Large-scale identification of undiagnosed hepatic steatosis using natural language processing

Carolin V. Schneider,[a,b,*,i] Tang Li,[c,i] David Zhang,[a] Anya I. Mezina,[d] Puru Rattan,[d] Helen Huang,[a] Kate Townsend Creasy,[a] Eleonora Scorletti,[a] Inuk Zandvakili,[a,d,e] Marijana Vujkovic,[a,c,f] Leonida Hehl,[b] Jacob Fiksel,[c] Joseph Park,[a] Kirk Wangensteen,[g] Marjorie Risman,[a] Kyong-Mi Chang,[d,f] Marina Serper,[d,f] Rotonya M. Carr,[h] Kai Markus Schneider,[b] Jinbo Chen,[c,j] and Daniel J. Rader[a,j]

[a]Division of Translational Medicine and Human Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[b]Department of Medicine III, RWTH Aachen University, Aachen, Germany
[c]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[d]Division of Gastroenterology and Hepatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[e]Division of Digestive Diseases, Department of Internal Medicine, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA
[f]Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA 19104, USA
[g]Department of Medicine, Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN 55902, USA
[h]Department of Medicine, Division of Gastroenterology, University of Washington, Seattle, WA 98195, USA

eClinicalMedicine
2023;62: 102149
Published Online xxx
https://doi.org/10.1016/j.eclinm.2023.102149

Summary

Background Nonalcoholic fatty liver disease (NAFLD) is a major cause of liver-related morbidity in people with and without diabetes, but it is underdiagnosed, posing challenges for research and clinical management. Here, we determine if natural language processing (NLP) of data in the electronic health record (EHR) could identify undiagnosed patients with hepatic steatosis based on pathology and radiology reports.

Methods A rule-based NLP algorithm was built using a Linguamatics literature text mining tool to search 2.15 million pathology report and 2.7 million imaging reports in the Penn Medicine EHR from November 2014, through December 2020, for evidence of hepatic steatosis. For quality control, two independent physicians manually reviewed randomly chosen biopsy and imaging reports (n = 353, PPV 99.7%).

Findings After exclusion of individuals with other causes of hepatic steatosis, 3007 patients with biopsy-proven NAFLD and 42,083 patients with imaging-proven NAFLD were identified. Interestingly, elevated ALT was not a sensitive predictor of the presence of steatosis, and only half of the biopsied patients with steatosis ever received an ICD diagnosis code for the presence of NAFLD/NASH. There was a robust association for *PNPLA3* and *TM6SF2* risk alleles and steatosis identified by NLP. We identified 234 disorders that were significantly over- or underrepresented in all subjects with steatosis and identified changes in serum markers (e.g., GGT) associated with presence of steatosis.

Interpretation This study demonstrates clear feasibility of NLP-based approaches to identify patients whose steatosis was indicated in imaging and pathology reports within a large healthcare system and uncovers undercoding of NAFLD in the general population. Identification of patients at risk could link them to improved care and outcomes.

Funding The study was funded by US and German funding sources that did provide financial support only and had no influence or control over the research process.

Keywords: Liver disease; NAFLD; Biopsy; EHR; Natural language processing

*Corresponding author. RWTH Aachen University, Pauwelsstr.30, Aachen 52074, Germany.
E-mail address: cscheider@ukaachen.de (C.V. Schneider).

[i]Authors share the first authorship.

[j]Authors share the last authorship.

Research in context

Evidence before this study

Non-alcoholic fatty liver disease (NAFLD) is a major cause of fibrotic liver disease and cirrhosis, but a large proportion of NAFLD cases are undiagnosed. Can Natural Language Processing (NLP) be used to identify undiagnosed patients with hepatic steatosis in electronic health records (EHR) by applying it to imaging and liver biopsy reports?

Added value of this study

Using an NLP approach to liver biopsy pathology reports and imaging studies, this study identified 3007 patients with biopsy-proven steatosis and 42,083 with steatosis present on imaging from the EHR, of whom only ~35% had a diagnosis code in their chart for NAFLD or related conditions. PheWAS and LabWAS analyses found 234 phenotypic traits significantly over- or underrepresented in subjects with steatosis and nine serum markers associated with presence of steatosis. Genetic association analysis revealed robust associations with genetic variants known to be associated with steatosis.

Implications of all the available evidence

This study demonstrates the feasibility of NLP at a population-based scale in identifying undiagnosed patients with steatosis.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is a silent epidemic that is estimated to affect nearly two billion people worldwide[1] and is a major cause of fibrotic liver disease and cirrhosis.[2] Hepatic steatosis, the first stage in NAFLD, is defined as hepatic lipid droplet accumulation in more than 5% of the hepatocytes in the absence of other known inducers of steatosis (such as chronic viral hepatitis and chronic alcohol abuse).[3] Obesity is the most important risk factor for the development of NAFLD, but genetics also play a major role.[4] Identifying NAFLD has clinical value because it is a prevalent condition that can lead to serious liver problems if left untreated. Early identification can lead to important lifestyle changes and clinical follow up that may prevent progression.[5]

While hepatic steatosis is common, estimated to be present in approximately 25% of the US population,[6] less than a quarter of these individuals progress to non-alcoholic steatohepatitis (NASH),[7] the inflammatory, progressive form of NAFLD,[8] and even fewer to liver fibrosis, cirrhosis and hepatocellular carcinoma.[9] Elevated blood alanine aminotransferase (ALT) is often used as a proxy for NAFLD but normal levels of ALT do not exclude the presence of steatosis or even NASH.[4,10] Imaging modalities can identify hepatic steatosis, often as an incidental finding. The gold standard for NAFLD and NASH diagnosis is liver biopsy.[11] Overall, NAFLD is underdiagnosed and there are no systematic screening protocols.[12–14] Despite the high prevalence of hepatic steatosis and NAFLD, research on their natural history using electronic health record (EHR) data has been limited due to the requirement of long-term follow up as well as the underdiagnosis of these conditions.

EHR systems include structured data (such as ICD codes) and unstructured data (clinical documentation such as biopsy reports, and imaging reports). EHRs are emerging as a powerful data source for clinical and translational research studies.[15] However, phenotypes of interest need to be accurately defined.[16] In the last few years, artificial intelligence has shown increasing promise when applied to the prediction of medical outcomes.[17,18] Natural language processing (NLP) is a computational method that can be used to break down sentences and apply linguistic rules to interpret the meaning of a sentence fragment. The large-scale identification of patients with NAFLD by applying NLP to unstructured text in the EHR is one potential method to address the gap between using the EHR for patient care and leveraging it for translational research.

In this study, we applied NLP to unstructured data in a large academic health system EHR to identify patients with hepatic steatosis and NAFLD that could not be identified by a search of structured diagnosis codes. We used NLP to search 2.15 million pathology reports and 2.67 million radiology reports and identified a large number of patients with biopsy-proven NAFLD/NASH or imaging-identified steatosis who did not have ICD codes for these conditions in their EHR reports. We performed a number of analyses that demonstrated excellent accuracy of the NLP approach. Our study demonstrates that the rapid and accurate identification of steatosis by NLP from unstructured text is a potential method to bridge the gap between this incidental finding with high prevalence in the general population and the low detection of NAFLD using ICD-9/10 codes in EHR systems.

Methods

Study setting, population and data processing

The study was conducted at the University of Pennsylvania and used the data resources of the Penn Medicine EHR and the Penn Medicine Biobank. Protocols for this study were approved by the Institutional Review Board (IRB) at Penn Medicine (#813913). The requirement for explicit informed consent was waived due to the retrospective nature and the use of de-identified data. Throughout the study, all data were anonymized to ensure confidentiality.

We extracted 2.15 million pathology reports and 2.67 million imaging reports from Penn Medicine EHRs starting November 27, 2014, through December 31, 2020. Of those images 250,329 liver-related key-words in 200,132 images that included the liver were found. We applied NLP (described below) to create mutually exclusive datasets of imaging-identified steatosis cases, biopsy-proven steatosis/NAFLD cases, and PMBB controls. If multiple biopsies/images were available on one patient, we chose the most recent biopsy or image for analysis. If a patient had both a biopsy and a liver image, they were only included in the biopsy group.

The Penn Medicine Biobank (PMBB) is a resource that provides controlled access to genetic information, clinical data and biospecimens from consented participants from the Penn Medicine Health System. We identified a group of control patients who are free of NAFLD from PMBB as patients without imaging-identified steatosis or biopsy-proven steatosis and all liver diseases (ICD10: B18, K7 and F10; ICD9: 155.0, 197.7, 570–573). In total 21,195 controls were identified.

