world especially in developing countries.because of the serious impact of the disease in public health and socio-economical status, controling the disease is very important in developing countries. The purpose of this article is to identify hidden patterns and relations between brucellosis patients which can be benefitial for physicians in diagnosis process.

Material and Methods: This study is a retrospective study of data collected from brucellosis Khorasan Razavi province recorded at the health center, have been used. Due to differences in format and number of features collected during different years, before processing operations carried out in several stages to the same data. Fields associated with different methods and with expert opinion was converted into discrete fields and fields lost was estimated using the EM algorithm. APPIORI algorithm analysis was performed using the hidden relationships between data found that significant relationships were infected with expert opinion.

Results: Among the 163 relationship with over 7.0 Conficence rate which Weka software was discovered, by the application in consultation with an infectious disease expert, 10 clinically significant relationship was reported.

Conclusion: Diagnosig brucellosis is realy difficult to physicions because of its vagious nature and symptoms. Because many unknown relationships between risk factors and demographic characteristics of the patients, the use of data mining concepts, especially in the medical data is beneficial because usually high volume assumptions are available. further studies can test the validity of these rules like Randomize Control Trial studies.

Article History
Received: 2016-03-22
Accepted: 2016-05-31
Published: 2016-07-15

Keywords
Brucellosis
Data Mining
Appriori

ⓘ*Please cite this article as:*
Hashtarkhani S, Heidari AA, Etminani K. Detecting Hidden Patterns from Brucellosis Patients' Data in Khorasan Razavi Province Using Appriori Algorithm. Iran J Med Inform. 2016; 5(1): 24- 27. DOI: 10.24200/ijmi.v5i0.113

* Corresponding Author: K Etminani, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Science, Mashhad, Iran (Email: EtminaniK@mums.ac.ir)

INTRODUCTION

Brucellosis or brucellosis is the most common infection between humans and livestock that is transmitted through contaminated animals and their products. Each year, more than 500,000 new cases of illness occur in the world, with different incidence rates in different regions [1, 2]. According to the World Health Organization, 500,000 human cases of brucellosis are reported annually, with only 5 to 10 percent of cases reported even in advanced countries. The average reported annual incidence of brucellosis in the country has been 43/44 per 100,000 since 1991 to 1998 [3].

In this study, data mining techniques were used to explore the risk factors and hidden patterns of the disease. Data mining is a process for extracting knowledge and rules from a bulky data set. Data mining techniques are known as a successful solution in the field of medical sciences [4]. Extracting associative rules is one of the most important data mining techniques for generating strong rules from databases. This technique has various algorithms including Apriori algorithm, Eclat algorithm and FP-growth algorithm [5]. In the field of data mining, data on patients with brucellosis has not been studied yet, but different studies have been conducted for other diseases. Desikan et al. In a study of the application of data mining in health centers and its impact on health information management. The author describes important models for discovering various diseases and important relationships in hospital data, including data used in HL7, EHR, EMR, and so on. Data mining can also help in detecting financial mismanagement, managing financial resources and discovering and treating various diseases [6].

In a study by Mao et al., Data from patients in the intensive care unit was implemented to predict the hazard. In this study, using a combination of machine learning methods, including logistic regression, 42% of patients transferred to the intensive care unit and 55% of the patients died of the disease were warned [7]. Another study on the SEER (Surveillance Epidemiology and End Results) data collection was developed to extract association's rules using the Apriori algorithm. In this study, the relationship between the therapeutic and mortality characteristics of patients with breast cancer was analyzed based on other medical characteristics of patients and significant and meaningful relationships were reported [8].

