# Rapid Identification and Phenotyping of Nonalcoholic Fatty Liver Disease Patients Using an Automated Algorithmic Approach in Diverse, Urban Healthcare Systems

**Short title:** Rapid Risk Assessment in NAFLD: Targeting Care Using a System-Wide Automated Approach

Anna O. Basile[1,2], Anurag Verma[3], Leigh Anne Tang[4], Marina Serper[5], Andrew Scanga[6], Ava Farrell[7], Brittney Destin[7], Rotonya M. Carr[5], Anuli Anyanwu-Ofili[8], Gunaretnam Rajagopal[8], Abraham Krikhely[7], Marc Bessler[7], Muredach P. Reilly[9,10], Marylyn D. Ritchie[3], Joshua Denny[11], Nicholas P. Tatonetti[1*], Julia Wattacheril[12*]

**Affiliations:**
[1] Department of Biomedical Informatics Columbia University NY, NY
[2] New York Genome Center NY, NY
[3] Department of Genetics and Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[4] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
[5] Division of Gastroenterology and Hepatology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[6] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[7] Department of Medicine, Center for Liver Disease and Transplantation, Columbia University Irving Medical Center NY, NY
[8] Janssen Pharmaceutical Spring House, PA
[9] Irving Institute for Clinical and Translational Research, Columbia University, New York, NY 10032, USA
[10] Division of Cardiology, Department of Medicine, Columbia University Irving Medical Center, NY, NY
[11] *All of Us* Research Program, National Institutes of Health, Bethesda, MD
[12] Division of Digestive and Liver Diseases, Department of Medicine, Center for Liver Disease and Transplantation, Columbia University Irving Medical Center, NY, NY

**Grant Support:**
Funding was provided by Janssen Research and Development in collaboration with Columbia University Irving Medical Center. The sponsor was involved in study concept and design.

This publication was supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant Number UL1TR001873. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Health (NIH).

JD's involvement in this project was as faculty at Vanderbilt University Medical Center prior to joining the NIH.

NPT is supported by R35GM131905.

**Abbreviations:**

EHR Electronic Health Record, NAFLD nonalcoholic fatty liver disease, NASH nonalcoholic steatohepatitis, OMOP Observational Medical Outcomes Partnership

*** Correspondence:**
Julia Wattacheril, MD, MPH
Director, Nonalcoholic Fatty Liver Disease Program
Associate Professor of Medicine
Center for Liver Disease and Transplantation
Columbia University - NY Presbyterian Hospital
622 West 168th Street, PH 14 105-D
New York, NY 10032
212.305.0660 TEL
212.305.9139 FAX
jjw2151@cumc.columbia.edu

Nicholas P. Tatonetti, PhD
Department of Biomedical Informatics
Department of Systems Biology
Department of Medicine
Herbert Irving Comprehensive Cancer Center
Institute for Genomic Medicine
622 West 168th Street
New York, NY 10032
212.305.9104 TEL
nick.tatonetti@columbia.edu

**Disclosures:**

Patent for algorithm to Columbia University Trustees; © 2021 The Trustees of Columbia University in the City of New York. The owner has no objection to reproduction of the work for academic non-commercial purposes, but otherwise reserves all copyright rights whatsoever.

JW, NPT and AOB are co-inventors.

JW has received research support from Janssen, Galectin, Intercept, Genfit, Shire, Conatus, Zydus, and is on the advisory board for Astra Zeneca/MedImmune, AMRA.

RMC has received research support from Intercept Pharmaceuticals and Merck, Inc.

MS is a consultant for Gilead, Inc.

AK is a speaker and proctor for Intuitive, reviewer for surgical videos for Crowd Sourced Assessment of Technical Skills (CSATs), and a consultant for Johnson and Johnson and Surgical Specialties Corporation.

AV, LT, AS, AF, BD, AA, GR, AK, MB, MPR, MDR, JD, and NPT have nothing to disclose.

