Electronic Medical Records for Discovery Research in Nonalcoholic Fatty Liver Disease

Uri Kartoun (uri.kartoun@ibm.com)
IBM Research https://orcid.org/0000-0003-0988-8037
Rahul Aggarwal
Boston University
Adam Perer
CMU
Yoonyoung Park
IBM
Ping Zhang
OSU
Heng Luo
BenevolentAI
Sanjoy Dey
IBM
Kathleen Corey
Massachusetts General Hospital
Kenney Ng
IBM

Research article

Keywords: Clinical informatics; Predictive modeling; Electronic medical records; Data mining; Patient outcomes; Nonalcoholic fatty liver disease

Posted Date: January 2nd, 2020

DOI: https://doi.org/10.21203/rs.2.11644/v3

Abstract

Background: Nonalcoholic fatty liver disease (NAFLD) is a highly prevalent yet under-diagnosed and under-discussed disease. Given that NAFLD has not been explored sufficiently compared with other diseases, opportunities abound for scientists to discover new biomarkers (such as laboratory observations, current comorbidities, behavioral descriptors) that can be linked to the development of conditions and complications that may develop at a later stage of the patient's life. Methods: We analyzed IBM Explorys, a repository that contains electronic medical records (EMRs) of more than 60 million individuals. We used a classification algorithm that members of our group have previously validated to identify patients at a high probability for NAFLD. The algorithm identified more than 80,000 patients with a high probability for NAFLD who had at least 5 years of follow-up. We applied standard statistical methods (such as logistic regression and bootstrapping) and used Clinical Classifications Software (CCS) definitions to identify associations between a variety of covariates and disease outcomes. Results: Our methodology identified several thousand strongly statistically significant associations between covariates and outcomes in NAFLD. Most of the associations are known, but others may be new and require further investigation in subsequent studies. Conclusions: A discovery mechanism composed of standard statistical methods applied on a large collection of EMRs, confirmed known associations and identified potentially new associations that can act as biomarkers that might merit further research.

Background

Nonalcoholic fatty liver disease (NAFLD) is characterized by the accumulation of excess fat within the liver and is associated with several risk factors, including obesity, type 2 diabetes mellitus (T2DM), and metabolic syndrome. Furthermore, NAFLD can develop in those with normal body mass index (BMI), and the risk factors for this condition are poorly understood. A NAFLD diagnosis has important health and clinical implications because it is a risk factor for the development of diseases such as T2DM and is an independent risk factor for cardiovascular-related mortality and all-cause mortality [1,2]. Nonalcoholic steatohepatitis, the progressive form of NAFLD, can result in cirrhosis and hepatocellular carcinoma and is expected to become the leading indication for transplant in the United States by 2020 [3].

The combination of increased computational power and the availability of experiential data has facilitated the more rapid development and use of algorithms to identify patients at high risk for disease complications, to discover new biomarkers, and to improve the understanding of NAFLD. Most of the published research to assess clinical outcomes in patients with NAFLD has focused on one or only a few outcomes per study [4,5]. Developing systematic discovery mechanisms capable of assessing a large collection of disease outcomes either individually or in combination is expected to yield a better understanding of the interplay between the development or progression of NAFLD and other comorbidities. Furthermore, such discovery mechanisms will be able to identify novel variables (such as medications or laboratory values) that are strongly associated with a disease, whether positively or negatively correlated.

A discovery engine can identify known associations as well as propose potentially new ones; analyzing large collections of electronic medical records (EMRs) from a population at a high probability for NAFLD allows the evaluation of a variety of associations between laboratory observations, comorbidities, and behavioral covariates. Increased creatinine level, for instance, has been widely reported to be associated with measuring renal functioning in the general population [12,13]. Regarding NAFLD, increased creatinine and reduced estimated glomerular filtration rate (eGFR) have been studied extensively and have been found to be linked to kidney malfunction [2,14]. NAFLD's associations with high levels of HbA1c (Hemoglobin A1c) and a high prevalence of T2DM are also well known [15].

Increased HbA1c and cardiovascular conditions are well-known associations [16]. A discovery engine can help in studying the correlation of HbA1c with different types of cancers (e.g., prostate cancer [17]). In particular it could help in identifying inverse associations with developing cancers given an increased HbA1c, as well as other factors such as an increased BMI or comorbid conditions related to the metabolic syndrome [18–20].

Another use of a discovery engine is to identify benefits in behavior that is well known to be harmful, such as tobacco use. For example, smoking could be identified as associated with the development of a variety of diseases but protective against others; at least one study reported on the potential protective effect of nicotine against glaucoma [28].

The aim of the present study was to develop a discovery mechanism to assess associations between patient-level covariates (diagnosis codes, laboratory measurements, and demographic descriptors) and disease outcomes in patients suspected with NAFLD. Our discovery engine confirmed known associations and identified new associations that can act as potential biomarkers that might merit further research.

Methods

2.1. Study population

We used the IBM Explorys clinical data set, which has more than 60 million patient records pooled from different health-care systems with EMRs (The IBM Explorys Network) [6]. The data were standardized and normalized using common ontologies, searchable through a HIPAA-enabled, de-identified cloud-computing platform. Patients were seen in multiple health-care systems between January 1, 1999, and December 31, 2015, with a combination of data from clinical EMRs, outgoing health-care system bills, and adjudicated payor claims.

We first defined a broad cohort of patients suspected of having liver disease (any diagnosis, test, or procedure related to the liver) or at least one measurement of an elevated triglyceride laboratory test (a core component of the NAFLD classification algorithm that we applied described further in the text). This definition resulted in a population of approximately 5 million patients; those included patients diagnosed with a liver-related disease as well as patients who went through liver evaluations (e.g., biopsy, computed

tomography) with no liver-related diseases found. Given our objective to identify physiological and comorbid biomarkers, we excluded all patients who had an indication either by a diagnosis code or a lab result indicative of viral hepatitis or human immunodeficiency virus (HIV); consequently, approximately 11,000 patients were removed. Patients at high probability for NAFLD were identified using a classification algorithm that members of our group have previously validated [7–9, 45]. In brief, the algorithm relies on three components for each patient: 1) total number of NAFLD International Classification of Diseases (ICD) codes, 2) total number of notes each contains at least one mention of NAFLD, including negations, and 3) most recent level of triglyceride. The algorithm calculates probability for NAFLD per patient, and if the probability exceeds a 0.85 threshold, then the patient is labeled as “high probability for NAFLD” patient. The probability threshold was selected as 0.85 to provide an optimal combination of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). To exceed the threshold, for instance, a patient in the base cohort needed to have a triglyceride measurement of 367mg/dL or above. By applying this algorithm to the EMR database of patients suspected with a liver-disease-related finding (the population of 5 million), we identified the first date on which each patient developed a high probability for NAFLD (index date), resulting in a population of 334,258 patients. We selected only patients who were over the age of 18 at the index date and had at least 5 years of follow-up after the index date using encounter entries (e.g., office visit, admission, emergency room visit, or observation). This identification yielded a population of 81,911 patients with a high probability for NAFLD. Table 1 presents patient characteristics for our cohort.

