International Journal of Medical Informatics 129 (2019) 334–341

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression

Tielman T. Van Vleck[a,*], Lili Chan[b], Steven G. Coca[b], Catherine K. Craven[c,d], Ron Do[a], Stephen B. Ellis[a], Joseph L. Kannry[e], Ruth J.F. Loos[a], Peter A. Bonis[f], Judy Cho[a,g,h,1], Girish N. Nadkarni[a,b,*,1]

[a] The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, USA
[b] Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, USA
[c] Institute for Healthcare Delivery Science, Dept. of Pop. Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, USA
[d] Clinical Informatics Group, IT Department, Mount Sinai Health System, New York, USA
[e] Information Technology, Mount Sinai Medical Center, New York, USA
[f] Division of Gastroenterology, Tufts Medical Center, Boston, USA
[g] Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, USA
[h] Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, USA

ARTICLE INFO

Keywords:
Natural language processing
Patient safety
NAFLD

ABSTRACT

Objective: Electronic health record (EHR) systems contain structured data (such as diagnostic codes) and unstructured data (clinical documentation). Clinical insights can be derived from analyzing both. The use of natural language processing (NLP) algorithms to effectively analyze unstructured data has been well demonstrated. Here we examine the utility of NLP for the identification of patients with non-alcoholic fatty liver disease, assess patterns of disease progression, and identify gaps in care related to breakdown in communication among providers.

Materials and Methods: All clinical notes available on the 38,575 patients enrolled in the Mount Sinai Bio*Me* cohort were loaded into the NLP system. We compared analysis of structured and unstructured EHR data using NLP, free-text search, and diagnostic codes with validation against expert adjudication. We then used the NLP findings to measure physician impression of progression from early-stage NAFLD to NASH or cirrhosis. Similarly, we used the same NLP findings to identify mentions of NAFLD in radiology reports that did not persist into clinical notes.

Results: Out of 38,575 patients, we identified 2,281 patients with NAFLD. From the remainder, 10,653 patients with similar data density were selected as a control group. NLP outperformed ICD and text search in both sensitivity (NLP: 0.93, ICD: 0.28, text search: 0.81) and F2 score (NLP: 0.92, ICD: 0.34, text search: 0.81). Of 2281 NAFLD patients, 673 (29.5%) were believed to have progressed to NASH or cirrhosis. Among 176 where NAFLD was noted prior to NASH, the average progression time was 410 days. 619 (27.1%) NAFLD patients had it documented only in radiology notes and not acknowledged in other forms of clinical documentation. Of these, 170 (28.4%) were later identified as having likely developed NASH or cirrhosis after a median 1057.3 days.

Discussion: NLP-based approaches were more accurate at identifying NAFLD within the EHR than ICD/text search-based approaches. Suspected NAFLD on imaging is often not acknowledged in subsequent clinical documentation. Many such patients are later found to have more advanced liver disease. Analysis of information flows demonstrated loss of key information that could have been used to help prevent the progression of early NAFLD (NAFL) to NASH or cirrhosis.

Conclusion: For identification of NAFLD, NLP performed better than alternative selection modalities. It then facilitated analysis of knowledge flow between physician and enabled the identification of breakdowns where key information was lost that could have slowed or prevented later disease progression.

* Corresponding author at: Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1243, New York, NY 10029, USA.
E-mail addresses: tielman.vanvleck@mssm.edu (T.T. Van Vleck), girish.nadkarni@mountsinai.org (G.N. Nadkarni).

[1] Equal Contribution.

https://doi.org/10.1016/j.ijmedinf.2019.06.028
Received 12 March 2019; Received in revised form 20 May 2019; Accepted 28 June 2019

1. Introduction

Artificial intelligence has shown increasing promise when applied identification and prediction of countless medical outcomes. When applied to clinician workflow, it provides a form of augmented intelligence, aiding clinicians with decision support and error avoidance. These applications have predominantly been built leveraging only structured data because of its availability and ease of interpretation, however unstructured data (such as dictated notes) contain critical clinical information and thereby offer the potential to greatly enhance clinical insights than can be derived from use of structured data alone. Several public and proprietary approaches have been taken to developing natural language processing (NLP) systems to make sense of unstructured clinical data [1,2]. NLP approaches have been used successfully for biomedical research such as accurate phenotyping of complex diseases and for clinical tasks including identification of patients with NAFLD [9,10]. Here we examine the use of one SNOMED-based NLP tool for extracting patient features related non-alcoholic fatty liver disease (NAFLD). NAFLD represents a spectrum of liver diseases characterized histologically by macrovesicular fat and ranging in severity from nonalcoholic fatty liver (NAFL) to non-alcoholic steatohepatitis (NASH) [3,4]. A subset of patients with NAFLD progresses to cirrhosis and has an increased risk of hepatocellular carcinoma and liver-related mortality [3]. NAFLD is emerging as one of the most common causes of liver failure in the United States [5]. Multiple professional societies have published guidelines for the diagnosis and management of patients with NAFLD [6]. NAFLD is suspected in patients with metabolic syndrome, hepatomegaly, or mild elevations in aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels. However, normal levels of AST and ALT do not exclude the presence of NAFLD and hepatomegaly is found only in approximately 20% of patients [7,8]. A key component in the diagnosis of NAFLD is evidence of hepatic steatosis on imaging or biopsy. Many patients have abdominal or chest imaging performed for unrelated disorders which may incidentally find hepatic steatosis and allow for additionally workup for NAFLD including exclusion of other chronic liver diseases and alcohol consumption. The rapid and accurate identification of NAFLD by NLP from unstructured text such as radiology reports is one potential method to address the gap between incidental findings and patient care.

