Extracting Diagnosis Patterns in Electronic Medical Records using Association Rule Mining

Stephen M. Kang'ethe
School of Computing and Informatics, University of Nairobi
P. O. Box 30197 – 00100 Nairobi, Kenya

Peter W. Wagacha
School of Computing and Informatics, University of Nairobi
P. O. Box 30197 – 00100 Nairobi, Kenya

ABSTRACT
Data mining technologies have been used extensively in the commercial retail sectors to extract data from their "big data" warehouses. In healthcare, data mining has been used as well in various aspects which we explore. The voluminous amounts of data generated by medical systems form a good basis for discovery of interesting patterns that may aid decision making and saving of lives not to mention reduction of costs in research work and possibly reduced morbidity prevalence. It is from this that we set out to implement a concept using association rule mining technology to find out any possible diagnostic associations that may have arisen in patients' medical records spanning across multiple contacts of care. The dataset was obtained from Practice Fusion's open research data that contained over 98,000 patient clinic visits from all American states.

Using an implementation of the classical apriori algorithm, we were able to mine for patterns arising from medical diagnosis data. The diagnosis data was based on ICD-9 coding and this helped limit the set of possible diagnostic groups for the analysis. We then subjected the results to domain expert opinion. The panel of experts validated some of the most common associations that had a minimum confidence level of between 56-76% with a concurrence rate of 90% whereas others elicited debate amongst the medical practitioners. The results of our research showed that association rule mining can not only be used to confirm what is already known from health data in form of comorbidity patterns, but also generate some very interesting disease diagnosis associations that can provide a good starting point and room for further exploration through studies by medical researchers to explain the patterns that are seemingly unknown or peculiar in the concerned populations.

Keywords
Medical Diagnosis Patterns, Electronic Medical Records, Health Informatics, Association Rule Mining, Apriori.

1. INTRODUCTION
The health sector worldwide has been involved in automation of medical records worldwide. Medical practitioners have had to learn new ways of capturing their findings and treatment plans on their patients after having had years of the same on paper. Different health institutions who have adopted Electronic Medical Record systems (EMR) have done it in their own ways before owing to the lack of standardization of such implementations in the years past. In recent times however, world governing institutions like WHO and ISO have embraced the advent of Health Information Systems (HIS) and spearheaded the development of standards that were hitherto unavailable to implementers of health systems. These standards make it easy not only to capture and share data across multiple and seemingly disparate implementations, but to also query, analyze and extract useful statistics from data entered in the same systems.

The need to have EMR systems has been influenced by some factors including complex medical data, the influx of patients and the need to have proper recording of health data. When EMR systems are well developed, they are likely to positively impact the quality and reliability of health data, as well as standardized reporting[1].
The standards that will be of particular interest in our research are the International Classification of Diseases (ICD) standards, (both ICD-9 and ICD-10) and HL7 health information interchange standards.

In their work, *Fast algorithms for mining association rules in large databases,* [2], the authors presented an algorithm, known as Apriori, for discovering association rules within large, primarily transactional, sales databases. This algorithm was a development of previously known algorithms for itemset mining and association rules discovery. We have a brief look at how this algorithm works and its known uses in the commercial, particularly retail sales databases, for which the authors admit the algorithm was originally conceived. We will also explore the benefits accrued by using this algorithm over other known algorithms for association rules mining.

The availability of standardized medical data creates a large pool of data with a lot of hidden and potentially useful information. Using association rule mining and the apriori algorithm in particular, we seek to unravel the hidden diagnosis patterns that could be present within the data availed by these systems. We also intend to generate and discover strong rules (relationships) that indicate multimorbidity trends from the EMR data with varying measures of interestingness.

2. LITERATURE REVIEW
2.1 International Statistical Classification of Diseases (ICD)
International Statistical Classification of Diseases is the standard diagnostic tool for epidemiology, health management and clinical purposes. [3]. It contains standard diagnostic codes that attempt to cover all known morbidity and mortality causes statistics.

ICD-10 is the current standard and is a replacement of the widely used ICD-9. The latest version is the 2010 version. ICD 9 has also been in use for a while and is in the process of being replaced by ICD 10. WHO also state that the 11th revision of the classification (ICD-11) is in place and is set to go on until the year 2017 [4].

ICD-9, codes are three to five digits. The first digit is either numeric or alpha (the letters E or V only) and all other digits are numeric.

In ICD-10-CM, however, codes can be up to seven digits. The first digit is always alpha (it can be any letter except U), the second digit is always numeric, and the remaining five digits can be any combination [5].

For our research, any EMR that uses either of the two codes will suffice provided they are used consistently across the implementation.

