Results on mining NHANES data: A case study in evidence-based medicine

246532||

http://dx.doi.org.ezproxy.lib.uwf.edu/10.1016/j.compbiomed.2013.02.018

Abstract

The National Health and Nutrition Examination Survey (NHANES), administered annually by the National Center for Health Statistics, is designed to assess the general health and nutritional status of adults and children in the United States. Given to several thousands of individuals, the extent of this survey is very broad, covering demographic, laboratory and examination information, as well as responses to a fairly comprehensive health questionnaire. In this paper, we adapt and extend association rule mining and clustering algorithms to extract useful knowledge regarding diabetes and high blood pressure from the 1999–2008 survey results, thus demonstrating how data mining techniques may be used to support evidence-based medicine.

Keywords

- Medical data mining;
- Observational study;
- Evidence-based medicine;
- NHANES

1. Introduction

Evidence-based medicine is an attempt at enriching the decision-making process of healthcare professionals by collecting and analyzing data (e.g., clinical trials), making relevant results readily available for use in diagnoses, prescriptions and treatments. In this way, the evidence collected through systematic research by the larger medical community can be used to complement and extend an individual practitioner's clinical expertise [1].

With the digitization of medical publications, and the deployment of standards and tools for the systematic collection of healthcare patient data, greater opportunities are now available for evidence-based medicine. Over the past 20 years, for example, a number of researchers have focused their attention on literature-based discovery, i.e., the discovery of interesting medical facts from the medical literature [2], [3], [4], [5] and [6]. A variety of techniques, such as simple co-occurrence counts, information retrieval measures, and association rules (see [7] for a recent and fairly comprehensive survey), have been used to generate a handful of valuable hypotheses from Medline following Swanson's ABC model of discovery [2]. Most recently, literature-based discovery has been shown to be more reliable and more timely than regulatory agencies at identifying dangerous adverse drug reactions [8]. In the past decade, with the availability of more structured data and the development of novel algorithms, other researchers have increasingly turned their attention to medical data mining, i.e., the discovery of interesting patterns from observational healthcare patient data [9] and [10].

In this paper, we focus on the latter in the context of health data and statistical information collected by the National Center for Health Statistics (NCHS) with a view to improve public health. Since 1971, except for a short 4-year gap from 1994 to 1999, the NCHS has been conducting the annual National Health and Nutrition Examination Survey (NHANES) to assess the health and nutritional status of adults and children in the United States. Since 1999, the survey has been administered systematically every other year to approximately 10,000 individuals of all ages, with a response rate of about 80%. Results of the survey are released in two-year blocks, with the five blocks of data for the years 1999–2008 available electronically from the website of the Centers for Disease Control and Prevention (CDC).[2]

Because of its coverage of a broad range of health-related issues, and its combination of self-reported questionnaire data with lab results and examinations, NHANES has been a rich source of data for the investigation of specific health questions. Most NHANES-derived findings are the result of sophisticated statistical analyses. To the best of our knowledge, the application of data mining to NHANES—and indeed in the public health domain in general—has been limited. Here, we mine the 1999–2008 NHANES data for knowledge on diabetes and high blood pressure.

The motivation for our choice is two-fold. First, it is well known that data mining is more than blindly applying algorithms to data hoping that something useful will be discovered. Successful application of data mining requires domain knowledge and a clearly articulated objective or area of interest [11]. Furthermore, the application of data mining to new domains often raises interesting issues that current techniques do not handle well, and thus provides opportunities for useful algorithmic development and extensions. Second, diabetes and high blood pressure, in addition to being specific, are two significant and increasingly concerning modern health issues. Diabetes is a disease characterized by high levels of blood glucose. It is often accompanied by other serious health complications and may lead to premature death. According to the National Diabetes Information Clearinghouse's 2011 statistics, diabetes is one of the top ten leading causes of death in the United States, and 25.8 million people (8.3%) of all ages have diabetes in the United States [12]. Likewise, high blood pressure is a significant health concern. According to the American Heart Association, about 76.4 million people in the United States age 20 or older have high blood pressure [13], and the death rate due to associated complications, such as heart disease and stroke, increased 25.2% from 1995 to 2005 [14]. Hence, finding related symptoms and quantifying their relationships to high blood pressure and diabetes are valuable efforts.

To illustrate the scope of data mining technology, we make use of several complementary approaches in our analyses. We first look at simple correlations between our selected health issues and other health conditions. We then take a more global view in which we adapt and extend the MSapriori algorithm [15] to apply association rule mining effectively to our data in order to highlight which conditions are more likely to occur with each other. Finally, we propose an original definition of distance between health conditions based on the frequency of co-occurrence of Yes values among health indicators and use it as a basis for clustering, thus bringing out further interesting relationships. In all cases, we check our findings against the medical literature. We find that most of them are supported by existing medical knowledge, which validates the data mining approach. By extension, other findings through data mining can be afforded some credibility. In particular, those rules for which we find no, or little, support in the current literature may offer possibly interesting avenues of further medical investigation. While we certainly do not make any claim of definitiveness about our results, we do highlight some interesting relationships and illustrate the value of data mining in the public health domain.

The paper is organized as follows. Section 2 briefly describes several previous studies of the NHANES data, highlighting the difference with our data mining approach. Section 3 provides basic statistical information about the data, and describes the methodology followed for the important preprocessing phase of the data mining process. Section 4 describes our overall experimental framework, and discusses the results and health-related knowledge uncovered through the use of various data mining strategies. Finally, Section 5 concludes the paper.

2. Related work

While medical and health informatics have been growing rapidly in the past couple of decades with work ranging from standardization (e.g., ICD-9) to electronic health records (EHR) to Web and mobile health applications (e.g., health forums, iPhone apps, blogs), the take up of medical data mining has been somewhat slower, yet steady [9] and [10]. Unsupervised approaches, as we use here, were first introduced to the healthcare field by [16], and have since been used by a number of other researchers in a variety of subfields [17], [18], [19], [20] and [21].

Annually, diverse studies with NHANES and other similar public health data are carried out in order to gain insight in current health behavior and trends. We briefly describe a few of the most recent ones here for purposes of illustration. Bethene Ervin examined the prevalence of individual risk factors for metabolic syndrome in NHANES 2003–2006 (two consecutive 2-year cycles) and found that non-Hispanic white male individuals who are older and overweight are more likely to meet the criteria for metabolic syndrome [22]. Wright et al. analyzed NHANES 1999–2002 (two consecutive 2-year cycles) for people who had been given recommendations for reducing the incidence of cardiovascular disease and found that only about one-third of adults complied with most of the recommendations (i.e., 6 out of 9), with higher-income individuals and those age 60 year or older more likely to comply than lower-income individuals and younger people [23]. They also showed that even among high compliance individuals, most people do not follow the daily fruit intake recommendation. Fryar et al. recently presented an analysis of NHANES 1999–2004 (three consecutive 2-year cycles) for US adolescents' self-reported health risk behavior associated with alcohol, smoking and illicit drugs [24]. One key result is that, in spite of NHANES being a relatively general health survey and its content being self-reported, the behavior found in NHANES for these 12–17 year old youth is consistent with that derived from domain-specific surveys such as the National Survey on Drug Use and Health (NSDUH). Ventura analyzed data about pregnancy among unmarried women between 2002 and 2007 (three consecutive 2-year cycles) and found that the percentage of pregnancy has been increasing since 2002 and that the number of births to unmarried women was 26% higher in 2007 than in 2002 [25]. Nearly four out of 10 US births were to unmarried women in 2007. Finally, McDowell et al. examined the blood folate level—an essential vitamin for good health—from earlier NHANES data and found that, while it is known that blood folate levels tend to be low among women of childbearing age, which increases the risk of neural tube birth defects, overall, there were large increases in blood folate levels between 1988–1994 and 1999–2000 among women of childbearing age [26].

While these, and other, studies rely mostly on traditional statistical analysis techniques, in 2008, the *American Medical Informatics Association* essentially introduced NHANES to the data mining community by organizing a very open-ended *Data Mining Competition* based on the data from the NHANES 2005–2006 cycle. One significant difference with data mining approaches is that, other than deciding on the domain of focus, one does not start from any specific hypothesis to test. Rather, one lets the algorithms generate hypotheses from the data automatically. One of the authors participated in that competition [27]. The study presented here significantly extends, both in terms of data and algorithms, the results reported then. Interestingly, the winners of the competition also recently published a follow-up study in [28]. Like us, they use association mining and clustering. However, there are significant differences between their work and ours. Whereas they confine themselves to the 2005–2006 cycle and further constrain their data to patients 20 years old and older, we consider a much larger amount of data, with five consecutive 2-year cycles (including the latest available set of survey results), which, as stated above, increases the accuracy and relevance of our findings. Xing and Pei also focus on association patterns rather than association rules. We contend that in the medical domain in particular, rules are more insightful than associations as they provide direction. Furthermore, since they use any-cosine, a novel interestingness measure defined as the *maximum* of the cosines of the possible rules in an association pattern, they are liable to generating many spurious rules from their patterns. From a methodological standpoint, Xing and Pei's study is completely undirected, keeping all 26 possible diseases in their analysis. In contrast, we focus our attention on only a couple of conditions of interest, which is more in line with data mining process principles (e.g., see CRISP-DM [29]). As a result, they do have broader coverage, but their results lack the robustness of ours for the specific conditions we consider here. Finally, they cluster patients rather than diseases, which means that they need to post-process their results (e.g., via thresholding) to extract disease associations. By contrast, we cluster diseases with a novel similarity measure and thus obtain associations directly. We also use correlation to highlight strong correlations to our target diseases and thus focus attention on the most likely associations.