We first assess the accuracy of NLP against other simpler approaches to derive insights into the treating clinician's understanding of the patient. We then determine the ability of NLP to track the diagnostic process and identify potential breakdowns. We determine the proportion of patients with fatty liver documented in radiology reports in which the presumptive diagnosis was also documented in a progress note from a healthcare provider. Second, to identify communication breakdowns at the point of care, we examine patients where NAFLD was identified in radiology notes but never referenced in progress note. While we focused our approach on NAFLD, the methodology is broadly applicable to other disease processes.

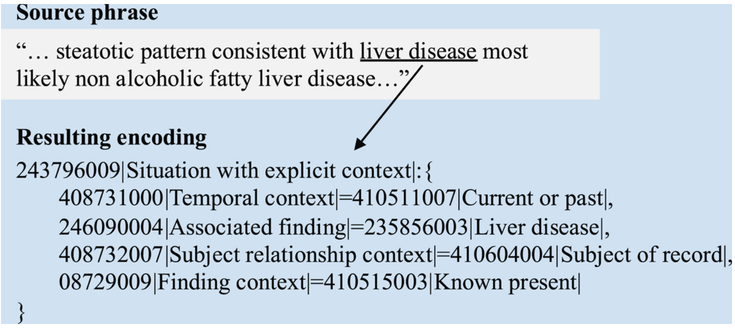
2. Materials & methods

2.1. Study setting and population

The study was conducted at the Icahn School of Medicine at Mount Sinai and used the data resources of the Bio*Me* Biobank at the Charles Bronfman Institute of Personalized Medicine [11]. The Bio*Me* Biobank is a prospective cohort study with over 40,000 ethnically diverse patients recruited from primary care and subspecialty clinics within the Mount Sinai Health System, used for a diverse range of associated studies [12–15]. Bio*Me* has no inclusion or exclusion criteria beyond that they must be able to enroll themselves. Open enrollment is conducted at multiple Mount Sinai Hospitals around greater New York City, resulting in one of the most genetically diverse biobanks created. The Institutional Review Board approved the study and informed consent was obtained for all subjects. All patient notes are uploaded weekly to Bio*Me* servers, with notes dating back to 2007.

2.2. Data preparation: natural language processing

We extracted the clinical documentation for all Bio*Me* participants from the centralized DataMart up to December 31, 2017, with the sample starting June 11, 2009. Enrollment times for each patient was relatively random as it was a continuous process. Children remained enrolled until their 18th birthday (at which point they have to re-enroll) or they died. Clinical documentation was comprised of progress notes, radiology reports, discharge summaries and pathology reports. We then applied the CLiX clinical NLP engine produced by Clinithink (see Appendix for more information) to this cohort, an algorithm recently demonstrated to be useful in patient phenotyping [16]. CLiX is a general-purpose stochastic parser, which maps patient facts described in clinical narrative to post-coordinated SNOMED expressions, thereby creating a highly descriptive, standardized data layer capturing all identified clinical facts across domains and clinical contexts [17]. SNOMED expressions are based on SNOMED CT, a granular, hierarchical, general-purpose clinical terminology combining the most comprehensive single English terminology for medicine. The SNOMED compositional grammar then specifies how concepts should be combined as expressions, making it possible to describe the clinical context around the core concept with modifiers such as laterality, certainty, tense, negation or the person being discussed. The combinatorial effect allows for substantially more expressivity than possible with single concepts with post-coordination [17]. For example, a condition mentioned as part of a differential diagnosis will be assigned a *Finding context* of *Probably present*, while when expressed definitively, it will be assigned *Known present*. While SNOMED (the 2016 US edition) has approximately 430,000 concepts, the NLP system identified 6,665,726 unique expressions to describe 420,181,346 total findings on the



Source phrase

"… steatotic pattern consistent with liver disease most likely non alcoholic fatty liver disease…"

Resulting encoding

```
243796009|Situation with explicit context|:{
      408731000|Temporal context|=410511007|Current or past|,
      246090004|Associated finding|=235856003|Liver disease|,
      408732007|Subject relationship context|=410604004|Subject of record|,
      08729009|Finding context|=410515003|Known present|
}
```

Fig. 1. A SNOMED expression output by the NLP system on the phrase "liver disease", including associated metadata describing the context. When used as a query, this expression would identify all references to any liver disease known present or previously present.

patients in this cohort. Supplementary Appendix Fig. 1 demonstrates the SNOMED expressions identified for a sample phrase, and how the core concepts fit into the SNOMED CT terminology.

To identify patients matching specific criteria, we used a SNOMED query engine (a second component of CLiX) to perform hierarchical subsumption queries identifying relevant SNOMED expressions, meaning the identification of all SNOMED expressions found for the patient that were logical descendants (according to the SNOMED CT hierarchy) of each query expression. For example, according to SNOMED, *NAFLD* is a great-grandchild of *Disease of liver* and the temporal context of *Current or past* is a child of *Temporal context value*. The complete query for patients known to have had some form of liver disease is shown in Fig. 1.