Among the important studies that have been done in discovering association affairs in Iran, we can use data mining to explore the risk factors for gastric cancer by Mahmoudi et al. Apriori was used in this study and it was shown that people with cardiovascular disease are less likely to develop gastric cancer [9]. Another study by Attici et al. Used the Apriori algorithm to discover hidden patterns in the data set of patients with breast cancer. In this study, the EM algorithm was used to estimate missing items, and 100 relationships with a high confidence coefficient of 0.9 were detected, of which 10 were detected by a significant physician [9]. Maltese fever is one of the most common diseases in our country, and due to different clinical faces, doctors have difficulty diagnosing. Therefore, more accurate and scientific methods are needed. To explore the hidden relationships and risk factors of the disease, researchers have suggested using these techniques to help better recognize and diagnose disease, given the power of data mining techniques.

MATERIAL AND METHODS

In this study, the data of patients with brucellosis were collected from the beginning of April 2009 to the end of March 2013 to health centers, clinics, outpatient clinics and hospitals in Khorasan Razavi province. After the tests Necessary and confirmation of illness in them, their information was registered in special forms and required treatment. The entry requirement for the study was based on the national standard definition, all subjects with suspected clinical symptoms that had a Wright tittle or coombsorth greater than or equal to 1.80, or a 2Me header greater than or equal to 1.40. Patient information fields are shown in Table 1.

This study was carried out in four steps:

In the first stage, the pre-processing operation was performed and the data was cleaned. At this point, some features such as the year of the disease, the city, the date of the incident, and other issues that were unimportant for the discovery of laws were removed. Some features, like nationality, were abnormally abnormal (only 4 Afghan patients and the rest of Iran). All features with continuous numerical value were categorized into discrete categories, for example, the age attribute according to similar studies and expert opinion divided into three categories and the interval between incidence and diagnosis of the disease in two categories below one month and more than one month It was decomposed. Characteristics such as the amount of non-pasteurized dairy products and clinical symptoms have been converted into several binary variables. Also, some fields have been recorded in different formats or units during the years of the study, making the units homogeneous and converting for these variables. Another common problem in data analysis is the missing variables. Fields with more than 15% lost items were excluded. This phase was performed with IBM SPSS 21 software and the study variables are shown in Table 1.

Table 1: Information fields of patients with brucellosis in Khorasan Razavi province

Row	Name	Domain
1	Address	Rural, urban
2	Age	Less than 40- Between 40 and 60- Over 40
3	Gender	Male- female
4	pregnancy	Yes- No
5	Update interval to detection	Less than a month - more than a month
6	job	Housewife-Farmer-Farmer-Slaughterhouse worker-Butcher-employee and etc.
7	The history of animal contact in the past year	Yes- No
8	History of non-pasteurized milk	Yes- No
9	History of non-pasteurized cheese	Yes- No
10	History of non-pasteurized oatmeal	Yes- No
11	History of non-pasteurized ice-cream	Yes- No

12	History of non-pasteurized creamy	Yes- No
13	History of non-pasteurized butter	Yes- No
14	History of non-pasteurized Ecstasy	Yes- No
15	fever	Yes- No
16	Descendants	Yes- No
17	The size of the spleen	Yes- No
18	The size of the liver	Yes- No
19	Muscle and bone pain	Yes- No
20	Depression	Yes- No
21	Adenopathy	Yes- No
22	Weight Loss	Yes- No
23	The result of Wright's diagnostic test	۱/۲۰or۱/۴۰ or ۱/۸۰ ۱/۱۶۰or ۱/۳۲۰
24	The result of the Coombs Wright diagnostic test	۱/۲۰or۱/۴۰ or ۱/۸۰ or ۱/۱۶۰or۱/۳۲۰
25	The result of the 2nd May diagnostic test of the disease	۱/۲۰or۱/۴۰ or ۱/۸۰ or ۱/۱۶۰or۱/۳۲۰
26	The result of the 2nd May diagnostic test of the disease	New- Treatment failure
27	Hospitalization	Yes- No
28	Other family members in the past year	Yes- No

In the second step, the data was processed using WEKA software and APPRIORI algorithm. The apriori algorithm was first introduced by Agrawal and Sirkant. This algorithm provides repetitive itemsetting according to the minimum backup level. In the first transition, the algorithm makes candidate candidate 1-itemset. Then, those who repeat themselves are less than the minimum backup level. Then, the 1-item algorithm combines candidate items to create 2-itemset and re-greed again. These steps are repeated in the same way so that no items can be produced (10). In the implementation of the algorithm, the degree of assurance and the minimum assurance were used 0.2 and 0.7.