3. Data preprocessing

The NHANES data for each 2-year cycle consists of a collection of four distinct components, each addressing complementary aspects of health issues and behavior:

- Indicators and measurements taken during *physical examinations* by medical professionals, such as audiometry, ophthalmology, body measurements, cardiovascular fitness, oral health, and vision.
- Results from *laboratory analyses*, such as blood, urine, diabetes profile, infectious disease profile, nutritional biochemistries, miscellaneous laboratory assays, and environmental disease profile.
- Answers to a comprehensive *patient health questionnaire*, including questions about topics such as acculturation, allergy, blood pressure, dietary supplements, diet behavior and nutrition, diabetes, immunization, kidney conditions, occupation, physical activity and physical fitness, respiratory health and disease, sleep disorders, smoking and tobacco use, weight history, alcohol, tobacco, drugs, sexual behavior, and reproductive health.
- *Demographic information*, such as gender, age, and ethnicity.

In the present study, we ignore laboratory and examination data, and focus exclusively on self-reported responses to the questionnaire, since these contain explicit symptoms and disease status. The other data sources capture only indirect disease information, often consisting of real values (e.g., amount of certain chemicals), and are thus not useful in finding association among health conditions. Table 1 shows for each 2-year cycle, the number of respondents to the survey and the number of items in the questionnaire.

Table 1.

Basic dataset statistics.

Cycle	# Respondents	# Items
1999–2000	9965	1030
2001–2002	11,039	1007
2003–2004	10,122	1121
2005–2006	10,348	1089
2007–2008	10,149	905

Our motivation for considering data over all of these cycles rather than a single cycle at a time is found in the following NCHS statement: "For two-year cycles, the sample size may be too small to produce statistically reliable estimates for very detailed demographic sub-domains (e.g., sex-age-race/ethnicity groups) or for relatively rare events. The sample design for NHANES makes it possible to combine two or more cycles to increase the sample size and analytic options." [30].

Rather than all questionnaire items being in a single file, they are organized in separate files, by topic, as listed above. Because data mining algorithms typically expect all data to be in one file, we must construct such a file from disparate questionnaires across the five 2-year cycles. To do so, it is necessary to (1) join all topic-based files for each cycle, (2) select items (or attributes) relevant to our domain of interest (i.e., high blood pressure and diabetes), and (3) concatenate the resulting files to obtain a final single file for the entire study period. The procedure is broken down into the following steps:

1.

Join all 44 topic-based files in the 2005–2006 cycle into a file F_0, using the provided sequence ID (There is nothing special about starting from this cycle, other than it is where earlier efforts had focused [27].)

- Let I=∅I=∅
- For each item *i* in F_0
  - If *i* may serve as an indicator of, or proxy for, a symptom or ailment (e.g., “have you ever been told by a doctor or health professional that you had weak or failing kidneys?”, “do you now smoke cigarettes?”, etc.) then add *i* to *I*
- Let D_0 be the projection of F_0 onto *I*

2.

For each of the other four cycles

- Join all topic-based files into a file F_k, using the provided sequence ID
- For each item *i* in *I*
  - If *i* is in F_k under a different name, rename the corresponding item to *i* in F_k (Note that due to some minor changes over the years, not all items have the same name across cycles. We thus had to perform entity resolution, where we simply compared the items’ descriptions and matched them on that basis.)
  - If *i* is not in F_k, create an item *i* in F_k and set its value to “missing” for all respondents
- Let D_k be the projection of F_k onto *I*

3.

Let $D = \bigcup_{k=0}^{k=4} D_k$ D=⋃k=0k=4Dk

At the completion of this process, *D* contains the responses of 51,623 individuals to the 31 health-related indicators described in Table 2.

Table 2.

Selected health indicators.

Topic	Indicator
Audiometry	Have you ever worn a hearing aid? (AUQ150)
Blood pressure and cholesterol	Are you now taking prescribed medicine for HBP? (BPQ050A)
Cardiovascular disease	Have you had shortness of breath either when hurrying on the level or walking up a slight hill? (CDQ010)
Diet behavior and nutrition	In the past 12 months, did you go to a community program or senior center to eat prepared meals? (DBQ330)
Diabetes	Have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes? (DIQ010)
Early childhood	Did biological mother smoke at any time while she was pregnant with child? (ECQ020)
	Did child receive any newborn care in an intensive care unit, premature nursery, or any other type of special care facility? (ECQ060)
Kidney conditions: urology	Have you ever been told by a doctor or health professional that you had weak or failing kidneys? (KIQ022)
Medical conditions	Has a doctor or other health professional ever told you that you have asthma? (MCQ010)
	During the past 3 months, have you been on treatment for anemia, sometimes called “tired blood” or “low blood”? (MCQ053)
	Has a doctor or other health professional ever told you that you were overweight? (MCQ080)
	Have you ever received a blood transfusion? (MCQ092)

Topic	Indicator
	Has a doctor or other health professional ever told you that you had arthritis? (MCQ160A)
	Has a doctor or other health professional ever told you that you had congestive heart failure? (MCQ160B)
	Has a doctor or other health professional ever told you that you had coronary heart disease? (MCQ160C)
	Has a doctor or other health professional ever told you that you had angina, also called angina pectoris? (MCQ160D)
	Has a doctor or other health professional ever told you that you had heart attack (also called myocardial infarction)? (MCQ160E)
	Has a doctor or other health professional ever told you that you had a stroke? (MCQ160F)
	Has a doctor or other health professional ever told you that you had emphysema? (MCQ160G)
	Do you still have chronic bronchitis? (MCQ170K)
	Do you still have any kind liver condition? (MCQ170L)
	Have you ever been told by a doctor or health professional that you had cancer or a malignancy of any kind? (MCQ220)
Respiratory health and disease	In the past 12 months have you had wheezing or whistling in your chest? (RDQ070)
Reproductive health	Have you had a hysterectomy that is, surgery to remove your uterus or womb? (RHD280)
Sleep disorders	Have you ever been told by a doctor or health professional that you have a sleep disorder? (SLQ060)
Smoking: cigarette use	Do you now smoke cigarettes? (SMQ040)
Sexual behavior	Has a doctor or other health professional ever told you that you had genital herpes? (SXQ260)
	Has a doctor or other health professional ever told you that you had genital warts? (SXQ265)
	In the past 12 months, has a doctor or other health professional told you that you had gonorrhea, sometimes called GC or clap? (SXQ270)
	In the past 12 months, has a doctor or other health professional told you that you had chlamydia? (SXQ272)
Vision	At the present time, would you say your eyesight, with glasses or contact lenses if you wear them is. (VIQ031)

We do not claim that our selection mechanism gives a complete and accurate view of a person's health, nor that the removed items are useless. In fact, many of them are essential to the traditional and highly specialized statistical studies that focus on specific issues (e.g., osteoporosis). However, as stated earlier, we are here focusing only on gaining insight about possible high-level interactions across health components, in which case summarized information, in the form of simple indicators, may be sufficient. Interestingly, this very approach is one of the advantages of data mining techniques over statistical ones. Data mining is an analytic process designed to explore data in order to uncover patterns and/or systematic relationships among attributes. Hence, data mining techniques applied to NHANES data can reveal certain dependencies among attributes and patterns existing across several related attributes.

Finally, before turning to our analysis, we include some general statistics about the population under study. As far as gender is concerned, the respondents are spread evenly across the two possible values, with 26,306 (51%) female and 25,317 (49%) male respondents. The distribution of ethnicities is given in Table 3, while Fig. 1 shows the distribution of ages.

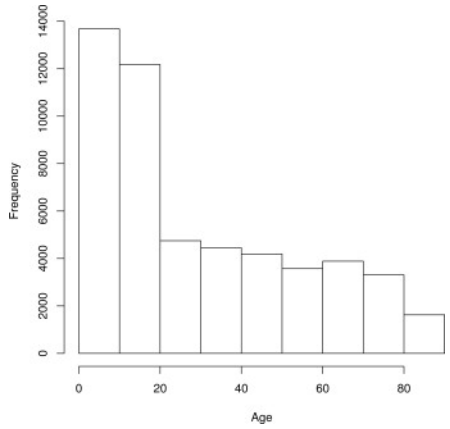
Table 3.