We used a SNOMED browser to identify other concepts critical to the identification of patients with NAFLD that were not SNOMED descendants of NAFLD to ensure that our queries captured patients with all concepts appropriate for the analysis. This led us to conclude that the optimal method would be to search for patients with the NAFLD parent, *Steatosis of liver*, and exclude diagnoses other than NAFLD. As a result, we excluded patients with Wilson disease, abetalipoproteinemia, alcoholic fatty liver, alcoholism (e.g. problem drinker, very heavy drinker, heavy drinker) and hepatitis B and C using a combination of NLP and ICD9/10 criteria.

2.3. Manual review study design

In order to assess the accuracy of NAFLD patient selection, we performed a manual review on three different methods for patient selection. The study design is shown in Fig. 2.

After parsing the notes using NLP queries as described above, we conducted a simple text search for pre-defined phrases pertaining to NAFLD.

2.3.1. ICD codes and simple text search to identify NAFLD

We compared NLP-based approaches to searching structured data and simpler text-based approaches for identifying NAFLD within the EHR. As there is not ICD code for NAFLD itself, structured search looked for patients with ICD-9-CM and ICD-10-CM codes commonly used for NAFLD [18–20], including:

- ICD-9-CM 571.8: Other chronic nonalcoholic liver disease
- ICD-9-CM 571.9: Unspecified chronic liver disease without mention of alcohol
- ICD-10-CM K76.0: Fatty (change of) liver, not elsewhere classified
- ICD-10-CM K75.81: Nonalcoholic steatohepatitis (NASH)
- ICD-10-CM K75.89: Other specified inflammatory liver diseases

Unstructured search used SQL on raw patient notes to identify patients with notes containing phrases indicative of NAFLD to see how well patients could be identified from notes without full NLP assessing negation, context, etc. Text search was performed for “NAFLD”, “NASH”, “fatty liver”, “steatosis”, “steatohepatitis”, “fatty infiltration of the liver” and “fatty infiltration of liver”.

2.3.2. Comparison with manual validation through chart abstraction

We compared all approaches (NLP/text search/ICD) to manual validation using blinded manual chart review. Two physicians independently, without knowing case/control status, reviewed all records on 200 patients, 100 case patients identified as having NAFLD and 100 randomly (from a similar cohort, as explained below) selected patients identified as controls. Patients were classified as cases or controls based on clinical criteria. A clinical diagnosis of NAFLD required 1) the presence of fatty infiltration of the liver on imaging, 2) exclusion of hepatitis C infection, 3) absence of documented alcohol abuse. For controls, we randomly selected patients from a pool of patients with a similar frequency of imaging reports and progress notes to the case cohort. Factoring in report frequency alleviated the comparison with patients with few or no imaging results, as well as those with dramatically more than the average case patient. The two physician raters agreed in 95% of cases. Discordant cases were reviewed until consensus was achieved for all patients. We also estimated the prevalence of NAFLD in the Bio*Me* cohort by each approach and compared them to prevalence estimates that have been reported in similar populations [21].

2.4. NAFLD progression analyses

As a methodological proof of concept, we then sought to analyze progression time from early stage NAFLD without NASH to the point when the treating physician thought the patient to likely have progressed to NASH or cirrhosis. Any scientific evaluation of progression to NASH would require a biopsy confirming NASH, but the goal here was to test a generalizable methodology for understanding progression of a patient condition in the eyes of the physician in order to track the flow of patient knowledge between physicians. We ordered narrative documents chronologically and then stepped through them, observing document type along with the presence of NAFLD and NASH/cirrhosis to observe how references moved between note types over time along with progression. From the first note on each patient, we stepped through looking for notes with a NAFLD reference and documented whether it was a radiology note or a clinic note (a provider note or a discharge summary). Once a radiology note had been identified, we tracked how long it took for a non-radiology clinical note to reference NAFLD. If no further non-radiology notes mentioned it, we confirmed that there were additional clinical notes where NAFLD was not mentioned to ensure that lack of further reference was not due to a

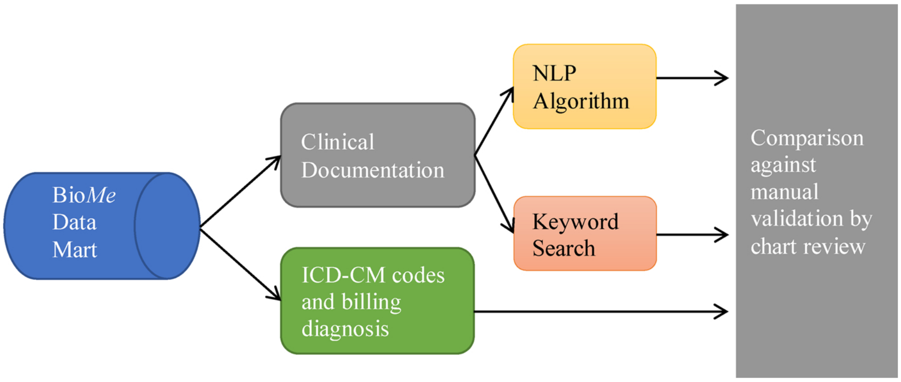


Fig. 2. Study design for data extraction and manual chart review assessing automated identification of patients with NAFLD.

discontinuation of care.