We extracted all disease outcomes during the 5-year follow-up and mapped them using the Clinical Classifications Software (CCS) categories based on the International Classification of Diseases (ICD) 9th and 10th revisions [10]. We defined a prevalence cut-off threshold of 1% for analyzing the outcomes to assure statistical robustness; keeping only those that were prevalent in at least 1% of the NAFLD population yielded 174 unique disease categories to be used as binary-variable outcomes in our analyses. All correlations between covariates and outcomes considered are presented in Supplementary Table 1 and Supplementary Figs. 1 and 2.

Table 1. Characteristics of high probability for NAFLD cohort

Variable and category	Overall (n = 81,911)
Age (years); Mean (SD)	53.5 (12.5)
Gender (%)	
Male	60.1
Female	39.9
Ethnicity (%)	
Caucasian	87.8
African American	7.5
Other / Unknown	4.7
Top 10 comorbidities in prevalence (%)	
Disorders of lipid metabolism	33.5
Essential hypertension	29.2
Diabetes mellitus without complications	17.9
Other connective tissue disease	14.5
Spondylosis; intervertebral disc disorders; other back problems	13.9
Other upper respiratory infections	12.2
Other nutritional; endocrine; and metabolic disorders	12.0
Other non-traumatic joint disorders	11.2
Esophageal disorders	10.9
Other lower respiratory disease	10.8

2.2. Discovery engine focused on NAFLD

To identify associations between covariates and outcomes in a follow-up window of 5 years after the index date, we developed a process that extracts a large collection of structured variables from the EMRs. The variables included demographics (e.g., gender, ethnicity, age), comorbidities, laboratory measurements (e.g., albumin, sodium, BMI), and behavioral descriptors (e.g., smoking status, alcohol use). For laboratory variables, we used the most recent values found in the 12 months preceding the index date. We determined the existence of a comorbidity if at least one CCS code for this comorbidity was found in the patient's problem list before the index date. The disease outcomes in the 5-year follow-up window consisted of the 174 CCS categories described earlier. To assess the potential associations between these variables and disease outcomes, we excluded patients for each outcome who had a diagnosis for that specific outcome disease before the index date. In this way, we assessed only the newly developed disease outcomes after the index date. We imputed the missing values with the mean of the available data for each variable and performed all programming using Python (libraries: pandas.io.sql [for data manipulation and analysis], pyodbc [for accessing databases], time [to handle dates and times]) and R (libraries: ggplot2 [for creating plots], Hmisc [a variety of functionalities], stringr [for string manipulation]).

We applied the process of NAFLD index date extraction and feature selection, (Fig. 1). We followed this methodology for each of the considered outcomes, resulting in 174 different separate experimental sets. To select a subset of the potentially most predictive variables, we first applied univariate analyses to all

covariates related to each outcome. We compared categorical variables using a chi-squared test and compared the differences in the means of continuous variables using a *t*-test or Wilcoxon rank sum test as appropriate. All statistical tests were two-sided, with Bonferroni corrections for the 314 comparisons; the adjusted *P* value threshold for statistical significance was 1.6×10^{-4} for each comparison. Three hundred and nine of the 314 covariates were continuous (including 281 CCS-defined comorbidities and 28 laboratory observations). Five of the 314 covariates were categorical, including smoking status, alcohol use, gender, and ethnicities (White, African American).

We used all covariates that were statistically significant in the univariate comparison to train a multivariate logistic regression model. The model with this smaller set of covariates yielded odds ratios (ORs) and *P* values for each covariate. We then took a stringent approach for feature selection and used the variables that were statistically significant ($P < 0.05$) to train another multivariate logistic regression model. This two-step training provided an increased level of confidence for the reliability of the significance of the selected variables. Finally, to account for variability, we applied bootstrapping to the reduced set of covariates, excluding those with confidence intervals that crossed an OR of 1. This methodology was capable of identifying a subset of covariates that were strongly correlated with each outcome.

2.3. General approach to extend the discovery engine

Although our manuscript focuses primarily on exploring NAFLD, we believe that the scientific community may be interested in using our proposed methodology and applying it to additional conditions. Fig. 2 illustrates a general representation of the steps required to assess associations, not necessarily restricted to NAFLD, between covariates and outcomes. The first step requires identifying a population with a disease of interest and the date of first diagnosis of the disease for each patient. For certain diseases, assessing when the patient is at a high risk for the disease is straightforward; for example, patients with high blood pressure levels (a commonly measured observation) may be candidates for the engine to explore hypertension outcomes. Other conditions are known to be under-documented and underdiagnosed; thus, a computational algorithm can help assess the probability of disease occurrence (e.g., 41).

The second step is to define a large collection of distinct outcomes: for example, diagnoses, procedures, uncontrolled laboratory observations, and mortality. While not mandatory, a follow-up period should also be defined (e.g., 30 days, 5 years).

The third step is to define and extract covariates; these could be comorbidities defined by ICD or CCS codes, laboratory values, demographic details, and covariates extracted from the clinical narrative notes (e.g., smoking status, alcohol use, nonadherence, family history of cardiovascular disease) as well as more detailed covariates such as genetics-related factors and measurements captured from wearable devices or edible sensors. An observation window for the covariates should be defined (e.g., most recent value for a laboratory value within the preceding 12 months, a history of hospital admissions unrestricted

by time, whether colonoscopy was performed within the past decade, whether an immunization was performed over the past 12 months). Another definition may include whether the disease outcome has been observed for the first time after the index date or if it is a recurrence of a preexisting condition.

Once index dates, covariates, and outcomes are defined, the forth step is to extract the actual data from the EMR database; this will result in a table for each outcome with different covariate values, all relative to the index date of each patient and the subsequent presence or absence of the outcome during the follow-up window. For each such data frame, the engine then applies a feature selection algorithm. In the NAFLD use case, we followed a traditional epidemiological approach to select features (i.e., applying a univariate analysis on the covariates and outcome, filtering out covariates with no statistical significance [given a predefined threshold], and then applying a logistic regression model on the statistically significant covariates and the outcome). Although in the context of NAFLD we assessed levels of association by following a statistical approach (i.e., using *P* values and ORs), alternative approaches may be possible as well (e.g., relying on importance scores calculated by a machine learning algorithm). More advanced feature selection methods may result in more trustable linkages; however, there is no guarantee that such methods may hold any significant difference compared to following a standard statistical approach. Thus, which feature selection method is the most preferable is an open research question within the context of discovery mechanisms focused on EMRs.

Once the engine has provided levels of association between each outcome and covariates, there could be several potential approaches to interpretation, as Fig. 2 shows. In a desirable scenario, the engine is capable of validating results that are already known (e.g., tobacco use is harmful). In another scenario, also possibly desirable, the engine is capable of identifying a linkage that has not yet been reported. Such a scenario may trigger the scientific community to evaluate the correctness of the linkage: for example, by designing and applying an experiment in a wet laboratory (e.g., evaluating the potential linkage between covariates and outcome in mice or in zebra fish [e.g., 46]). Additional possible approaches to test the linkage would be to extract the outcome and covariates of interest at a different medical site (using other EMRs or claims-based data) to evaluate whether the linkage is valid in additional layouts. In another scenario the engine may identify an association; however, the association may be in a different order in time. For example, the engine finds a linkage between carrying disease A and the future development of disease B, consistent with the literature; however, the engine indicates that disease B is actually associated with a subsequent development of disease A. Such scenarios may be possible (we provide examples in the discussion) and may stimulate further debate by the scientific community regarding the interplay between the two diseases.