Distribution of ethnicity.

Ethnicity	Number (%)
Mexican-American	20,149 (39.0)
Other Hispanic	2292 (4.4)
Non-hispanic White	12,492 (24.2)
Non-hispanic Black	13,692 (26.5)
Others (incl. multi-racial)	2997 (5.8)

Fig. 1.

Respondents' age distribution.



Interestingly, and important to the generalization of any results from this data, including our own, to the general population, the NCHS claims that “each two-year cycle and any combination of those two years cycles is a nationally representative sample” [30]. Since our preprocessing involves only projections and no selections (in the database sense), all individuals are retained in our final sample, with only a subset of their possible attributes. As a result, we can infer that our sample is representative of the general US population.

4. Experimental results

Data mining techniques include data visualization [31], data pre-processing, such as feature selection and/or extraction [32], classification [33], regression [34], association rule mining [35], and clustering [36]. These techniques, while interesting and useful in their own right, can also often complement each other.

Such is the case in the present study. The NHANES data is descriptive in nature, and thus contains no explicit target as would be the case in a typical classification task. Hence, we take an exploratory approach, and focus on unsupervised data mining techniques to uncover useful facts from the data. In particular, we show results using a correlation-based approach, association rule mining and clustering.

4.1. Correlation-based mining

Our first approach at exploring the data consists in using feature selection to obtain a list of health conditions that correlate with diabetes and high blood pressure. Specifically, by arbitrarily setting a given attribute as the target class, supervised feature selection can be applied to discover those other attributes that are most relevant to the target attribute, or strongly correlated with it.

Among the several available feature selection techniques, we choose Correlation-based Feature Selection (CFS) [37] and [38]. CFS focuses on evaluating subsets of attributes rather than individual attributes. In order to evaluate competing subsets of attributes, it takes into account both the predictive power of individual attributes and the inter-correlation among them, assigning higher scores to subsets whose attributes show little correlation among themselves but high correlation with the target class. A heuristic “merit” metric that applies to a feature subset *S* containing *k* features is designed to quantify this idea, as follows:

MeritS=krcf¯k+k(k−1)rff¯

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

where $\overline{r_{cf}}$ rcf¯ is the average feature–class correlation and $\overline{r_{ff}}$ rff¯ is the average feature–feature correlation. The numerator can be thought of as giving an indication of how predictive a group of features is, and the denominator as a measure of how much redundancy there is among them.

After computing a correlation matrix, CFS often uses some kind of search procedure to find a good subset of features, which can then be ranked in terms of their contribution to the goodness of the set. CFS handles irrelevant, as well as redundant features, naturally, since these are either poor class predictors or correlated with other features. Experiments show that, for moderate levels of interaction, CFS can effectively identify useful attributes [38].

We use the implementation of CFS found in Weka [39]. As we are interested in finding possibly valuable relationships among attributes, rather than the very best subset of attributes for some target classification task—the typical setting for CFS, we did not optimize for all parameters. In this case, Weka's default parameter setting is adequate and leads to reasonable efficient computation. In particular, (1) the locallyPredictive parameter is set to true (i.e., attributes with the highest correlation with the class are added iteratively as long as there does not already exist an attribute in the subset that has higher correlation with the attribute in question), (2) the missingSeparate parameter is set to false (i.e., missing value counts are distributed across other values in proportion to their frequency), and (3) the Search Method parameter is set to Best-First, with forward direction and a maximum of five backtracking steps (i.e., the space of attribute subsets is searched by greedy hillclimbing augmented with backtracking, staring with the empty set).

When we set our diabetes indicator (DIQ010) as the target class, and run CFS as described, the extracted subset of features is as follows:

CFS(DIQ010)=
{MCQ053,MCQ080,MCQ092,MCQ160A,MCQ160B,MCQ160C,MCQ160D,MCQ160F}CFS(DIQ010)={MCQ053,MCQ080,MCQ092,MCQ160A,MCQ160B,MCQ160C,MCQ160D,MCQ160F}

This result highlights an interesting connection between diabetes and anemia (MCQ053). It is possible that people with diabetes have blood tests more often, and anemia might be detected more in this group, and it may just be that it is under-detected in people without diabetes. Yet, there may also exist a more meaningful and deeper relationship between the two. Indeed, a recent medical paper states: “Diabetes and renal failure are thus tightly linked diseases, and so is anemia. However, whether anemia may be worsened and/or directly, at least in part, caused by diabetes is not clearly elucidated yet” [40].

Using the data, we can further compute: P(*anemia*|*diabetes*)=P(MCQ053=*Yes*|DIQ010=*Yes*)=194/2874=0.0675P(anemia|diabetes)=P(MCQ053=Yes|DIQ010=Yes)=194/2874=0.0675, while P(*anemia*)=P(MCQ053=*Yes*)=1326/51623=0.0257P(anemia)=P(MCQ053=Yes)=1326/51623=0.0257. Hence, it would appear that the presence of diabetes increases the chances of also suffering from anemia by a factor of $2.6\ (0.0675/0.0257)$ 2.6(0.0675/0.0257). Although we clearly cannot derive a definitive causal relationship from this observed correlation, this result suggests that the further elucidation alluded to above may indeed be warranted.

The observed connection between diabetes and overweight (MCQ080) simply confirms what is now a well-known fact that the risk of diabetes increases when a person is overweight (e.g., see [41]). This is also true of the connection between diabetes and heart conditions (MCQ160B-F), since it is also accepted that heart-related diseases are often linked with diabetes (e.g., see [42]). It is unclear why blood transfusion (MCQ092) appears in the set of selected features, although it may be due to the preeminence of No values in the dataset. In other words the correlation found may connect the fact that individuals who have not had a blood transfusion (the majority of our sample) do not suffer from diabetes (also the majority in our sample). We return to this issue later.

When we use our high blood pressure indicator (BPQ050A) as the target class, and as with DIQ050, run CFS, we obtain the following subset:

CFS(BPQ050A)={DIQ010,MCQ092,MCQ160A,MCQ160B,MCQ160C,MCQ160F,SMQ040}CFS(BPQ050A)={DIQ010,MCQ092,MCQ160A,MCQ160B,MCQ160C,MCQ160F,SMQ040}

In this case, the highlighted correlations are mostly the reflection of known medical facts. For example, blood transfusion (MCQ092) may be accompanied by complications including high blood pressure [43]; heart conditions and stroke (MCQ160B-F) are directly related to high blood pressure; and smoking (SMQ040) is likely to cause increased blood pressure [44].

The above results suggest that a simple correlation-based feature selection approach may be useful in finding interesting patterns in the data. While it can be finely tuned, such fine tuning requires human intervention to choose the target as well as the features of interest in the features returned by the CFS algorithm. On the one hand, this improves the chances of relevance

and validity for the extracted knowledge; on the other hand, it is limited by the user's expertise. Furthermore, CFS only captures pairwise correlations, and its results may, as noted above, be negatively affected by the over-abundance of No values. In an attempt at overcoming these limitations, we now turn to the more advanced techniques of association rule mining and clustering to generate new relationships.

4.2. Association rule mining

Our second approach at exploring the data consists in using the idea of association rule mining. Association rules are statements that relate the presence of certain feature values with that of others, such as $smoking=daily \wedge exercise=none \Rightarrow health=poor$smoking=daily∧exercise=none⇒health=poor, which states that the presence of a habit of daily smoking together with a lack of exercise in an individual's life likely indicates the presence of poor health. The typical approach to association rule mining is to first extract what are called frequent itemsets, i.e., sets of feature-value conditions (or items) that co-occur sufficiently often. The minimum degree of co-occurrence, called *minimum support*, is set by the user. Once all itemsets whose support is above the minimum threshold have been generated, all possible rules are constructed and their quality gauged based on whether they are above a pre-defined *minimum confidence*, where confidence is measured as the probability of the consequent given the antecedent of the rule. The minimum confidence value is also set by the user.

As pointed out in the previous section, one of the characteristics of our data is that the No values tend to be the most prevalent ones in our health indicators, i.e., most people are healthy. As a result, correlations and other association rules that may be discovered tend to relate the absence of some conditions to the absence of others (i.e., rules where the antecedent and consequent are of the form $condition_{i_1}=No \wedge \cdots \wedge condition_{i_n}=No \Rightarrow condition_{j_1}=No \wedge \cdots \wedge condition_{j_m}=No$conditioni1=No∧⋯∧conditionin=No⇒conditionj1=No∧⋯∧conditionjm=No. While there may be value in some of these connections, the more interesting ones clearly have to do with the co-occurrence of health conditions, i.e., Yes values, rather than their absence. Note that this is analogous to what is typically done in market-basket analysis, where one is concerned with those things that are bought rather than those that are not.

