Identifying HotSpots in Five Year Survival Electronic Health Records of Older Adults

Ankit Agrawal[1], Jason Scott Mathias[2], David Baker[2], Alok Choudhary[1]
[1]Department of Electrical Engineering and Computer Science
[2]Northwestern Memorial Hospital
Northwestern University
2145 Sheridan Rd, Evanston, IL 60201, USA
Email: ankitag@eecs.northwestern.edu

***Abstract*—Understanding the prognosis of older adults is a big challenge in healthcare research, especially since very little is known about how different comorbidities interact and influence the prognosis. Recently, a electronic healthcare records dataset of 24 patient attributes from Northwestern Memorial Hospital was used to develop predictive models for five year survival outcome. In this study we analyze the same data for discovering hotspots with respect to five year survival using association rule mining techniques. The goal here is to identify characteristics of patient segments where the five year survival fraction is significantly lower/higher than the survival fraction across the entire dataset. A two-stage post-processing procedure was used to identify non-redundant rules. The resulting rules conform with existing biomedical knowledge and provide interesting insights into prognosis of older adults. Incorporating such information into clinical decision making could advance person-centered healthcare by encouraging optimal use of healthcare services to those patients most likely to benefit.**

***Keywords*-Association rule mining; hotspots; prognosis, older adults;**

I. Introduction

Professional guidelines recommend that providers consider patient prognosis in order to optimize evidence-based, patient-centered preventive service use. The American Geriatrics Society recommends that patients and providers consider prognosis when making decisions about how aggressively to treat hyperglycemia, blood pressure, or lipid disorders in patients with diabetes [1]. Several other professional society guidelines recommend that patients and providers consider prognosis when making decisions about cessation of cancer screening [2]. In the absence of accurate prognostic information, providers often ignore these recommendations, instead relying on age-based cutoffs to make medical decisions [3]. This ultimately leads to poor quality care. For example, consider the following patients:

- **Patient A**: A 70 year-old man with hypertension, kidney disease, diabetes, a consistently uncontrolled blood pressure, and a slightly elevated Creatinine.
- **Patient B**: A 80 year-old man in excellent health with no chronic medical conditions, excellent blood pressure, and normal laboratory studies.

Patient B is healthy, and despite his advanced age, is likely to live long enough to benefit from cancer screening. Patient A on the other hand is ill, and therefore likely to suffer more harms than receive benefits from cancer screening. In current practice, however, Patient A is more likely to receive colon cancer screening than Patient B simply because Patient A is younger. Using age- and population-based guidelines leads to both overuse of cancer screening (Patient A) and underuse of cancer screening (Patient B), representing a significant quality problem.

It is therefore necessary to have a better understanding of the prognosis of an individual patient while making decisions about the course of healthcare delivery. This is even more critical for older adults, since advanced age is many times accompanied by multiple healthcare conditions which can interact among themselves in a myriad of ways and influence the overall prognosis of the patient. Such complex interactions and their impact of the overall survival is far from being well-understood.

In this era of "big data", huge amounts of various personalized information such as patients' electronic health records (EHR) is increasingly becoming available. In general, our ability of generate, collect, and store more and more data has advanced tremendously, but the analytical capability has not been able to keep pace with it. This is true in practically all fields, and the field of medicine and healthcare is no exception to it, where the Fourth paradigm of science (data-driven analytics) is increasingly becoming popular and has led to the emergence of the new field of healthcare informatics. The Fourth paradigm of science [4] unifies the previous three paradigms of science – namely theory, experiment, and simulation/computation. The need for data science in healthcare has also been emphasized by large-scale federal initiatives both in the US and elsewhere.

The rich clinical data available within the EHR allows for a more comprehensive assessment of an individual patient's prognosis. There have been many works dealing with application of data-driven analytics in healthcare [5], [6], [7], [8], [9], [10]. In one such work, a high-dimensional EHR database from Northwestern Memorial Hospital was analyzed to build predictive models of five year survival using ensemble predictive mining techniques [9] and an online five year life expectancy calculator was developed deploying those models [11] using 24 patient attributes. The ensemble predictive model was shown to outperform other better known prognostic indices, like the Charlson Comorbidity Index [12] and Walter Life Expectancy Index [13].