2.5. Knowledge transfer analysis

Finally, we explore the relationship between documentation on NAFLD on an imaging test, contemporaneous acknowledgment of the possible diagnosis in a clinic note (suggesting that a responsible provider acknowledged the possible diagnosis and considered further evaluation or management), and subsequent documentation of NASH or cirrhosis in a provider note or discharge summary (potentially suggesting disease progression during the observation period). With the hypothesis that at least some instances of NASH or cirrhosis could be avoided if clinicians took it more seriously, we assess how frequently NAFLD is identified but later dropped from the record. Then from this set, we identify patients who later developed NASH or cirrhosis, indicating that real damage could have occurred due to dropping the subject of NAFLD.

2.6. Statistical analyses

We calculated summary statistics to determine precision, recall, F1 (a measure of test accuracy that considers precision and recall) and F2 scores (placing more emphasis on recall and thus emphasizing false negatives more than false positives) [22]. Note that F2 is a derivative of the F1 combining PPV and Sensitivity, but weighted to prioritize PPV in scenarios like this where sensitivity is more important than specificity. We assessed ICD-9/10 codes, NLP, and text search for their ability to accurately identify NAFLD patients against the reference-standard (manual abstraction), calculating precision, recall, false positive rate, F1 and F2 scores for each method. We compared estimates of F1 and F2 scores between algorithms using the McNemar test. We used generalized score statistics to compare precision, recall and the false positive rate. All data analysis was performed in Python (v. 3.6.4 with standard packages). The significance threshold for analyses of differences was calculated as a two-sided significance p-value of < 0.05.

3. Results

We included 7,766,654 notes of 38,575 Bio*Me* enrollees from July 8, 2002 through December 31, 2017. Parsing with NLP yielded 428,469,717 post-coordinated SNOMED expressions describing clinical concepts and related context. Fig. 3 shows the queries (or query clusters, in the case of alcohol users) for the identification of case and control patient cohorts.

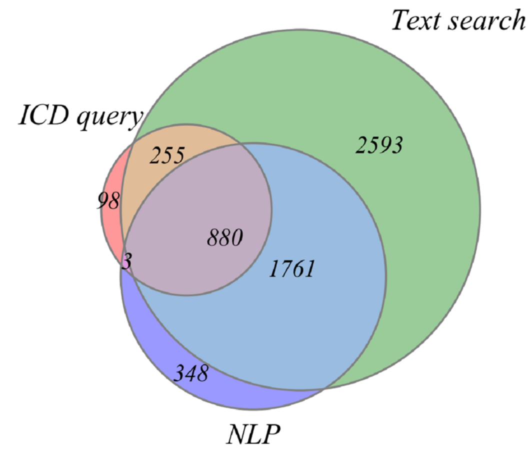


Fig. 4. Overlap of patients identified by three different approaches.

3.1. Identification of NAFLD by different approaches

We identified 2281 cases of NAFLD using NLP and 10,653 patients appropriate as controls for manual review. We also identified 1232 patients by ICD codes and 5489 patients by text search. The overlap of patients identified by each approach is shown in Fig. 4.

3.1.1. Comparison between different approaches

The summary statistics for a manual validation comparing the three patient selection methodologies on 100 cases and 100 controls are shown in Table 1.

Of the individual approaches, the NLP approach had the best summary statistics with high precision (0.89), recall (0.93) and F1 scores (0.91) and low FPR (0.11). Precision of the ICD approach was higher but recall and the F1 and F2 scores were significantly lower. The text-based approach was significantly worse than the NLP approach with respect to all parameters (p < 0.05), though the difference was far greater in specificity than sensitivity. Combining the ICD results with the NLP findings increased sensitivity (.96) no reduction in specificity (.89).

3.1.2. Baseline characteristics for NAFLD case and control cohorts

As the NLP-based approach had the best overall summary statistics, we used the cases and controls identified by this approach for further analyses. The baseline characteristics of the 2281 cases and 10,653

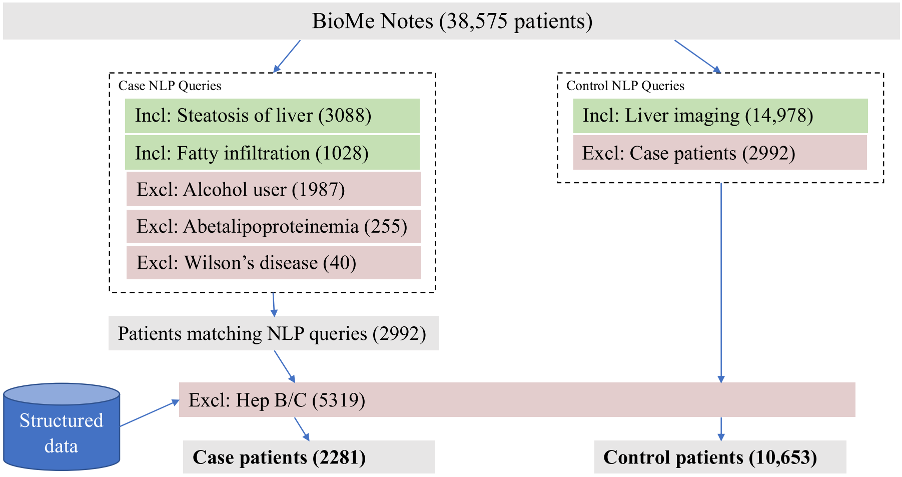

BioMe Notes (38,575 patients)
Case NLP Queries
Incl: Steatosis of liver (3088)
Incl: Fatty infiltration (1028)
Excl: Alcohol user (1987)
Excl: Abetalipoproteinemia (255)
Excl: Wilson's disease (40)
Control NLP Queries
Incl: Liver imaging (14,978)
Excl: Case patients (2992)
Patients matching NLP queries (2992)
Structured data
Excl: Hep B/C (5319)
Case patients (2281)
Control patients (10,653)

Fig. 3. NLP SNOMED queries for case and control cohorts.

