Discovering Multifactorial Associations with the Development of Age-Related Cataract Using Contrast Mining

Murugesan Raju, PhD
MU Informatics Institute
University of Missouri
Columbia, MO, USA
rajum@health.missouri.edu

Danlu Liu, MS
Dept of Computer Science
University of Missouri
Columbia, MO, USA
dltb9@mail.missouri.edu

Frederick W.Fraunfelder, MD
Department of Opthalmology
University of Missouri
Columbia, MO, USA
fraunfelderf@health.missouri.edu

Chi-Ren Shyu, PhD
MU Informatics Institute
University of Missour
Columbia, MO, USA
shyuc@missouri.edu

Abstract

Cataract is a cloudiness of eye lens and studies have reported many risk factors for the development of cataract. However, the cumulative effect of multiple factors along with clinical and systemic disease conditions have not been adequately tested due to a limitation in methodology. The collection of a large volume of Electronic Health Records (EHR) offers an opportunity to apply computational tools for knowledge discovery in databases (KDD) process which enable to discover and extract hidden patterns and relationships among a large number of variables. This approach is possible because of the computational friendly EHR database such as the Cerner Health Facts Database. The main goal of this paper is to investigate the factors which are associated with the development of age-related cataracts using EHR data. This study demonstrates the potential of applying data mining tools for risk assessments using large-scale EHR data.

Keywords: Eye diseases, Cataract Electronic Health Records and Data Mining

I. INTRODUCTION

A discovery of knowledge from a database is an interactive process between human and computer. First, we (Human) will select the data and define analysis goals and then ask computer algorithms to search the data for models and pattern within the defined goal of analysis and then human will interpret the results and its relevance to the clinical application or human wellbeing in general.

Cataract is the leading cause of blindness in the United States, but very little is known about the factors that are associated with the development of cataracts [1]. The juvenile cataract has strong hereditary component, but age-related cataract has mixed components including environmental factors, background radiation, and UV exposure [2]. Further, common health conditions such as diabetics, glaucoma, arthritis, hypertension, autoimmune disease, and medications including the use of corticosteroids have been reported to be associated with the development of age related cataracts.

Health Facts database captures and stores de-identified, longitudinal information on patients including demographics, type of encounter, diagnosis, medications, procedures, laboratory tests, hospital information, and billing details. The database contains more than 60 million patient records. Using the Health Facts database, we have reported an association between lacrimal disorder and development of cataract in our previous study [3]. Further to validate this association, we are applying data mining methods in this present study. Association and contrast mining have been widely employed in many research areas [4,5]. The association rule mining has been successfully implemented by the retailor applications to determine frequently selling item(s) for better business model. The core concepts can be applied to many sectors including healthcare systems. The association rule mining has several advantages over classical statistical model including the analytics results that can be easily interpretable and provides better characterization of an association with multiple factors [6,7]. In this study, we demonstrated the application of data mining methods to investigate risk factors and its association with the development of age-related cataracts.

II. METHODS

Knowledge Discovery Process: The contrast mining algorithm [8] basically shows patterns and models with the frequently seen condition in one group but relatively

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

rare in the other group. For example, in cataract dataset, the contrasts between cataract and non-cataract phenotype is our interest. we also performed subgroup analysis to see the difference between lacrimal disorders with cataract phenotype and lacrimal disorder without cataract because the preliminary data showed a strong association between cataract phenotype and lacrimal disorders [3].

To explain our method more clearly and easily to understand, some terms are defined as follow. Given a class D, a pattern (P) in D is defined as the combination of the items in the records. The frequency of pattern P in D is calculated by using statistical methods. The $Support$ of P in D is defined as the proportion of records in D which contains P with total records in D. A frequent pattern P is a pattern whose support is greater than a threshold. For two classes D_1 and D_2, a contrast pattern is the pattern whose $Support$ differs greatly between these two classes. The degree of their differences can be measured via the $Growth$ which is the ratio of supports between two classes $Growth = \ Support_1/Support_2$.

The eye visit dataset contains a total of 18 attributes along with demographic information. Data in Health Facts is extracted directly from the EHR from hospitals in which Cerner has a data use agreement. Patient's data from EHR was preprocessed for data mining analysis. We also applied contrast data mining method to identify differential growth between the subset of patients' groups. The dataset was divided into two groups such as class 1 (*D1*) cataract group and class 2 (*D2*) non-cataract group. There are 152,183 patients' data in *D1* and 677,943 patients' data in *D2*. Using Big Data technologies in pattern mining under a distributed computing environment, we perform a series of exploratory mining processes to identify subgroups that are highly contrasted. For example, in the subgroup analysis, there are two sub-groups highlighted by the method: sub-class 1 (lacrimal disorder, no cataract, and no smoking history) and sub-class 2 (lacrimal disorder, cataract, and smoking history).

In order to discover the contrast patterns, we apply the frequent pattern mining first and then calculate the growth of these frequent patterns. The growth of them can be calculated via the formula we talked above. In the pattern mining area, the Apriori algorithm is mostly used in the frequent pattern mining area. Thus, we use the Apriori method to analyze the cataract data. This general idea of this algorithm is to scan the data to discover the frequent patterns with the length *1*. In the next iteration, it generates the patterns with the length *2* based on the previous length *1* patterns. Repeat these work until no new pattern is found. In our method, we use it to find the frequent patterns and then calculate the growth of those frequent patterns. The institutional review board (IRB) at the University of Missouri approved the study protocol (IRB #2006793 HS) The subsequent section describes the results obtained from contrast mining analysis.