In this work, we analyze the same dataset used in [9] for the purpose of identifying hotspots with respect to five year survival using association rule mining techniques. The goal here therefore, is to automatically discover segments of this data where the survival fraction is significantly higher or significantly lower than the survival fraction across the entire dataset. The application of association rule mining techniques can result in a large number of association rules, but many of them may be redundant. We used a two-step semi-manual post-processing procedure to eliminate the redundant rules. The non-redundant rules discovered as a result of the current analysis represent interesting insights into the prognosis of older adults.

The rest of the paper is organized as follows: Section II gives an overview of association rule mining and describes the HotSpot algorithm used in this work, followed by a description of the data and its attributes in Section III. Experiments and results are presented in Section IV, and finally the conclusion and future work in Section V.

II. Association Rule Mining

Association rule mining is a class of machine learning techniques that are useful to discover patterns in the data. In contrast to predictive modeling where the goal is to predict the value of a target variable based on the values of input variables, the goal in association rule mining is bottom-up discovery of associations among the attributes.

It is formally stated as follows [14]: Let I be a set of n binary attributes called items. Let T be a set of transactions. Each transaction in T contains a subset of the items in I. A rule is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \phi$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively. A target/outcome attribute, if available, is fixed as the RHS set Y, and thus association rule mining would discover segments X where the average value of the target attribute (or fraction of the instances with the target attribute having the value of interest, in case the target attribute is nominal) is significantly higher or lower from that across the entire dataset. A commonly given example from market basket analysis is of the rule $\{Bread\} \Rightarrow \{Butter\}$, meaning that customers who buy bread also buy butter.

Some of the association rule mining algorithms only work with binary attributes, indicating the presence/absence of the item in the transaction. To be able to use nominal attributes (having multiple but finite possible values) and numeric attributes (having continuous range of numerical values) with such algorithms, it is often necessary to derive binary attributes first for the purpose of association rule mining. Popular algorithms for association rule mining include Apriori [15], Eclat [16], FP-Growth [17].

Note that the problem of association rule mining is tantamount to the inverse question of retrieval in databases. In database retrieval, the input query is the segment definition in terms of attribute values, and the database system returns the segment, whose properties can subsequently be analyzed. However, such database retrieval cannot automatically discover segments with high fraction of a target attribute value of interest, which is exactly what association rule mining can do. Let us see it in context of the current EHR data. In this case, we have several patient attributes including an binary outcome/target attribute (five year survival). Let us say the average fraction of patients who did not survive at least five years is f. It would then be of interest to automatically discover from the data under what conditions – as defined by the combination of patient attribute-values – is the survival time f' significantly higher or significantly lower than f.

HotSpot Algorithm

This is an association rule mining algorithm where the RHS or consequent is fixed to the target attribute. It is a simple yet powerful algorithm which can be used for segmentation with both nominal and numeric target attributes. It uses a greedy approach to construct the association rules in a tree-like fashion, where the depth-first search is constrained by the three parameters. The root of the tree consists of the entire data where the average fraction of the target attribute is f. A branch is added if the algorithm is able to find segment defined by fixing the value of a single attribute such that the resulting fraction of the target attribute in that segment f' is significantly higher or lower.

1) **Maximum branching factor**: This represents the maximum number of children nodes to consider at each node, and controls the amount of search performed, since the algorithm uses a greedy search. A higher value of this parameter would lead to a greater amount of search and potentially more number of rules.
2) **Minimum improvement in target value**: The algorithm must be able to find at least this much improvement in the target value of the resulting segment in order to add a new branch. The improvement in the target value can be defined as either an increase or a decrease in the average target value (in case of numeric targets) or target fraction (in case of nominal targets). Higher values would contrain the search and result in less number of rules.
3) **Minimum segment size**: The size of the resulting segment must be at least this much in order to add a new branch. This relates to the support or generalizability of the rule. Lower values of support would result in discovery of more rules. Very low values would give many noisy rules that are applicable to only a few instances, while very high values may give very few or no rules at all.

