**Title:** Highly accurate natural language algorithm for grading fatty liver disease histopathology identifies accelerating risk of advanced liver disease with later fibrosis stage

**Authors:** Marc S. Sherman MD, PhD1,2,3, Prasanna K. Challa1, Ashley T. Ott MD4, Eric M. Przybyszewski MD1, Eugenia Uche-Anya MD1, Robert Wilechansky MD1, Jessica McGoldrick1, Wolfram Goessling MD, PhD1,2,3, Hamed Khalili MD1,3, Tracey G. Simon MD1,3.

1. Division of Gastroenterology, Massachusetts General Hospital, Boston, MA, United States.
2. Division of Genetics, Brigham and Women’s Hospital, Boston, MA, United States.
3. Harvard Medical School, Boston, MA, United States.
4. Department of Medicine, Massachusetts General Hospital, Boston, MA, United States.

**Author contributions:** Conceptualization, MSS, PKC, and TGS; Methodology, MSS, PKC, and TGS; Investigation, MSS, PKC, ATO, EMP, EU, RW, JM, and TGS; Writing – Original Draft, MS and TGS; Writing –Review & Editing, MSS, PKC, ATO, EMP, EU, RW, JM, WG, HK and TGS; Funding Acquisition, WG, TGS; Resources, HK, TGS; Supervision, MS and TGS.

**Address for Correspondence:**

Tracey G. Simon

Division of Gastroenterology

55 Fruit St, Wang 5, Gastroenterology

Boston, MA 02114

[tgsimon@mgh.harvard.edu](mailto:tgsimon@mgh.harvard.edu)

**Financial support and sponsorship. \*\*\***

**Conflicts of interest:** Nothing to report.

**Keywords:** natural language processing, non-alcoholic fatty liver disease, hepatocellular carcinoma, electronic medical record, clinical epidemiology

**List of Abbreviations:** Non-alcoholic fatty liver disease (NAFLD), non-alcoholic steatohepatitis (NASH), electronic medical record (EMR), natural language processing (NLP), hepatocellular carcinoma (HCC), Mass General Brigham (MGB), Research Patient Database Registry (RPDR), non-alcoholic fatty liver (NAFL), fibrosis stage 1-4 (F1-F4).

**Abstract**

**Background & Aims:** Histopathology remains the gold standard for diagnosing and staging non-alcoholic fatty liver disease (NAFLD). The feasibility of studying NAFLD progression in electronic medical records based on histological features is limited by the free-text nature of pathology reports. Here we introduce a natural language processing (NLP) algorithm to automatically score NAFLD histology features.

**Approach & Results:** From the Mass General Brigham Healthcare system electronic medical record we identified all patients (1987-2021) with steatosis on index liver biopsy after excluding excess alcohol use and other etiologies of liver disease. An NLP algorithm was constructed in Python to detect steatosis, lobular inflammation, ballooning degeneration, and fibrosis stage from pathology free-text and manually validated in >1200 pathology reports. Patients were followed from the index biopsy to incident cirrhosis or HCC accounting for covariates. The NLP algorithm demonstrated positive and negative predictive values from 93.5-100% for all histologic concepts. Among 3.547 patients with biopsy-confirmed NAFLD followed for 23,070 person-years (PY), rates of the composite endpoint increased monotonically with worsening index NAFLD histology (p for linear trend<0.005). Compared to simple steatosis (incidence rate, 9.36/1000PY), the multivariable-adjusted hazard ratios for cirrhosis or HCC were 1.08 (0.74-1.59) for NASH/F0, 1.07 (0.8-1.41) for NASH/F1, 1.83 (1.34-2.5) for NASH/F2, and 4.45 (3.4-5.83) for NASH/F3.

**Conclusions:** The NLP algorithm accurately scores histological features of NAFLD from pathology free-text. This algorithm enabled construction of a large and high-quality NAFLD cohort across a multihospital healthcare system and disclosed an accelerating risk for cirrhosis based on index NAFLD fibrosis stage.

**Introduction**

Non-alcoholic fatty liver disease (NAFLD) and its progressive counterpart non-alcoholic steatohepatitis (NASH) represent the fastest growing clinical burden in hepatology1,2. Large cohorts are needed to accurately quantify disease incidence, estimate rates of disease progression, and identify risk factors3,4. Assembling such cohorts is increasingly tenable due to the wide availability of large electronic medical record (EMR) databases5. However, studies employing such datasets tend to estimate NAFLD diagnoses using non-invasive, non-narrative-based diagnostics, ignoring radiology and liver pathology because the free-text cannot be processed accurately in an automated fashion. However, an appropriate diagnosis of NAFLD requires either imaging or histopathology, with the latter representing the gold standard for both diagnosis and staging disease severity6. The omission of such diagnostic data fundamentally limits the granularity and specificity of conclusions that one may draw from large cohort studies.

An automated approach to analyzing the free-text describing liver biopsy results represents an appealing and powerful potential tool for increasing the fidelity and scope of EMR-based cohorts of NAFLD. While such data are often available at the center-level, and occasionally at the state-level5, their interpretation and collation for cohort construction poses significant challenges. Not all features of NAFLD or NASH are explicitly noted in every document, and there is wide inter- and intra-person stylistic variability in the reporting of each histologic feature7. Pathologists may emphasize the absence of features, or list differential diagnoses that ought to be considered or excluded. Furthermore, the currently-accepted scoring rubric for NASH was established in 20058, resulting in evolution over time in descriptions of key histopathologic features. These challenges make simple free-text matching and if-then logic inadequate for high-quality automated data extraction.