Table 1
Accuracy of NLP identification of NAFLD patients relative to ICD codes and basic text search.

	NLP	ICD Selection	Text Search	NLP & ICD
Sensitivity	0.93	0.32	0.81	.96
Specificity	0.89	0.99	0.84	.89
PPV	0.89	0.97	0.83	.90
FPR	0.11	0.03	0.17	.10
F1	0.91	0.48	0.82	.93
F2	0.92	0.37	0.81	.95

control participants are shown in Table 2.

3.1.3. Reasons for misidentification of NAFLD

We analyzed reasons for false positives and false negatives of the NLP algorithm. Most false positives were due to hypothetical or otherwise uncertain references. Three examples are shown in Table 3.

The low precision and F1 score of the ICD-based approach occurred because ICD codes are more specific than sensitive. An ICD code is only applied if the physician and hospital coders believe that the diagnosis is justifiable as a billing code, but there are many reasons (often non-medical) that a problem might not be coded despite its presence. Text search had a high false-positive rate due to negations, references in templates and other references not indicative of the patient having the problem. The major limitations of each approach are explained in Table 4.

3.2. Progression from early-stage NAFLD to NASH/cirrhosis

Among 2281 patients identified as having NAFLD, 486 (21.3%) were identified as probably having NASH. Another 187 were identified as developing cirrhosis, not specifically due to NAFLD. Among the 486 documented as having both NAFLD and NASH, 310 patients had NAFL and NASH identified at the same time. For the remaining 176 where NAFLD was identified prior to NASH, the median progression time to documentation of NASH was 410 days.

3.3. Evolution analyses of information flow from radiology to clinical notes

Of the 2281 patients with NAFLD identified in notes, 619 had NAFLD identified in only radiology notes (excluding pathology), 1020 had it identified in only clinical documentation and 619 again identified it in both radiology and clinical notes. A small number of patients (23) had NAFLD identified only in pathology notes, but these references were excluded from all analyses (Fig. 5).

Of the 619 patients with NAFLD (fatty infiltration/steatosis, not NASH) identified only in radiology notes, 170 (27.5%) were later identified as likely having NASH or cirrhosis. We observed a temporal gap averaging 1057 days (range 4–4324 days). Of these, 105 were presumed to have developed NASH while the remaining 65 were described as having cirrhosis.

4. Discussion

We assessed several informatics approaches to identify NAFLD within the EHR data compared to manual validation by clinicians in a large, multiethnic cohort. Our observations suggest that NLP approaches had the best overall performance compared to ICD and text search-based approaches, though there were numerous patients identified by ICD that were missed by NLP. In addition, the prevalence of NAFLD (˜18% in those patients with imaging data) identified by the NLP-based approach was similar to population prevalence using nationally representative data, especially considering the ethnic minority predominant demographics of the Bio*Me* Biobank [4,21].

The widespread availability of EHRs in hospital systems provides an opportunity for clinical and genomic research, population health analytics, and improvement of patient care through clinical decision support [23,24]. However, appropriate use of the large-scale, granular information in EHRs depends on accurate and rapid identification of patients with the disease of interest. This is especially relevant to the study of NAFLD where the study of its natural history has been restricted by the lack of large, longitudinal cohorts with most cohorts comprising a few hundred radiological/histologically confirmed NAFLD patients [25,26]. High-throughput identification of NAFLD with “electronic” follow-up through the EHR, could aid in understanding the risk factors for progression to cirrhosis. Analysis of progression to NASH was complicated by the fact that due to the complexity of diagnosing NASH, a firm diagnosis is often not made. We believe that the methodology could be applied better to problems where diagnosis is more certain.

Previous studies have attempted to create algorithms to identify NAFLD through the EHR. A study by Corey et al. used limited NLP approaches to define NAFLD within the EHR confirmed through radiology reports in combination with ICD codes [27]. They demonstrated that the PPV (89%) and NPV (56%) was superior to an approach utilizing ICD-9 coding alone or a model incorporating AST/ALT laboratory values. However, the language parsing approach used counted only the occurrences of pre-defined terms related to NAFLD without considering critical issues in NLP including negation, context, spelling, and acronyms [28,29]. In contrast, we used an integrated NLP approach that fully accounted for these issues as applied to EHR documentation. On manual validation, we demonstrated improved summary statistics compared to not only ICD-9/10 codes, but also to simple text search. We did not attempt a formal comparison of NLP techniques, but rather assessed the viability of one particular strategy for two research applications that would be difficult without NLP. NLP performed on physicians' notes will never identify certain truths on the patient, rather a representation of what the treating physician believed to be true at the time of authoring the note. These should not be confused and hunches should be confirmed by lab tests when possible, but this is not possible when looking at data on thousands of patients in a retrospective chart review, so one must settle for the best data captured at the point of care.

While conducting analyses of the information flow between several different types of documentation where NAFLD may be identified, we found that NAFLD discovered in radiology notes was not acknowledged in clinical documentation in approximately one-half of the cases. One in ten patients who had NAFLD identified in radiology notes but which was not acknowledged in clinical documentation were later documented as having NASH without any acknowledgement of other NAFLD in the intervening clinical encounters or progress notes.