2.2 Association Rule Mining and the Apriori Algorithm

Association rule mining has been used extensively in the commercial industry particularly in the retail sector. It has mainly been used to do market basket analysis where the focus is on analyzing the contents of the customer's "basket". As [6] explain, Market basket analysis provides insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to promotion. Association rules identify strong relations that exist in databases using several measures of interestingness (usually based on minimum support and minimum confidence) [7].

The patterns discovered may have different uses in nature and they may be categorized as actionable rules (contain high-quality, actionable information), trivial rules (already known by anyone at all familiar with the business) or inexplicable rules (these seem to have no explanation and do not suggest a course of action) [6].

When large databases are involved, an efficient algorithm to find frequently items that exist together (frequent itemsets) and find any patterns amongst these is needed. [2] present an algorithm (Apriori) that aims at discovering association rules between items in a large database of sales transactions. The algorithm is simple in concept and is split into two main sub problems:

1. Find all sets of items (itemsets) that have transaction support above minimum support. The support for an itemset is the number of transactions that contain the itemset. Itemsets with minimum support are called large itemsets, and all others small itemsets.
2. Use the large itemsets to generate the desired rules [2].

The minimum support and confidence are given as follows [8]:

Support:

$$supp(X) = \frac{no.of\ transactions\ which\ contain\ the\ itemset\ X}{total\ no.of\ transactions}$$

Confidence:

$$conf(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

2.3 Association Rule Mining in Healthcare

In health informatics, a lot of work has gone in the use of data mining to previously commercial-only applications. Key amongst the uses has been in matching patient diagnosis with symptoms which intertwines a lot with the use of knowledge based systems. It is difficult to induce reliable diagnostic rules from amongst a set of possibly infinite permutations of symptoms since the resulting hypotheses may have unsatisfactory prediction accuracy [9].

However, other researchers have come up with further refinements by using association rules to improve the prediction level claimed at 90% by [10] by combining it with supervised learning methods. The researchers applied their work to cancer but they claim that this can be extended to other disease diagnosis.

Association analysis as it is also called has been used to give probabilistic statements such as "If patients undergo treatment A, there is a 0.35 probability that they will exhibit symptom Z" [11]. These can be useful when establishing relationships that affect effectiveness of particular patient treatment plans.

2.4 Related and Specific Applications in EMR Implementations

In this research's specific field, some work has been done to take advantage of association rule mining in general and the Apriori algorithm in particular. Most of it centers on mining patterns in relation to a specific disease or diagnostic factor.

[12] used association rule mining to discover associations from data obtained from the National Health Insurance Database of Taiwan. Their work was intended not only to discover the comorbidity patterns of Attention Deficiency Hyperactive Disorder (ADHD) but to also examine the application of association rule mining in clinical databases.

The database used ICD-9 diagnosis coding and drew a sample of about 18,000 patients aged 18 and below with a diagnosis of ADHD. The researchers then made comparisons using Apriori algorithm to check the strength of associations amongst comorbidity rates and relative risk (RR) ratios of both groups of each diagnosis which were compared to one another. The results were published along with the resultant levels of interestingness.

More work was also done [13] to analyze comorbidity in patients with type 2 diabetes mellitus (T2DM). The data was obtained from a medical center in Korea with an EMR that uses ICD-10 coding for the clinical diagnosis. The researchers developed a tool that uses Apriori algorithm to generate the strongest rules (diagnosis) that are associated with the T2DM. They then published the results of their findings with the resultant support and confidence levels.

Another prototype namely Clinical State Correlation Prediction (CSCP) was developed in order to predict the correlation(s) amongst the primary disease (the disease for which the patient visits the doctor) and secondary disease/s (which is/are other associated disease/s carried by the same patient having the primary disease [14]. The system developed uses the Apriori algorithm as well and checks the correlation between the primary disease and other secondary diseases. The CSCP is built on top of the transaction based health system which they base on and refer to as the OLTP. The diagnoses are not based on any diagnosis group like ICD.

They also use data from this health OLTP, and pass the algorithm over data selected for different age groups and sex. The results of the top two-item itemsets are then analyzed for any meaningful information.

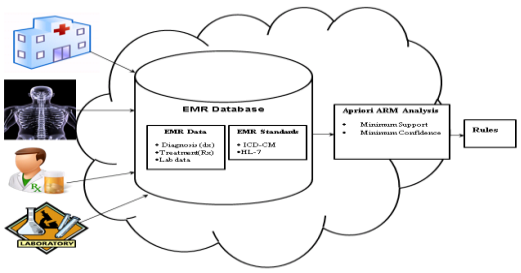

Figure 2.3 Illustration of research context

3. METHODOLOGY

3.1 Data Collection

We obtained the data necessary for this research. The data met the standards defined earlier herein in order for it to be usable. For this research, we obtained our data from Practice Fusion research data that is one of the leading EMR implementers in the United States of America. The dataset availed contains over 10,000 de identified patient records [15] that contain over 98,000 individual contact points from over 150,000 medical practitioners across the country, in ICD-9 diagnosis codes. The larger dataset from which they extracted this contains over 30 million records and has already been used to track the spread of H1N1 to help physicians obtain vaccines, and Practice Fusion's Research Division is partnering with leading academic institutions and public health agencies to pursue ambitious new health studies [16].