Hence, we now focus our attention on Yes values only. This requires a minimal amount of extra data pre-processing. First, we replace the three values of SMQ040 ("Do you now smoke cigarettes") into two values: Yes for "Everyday" and "Some Days," and No for "Not at all." Second, we remove from our dataset all individuals with No values on all indicators, as these cannot contribute anything to the association rules we wish to induce. The resulting dataset contains 39,086 respondents.

The classic algorithm for learning association rules is the Apriori algorithm, which uses breadth-first search and a tree structure to count candidate itemsets efficiently [45]. Apriori generates candidate itemsets of size k from itemsets of size k−1k−1, and prunes the candidates which have infrequent sub-patterns. The strength of rules is characterized by both support and confidence, and only rules that meet user-defined minimum thresholds for these metrics are returned. One of the limitations of Apriori is that it uses a single minimum support value for all items. As a result, it does not do well with rare items. Indeed, it is either unable to extract rules for rare items, when the minimum support is set too high, or it extracts too many rules, making it unfeasible, when the minimum support is set too low.

In addition to the Yes values being a minority relative to the No values in our dataset, Yes values are also generally few in the absolute for most indicators considered. Hence, even in the transformed dataset, Yes values constitute rare items. Furthermore, there are noticeable variations among the various health indicators as to their frequency. It follows that Apriori is not the best choice for mining our datasets.

Recently, an extension to Apriori, known as MSapriori, was proposed, which handles both rare items and multiple minimum support thresholds [15]. A faster implementation was proposed in [46], but we are not concerned about performance here. We started with the only implementation we could find available online, from Bodon.[3] We fixed it and extended it in the following ways:

- We fixed a few bugs in the code that caused the algorithm to crash in certain circumstances.
- We added a pre-processing step so that arbitrary nominal attributes could be used, rather than binary value-only attributes.
- We added a parameter for the user to select the maximum size of rules to generate. This is useful as in many instances only the smaller rules are interesting, since they are more general and more comprehensible. It also allows a reduction of the computation time.
- We recoded the rule generation part of the program. The current implementation produces only rules with a single item in the antecedent, which is clearly not appropriate. We extended this to all possible and valid rules.
- We set the minimum item support (MIS) values of each item based on the improved formula proposed in [47]: $MIS(i)=\max\{S(i)-SD, LS\}$MIS(i)=max{S(i)−SD,LS}, where $S(i)$ is the support (or frequency) of item i , LSLS is a user-defined (default) minimum support value, and $SD=\lambda(1-\alpha)$SD=λ(1−α), with $0 \leq \alpha \leq 10 \leq \alpha \leq 1$ and λλ a parameter of the distribution of item supports, such as average, mode or maximum.
- We display the rules in descending order of confidence value, and add to each rule, its lift value. For an association rule A⇒BA⇒B, confidence is P(B|A)P(B|A) and lift is the ratio of confidence over P(B)P(B), i.e., a measure of how much the presence of A affects the presence of B.

We apply this improved version of MSapriori to our dataset,[4] with the parameter values shown in Table 4. The result is a set of 156 association rules. For further analysis, we select only rules whose confidence is greater than 0.5, which leaves us with only 37 association rules.

Table 4.

Parameter values for MSapriori.

Parameter	Value
Maximum rule size	4
LS	0.01
αα	0.25
λλ	Mean of item supports

Not surprisingly, given the results from the previous section, 19 of these 37 rules involve relationships between heart disease and stroke, and high blood pressure, as well as between

heart disease and stroke, and arthritis. In particular, we have the rules:

- MCQ160B⇒BPQ050A,MCQ160C⇒BPQ050A,MCQ160D⇒BPQ050A,MCQ160E⇒BPQ050A,MCQ160F⇒BPQ050A with confidence values ranging between 0.58 and 0.64, and lift values ranging between 4.7 and 5.2. Recall that BPQ050A stands for whether the patient is now taking prescribed medicine for high blood pressure, not whether he/she suffers from high blood pressure. MCQ160B-F on the other hand code for whether the patient has, in the past, suffered from a heart condition or a stroke. Given that high blood pressure is known to increase the risk for heart disease and stroke (e.g., see [48]), it is likely that this rule captures the fact that doctors probably prescribe blood pressure regulating medicine to heart patients to limit further cardiac accidents and strokes.

$$\begin{aligned}\mathrm{MCQ160B} &\Rightarrow \mathrm{BPQ050A},\ \mathrm{MCQ160C} \Rightarrow \mathrm{BPQ050A},\ \mathrm{MCQ160D}\\ &\Rightarrow \mathrm{BPQ050A},\ \mathrm{MCQ160E} \Rightarrow \mathrm{BPQ050A},\ \mathrm{MCQ160F}\\ &\Rightarrow \mathrm{BPQ050A}\end{aligned}$$

- MCQ160B⇒MCQ160A,MCQ160C⇒MCQ160A,MCQ160D⇒MCQ160A,MCQ160E⇒MCQ160A,MCQ160F⇒MCQ160A with confidence values ranging between 0.55 and 0.60, and lift values ranging between 4.1 and 4.5. These rules denote the known connection between heart disease and stroke, and arthritis. Indeed, “arthritis and heart diseases often occur simultaneously…a recent study found that arthritis affects 57% of adults with heart disease” [49].

$$\begin{aligned}\mathrm{MCQ160B} &\Rightarrow \mathrm{MCQ160A},\ \mathrm{MCQ160C} \Rightarrow \mathrm{MCQ160A},\ \mathrm{MCQ160D}\\ &\Rightarrow \mathrm{MCQ160A},\ \mathrm{MCQ160E} \Rightarrow \mathrm{MCQ160A},\ \mathrm{MCQ160F}\\ &\Rightarrow \mathrm{MCQ160A}\end{aligned}$$

The other nine rules further confirm the medical knowledge by establishing various other connections among these conditions, e.g., MCQ160A∧MCQ160B⇒BPQ050AMCQ160A∧MCQ160B⇒BPQ050A and MCQ160C⇒MCQ160EMCQ160C⇒MCQ160E.

When considering our second target condition, our result set contains four rules involving diabetes, namely:

1. DIQ010(diabetes)∧MCQ160A(arthritis)⇒BPQ050A(highbloodpressure), conf: 0.67, lift: 5.4.

$$\mathrm{DIQ010}(\mathit{diabetes}) \wedge \mathrm{MCQ160A}(\mathit{arthritis}) \Rightarrow \mathrm{BPQ050A}(\mathit{high\ blood\ pressure})$$

2. DIQ010(diabetes)∧VIQ031:2(eyesightisgood)⇒BPQ050A(highbloodpressure), conf: 0.60, lift: 4.8.

$$\mathrm{DIQ010}(\mathit{diabetes}) \wedge \mathrm{VIQ031:2}(\mathit{eyesight\ is\ good}) \Rightarrow \mathrm{BPQ050A}(\mathit{high\ blood\ pressure})$$

3. DIQ010(diabetes)⇒BPQ050A(highbloodpressure), conf: 0.57, lift: 4.6.

$$\mathrm{DIQ010}(\mathit{diabetes}) \Rightarrow \mathrm{BPQ050A}(\mathit{high\ blood\ pressure})$$

4. BPQ050A(highbloodpressure)∧DIQ010(diabetes)⇒MCQ160A(arthritis), conf: 0.54, lift: 4.0.

$$\mathrm{BPQ050A}(\mathit{high\ blood\ pressure}) \wedge \mathrm{DIQ010}(\mathit{diabetes}) \Rightarrow \mathrm{MCQ160A}(\mathit{arthritis})$$

As with MCQ160B-F and BPQ050A, the first three rules are likely to correspond to the prescription of blood pressure regulating medicine to diabetes patients to prevent further “development and worsening of many complications of diabetes” caused by high blood pressure [50]. The fourth rule suggests that the presence of arthritis is increased in the presence of diabetes and high blood pressure. This is interesting as it has been argued that although diabetes and arthritis are not directly related, they often overlap. “In fact, recent reports from the Centers for Disease Control and Prevention (CDC) found that more than half (52%) of people with diabetes also have arthritis. The two diseases have several other commonalities depending on the different chemicals in the body that reduce glucose levels” [51].

While they only relate indirectly to our target conditions (esp. high blood pressure), we mention an interesting set of six rules involving hysterectomy:

1. $\mathrm{MCQ092}(\mathit{blood\ transfusion}) \wedge \mathrm{RHD280}(\mathit{hysterectomy}) \Rightarrow \mathrm{MCQ160A}$

   MCQ092(bloodtransfusion)∧RHD280(hysterectomy)⇒MCQ160A(arthritis), conf: 0.64, lift: 4.7.