Patients with other causes and overlapping causes of hepatic steatosis were excluded from our study. Specifically, any patient with one of the following diagnoses (on the biopsy/imaging report or as ICD code), was excluded: viral hepatitis, hemochromatosis, primary biliary cholangitis, autoimmune hepatitis, secondary biliary cirrhosis, Wilson's disease, alcohol use disorder, alcoholic liver disease, alcoholic hepatitis and/or ascites, alcoholic fibrosis and sclerosis of liver, alcoholic cirrhosis of liver and/or ascites, alcoholic hepatic failure and/or coma, and unspecified alcoholic liver disease, biopsies/imaging after liver transplantation. Moreover, we excluded patients with age <18 at the time of biopsy or imaging and those with BMI <12 or >200 kg/m^2 since these BMI values were likely typographical errors. A total of 15,667 (31%) unique imaging cases as well as 3550 (54%) unique biopsy cases were excluded.

For our identified NAFLD cohort, we extracted their longitudinal data from Penn Medicine EHRs, including 23.7 million ICD-10 codes, 3.1 million laboratory values and more than 1.2 million BMI measurements. BMI and serum values of available biomarkers strictly within 183 days prior to or after the biopsy were evaluated. We excluded lab observations which were negative, or >10xULN or <0.1xLLN.

Natural language processing using Linguamatics I2E

We used Linguamatics, an NLP software, to identify patients with biopsy-proven steatosis as well as imaging reports mentioning hepatic steatosis. The query was built using Linguamatics' I2E, a literature text mining tool based on natural language processing and linguistic analytics[19] applies logic concepts to build a rules-based NLP algorithm. The Linguamatics algorithm is particularly good at detecting negations, which is a common problem when using NLP. I2E uses linguistic patterns, grammatical rules, and pre-built domain-specific ontologies to identify and extract entities, relationships, and associations within the text.

The algorithm was developed as follows:

1. Query formulation: The I2E platform offers an intuitive, visual query-building interface, making it easy for users to design complex queries. The tool was adapted to our specific problem by creating a custom query to identify key terms related to NAFLD diagnosis and steatosis severity. This query was formulated based on expert knowledge, literature review, and a list of predefined terms relevant to liver disease. In summary, 12 keywords/phrases were used to identify biopsy-proven NAFLD cases, five keywords for NASH cases, and ten keywords/phrases for imaging-identified steatosis cases (Supplementary eTables S1–S3). The NLP tool used several key terms considered diagnostic for NAFLD/NASH, including "hepatic steatosis", "fatty liver" and "non-alcoholic steatohepatitis".
2. Text processing: I2E processes the input text data by tokenizing it and parsing the sentences to determine their grammatical structure.
3. Information extraction: Using the formulated queries, I2E searches the processed text and extracts the relevant information. This process allows the platform to transform unstructured data into structured, actionable insights.
4. Manual check for a subset (n = 200) and algorithm re-evaluation
5. Multiple rounds of 1–4 until the final algorithm was developed, when no patient in the subset (n = 200) was falsely identified

The sample size in this study was determined based on the availability of data within the Penn Medicine EHR system, as it serves as a representative example of a tertiary, academic healthcare system.

Quality control: manual validation by two qualified physicians

After manual review (described in results), we scored the presence or absence of ICD-9/10 codes for NAFLD/NASH (571.8/9, K76.0, K75.8) in patients found to have steatosis by biopsy or imaging, as well as the elevation of ALT (>35 U/L for males, >25 U/L for females, if sex was not available/unclear we used >35U/L, 0.63% of steatosis on biopsy cohort have unclear sex, 0.09% of steatosis in imaging cohort have unclear sex, no unclear sex exist in controls).[20]

Severity of steatosis

Hepatic steatosis is defined by the presence of greater than 5% of lipid droplets in the liver, and NAFLD is defined by hepatic steatosis in the absence of other acute or chronic liver disease and the absence of excessive alcohol use in the two years prior to diagnosis.[20] The NAFLD Activity (NAS) score was rarely used in the pathology reports in our EHR, but the use increased with time. NASH is defined as NAFLD with the presence of predominantly macrovesicular steatosis along with hepatocyte ballooning and inflammation (equaling NAS ≥ 5, Supplementary eTable S4).

Borderline NASH is defined as steatosis with fibrosis or inflammation equaling NAS < 5, therefore not fully reaching the cutoff for NASH (e.g., presence of only fibrosis or only ballooning). Therefore, biopsy-proven steatosis included all cases of biopsy-proven NAFLD, borderline NASH and biopsy-proven NASH. To determine the severity of fibrosis in our patient cohort, we employed two widely-accepted and non-invasive fibrosis scores: the Fibrosis-4 (FIB-4) index and the Aspartate Aminotransferase-to-Platelet Ratio Index (APRI). The FIB-4 index incorporates patient age, AST levels, alanine aminotransferase (ALT) levels, and platelet counts into its calculation, while the APRI is calculated using AST levels and platelet counts.[21,22]

Statistical analysis

After normality testing by Kolmogorov–Smirnov test, we summarized continuous baseline variables by median and interquartile range (IQR), and applied Mann–Whitney U test for between-group comparison. All categorical variables were displayed as counts and relative frequencies (%), and the Chi-square or Fisher's exact test was performed for between-group comparison depending on whether all cell counts were above 5. We additionally reported s-values,[23–25] which convert p-values into a more interpretable measure of evidence against the null hypothesis. The analyses were performed using R software version 4.0.2 (R Foundation for Statistical Computing; Vienna, Austria, all Tables) and SPSS Statistics version 26 (IBM; Armonk, NY, USA). All tests were two-sided at significance level 0.05. Fig. 1A and the graphical abstract was created with biorender.com.

Genetic analysis

To validate the imaging-identified steatosis cohort, we leveraged available genomic data in PMBB. Among the 44,076 participants in PMBB with available genomic data, there were 2840 cases of imaging-identified steatosis cohort (6.75% of all imaging steatosis cases) (Supplementary eTable S5). For each PMBB participant, whole exome sequences (WES) were generated by the Regeneron Genetics Center from DNA extracted from stored buffy coats. These sequences were mapped to GRCh37. We removed samples with low exome sequencing coverage (less than 75% of targeted bases achieving 20 × coverage), high missingness (greater than 5% of targeted bases), high heterozygosity, dissimilar reported and genetically determined sex, genetic evidence of sample duplication, and cryptic relatedness (closer than third-degree relatives). We performed ancestry-specific NAFLD analyses in the PMBB using WES data that had been filtered using a series of quality control filters known as Goldilocks filter.[26] First, we used a causal directed acyclic graph to identify a minimally sufficient set of confounders and evaluated the linearity of the confounders for usage in our logistic regression models (Supplementary eFigs. S2 and S3). The variance explained by the different PCs can be found in Supplementary eFig. S4. We tested all single nucleotide variants in the exome for association with NAFLD through logistic regression assuming an additive model and excluded synonymous variants and variants with less than ten African and/or European ancestry carriers. Covariates in the regression model were age, sex, BMI, diabetes and the first five principal components of genetic ancestry for African-specific and the first ten principal components for European-specific analyses. We aggregated summary statistics from the African and European ancestry-specific analyses and performed a multi-ancestry meta-analysis using the inverse variance method for pooling. The R package "meta" was used to perform all fixed effects meta-analyses. Variants were considered exome-wide significant if they passed our FDR-adjusted p-value threshold ($p < 1.53 \times 10^{-6}$). All association analyses were performed using plink 2.0.

PheWAS

To validate the biopsy- and imaging-identified steatosis cases we performed a phenome-wide association study (PheWAS). The coding for clinical diagnoses in our data set followed the WHO's International Classification of Diseases (ICD) coding systems. All of the 21,195 PMBB control patients who had exome-wide sequencing data available, all of the 3007 biopsy steatosis patients, and 99.7% (41,972) imaging steatosis patients had at least one ICD-10 code available. For each patient, ICD-9/10 codes from the EHR diagnoses throughout the study period were collated with duplicates removed. ICD-9 codes were translated into ICD-10 codes. We converted the ICD-10 codes to 1847 associated Phe-codes using the R software PheWAS package, a method to bin similar ICD-codes into single coherent phenotypes. Only phenotypes with at least ten cases per group were evaluated. Only Phecodes that were diagnosed prior to the date of imaging/biopsy/study inclusion were analyzed. R package "PheWAS" was used for the analyses.