2.4. Visualization

We were also interested in providing a high-level overview of how the covariates and outcomes were connected so interesting relationships could be revealed across all 174 experiments. As illustrated in Fig. 3, we created a network visualization that shows covariates and outcomes as nodes (rendered as a circle) with edges (rendered as a line) connecting them if they had a statistically significant OR relationship. We

colored covariate nodes gray, whereas outcome nodes each received a unique color. We sized nodes proportionally to their degree centrality, so those with more connections were larger, and nodes with few connections were smaller. We colored the edges connecting nodes according to the nodes to which they were connected. Thick edges had an OR greater than 1 (a positive association between the covariate and the outcome), whereas thin edges had an OR less than 1 (a negative association between the covariate and the outcome). Because of space limitations, we only show a partial network in Fig. 3, illustrating the connections between several outcomes related to diseases of the circulatory system. The nodes on the border of the figure show the covariates associated with only a single outcome, whereas the nodes in the center are shared across multiple outcomes. Such a visualization allows researchers to see how certain covariates can have different relationships among different outcomes.

2.5. Availability of data and materials

IBM's Data Access and Compliance Board approved this study and all its methods, including the EMR cohort assembly, data extraction, and analyses. Data contain potentially identifying information and may not be shared publicly. The data sets used and/or analyzed during the current study, as well as the source code used to develop the engine, are available from the corresponding author on reasonable request (address: 75 Binney St, Cambridge, MA 02142, USA; telephone: 857-500-2425; uri.kartoun@ibm.com).

Results

Our engine identified a variety of strongly statistically significant associations between covariates and outcomes in NAFLD. In this section, we decided to present a subset of the results; those related to covariates that are commonly available in the EMR to diagnose or monitor highly prevalent diseases. Note that all associations are available in Supplementary Table 1 and Supplementary Figs. 1 and 2. We stratified the covariates into subtypes—physiological (laboratory observations) and observational (comorbidities)—given the different methods required to collect such covariates (i.e., laboratory observations are gathered by taking urine and blood samples from the patient, whereas comorbidities are diagnosed and reported by the clinical staff). A unique covariate that required special attention and interest is the smoking status covariate, a behavioral covariate whose harmful effects on a broad range of diseases are known but which potentially has beneficial effects on a small number of diseases. In a subsequent subsection focused on visualization, we emphasize the importance of visualizing associations between covariates and outcomes for rapidly observing a large number of associations.

3.1 Association between covariates and outcomes

3.1.1. Laboratory covariates

Many of the laboratory and disease associations were consistent with findings reported in the literature. For instance, increased levels of creatinine were associated with the development of a variety of diseases related to kidney function, including chronic kidney disease, diseases of the kidney and ureters, and nephritis. HbA1c was positively correlated with diabetes with or without complications. HbA1c was also

positively correlated with other known associated outcomes (Fig. 4a). For instance, increased HbA1c was positively correlated with retinal defects, nephritis, acute MI, and glaucoma. Increased levels of creatinine and HbA1c were negatively correlated with several conditions. Increased creatinine, in contrast, was negatively correlated with prolapse and menopausal disorders, and increased HbA1c was negatively correlated with ulcerative colitis and prostate cancer. These are candidates for future confirmatory research that may potentially lead to new discoveries.

Interestingly, only two covariates were associated with the development of phlebitis: international normalized ratio (INR) and albumin. Although our engine found that INR was associated with a series of cardiovascular complications in NAFLD, consistent with another study published by members of our group [8], it has not yet been reported as an independent covariate that may predict phlebitis unrestricted to any specific disease.

Increased hemoglobin was correlated with decreased fatigue and decreased anemia, consistent with well-known associations of increased prevalence of anemia/fatigue with lower hemoglobin (Fig. 4b). Additionally, increased hemoglobin was also associated with benign prostatic hyperplasia and non-epithelial cancers; total bilirubin was associated with hemorrhoids and diverticulosis; increased calcium was correlated with decreased gastritis and gastroduodenal ulcers; and chloride was positively correlated with urinary stones. We found various other correlations with lab values, but given the short space, we present only the outcomes associated with two of the most commonly measured labs, HbA1c and hemoglobin (Fig. 4(a–b)).

3.1.2. Comorbidity covariates

Our engine identified a variety of known links between different diseases as well as several unreported associations. Our engine found, for instance, that the development of acute MI was associated with known factors such as preceding coronary atherosclerosis, being a smoker, and being male. One unexpected association was that acute MI was associated with a preceding vehicle accident.

Varicose veins were positively correlated with age and BMI. Interestingly, the strongest association with varicose veins was Parkinson’s disease. As expected, osteoporosis was negatively correlated with being male and was positively correlated with multiple myeloma, as expected; interestingly, however, it was also positively correlated with sepsis and multiple sclerosis. Transient ischemic attacks were correlated with traditional risk factors such as coronary disease and peripheral atherosclerosis but negatively correlated with diastolic blood pressure and albumin. As expected, Parkinson’s disease was associated with a subsequent indication for falls and with a wide range of mental conditions including mood disorders, dementia, and schizophrenia. In addition, systemic lupus erythematosus (SLE) has broadly been described as being linked to the development of cervical cancer [11]. Indeed, our engine highlighted a strong association between the two diseases. The link, however, had a different order of association, with the occurrence of cervical cancer preceding the development of SLE.

3.1.3. Smoking status covariate

The covariate indicating status as a current smoker was positively associated with 48 of the outcomes. These included behavioral conditions such as screening for and history of mental health and substance abuse, substance-related disorders, and mood and anxiety disorders. Several respiratory conditions (e.g., chronic obstructive pulmonary disease [COPD], respiratory failure, asthma, and bronchitis) were positively correlated with being a smoker as well. Several cardiovascular outcomes were associated with being a smoker, including peripheral and visceral atherosclerosis and acute MI. Cataracts and glaucoma were the only covariates that were negatively associated with being a smoker.

3.2. Visualization

Our engine identified several thousand strongly statistically significant links between covariates and outcomes. To help in observing and interpreting the results, we created network visualizations, as shown in Fig. 3 (because of space limitations, we show a partial network, illustrating the connections between several outcomes related to diseases of the circulatory system). Such visualizations can help researchers quickly review which covariates are linked to multiple diseases (e.g., age, INR, diabetes) and which are linked to only one disease (such as smoking to MI and carditis to congestive heart failure [CHF]), and the strength of the associations. One specific example is the alanine aminotransferase test (ALT) covariate (located near the top center of Fig. 3), which had a positive association with essential hypertension but negative associations with acute MI and CHF.