3.2 Data Preparation and Processing

The data we obtained was then prepared and processed by the following steps:

i. Extracting the major diagnostic groups for each ICD-9 diagnosis for every patient record. This is due to the fact that every sub-diagnosis group after the period (.) still represents the anatomic site or severity of the specific disease category. For example the ICD-9 code 473 represents "Chronic Sinusitis". Others under this would be 473.1 – "Chronic Sinusitis- Maxillary Antritis (chronic)", 473.1 "Chronic Sinusitis Frontal", 473.2 "Chronic Sinusitis – Ethmoidal" all the way to 473.9 - "Unspecified Chronic Sinusitis".

ii. We are therefore still able to obtain interesting associations on a higher level without losing meaning from where we can dig further.

iii. We then filtered out the codes that begin with E & V since these represent External causes of injury (e.g. accidents) and Supplementary classification of factors influencing health status and contact with health services respectively [17]

iv. After this, we transformed the data into itemsets where each patient has all the associated diagnosis combined into a single comma separated record. We have done this by use of an SQL script (Appendix B). A sample record would look like: (PID0001 | 420,421,618).

v. We are able to filter from the beginning if we want to observe chronic diagnosis only, acute only or all combined to give us the potential patterns we want.

3.3 Prototype Development

We developed a prototype that implements the Apriori algorithm. The prototype borrows some implementations that have been used in market basket analysis. The prototype is capable of taking the data and finding associations based on the user defined values for the minimum support and confidence level. It was implemented in C# and relies on a backend database of MS-SQL server.

This follows the classical Apriori algorithm with the steps as explained above in section 2.3

3.4 Prototype Testing and Implementation

The prototype was developed and tested using the de-identified patient records. We were able to run the data and observe the association rules generated. Using varying minimum support and confidence values generated a number of rules. The top rules were those with the strongest confidence level above a support threshold.

Since there is no globally accepted minimum support (as this is a custom user generated variable that depends on what they want to achieve, and how far deep they want to dig into the associations), we varied these values to observe the results and recorded each observation. Just as in other works using the values of support and confidence in this mining for strong associations like in [13] and that of [12], we vary the same measures and indicate the values of support and confidence for each rule.

3.5 Analysis, Validation and Presentation of Results

Based on the rules observed, we compare this with the demographic data and select the demographic distribution of the top associations. These are mainly age groups and gender.

We then use measures of central tendency (as appropriate for the nominal and ordinal variables) and classify the data into the different categories that they fall in.

We also used a panel of experts drawn from the medical field who gave their opinion over the results.

We used the Likert scale to gather expert opinion and listened to their overall advice while noting explanations to some of their responses.

4. RESULTS AND DISCUSSION

After execution of the runs and aggregation of the same, we were able to come up with a number of rules based on a support factor of 10% and a minimum confidence of 50% for the first instance in order to see if we would obtain less than 10 rules, from which we expected to see rules already known by anyone in the medical field, also known as trivial rules [6].

4.1 Potentially Trivial Associations

Two sample results of the ten runs are shown in Fig 4.1. In each are the top associations and the associated percentage confidence levels.

Run 1		
	250,401 --> 272	80.25
	250,272 --> 401	76.83
	401 --> 272	61.73
	250 --> 272	59.85
	250 --> 401	57.30
	272 --> 401	56.71
	530 --> 272	54.84

Run 2		
17	250,272 --> 401	71.02
19	250,401 --> 272	71.02
4	530 --> 272	61.40
32	272 --> 401	59.79
28	250 --> 272	57.33
30	250 --> 401	57.33
33	401 --> 272	51.90

Figure 4.1 Sample Runs

These aggregate for each was obtained and are listed in Table 1:

Table 1 Top Associations (note the bottom two as well)

ICD9 Disease Description	Average confidence
Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism	74.22
Diabetes Mellitus, Disorders of Lipoid metabolism-> Essential Hypertension	75.66
Essential Hypertension->Disorders of Lipoid metabolism	57.95
Diabetes Mellitus->Disorders of Lipoid metabolism	56.34
Diabetes Mellitus->Essential Hypertension	57.45
Disorders of Lipoid metabolism->Essential Hypertension	58.23
Esophageal disease->Disorders of Lipoid metabolism	34.68
Esophageal disease->Essential Hypertension	6.28

There were two that met the minimum support but not the confidence levels (the bottom two in Fig 4.1) and in order to test if they were inexplicable rules, we allowed them to be in the survey and they were proved to be so as shown in the survey interpretation later in this chapter.

A graphical representation is also shown in Fig 4.3.