Role of funding

The research described in this study received support from various funding sources, which played a crucial role in enabling the execution and completion of the research. The funding sources mentioned at the end of

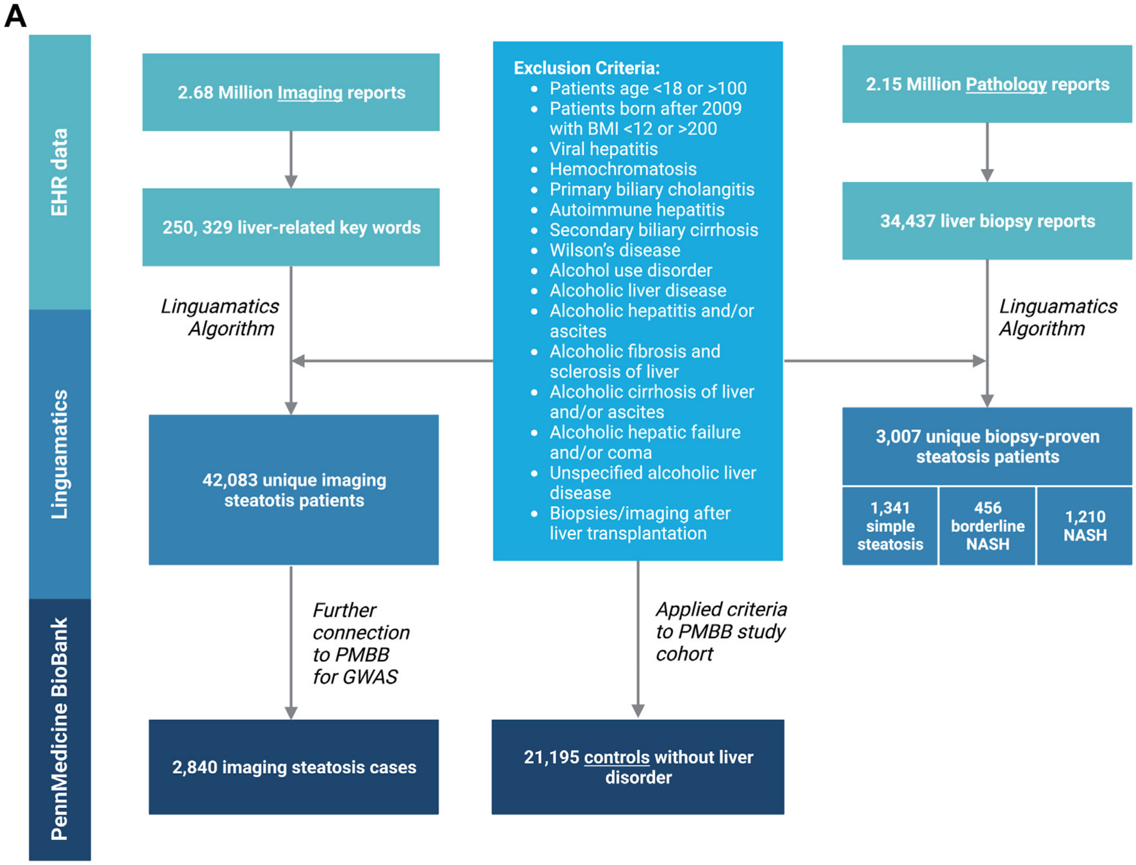

A
EHR data
Linguamatics
PennMedicine BioBank
2.68 Million Imaging reports
250, 329 liver-related key words
Linguamatics Algorithm
42,083 unique imaging steatotis patients
Further connection to PMBB for GWAS
2,840 imaging steatosis cases
Exclusion Criteria:
Patients age <18 or >100
Patients born after 2009 with BMI <12 or >200
Viral hepatitis
Hemochromatosis
Primary biliary cholangitis
Autoimmune hepatitis
Secondary biliary cirrhosis
Wilson's disease
Alcohol use disorder
Alcoholic liver disease
Alcoholic hepatitis and/or ascites
Alcoholic fibrosis and sclerosis of liver
Alcoholic cirrhosis of liver and/or ascites
Alcoholic hepatic failure and/or coma
Unspecified alcoholic liver disease
Biopsies/imaging after liver transplantation
Applied criteria to PMBB study cohort
21,195 controls without liver disorder
2.15 Million Pathology reports
34,437 liver biopsy reports
Linguamatics Algorithm
3,007 unique biopsy-proven steatosis patients
1,341 simple steatosis
456 borderline NASH
1,210 NASH

Fig. 1: (A) Flowchart of the selection process of steatotic individuals using NLP and controls. We applied an NLP algorithm to 2.67 million imaging reports and, after applying several exclusions, we identified 42,083 discrete patients in whom the presence of hepatic steatosis was specifically reported by the radiologist. We also searched 2.15 million pathology reports to find 34,437 liver biopsy reports. After applying exclusions for other known causes of steatosis and liver disease, we identified a total of 3007 discrete patients with biopsy-proven NAFLD. Among these patients, 1210 patients met criteria for NASH, 456 patients had borderline NASH, and 1341 patients had steatosis (B) Genetic analysis of steatosis on imaging patients compared to controls. Manhattan plot of genome-wide markers for imaging-identified steatosis (2840 cases and 21,195 controls). Logistic regression analysis performed ancestry specific EWASs in the PMBB using WES data that had been filtered using a series of quality control filters known as the Goldilocks filter assuming an additive genetic model, adjusted for age, sex, BMI, and genetic ancestry. Results are plotted as $-\log_{10}$ p values on the y-axis by position in chromosome (x-axis) (NCBI build 37).

the manuscript had no involvement in the study design, the collection, analysis, and interpretation of data, the writing of the report, or in the decision to submit the paper for publication. The authors declare that the funding agencies provided financial support only and had no influence or control over the research process or the content of the manuscript.

Results

Patients with imaging-identified steatosis and biopsy-proven NAFLD/NASH identified using NLP

We applied an NLP algorithm to 2.67 million imaging reports involving 1.51 million patients and, after applying several exclusions (see Methods). The different imaging modalities evaluated are highlighted in Supplementary eTable S6. Prior to applying exclusion criteria, we identified 6557 (22% of all evaluated distinct cases) biopsy-proven steatosis cases and 50,104 (3% of all evaluated images, 25% of all distinct cases where images were showing the liver) imaging-identified steatosis cases. Finally, we identified 42,083 discrete patients in whom the presence of hepatic steatosis was specifically reported by the radiologist (Fig. 1). Only 34% of these individuals had an ICD code for NAFLD/NASH in the EHR (Table 1). We also searched 2.15 million pathology reports to find 34,437 liver biopsy reports involving 30,215 patients and applied a similar NLP algorithm (Fig. 1). We ranked the top features that contributed to the identification of cases by the number of times they were found. The most influential features, in order of importance, were among both biopsy and imaging cases "(hepatic) steatosis" as well as "steatohepatitis". After applying exclusions for other known

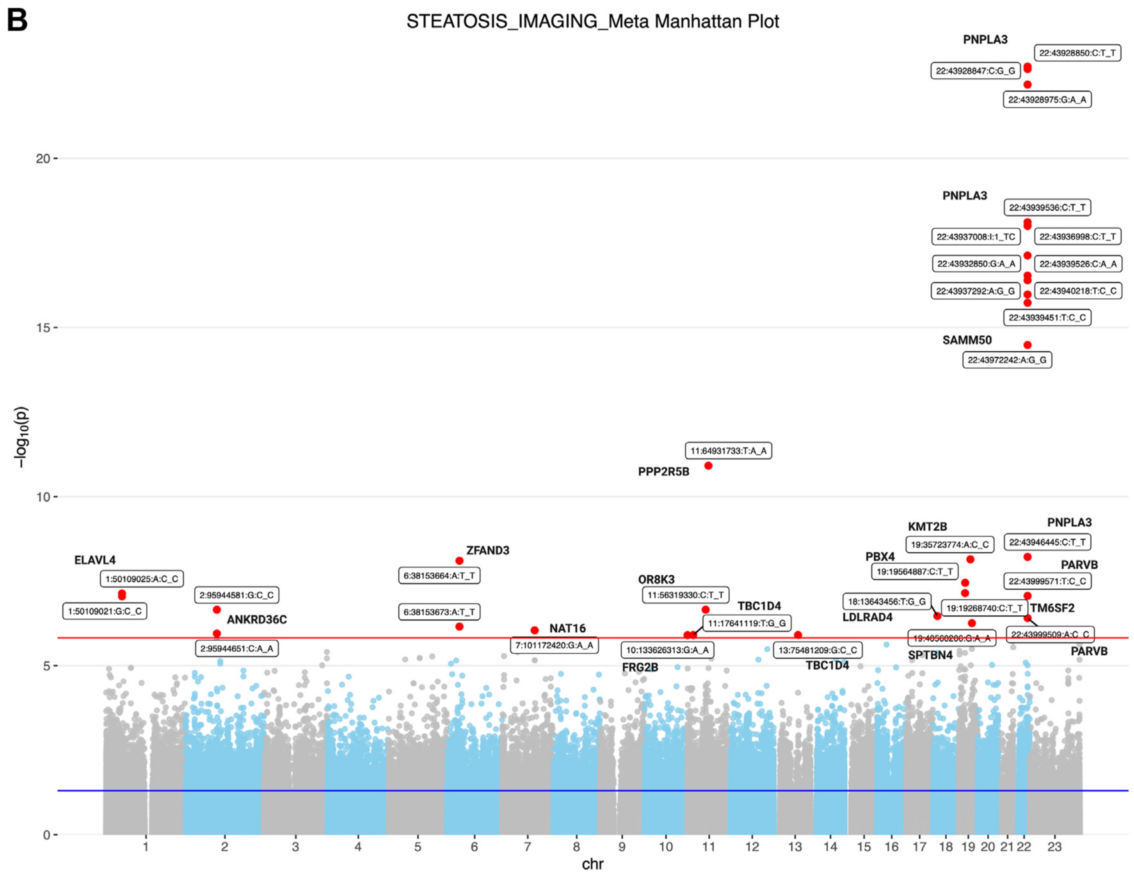


Fig. 1: Continued.

causes of steatosis and liver disease (see Methods), we identified a total of 3007 discrete patients with biopsy-proven NAFLD. Among these patients, 1210 patients met criteria for NASH, 456 patients had borderline NASH, and 1341 patients had steatosis (Fig. 1). Remarkably, only 53% of these individuals had an ICD code for NAFLD or NASH in the EHR. We generated a control group of 21,195 patients drawn from the same EHR who had no evidence of steatosis on imaging and no liver-related ICD codes (Fig. 1).