**Author Contributions:**

**Anna O. Basile:** drafting and reviewing of the manuscript; critical revision of the manuscript for important intellectual content; statistical analysis; analysis and interpretation of data; study concept and design

**Anurag Verma:** analysis and interpretation of data; acquisition of data; drafting of the manuscript

**Leigh Anne Tang:** analysis and interpretation of data; acquisition of data; drafting of the manuscript

**Marina Serper**: acquisition of data; drafting of the manuscript

**Andrew Scanga**: acquisition of data; drafting of the manuscript

**Ava Farrell**: acquisition of data; drafting of the manuscript

**Brittney Destin**: acquisition of data; drafting of the manuscript

**Rotonya M. Carr**: drafting of the manuscript; acquisition of data

**Anuli Anyanwu-Ofili**: obtained funding; drafting of the manuscript

**Gunaretnam Rajagopal**: obtained funding; drafting of the manuscript; study concept and design

**Abraham Krikhely**: acquisition of data; drafting of the manuscript

**Marc Bessler**: acquisition of data; drafting of the manuscript

**Muredach P. Reilly**: critical revision of the manuscript for important intellectual content; obtained funding

**Marylyn D. Ritchie**: drafting and critical revision of the manuscript for important intellectual content

**Joshua Denny**: drafting and critical revision of the manuscript for important intellectual content

**Nicholas P. Tatonetti**: drafting and critical revision of the manuscript for important intellectual content; statistical analysis; analysis and interpretation of data; study concept and design

**Julia Wattacheril**: study concept and design; acquisition of data; analysis and interpretation of data; drafting of the manuscript; critical revision of the manuscript for important intellectual content; obtained funding; technical, or material support; study supervision

**Data Transparency Statement:**

Algorithmic code is available for academic, non-commercial collaborations by request to the corresponding authors.

**What You Need to Know:**

**Background and Context**: NAFLD is the leading form of liver disease worldwide with a rising prevalence in the population. Current means of identification are complex and dependent on provider recognition of clinical risk factors.

**New Findings**: We present an accurate (mean PPV=84%) and cross-institution validated, rule-based algorithm for the high-throughput, rapid identification of NAFLD patients across diverse EHR systems comprising approximately 12.1 million patients. The majority of patients were previously unidentified.

**Limitations**: Inaccessible imaging and histologic data (performed outside the healthcare system) limited our ability to verify hepatic steatosis and resulted in low sensitivity for the final step of the algorithm.

**Impact**: Our NAFLD algorithm provides an accurate means of rapidly identifying NAFLD in large EHR systems to target patients at greatest risk for disease progression and clinical outcomes towards diagnostic and therapeutic interventions.

## Short Summary

NAFLD, the leading cause of liver disease globally, is often under-recognized in at-risk individuals. Here we present a rapid, non-invasive algorithm for identifying patients within large health systems who are at greatest risk for disease progression and clinical decompensation for diagnostic and therapeutic intervention.

## Abstract

**Background** and Aims**:** Nonalcoholic Fatty Liver Disease (NAFLD) is the most common global cause of chronic liver disease. Therapeutic interventions are rapidly advancing for its inflammatory phenotype, nonalcoholic steatohepatitis (NASH). Diagnosis codes alone fail to accurately recognize at-risk patients. The objective of the present work is to identify NAFLD patients within large electronic health record (EHR) databases for targeted intervention based on clinically relevant phenotypes.

**Methods:** We present a rule-based phenotype algorithm for the rapid identification of NAFLD patients developed using EHRs from 5.8 million adult patients at Columbia University Irving Medical Center (CUIMC). The algorithm was developed using the Observational Medical Outcomes Partnership (OMOP) Common Data Model, and queries multiple structured and unstructured data elements, including diagnosis codes, laboratory measurements, radiology and pathology modalities.