Natural language processing (NLP) is a rapidly maturing discipline of machine learning that aims to interpret human prose and offers a powerful potential solution to interpretation of pathology free-text. Early efforts demonstrated that NLP was capable of automatically scoring tumor histology grade and stage9–11 or morphology12 from surgical pathology reports for solid tumors. While accurate, these successes required niche computational linguistic expertise in sentence parsing, parts of speech labeling, feature identification, and context-specific training10,12, thus limiting their broad application. The recent development of more generalized NLP algorithms pre-trained on medical text which incorporate these prerequisite programmatic routines13–16 offers an outstanding opportunity to understand medical documentation without having to train a new algorithm for each new text corpus, and has been applied to annotate radiology17, progress15,16,18,19 and colonic anatomical pathology reports16.

In NAFLD, much of the NLP-directed efforts to date have focused on detecting NAFLD from narrative clinical documentation like progress notes18,20. However, none have developed and validated an NLP algorithm to ascertain and stage both NAFLD and NASH from liver histology free-text. Here we asked whether a pre-trained NLP algorithm could be applied to interpreting a liver biopsy report according to NASH Clinical Research Network diagnostic criteria for NAFLD/NASH and Brunt fibrosis staging21. We developed a computational infrastructure in Python to identify the diagnostic component of liver biopsy reports from our multi-hospital healthcare system, and then adapted a family of pre-trained algorithms to identify key NAFLD and NASH concepts. We manually validated more than 1200 liver biopsy reports to benchmark and establish the performance characteristics of our algorithm. We find the algorithm competes, and in some cases exceeds the fidelity of our human validators. As proof of concept, we use this tool to construct a large, system-wide cohort of patients with biopsy-proven NAFLD and estimate rates of progression to advanced liver disease and hepatocellular carcinoma (HCC).

**Methods**

*Study population and histopathology dataset*

The study population includes patients seen at a Mass General Brigham (MGB) hospital, a Boston-based regional healthcare system that includes Massachusetts General, Brigham and Women’s, Newton-Wellesley, Brigham and Women’s Faulkner, Cooley-Dickinson, Salem, Spaulding, Martha’s Vineyard, Nantucket Cottage, Wentworth-Douglass hospitals, and associated community health centers. The MGB includes tertiary-care academic medical centers, satellite teaching hospitals, and community hospitals which are primarily urban and suburban. Patient data were obtained from the Research Patient Database Registry (RPDR), a system-wide repository for all retrospective electronic healthcare records from the MGB system. This study was approved by the MGB Institutional Review Board. The RPDR was queried on 3/22/2021 for all patients between the years 1987-2021 with a pathology report from one of the MGB hospitals containing the term “hepatic” or “liver”. From each histopathology report, headers and footers containing demographic information, clinical context, and gross specimen detail were parsed for removal using custom Python code, leaving only the diagnostic text corresponding to findings and impressions or conclusions. These reports were further filtered using simple text matching criteria to identify *bona fide* liver biopsies, by excluding reports describing biopsies of other organs that also mentioned “liver” or “hepatic”, as well as liver fine needle aspirations and biopsies of hepatic tumors (Fig. 1).

*Natural Language Processing (NLP) algorithm*

The NLP algorithm was created in Python drawing from spacy14,22 and negex/negspacy23,24 NLP packages, and using the United Medical Language System corpus13. The algorithm parses the diagnostic component of a pathology report and returns paired outputs containing (1) a medical concept, and (2) the presence or absence of that feature. Importantly, the concepts were selected automatically from the curated UMLS medical dictionary13, and not manually ascribed, rendering the output an unbiased representation of the pathologist’s assessment. To interpret this raw output, keywords and text search were used to identify key NAFLD concepts among the medical concepts (output 1), and to identify biopsies where NAFLD was absent or coincident with a second disease (MS / PC \*\*\* github link).

*Definitions of NAFLD/NASH for NLP algorithm and manual validation*

NAFLD and NASH are diagnosed by the presence of specific histopathologic features. Non-alcoholic fatty liver (NAFL) or simple steatosis is defined by the presence of steatosis, or lipid droplet accumulation in the liver. NASH, the progressive form of the disease, is defined by a pattern of liver injury superimposed on steatosis which includes ballooning degeneration and lobular inflammation6,8,21. Brunt *et al* further established a staging system for fibrosis for which pericellular fibrosis in Zone 3 is defined as fibrosis Stage 1 (F1), pericellular fibrosis in Zone 3 and fibrosis in Zone 1 is defined as stage 2 (F2), bridging fibrosis is defined as stage 3 (F3), and nodule formation is defined as stage 4 (F4) and equivalent to cirrhosis25.

To adapt NAFLD feature definitions and staging to our dataset, a scoring rubric was developed by lead authors (MSS and TGS) to guide both algorithm and validators. Four core NAFLD terms were defined as present or absent: steatosis, lobular inflammation, ballooning degeneration, and NASH/steatohepatitis (as a summative term). For these terms and their synonyms, any assertion of the pathologist of any of these concepts’ presence or absence was scored as such, while omissions were recorded as absent. Descriptive diminutives like “minimal”, “patchy”, and “borderline” were scored as present. Portal inflammation and synonyms were not considered to reflect lobular inflammation, whereas lobular or Zone 3 inflammation were scored as positive for lobular inflammation. For all cohort and outcomes analyses, NASH was defined by the presence of either NLP-defined NASH being present, or the presence of NLP-defined steatosis, lobular inflammation and ballooning degeneration.