2. BPQ050A(highbloodpressure)∧RHD280(hysterectomy)⇒MCQ160A(arthritis), conf: 0.63, lift: 4.6.

$$\mathrm{BPQ050A}(\mathit{high\ blood\ pressure}) \wedge \mathrm{RHD280}(\mathit{hysterectomy}) \Rightarrow \mathrm{MCQ160A}(\mathit{arthritis})$$

3. RHD280(hysterectomy)∧VIQ031:2(eyesightisgood)⇒MCQ160A(arthritis), conf: 0.56, lift: 4.2.

$$\mathrm{RHD280}(\mathit{hysterectomy}) \wedge \mathrm{VIQ031:2}(\mathit{eyesight\ is\ good}) \Rightarrow \mathrm{MCQ160A}(\mathit{arthritis})$$

4. MCQ160A(arthritis)∧RHD280(hysterectomy)⇒BPQ050A(highbloodpressure), conf: 0.56, lift: 4.5.

$$\mathrm{MCQ160A}(\mathit{arthritis}) \wedge \mathrm{RHD280}(\mathit{hysterectomy}) \Rightarrow \mathrm{BPQ050A}(\mathit{high\ blood\ pressure})$$

5. RHD280(*hysterectomy*)⇒MCQ160A(*arthritis*)RHD280(hysterectomy)⇒MCQ160A(arthritis), conf: 0.55, lift: 4.0.

6. $\mathrm{RHD280}(\mathit{hysterectomy}) \wedge \mathrm{VIQ031:2}(\mathit{eyesight\ is\ good}) \Rightarrow \mathrm{BPQ050A}(\mathit{high\ blood\ pressure}),$

   RHD280(hysterectomy)∧VIQ031:2(eyesightisgood)⇒BPQ050A(highbloodpressure), conf: 0.53, lift: 4.3.

These rules point to what appears to be a strong relationship between arthritis and hysterectomy. Interestingly, it is easy to find statements aimed at the general public, for example, “women who have had hysterectomies are more at risk for heart disease, arthritis and osteoporosis” [52], or similarly that sex hormones, which are clearly affected by hysterectomy, play a critical role in arthritis. However, we were not able to find definitive medical papers on this connection. As a matter of fact, a short clinical study from 1938 suggests that hysterectomy may, in some cases, actually alleviate arthritis [53].

We note here the synergy alluded to in the introduction between data mining and literature-based discovery. For some of the rules found above (e.g., arthritis and diabetes, arthritis and hysterectomy), the evidence from the literature is rather small. In such cases, discovering rules, as we have, that exhibit reasonable support, and relatively high confidence and lift, over large amounts of observational data, confirms said evidence and encourages further investigation.

4.3. Clustering

We complete our study with a third approach at analyzing data, based on clustering, i.e., the automatic discovery of groups of similar items. There are a number of available clustering algorithms. The most popular ones are distance-based in that they rely on some notion of distance, or reciprocally similarity, among items to construct clusters. These algorithms generally fall into two classes: partitional methods, which produce a single clustering through iterative re-assignment to centroid-based clusters, and hierarchical methods, which produce a series of nested clusterings through iterative proximity-based merging. Examples of the former include the well-known *k*-means algorithm [54] and [55], while examples of the latter include hierarchical agglomerative clustering [56] and [57]. Here, we will focus on hierarchical agglomerative clustering, as it provides more information about the data, in particular the order in which clusters are formed. The graphical representation, in the form of a dendrogram, is also rather intuitive.

As with association rule mining, we focus on Yes values only, and propose an original definition of distance between health conditions based on the frequency of co-occurrence of Yes values among health indicators. This approach is equivalent to counting the number of respondents suffering from multiple health conditions. The basic idea is that given two conditions C_1 and C_2, if there are many respondents suffering from both, then C_1 and C_2 are related.

Formally, Let *i* and *j* be two indicator attributes. The distance, *Dist*(i,j)Dist(i,j), between *i* and *j* is given by the following Jaccard index-like equation:

Sim(i,j)=value(i,j,Y)value(i,Y)+value(j,Y)−value(i,j,Y)

$$Sim(i,j) = \frac{value(i,j,Y)}{value(i,Y)+value(j,Y)-value(i,j,Y)}$$

Dist(i,j)=1−*Sim*(i,j)Dist(i,j)=1−Sim(i,j)

where *value*(i,j,Y)value(i,j,Y) represents the number of respondents who answered Yes for both indicator attributes *i* and *j* , and *value*(i,Y)value(i,Y) is the number of respondents who answered Yes for indicator attribute *i*.

The results of our correlation and association analyses suggest that some of the symptoms and diseases included in our set of indicators (see Table 2) have stronger relations among them than others. The complementary analysis with clustering described here is an attempt at isolating these stronger interactions, and discovering new ones that may have been occulted from the earlier results due to the presence of the additional, less relevant indicators. As a result, we focus on 16 indicators of well-defined symptoms/diseases that were highlighted as highly relevant in the previous analyses. This is rather standard practice in data mining studies. The 16 indicators used in clustering the data are as follows:

- Heart disease and stroke: MCQ160B, MCQ160C, MCQ160D, MCQ160E, MCQ160F.
- Arthritis: MCQ160A.
- Diabetes: DIQ010.
- Blood pressure: BPQ050A.
- Early childhood: ECQ020, ECQ060.
- Other conditions: MCQ010 (asthma), MCQ053 (anemia), MCQ170K (chronic bronchitis), RDQ070 (wheezing), SLQ060 (sleep disorder), KIQ022 (weak/failing kidneys).

Fig. 2 shows the clustering structure obtained from our data, using complete link hierarchical agglomerative clustering with the *agnes* function from the *cluster* package of R [58].

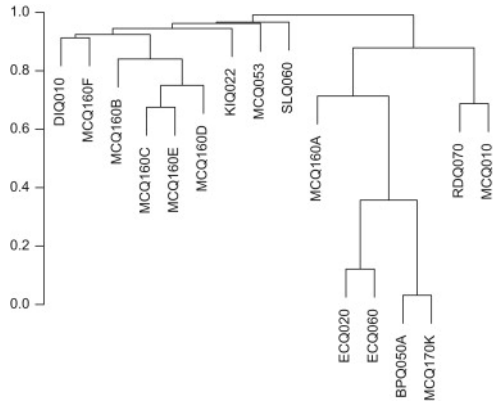


Fig. 2.

Health conditions clustering.

Interestingly, the high-level structure of the dendrogram reveals a fairly clean separation between heart-related conditions to the left and lung-related conditions to the right. As nothing was done to point the algorithm in that direction, this demonstrates the value of clustering techniques in finding natural groupings in data.

Looking at the rest of the clustering, many parts are without surprise. For example, to the left of the dendrogram, all heart conditions are grouped together first (MCQ160B-E), followed by stroke (MCQ160F) and then diabetes (DIQ010), in line with common sense and the earlier highlighted connection between diabetes and heart disease/stroke. This group is then linked in close succession with conditions having to do with failing kidneys (KIQ022), anemia (MCQ053) and sleep disorder (SLQ060). While it may be difficult to explain all of these, the co-occurrence of several has been reported in the medical literature. We list a few interesting, less evident examples here.

- Diabetes, kidney weakness/failure and anemia:

  Anemia occurs when the level of hemoglobin in the blood falls below normal levels. According to the National Anemia Action Council, a significant proportion of individuals (about 1 in 3) with a long history of type 1 diabetes (over 15 years) end up developing kidney disease, and a number of people suffering from kidney disease become subject to anemia [59]. Hence, the relationship between diabetes and anemia, through kidney disease, seems also rather strong.

- Sleep disorder and kidney weakness/failure:

  There are many specific symptoms regarding sleep disorder, which makes it difficult to conclude with confidence. There may also be more indirect connections between these conditions. For example, people who are overweight (not included here) often suffer from co-morbidities including diabetes and sleep disorder, so that a relationship may exist between sleep disorder and kidney weakness/failure through diabetes. Of the respondents who said they suffered from both sleeping disorder and kidney weakness/failure, 72% also said they were overweight. One recent study, however, does show that the risk of sleep apnea, one of several sleep disorders, is increased in patients with early chronic kidney disease [60].

As to the right-hand side of the dendrogram, it also exhibits both rather well-known connections, such as the relationships between ECQ020 and ECQ060, and between MCQ010 and RDQ070, as well as possibly less well-known ones, such as the relationship between chronic bronchitis and high blood pressure. Yet, all of these have been the object of prior medical studies. It is, for example, widely recognized that smoking during pregnancy (ECQ020) is likely to have malignant effects on the newborn (ECQ060). A recent study shows that the newborns of mothers who continue to smoke through their pregnancy have a significantly higher risk of suffering from infantile colic than their counterparts whose mothers stopped smoking during pregnancy and the postpartum period [61]. Similarly, wheezing or whistling in the chest (RDQ070) is a common breathing problem related with a number of lung diseases, such as asthma (MCQ010) [62]. Finally, high blood pressure (BPQ050A) has been identified as one of the likely complications of chronic bronchitis, since it is often accompanied by a constriction of certain blood vessels [63].

Smoking is one of the causes of chronic bronchitis, likely explaining why the corresponding conditions merge next in the dendrogram, and both chronic bronchitis and smoking are serious causes of wheezing, which again probably explain why these all also join together on the same side of the dendrogram. Finally, the presence of arthritis on the lung disease side of the dendrogram is likely due to their connection through rheumatoid lung disease [64].