Discussion

Our discovery engine can identify known associations as well as propose potential new ones. By analyzing large collections of EMRs from a population at a high probability for NAFLD, we were able to evaluate associations between laboratory observations, comorbidities, and behavioral covariates. We took an unbiased approach for feature selection, a data-driven approach in which a large collection of covariates were considered as candidates for selection without the need for a human domain expert. We used all possible comorbidity covariates defined by the Agency for Healthcare Research and Quality (i.e., CCS codes). Additionally, the covariates included a comprehensive collection of common laboratory covariates as well as traditional factors such as age, gender, and ethnicity. This approach for unbiased feature selection combined with a stringent variable-filtering process provided increased confidence in the results. Our engine could be used for fast screening of a variety of covariates, risk factors, and their associations and could further be used as a hypothesis generator for additional experiments to be carried out by other researchers.

Our engine identified several thousand strongly statistically significant associations between covariates and outcomes in NAFLD. Most of the associations are known, but many others may be new and require further investigation in subsequent studies. Within the scope of a short scientific paper, it was not straightforward for us to decide which findings to describe because our engine found so many interesting correlations. We thus highlighted only a few findings.

Increased creatinine level, for instance, has been widely reported to be associated with measuring renal functioning in the general population [12,13]. Regarding NAFLD, increased creatinine and reduced eGFR have been studied extensively for their links to kidney malfunction [2,14]. It is also well known that NAFLD is associated with high levels of HbA1c and a high prevalence of T2DM [15]. Consistently, increased levels of HbA1c were associated with T2DM complications (e.g., retinal defects, nephritis, glaucoma). Our findings that T2DM complications (e.g., hyperglycemia or diabetic ketoacidosis) were strongly associated with HbA1c seem reasonable given the broad literature; however, these complications have not yet sufficiently been explored in NAFLD patients.

Increased HbA1c and cardiovascular conditions are well-known associations [16], and our engine was able to identify this association as well (e.g., Acute MI in Fig. 4a). We also identified a negative correlation between HbA1c and ulcerative colitis, something not captured in the literature. More interestingly, increased levels of HbA1c were correlated with decreased prostate cancer [17], but not with a broad range of other types of cancer. This potential for inverse association with developing cancers, given increased HbA1c, increased BMI, or comorbid conditions related to metabolic syndrome, has already been reported [18–20], but further studies are required to assess such associations more precisely. Notable was the correlation of increased creatinine with decreased diseases of the female genital organs (e.g., prolapse and menopausal disorders). Such an association has not yet been reported and therefore requires further investigation.

As further reassurance that our findings have the potential to identify real-life connections, increased hemoglobin was correlated with decreased fatigue. This finding is sound from a biological plausibility standpoint because hemoglobin is the oxygen-carrying molecule of the body, so its increase in it would be expected to decrease fatigue. These associations are well known in studies focused on specific populations [21] but not in NAFLD. Increased hemoglobin was also associated with benign prostatic hyperplasia and non-epithelial cancers. These findings are an example of a topic for further research because the link between them is not well studied. Additional findings identified by our engine were consistent with the literature, such as those for calcium [22] and chloride [23].

SLE has been associated with the development of cervical cancer, and our results were similar [11]. Our results suggest that cervical cancer may precede the development of SLE, contrary to proposals that SLE predisposes a patient to cervical cancer. Autoimmune diseases are well known to have environmental factors that increase the likelihood of disease development. It is possible that an infection with human papillomavirus, one of the well-known risk factors for cervical cancer development, is an environmental trigger [24] that increases the likelihood of developing SLE. Further characterization of this correlation could reveal valuable insight into disease pathogenesis.

Beyond known cardiovascular-related comorbidities and smoking associated with the development of MI, our engine identified that an injury related to a vehicle, train, or motorcycle accident, defined by CCS codes as “Motor Vehicle Traffic (MVT),” may be a powerful predictor (OR, 1.92; 95% CI, 1.31–2.74). Although this association had already been proposed [25], it has usually been reported in the lay press, unrelated to

NAFLD. Notably, this covariate was not associated with other cardiovascular outcomes in NAFLD. In clinical practice, acute MI is rare after MVT crashes, and cardiac contusion is more likely to result in high cardiac enzymes being miscoded as acute MI. Our observation thereby should be interpreted with caution because it may be simply a coding artifact rather than clinically meaningful.

Parkinson's disease diagnosis. This finding has not yet been reported. This association offers a variety of possible ideas to explore. For example, maybe there is a subset of Parkinson's patients who are at increased risk for venous insufficiency. Another possibility is that maybe a subset of patients experience this as an adverse reaction of medication. It is even possible that only those with NAFLD and Parkinson's experience this feature. It may also be possible to have a common genetic underpinning for comorbid conditions. Our engine can generate such hypotheses relating covariates to NAFLD outcomes, which can further be used for genetic studies.

As expected, being a woman with no history of osteoporosis was strongly associated with the development of osteoporosis. Another known association was that a history of multiple myeloma was associated with the development of osteoporosis [27]. Interestingly, being a patient with established sepsis or multiple sclerosis was associated with the development of osteoporosis. Other known associations, such as the fact that transient ischemic attacks were positively associated with the development of traditional cardiovascular-related risk factors but negatively correlated with albumin, contribute to the increased confidence of the accuracy of our engine.

Although our engine identified associations between smoking and the development of a variety of diseases, it found that smoking was protective against glaucoma and cataracts. These findings are surprising, though at least one study has reported similar results regarding glaucoma and the potential protective effect of nicotine [28]. Determining a patient's smoking status (past, present, none) accurately is challenging. We extracted the statuses by using the social history table; however, it could be that the prevalence of current smoking status used in our study was underestimated. Smoking status may be stored in additional resources such as Systematized Nomenclature of Medicine diagnosis codes. Clinical narrative notes also serve as a primary source to document patient smoking status, presenting additional challenges for extracting the statuses accurately [29,30].

In the univariate analysis, we found that osteoporosis was positively correlated with the development of a urinary tract infection (UTI) (2.6% vs. 1.0%; $P < 0.000001$) among patients with NAFLD. In the multivariate regression, however, the direction of association changed (OR, 0.84; 95% CI, 0.71–0.98). This suggests that interpreting the results from our engine depends on the clinical context and thus should be based on general clinical knowledge as well as prior research results. A similar example is the negative association observed between prostate cancer and coronary atherosclerosis. Several publications have reported on results that seem to confirm this reduced risk. Although this association may only be minor and of limited statistical significance [31,32], it may have more of an impact in the NAFLD population.

Conclusion

Our discovery engine has several benefits for future research studies. First, rapidly implementing such a methodology on EMRs is relatively simple. With the increased availability of EMRs, this type of study can be used in a variety of clinical research centers. Second, our approach can investigate and synthesize knowledge from a large number of patients. For example, our starting point for this analysis was approximately 60 million patients and resulted in data from over 80,000 patients with a high probability for NAFLD, yielding a tremendously large cohort. Third, this type of study allows for better recognition of any type of interaction between one disease and another. For example, our study could be used to explore whether having NAFLD changes the trajectory of diabetes compared with diabetic patients without NAFLD. Such findings could be valuable as medicine moves away from a one-size-fits-all model to one of increased precision and personalization.