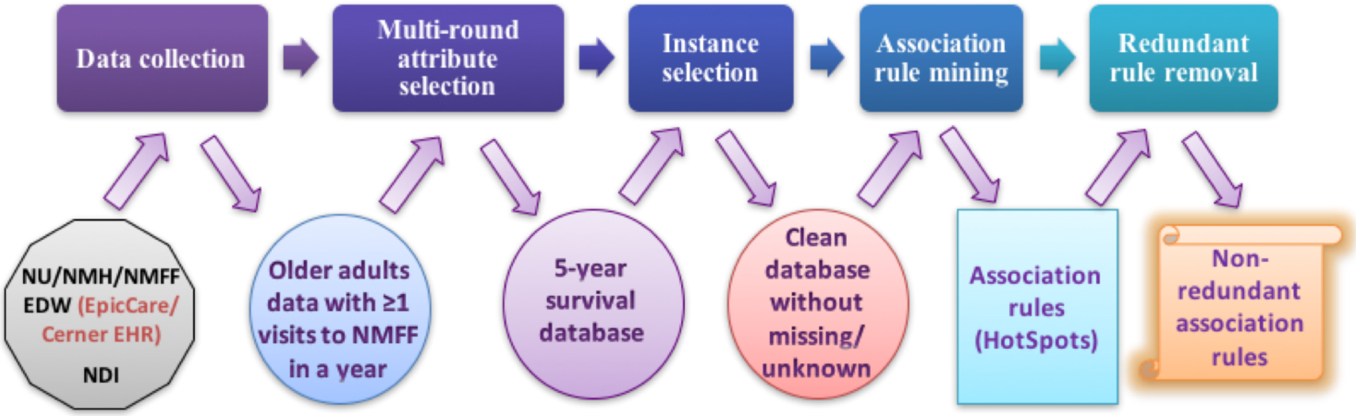
Data collection
Multi-round attribute selection
Instance selection
Association rule mining
Redundant rule removal
NU/NMH/NMFF EDW (EpicCare/ Cerner EHR)
NDI
Older adults data with ≥1 visits to NMFF in a year
5-year survival database
Clean database without missing/ unknown
Association rules (HotSpots)
Non-redundant association rules

Fig. 1. The data mining workflow used in this work. EHR data is extracted for patients with at least one visit to NMFF in 2003, nearly 1000 attributes were derived, in a bid to provide domain knowledge to the predictive model. Multiple rounds of automatic attribute selection interleaved with manual inspection and selection were performed to identify a small non-redundant attribute set. All instances with any missing/unknown values are then removed to get the dataset for association rule mining. HotSpot algorithm was used for association rule mining, and redundant rule removal procedure performed to get the final list of non-redundant association rules.

The HotSpot algorithm itself is quite straightforward. It begins with the entire dataset at the top, and goes down the data in a depth-first fashion using a greedy approach, i.e., by identifying an attribute which would give the maximum improvement in target value subject to the above constraints and subsequently branching on that to create a new child node. It then tries the same thing at every node recursively. Each node of the resulting tree corresponds to an association rule represented by the corresponding segment. The HotSpot algorithm has previously been used for finding association rules in a lung cancer dataset [8], [18].

We use the implementation of the HotSpot algorithm provided in the WEKA data mining toolkit [19].

III. EHR Data of Older Adults used for Association Rule Mining

In this work, we have used the same dataset as used in [9]. Some key features of the data are again described here. Patient-level EHR data was extracted from the Enterprise Data Warehouse (EDW) implemented by Northwestern University (NU), Northwestern Memorial Hospital (NMH), and Northwestern Medical Faculty Foundation (NMFF) using Cerner and Epic EHR. Those patients were selected with at least one visit to NMFF in 2003 aged 50 years or more. This was linked with the National Death Index for the years 2003-2008, to be able to model 5-year survival. Five year survival is typically recommended to be considered while making because decisions about preventive service use (eg, cancer screening, aggressive glucose control) [20].

A total of 980 predictive attributes for 7463 patients were derived. The attributes were derived from all a priori plausible predictors of mortality available within the EHR. These included: 11 sociodemographic attributes, 117 comorbidities, 20 vital signs, 120 laboratory results, 664 possible medications, and 48 healthcare utilization attributes. Please refer to [9] for details. Multiple rounds of feature selection techniques were used to find a subset of 23 features, to which sex was added for a final set of 24 input attributes. The target/outcome attribute denoted whether or not the patient survived at least 5 years. For the current analysis, only those patient instances were included that did not have missing or unknown values for any of the attributes, since we are interested in finding segments with precise attribute definitions. The resulting database consisted of 4262 instances, 24 input attributes, and the target attribute.