5. Discussion and conclusions

Randomized Controlled Trials (RCTs) have long been the norm in medical research. In a typical RCT, researchers focus on a single health condition wherein two or more groups of individuals are constituted such that the groups are identical in every way, except in the treatment they receive relative to the condition under study. The main advantage of RCTs, and the reason for their popularity in the medical field, is that they make it possible to establish statistical significance and causality between condition and treatment. On the other hand, RCTs require dedicated resources and careful experimental design (e.g., volunteers, control groups, etc.), and may, in some instances, be entirely unsuitable [65].

By contrast, evidence-based medicine takes advantage of the vast amount of information collected in standard medical encounters (e.g., visits to the doctor, results of laboratory analyses, etc.) or targeted health questionnaires (e.g., NHANES), which can be analyzed with statistical and data mining tools (e.g., see [66], [67], [68] and [69]). Online activities, such as writing blogs or exchanging messages in social media applications (e.g., Facebook, Twitter), as well as increased use of mobile technology, have also been shown to leave behind "digital breadcrumbs—tiny records of daily experiences" that, when mined and analyzed, can provide insight into health behavior and health outcomes [70] and [71]. Because they rely on data collected through patients' experiences or during routine care, the observational studies of evidence-based medicine reduce the risk of analysis bias [72] and publication bias [73]. Furthermore, whereas causality may not strictly be inferred from observations, the existence of strong correlations in large volumes of data certainly suggests the possibility of causation, which may then be further investigated through clinical trials. Hence, observational studies provide a valuable complement to traditional clinical trials. As pointed out by Johnson and Ambrose, "health care researchers would do well to tap into [the] bountiful pool of experiential information, to supplement their more structured research, to gain insights that might otherwise be lost...Such insights may be invaluable for medical research, where the causal factors of a disease, or the pattern of adverse drug interactions, are difficult to establish" [74]. Conversely, results from clinical trials may be used, for example, to guide problem formulation and feature selection in observational studies.

While observational studies are data-driven, they are not blind. It is well accepted that successful applications of data mining require domain knowledge and a clearly articulated objective or area of interest [11], and consist not only of a single activity but also a complete, and generally iterative, process involving data transformation, algorithm adaptation, model creation, results evaluation and dissemination [29]. We have followed such a process here in the context of mining the 1999–2008 NHANES data. In doing so, we have shown how the selection of specific health conditions (here, diabetes and high blood pressure) leads to a number of relevant "discoveries," most of which have been validated by extant medical literature. In that sense, we have demonstrated that mining techniques may be a reliable means of obtaining valuable medical information, and certainly where data is available in sufficient quantity and quality, it can beneficially supplement the more traditional RCTs that form the basis of most published medical results. In a few instances, in particular in the relationships between diabetes and anemia, and between arthritis and hysterectomy, our results also bring out mostly undocumented, but possibly interesting, connections that may inform further medical research.

These results have been made possible through the synergistic combination of several data mining approaches. We used correlation mining with our two selected conditions as target to discover which other conditions and symptoms are most likely to be predictive of these conditions. We then went beyond traditional correlational analysis and took a more global view by (1) adapting and extending the MSapriori association rule mining algorithm and (2) using an original definition of distance between health conditions based on the frequency of co-occurrence of Yes values among health indicators as the basis for clustering. The extension to MSapriori is such that the algorithm can now be used with arbitrary nominal attributes to produce all valid rules of user-defined fixed size based on automatically set minimum supports for each item as per the work of Kiran and Reddy [47]. As far as clustering is concerned, our novel measure of distance makes it possible to mine clusters of health conditions directly rather than have to go through the clustering of patients first as in [28].

Our analysis of the NHANES data demonstrates how the use of data mining in novel application domains has the potential to contribute both to the domain to which it is applied as previously unknown nuggets of information may be extracted and to the technology itself where the specific context often leads to the adaptation, extension and development of algorithmic approaches. With the recent approval of the medical subspecialty on clinical informatics [75], we expect that studies like ours will multiply in the biomedical domain, and that evidence-based medicine will become a valuable complement to more traditional RCTs.

6. Summary

With the digitization of medical publications, and the deployment of standards and tools for the systematic collection of healthcare patient data, greater opportunities are now available for evidence-based medicine. The National Health and Nutrition Examination Survey (NHANES), administered annually by the National Center for Health Statistics, is designed to assess the general health and nutritional status of adults and children in the United States. Given to several thousands of individuals, the extent of this survey is very broad, covering demographic, laboratory and examination information, as well as responses to a fairly comprehensive health questionnaire, thus providing a rich collection of health data. While several statistical studies with NHANES and other similar public health data have been carried out in order to gain insight in current targeted health behavior and trends, to the best of our knowledge, the application of data mining to NHANES—and indeed in the public health domain in general—has been limited.

In this paper, we mine the 1999–2008 NHANES data for knowledge on diabetes and high blood pressure. We make use of several complementary approaches in our analyses. We first look at simple correlations between our selected health issues and other health conditions. We then take a more global view in which we adapt and extend the MSapriori algorithm to apply association rule mining effectively to our data in order to highlight which conditions are more likely to occur with each other. Finally, we propose an original definition of distance between health conditions based on the frequency of co-occurrence of Yes values among health indicators and use it as a basis for clustering, thus bringing out further interesting relationships. In all cases, we check our findings against the medical literature. We find that most of them are supported by existing medical knowledge, which validates the data mining

approach. By extension, other findings through data mining can be afforded some credibility. In particular, those rules for which we find no, or little, support in the current literature may offer possibly interesting avenues of further medical investigation. While we certainly do not make any claim of definitiveness about our results, we do highlight some interesting relationships and illustrate the value of data mining in the public health domain.

Conflict of interest statement

None declared.

Acknowledgments

We wish to thank Yao Huang Lin, David Wilcox, Matthew Smith, Udip Pant and Kevin Murdock for valuable discussions, assistance in running experiments and comments on the paper. We also wish to thank the organizers of the American Medical informatics Association 2008 Data Mining Competition for bringing this data to our attention and encouraging our efforts.

Appendix A. Association rules

Table 5 shows all 37 association rules with confidence greater than 0.5 obtained by MSapriori, ordered in decreasing value of confidence.

Table 5.

Association rules.

Rule	Confidence	Lift	Support
MCQ160A, MCQ160B ⇒⇒ BPQ050A	0.6907	5.5927	0.0076
MCQ160A, MCQ160F ⇒⇒ BPQ050A	0.6672	5.4024	0.0077
DIQ010, MCQ160A ⇒⇒ BPQ050A	0.6652	5.3862	0.0172
MCQ160A, MCQ160C ⇒⇒ BPQ050A	0.6641	5.3773	0.0082
MCQ160A, MCQ160E ⇒⇒ BPQ050A	0.6497	5.2607	0.0087
BPQ050A, MCQ160B ⇒⇒ MCQ160A	0.6496	4.8047	0.0076
MCQ160B ⇒⇒ BPQ050A	0.6382	5.1676	0.0117
MCQ092, RHD280 ⇒⇒ MCQ160A	0.6351	4.6975	0.0107
MCQ170K ⇒⇒ RDQ070	0.6328	4.9709	0.0092
MCQ160F ⇒⇒ BPQ050A	0.6282	5.0866	0.0129
BPQ050A, RHD280 ⇒⇒ MCQ160A	0.6274	4.6405	0.0162
BPQ050A MCQ160E ⇒⇒ MCQ160A	0.6149	4.5481	0.0087
MCQ160C ⇒⇒ BPQ050A	0.6129	4.9628	0.0138
MCQ160D ⇒⇒ BPQ050A	0.6033	4.8850	0.0106
MCQ160D ⇒⇒ MCQ160A	0.6033	4.4623	0.0106
MCQ160B ⇒⇒ MCQ160A	0.6002	4.4393	0.0110
BPQ050A, MCQ160C ⇒⇒ MCQ160A	0.5977	4.4209	0.0082
DIQ010, VIQ031=2 ⇒⇒ BPQ050A	0.5973	4.8364	0.0128
BPQ050A, MCQ160F ⇒⇒ MCQ160A	0.5949	4.4001	0.0077
MCQ092, MCQ220 ⇒⇒ MCQ160A	0.5867	4.3395	0.0085
MCQ160E ⇒⇒ BPQ050A	0.5816	4.7093	0.0141
SLQ060 ⇒⇒ MCQ080	0.5808	5.8845	0.0096
MCQ170K ⇒⇒ MCQ160A	0.5728	4.2367	0.0083
DIQ010 ⇒⇒ BPQ050A	0.5706	4.6202	0.0318
BPQ050A, MCQ220 ⇒⇒ MCQ160A	0.5627	4.162	0.0115
RHD280, VIQ031=2 ⇒⇒ MCQ160A	0.5625	4.1605	0.0132