We have developed a large-scale discovery engine that can analyze large volumes of EMR data to validate known associations between covariates and outcomes and propose likely candidates for future research investigations. We demonstrated the effectiveness of our approach on a large EMR population of patients with a high probability for NAFLD. Our unbiased approach for feature selection, which includes a stringent approach to filtering out features with limited significance, may prove to be a useful tool for the efficient screening of clinical hypotheses. Our engine differs in several ways from other methodologies investigating the importance of features in medical knowledge discovery (e.g., 33–35). First, our study reports on a probabilistic approach to identify patients with suspected NAFLD diagnoses. It has been well reported that relying on diagnosis codes alone underrepresents the true prevalence of the NAFLD population [7, 8], and our engine attempts to address this limitation. Our engine further relies on defining disease comorbidities in a highly unbiased paradigm: relying on standard groups of diagnosis codes (CCS) to define each disease and attempting to capture every possible condition while preserving statistical power (as opposed to defining each disease as a single diagnosis code, which is expected to reduce statistical power). Another novelty of our engine is its ability to test any number of hypotheses (174 outcomes were tested); this approach differs substantially from that of traditional EMR-based studies in which one or a small number of hypotheses are evaluated (e.g., 36).

A potential improvement of our engine would be to integrate into it more advanced feature selection methods. In its current version the engine relies on a two-step feature selection method that is commonly used in epidemiological studies (i.e., a univariate analysis followed by logistic regression). This approach has resulted in a substantial number of scientific works, including analyses focused on T2DM (e.g., 37) and NAFLD (e.g., 8). While it is not guaranteed, our engine may benefit from other methods for selecting features, especially those that better capture correlations between covariates (e.g., 38, 39), as such methods were found highly useful in a variety of studies (e.g., 40, 41, 36).

Limitations

Our study has several limitations. First, it is a retrospective analysis that has the potential for confounding effects. Although our engine confirmed many known associations between covariates and outcomes, subsequent studies must further assess the validity of our results and consider different age

ranges, coding systems, and data-collection methods. Second, our engine reported on potentially contradictory results or counterintuitive associations. This could be a result of confounding variables from unmeasured or unadjusted factors, an observed variable acting as a proxy for a different factor, or a spurious association arising from the use of a multiple testing data-driven approach using EMRs. Also, while we identified many associations, future large-scale analyses and methods will benefit if they would be capable of identifying the possible associations for disease linkages that are the most clinically meaningful. We believe that with advances in computation (e.g., novel security methods, data access methods, and algorithms) this will be increasingly common. Another potential limitation of our results is the use of the mean imputation technique, one of the simplest and computationally efficient imputation methods. The scientific community constantly debates whether advanced imputation techniques are preferable to simple ones. Many studies often advise against using mean-based imputation techniques (e.g., 42). Other studies, however, report on the minor importance of imputation types (e.g., 43). The trade-off of applying simple vs. advanced imputation techniques in prediction modeling has also been discussed in (44), which suggests the need for further investigation. Another challenge in evaluating outcomes of sub-populations would be to decide on a prevalence cut-off threshold to assure the analyses of statistically meaningful cohorts; while we decided to analyze sub-populations with outcome prevalence higher than 1%, future studies may benefit from defining other threshold values to potentially identify linkages between covariates and outcomes in sub-populations that only rarely associated with the explored outcome. Also, for our analysis, we excluded viral hepatitis and HIV to increase specificity for identifying the high probability for NAFLD population. However, we did not exclude every possible liver pathology since we have used our validated algorithm for NAFLD (7). Not excluding rare diseases is less concerning as the prevalence of these diseases is much smaller compared to major diseases like viral hepatitis and HIV. Finally, we were primarily interested in NAFLD for adult populations. We preferred this in case of differing phenotypes for NAFLD in children vs. adults. We do believe that a study such as this evaluating NAFLD in younger ages would be of interest as well and is something to definitely be considered for future work.

Despite these limitations, the scientific community can consider the new linkages our engine has identified as a starting point for additional exploration and validation. Such potential validations could be carried out in forms other than the analysis of medical databases (e.g., wet labs, clinical trials). Results derived from our engine should be interpreted as possibly correct only in the case of broad agreement by independent scientists. Finally, while our methodology is relatively simple to follow (including initial population selection, deployment of the NAFLD identification algorithm, defining index dates, extracting covariates, defining observational windows, applying logistic regression, etc.), applying the same methodology to data sources other than Explorys may require extensive data science expertise; we believe, however, that our manuscript provides sufficient guidance on how to achieve that with a hope that the scientific community may also be capable of following our methodology and applying improvements to it.

Abbreviations

ALT: Alanine aminotransferase test

BMI: Body mass index

CCS: Clinical Classifications Software

CHF: Congestive heart failure

COPD: Chronic obstructive pulmonary disease

eGFR: Estimated glomerular filtration rate

EMR: Electronic medical record

HbA1c: Hemoglobin A1c

HIPAA: Health Insurance Portability and Accountability Act

HIV: Human immunodeficiency virus

ICD: International Classification of Diseases

INR: International normalized ratio

MI: Myocardial infarction

MVT: Motor vehicle traffic

NAFLD: Nonalcoholic fatty liver disease

NPV: Negative predictive value

OR: Odds ratio

PPV: Positive predictive value

SLE: Systemic lupus erythematosus

T2DM: Type 2 diabetes mellitus

UTI: Urinary tract infection

Declarations

Ethics approval and consent to participate: IBM's Data Access and Compliance Board (DACB) approved this study and all its methods, including the EMR cohort assembly, data extraction, and analyses. IBM

obtained all required consents in order to provide the research detailed in herein; IBM reviewed the manuscript for compliance with applicable export regulations and obtained all required consents.

Consent for publication: Data contain potentially identifying information and may not be shared publicly. Written informed consent to publish analyses that rely on the data was obtained from the study participants.

Availability of data and material: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request (Address: 75 Binney St, Cambridge, MA 02142, USA, Telephone: 857-500-2425; uri.kartoun@ibm.com).

Competing interests: Ping Zhang is a member of the editorial board (Associate Editor) of this journal. All other authors declare that they have no competing interests.

Funding: No funding was obtained for this study.

Authors' contributions: Study Conception and Design: UK, AP, YP, KC, KN. Acquisition of Data: UK, KN. Analysis and Interpretation of Data: UK, RA, AP, YP, PZ, HL, SD, KC, KN. Drafting of Manuscript: UK, RA, AP, YP, PZ, HL, SD, KC, KN. Critical Revision: UK, RA, AP, YP, PZ, HL, SD, KC, KN. All authors read and approved the final manuscript.

Acknowledgements: We would like to thank Jason Gilder (IBM Watson Health, Present Company: Syapse) and Anil Jain (IBM Watson Health) for providing valuable feedback toward forming this manuscript.

References

1. Musso G, Gambino R, Cassader M, Pagano G. Meta-analysis: natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of noninvasive tests for liver disease severity. Annals of Medicine 2011;43:617–49.
2. Byrne CD, Targher G. NAFLD: a multisystem disease. Journal of Hepatology 2015;62:S47–S64.
3. Charlton M. Cirrhosis and liver failure in nonalcoholic fatty liver disease: Molehill or mountain? Hepatology 2008;47:1431–3.
4. Simon TG, Kartoun U, Zheng H, Chan AT, Chung RT, Shaw S, Corey KE. MELD-Na score predicts incident major cardiovascular events, in patients with nonalcoholic fatty liver disease. Hepatol Commun 2017;1(5):429–38.
5. Kim D, Touros A, Kim WR. Nonalcoholic fatty liver disease and metabolic syndrome. Clin Liver Dis 2018;22(1):133–140.
6. The IBM Explorys Platform / Solution Brief. IBM Watson Health. IBM Corporation 2016.
7. Corey KE, Kartoun U, Zheng H, Shaw SY. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Digestive Diseases and Sciences 2016;61(3):913–9.