Fibrosis was scored according to the Brunt system25. The algorithm first identified explicit labeling of fibrosis stage from the report text, and if absent sought to stage based on the presence of component parts and their synonyms (e.g., if no explicit fibrosis stage stated, the presence of perisinusoidal and periportal fibrosis would be scored as F2. As per Brunt and Tiniakos25, we took the highest fibrosis stage present, even if the pathologist indicated a range, such that, for example, “F1-F2” becomes F2, and “F2 with focal areas of bridging” becomes F3.

*Structured validation of the NLP algorithm*

Validation was structured in two phases, both as double-blinded manual validation for the four NASH concepts plus the five Brunt/fibrosis stages. In phase 1, biopsies were validated from MGH, which comprised over 80% of reports. At least 100 positive examples of each of the nine concepts were generated randomly, balanced by decade 1990-2021. Five expert clinicians, blinded to the NLP algorithm interpretation, manually scored a total of 1224 reports. In parallel, another author (PC), blinded to the validation results, produced and collated the NLP output for the same reports. One coding error which led to misclassification of cirrhosis was identified after this first round of manual validation (affecting <1% of cases). This was corrected, and the final algorithm was deployed in the second phase, where again at least 100 examples of each of the 9 concepts were generated from liver biopsy reports from 4 additional hospitals within the healthcare system, each with their own pathologists: Brigham and Women’s Hospital, Faulkner Hospital, Newton-Wellesley Hospital, and North Shore Medical center.

*Analysis of algorithmic performance and algorithm-clinician disagreement*

Results from the manual interpretation and NLP interpretation for each report in the validation cohort were compared, with calculation of performance metrics, including sensitivity, specificity, negative predictive value, positive predictive value, F1 scores, and support. F1 scores are a machine learning statistic for overall algorithm fidelity and represent the harmonic mean of precision and recall10. Support grades the adequacy of concept coverage in the validation cohort and is equivalent to the count of true positives for each concept. A structured secondary analysis with blinded rescoring by an additional study member (MS) was then conducted to identify algorithm and validator error rates, in instances of algorithm-validator disagreement. A qualitative assessment was also conducted to understand patterns in algorithm failure.

*NAFLD cohort construction, exclusions, and comorbidities*

Once validated, the NLP algorithm was applied to all available liver biopsies within the MGB system to construct a comprehensive, system-wide cohort with biopsy-confirmed NAFLD between 1987 and 2021. The cohort inclusion and exclusion criteria are summarized in Figure 1, with detailed definitions in the Supplement. Briefly, among all NLP-identified simple steatosis or NASH biopsies, we excluded the presence of other liver diseases, alcohol use, exposure to steatogenic medications, history of bariatric surgery and post-transplantation status. Matched data from our original histopathology query was used to obtain patient demographics, comorbidities and outcomes indexed to 12 months prior to 6 months after the initial diagnosis. Comorbidities examined included smoking, diabetes, hypertension, BMI, and hyperlipidemia. Covariates included aspirin therapy and statin therapy (see Supplement).

*Definitions of outcomes*

The primary outcome was a diagnosis of advanced liver disease, a composite including incident cirrhosis or a liver decompensation event, including ascites, spontaneous bacterial peritonitis, bleeding esophageal or gastric varices, or hepatic encephalopathy, using validated ICD algorithms with established PPVs (see Supplement)26. In a secondary, pre-specified analysis we also ascertained incident HCC using analogous, validated ICD algorithms (Supplement).

*Outcomes statistical analysis*

The primary outcome examined was time to occurrence of cirrhosis or advanced liver disease. Index time was considered the date of the first liver biopsy meeting criteria for NAFLD/NASH without cirrhosis, and the endpoint was documented as the first occurrence of cirrhosis or advanced liver disease. Cumulative incidence curves were used to calculate incidence rates bounded by 95% confidence intervals and results were reported in events per 1000 patient years. Cox proportional hazard models were used to estimate multivariable-corrected hazard rates which accounted for age at the index date, sex, BMI, diabetes, dyslipidemia, smoking status, aspirin therapy and statin therapy. Data were prepared in Python as above, and then statistical analysis was conducted in R using previously published scripts5.

Multiple sensitivity analyses were conducted to ascertain the robustness of the data. The cirrhosis outcome was examined in isolation to determine whether the addition of clinical decompensation events altered incidence rates. Each outcome was examined excluding events occurring within 30 days after the index biopsy. Finally, we re-analyzed our data excluding patients with missing BMI to assess whether BMI missingness affects the multivariate corrected hazard rates.

