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

Harvard Medical School

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: August 4th, 2019

DOI: <https://doi.org/10.21203/rs.2.11644/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

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 50 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 the 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 both individually or in combination is expected to yield a better understanding of the interplay between the development or the 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 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 linked to kidney malfunction [2,14]. It is also well known that NAFLD is associated with high levels of HbA1c (Hemoglobin A1c) and a high prevalence of T2DM [15].

Increased HbA1c and myocardial infarction (MI) are well-known associations [16]. A discovery engine can help studying the correlation of HbA1c with different types of cancers (e.g., prostate cancer [17]). Especially it could help 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 its ability to identify benefits in behavior that are well known as 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 within the context of 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 with NAFLD. Our discovery engine confirmed known associations and identified potentially 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 50 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 related to the liver) or at least one measurement of an elevated triglyceride lab test. Patients at high probability for NAFLD were identified using a classification algorithm that members of our group have previously validated [7–9].

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. Application of this algorithm to the EMR database yielded 334,258 patients with a high probability of underlying NAFLD. Our algorithm was able to identify the first date at which each patient developed a high probability for NAFLD (“index date”). We selected only patients 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. Patient characteristics for our cohort are presented in Table 1.

We extracted all disease outcomes during the 5-year follow-up and mapped them using the Clinical Classifications Software (CCS) categories based on the ICD 9th and 10th revisions [10]. Keeping only those that were prevalent in at least 1% of the NAFLD population yielded 173 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

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, and age), comorbidities, laboratory measurements (e.g., albumin, sodium, body mass index), and behavioral descriptors (e.g., smoking status, alcohol use). For laboratory variables, we used the most recent values found within the preceding 12 months of 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 173 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 prior to 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 and R.

We applied the process of NAFLD index date extraction and feature selection, as shown in Figure 1. We followed this methodology for each of the considered outcomes, resulting in 173 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 we compared the differences in the means of continuous variables using a t-test or Wilcoxon rank sum test, as appropriate. All statistical tests were 2-sided, with Bonferroni corrections for the 314 comparisons, and the adjusted P value threshold for statistical significance was 1.6×10^{-4} for each comparison.

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 double 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, and we excluded those with confidence intervals that crossed an OR of 1.

2.3. 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 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).

Results

3.1. Laboratory covariates

Many of the laboratory covariate 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 (Figure 2a). 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 (Figure 2b). 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. Various other correlations with lab values were found, but given the short space, we only present the outcomes associated with two of the most commonly measured labs: HbA1c and hemoglobin (Figure 2(a-b)).

3.2. 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, or 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. Osteoporosis was negatively correlated with being male and multiple myeloma, as expected, but interestingly it was also associated with sepsis and multiple sclerosis. Transient ischemic attacks were associated 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 to a wide range of mental conditions including mood disorders, dementia, and schizophrenia. In addition, systemic lupus erythematosus (SLE) has broadly been described as 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.

3.3. 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, 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.

3.4. 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 Figure 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]). One specific example is the alanine aminotransferase test (ALT) covariate (located near the top center of Figure 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 potentially 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 was 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 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 selected 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 linked 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 MI are well-known associations [16], and our engine was able to identify this association as well (Figure 2a). 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 for 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 the metabolic syndrome has already been reported [18–20], but further studies are required to more precisely assess such associations. 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 an 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 these is not well studied. Additional findings of our engine identified 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.

Although a variety of findings such as age and BMI were associated with the development of varicose veins [26], a surprising covariate was a preceding 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 this feature is only experienced by those with NAFLD and Parkinson's. It may also be possible to have a common genetic underpinning for comorbid conditions. Our engine can generate such hypotheses relating covariates with 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 diastolic blood pressure and albumin contribute to the increased confidence of the accuracy of our engine.

Although our engine identified that smoking was associated with the development of a variety of diseases, it was protective against glaucoma and cataracts. These findings are surprising, and at least one study reported similar results within the context of glaucoma and the potential protective effect of nicotine [28]. Determining 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 a primary source to document patient smoking status holding additional challenges regarding the accuracy of extracting the statuses correctly [29,30].

In the univariate analysis, we found that osteoporosis was positively correlated with the development of a 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 interpretation of the results from our engine depends on the clinical context and thus should be made 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 be more impactful in the NAFLD population.

Conclusion

Our discovery engine has several benefits for future research studies. First, it is easy to rapidly implement such a methodology on EMRs. 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 50 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, this study could be used to explore whether having NAFLD changes the trajectory of diabetes compared to diabetic patients without NAFLD. Such findings could be very valuable as medicine moves away from a “one-size-fits-all” model to one of increased precision and personalization.

Our study has several limitations. First, it is a retrospective analysis that has the potential for confounding effects. Even though 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 counter-intuitive associations. It could be either 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 data-driven multiple testing approach.

We 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 be a potentially useful tool for the efficient screening of clinical hypotheses.

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

INR: International normalized ratio

MI: Myocardial infarction

MVT: Motor vehicle traffic

NAFLD: Nonalcoholic fatty liver disease

OR: Odds ratio

T2DM: Type 2 diabetes mellitus

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 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. pii: S0895-4356(18)30961-2. doi: 10.1016/j.jclinepi.2019.01.001. [Epub ahead of print] [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.

Figures

Figure 1

Methodology to identify associations between covariates and outcomes in NAFLD.

Figure 2

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.

Figure 3

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

Supplementary Files

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

- [supplement1.pdf](#)
- [supplement2.pdf](#)
- [supplement3.pdf](#)