8. Corey KE, Kartoun U, Zheng H, Chung RT, Shaw SY. Using an electronic medical records database to identify nontraditional cardiovascular risk factors in nonalcoholic fatty liver disease. Am J Gastroenterol 2016;111(5):671–6.
9. Kartoun U. A glimpse of the difference between predictive modeling and classification modeling. J Clin Epidemiol 2019 May;109:142. doi: 10.1016/j.jclinepi.2019.01.001. Epub 2019 Jan 10.
10. Wei WQ, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, Cox NJ, Roden DM, Denny JC. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. PLoS One. 2017;12(7):e0175508. doi: 10.1371/journal.pone.0175508. eCollection 2017.
11. Feldman CH, Liu J, Feldman S, Solomon DH, Kim SC. Risk of high-grade cervical dysplasia and cervical cancer in women with systemic lupus erythematosus receiving immunosuppressive drugs. Lupus 2017;26(7):682–9.
12. Levey AS, Perrone RD, Madias NE. Serum creatinine and renal function. Annu Rev Med 1988;39:465–90.
13. Perrone RD, Madias NE, Levey AS. Serum creatinine as an index of renal function: new insights into old concepts. Clin Chem 1992;38(10):1933–53.
14. Marcuccilli M, Chonchol M. NAFLD and chronic kidney disease. Int J Mol Sci 2016;17(4):562.
15. Hazlehurst JM, Woods C, Marjot T, Cobbold JF, Tomlinson JW. Non-alcoholic fatty liver disease and diabetes. Metabolism 2016;65(8):1096–108.
16. Kwak MS, Yim JY, Kim D, Park MJ, Lim SH, Yang JI, Chung GE, Kim YS, Yang SY, Kim MN, Lee CH, Yoon JH, Lee HS. Nonalcoholic fatty liver disease is associated with coronary artery calcium score in diabetes patients with higher HbA1c. Diabetology & Metabolic Syndrome 2015;7:28.
17. Ohwaki K, Endo F, Muraishi O, Yano E. Relationship between changes in haemoglobin A1C and prostate-specific antigen in healthy men. Eur J Cancer 2011;47(2):262–6.
18. Libby G, Donnelly LA, Donnan PT, Alessi DR, Morris AD, Evans JM. New users of metformin are at low risk of incident cancer: a cohort study among people with type 2 diabetes. Diabetes Care 2009;32(9):1620–5.
19. Tande AJ, Platz EA, Folsom AR. The metabolic syndrome is associated with reduced risk of prostate cancer. Am J Epidemiol 2006;164(11):1094–102.
20. Hwang YC, Ahn HY, Park SW, Park CY. Nonalcoholic fatty liver disease associates with increased overall mortality and death from cancer, cardiovascular disease, and liver disease in women but not men. Clin Gastroenterol Hepatol 2018;16(7):1131–37.
21. Holzner B, Kemmler G, Greil R, Kopp M, Zeimet A, Raderer M, Hejna M, Zöchbauer S, Krajnik G, Huber H, Fleischhacker WW, Sperner-Unterweger B. The impact of hemoglobin levels on fatigue and quality of life in cancer patients. Ann Oncol 2002;13(6):965–73.
22. Wood RJ, Serfaty-Lacrosniere C. Gastric acidity, atrophic gastritis, and calcium absorption. Nutr Rev 1992;50(2):33–40.

23. Parvin M, Shakhssalim N, Basiri A, Miladipour AH, Golestan B, Mohammadi Torbati P, Azadvari M, Eftekhari S. The most important metabolic risk factors in recurrent urinary stone formers. Urol J 2011;8(2):99–106.
24. Bosch FX, Manos MM, Muñoz N, Sherman M, Jansen AM, Peto J, Schiffman MH, Moreno V, Kurman R, Shah KV. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. J Natl Cancer Inst. 1995;87(11):796–802.
25. Mackintosh AF, Fleming HA. Cardiac damage presenting late after road accidents. Thorax. 1981;36(11):811–3.
26. Campbell B. Varicose veins and their management. The BMJ 2006;333(7562):287–92.
27. Edwards CM, Zhuang J, Mundy GR. The pathogenesis of the bone disease of multiple myeloma. Bone 2008;42(6):1007–13.
28. Law SM, Lu X, Yu F, Tseng V, Law SK, Coleman AL. Cigarette smoking and glaucoma in the United States population. Eye (Lond) 2018;32(4):716–25.
29. Uzuner, O., Goldstein, I., Luo, Y., and Kohane, I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association 2008;15(1):14–24.
30. Kartoun U. Text nailing: an efficient human-in-the-loop text-processing method. ACM Interactions 2017;24(6):44–9.
31. Overman M, Wang C, Detrano R, Douglas-Escobar M, Ipp E, Hara B, Layos E, Swerdloff R, Berman N, Chlebowski R. Cardiovascular risk and sub-clinical atherosclerosis in prostate cancer: patient with and without androgen ablation. Journal of Clinical Oncology 2004, 22(14):4605.
32. Omalu BI, Hammers JL, Parwani AV, Balani J, Shakir A, Ness RB. Is there an association between coronary atherosclerosis and carcinoma of the prostate in men aged 50 years and older? An autopsy and coroner based post-mortem study. Niger J Clin Pract 2013;16(1):45–8.
33. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA. 2007 May 22;104(21):8685–90.
34. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science 2015;347(6224):1257601. doi: 10.1126/science.1257601.
35. Zhou X, Menche J, Barabási AL, Sharma A. Human symptoms-disease network. Nat Commun. 2014 Jun 26;5:4212. doi: 10.1038/ncomms5212.
36. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S, Szolovits P, Churchill S, Murphy S, Kohane I, Karlson EW, Plenge RM. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010;62(8):1120–7.
37. Kartoun U, Iglay K, Shankar RR, Beam A, Radican L, Chatterjee A, Pai JK, Shaw S. Factors associated with clinical inertia in type 2 diabetes mellitus patients treated with metformin monotherapy. Curr Med Res Opin 2019 Dec;35(12):2063-2070. doi: 10.1080/03007995.2019.1648116. Epub 2019 Sep 6.

38. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006; 101:1418–29.
39. Breiman L. Random forests. Machine Learning 2001;45(5–32).
40. Kartoun U, Corey KE, Simon TG, Zheng H, Aggarwal R, Ng K, Shaw SY. The MELD-Plus: A generalizable prediction risk score in cirrhosis. PLOS ONE 2017;12(10):e0186301. doi: 10.1371/journal.pone.0186301. eCollection 2017.
41. Kartoun U, Aggarwal R, Beam AL, Pai JK, Chatterjee AK, Fitzgerald TP, Kohane IS, Shaw SY. Development of an algorithm to identify patients with physician-documented insomnia. Sci Rep 2018;8(1):7862. doi: 10.1038/s41598-018-25312-z.
42. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, Petersen I. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol 2017;9:157– doi: 10.2147/CLEP.S129785. eCollection 2017.
43. Beaulieu-Jones BK, Moore J H. Missing data imputation in the electronic health record using deeply learned autoencoders. PacSymp Biocomput 2017;22:207–18.
44. Kartoun U. The trade-off of applying simple vs. advanced imputation techniques in prediction modeling. J Med Syst 2019;43(5):142. doi: 10.1007/s10916-019-1274-9.
45. Kartoun U, Kumar V, Cheng SC, Yu S, Liao K, Karlson E, Ananthakrishnan A, Xia Z, Gainer V, Cagan A, Savova G, Chen P, Murphy S, Churchill S, Kohane I, Szolovits P, Cai T, Shaw SY. Demonstrating the advantages of applying data mining techniques on time-dependent electronic medical records. Proc. of American Medical Informatics Association 2015 Annual Symposium. Nov. 2015, San Francisco, CA.
46. Ricciotti E, Haines PG, Beerens M, Kartoun U, Lahens NF, Wang T, Shaw SY, Macrae CA, Fitzgerald GA. Cyclooxygenase-2 and heart failure with preserved ejection fraction. American Heart Association Scientific Sessions, Philadelphia, Pennsylvania, 2019.

Figures

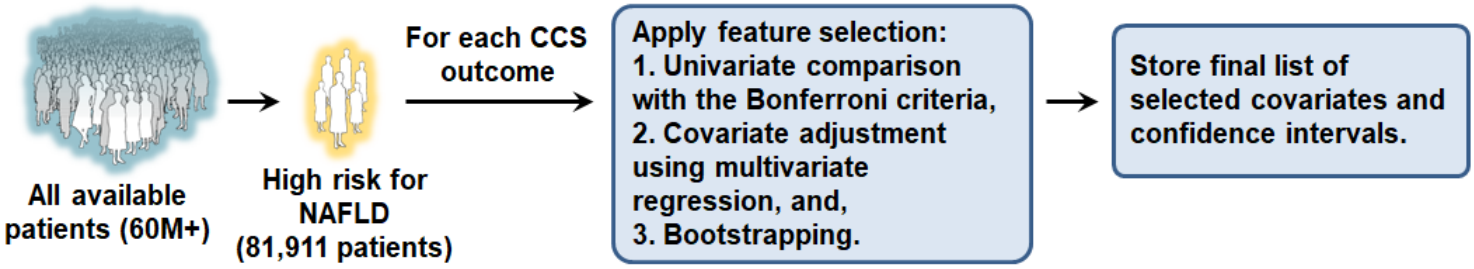

Figure 1

Methodology to identify associations between covariates and outcomes in NAFLD.

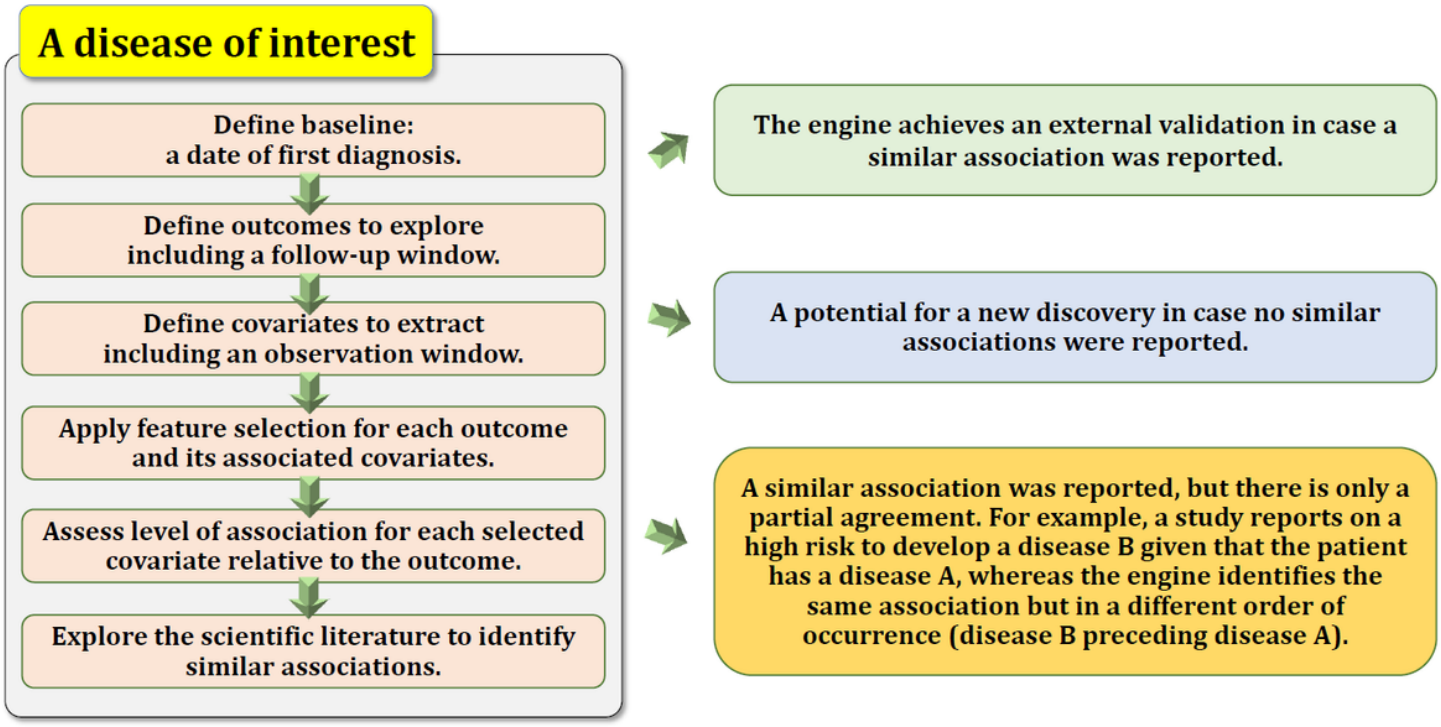

Figure 2

A general methodology to identify associations between covariates and outcomes using observational data.