In the third step, the rules extracted from the previous step to refine the rules that are not of interest are refined. For example, the rules on the right of their type of place of residence or gender will certainly not have clinical application. So, using the Microsoft Excel 2013 software, a code was written to exclude the rules to the right of them containing the fields of interest. Thus, out of 1,000 laws produced at the previous stage, 879 laws were removed and 121 remained in force.

In the fourth stage, the rules were chosen for the infectious specialist doctor to select the most meaningful and relevant rules. 9 laws were chosen as meaningful among 121 acts.

RESULTS

Information on 5743 cases of mumps has entered the data mining stage during the five years from the preprocessing stage. Among these data, 85% of the patients were in the village and 15% were in the city. Women (43.1%) and men (56.9%) had cases. Housewives (33.9%) and farmer-livestock (27.8%) had the highest incidence and the average age of the patients was 33.04 ± 18.1. 77.2% had a history of non-pasteurized dairy consumption It was found that milk (91.4%) and cheese (21.4%) were the most consumed.

Association rules are a powerful method in data mining that can detect hidden rules in data. Rules derived from an algorithm approved by a specialist physician are shown in Table 2. According to these rules, most patients with fever, loss of appetite and weight loss had muscle and bone pain. Most patients aged 11 to 20 are men. The majority of patients with a 2ME titre of 1.80 were the result of a Wright titre of 1.160. Most patients who were newly ill were not hospitalized. Most patients who did not cure had no weight loss. Most patients admitted to the hospital had symptoms of weight loss, adenopathy, and liver enlargement. Patients with febrile seizures often had fever. Patients who had large spleen often had liver enlargement. Patients diagnosed with their illness for more than a month usually develop fever and musculoskeletal pain.

Table 2: Final rules extracted from patients with brucellosis

Row	Rule	Result	Confident Degree	Lift	Lev.	Conv.
1	If the patient has fever, loss of appetite, weight loss	Musculoskeletal pain	0.78	1.57	0.04	2.29
2	Age from 11 to 20 years	Gender (Male)	0.71	1.25	0.03	1.47
3	The result of a 2Me test is 1.80	The result of the Wright test 1.160	0.7	2.64	0.09	2.37
4	If the patient has not been hospitalized	A new case of disease. (No treatment failure)	0.98	۱	۰	1.05
5	If the disease does not lose weight	There is no treatment failure	0.98	۱	۰	1.16
6	If the patient has clinical signs of weight loss, adenopathy has a large liver	Hospitalization	۱	۱	۰	4.18
7	If the patient has an insult.	fever in the patient	0.85	1.37	0.09	2.51
8	A disease that affects the size of the spleen	They also have a large liver	۱	۱	۰	4.18
9	If the incidence is longer than the diagnosis of the disease for more than a month	The patient has fever and musculoskeletal pain.	0.79	1.3	0.05	1.87

DISCUSSION

Current study on actual data of brucellosis patients over 5 years with the aim of familiarizing the medical community is with one of the methods of knowledge extraction. The use of association rules when there is no hypothesis about relationships between variables can be very useful and reveal laws that are hidden from the viewpoint of physicians and health professionals. No similar studies have been found to use data mining methods to explore

the knowledge of patients with brucellosis. One of the reasons for this can be due to the newness of these techniques in the medical community, and the other reason is the obsolete disease in the developed world. One of the advantages of association laws is that the probability of the occurrence and strength of each law is mentioned along with that which helps the physician in choosing meaningful rules. In this study, the rules were arranged with a descending confidence level and written in a colloquial language to be easily understood and valued comfortably.