Table 2
Baseline characteristics of NAFLD cases and controls, all values: $p \le 0.01$.

	Cases (n = 2281)	Controls (n = 10,653)
Mean Age	59.8	59.5
% Male	42%	39%
Race		
African American	20%	30%
Caucasian/European	23%	22%
Asian	2%	2%
Hispanic	48%	41%
Other	6%	5%
Mean liver serology at baseline		
Aspartate Aminotransferase	45.7	39.7
Alanine Aminotransferase	48.5	35.9
Baseline Comorbidities		
Diabetes Mellitus, n (%)	1233 (40.7%)	3524 (27.6%)
Hypertension, n (%)	2079 (68.6%)	7898 (61.9%)
Mean Body Mass Index	31.8	29.2

Table 3
Review of NLP false-positives.

Problem	Example
Hypothetical reference	Nonspecific hepatic parenchymal change, such as, but not limited to, fatty liver
Mis-hyphenated word	Associated enhancing right paravertebral 5.6 cm soft issue with calcific rim at the L1/L2 level effacing the peri-hepatic fat
Uncertain Reference	Liver is nodular in contour, suggestive of cirrhosis. There is diffuse hypoattenuation of the liver likely representing hepatic steatosis

Previous studies have demonstrated that findings on radiology reports are not uniformly acknowledged or pursued by relevant providers [30,31]. This may in part reflect the deluge of biomedical information available within EHR systems, which may lead to a breakdown in information flow [32]. Although progress has been made on “closing the loop” on these non-emergent radiology findings [33], it still is an active area of both research and clinical improvement. Our research demonstrates how important findings may not be acknowledged or noticed by physicians involved in direct patient care. Further research will look at medical problems where follow-up is more critical and inaction could have detrimental consequences for the patient. If such breakdowns can be identified, it is critical that the treating physician be notified as it could constitute a medical error. How this notification best happens should be the subject of additional future work.

Past studies have demonstrated that NLP can be used to obtain valuable data for research that can be more accurate than ICD codes [9,34–36]. This study supports these findings, identifying NLP as clearly superior for individual phenotype algorithms. As data volume and accuracy are critical for big data initiatives, it stands to reason that NLP-derived features will yield superior models for these endeavors. Further research is in progress testing the use of these features in deep learning models to predict various clinical outcomes, such as the identification of NAFLD patients before the treating physician has mentioned the possibility in the notes.

Our research should be interpreted in the context of its limitations. First, we used only one NLP tool for analyses. Future work should attempt a comprehensive evaluation and comparison of different NLP software for various use cases. We anticipate that the best software would depend on the use case and setting. This approach could be accomplished with any NLP software capable of accurately mapping to and querying any clinical terminology as expressive as SNOMED CT/DL, however no other tool capable of this task was available at the time of publication. Second, we used data from only one medical center thus the applicability of our findings to other settings remains to be determined. That said, the Mount Sinai medical system has a large network of providers from different specialties and each with their own unique writing style, therefore we believe this electronic phenotyping approach can be successfully used across multiple healthcare systems [37]. Third, in the analyses of information flow from radiology to clinical documentation, it is possible that patients may have been receiving care outside of our hospital system and thus not be captured by our EHR. However, we limited this possibility by conducting a subset in which patients had at-least one clinical encounter after the radiology identification of NAFLD. Lastly, the gold standard for diagnosis of NASH is liver biopsy, however only 22 patients in our cohort had one performed. Our progression analysis therefore can only be reflective of the physician’s interpretation of the patient’s NAFL to NASH progression. From the perspective of studying actual disease progression, these methods are clearly limited by the physicians’ understanding of their patients and the availability of all patient notes.

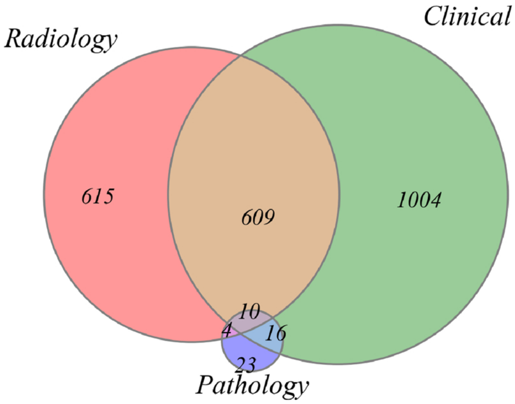


Fig. 5. Distribution of note types referencing NAFLD, counted by patient with at least one note of each type referencing NAFLD.

5. Conclusions

In summary, we demonstrate that NLP-based approaches have superior accuracy in identifying NAFLD within the EHR compared to ICD/text search-based approaches. There is lack of acknowledgement in clinical documentation of NAFLD findings in radiology reports and a significant number of these patients are later reported to have NASH. As medical practice becomes more specialized and patient care is provided by more physicians, the opportunities for information loss at patient handoffs increase. Our observations suggest that NLP-based approaches have the potential to identify dropped observations in EHR data that warrant additional follow-up.

Table 4
Strengths and limitations of patient selection modalities.