Quality control

Linguamatics can export the sentence where the searched phrases are located. One physician screened these sentences for most identified patients. For quality control, we randomly selected 100 biopsy reports demonstrating steatosis, 80 imaging reports evidencing steatosis, and 20 controls (patients devoid of steatosis), composed of ten patients with imaging and ten with biopsy. These biopsy and imaging reports were evaluated by another pair of physicians. The manual review confirmed that the NLP algorithm adeptly identified every participant, accurately classifying them according to the findings of this manual review. To augment the precision in pinpointing potential misclassifications, we randomly selected patients from a subset of our data set that had the greatest likelihood of being misclassified—those devoid of a NAFLD/NASH ICD10 diagnosis and with no history of elevated ALT. This selected sub-cohort comprised 46 additional patients with biopsy-validated steatosis, ten randomly chosen patients with biopsy-confirmed absence of steatosis, 79 patients with imaging-validated steatosis, and 18 patients with imaging-confirmed nonexistence of steatosis. Two additional physicians conducted a comprehensive chart review for these patients. Once again, the manual review identified all participants to be correctly identified into their subsequent group by our algorithm. Only in a singular case, a recent imaging report was copied and therefore duplicated within the biopsy report's results. While the term 'steatosis' was correctly flagged by the algorithm as present in the results section, it was missing from the biopsy report. However, an in-depth chart review confirmed this patient had clinical evidence of NAFLD, validating the accuracy of the algorithm. Still, in total among the 353 evaluated biopsy reports sensitivity was 100%, and specificity 98%.

	NAFLD on biopsy (n = 3007)	Steatosis in Imaging (n = 42,083)	p[c] steatosis biopsy vs. steatosis imaging	s[d] steatosis biopsy vs. steatosis imaging	Controls (n = 21,195)	p steatosis biopsy vs. controls	s steatosis biopsy vs. controls	p steatosis imaging vs. controls	s steatosis imaging vs. controls
Age at image/biopsy/inclusion (years), median (IQR)	55 (43–64)	57 (45–67)	<0.001	40.4	57 (41–68)	<0.001	19.4	<0.001	14.3
Male, n (%)	1366 (45.7)	19,570 (46.5)	0.39	1.4	10,375 (49)	<0.001	10.0	<0.001	26.4
White, n (%)	2080 (74.1)	29,694 (70.6)	0.0001	13.1	14,296 (69.9)	<0.001	17.6	0.04	4.5
Hispanic Latino, n (%)	105 (3.7)	2399 (5.7)	<0.001	16.2	511 (2.4)	<0.001	14.5	<0.001	254.1
Black, n (%)	379 (13.5)	7952 (18.9)	<0.001	39.8	5020 (24.5)	<0.001	125.3	<0.001	194.2
BMI (kg/m2), median (IQR)	30 (26–37)	32 (27–37)	<0.001	40.5	28 (25–33)	<0.001	144.9	0	Inf
Diabetes (ICD), n (%)	893 (29.7)	12,969 (30.8)	0.21	2.3	4209 (20)	<0.001	114.1	<0.001	622.3
Serum markers									
ALT (U/L), median (IQR)	35 (20–75)	24 (16–40)	<0.001	295.8	18 (13–25)	0	Inf	0	Inf
ALT > ULN, n (% of patients with elevated ALT)	1219 (57.5)	11,151 (37.92)	<0.001	233.3	1666 (14.39)	0	Inf	0	Inf
AST (U/L), median (IQR)	32 (21–62)	22 (17–32)	<0.001	502	19 (16–24)	0	Inf	<0.001	705.9
GGT (U/L), median (IQR)	78 (41–163)	50 (25.9–128)	<0.001	21.4	24 (15–49)	<0.001	69.6	<0.001	47.5
ALP (U/L), median (IQR)	82 (64–117.5)	73 (59–93)	<0.001	149.9	65 (52–81)	<0.001	524.6	<0.001	759.5
Albumin (g/dl), median (IQR)	4.0 (3.6–4.3)	4.2 (3.8–4.4)	<0.001	67.7	4.1 (3.9–4.4)	<0.001	79.2	0.586	0.8
Triglycerides (mg/dl), median (IQR)	126 (88–178)	133 (94–191)	<0.001	12.0	100 (71–145)	<0.001	92.1	<0.001	906.8
Cholesterol (mg/dl), median (IQR)	172 (144.3–205.8)	175 (146–205)	0.488	1.0	174 (146–204)	0.977	0.03	0.089	3.5
LDL (mg/dl), median (IQR)	97.2 (76–123)	96 (71–123)	0.099	3.3	98 (75–123)	0.866	0.2	<0.001	11.6
HDL (mg/dl), median (IQR)	44 (36–53)	45 (37–54)	0.005	7.6	49 (40–60)	<0.001	81.2	<0.001	295.4
HbA1C (%), median (IQR)	6.0 (5.5–6.8)	6.2 (5.7-7.2)	<0.001	28.6	5.8 (5.4–6.7)	0.005	7.8	<0.001	108.1
WBC (× 10^9/L), median (IQR)	7.4 (5.5–9.9)	7.5 (5.9-9.9)	<0.001	12.2	6.9 (5.6–8.7)	<0.001	26.1	<0.001	459.9
FIB-4, median (IQR)	1.45 (0.86–2.76)	1.09 (0.72–1.7)	<0.001	199.7					
FIB-4, n (%)									
F0 (<1.3)	872 (43.86)	17,021 (60.91)	<0.001	313.5					
F2/3 (1.3–2.67)	593 (29.83)	7771 (27.81)							
Probable F4 (>2.67)	523 (26.31)	3151 (11.28)							
APRI, median (IQR)	0.40 (0.23–1.01)	0.24 (0.17–0.39)	<0.001	506.2					
ICD									
NAFLD ICD9/10 code (%)	1603 (53.3)	14,035 (33.4)	<0.001	360.1	0 (0)	0	Inf	0	Inf
NASH ICD10 code[a] (%)	684 (22.7)	1831 (4.4)	0	Inf	0 (0)	0	Inf	<0.001	689.2
NAFLD or NASH ICD9/10 code[b] (%)	1710 (56.9)	14,490 (34.4)	<0.001	446.9	0 (0)	0	Inf	0	Inf

BMI, body mass index; ALT, alanine transaminase; AST, aspartate transaminase; AP, alkaline phosphatase; GGT, gamma-glutamyl transferase; HbA1c, hemoglobin A1c, LDL, low-density lipoprotein; HDL, high-density lipoprotein; INR, international rationalized ratio; WBC, white blood cell; NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis. [a]K57.8. [b]K67.0 and 571.8/9. [c]p; p-value obtained from Mann–Whitney U test for continuous variables and Pearson's chi-squared test for categorical variables. All continuous variables were non-normally distributed. [d]s; s-value obtained from formula: s-value = $-\log_2$ (p-value).

Table 1: **Baseline characteristics of biopsy-proven NAFLD and imaging-identified steatosis cases identified by NLP and controls.**

Table 1 shows the comparison of selected clinical and laboratory data for the three groups (imaging-identified steatosis, biopsy-proven NAFLD/NASH, and controls). Compared with the controls, both steatosis/NAFLD groups had higher median BMIs (Body Mass Index) and higher prevalence of diabetes. We examined laboratory values drawn within one year of the liver biopsy or the imaging study showing steatosis (Table 1). Among the three groups, biopsy-proven NAFLD patients had the highest median levels of ALT, AST, AP and GGT, with controls the lowest (Table 1). Interestingly, while the median ALT of patients with imaging-identified steatosis of 24 U/L (IQR) was significantly higher than the median ALT in the control group of 18 U/L, it was still within the normal range. As expected APRI and FIB-4 scores were the highest in biopsy proven steatosis (Table 1). A total of 58% of patients with biopsy-proven steatosis and 38% of patients with imaging-identified steatosis had median ALT levels above the upper limit of the normal range (ULN) within one year of the biopsy/image. Only 53% of patients with biopsy-proven NAFLD and 33% of patients with imaging-identified steatosis had ever been diagnosed with either a NAFLD or NASH ICD-10 code (Table 1). Differences in medication prescription and the completeness of the data for serum markers are reported in Supplementary eTables S7–S9.