Since brucellosis is a disease that has different faces and various epidemiological factors affect the incidence of disease and severity and symptoms, it should be noted that the results are exclusive to Khorasan Razavi province and similar studies should be done for other areas and Compare with the rules of this study.

The Apriori algorithm is one of the most important data mining algorithms in the domain of the discovery of associative rules. The rules have shown that patients who have been admitted to the hospital because of dangerous complications are mostly ill for the first time and have not been treated for failure. This is a result of the fact that the dangerous complications of hospitalization are less common in patients with failure of treatment.

In the extracted rules, defoliation and weight loss were identified as very important and influential variables. Weight loss and euthanasia were common side effects in patients with fever, hospitalized, and treatment failure.

In Law No. 5, we conclude that there seems to be a relationship between weight loss and treatment failure, so that in the majority of cases where the patient did not cure, he did not lose weight. This law can be considered as a hypothesis in scientific research, and no study has been conducted in this regard.

In the Law, six important risk factors have been extracted in hospital admissions patients. As it is seen, adenopathy, liver enlargement and weight loss are one of the important risk factors for hospitalization in hospitals, which should be considered in clinical examinations.

CONCLUSION

The fever and musculoskeletal pain complications were seen in accordance with No. 9 in patients with a diagnosis until treatment for more than a month. This suggests that in cases when the bacteria is chronic in the long term in the body, there are side effects of musculoskeletal pain and fever in patients that should be considered in clinical examinations. The use of data mining concepts, especially in medical data, is particularly useful when it is usually of a high volume. These rules, in fact, determine the assumptions of subsequent studies and, by conducting further studies in other ways, including performing RCTs It are possible to reject or prove these assumptions.

REFERENCES

1. Skalsky K, Yahav D, Bishara J, Pitlik S, Leibovici L, Paul M. Treatment of human brucellosis: Systematic review and meta-analysis of randomised controlled trials. BMJ. 2008; 336(7646): 701-4. PMID: 18321957 DOI: 10.1136/bmj.39497.500903.25 [PubMed]
2. Seleem MN, Boyle SM, Sriranganathan N. Brucellosis: a re-emerging zoonosis. Vet Microbiol. 2010; 140(3-4): 392-8. PMID: 19604656 DOI: 10.1016/j.vetmic.2009.06.021 [PubMed]
3. Mostafavi E, Asmand M. Trend of brucellosis in Iran from 1991 to 2008. Iranian Journal of Epidemiology. 2012; 8(1): 94-101.
4. Richards G, Rayward-Smith VJ, Sönksen P, Carey S, Weng C. Data mining for indicators of early mortality in a database of clinical records. Artif Intell Med. 2001; 22(3): 215-31. PMID: 11377148 [PubMed]
5. Jain D, Gautam S. Implementation of Apriori algorithm in health care sector: A survey. International Journal of Computer Science and Communication Engineering. 2013; 2(4): 22-8.
6. Desikan P, Hsu K-W, Srivastava J. Data mining for healthcare management. Proceeding of International conference of Data Mining; USA. 2011.
7. Mao Y, Chen Y, Hackmann G, Chen M, Lu C, Kollef M, et al., editors. Medical data mining for early deterioration warning in general hospital wards. 11th International Conference on Data Mining Workshops; IEEE. 2011.
8. Fan Q, Zhu C-J, Xiao J-Y, Wang B-H, Yin L, Xu X-L, et al., editors. An application of apriori algorithm in SEER breast cancer data. Proceeding of International Conference on Artificial Intelligence and Computational Intelligence (AICI); IEEE. 2010.
9. Mahmoodi SA, Mirzaei K, Mahmoodi SM. Using association rules for the detection of risk factors in gastric cancer. Journal of Health and Biomedical Informatics. 2015; 1(2): 95-103.
10. Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining: A general survey and comparison. ACM SIGKDD Explorations Newsletter. 2000; 2(1): 58-64.