The overall data-driven process is depicted as a block diagram in Figure 1. The 25 attributes present in the dataset are as follows:

1) *Age*: Numeric age of the patient
2) *Sex*: Gender of the patient (male/female)
3) *Heart failure diagnosis*: Whether or not patient has been diagnosed with heart failure (yes/no)
4) *Atrial fibrillation diagnosis*: Whether or not patient has been diagnosed with atrial fibrillation (yes/no)
5) *Any kidney disease diagnosis*: Whether or not patient has been diagnosed with any kidney disease (yes/no)
6) *Any cardiovascular disease diagnosis*: Whether or not patient has been diagnosed with any cardiovascular disease (yes/no)
7) *Dementia diagnosis*: Whether or not patient has been diagnosed with dementia (yes/no)
8) *Metastatic cancer diagnosis*: Whether or not patient has been diagnosed with metastatic cancer (yes/no)
9) *Anemia diagnosis*: Whether or not patient has been diagnosed with anemia (yes/no)
10) *Chemotherapy diagnosis*: Whether or not patient has been given chemotherapy (yes/no)
11) *Comorbidity count*: Number of comorbidities
12) *Number of visits for any cancer diagnosis (0-1 year before)*: Number of doctor visits for any cancer diagnosis in the past year.
13) *Number of primary care physician visits (0-1 year before)*: Number of primary care physician visits in the

past year.

14) *Number of hospitalizations (0-1 year before)*: Number of hospitalizations in the past year.
15) *Number of hospitalizations (1-2 years before)*: Number of hospitalizations between one and two years prior to the current date.
16) *Mean diastolic blood pressure*: Numeric value of mean diastolic blood pressure (in mm Hg).
17) *Mean albumin*: Numeric value of mean albumin (in g/dL).
18) *Highest blood urea nitrogen*: Numeric value of highest blood urea nitrogen (in mg/dL).
19) *Mean creatinine*: Numeric value of mean creatinine (in mg/dL).
20) *Lowest sodium*: Numeric value of lowest sodium (in mEq/L).
21) *Highest bicarbonate*: Numeric value of highest bicarbonate (in mEq/L).
22) *Lowest calcium*: Numeric value of lowest calcium (in mg/dL).
23) *Digoxin prescription*: Whether or not the patient has been prescribed digoxin (yes/no).
24) *Loop diuretic prescription*: Whether or not the patient is on loop diuretic prescription (yes/no).
25) *Death*: Whether the patient died within 5 years of the current hospital visit (yes/no).

IV. Experiments and Results

Of the 4262 patients in the dataset, 653 died within five years and 3609 survived at least five years. The not survived fraction f, therefore, is 15.32%, and the goal of association rule mining is to find segments where this fraction f' is significantly higher or lower than f. Figure 2 shows the histograms of all the 24 attributes color-coded by the outcome attribute.

We performed two independent analyses to find segments in which the fraction of patients not survived after five years was higher and lower than the fraction across the entire dataset. Several combinations of algorithm parameters (maximum branching factor, minimum improvement in target value, and minimum segment size) were tried. Here we report the results with the same parameters as used in [8]: maximum branching factor = 3, minimum improvement in target value = 1%, and minimum segment size = 100.

It is well known that association rule analysis can lead to the discovery of a large number of redundant rules, which need to be post-processed. Here we used a two-stage semi-manual procedure to remove redundant rules:

1) **Stage I**: As described before, the way HotSpot algorithm works is to try to go deeper and deeper into the data to make the rule tree as long as it is able to improve the target value. A natural consequence of this is that the leaf nodes would have the best target value compared to all the nodes on its path. Therefore, we discard all the rules corresponding to the non-leaf nodes, and retain only the ones corresponding to the leaf nodes. This stage can be easily automated and does not require manual intervention.
2) **Stage II**: Even after Stage I, there can still remain quite a few redundant rules, the removal of which require domain expertise and manual supervision.

Lift of a rule is the relative improvement in the target (here the fraction of patients who did not survive at least five years) as compared to the average value of the target across the entire dataset. Therefore, lift for the two modes can be defined as follows:

$$L_{high} = f'_{high}/f$$

$$L_{low} = f/f'_{low}$$

where L_{high} and L_{low} are the lift values for the high and low modes respectively; f'_{high} and f'_{low} are the fraction of not survived patients in the segments resulting from hotspot analysis with high and low modes respectively, and f is the fraction of not survived patients across the entire dataset.