Genetic association analysis of imaging-identified steatosis

We used a genetic approach to help validate the identification of steatosis by NLP of radiology reports. Among the patients identified with imaging-identified steatosis, 2840 (6.75%) had genomic data available. We performed a WES (whole exome sequencing) analysis comparing these 2840 imaging-identified steatosis patients to the 21,195 controls, all of whom had genomic data available (Fig. 1). Our QQ plot shows that there is no p-value inflation or deflation, and the divergent tail of the plot represents significant associations for the variants that are in linkage disequilibrium with a causal polymorphism for our phenotype steatosis (Supplementary eFig. S1). Thirty-two single nucleotide variants (SNVs) were significant (Fig. 1), including 14 SNVs that have reported to be associated with NAFLD (Supplementary eTable S10). Consistent with previous literature,[23–25] a robust association was detected for the *PNPLA3* gene cluster (Fig. 1 and Supplementary eTable S10). At the *PNPLA3-SAMM50* region, thirteen SNVs, including rs738409, rs738408, and rs3747207, showed the strongest association with steatosis (best SNP rs738408 $p < 10^{-23}$, Fig. 1).

PheWAS reveals both expected and unexpected associations

PheWAS was employed to scrutinize the underlying causes associated with imaging and biopsy observations of steatosis. We compared imaging-identified steatosis patients to controls and biopsy-proven NAFLD patients to controls (Fig. 2 and Supplementary eTables S11 and S12). 832 phecodes were significantly more common in patients with imaging-identified steatosis compared with controls (Fig. 2 and Supplementary eTable S11). These included a number of liver-related phecodes, as well as obesity, diabetes, and hypertension related diagnoses. In patients with biopsy-proven NAFLD, 275 phecodes were significantly more common compared with controls (Fig. 2 and Supplementary eTable S12). These included a number of expected phecodes such as cholelithiasis and ascites, as well as some unexpected associations such as systemic inflammatory response syndrome (SIRS) and neutropenia. The Venn diagram of significantly associated phecodes showed substantial overlap, with 234 phecodes over-represented in both the imaging-steatosis and biopsy-NAFLD cohorts compared with controls (Fig. 2). For both cohorts, the presence of abnormal serum enzymes preceding the identification of steatosis by imaging or biopsy was one of the most strongly associated phecodes (Odds ratio (OR) 11 as well as 143, respectively). We also confirmed an increased number of metabolic, gastrointestinal, and vascular comorbidities in both steatosis groups compared with controls.

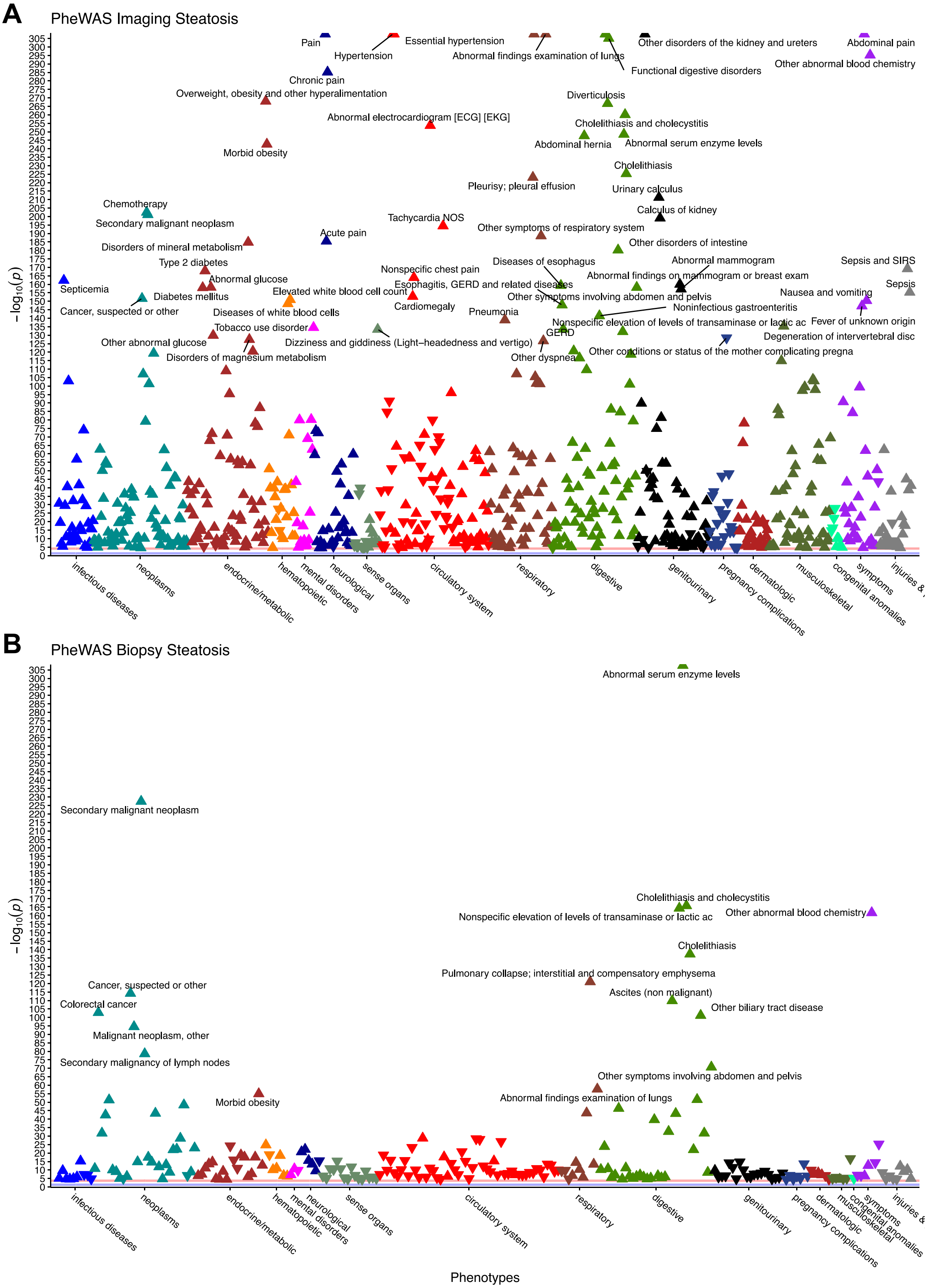
Liver histology findings in biopsy-proven NAFLD comparing NASH to steatosis

Among the biopsy-proven NAFLD patients, we grouped them based on the histology findings into definite NASH, borderline NASH, and steatosis groups (Table 2). All reports that mentioned steatosis/steatohepatitis related key words at least once were included in this study, but qualitative or quantitative measurements of the amount of steatosis were rare. Steatosis grade was mentioned only in 272 patients with steatosis, and most patients were steatosis grade 1. Patients with NASH had significantly more severe steatosis, both quantitatively (29.5%) and qualitatively (9.8%), compared with the patients with steatosis, with borderline NASH intermediate for steatosis (Table 2). The vast majority of the steatosis in all groups was macrovesicular in the qualitative assessment.

By definition, patients with NASH had steatohepatitis that was absent in steatosis patients. However, grading of steatohepatitis in NASH was rarely present in the reports (Supplementary eTable S9, 55.4% missing). Patients with NASH also tended to have more fibrosis in all grades compared to patients with steatosis (Table 2, p-value 0.046, s-value 4.5). The specific fibrosis grade was rarely specified in histology reports (Supplementary eTable S9, 91.1% missing).

Comparison of clinical data in patients with biopsy-proven NASH to biopsy-proven steatosis

While hepatic steatosis is very common, only a fraction of patients with steatosis progress to NASH and fibrotic