Methodology	Strengths	Weaknesses
ICD query	Very high specificity Can be standardly used	Low sensitivity, can’t differentiate between NASH and simpler NAFL
Text search	Moderately high sensitivity and specificity	Unpredictable as it depends on identifying the exact phrases used by note authors Scalability issues since manual verification of exact search strings to be used is must be performed
NLP	High sensitivity and specificity	Requires access to NLP infrastructure, especially one capable of post-coordinating SNOMED expressions

Authors' contributions

Conception and design: TTVV, LC, GNN
Analysis and interpretation: TTVV, GNN
Data collection: TTVV, LC, GNN
Writing the article: TTVV, LC, GNN
Critical revision of the article: PB, CKC, JLK, SBE, RD, RL
Final approval of the article: TTVV, GNN
Statistical analysis: TTVV, GNN
Obtained funding: JC, SGC, GNN
Overall responsibility: TTVV, LC, GNN

Funding

SGC is supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (grant no. R01DK096549 to SGC). GNN is supported by a career development award from the National Institutes of Health (NIH) (K23DK107908) and is also supported by R01DK108803, U01HG007278, U01HG009610, and 1U01DK116100-01 grants. SGC and GNN are members and are supported in part by the Chronic Kidney Disease Biomarker Consortium (U01DK106962). SGC is also supported by R01DK106085, R01HL85757, R01DK112258, and U01OH011326 grants. LC is supported in part by the NIH (5T32DK007757-18) and the recipient of a research grant from the Renal Research Institute.

Summary table

What was already known on the topic

- Notes contain deep clinical information not captured by structured data.
- NLP can be used extract data from clinical notes not captured in structured data.
- SNOMED is a highly descriptive and well organized terminology that can be leveraged for all sorts of medical logic.
- NAFLD is a growing public health problem, not always taken seriously.

What this study added to our knowledge

- Even when NAFLD is known to be present, doctors code for it infrequently. SNOMED-based NLP is good at identifying those instances where they do not.
- NAFLD is frequently not mentioned in notes even when identified by a previous physician. This can lead to the development of more serious diseases that could have been avoided.
- Using NLP, we can identify breakdowns in communication at the point of care that can lead to patient problems and/or medical errors if left unchecked.

Declaration of competing interest

TTVV was part of launching Clinithink and retains a financial interest in the company. GNN is cofounder of Renalytix AI and owns equity in that company.

Acknowledgements

The Bio*Me* healthcare delivery cohort at Mount Sinai was established and maintained with a generous gift from the Andrea and Charles Bronfman Philanthropies. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijmedinf.2019.06.028.

References

[1] M. Jiang, Y. Huang, J. Fan, et al., Parsing clinical text: how good are the state-of-the-art parsers? BMC Med. Inform. Decis. Mak. 15 (Suppl 1) (2015) S2, https://doi.org/10.1186/1472-6947-15-S1-S2.
[2] C. Friedman, G. Hripcsak, I. Shablinsky, An evaluation of natural language processing methodologies, Proc AMIA Symp (1998) 855–859.
[3] R. Loomba, A.J. Sanyal, The global NAFLD epidemic, Nat. Rev. Gastroenterol. Hepatol. 10 (2013) 686–690, https://doi.org/10.1038/nrgastro.2013.171.
[4] Z. Younossi, Q.M. Anstee, M. Marietti, et al., Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention, Nat. Rev. Gastroenterol. Hepatol. 15 (2018) 11–20, https://doi.org/10.1038/nrgastro.2017.109.
[5] N. Chalasani, Z. Younossi, J.E. Lavine, 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, Am. J. Gastroenterol. 107 (2012) 811–826, https://doi.org/10.1038/ajg.2012.128.
[6] F. Nascimbeni, R. Pais, S. Bellentani, et al., From NAFLD in clinical practice to answers from guidelines, J. Hepatol. 59 (2013) 859–871, https://doi.org/10.1016/j.jhep.2013.05.044.
[7] D. Amarapurkar, P. Kamani, N. Patel, et al., Prevalence of non-alcoholic fatty liver disease: population based study, Ann. Hepatol. 6 (2007) 161–163.
[8] P. Angulo, J.C. Keach, K.P. Batts, et al., Independent predictors of liver fibrosis in patients with nonalcoholic steatohepatitis, Hepatol Baltim Md 30 (1999) 1356–1362, https://doi.org/10.1002/hep.510300604.
[9] K.P. Liao, T. Cai, G.K. Savova, et al., Development of phenotype algorithms using electronic medical records and incorporating natural language processing, BMJ 350 (2015) h1885.
[10] J.S. Redman, Y. Natarajan, J.K. Hou, et al., Accurate identification of fatty liver disease in data warehouse utilizing natural language processing, Dig. Dis. Sci. 62 (2017) 2713–2718, https://doi.org/10.1007/s10620-017-4721-9.
[11] BioMe BioBank Program | Icahn School of Medicine. Icahn Sch. Med. Mt. Sinai. https://icahn.mssm.edu/research/ipm/programs/biome-biobank (accessed 16 May 2019).
[12] G.M. Belbin, J. Odgis, E.P. Sorokin, et al., Genetic identification of a common collagen disease in Puerto Ricans via identity-by-descent mapping in a health system, eLife 6 (2017), https://doi.org/10.7554/eLife.25060.
[13] M.A. Levin, T.T. Joseph, J.M. Jeff, et al., iGAS: a framework for using electronic intraoperative medical records for genomic discovery, J. Biomed. Inform. 67 (2017) 80–89, https://doi.org/10.1016/j.jbi.2017.02.005.
[14] M.R. Smith, B.S. Glicksberg, L. Li, et al., Loss-of-function of neuroplasticity-related genes confers risk for human neurodevelopmental disorders, Pac Symp Biocomput Pac Symp Biocomput 23 (2018) 68–79.
[15] A. Tin, G. Nadkarni, A.M. Evans, et al., Serum 6-Bromotryptophan levels identified as a risk factor for CKD progression, J Am Soc Nephrol JASN 29 (2018) 1939–1947, https://doi.org/10.1681/ASN.2017101064.
[16] M.M. Clark, A. Hildreth, S. Batalov, et al., Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation, Sci. Transl. Med. 11 (2019), https://doi.org/10.1126/scitranslmed.aat6177 eaat6177.
[17] K.A. Spackman, K.E. Campbell, R.A. Côté, SNOMED RT: a reference terminology for health care, Proc Conf Am Med Inform Assoc AMIA Fall Symp (1997) 640–644.
[18] A.M. Allen, H.K. Van Houten, L.R. Sangaralingham, et al., Healthcare cost and utilization in nonalcoholic fatty liver disease: real-world data from a large U.S. Claims database, Hepatology 68 (2018) 2230–2238, https://doi.org/10.1002/hep.30094.
[19] K.E. Corey, U. Kartoun, H. Zheng, et al., Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record, Dig. Dis. Sci. 61 (2016) 913–919, https://doi.org/10.1007/s10620-015-3952-x.
[20] V.F. Williams, S.B. Taubman, S. Stahlman, Non-alcoholic fatty liver disease (NAFLD), active component, U.S. Armed forces, 2000–2017, Mil. Health Syst. (2019) (accessed 10 May 2019), http://health.mil/News/Articles/2019/01/01/NAFLD.
[21] Devaki P. Le MH, N.B. Ha, et al., Prevalence of non-alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States, PLoS One 12 (2017), https://doi.org/10.1371/journal.pone.0173499.
[22] Y. Sasaki, The truth of the F-measure, Teach Tutor Mater 151–5 (2007) 5.
[23] O. Gottesman, H. Kuivaniemi, G. Tromp, et al., The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future, Genet Med Off J Am Coll Med Genet 15 (2013) 761–771, https://doi.org/10.1038/gim.2013.72.
[24] R.H. Miller, I. Sim, Physicians' use of electronic medical records: barriers and solutions, Health Aff. (Millwood) 23 (2004) 116–126, https://doi.org/10.1377/hlthaff.23.2.116.
[25] M. Ekstedt, L.E. Franzén, U.L. Mathiesen, et al., Long-term follow-up of patients with NAFLD and elevated liver enzymes, Hepatol Baltim Md 44 (2006) 865–873, https://doi.org/10.1002/hep.21327.
[26] S. Dam-Larsen, M. Franzmann, I.B. Andersen, et al., Long term prognosis of fatty liver: risk of chronic liver disease and death, Gut 53 (2004) 750–755.
[27] K.E. Corey, U. Kartoun, H. Zheng, et al., Development and validation of an