Rule	Confidence	Lift	Support
MCQ160F ⇒⇒ MCQ160A	0.5601	4.1428	0.0115
MCQ160A, RHD280 ⇒⇒ BPQ050A	0.5580	4.5182	0.0162
MCQ160C ⇒⇒ MCQ160A	0.5517	4.0806	0.0124
MCQ160E ⇒⇒ MCQ160A	0.5504	4.0710	0.0133
RHD280 ⇒⇒ MCQ160A	0.5462	4.0399	0.0291
KIQ022 ⇒⇒ BPQ050A	0.5461	4.4219	0.0080
BPQ050A, DIQ010 ⇒⇒ MCQ160A	0.5402	3.9956	0.0172
RHD280, VIQ031=2 ⇒⇒ BPQ050A	0.5271	4.2680	0.0124
MCQ160A, MCQ220 ⇒⇒ BPQ050A	0.5152	4.1717	0.0115
AUQ150 ⇒⇒ MCQ160A	0.5129	3.7936	0.0123
MCQ160C ⇒⇒ MCQ160E	0.5095	21.0537	0.0114

References

1.
   - [1]
   - D.L. Sackett, W.M.C. Rosenberg, J.A. Muir Gray, R.B. Haynes, W.S. Richardson
   - Evidence-based medicine: what it is and what it isn't
   - Br. Med. J., 312 (7023) (1996), pp. 71–72
   - [SD-008]
2.
   - [2]
   - D.R. Swanson
   - Medical literature as a potential source of new knowledge
   - Bull. Med. Libr. Assoc., 78 (1) (1990), pp. 29–37
   - [SD-008]
3.
   - [3]
   - M.D. Gordon, S. Dumais
   - Using latent semantic indexing for literature based discovery
   - J. Am. Soc. Inf. Sci., 49 (8) (1998), pp. 674–685
   - [SD-008]
4.
   - [4]
   - P. Srinivasan
   - Text mining: generating hypotheses from medline
   - J. Am. Soc. Inf. Sci. Technol., 55 (5) (2003), pp. 396–413
   - [SD-008]
5.
   - [5]
   - M. Weeber, J.A. Kors, B. Mons
   - Online tools to support literature-based discovery in the life sciences
   - Briefings Bioinform., 6 (3) (2005), pp. 277–286
   - [SD-008]

6.
- [6]
- M. Yetisgen-Yildiz, W. Pratt
- Using statistical and knowledge-based approaches for literature-based discovery
- J. Biomed. Inf., 39 (6) (2006), pp. 600–611
- [SD-008]

7.
- [7]
- M.C. Ganiz, W.M. Pottenger, C.D. Janneck, Recent Advances in Literature based Discovery, Technical Report LU-CSE-05-027, Lehigh University, CSE Department, 2005.
- [SD-008]

8.
- [8]
- K.D. Shetty, S. Dalal
- Using information mining of the medical literature to improve drug safety
- J. Am. Med. Inf. Assoc., 18 (2011), pp. 668–674
- [SD-008]

9.
- [9]
- K.J. Cios, Medical Data Mining and Knowledge Discovery, Studies in Fuzziness and Soft Computing, vol. 60, Springer, 2001.
- [SD-008]

10.
- [10]
- J.H. Harrison Jr.
- Introduction to the mining of clinical data
- Clin. Lab. Med., 28 (1) (2008), pp. 1–7
- [SD-008]

11.
- [11]
- A. Montgomery, Data mining: business hunching, not just data crunching, in: Proceedings of the Second International Conference on Practical Applications of Knowledge Discovery and Data Mining, 1998, pp. 39-48.
- [SD-008]

12.
- [12]
- Centers for Disease Control and Prevention. National Diabetes Fact Sheet: National Estimates and General Information on Diabetes and Prediabetes in the United States, 2011. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, Online at 〈〈http://diabetes.niddk.nih.gov/DM/PUBS/statistics/〉〉 (retrieved 10 August 2012), 2011.
- [SD-008]

13.
- [13]
- American Heart Association. About High Blood Pressure, Online at 〈〈http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/AboutHighBloodPressure/About-High-Blood-Pressure_UCM_002050_Article.jsp〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

14.
- [14]
- Medical Disability Advisor. High Blood Pressure, Benign, Online at 〈〈http://www.mdguidelines.com/high-blood-pressure-benign〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

15.
- [15]
- B. Liu, W. Hsu, Y. Ma, Mining association rules with multiple minimum supports, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999, pp. 337–341.
- [SD-008]

16.
- [16]
- S. Stilou, P.D. Bamidis, N. Maglaveras, C. Pappas
- Mining association rules from clinical databases: an intelligent diagnostic process in healthcare
- Stud. Health Technol. Inf., 84 (Pt 2) (2001), pp. 1399–1403
- [SD-008]

17.
- [17]
- S. Doddi, A. Marathe, S.S. Ravi, D.C. Torney
- Discovery of association rules in medical data
- Med. Inf. Internet Med., 26 (1) (2001), pp. 25–33
- [SD-008]

18.
- [18]
- C. Creighton, S. Hanash
- Mining gene expression databases for association rules
- Bioinformatics, 19 (1) (2003), pp. 79–86
- [SD-008]

19.
- [19]
- A.M. Berger, C.R. Berger
- Data mining as a tool for research and knowledge development in nursing
- Comput. Inf. Nursing, 22 (3) (2004), pp. 123–131
- [SD-008]

20.
- [20]
- T. Mikos, N. Maglaveras, K. Pantazis, D. Goulis, J. Bontis, J. Papadimas
- The use of data mining in the categorization of patients with azoospermia
- Hormones (Athens Greece), 4 (4) (2005), p. 214
- [SD-008]

21.
- [21]
- T. Imamura, S. Matsumoto, Y. Kanagawa, B. Tajima, S. Matsuya, M. Furue, H. Oyama
- A technique for identifying three diagnostic findings using association analysis
- Med. Biol. Eng. Comput., 45 (2007), pp. 51–59
- [SD-008]

22.
- [22]
- R. Bethene Ervin, Prevalence of metabolic syndrome among adults 20 years of age and over, by sex, age, race and ethnicity, and body mass index. National Health Statistics Report, Number 13, 2009.
- [SD-008]

23.
- [23]
- J.D. Wright, R. Hirsch, C.-Y. Wang, One-third of US adults embraced most heart healthy behavior in 1999–2002. NCHS Data Brief, Number 17, 2009.
- [SD-008]

24.
- [24]
- C.D. Fryar, M.C. Merino, R. Hirsch, K.S. Porter, Smoking, alcohol use, and illicit drug use reported by adolescents aged 12–17 years: United States, 1999–2004. National Health Statistics Report, Number 15, 2009.
- [SD-008]

25.
- [25]
- S.J. Ventura, Changing patterns of nonmarital childbearing in the united states. NCHS Data Brief, Number 18, 2009.
- [SD-008]

26.
- [26]
- M.A. McDowell, D.A. Lacher, C.M. Pfeiffer, J. Mulinare, M.F. Picciano, J.I. Rader, E.A. Yetley, J. Kennedy-Stephenson, C.L. Johnson, Blood folate levels: the latest NHANES results. NCHS Data Brief, Number 6, 2008.
- [SD-008]

27.
- [27]
- H.Y. Lin, J. Lee, M. Smith, Dependency mining on the 2005–2006 national health and nutrition examination survey data, in: Notes of a Special Session on Data Mining Competition at the Annual Symposium of the American Medical Information Association, 2008.
- [SD-008]

28.
- [28]
- Z. Xing, J. Pei
- Exploring disease association from the NHANES data: data mining, pattern summarization, and visual analytics
- Int. J. Data Warehousing Mining, 6 (3) (2010), pp. 10–27
- [SD-008]

29.
- [29]
- P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: Step-by-step Data Mining Guide. Technical Report, SPSS, Inc., 2000.
- [SD-008]

30.
- [30]
- National Center for Health Statistics. Analytic and reporting guidelines: the national health and nutrition examination survey (NHANES), Online at: 〈〈http://www.cdc.gov.ezproxy.lib.uwf.edu/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf〉〉, 2005.
- [SD-008]

31.
- [31]
- J. Heer, M. Bostock, V. Ogievetsky
- A tour of the visualization zoo
- Commun. ACM, 53 (6) (2010), pp. 50–67
- [SD-008]

32.
- [32]
- I. Guyon, A. Elisseeff
- An introduction to variable and feature selection
- J. Mach. Learn. Res., 3 (2003), pp. 1157–1182
- [SD-008]

33.
- [33]
- S.B. Kotsiantis
- Supervised machine learning: a review of classification techniques
- Informatica, 31 (2007), pp. 249–268
- [SD-008]

34.
- [34]
- J. Fox
- Applied Regression Analysis, Linear Models, and Related Methods
- SAGE Publications (1997)
- [SD-008]