III. RESULTS AND DISCUSSION

We selected eye visit dataset from Health Facts database and preprocessed the data and grouped into cataract verses non-cataract for analysis. This dataset was analyzed using our customized development of contrast mining using Apache Spark to identify highly contrasted subpopulation groups. Patterns between the two groups were extracted through association/ co-occurrence of encounter, medication, diagnosis, laboratory orders, pharmacy, procedure, and other selected attributes. Among all the highly contrasting patterns, those with statistically significant between sub-populations were considered. Though we extracted hundreds of contrast groups, we listed only top 10 association set in Table 1. Each rule has "credentials" for the quality of contrast, such as support, confidence lift, growth rate, etc. For example: {Lacrimal Disorder=Yes, Glaucoma=Yes, Type II Diabetes=Yes, } => {CATARACT=Yes}. Similar to our previous study, lacrimal disorder showed a strong association with cataract phenotype [3].

For subgroup analysis, we divided the patient's data into two classes: Class 1 lacrimal disorder, no cataract, and no smoking history; Class 2 lacrimal disorder, cataract, smoking history. The contrast mining algorithm extracted highly contrasted subpopulation groups with the frequently seen condition in one group but relatively rare in the other group. In this analysis, Class 2 (patients diagnosed with lacrimal disorder, cataract, and smoking history) have high contrasting patterns with glaucoma and hypertension (growth 24.702; Confidence 0.873 and support 0.109) compared to Class 1 (growth 0.040; Confidence 0.126 and support 0.004) suggesting that patients diagnosed with lacrimal disorder along with glaucoma and hypertension have much higher chance of developing cataract.

Table 1: Association rules “cataract vs non-cataract”:

Rules	Class1_Support	Class2_Support	Growth	Support diff.
Glaucoma, Alcohol, Hypertension, Lacrimal disorder	0.010907	0.001298	8.403350	0.0096098
Lacrimal disorder, Alcohol, Glaucoma	0.013838	0.001678	8.244097	0.01215
Lacrimal disorder, Alcohol, Caucasian, Glaucoma	0.010388	0.001385	7.500553	0.00900
Female, Lipoid metabolism disorder, Lacrimal disorder, Glaucoma	0.010684	0.001460	7.316651	0.00922
Glaucoma, Lacrimal disorder, Type two	0.010171	0.001398	7.274274	0.00877
Hypertension, Lacrimal disorder, Glaucoma, lipoid metabolism disorder	0.013792	0.001908	7.226120	0.01188
Glaucoma, Lipoid metabolism disorder, Lacrimal disorder	0.015777	0.002196	7.183308	0.01358
Hypertension, Lacrimal disorder, Glaucoma	0.018688	0.002687	6.953576	0.01600
Hypertension, Lacrimal disorder, Female, Glaucoma	0.012609	0.001846	6.828065	0.01076
Glaucoma, Lacrimal disorder	0.027650	0.004223	6.547589	0.02342

IV. Conclusion

This study suggests that development of age-related cataract and lacrimal disorder may have a common pathway for disease manifestation. Further, this study demonstrated that Lacrimal disorder, Glaucoma, hypertension, Lipoid metabolism disorder and alcohol history have a strong association with cataract phenotype. The study also demonstrates that data mining tools can be effectively applied to investigate risk factors from electronic health records.

Acknowledgement

Funding sources for this work include NIH T32 training grants 5T32LM012410-02 (MR), and Shumaker Endowment for Biomedical Informatics (DL, CRS). The high performance computing infrastructure is supported by the National Science Foundation under grant number CNS-1429294.

References

[1] WHO, Global Data on Visual Impairments 2010 (WHO/NMH/PBD/12.01) world Health Organization, Geneva. 2010.

[2] West, S.K., et al., Model of risk of cortical cataract in the US population with exposure to increased ultraviolet radiation due to stratospheric ozone depletion. Am J Epidemiol, 2005. 162(11): p. 1080-8.

[3] M. Raju, M. Chisholm, A. S. M. Mosa, C. R. Shyu and F. W. Faunfelder. Investigating Risk Factors for Cataract Using the Cerner Health Facts® Database. *Journal of Eye & Cataract Surgery*, 2017, 3:19

[4] Pudil P., Bláha S., Novoičová J. (1988) Credits — Software package for solving pattern recognition and diagnostic problems. In: Kittler J. (eds) *Pattern Recognition. Lecture Notes in Computer Science*, vol 301. Springer, Berlin, Heidelberg

[5] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Database," *VLDB* 94, Morgan Kaufmann, 1994, pp. 407-419

[6] Witten I.H., Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*, Second Edition (Morgan Kaufmann Series in Data Management Systems) Morgan Kaufmann Publishers Inc.; San Francisco, CA, USA: 2005

[7] Sheets, L, Petroski, G., Zhuang, Y., Phinney, M. A., Ge, B, Parker, J., Shyu, C. R., Combining Contrast Mining with Logistic Regression to Predict Healthcare Utilization in a Managed Care Population, *Applied Clinical Informatics*, 2017, 8: 430-446

[8] G. Dong and J. Bailey, Contrast Data Mining: Concepts, Algorithms, and Applications. Chapman & Hall/CRC, 2012, p. 434.