Table I presents the number of rules generated by the HotSpot algorithm, and the rules after each stage of redundant rule removal. Interestingly, the low mode (which corresponds to finding segments with low death rate, i.e. healthy patients) generated over a thousand rules, which were reduced to less than half by Stage I redundant rule removal. It was found during Stage II redundant rule removal that all these rules were essentially different combinations of attributes negating different comorbidities, which is quite obvious and thus redundant from the domain perspective. The rule with the highest lift value was selected as the representative rule for the applied parameter combination.

TABLE I
Number of association rules

Mode	*HotSpot algorithm*	*Redundant rule removal - I*	*Redundant rule removal - II*
High	40	18	6
Low	1150	482	1

TABLE II
Non-redundant association rules denoting segments where fraction of not survived patients is significantly higher than f=15.32%

Segment description	f'_{high} (%)	*Segment size*	*Lift*
#Hospitalizations0-1yr > 2, HighestBloodUreaNitrogen > 16, MeanAlbumin <= 3.9333	59.17	100	3.86
Age > 81, #Comorbidities > 2, HighestBloodUreaNitrogen > 10, MeanAlbumin <= 4.08, MeanDiastolicBloodPressure > 56.1667, #Comorbidities <= 10	59.17	100	3.86
Age > 81, #Comorbidities > 2, HighestBloodUreaNitrogen > 10, MeanDiastolicBloodPressure > 56.1667, #Comorbidities <= 10, LowestSodium > 125	59.17	100	3.86
#Hospitalizations0-1yr > 2, MeanDiastolicBloodPressure <= 89, MeanAlbumin <= 3.9333, Dementia=0	58.82	100	3.84
#Hospitalizations0-1yr > 2, Age > 51, MeanAlbumin <= 3.9333	58.14	100	3.79
MeanAlbumin <= 2.77, MeanDiastolicBloodPressure > 52, MeanAlbumin > 1.1714	56.5	100	3.69