algorithm to identify nonalcoholic fatty liver disease in the electronic medical record, Dig. Dis. Sci. 61 (2016) 913–919, https://doi.org/10.1007/s10620-015-3952-x.

[28] K.B. Cohen, F.R. Goss, P. Zweigenbaum, et al., Translational morphosyntax: distribution of negation in clinical records and biomedical journal articles, Stud. Health Technol. Inform. 245 (2017) 346–350.

[29] S. Moon, B. McInnes, G.B. Melton, Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain, Healthc. Inform. Res. 21 (2015) 35–42, https://doi.org/10.4258/hir.2015.21.1.35.

[30] J.L. Kwan, H. Singh, Assigning responsibility to close the loop on radiology test results, Diagn Berl Ger 4 (2017) 173–177, https://doi.org/10.1515/dx-2017-0019.

[31] B. Lumbreras, L. Donat, I. Hernández-Aguado, Incidental findings in imaging diagnostic tests: a systematic review, Br. J. Radiol. 83 (2010) 276–289, https://doi.org/10.1259/bjr/98067945.

[32] N. Menachemi, T.H. Collum, Benefits and drawbacks of electronic health record systems, Risk Manag. Healthc. Policy 4 (2011) 47–55, https://doi.org/10.2147/RMHP.S12985.

[33] E.H. Dibble, D.W. Swenson, C. Cobb, et al., The RADCAT-3 system for closing the loop on important non-urgent radiology findings: a multidisciplinary system-wide approach, Emerg. Radiol. 24 (2017) 119–125, https://doi.org/10.1007/s10140-016-1452-8.

[34] E.R. McPeek Hinz, L. Bastarache, J.C. Denny, A natural language processing algorithm to define a venous thromboembolism phenotype, AMIA Annu Symp Proc AMIA Symp 2013 (2013) 975–983.

[35] C.-I. Wi, S. Sohn, M. Ali, et al., Natural language processing for asthma ascertainment in different practice settings, J. Allergy Clin. Immunol. Pract. 6 (2018) 126–131, https://doi.org/10.1016/j.jaip.2017.04.041.

[36] N. Afzal, V.P. Mallipeddi, S. Sohn, et al., Natural language processing of clinical notes for identification of critical limb ischemia, Int. J. Media Inf. Lit. 111 (2018) 83–89, https://doi.org/10.1016/j.ijmedinf.2017.12.024.

[37] K.M. Newton, P.L. Peissig, A.N. Kho, et al., Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network, J Am Med Inform Assoc JAMIA 20 (2013) e147–154, https://doi.org/10.1136/amiajnl-2012-000896.