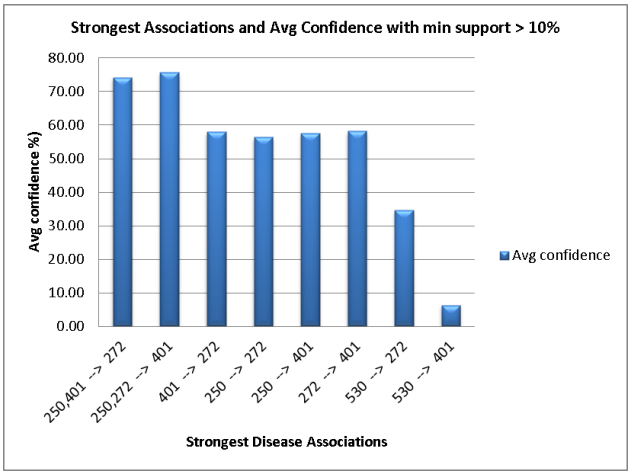

Figure 4.3 Strongest Associations with support above 10%

4.2 Potentially Actionable Rules

According to [6] actionable rules as those that contain high-quality, actionable information. We intended to obtain these and us them in our comparisons.

Since the support of 10% only generated eight associations with a confidence level above 50%, reducing the minimum support to 5% but still maintaining a confidence level above 50% in order to avoid compromising on the quality of associations gave a number of interesting rules. They are listed below:

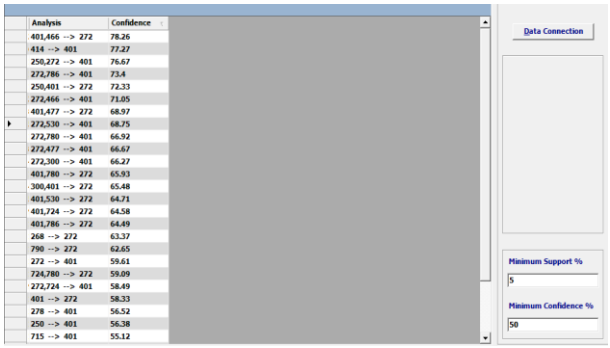

Figure 4.4 Other Association Codes

These translate to ICD descriptions as shown in Table 2.

Table 2. ICD 9 Descriptions

Essential hypertension,Allergic rhinitis --> Disorders of lipoid metabolism
Disorders of lipoid metabolism,Diseases of esophagus (excludes esophageal varices) --> Essential hypertension
Disorders of lipoid metabolism,Allergic rhinitis --> Essential hypertension
Disorders of lipoid metabolism,Anxiety, dissociative and somatoform disorders --> Essential hypertension
Anxiety, dissociative and somatoform disorders,Essential hypertension --> Disorders of lipoid metabolism
Essential hypertension,Diseases of esophagus (excludes esophageal varices) --> Disorders of lipoid metabolism
Essential hypertension,Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metabolism
Essential hypertension,Symptoms involving respiratory system and other chest symptoms --> Disorders of lipoid metabolism
Vitamin D deficiency --> Disorders of lipoid metabolism
Nonspecific abnormal findings on examination of blood (see excludes 2) --> Disorders of lipoid metabolism
Disorders of lipoid metabolism --> Essential hypertension
Disorders of lipoid metabolism,Other and unspecified disorders of back (see excludes 1) --> Essential hypertension
Essential hypertension --> Disorders of lipoid metabolism
Overweight, obesity and other hyperalimentation --> Essential hypertension
Diabetes mellitus --> Essential hypertension
Osteoarthrosis and allied disorders --> Essential hypertension
Edema--> Essential hypertension
Excludes: 1. Excludes collapsed vertebra (code to cause, e.g., osteoporosis) conditions due to intervertebral disc disorders, spondylosis 2. Excludes abnormality of: platelets , thrombocytes, white blood cells

The diagnosis distribution amongst the group of "Disorders of lipoid metabolism" coded as 272 is shown below in Figure 4.6. It is important to note that "Mixed Hyperlipidemia" and "Other Unspecified hyperlipidemia" formed the bulk (76.27%) of the "Disorders of lipoid metabolism"

Code 272 Diagnosis Breakdown	Diagnosis (%)
Mixed hyperlipidemia	51.90%
Other and unspecified hyperlipidemia	24.37%

Pure hypercholesterolemia	18.73%
Pure hyperglyceridemia	4.41%
Unspecified disorder of lipoid metabolism	0.19%
Lipoprotein deficiencies	0.19%
Lipodystrophy	0.06%
Other disorders of lipoid metabolism	0.06%
Hyperchylomicronemia	0.04%
Disorders of lipoid metabolism	0.04%
Grand Total	**100.00%**

Figure 4.6 Distribution for "Disorders of lipoid metabolism"

4.3 Demographic Comparison

We compared the diagnoses with the demographic prevalence where they were comorbid and these were shown to be consistent with expectations. It is worth noting that the average age of the population was age 52.