**Results**

*Construction and validation of an NLP algorithm*

Our query for patients with liver biopsy reports from the RPDR database resulted in 95,150 free text pathology reports, of which 54,858 were determined to be parenchymal liver biopsies. During construction of the NLP algorithm and finalizing of the text search terms for post-NLP output filtering, the algorithm appeared to be highly effective at positively identifying cases of steatosis. We therefore used the NLP algorithm to enrich for cases where steatosis was present (Fig. 1) for validation purposes, and set aside 200 random cases for manual validation of the algorithm’s negative predictive value for the “steatosis” term. Thus, the initial filter for cohort construction and validation contained only cases where the NLP algorithm already detected reports positive for “steatosis” (Fig. 1). Subsequent filtering for alcohol use, post-transplant, other liver diseases, and excluding history of amiodarone, prednisone and methotrexate resulted in a cohort of 8,351 patients, of which 7,945 were index biopsies. The validation cohort selected from amongst this group resulted in 517 biopsy reports from MGH (phase 1 validation), 409 reports from Brigham and Women’s hospital, and 98 reports from Newton Wellesley Hospital, Wentworth Douglas Hospital and North Shore Hospital (phase 2 validation).

Agreement was high between the NLP algorithm and manual validators for all categories (Table 1). Positive and negative predictive values ranged from 93.5-100%, and >97% for 7 of 9 concepts. Of the four histological features, ballooning degeneration (PPV 100%, NPV 98.5%) and steatosis (PPV 100%, NPV 97.5%) showed the best agreement with validators, while lobular inflammation was the poorest (PPV 99%, NPV 95%). Among Brunt/fibrosis stages, only F2 showed slight disagreement (PPV 93.5%, NPV 99.4%), while the remainder of stages including an absence of fibrosis (F0) were accurately ascertained by the NLP algorithm. Although the algorithm was primarily created using MGH pathology reports as a reference, no clear failure pattern emerged when examining BWH validation or other hospital validation separately (Supplement Tables S1-S3). Finally, reports drawn from the parenchymal liver biopsy cohort which were negative for NLP-detected steatosis showed extraordinary negative predictive value (Table S4, NPV 99.5%).

Algorithm-validator disagreement, as arbitrated by a second blinded validator, highlighted algorithmic gaps and quantified algorithm and human validator error rates (Tables S5-7). Cumulatively across all validation sets with algorithm-validator disagreement, the algorithm was correct in 36% of cases, the first validator was correct in 62% of cases, and in 2% of cases all three disagreed. The overall apparent algorithmic error rate across all categories was 1.7%, while the overall human error rate across all categories was 1.1%. Interestingly, errors occurred at a rate of ~4:1 favoring the validators in the first 517 cases (Table S6), but occurred at a rate 1:1 in the second 507 cases (Table S7). Most algorithm-validator disagreements originated from the lobular inflammation category (n=49, 28%) and fibrosis staging (n=60, 34%). On secondary validator review, this subjectively seemed to derive from more ways to describe both lobular inflammation and the nature of fibrosis, whereas steatosis and ballooning notation was generally binary and with few synonyms.

*Creation of a large, high-quality NAFLD cohort from histopathology free text*

We then applied this highly accurate, validated NLP algorithm to construct a cross-sectional NAFLD cohort of patients across our entire healthcare system (Table 2). The algorithm-generated cohort identified a steady rise of AST and ALT, with ALT>AST across fibrosis stage accompanied by a relative decline with the onset of cirrhosis. Cirrhosis was marked by an uptick in PT-INR, total bilirubin, and a reduction in BMI and hypertension. Dyslipidemia was markedly reduced at the onset of cirrhosis.

*Progression to cirrhosis or advanced liver disease by NAFLD/NASH fibrosis stage*

After narrowing the cross-sectional cohort (Table 2) to include only index biopsies, excluding patients with inadequate follow up (Fig. 1), and excluding patients with liver biopsies performed at the time of bariatric surgery (n=3,152), 3,547 patients were included in the risk analysis with baseline characteristics in Table 3. Among patients with biopsy-proven NAFLD at index liver biopsy, 35.8% had pre-cirrhotic NASH. Of note, (\*\*\* TGS: anything else to note about our baseline demographics? Race, gender, things that are reassuring or provocative?).

In our risk analysis, we documented 524 cases of cirrhosis or advanced liver disease (23,053 PY). Risk of cirrhosis or advanced liver disease increased from steatosis to NASH/F0, through F3 (trend, p<.005, Table 4). Hazard rates for development of the primary outcome did not significantly differ from steatosis until stage F2, though F1 trended toward significance (p=0.18). Risk of cirrhosis nearly doubled from F0-F1 to F2, and more than doubled from F2 to F3. Absolute incidence rates were ~20-30% lower by excluding outcomes occurring within 30 days of index biopsy, without changing the overall pattern, however, hazard rates were not materially changed in this sensitivity analysis. Cumulative incidence curves highlight that outcomes are only modestly different for steatosis through NASH/F1, but clearly stratify with increasing occurrence of the primary outcome for F2 and F3 (Fig. 2).

Because of the heterogeneous nature of the primary endpoint which included biopsy-proven, NLP-detected cirrhosis, validated ICD9/10 coded cirrhosis, or evidence of decompensated cirrhosis, we also examined cirrhosis in isolation. Here, the incidence rates of NASH/F0 were not significantly changed compared to the combined primary endpoint analysis, except steatosis which was more definitively less severe than NASH F0 and NASH F1. This did not reach statistical significance (Table S\*), however, in sensitivity analysis further excluding events occurring within 30 days of index biopsy, NASH/F1 reached statistical significance for increased risk of cirrhosis as compared to steatosis.