A
PheWAS Imaging Steatosis
305 300 295 290 285 280 275 270 265 260 255 250 245 240 235 230 225 220 215 210 205 200 195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100 95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
$-\log_{10}(p)$
Pain
Essential hypertension
Other disorders of the kidney and ureters
Abdominal pain
Hypertension
Abnormal findings examination of lungs
Other abnormal blood chemistry
Functional digestive disorders
Chronic pain
Overweight, obesity and other hyperalimentation
Diverticulosis
Abnormal electrocardiogram [ECG] [EKG]
Cholelithiasis and cholecystitis
Abdominal hernia
Abnormal serum enzyme levels
Morbid obesity
Cholelithiasis
Pleurisy; pleural effusion
Urinary calculus
Chemotherapy
Calculus of kidney
Tachycardia NOS
Secondary malignant neoplasm
Other symptoms of respiratory system
Acute pain
Disorders of mineral metabolism
Other disorders of intestine
Type 2 diabetes
Diseases of esophagus
Abnormal mammogram
Sepsis and SIRS
Nonspecific chest pain
Abnormal glucose
Abnormal findings on mammogram or breast exam
Septicemia
Esophagitis, GERD and related diseases
Sepsis
Elevated white blood cell count
Nausea and vomiting
Diabetes mellitus
Other symptoms involving abdomen and pelvis
Cardiomegaly
Noninfectious gastroenteritis
Cancer, suspected or other
Diseases of white blood cells
Pneumonia
Nonspecific elevation of levels of transaminase or lactic ac
Fever of unknown origin
Tobacco use disorder
GERD
Degeneration of intervertebral disc
Other abnormal glucose
Dizziness and giddiness (Light−headedness and vertigo)
Other conditions or status of the mother complicating pregna
Disorders of magnesium metabolism
Other dyspnea
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries &
B
PheWAS Biopsy Steatosis
305 300 295 290 285 280 275 270 265 260 255 250 245 240 235 230 225 220 215 210 205 200 195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100 95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20 15 10 5 0
$-\log_{10}(p)$
Abnormal serum enzyme levels
Secondary malignant neoplasm
Cholelithiasis and cholecystitis
Nonspecific elevation of levels of transaminase or lactic ac
Other abnormal blood chemistry
Cholelithiasis
Pulmonary collapse; interstitial and compensatory emphysema
Cancer, suspected or other
Ascites (non malignant)
Colorectal cancer
Other biliary tract disease
Malignant neoplasm, other
Secondary malignancy of lymph nodes
Other symptoms involving abdomen and pelvis
Morbid obesity
Abnormal findings examination of lungs
infectious diseases
neoplasms
endocrine/metabolic
hematopoietic
mental disorders
neurological
sense organs
circulatory system
respiratory
digestive
genitourinary
pregnancy complications
dermatologic
musculoskeletal
congenital anomalies
symptoms
injuries &
Phenotypes

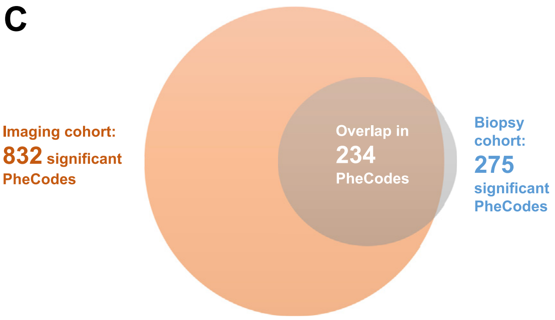


Fig. 2: Continued.

liver disease. One of the major questions regarding NAFLD is what factors determine who progresses from steatosis to NASH and cirrhosis. To better understand this question, we compared patients with biopsy-proven NASH to those with biopsy-proven steatosis (Supplementary eTable S8). The two groups were similar in age, sex, and BMI. Patients with NASH were significantly more likely to have T2DM (36% vs 27%, s-value 20.8.) and had higher levels of HbA1c, glucose, and insulin. Patients with NASH also had higher ALT and AST levels compared to those with steatosis (Supplementary eTable S8). Triglycerides were significantly higher and HDL-C significantly lower in NASH patients compared to steatosis (s-values 14.8 and 11.2 for Triglycerides and HDL-C). Patients with borderline NASH had intermediate ALT and AST levels (Supplementary eTable S8). FIB-4 scores and APRI scores differed between the different biopsy-proven steatosis groups. NASH showed the highest APRI and FIB-4 scores, followed by borderline NASH and then steatosis (Supplementary eTable S8).

Discussion

In this study we demonstrate that NLP applied to imaging reports in an academic health system identified a large number (>42,000) patients with hepatic steatosis and NLP applied to liver biopsy pathology reports identified >3000 patients with NAFLD/NASH, all with extensive longitudinal EHR data available. Less than 40% of these patients had ICD codes for NAFLD or NASH. By contrast, a recent publication from the NASH Clinical Research Network reported on a median longitudinal follow-up of four years in 1773 patients with biopsy-proven NAFLD.[27] Applying NLP, the algorithm led to the identification of steatosis in 22% of all evaluated biopsy cases and 25% of all evaluated images that included the liver. These numbers are lower than the estimated prevalence of NAFLD of 30% of the general population.[28] We hypothesize that these differences arise due to the underdiagnosis of NAFLD and the low percentage of all imaging studies across the EHR that capture the liver parenchyma. Still, our study shows how large-scale data mining can dramatically enhance the identification of patients with hepatic steatosis compared to using ICD codes. Our research highlights the undercoding of NAFLD ICD codes in the general population, which leads to less care received by the affected patients, and our results point toward an effective approach for case identification to enhance the use of the EHR to better understand the factors that contribute to disease progression. However, validation in other large-scale cohorts is needed and we acknowledge that a proportion of individuals with NAFLD were missed using the NLP method. Our findings highlighting the undercoding of steatosis certainly shed light on the need for improvements in disease coding practices. By using NLP algorithms to identify patients with steatosis from EHRs, we hope to demonstrate the potential for automated systems to assist in improving diagnostic accuracy and disease coding. While the implementation of such improvements would require further steps, including policy changes and training, our study contributes to the body of evidence supporting this direction.

The widespread availability of EHRs in healthcare systems provides an excellent opportunity for clinical research and advanced patient care through clinical decision support.[29] Nevertheless, proper use of the large-scale data in EHRs relies on the objective and prompt identification of patients with the disease of interest.[30] Integrating data extracted by NLP into a phenotyping algorithm has various advantages. First, NLP can provide data that is not available in structured EHR databases or in contexts where the accuracy of the structured data is low. For example, before 2012, no specific ICD-9 code existed for NASH. Secondly, NLP can systematically connect multiple terms to a concept. NLP differs from a “find” command because the algorithm can be trained to recognize that the terms “steatosis,” “NAFLD,” and “NAS-Score” are all related to the concept of hepatic steatosis.

Here we have identified the remarkable underuse of steatosis diagnosis codes, even after a radiologist has reported an incidental finding of steatosis on imaging or a pathologist has reported steatosis or NAFLD on a

Fig. 2: Comorbidity PheWAS analysis for patients with (A) steatosis on imaging and (B) biopsy proven steatosis compared to controls. This analysis only includes diagnoses that were diagnosed prior to the imaging. Manhattan plot of adjusted $-\log_{10}$ (p-values) for all PheCodes comparing their occurrence. Highlighted are associations results with p-values $< 3 \times 10^{-4}$. Upwards/downwards pointing trials refer to PheCodes that are over-/underrepresented. (C) Venn Diagram showing overlap of phecodes that were diagnosed prior to the imaging/biopsy for biopsy and imaging identified steatosis.

Biopsy characteristics	NASH (n = 1210)	Borderline NASH (n = 456)	Steatosis (n = 1341)	p[a] steatosis vs NASH	s[b] steatosis vs NASH	p Borderline NASH vs NASH	s Borderline NASH vs NASH	p Steatosis vs Borderline NASH	s Steatosis vs Borderline NASH
Steatosis assessment									
Quantitative Steatosis grade, n (%)									
1	162 (47.2)	70 (79.5)	238 (87.5)	<0.001	78.7	<0.001	21.9	0.17	2.6
2	80 (23.3)	11 (12.5)	19 (7)						
3	101 (29.5)	7 (8)	15 (5.5)						
Qualitative Steatosis assessment									
Macrovesicular, n (%)	324 (26.8)	112 (24.6)	294 (21.9)	0.005	7.7	0.39	1.4	0.27	1.9
Microvesicular, n (%)	21 (1.7)	8 (1.8)	24 (1.8)	1	<0.001	1	<0.001	1	<0.001
Degree of steatosis, n (%)									
mild	417 (55.9)	166 (76.5)	484 (84.4)	<0.001	89.4	<0.001	22.9	0.03	5.1
moderate	256 (34.3)	45 (20.7)	76 (13.3)						
severe	73 (9.8)	6 (2.8)	13 (2.3)						
Steatohepatitis assessment									
NLP identified key word: Steatohepatitis, n (%)	1210 (100)	0 (0)	0 (0)						
Degree of steatohepatitis, n (%)									
mild	402 (74.4)	0 (0)	0 (0)						
moderate	122 (22.6)	0 (0)	0 (0)						
severe	16 (3)	0 (0)	0 (0)						
NLP identified key word: Ballooning, n (%)	193 (16)	67 (14.7)	0 (0)			0.58	0.8		
Fibrosis grade, n (%)									
1	57 (52.8)	5 (62.5)	16 (47.1)	0.046	4.5	0.39	1.4	1	0
2	23 (21.3)	3 (37.5)	13 (38.2)						
3	26 (24.1)	0 (0)	3 (8.8)						
4	2 (1.8)	0 (0)	2 (5.9)						
NLP identified key word: Fibrosis, n (%)	146 (12.1)	398 (87.3)	0 (0)			<0.001	617.3		
NLP identified key word: Cirrhosis, n (%)	220 (18.2)	0 (0)	0 (0)						

NAFLD, non-alcoholic fatty liver disease; NASH, non-alcoholic steatohepatitis. [a]p; p-value obtained from Pearson's chi-squared test for categorical variables; with exception for fibrosis grade, the p-value obtained from Fisher's exact test, as more than 20% of the cells have frequency <5. All variables were non-normally distributed. [b]s; s-value obtained from formula: s-value = $-\log_2$ (p-value).