35.
- [35]
- J. Hipp, U. Guntzer, G. Nakhaeizadeh
- Algorithms for association rule mining: a general survey and comparison
- SIGKDD Explorations, 2 (1) (2000), pp. 58–64
- [SD-008]

36.
- [36]
- R. Xu, D. Wunsch II
- Survey of clustering algorithms
- IEEE Trans. Neural Networks, 16 (3) (2005), pp. 645–678
- [SD-008]

37.
- [37]
- M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the Seventeenth International Conference on Machine Learning, 2000, pp. 359–366.
- [SD-008]

38.
- [38]
- M.A. Hall, G. Holmes
- Benchmarking attribute selection techniques for discrete class mining
- IEEE Trans. Knowl. Data Eng., 15 (2003), pp. 1437–1447
- [SD-008]

39.
- [39]
- I.H. Witten, F. Eibe
- Data Mining: Practical Machine Learning Tools and Techniques
- (second ed.)Morgan Kaufmann, San Francisco, CA (2005)
- [SD-008]

40.
- [40]
- G. Deray, A. Heurtier, A. Grimaldi, V. Launay Vacher, C. Isnard Bagnis
- Anemia and diabetes
- Am. J. Nephrol., 24 (5) (2004), pp. 522–526
- [SD-008]

41.
- [41]
- American Diabetes Association. Diabetes basics: overweight, Online at 〈〈http://www.diabetes.org/diabetes-basics/prevention/checkup-america/overweight.html〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

42.
- [42]
- NewYork-Presbyterian Hospital. Diabetes and heart disease, Online at 〈〈http://nyp.org/health/diabetes-heart.html〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

43.
- [43]
- E. Martin, How to lower high blood pressure after blood transfusion. eHow, Online at 〈〈http://www.ehow.com/how_5614374_lower-pressure-after-blood-transfusion.html〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

44.
- [44]
- WebMD. High blood pressure and smoking, Online at 〈〈http://www.webmd.com/hypertension-high-blood-pressure/guide/kicking-habit〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

45.
- [45]
- R. Agrawal, T. Imielinski, A.N. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the ACM International Conference on the Management of Data (SIGMOD), 1993, pp. 207–216.
- [SD-008]

46.
- [46]
- I.N. Kouris, C.H. Makris, A.K. Tsakalidis, An improved algorithm for mining association rules using multiple support values, in: Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, 2003, pp. 309–313.
- [SD-008]

47.
- [47]
- R.U. Kiran, P.K. Reddy, An improved multiple minimum support based approach to mine rare association rules, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, 2009, pp. 340–347.
- [SD-008]

48.
- [48]
- WebMD. High blood pressure and stroke, Online at 〈〈http://www.webmd.com/hypertension-high-blood-pressure/guide/hypertension-high-blood-pressure-stroke〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

49.
- [49]
- Arthritis Foundation. Link between arthritis and heart disease, Online at 〈〈http://www.arthritis.org.ezproxy.lib.uwf.edu/heart-disease-connection.php〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

50.
- [50]
- WebMD. Diabetes and high blood pressure, Online at 〈〈http://www.webmd.com/hypertension-high-blood-pressure/guide/high-blood-pressure〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

51.
- [51]
- Arthritis Foundation. Arthritis and diabetes, Online at 〈〈http://www.arthritis.org.ezproxy.lib.uwf.edu/arthritis-and-diabetes.php〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

52.
- [52]
- eHow Contributor, How to recognize the side effects of a hysterectomy, Online at 〈〈http://www.ehow.com/how_2107627_recognize-hysterectomy-side-effects.html〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

53.
- [53]
- B. Solomons
- Hysterectomy for rheumatoid arthritis
- Irish J. Med. Sci., 13 (1) (1938), p. 28
- [SD-008]

54.
- [54]
- J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1967, pp. 281–297.
- [SD-008]

55.
- [55]
- L. Kaufman, P.J. Rousseeuw
- Finding Groups in Data: An Introduction to Cluster Analysis
- John Wiley & Sons, Inc. (1990)
- [SD-008]

56.
- [56]
- S.C. Johnson
- Hierarchical clustering schemes
- Psychometrika, 2 (1967), pp. 241–254
- [SD-008]

57.
- [57]
- B. Fung, K. Wang, M. Ester, Hierarchical document clustering using frequent itemsets, in: Proceedings of the SIAM International Conference on Data Mining, 2003, pp. 59–70.
- [SD-008]

58.
- [58]
- R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.
- [SD-008]

59.
- [59]
- National Anemia Action Council. Handouts: anemia & diabetes, Online at 〈〈http://www.anemia.org/patients/information-handouts/diabetes/〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

60.
- [60]
- J.J. Sim, S.A. Rasgon, D.A. Kujubu, V.A. Kumar, L.A. Liu, J.M. Shi, T.T. Pham, S.F. Derose
- Sleep apnea in early and advanced chronic kidney disease
- Chest, 135 (3) (2008), pp. 710–716
- [SD-008]

61.
- [61]
- C. Sondergaard, T.B. Henriksen, C. Obel, K. Wisborg
- Smoking during pregnancy and infantile coli
- Pediatrics, 108 (2) (2001), pp. 342–346
- [SD-008]

62.
- [62]
- MedlinePlus. Wheezing, Online at 〈〈http://www.nlm.nih.gov.ezproxy.lib.uwf.edu/medlineplus/ency/article/003070.htm〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

63.
- [63]
- PDRhealth. Chronic bronchitis symptoms, Online at 〈〈http://www.pdrhealth.com/diseases/chronic-bronchitis/symptoms〉〉 (retrieved 10 August 2012), 2012.
- [SD-008]

64.
- [64]
- C. Eustice, What is rheumatoid lung disease? About.com Guide, Online at 〈〈http://arthritis.about.com/od/rheumatoidarthritis/a/rheumatoidlung.htm〉〉 (retrieved 10 August 2012), 2011.
- [SD-008]

65.
- [65]
- G.C.S. Smith, J.P. Pell
- Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials
- Br. Med. J., 327 (2003), pp. 1459–1461
- [SD-008]

66.
- [66]
- N. Aoki, M.J. Wall, J. Demsar, B. Zupan, T. Granchi, M.A. Schreiber, J.B. Holcomb, M. Byrne, K.R. Liscum, G. Goodwin, J.R. Beck, K.L. Mattox
- Predictive model for survival at the conclusion of a damage control laparotomy
- Am. J. Surg., 180 (6) (2000), pp. 540–545
- [SD-008]

67.
- [67]
- D.P. Strum, A.R. Sampson, J.H. May, L.G. Vargas
- Surgeon and type of anesthesia predict variability in surgical procedure times
- Anesthesiology, 92 (5) (2000), pp. 1454–1466
- [SD-008]

68.
- [68]
- J. Bocsi, J. Hambsch, P. Osmancik, P. Schneider, G. Valet, A. Tarnok
- Preoperative pprediction of pediatric patients with effusions and edema following cardiopulmonary bypass surgery by serological and routine laboratory data
- Critical Care (London, England), 6 (3) (2002), pp. 226–233
- [SD-008]

69.
- [69]
- A. Kusiak, C.A. Caldarone, M.D. Kelleher, F.S. Lamb, T.J. Persoon, A. Burns
- Hypoplastic left heart syndrome: knowledge discovery with a data mining approach
- Comput. Biol. Med., 36 (1) (2006), pp. 21–40
- [SD-008]

70.
- [70]
- A. Pentland, D. Lazer, D. Brewer, T. Heibeck
- Using reality mining to improve public health and medicine
- Stud. Health Technol. Inf., 149 (2009), pp. 93–102
- [SD-008]

71.
- [71]
- D. Estrin, I. Sim
- Health care delivery open mhealth architecture: an engine for health care innovation
- Science, 330 (6005) (2010), pp. 759–760
- [SD-008]

72.

- [72]
- R. Peto, C. Baigent
- Trials: the next 50 years
- Br. Med. J., 317 (7167) (1998), pp. 1170–1171
- [SD-008]

73.

- [73]
- M.A. Maggard, L.R. Shugarman, M. Suttorp, M. Maglione, H.J. Sugerman, E.H. Livingston, N.T. Nguyen, Z. Li, W.A. Mojica, L. Hilton, S. Rhodes, S.C. Morton, P.G. Shekelle
- Meta-analysis: surgical treatment of obesity
- Ann. Internal Med., 142 (7) (2005), pp. 547–559
- [SD-008]

74.

- [74]
- G.J. Johnson, P.J. Ambrose
- Neo-tribes: the power and potential of online communities in health care
- Commun. ACM, 49 (1) (2006), pp. 107–113
- [SD-008]

75.

- [75]
- R.M. Gardner, J.M. Overhage, E.B. Steen, B.S. Munger, J.H. Holmes, J.J. Williamson, D.E. Detmer
- Core content for the subspecialty of clinical informatics
- J. Am. Medical Inf. Assoc., 16 (2) (2009), pp. 153–157
- [SD-008]