4.4 Validation of Results

In this stage we are able to compare the results of our prototype and the opinion of experts regarding whether the associations obtained here are known to them or not, and if not whether they agree that they could be linked (probably indirectly) and by how much (strongly or otherwise). This is done through a questionnaire survey (see appendix I).

Each of the questions can be scored as follows:

Question	LK	HC	JM	DO	JA	BA	Median
1	5	4	5	4	5	3	4.5
2	5	3	5	4	5	2	4.5
3	5	4	5	4	5	4	4.5
4	5	4	5	4	5	4	4.5
5	5	4	5	5	5	3	5
6	5	4	5	4	5	4	4.5
7	3	1	1	2	1	2	1.5
8	3	1	1	2	1	2	1.5
9	3	1	3	-	-	2	2.5
10	5	4	5	1	5	4	4.5
11	5	3	5	5	5	3	5
12	3	4	4	1	4	4	4
13	5	4	5	4	5	3	4.5
14	3	3	3	1	-	2	3
15	3	1	3	2	-	3	3
16	3	2	3	1	-	2	2
17	3	2	3	2	-	2	2
18	4	2	3	2	4	3	3
19	4	2	3	1	4	3	3
20	3	2	3	1	-	3	3
21	2	2	3	1	4	3	2.5
22	2	2	4	1	4	4	3
23	1	1	4	1	2	5	1.5
24	1	2	2	1	-	2	2
25	5	3	5	4	5	4	4.5
26	3	3	3	2	4	4	3
27	5	4	5	4	5	2	4.5
28	4	5	5	4	5	5	5
29	5	3	5	4	5	2	4.5
30	3	3	-	3	2	2	3
31	4	2	4	2	5	2	3

Figure 4 .7 Likert Scale Scores

Each of the questions has a score associated that is calculated from the median of responses from all experts (since the scale consists of ordinal values).

4.5 Discussion of Results

After running through the dataset, we were able to generate several associations that differed based on what we set as the minimum support and confidence level. We did not find a universally applicable or acceptable threshold for minimum support and confidence, as this seems to be applicable in different ways to different domains, depending on what patterns the end user intends to accept or reject. As earlier discussed, as in other works using the values of support and confidence in this mining for strong associations like in [13] and that of [12], we vary the same measures and indicate the values of support and confidence for each rule.

It is possible to obtain a very large number of rules since these increase as the values of minimum support and minimum confidence are decreased and are approaching zero. The outliers in the data in this case will be the rules that may not necessarily meet the selected user-specific threshold for minimum support and confidence. It is therefore up to the user to decide what the most acceptable values for minimum support and confidence

are, and what criteria to use to discard or accept the generated associations.

We observed that some rules were generated which happened to be consistent with common knowledge amongst the members of the medical fraternity, for example the link between Essential Hypertension and Disorders of the Lipoid Metabolism, or Diabetes Mellitus (as shown in the first six rules of Table 1). The panelists accepted this with a concurrence of 4.5/5 translating to a 90% nod. These known associations also had all high measures of confidence (between 56.34-75.66% from our system) as shown in Table 1. Some diagnosis were also consistent with some of previous specific research like that of [13] that indicate the strongest link between Type 2 Diabetes mellitus and Essential Hypertension with a confidence of 34.86%. This is captured as rule 5 in our results with a confidence of 57.45%.

There are other rules which most of the panelists chose to neither agree nor disagree. They attributed this to the fact that some of the associations may be incidental to some specific patients and it may be observed in a few cases but not necessarily a majority of the cases. The presence of one qualifying diagnosis from amongst the set on the left being linked to that on the right also caused a mixed reaction in most of the practitioners, an example being that of:

Essential hypertension, Allergic rhinitis --> Disorders of lipoid metabolism.

In such a case, the panelists argued that it is the link of Essential hypertension to Disorders of lipoid metabolism and not the Allergic rhinitis that would trigger the association.

We also observe that some associations were out rightly rejected by the same panel of experts as expected (e.g. the association between *Esophageal disease->Disorders of Lipoid metabolism*). These were listed despite having very low confidence levels (as low as 6.28%) in order to compare the responses with the others that had higher values for confidence.

Some experts indicated that some of the associations could be comorbid but not necessarily linked, that is without a cause-effect relation and that some conditions coexist but are not very frequent. This, they said, also determined how they scaled the associations.

However, particularly with the second run where the minimum support was lowered to 5%, but the confidence level maintained, it is interesting to note that there was mixed opinion, or outright rejection of some interesting associations. An association that seemed interesting to the researcher that got a "Strongly Disagree" despite having a high confidence level (63.37%) is that of (*Vitamin D deficiency -> Disorders of lipoid metabolism*). As shown in Figure 4.6, more than 75% of this was hyperlipidemia (mixed and unspecified).