Table 2: **Biopsy Characteristics of patients with biopsy-proven NASH, Borderline NASH, and steatosis.**

biopsy. Surprisingly, even biopsy-proven NASH was reflected in diagnosis codes less than half of the time. This is clinically important, because individuals with steatosis may warrant further assessment of disease severity, monitoring for disease progression, and benefit from medical optimization of metabolic risk factors.[31]

Various approaches to identifying NAFLD patients using EHRs have been described and the majority of NLP cohorts have comprised a few hundred radiological/histologically confirmed NAFLD patients.[17,32,33] An earlier study used NLP to identify NAFLD within the EHR through radiology reports combined with ICD codes and found that the combination had a PPV (89%) and NPV (56%), which was superior to a method utilizing ICD coding alone or a model including AST/ALT serum measurements.[34] Some studies used deep learning to predict steatosis on images or histopathology images.[35,36] Still, other NLP approaches counted only the occurrences of pre-defined terms related to NAFLD without considering crucial issues in NLP, including negation, context, spelling, and acronyms.[37,38] Nonetheless, our data support the findings that steatosis-related ICD codes are vastly underutilized in the EHR.

Interestingly, sex-specific ALT was only elevated in ~40% of NAFLD patients, identifying NLP as clearly superior for unique phenotype algorithms. This is especially important in the context of elevated ALT as a proxy for steatosis, as this approach would miss 50% of the steatosis population. A recent paper estimated a 25% rate of missed diagnosis by using ALT as a proxy based on diagnosed NAFLD patients,[39] but our results suggest this is an underestimate. The same is true for GGT levels; while GGT is a consistent marker for NAFLD, elevations can be attributed to metabolic phenotypes in NAFLD patients and are also seen in unrelated liver conditions such as cholestatic and alcohol-induced liver disease.[40] Other recent studies showed abnormal GGT and ALT in only 46% of image-confirmed NAFLD patients,[41] demonstrating the superiority of NLP as supported by our study. Furthermore, we saw a difference between biopsy-proven and imaging-identified patients. This could be due to the severity of liver disease, since patients undergoing imaging for a non-liver indication were more likely to have an incidental finding of steatosis, whereas patients with liver biopsies were more likely to have medical complications that justified an invasive procedure. This is also reflected in the differences in the baseline demographics of the two groups. These findings support the use of NLP of the EHR, especially of imaging and pathology reports, as a valuable tool for the early detection of steatosis to initiate accurate patient care.

We found that NLP of imaging reports in identifying hepatic steatosis is highly accurate. Furthermore, we used available genomic data and confirmed associations with well-established genetic variants associated with steatosis and NAFLD.[42–44] Finally, we used PheWAS and found many highly significant expected associations with imaging-identified and biopsy-proven steatosis. Given the substantial number of identified individuals and a precise collection of disease phenotypes, we were able to gain new insights and identified 234 comorbidities that are significantly associated with the presence of steatosis. We confirmed a strong relationship with specific disorders such as hypertension, diabetes mellitus or obesity, which are well documented in the literature.[45,46] We confirmed earlier findings associating steatosis to metabolic comorbidities.[2,45] Our analyses suggest that imaging-identified steatosis was associated with more comorbidities but less severe elevation of steatosis-associated serum biomarkers, which might also reflect the clinical context for the imaging study. Consequently, the PheWAS results should be interpreted as reflecting the range of clinical indications for obtaining imaging in this cohort, since studies were typically not performed for evaluation of liver disease specifically.

High-throughput identification of steatosis with electronic follow-up through the EHR could aid in understanding the risk factors for progression to NASH in the future. Analysis of progression from steatosis to NASH is complicated, particularly when a NASH ICD diagnosis is often not made, as we reported in this study. A robust scientific evaluation of progression from NAFLD to NASH would require biopsy confirmation of NASH. In this study we selected the latest available biopsy specimens for analysis, but future studies could examine patients with interval biopsies to identify markers associated with progressive disease.

Our study should be interpreted in the context of its limitations. First, we used NLP for our analyses, which along with ICD coding errors or omissions, could have erroneously labelled patients as having steatosis. We did perform a manual analysis of a subset of reports to help negate this potential source of error. The limitations of the study design and analysis including unmeasured confounding have to be mentioned. The PheWAS analysis is well suited to identify an extensive repertoire of steatosis-associated conditions. The missing temporal information may introduce reverse causation bias. Still, as we have shown in this study, outcomes based on ICD codes suffer from misclassification and underdiagnosis. Another limitation of PheWAS analysis is sparse-data bias[47] which is an important limitation of EHR studies. Another limitation of this study is that some of the imaging studies and biopsies were obtained for evaluation of extrahepatic malignancies, which is among the leading causes of death among patients with NAFLD.[48] Therefore, a systematic approach to quantifying the various indications for both imaging and biopsy reports is needed in future studies. In addition, Penn patients may differ from patients in other healthcare systems and the algorithm may need adjustments. Finally, we do not claim that our analysis fully identified *all* undiagnosed NAFLD cases. Based on the QC data, we are confident that NLP largely identified those who were noted in the imaging and pathology reports. But under-diagnosed patients who never had imaging or pathology reports or whose condition was not noted in the reports are yet to be identified.

In conclusion, we reveal that NLP-based approaches can identify large cohorts of biopsy and imaging-proven steatosis patients. We have shown that NLP has superior accuracy in identifying biopsy-proven NAFLD and NASH within the EHR compared to ICD codes as well as ALT serum measurements within one year of the biopsy/imaging. There is a lack of acknowledgment in clinical documentation of NAFLD findings in radiology reports, and a considerable number of these patients are later reported to have NASH. Our observations suggest that NLP-based approaches at scale have the potential to identify patients with important undiagnosed conditions, like steatosis and NAFLD, that have clinical consequences and warrant additional follow-up.

Contributors

Acquisition of data: C.V.S, D.J.R; Analysis and interpretation of data: T.L., C.V.S., A.I.M., P.R.; Drafting of the manuscript: C.V.S, T.L, M.V., E.S, K.M.S, J.C, D.J.R; Critical revision of the manuscript for important intellectual content: all authors; Figures and tables: L.H, C.V.S; Statistical analysis: D.Z., T.L., C.V.S; Obtained funding: C.V.S, J.C, D.J.R; Study supervision: C.V.S, J.C, D.J.R.

Data sharing statement

The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research data is not available, but the data analysis plan, is available at request from the corresponding author after a signed data access agreement.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

C.V.S is supported NRW Rueckkehr Programme. K.M.S is supported by the Federal Ministry of Education and Research (BMBF) and the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) under the Excellence strategy of the federal government and the Laender. A.I.M. is supported by 5T32DK007066, K.J.W. is supported by R01-CA259201, T.L. and J.C are supported by R01-HL138306, UL1-TR001878 and PICORI. R.M.C. is supported by NIH 1R01AA026302. K.T.C. and D.J.R. are supported by NIH R01DK114291. The Penn Medicine Biobank (PMBB) is funded by the Perelman School of Medicine at the University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA Award Number UL1TR001878. Regeneron Genetics Center contributed to this work by providing PMBB the services for genetic analyses, which were used in this paper. M.V. and K.M.C. are supported by VA BX003362. M.V. is supported by R01-DK134575. I.Z. is supported by NIH 5T32DK007066.

C.V.S and T.L. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors agreed to submit the manuscript, read, and approved the final draft. Jacob Fiksel has received grants, stocks as well as travel support from Vertex Pharmaceuticals and royalties for Software for Cell-free DNA Fragmentation in Patients with Cancer, Johns Hopkins University Rotonya Carr has received consulting fees from Intercept, has a leadership role in American Gastroenterological Association, American Association for the Study of Liver Diseases, American Liver Foundation as well as Association of Black Gastroenterologists and Hepatologists and received research funding from Merck. Protocols for this study were approved by the Institutional Review Board (IRB) at Penn Medicine (#813913).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2023.102149.

References

1 Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. 2016;64(1):73–84. https://doi.org/10.1002/hep.28431.

2 Baratta F, Pastori D, Angelico F, et al. Nonalcoholic fatty liver disease and fibrosis associated with increased risk of cardiovascular events in a prospective study. *Clin Gastroenterol Hepatol*. 2019;18(10):2324–2331.e4. https://doi.org/10.1016/j.cgh.2019.12.026.

3 Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the American association for the study of liver diseases, American college of gastroenterology, and the American gastroenterological association. *Hepatology*. 2012;55(6):2005–2023. https://doi.org/10.1002/hep.25762.

4 Vujkovic M, Ramdas S, Lorenz KM, et al. A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation. *Nat Genet*. 2022;54(6):761–771. https://doi.org/10.1038/s41588-022-01078-z.

5 Dyson JK, Anstee QM, McPherson S. Non-alcoholic fatty liver disease: a practical approach to diagnosis and staging. *Frontline Gastroenterol*. 2014;5(3):211–218. https://doi.org/10.1136/flgastro-2013-100403.