**Discussion**

Histopathology remains the gold standard in the diagnosis and staging of human NAFLD6,8,21. Despite histopathology reports being widely available in center-level databases, these reports remain a largely untapped resource due to their free text nature. This study establishes that an NLP algorithm broadly trained on medical corpuses, without training, can automatically detect pathologist-described liver histopathology with performance characteristics that rival human validators, the traditional gold standard. Deployment of NLP has the potential to save time, reduce errors, and enable large-scale analysis of pathology data that would be infeasible manually.

As proof of concept, we deployed the algorithm to automatically score liver biopsies across the MGB, a multi-center hospital system, to create two large NAFLD cohorts. The first cohort, which included more than 8000 patients, provided a cross-sectional view of NAFLD patients at the time of their liver biopsy. Cohort characteristics matched known clinical phenomena in human steatosis and NASH, including steadily rising transaminases (ALT>AST) across fibrosis stage before declining in the setting of cirrhosis, and PT-INR which becomes abnormal at the cirrhosis stage. These data also confirm a reduction in dyslipidemia at the onset of cirrhosis, as previously described27,28.

The second cohort examined rates of progression of NAFLD to cirrhosis or end-stage liver disease. To our knowledge, this is the largest study of disease progression stratified by Brunt stage fibrosis status yet conducted, and offers patients and clinicians a more precise prognosis based on index liver biopsy staging. We find an accelerating risk of progression to cirrhosis with increasing fibrosis stage, mirroring and validating the results of a meta-analysis (n=1,495) which found a similar accelerating trend in mortality for patients with later fibrosis stages29. Other studies examined fibrosis progression by pooling paired liver biopsy studies30 or analyzing the placebo arms of clinical trials31, and estimated rates of 0.14 (n=116) or 0.03 (n=952) fibrosis stages per year, respectively. We find an incidence rate of 0.083 (0.072-0.10, 95% CI, Table 4) for single-stage advancement of F3 to F4, which is comparable to prior estimates (n=3,547). As an exercise for comparison purposes, and assuming linear progression by standardizing our rates on the scale of a single F-stage yields 0.016 per stage (F2), 0.006 per stage (F1), 0.0038 per stage (F0), and 0.0035 per stage (simple steatosis). Prior studies were not stratified by fibrosis stage, but single stage progression of patients with simple steatosis to F1 was previously estimated to be 0.0530, an order of magnitude higher than estimated here.

The stratification of risk by fibrosis stage has significant implications for the study of human NAFLD and NASH. The apparent disease acceleration at higher fibrosis stages, mirrored by accelerated mortality rates29, suggests clinical trials may benefit from focusing on patients with F3 fibrosis. Not only are patients with F3 at greatest risk of progression to cirrhosis, but their accelerated risk indicates fewer patients would need to be followed, and for less time, to achieve meaningful endpoints. In contrast, our NASH F0 absolute incidence rate of 15.18 per 1000 patient years corresponds to an average time-to-event of 66 years. Coupled with a mean age of biopsy for NASH/F0 being 47 years, patients with NASH/F0 are on average unlikely to progress to cirrhosis in their natural lifespan. Our analysis indicates the same may be true of both simple steatosis and even NASH/F1, and may explain the failure of many clinical trials to achieve significant endpoints. Our observation lends credibility to the theory of rapid progressors, a notion raised by paired biopsy studies in which ~14-20% of patients had >1 fibrosis stage progression in short interval paired biopsies30. While numerous biomarkers for NASH exist, additional scrutiny is needed to differentiate rapid progressors, especially among patients with early stages of disease. Finally, if confirmed at other centers, our results may revise epidemiologic projections of NASH burden on liver transplantation.

There are some limitations to this study. Our biopsy report dataset is intrinsically heterogeneous, covering more than three decades during which NAFLD definitions shifted8, and representing the assessments of several dozen pathologists. In part to address stylistic heterogeneity, the algorithm was also necessarily reductive, for example simplifying equivocal histologic features as positives, and reducing steatosis grade to a binary presence or absence even when grade was reported. Whether these drawbacks are offset by the large number of cases studied remains to be seen. In addition, our estimated single-stage rates may be lower than prior estimates because our primary outcome includes evidence of decompensated cirrhosis, whereas paired biopsy studies were more likely to capture evolution to compensated cirrhosis. This limitation is somewhat addressed in our sensitivity analysis by examining cirrhosis alone, which enriched for biopsy-confirmed cirrhosis and did not significantly alter incidence rates. Future analyses where our NLP algorithm is replicated at other centers may enable sufficient sample size to conduct the more definitive paired biopsy study analysis. Related, though we used standardized ICD codes to define our outcomes, this is an inherently less precise outcome than those defined by biopsy or radiology. As an example of uncertainty intrinsic to ICD-code based outcome ascertainment, our HCC cases were halved by excluding diagnoses within 30 days of biopsy, the majority of which derived from patients with a biopsy showing steatosis. Anecdotally these cases appear to be colorectal metastases, but for which the treating physician coded for HCC as the indication for liver biopsy. Our work offers extreme precision in the definition of our exposure, liver biopsy histologic status, but further attention is needed to more accurately annotate clinical endpoints. Finally, this study encompasses multiple hospitals but within a single hospital system in the Northeast; future studies encompassing a more representative sampling of national geographic demographics may further refine our incidence rate estimates.

In conclusion, open source NLP algorithms can be deployed, without training, to score NAFLD histopathology according to field-standard definitions. This automation proved reliable and helpful in the creation of high quality NAFLD/NASH cohorts and may be easily adapted to any free-text medical document. Our study identified an accelerating trend for developing cirrhosis by index fibrosis stage, suggesting that early stage patients are on average unlikely to progress to cirrhosis in their natural life span. We anticipate our specific NAFLD pathology NLP algorithm, and the overall methodology, to continue to disclose new insights into this important and growing disease.