This result presents an interesting dimension as the experts indicated little or no known association between the two. In ensuing discussions over the results, one panelist noted that this association could have different indirect associations that could potentially explain it. Hyperlipidemia (explained to the researcher as a condition resulting from elevated levels of lipids in the blood) could have been as an indirect result of vitamin D deficiency since people who lack in vitamin D may be those that tend to stay indoors most of the time (one of the major sources of Vitamin D being skin exposure to sunlight). This could arguably be in line with the average age of 52 for the patients in the sampled dataset. The hyperlipidemia therefore in his reasoning, comes not as a result of the Vitamin D deficiency, but as a result of the lifestyle likely to be found amongst patients with Vitamin D deficiency. That relation alone as a real possibility could be subject to investigation outside the scope of this research.

Another panelist was also keen to indicate that the associations that we seek to investigate can only be investigated as comorbidity patterns and causal relations may not necessarily be possible to state comprehensively at this level. This is what the research emphasizes as the output of its findings.

Findings to mining medical datasets requires a lot of domain expertise to interpret the rules as was reiterated by [18]. Most of them will be known but others may be less known while those that seem unusual may be discarded at a first look. However, output to this research may prove to be of utmost importance to curious specialists since some of the rules generated, however few, could be used as a starting point for future research by the domain experts. Of great interest would be to attempt to establish the reasons for comorbidity amongst our associations that seem unusual or unknown, a good example being the Vitamin D-> Hyperlipidemia association. These reasons could be causal links or outright co-existence due to the condition of the patient. As one panelist explained, a patient diagnosed separately with allergic rhinitis, bronchitis and eczema (dermatitis) will have allergic tendencies that make such conditions, whereas unrelated, to be present in the same patient over time. When this happens frequently in the sampled population, some associations like these will certainly emerge from our system, and only further investigation by domain experts will show that.

Comparisons with demographics showed some expected patterns like some disease prevalence being higher in older patients e.g. the combination of Hypertension and Diabetes Mellitus being found in patients with an average age of 63.5, presenting a distance of 11.5 years above our average age. This is true for the most common diagnosis associations from our results. Further demographic analysis could be done on individual sets of associations as far as one would desire to find more relevant demographic patterns and compare them with the expected patterns.

5. CONCLUSION

Using this prototype, we are now able to mine data from EMR systems that implement any standardized diagnosis coding guideline. In our case, it is the WHO recommended standard of ICD-CM coding. Multiple systems can exchange their data and we are therefore able to take advantage of big data and generate patterns from it based on user defined measures of interestingness on what suits one as the minimum support and confidence.

It is also key to note that the data used for mining the associations was primarily intended for other clinical purposes. In this research, we were able to take advantage and build our system to find interesting patterns that could arise from this kind of well-organized big data. This goes to demonstrate the power of having standardized clinical data across multiple implementations of electronic medical records systems.

We were able to see that although the medical practitioners agreed on some already known associations, it would not be prudent to expect them to agree on all previously unknown associations. This research would therefore prove to be key as input to another research on causation, and would be a good starting point for any medical researcher seeking to look for multi-morbidity trends amongst patients in any given patient population.

We are particularly encouraged by previous studies that seemed to suggest that Vitamin D deficiency is associated with Hypertension but the causal relationship is not known [19].

This is the same way in which there could be a (perhaps less prevalent but nonetheless unknown and important) relationship between Vitamin D deficiency and disorders of lipoid metabolism mostly hyperlipidemia (mixed and unspecified type).

This would ideally then be used as input to another study that seeks to dwell on the specific association and finding if there is any causal association.

The demographic prevalence of our associations showed no much difference with the expected outcome as discussed in the previous chapter.

6. REFERENCES

[1] "Standards and Guidelines for Electronic Medical Record Systems in Kenya," Ministries of Health, Government of Kenya, Health Information Policy, 2009.

[2] R. Agrawal and R. Srikant, "Fast Algorithms For Mining Association Rules In Datamining," in *20th International Conference on Very Large Data Bases*, Santiago, Chile, 1994, pp. 487–499.

[3] "WHO | International Classification of Diseases (ICD)," *WHO*. [Online]. Available: http://www.who.int/classifications/icd/en/. [Accessed: 03-Feb-2014].

[4] "WHO | World Health Organization," *WHO*, 2014. [Online]. Available: http://www.who.int/classifications/icd/revision/en/. [Accessed: 05-Feb-2014].

[5] "ICD-10 Conversion and Mapping - AAPC," 2014. [Online]. Available: http://www.aapc.com/icd-10/conversion-mapping.aspx. [Accessed: 06-Feb-2014].