6 Arshad T, Golabi P, Henry L, Younossi ZM. Epidemiology of non-alcoholic fatty liver disease in North America. *Curr Pharm Des*. 2020;26(10):993–997. https://doi.org/10.2174/1381612826666200303114934.

7 Hardy T, Oakley F, Anstee QM, Day CP. Nonalcoholic fatty liver disease: pathogenesis and disease spectrum. *Annu Rev Pathol*. 2016;11:451–496. https://doi.org/10.1146/annurev-pathol-012615-044224.

8 Wong VW-S, Chitturi S, Wong GL-H, Yu J, Chan HL-Y, Farrell GC. Pathogenesis and novel treatment options for non-alcoholic steatohepatitis. *Lancet Gastroenterol Hepatol*. 2016;1(1):56–67. https://doi.org/10.1016/S2468-1253(16)30011-5.

9 Singh S, Allen AM, Wang Z, Prokop LJ, Murad MH, Loomba R. Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis: a systematic review and meta-analysis of paired-biopsy studies. *Clin Gastroenterol Hepatol*. 2015;13(4):640–643. https://doi.org/10.1016/j.cgh.2014.04.014.

10 Serper M, Vujkovic M, Kaplan DE, et al. Validating a non-invasive, ALT-based non-alcoholic fatty liver phenotype in the million veteran program. *PLoS One*. 2020;15(8):e0237430. https://doi.org/10.1371/journal.pone.0237430.

11 Crawford AR, Lin XZ, Crawford JM. The normal adult human liver biopsy: a quantitative reference standard. *Hepatology*. 1998; 28(2):323–331. https://doi.org/10.1002/hep.510280206.

12 Kanwal F, Shubrook JH, Adams LA, et al. Clinical care pathway for the risk stratification and management of patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2021;161(5):1657–1669. https://doi.org/10.1053/j.gastro.2021.07.049.

13 Vieira Barbosa J, Lai M. Nonalcoholic fatty liver disease screening in type 2 diabetes mellitus patients in the primary care setting. *Hepatol Commun*. 2021;5(2):158–167. https://doi.org/10.1002/hep4.1618.

14 EASL-EASD-EASO Clinical Practice Guidelines for the management of non-alcoholic fatty liver disease. *J Hepatol*. 2016;64(6):1388–1402. https://doi.org/10.1016/j.jhep.2015.11.004.

15 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309(13):1351–1352. https://doi.org/10.1001/jama.2013.393.

16 Scriver CR. After the genome–the phenome? *J Inherit Metab Dis*. 2004;27(3):305–317. https://doi.org/10.1023/B:BOLI.0000031100.26546.6e.

17 Redman JS, Natarajan Y, Hou JK, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. *Dig Dis Sci*. 2017;62(10):2713–2718. https://doi.org/10.1007/s10620-017-4721-9.

18 Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH. Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology*. 2021;73(6):2546–2563. https://doi.org/10.1002/hep.31603.

19 Cormack J, Nath C, Milward D, Raja K, Jonnalagadda SR. Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. *J Biomed Inform*. 2015;58:S120–S127.

20 Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. *Hepatology*. 2018;67(1):328–357. https://doi.org/10.1002/hep.29367.

21 Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. *Hepatology*. 2006;43(6):1317–1325. https://doi.org/10.1002/hep.21178.

22 Loaeza-del-Castillo A, Paz-Pineda F, Oviedo-Cárdenas E, Sánchez-Avila F, Vargas-Vorácková F. AST to platelet ratio index (APRI) for the noninvasive evaluation of liver fibrosis. *Ann Hepatol*. 2008;7(4):350–357.

23 Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol*. 2020;20(1):244. https://doi.org/10.1186/s12874-020-01105-9.

24 Greenland S, Mansournia MA, Joffe M. To curb research misreporting, replace significance and confidence by compatibility: a Preventive Medicine Golden Jubilee article. *Prev Med*. 2022;164:107127. https://doi.org/10.1016/j.ypmed.2022.107127.

25 Mansournia MA, Nazemipour M, Etminan M. P-value, compatibility, and S-value. *Glob Epidemiol*. 2022;4:100085. https://doi.org/10.1016/j.gloepi.2022.100085.

26 Nicholls SM, Clare A, Randall JC. Goldilocks: a tool for identifying genomic regions that are 'just right.'. *Bioinformatics*. 2016;32(13):2047–2049. https://doi.org/10.1093/bioinformatics/btw116.

27 Sanyal AJ, Van Natta ML, Clark J, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. *N Engl J Med*. 2021;385(17):1559–1569. https://doi.org/10.1056/NEJMoa2029349.

28 Younossi ZM. Non-alcoholic fatty liver disease - a global public health perspective. *J Hepatol*. 2019;70(3):531–544. https://doi.org/10.1016/j.jhep.2018.10.033.

29 Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy*. 2011;4:47.

30 Kagawa R, Shinohara E, Imai T, Kawazoe Y, Ohe K. Bias of inaccurate disease mentions in electronic health record-based phenotyping. *Int J Med Inform*. 2019;124:90–96. https://doi.org/10.1016/j.ijmedinf.2018.12.004.

31 Rinella ME, Neuschwander-Tetri BA, Siddiqui MS, et al. AASLD Practice Guidance on the clinical assessment and management of nonalcoholic fatty liver disease. *Hepatology*. 2023;77(5):1797–1835. https://journals.lww.com/hep/Fulltext/2023/05000/AASLD_Practice_Guidance_on_the_clinical_assessment.31.aspx.

32 Walker RW, Belbin GM, Sorokin EP, et al. A common variant in PNPLA3 is associated with age at diagnosis of NAFLD in patients from a multi-ethnic biobank. *J Hepatol*. 2020;72(6):1070–1081. https://doi.org/10.1016/j.jhep.2020.01.029.

33 Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:h1885.

34 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. *Dig Dis Sci*. 2016;61(3):913–919.

35 Alshagathrh FM, Househ MS. Artificial intelligence for detecting and quantifying fatty liver in ultrasound images: a systematic review. *Bioengineering*. 2022;9(12):748. https://doi.org/10.3390/bioengineering9120748.

36 Roy M, Wang F, Vo H, et al. Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. *Lab Invest*. 2020;100(10):1367–1383. https://doi.org/10.1038/s41374-020-0463-y.

37 Cohen KB, Goss FR, Zweigenbaum P, Hunter LE. Translational morphosyntax: distribution of negation in clinical records and biomedical journal articles. *Stud Health Technol Inform*. 2017;245:346–350.

38 Moon S, McInnes B, Melton GB. Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. *Healthc Inform Res*. 2015;21(1):35–42.

39 Ma X, Liu S, Zhang J, et al. Proportion of NAFLD patients with normal ALT value in overall NAFLD patients: a systematic review and meta-analysis. *BMC Gastroenterol*. 2020;20(1):10. https://doi.org/10.1186/s12876-020-1165-z.

40 Neuman MG, Malnick S, Chertin L. Gamma glutamyl transferase - an underestimated marker for cardiovascular disease and the metabolic syndrome. *J Pharm Pharm Sci*. 2020;23(1):65–74. https://doi.org/10.18433/jpps30923.

41 Ma Q, Liao X, Shao C, et al. Normalization of γ-glutamyl transferase levels is associated with better metabolic control in individuals with nonalcoholic fatty liver disease. *BMC Gastroenterol*. 2021;21(1):215. https://doi.org/10.1186/s12876-021-01790-w.

42 Schneider CV, Fromme M, Schneider KM, Bruns T, Strnad P. Mortality in patients with genetic and environmental risk of liver disease. *Am J Gastroenterol*. 2021;116(8):1741–1745. https://doi.org/10.14309/ajg.0000000000001326.

43 Dongiovanni P, Petta S, Maglio C, et al. Transmembrane 6 superfamily member 2 gene variant disentangles nonalcoholic steatohepatitis from cardiovascular disease. *Hepatology*. 2015;61(2):506–514. https://doi.org/10.1002/hep.27490.

44 Wijarnpreecha K, Scribani M, Raymond P, et al. PNPLA3 gene polymorphism and liver- and extrahepatic cancer-related mortality in the United States. *Clin Gastroenterol Hepatol*. 2020;19(5):1064–1066. https://doi.org/10.1016/j.cgh.2020.04.058.

45 Stender S, Kozlitina J, Nordestgaard BG, Tybjærg-Hansen A, Hobbs HH, Cohen JC. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. *Nat Genet*. 2017;49(6):842–847. https://doi.org/10.1038/ng.3855.

46 Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology*. 2018;67(1):123–133. https://doi.org/10.1002/hep.29466.

47 Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ*. 2016;352:i1981. https://doi.org/10.1136/bmj.i1981.

48 Thomas JA, Kendall BJ, Dalais C, Macdonald GA, Thrift AP. Hepatocellular and extrahepatic cancers in non-alcoholic fatty liver disease: a systematic review and meta-analysis. *Eur J Cancer*. 2022;173:250–262. https://doi.org/10.1016/j.ejca.2022.06.051.