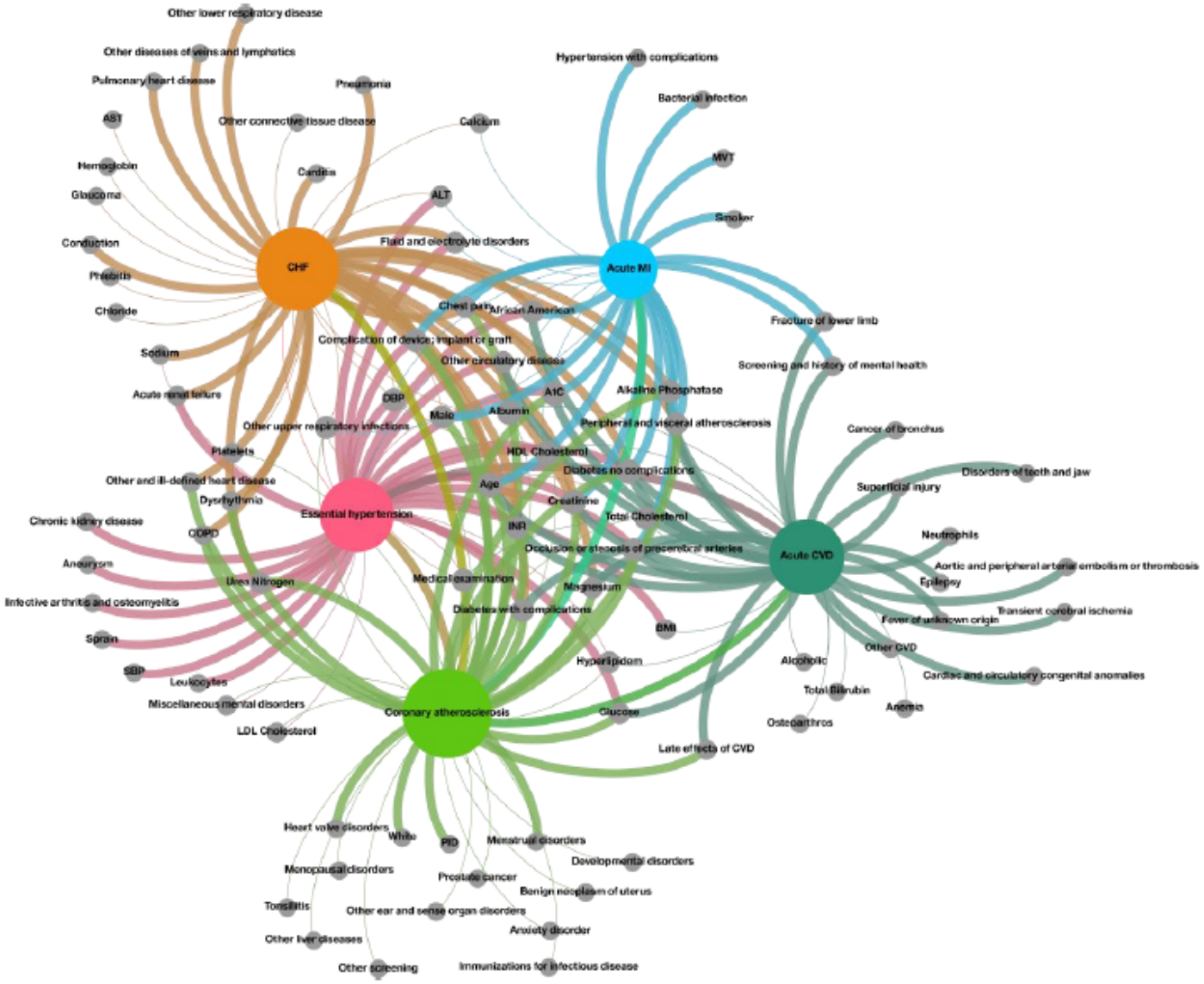

Figure 3

All statistically significant variables associated with diseases of the circulatory system developed within 5 years after the identification of high probability for NAFLD; because of space restrictions, we have chosen a limited subset of the outcomes for display.

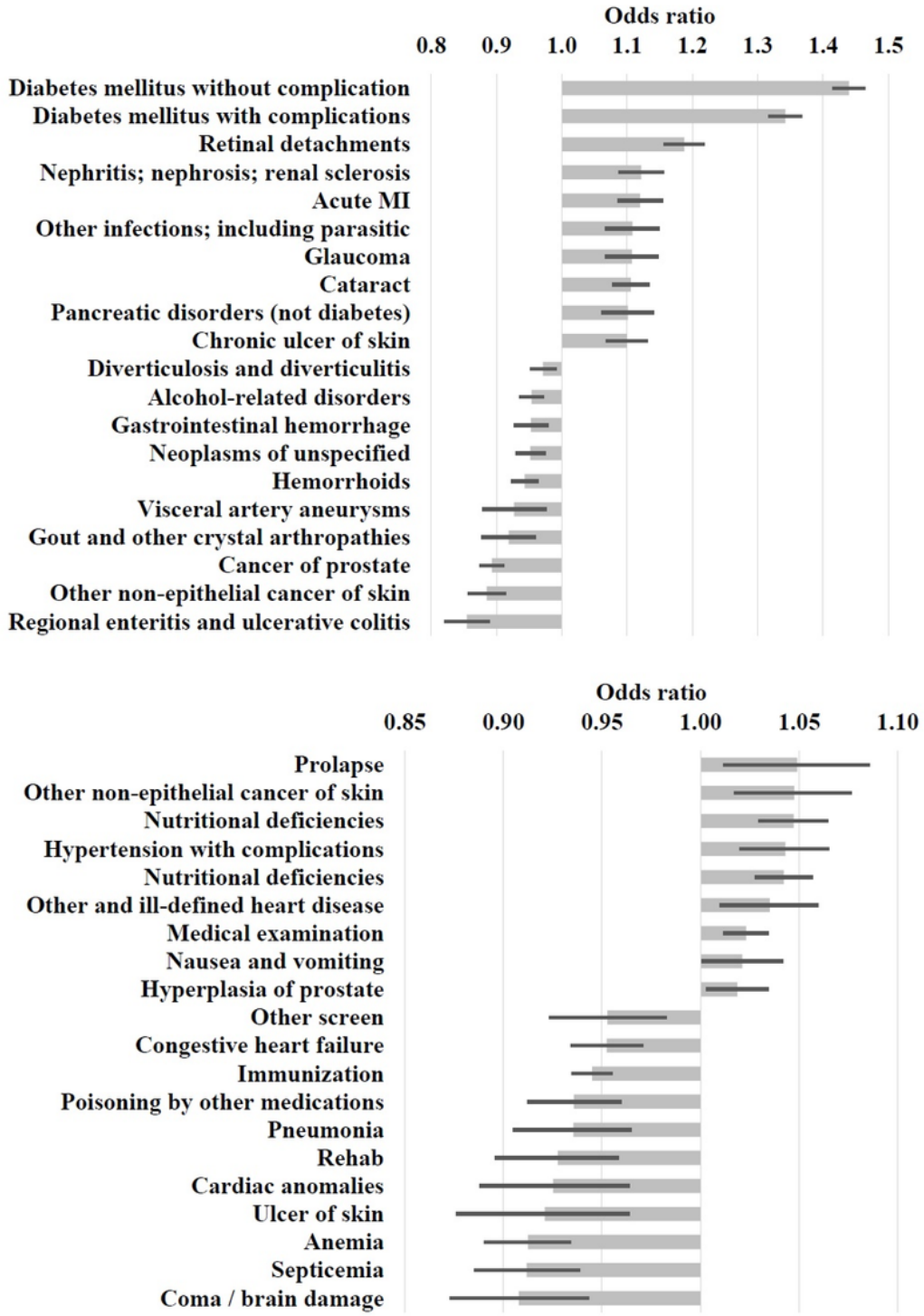

Figure 4

Statistically significant associations between lab covariates and outcomes developed within 5 years after the identification of high probability for NAFLD, presenting the top 10 and bottom 10 outcomes. (A) A1C as a covariate. (B) Hemoglobin as a covariate.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigs2.docx
- SupplementaryTable1.pdf
- SupplementaryFigs1.docx