1. Younossi, Z. M. *et al.* Epidemiology of chronic liver diseases in the USA in the past three decades. *Gut* **69**, 564–568 (2020).

2. Flemming, J. A., Djerboua, M., Groome, P. A., Booth, C. M. & Terrault, N. A. NAFLD and Alcohol-Associated Liver Disease Will Be Responsible for Almost All New Diagnoses of Cirrhosis in Canada by 2040. *Hepatology* **74**, 3330–3344 (2021).

3. Simon, T. G. *et al.* Association of Aspirin with Hepatocellular Carcinoma and Liver-Related Mortality. *N. Engl. J. Med.* **382**, 1018–1028 (2020).

4. Simon, T. G. *et al.* Daily Aspirin Use Associated With Reduced Risk For Fibrosis Progression In Patients With Nonalcoholic Fatty Liver Disease. *Clin. Gastroenterol. Hepatol.* **17**, 2776-2784.e4 (2019).

5. Simon, T. G., Roelstraete, B., Khalili, H., Hagström, H. & Ludvigsson, J. F. Mortality in biopsy-confirmed nonalcoholic fatty liver disease: results from a nationwide cohort. *Gut* **70**, 1375–1382 (2021).

6. Chalasani, N. *et al.* The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases: Hepatology, Vol. XX, No. X, 2017. *Hepatology* **67**, 328–357 (2018).

7. Brunt, E. M. *et al.* Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. *J. Hepatol.* **76**, 1030–1041 (2022).

8. Kleiner, D. E. *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).

9. Abedian, S. *et al.* Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports. *JCO Clin. Cancer Inform.* 1054–1061 (2021) doi:10.1200/CCI.21.00065.

10. Coden, A. *et al.* Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *J. Biomed. Inform.* **42**, 937–949 (2009).

11. Currie\*, A.-M. *et al.* Automated Extraction of Free-Text from Pathology Reports. *AMIA. Annu. Symp. Proc.* **2006**, 899 (2006).

12. Hammami, L. *et al.* Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. *J. Biomed. Inform.* **116**, 103712 (2021).

13. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).

14. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. (2020).

15. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications | Journal of the American Medical Informatics Association | Oxford Academic. https://academic.oup.com/jamia/article/17/5/507/830823.

16. Imler, T. D., Morea, J., Kahi, C. & Imperiale, T. F. Natural Language Processing Accurately Categorizes Findings From Colonoscopy and Pathology Reports. *Clin. Gastroenterol. Hepatol.* **11**, 689–694 (2013).

17. Yuan, J., Zhu, H. & Tahmasebi, A. Classification of Pulmonary Nodular Findings based on Characterization of Change using Radiology Reports. *AMIA Summits Transl. Sci. Proc.* **2019**, 285–294 (2019).

18. Corey, K. E., Kartoun, U., Zheng, H. & Shaw, S. Y. Development and Validation of an Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record. *Dig. Dis. Sci.* **61**, 913–919 (2016).

19. Van Vleck, T. T. *et al.* 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. *Int. J. Med. Inf.* **129**, 334–341 (2019).

20. Redman, J. S. *et al.* Accurate Identification of Fatty Liver Disease in Data Warehouse Utilizing Natural Language Processing. *Dig. Dis. Sci.* **62**, 2713–2718 (2017).

21. Brunt, E. M. *et al.* Nonalcoholic fatty liver disease (NAFLD) activity score and the histopathologic diagnosis in NAFLD: distinct clinicopathologic meanings. *Hepatology* **53**, 810–820 (2011).

22. Honnibal, M. & Johnson, M. An Improved Non-monotonic Transition System for Dependency Parsing. in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* 1373–1378 (Association for Computational Linguistics, 2015). doi:10.18653/v1/D15-1162.

23. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *J. Biomed. Inform.* **34**, 301–310 (2001).

24. Chapman, W. W. *et al.* Extending the NegEx Lexicon for Multiple Languages. *Stud. Health Technol. Inform.* **192**, 677–681 (2013).

25. Brunt, E. M. & Tiniakos, D. G. PATHOLOGICAL FEATURES OF NASH. 10.

26. Hagström, H. *et al.* Administrative Coding in Electronic Health Care Record-Based Research of NAFLD: An Expert Panel Consensus Statement. *Hepatology* **74**, 474–482 (2021).

27. Cicognani, C. *et al.* Serum Lipid and Lipoprotein Patterns in Patients With Liver Cirrhosis and Chronic Active Hepatitis. *Arch. Intern. Med.* **157**, 792–796 (1997).