[6] M. J. A. Berry and G. Linoff, *Data mining techniques for marketing, sales, and customer relationship management*, 2nd ed. Indianapolis: Wiley, 2004.

[7] C. J. Matheus, P. K. Chan, and G. Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases," *IEEE Trans Knowl Data Eng*, vol. 5, no. 6, pp. 903–913, Dec. 1993.

[8] B. Bhargavi, B. Venkanna, and V. H. Prasad, "Mining Frequent Items Using Directed Graphs," *Int. J. Sci. Res. Comput. Sci.*, vol. 1, no. 2, pp. 21–24, Sep. 2013.

[9] A. Rajak and M. K. Gupta, "Association rule mining-applications in various areas," in *Proceedings of International Conference on Data Management, Ghaziabad, India*, 2008, pp. 3–7.

[10] G. Serban, C. Istvan-Gergely, and C. Alina, "A Programming Interface For Medical diagnosis Prediction," *Stud. Univ. Babes - Bolyai Inform.*, vol. LI, pp. 21–30, 2006.

[11] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J. Healthc. Inf. Manag.*, vol. 19, no. 2, p. 65, 2011.

[12] Y.-M. Tai and H.-W. Chiu, "Comorbidity study of ADHD: applying association rule mining (ARM) to National Health Insurance Database of Taiwan," *Int. J. Med. Inf.*, vol. 78, no. 12, pp. e75–83, Dec. 2009.

[13] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, "Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining," *Korean J. Intern. Med.*, vol. 27, no. 2, pp. 197–202, Jun. 2012.

[14] M. A. Rashid, M. T. Hoque, and A. Sattar, "Association Rules Mining Based Clinical Observations," *ArXiv Prepr. ArXiv14012571*, 2014.

[15] "Analyze This! | Research Division," *Research- Practice Fusion*, 06-Jun-2012. [Online]. Available: http://www.practicefusion.com/research/analyze-this/. [Accessed: 06-Feb-2014].

[16] "Big Data Gets Put to Work for Public Health," *Practice Fusion*, 15-Mar-2012. [Online]. Available: http://www.practicefusion.com/pages/pr/big-data-public-health.html. [Accessed: 06-Feb-2014].

[17] CMS, "ICD-9 Code Lookup," *Centers for Medicare & Medicaid Services*, 2014. [Online]. Available: http://www.cms.gov/medicare-coverage-database/staticpages/icd-9-code-lookup.aspx. [Accessed: 06-Aug-2014].

[18] J. F. Roddick, P. Fule, and W. J. Graco, "Exploratory medical knowledge discovery: Experiences and issues," *ACM SIGKDD Explor. Newsl.*, vol. 5, no. 1, pp. 94–99, 2003.

[19] K. S. Vimaleswaran, A. Cavadino, D. J. Berry, R. Jorde, A. K. Dieffenbach, C. Lu, A. C. Alves, H. J. L. Heerspink, E. Tikkanen, J. Eriksson, A. Wong, M. Mangino, K. A. Jablonski, I. M. Nolte, D. K. Houston, T. S. Ahluwalia, P. J. van der Most, D. Pasko, L. Zgaga, E. Thiering, V. Vitart, R. M. Fraser, J. E. Huffman, R. A. de Boer, B. Schöttker, K.-U. Saum, M. I. McCarthy, J. Dupuis, K.-H. Herzig, S. Sebert, A. Pouta, J. Laitinen, M. E. Kleber, G. Navis, M. Lorentzon, K. Jameson, N. Arden, J. A. Cooper, J. Acharya, R. Hardy, O. Raitakari, S. Ripatti, L. K. Billings, J. Lahti, C. Osmond, B. W. Penninx, L. Rejnmark, K. K. Lohman, L. Paternoster, R. P. Stolk, D. G. Hernandez, L. Byberg, E. Hagström, H. Melhus, E. Ingelsson, D. Mellström, Ö. Ljunggren, I. Tzoulaki, S. McLachlan, E. Theodoratou, C. M. T. Tiesler, A. Jula, P. Navarro, A. F. Wright, O. Polasek, J. F. Wilson, I. Rudan, V. Salomaa, J. Heinrich, H. Campbell, J. F. Price, M. Karlsson, L. Lind, K. Michaëlsson, S. Bandinelli, T. M. Frayling, C. A. Hartman, T. I. A. Sørensen, S. B. Kritchevsky, B. L. Langdahl, J. G. Eriksson, J. C. Florez, T. D. Spector, T. Lehtimäki, D. Kuh, S. E. Humphries, C. Cooper, C. Ohlsson, W. März, M. H. de Borst, M. Kumari, M. Kivimaki, T. J. Wang, C. Power, H. Brenner, G. Grimnes, P. van der Harst, H. Snieder, A. D. Hingorani, S. Pilz, J. C. Whittaker, M.-R. Järvelin, and E. Hyppönen, "Association of vitamin D status with arterial blood pressure and hypertension risk: a mendelian randomisation study," *Lancet Diabetes Endocrinol.*, Jun. 2014.