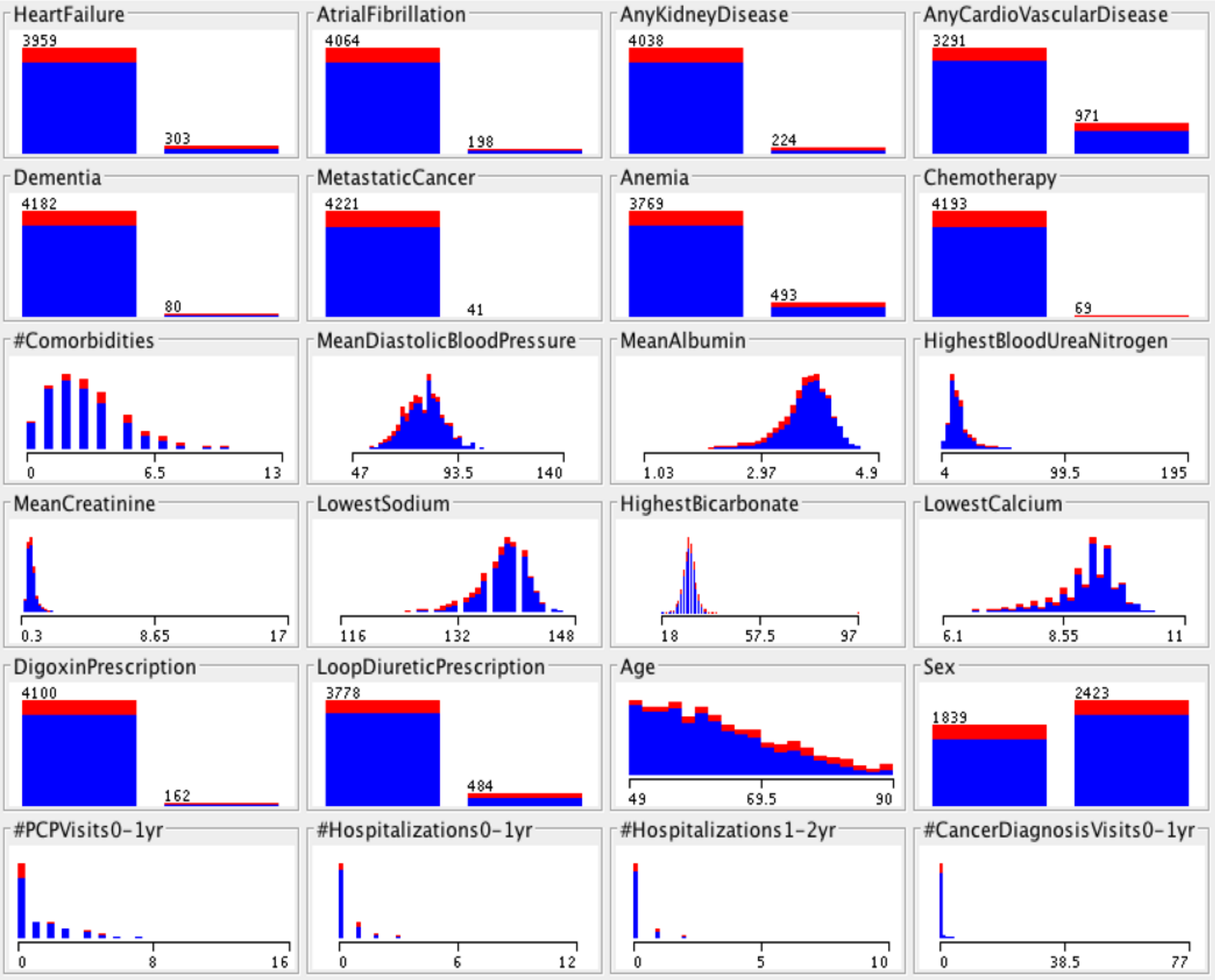

Fig. 2. Distribution (histograms) of all the 24 input patient attributes present in the dataset. The patients who did not survive at least five years are represented by red-colored fraction and those who did survive are represented by blue. For all binary (yes/no) attributes except sex, the first bar represents "No" and the second bar represents "Yes". For sex, the first bar represents "Male" and the second bar represents "Female".

TABLE III
NON-REDUNDANT ASSOCIATION RULES DENOTING SEGMENTS WHERE FRACTION OF NOT SURVIVED PATIENTS IS SIGNIFICANTLY LOWER THAN f=15.32%

Segment description	f'_{low} (%)	*Segment size*	*Lift*
AnyCardioVascularDisease=0, Anemia=0, HeartFailure=0, Sex=Female, Dementia=0, Chemotherapy=0, AtrialFibrillation=0, MetastaticCancer=0	6.31	100	2.43

Tables II and III present the non-redundant association rules/segments obtained with high and low mode respectively and offers some insight on the combined influence of multiple comorbidities on five year survival. Some of the factors associated with increased death rate were: high number of hospitalizations in the past year, high blood urea nitrogen levels, low albumin levels, advanced age, high diastolic blood pressure, more comorbidities, and high sodium. Among these discovered rules, we also see certain strange factors. For example, #Comorbidities<=10 appears in some rules. On looking closer at the data, we found that there were only 10 instances of patients having #Comorbidities>10, which might have influenced its frequent appearance. Another counter-intuitive risk factor appears in the fourth rule in Table II, which is not having dementia. Such results can result primarily due to artifacts of the specific data being analyzed. For the low mode, the representative association rule with the best value of lift suggests the following factors associated with a decreased death rate: no cardiovascular disease, no anemia, no heart failure, no dementia, no chemotherapy, no atrial fibrillation, and no metastatic cancer. Most of the rules obtained in both cases conform with existing biomedical knowledge and provide interesting insights into prognosis of older adults.

V. Conclusion and Future Work

In this paper, we performed association rule mining analysis on a EHR dataset of older adults to identify hotspots in the data, where the fraction of patients not survived after five years is significantly higher than and lower than the fraction across the entire dataset.

We believe that such analysis can be very useful to identify the factors affecting survival, create awareness, and aid doctors and patients to take appropriate proactive measures to avoiding the conditions which are known to reduce survival time, and encourage the conditions which are known to increase the survival time, whenever possible.

In the future, it would be good to investigate the effect of using different parameter combinations of HotSpot algorithm to find association rules. Identifying the best parameters for a given dataset is a challenging problem. Finally, similar data-driven analytics can be performed for other healthcare datasets, both disease-specific and general, and identify causative risk factors to enhance clinical decision support and informed patient consent.