**Results:** Our approach identified 16,060 CUIMC NAFLD patients with 170 having a biopsy-proven NASH diagnosis. Fibrosis scoring on patients without histology identified 943 with scores indicative of advanced fibrosis (FIB-4, APRI, NAFLD) in ≥2 of the scoring metrics. The algorithm

was validated at two independent healthcare systems, University of Pennsylvania Healthcare System (UPHS) and Vanderbilt Medical Center (VUMC), where 20,779 and 19,575 NAFLD patients were identified, respectively. Clinical chart review identified a high positive predictive value (PPV) for the algorithm across all healthcare systems: 91% at CUIMC, 75% at UPHS, and 85% at VUMC.

**Conclusions:** Our rule-based algorithm provides an accurate, automated approach for rapidly identifying and sub-phenotyping NAFLD patients within a large EHR system. This highlights the clinical potential algorithms have in discovering NAFLD patients at highest risk for disease progression for diagnostic and therapeutic intervention.

Keywords = NAFLD, NASH, phenotype algorithm, OMOP, computational, automation

## Introduction

Non-alcoholic fatty liver disease (NAFLD) is the most common form of chronic liver disease worldwide. NAFLD affects approximately 25-30% of the general adult population in industrialized countries[1]. While NAFLD, along with its inflammatory phenotype non-alcoholic steatohepatitis or NASH, is a chronic liver disease with rising incidence, it is often under-diagnosed[2] due to the cost and invasiveness of liver biopsy, the current gold standard of diagnosis. Identifying NAFLD patients is critically important to effective healthcare delivery, from preventative measures for diabetes to targeted diagnostics, specialist referral, and intervention for longitudinal assessment and treatment. This is particularly important given disease model projections of a doubling or tripling of end-stage liver disease patients by 2030 in many parts of the world[3]. Thus, prioritizing preventative care for groups at high risk of progression, such as those exhibiting the inflammatory NAFLD phenotype, non-alcoholic steatohepatitis (NASH), is crucial.

Emerging therapies for NASH will be limited in application if at-risk individuals remain difficult to identify in health care systems. Unfortunately, given current limitations inherent in diagnostic coding for this disease, the rapid identification of patients with NAFLD is problematic. Diagnostic advances in circulating and imaging biomarkers can assist in identifying patients, particularly those with advanced fibrosis and NASH. Electronic health record (EHR) phenotyping is another means by which patients can be targeted for diagnosis. Applied serially, this approach may also identify patients at risk for disease progression despite maximal medical and surgical therapy. EHRs and claims databases are convenient sources of large patient populations and can provide the data needed to phenotypically identify patients. EHR data are collected prospectively in a large-scale, long-term follow-up manner[4]. These properties, along with the inclusion of diverse aspects of patients' health-related information, make EHRs a valuable data source for constructing targeted intervention based on clinically actionable phenotypes. However, EHRs are limited by the completeness and accuracy of data which may have confounding effects if not properly addressed in the study design[5–7]. One approach in addressing these inaccuracies is to use a wide range of different data sources available in the EHR (including structured data, such as diagnosis/billing codes and laboratory measures, as well as unstructured elements such as imaging/radiology reports and provider notes) as a means of diagnostic confirmation. Additionally, quality control parameters can be implemented to reduce false-positive identifications.

Various approaches for identifying NAFLD patients using EHRs have been previously described[8–10]. These approaches have primarily focused on the use of limited data elements for NAFLD cohort discovery, such as 1) NAFLD diagnosis codes with little risk factor inclusion[9], and 2) unstructured, clinical notes[8], which fail to provide the full clinical picture of NAFLD. Herein, we describe a rule-based phenotype algorithm, developed at Columbia University Irving Medical Center (CUIMC), that utilizes a multitude of data sources (structured and unstructured) within

patient EHRs to identify NAFLD and NASH patients for clinical intervention. The algorithm queries over 400 diagnosis codes, about 100 laboratory and serology measurements, pathology and various radiology modalities. To demonstrate cross-institutional utility, the algorithm has been validated at two large independent medical centers and demonstrates high performance. We also performed fibrosis scoring on all identified NAFLD patients at CUIMC without histologically confirmed NASH to identify additional patients at highest risk for progressing to end-stage liver outcomes. As this algorithm was developed using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), it can be easily deployed at all healthcare institutions that support this CDM, which is presently 90 sites worldwide, and will include the National Institutes of Health *All of Us* Research Program[11].