APPENDIX

Appendix I: Questionnaire for Survey

In your opinion, how much do you feel the following disease diagnosis (ICD9) are associated with each other (the left, indicating a likely presence of that on the right)?

Main Associations	Strongly Agree<----------->Strongly Disagree				
➢ Diabetes Mellitus, Essential Hypertension-> Disorders of Lipoid metabolism	(1)	2	3	4	5
➢ Diabetes Mellitus, Disorders of Lipoid metabolism-> Essential Hypertension	(1)	2	3	4	5
➢ Essential Hypertension-> Disorders of Lipoid metabolism	(1)	2	3	4	5
➢ Diabetes Mellitus->Disorders of Lipoid metabolism	(1)	2	3	4	5
➢ Diabetes Mellitus-> Essential Hypertension	(1)	2	3	4	5
➢ Disorders of Lipoid metabolism-> Essential Hypertension	(1)	2	3	4	5
➢ Esophageal disease-> Disorders of Lipoid metabolism	1	2	3	4	(5)
➢ Esophageal disease-> Essential Hypertension	1	2	3	4	(5)
Other Associations	1	2	(3)	4	5
➢ Essential hypertension, Acute bronchitis and bronchiolitis --> Disorders of lipoid metabolism					
➢ Other forms of chronic ischemic heart disease --> Essential hypertension	(1)	2	3	4	5
➢ Diabetes mellitus, Disorders of lipoid metabolism --> Essential hypertension	(1)	2	3	4	5
➢ Disorders of lipoid metabolism, Symptoms involving respiratory system and other chest symptoms --> Essential hypertension	1	(2)	(3)	4	5
➢ Diabetes mellitus, Essential hypertension --> Disorders of lipoid metabolism	(1)	2	3	4	5
➢ Disorders of lipoid metabolism, Acute bronchitis and bronchiolitis --> Essential hypertension	1	2	(3)	4	5
➢ Essential hypertension, Allergic rhinitis -->Disorders of lipoid metabolism	1	2	(3)	4	5
➢ Disorders of lipoid metabolism, Diseases of esophagus (excludes esophageal varices) --> Essential hypertension	1	2	(3)	4	5
➢ Disorders of lipoid metabolism, Allergic rhinitis --> Essential hypertension	1	2	(3)	4	5
➢ Disorders of lipoid metabolism, Anxiety, dissociative and somatoform disorders --> Essential hypertension	1	2	(3)	4	5

In your opinion, how much do you feel the following disease diagnosis (ICD9) are associated with each other (the left, indicating a likely presence of that on the right)?

➢ Anxiety, dissociative and somatoform disorders, Essential hypertension --> Disorders of lipoid metabolism	1	2	(3)	4	5
➢ Essential hypertension, Diseases of esophagus (excludes esophageal varices) --> Disorders of lipoid metabolism	1	2	(3)	4	5
➢ Essential hypertension, Other and unspecified disorders of back (see excludes) --> Disorders of lipoid metabolism	1	2	(3)	4	5
➢ Essential hypertension, Symptoms involving respiratory system and other chest symptoms --> Disorders of lipoid metabolism	1	(2)	(3)	4	5
➢ Vitamin D deficiency --> Disorders of lipoid metabolism	1	(2)	3	4	5
➢ Nonspecific abnormal findings on examination of blood (see excludes 2) --> Disorders of lipoid metabolism	1	2	3	(4)	5
➢ Disorders of lipoid metabolism --> Essential hypertension	(1)	2	3	4	5
➢ Disorders of lipoid metabolism, Other and unspecified disorders of back (see excludes 1) --> Essential hypertension	1	2	(3)	4	5
➢ Essential hypertension --> Disorders of lipoid metabolism	(1)	2	3	4	5
➢ Overweight, obesity and other hyperalimentation --> Essential hypertension	(1)	2	3	4	5
➢ Diabetes mellitus --> Essential hypertension	(1)	2	3	4	5
➢ Osteoarthrosis and allied disorders --> Essential hypertension	1	2	3	4	5
➢ Edema--> Essential hypertension	1	(2)	3	4	5

Excludes:

1. Excludes: collapsed vertebra (code to cause, e.g., osteoporosis) conditions due to intervertebral disc disorders, spondylosis
2. Excludes: abnormality of: platelets, thrombocytes, white blood cells

Doctor's Name: DR. JAMES MOH Signature: [signature] Date 05/08/2014

Comments (if any):
- I have indicated a scale of '3' for those that have conditions that may be incidental in a patient. For instance, a person may have diabetes and as a consequence develop hypertension, but just by chance has allergic rhinitis; or a has Esophageal disease; or a back problem.
- Those indicated '2' may have a some association but not as strong.