Acknowledgment

This work is supported in part by the following grants: NSF awards CCF-1029166, IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330; AFOSR award FA9550-12-1-0458; NIST award 70NANB14H012.

References

[1] C. H. Foundation, A. G. S. P. in Improving, and C. Care for Elders with Diabetes, "Guidelines for improving the care of the older person with diabetes mellitus," *Journal of the American Geriatrics Society*, vol. 51, no. 5s, pp. 265–280, 2003.

[2] A. Wolf, R. C. Wender, R. B. Etzioni, I. M. Thompson, A. V. D'Amico, R. J. Volk, D. D. Brooks, C. Dash, I. Guessous, K. Andrews *et al.*, "American cancer society guideline for the early detection of prostate cancer: update 2010," *CA: a cancer journal for clinicians*, vol. 60, no. 2, pp. 70–98, 2010.

[3] D. A. Fisher, J. Galanko, T. K. Dudley, and N. J. Shaheen, "Impact of comorbidity on colorectal cancer screening in the veterans healthcare system," *Clinical Gastroenterology and Hepatology*, vol. 5, no. 8, pp. 991–996, 2007.

[4] T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[5] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on seer data," *Scientific Programming*, vol. 20, no. 1, pp. 29–42, 2012.

[6] K. Y. Bilimoria, Y. Liu, J. L. Paruch, L. Zhou, T. E. Kmiecik, C. Y. Ko, and M. E. Cohen, "Development and evaluation of the universal acs nsqip surgical risk calculator: A decision aid and informed consent tool for patients and surgeons," *Journal of the American College of Surgeons*, vol. 217, no. 5, pp. 833–842, 2013.

[7] A. Agrawal, J. Raman, M. J. Russo, and A. Choudhary, "Heart transplant outcome prediction using unos data," in *Proceedings of the KDD Workshop on Data Mining for Healthcare (DMH)*, 2013, pp. 1–6.

[8] A. Agrawal and A. Choudhary, "Identifying hotspots in lung cancer data using association rule mining," in *2nd IEEE ICDM Workshop on Biological Data Mining and its Applications in Healthcare (BioDM)*, 2011, pp. 995–1002.

[9] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," *Journal of the American Medical Informatics Association*, vol. 20, pp. e118–e124, 2013.

[10] A. Agrawal, R. Al-Bahrani, J. Raman, M. J. Russo, and A. Choudhary, "Lung transplant outcome prediction using unos data," in *IEEE BigData Workshop on Bioinformatics and Health Informatics (BHI)*, 2013, pp. 1–8.

[11] A. Agrawal, J. Mathias, D. Baker, and A. Choudhary, "Five year life expectancy calculator for older adults," 2016, available at http://info.eecs.northwestern.edu/FiveYearLifeExpectancyCalculator.

[12] A. J. Perkins, K. Kroenke *et al.*, "Common comorbidity scales were similar in their ability to predict health care costs and mortality," *Journal of clinical epidemiology*, vol. 57, no. 10, pp. 1040–1048, 2004.

[13] L. C. Walter and K. E. Covinsky, "Cancer screening in elderly patients: a framework for individualized decision making," *Jama*, vol. 285, no. 21, pp. 2750–2756, 2001.

[14] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD International Conference on Management of data*, ser. SIGMOD '93, 1993.

[15] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *International Conference on Very Large Data Bases (VLDB)*, 1994.

[16] M. J. Zaki, "Scalable algorithms for association mining," *IEEE Trans. on Knowledge and Data Engineering*, vol. 12, May 2000.

[17] J. Han, J. Pei, Y. Yin, and R. Mao, "Mining frequent patterns without candidate generation: A frequent-pattern tree approach," *Data Min. Knowl. Discov.*, vol. 8, pp. 53–87, January 2004.

[18] A. Agrawal and A. Choudhary, "Association rule mining based hotspot analysis on seer lung cancer data," *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)*, vol. 2, no. 2, pp. 34–54, 2011.

[19] M. Hall, E. Frank *et al.*, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[20] L. C. Walter, C. L. Lewis, and M. B. Barton, "Screening for colorectal, breast, and cervical cancer in the elderly: a review of the evidence," *The American journal of medicine*, vol. 118, no. 10, pp. 1078–1086, 2005.