**Methods:**

The NAFLD algorithm was developed using EHR data from patients within the Columbia University Irving Medical Center (CUIMC) health care center. CUIMC, a component of the New York Presbyterian health care system, serves the diverse population of New York City, and the healthcare system is composed of approximately 38% Hispanic patients, 37% European American, 21% African American, and 4% other ethnicities. At the time of the study, there were records for 6.4 million patients stored in the CUIMC clinical data warehouse (CDW)[12]. The CDW was converted to the Observational Medical Outcomes Partnership (OMOP)[13–15] model in March 2018, and the database is composed of longitudinal EHRs for inpatient and outpatient encounters dating back to 1985. The structured health record data (e.g., diagnoses, medications, procedures, and demographics) are all standardized to the OMOP common data model and are stored in MySQL.

Code for the NAFLD algorithm predominantly consists of SQL queries of the structured data coupled with unstructured data parsing. The workflow of the algorithm may be broken down into

three main steps, as seen in Figure 2. Step 1 is the inclusion of potential NAFLD patients and is composed of the identification of patients with NAFLD risk indicators (Step 1a) and that of patients with NAFLD diagnoses (Step 1b). In step 2, non-NAFLD patients meeting select exclusion criteria are removed from the cohort, and hepatic steatosis is verified in Step 3. Each stage of the algorithm flows consecutively so that a patient will not reach step 3 without meeting the criteria of preceding steps. The algorithm was further validated in two independent healthcare institutions: University of Pennsylvania Healthcare System (UPHS) in Philadelphia, Pennsylvania, and Vanderbilt University Medical Center (VUMC) in Nashville, Tennessee. Figure 1 provides an illustrative depiction of the NAFLD algorithm development and validation process.

***Step 1: Identification of NAFLD patients***

Step 1 identifies NAFLD patients and is broken down into 2 sub-steps. In Step 1a, NAFLD patients are identified by the presence of a NAFLD risk indicator, and in Step 1b, by the presence of a NAFLD diagnosis code. All diagnosis codes used in the algorithm and selection criteria for the NAFLD risk indicators are listed in Supplementary Table 1 (S1). The NAFLD diagnosis codes used for patient selection are listed in Supplementary Table 2 (S2). We required patients to be diagnosed with one risk indicator (Supplementary Table 1) or one NAFLD diagnosis code (Supplementary Table 2) for cohort inclusion, notably inclusive of cirrhosis. NAFLD risk indicators include diagnosis of the following: type 2 diabetes (Table S1a), obesity (Table S1b), abnormal liver enzymes (Table S1c), hyperlipidemia (Table S1d), or hypertension (Table S1e). For the abnormal liver enzyme category, we required patients to have an alanine aminotransferase (ALT) serum/plasma value ≥40 across at least 2 measurements taken at least 6 months apart for inclusion. Patients with one diagnosis of the specified International Classification of Diseases, Ninth and Tenth Revision, Clinical Modification (referenced as ICD-9/ICD-10 throughout the manuscript) codes were included in the cohort. For

laboratory measurements (coded using Logical Observation Identifiers Names and Codes, or LOINC codes), cutoff values for cohort inclusion are listed in the respective tables.

***Step 2: Exclusion of patients with confounding diagnoses***

Following identification of potential NAFLD patients, cases meeting specified exclusion criteria were removed in Step 2 of the algorithm. The exclusion criteria include diagnosis codes for excessive alcohol use, diagnosis of human immunodeficiency virus (HIV), viral hepatitis, type 1 diabetes, or other confounding liver or liver-affecting conditions that may result in secondary hepatic steatosis, including Alpha-1-antitrypsin deficiency, hemochromatosis, and cystic fibrosis. Patients prescribed a hepatotoxic medication associated with steatosis[16], such as an anti-retroviral, tamoxifen, or methotrexate, were also excluded. All patient exclusion criteria are listed in Supplementary Table 3. Patients meeting any of the exclusion criteria were removed from our cohort.