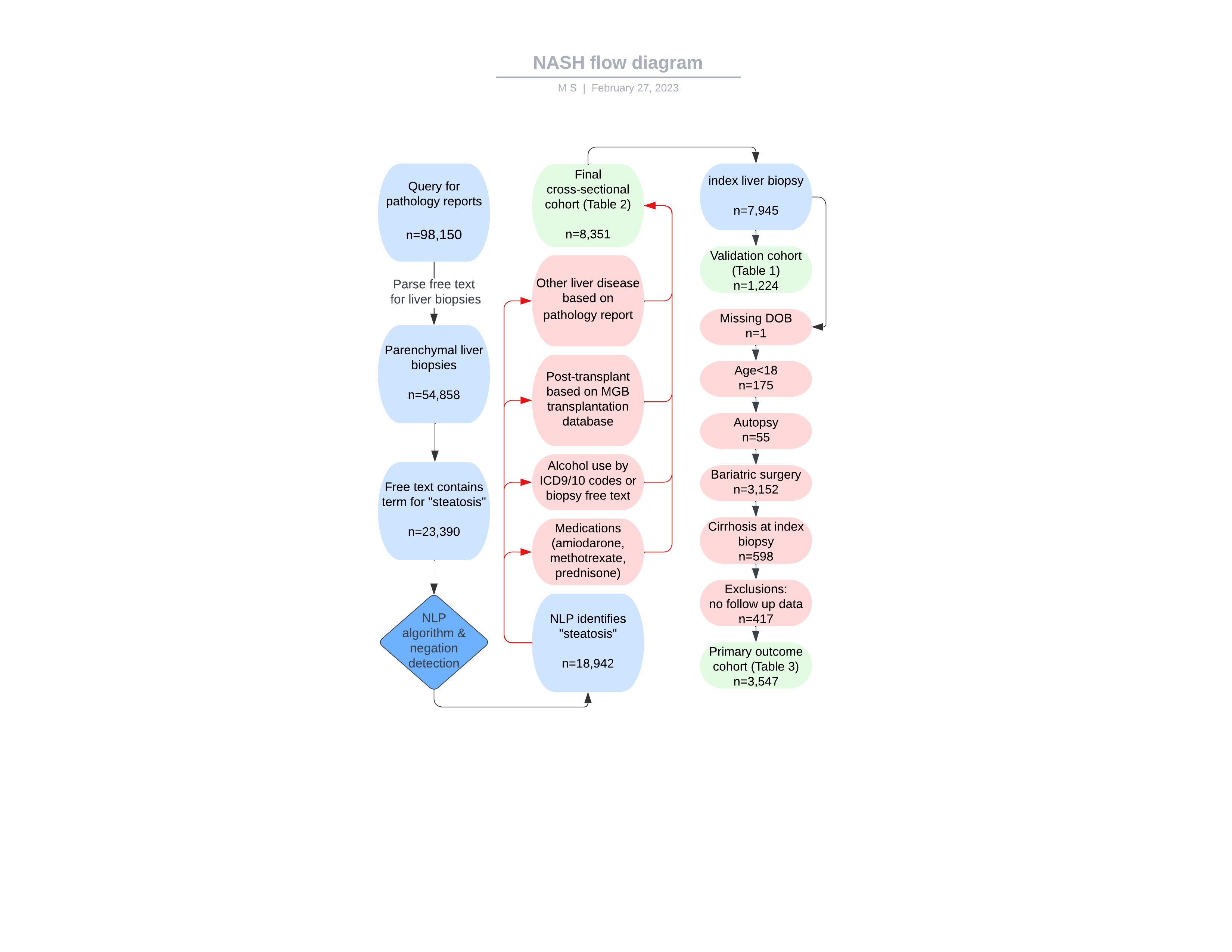
28. A. D’ARIENZO, F. M., G. SCAGLIONE, G. VICINANZA, R. BENNATO, G. MAZZACCA. Prognostic Value of Progressive Decrease in Serum Cholesterol in Predicting Survival in Child-Pugh C Viral Cirrhosis. *Scand. J. Gastroenterol.* **33**, 1213–1218 (1998).

29. Dulai, P. S. *et al.* Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: Systematic review and meta-analysis. *Hepatology* **65**, 1557–1565 (2017).

30. Singh, S. *et al.* Fibrosis Progression in Nonalcoholic Fatty Liver vs Nonalcoholic Steatohepatitis: A Systematic Review and Meta-analysis of Paired-Biopsy Studies. *Clin. Gastroenterol. Hepatol.* **13**, 643-654.e9 (2015).

31. Roskilly, A. *et al.* Fibrosis progression rate in a systematic review of placebo‐treated nonalcoholic steatohepatitis. *Liver Int.* **41**, 982–995 (2021).

Figures.



**Figure 1. Flow diagram for algorithm and cohort construction.** MGB-system wide pathology reports were requested and screened for parenchymal liver biopsies. After excluding other liver diseases and validating the NLP algorithm, a cross sectional cohort was created capturing patient characteristics at time of liver biopsy, and a second cohort was followed from index biopsy to incident cirrhosis or HCC.



**Figure 2. Cumulative incidence of cirrhosis based on fibrosis/NAFLD stage at index biopsy.**

**Tables.**

**Table 1. Performance characteristics of NLP algorithm for prediction of NAFLD concepts**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Concept | n | Sensitivity/  Recall % | Specificity % | PPV/  Precision% | NPV % | F1 Score | Support |
| “steatosis” | 1224 | 99.9 | 98.0 | 99.6 | 99.5 | 99.8 | 1021 |
| “lobular inflammation” | 1024 | 90.9 | 99.6 | 99.0 | 95.9 | 94.8 | 328 |
| “ballooning degeneration” | 1024 | 96.4 | 100.0 | 100.0 | 98.5 | 98.2 | 304 |
| “NASH” | 1024 | 98.6 | 97.8 | 97.1 | 99.0 | 97.9 | 444 |
| “cirrhosis” | 1024 | 98.5 | 99.4 | 97.6 | 99.6 | 98.0 | 203 |
| Brunt/Fibrosis F0 | 1024 | 96.0 | 99.2 | 97.3 | 98.9 | 96.6 | 223 |
| Brunt/Fibrosis F1 | 1024 | 93.5 | 99.0 | 96.2 | 98.3 | 94.8 | 216 |
| Brunt/Fibrosis F2 | 1024 | 97.4 | 98.4 | 93.5 | 99.4 | 95.4 | 193 |
| Brunt/Fibrosis F3 | 1024 | 98.4 | 99.5 | 97.9 | 99.6 | 98.2 | 191 |
| Brunt/Fibrosis F4 | 1024 | 98.0 | 99.3 | 97.0 | 99.5 | 97.5 | 198 |

**Table 2.** **Cross-sectional cohort characteristics for all NAFLD liver biopsies in the Mass General Brigham system 1995-2021**



Abbreviations: IQR – interquartile range; SD – standard deviation; BMI – body mass index; ALT – alanine aminotransferase; AST – aspartate aminotransferase

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Characteristic** | **All NAFLD**  *N=3547* | **Simple Steatosis**  *N=1122* | **NASH/F0**  *N=369* | **NASH/F1**  *N=934* | **NASH/F2**  *N=382* | **NASH/F3**  *N=327* |
| Female, % | 53.76 | 57.4 | 52.85 | 48.93 | 52.09 | 53.52 |
| Age at the index date, years (SD) | 51.13 (13.88) | 51.99 (14.27) | 46.95 (13.95) | 50.36 (13.5) | 51.09 (13.77) | 54.6 (12.43) |
| Race/ethnicity |  |  |  |  |  |  |
| • Asian | 4.71 | 2.76 | 6.23 | 6.75 | 3.66 | 3.67 |
| • Black | 3.27 | 3.57 | 3.79 | 3.1 | 2.88 | 2.45 |
| • Hispanic | 4.76 | 3.3 | 7.32 | 5.89 | 4.45 | 3.98 |
| • White | 75.98 | 80.04 | 68.02 | 73.88 | 76.44 | 78.9 |
| • Other | 4.62 | 2.58 | 6.78 | 5.03 | 6.54 | 4.89 |
| • Not recorded | 6.54 | 7.75 | 7.32 | 5.25 | 5.76 | 6.12 |
| Years of follow-up\*, median (IQR) | 4.23 (9.16) | 3.56 (10.88) | 5.49 (8.54) | 4.97 (9.1) | 4.76 (9.42) | 2.78 (6.98) |
| Start of follow-up, % |  |  |  |  |  |  |
| * 1987 - 1999 | 13.53 | 18.89 | 6.78 | 11.78 | 8.9 | 11.31 |
| * 2000 - 2009 | 30.9 | 34.22 | 30.89 | 30.3 | 31.15 | 29.36 |
| * 2010 - 2021 | 55.57 | 46.88 | 62.33 | 57.92 | 59.95 | 59.33 |
| Hospital, % |  |  |  |  |  |  |
| * Massachusetts General Hospital | 66.54 | 55.79 | 79.95 | 70.77 | 69.9 | 67.89 |
| * Brigham and Women’s Hospital | 28.47 | 39.04 | 16.8 | 24.2 | 24.08 | 23.24 |
| * Faulkner Hospital | 1.95 | 1.43 | 2.17 | 2.03 | 2.88 | 3.67 |
| * Newton-Wellesley Hospital | 1.55 | 2.5 | 0.27 | 1.39 | 2.09 | 1.53 |
| * North Shore Medical Center | 1.49 | 1.25 | 0.81 | 1.61 | 1.05 | 3.67 |
| BMI, % |  |  |  |  |  |  |
| * < 30 | 21.96 | 24.96 | 20.6 | 21.84 | 15.71 | 16.82 |
| * >=30 | 26.92 | 17.11 | 36.86 | 30.62 | 35.08 | 33.33 |
| * Missing | 51.11 | 57.93 | 42.55 | 47.54 | 49.21 | 49.85 |
| Diabetes, % | 18.3 | 14.62 | 12.47 | 20.34 | 21.47 | 33.03 |
| Dyslipidemia, % | 28.08 | 22.73 | 32.25 | 31.91 | 30.1 | 33.64 |
| Hypertension, % | 31.32 | 27.18 | 32.25 | 33.4 | 32.72 | 37.61 |
| Smoking Status |  |  |  |  |  |  |
| * Never | 22.41 | 18.0 | 26.29 | 23.77 | 26.18 | 22.02 |
| * Ever | 14.52 | 12.92 | 15.45 | 15.2 | 14.4 | 17.13 |
| * Missing | 63.07 | 69.07 | 58.27 | 61.03 | 59.42 | 60.86 |
| Aspirin therapy, % | 13.96 | 10.96 | 13.55 | 14.99 | 15.97 | 20.8 |
| Statin therapy, % | 22.36 | 19.52 | 21.41 | 23.77 | 24.87 | 26.61 |

­­­**Table 3.** **Longitudinal cohort characteristics for patients with non-cirrhotic NAFLD on index liver biopsy**

\*Years of follow-up: Considered any advanced liver disease, cirrhosis, or HCC

**Table 4.** **Longitudinal cohort risk analysis for primary endpoint (cirrhosis or advanced liver disease)**