***Step 3: Verification of hepatic steatosis***

Radiology and pathology reports, in the form of unstructured or free-text data, from 1980-2016 were used to verify hepatic steatosis in Step 3. Regular expressions, a powerful pattern search language, and tool[17], were used in conjunction with specific key terms to identify language and usage context indicative of hepatic steatosis in a string-matching approach. Language for an indicator of NAFLD, as well as of the inflammatory phenotype, NASH, were included. Supplementary Table 4 lists the various radiological modalities and the keywords that were queried in the respective reports. Supplementary Table 5 specifies the key terms used to identify hepatic steatosis from pathology reports obtained via liver biopsy.

### *Fibrosis Scoring*

Histologic confirmation is the current standard for verification of NASH. However, biopsies are often underutilized due to their invasive nature. NASH patients were identified from the total pool of patients with verified hepatic steatosis at CUIMC using NASH specific terminology from pathology records. To identify additional patients who may be **at risk** for fibrotic NAFLD, including NASH, we applied 3 common fibrosis scoring metrics on patients lacking histology. These validated metrics include the Fibrosis-4 (FIB-4)[18] calculation (Supplementary Equation 1), the aspartate transaminase (AST) to Platelet Ratio Index (APRI)[19] calculation (Supplementary equation 2), and the NAFLD Fibrosis score[20] (Supplementary equation 3). Data required for these calculations were extracted from patient clinical records. For each required variable, the mean of all measures within 1 year of the date of verified hepatic steatosis was used. For example, given a patient with verified hepatic steatosis on June 20, 2017, the alanine aminotransferase (ALT) value used in the scoring metric was the mean of all available ALT measures from June 20, 2016 to June 20, 2018. R base functions were used to calculate fibrosis scores, and any patient missing the required data elements was excluded from fibrosis scoring. As each of the fibrosis calculations has advantages and disadvantages, we required patients to exhibit a score suggestive of advanced fibrosis using at least 2 of the metrics. Scores indicative of advanced fibrosis are a FIB-4 > 3.25, an APRI >1.0, and a NAFLD FS > 0.675.

### *Quality Control*

To minimize EHR diagnosis code errors, we employed quality control (QC) measures requiring patients to have at least two NAFLD risk indicators (Step 1a), a risk indicator and a NAFLD diagnosis (Step 1b), or at least 3 unique occurrences of a single given NAFLD risk indicator diagnosis. The cohort was also restricted to patients who are 18 or older at the earliest date of hepatic steatosis confirmation by imaging or radiology.

### *Algorithm Validation*

The algorithm was validated at two external, independent healthcare systems, University of Pennsylvania Healthcare System (UPHS) and Vanderbilt Medical Center (VUMC). UPHS maintains and supports a data warehouse for translational research that combines data (both discrete data and unstructured text reports) from the five hospitals in the greater Philadelphia area and one in Princeton, New Jersey. The clinical data warehouse contains over four billion rows of discrete clinical data representing the care of 3 million patients dating back to 2005. This population is 64% European American, 24% African American, and 12% patients from other ethnicities. VUMC maintains a de-identified data warehouse that contains both structured (e.g., billing codes) and unstructured data (e.g., clinical notes) from the EHR. The warehouse dates back to 1990 and describes approximately 3.2 million patients, of which 82% are European American, 13% African American, 4% Hispanic, and 1% other ethnicities.

### *Chart Review and Algorithmic Performance*

Manual, retrospective chart review was performed to review data elements for cohort construction and to assess algorithmic accuracy. Random lists of patient Medical Record Numbers (MRNs), identified by the NAFLD algorithm, were used for the purpose of chart review verification. Provider notes, admission notes, discharge summaries, endoscopy records, diagnoses, pathology and radiology reports were all used in chart review. Manual chart review served as a critical component of algorithm development, allowing us to evaluate the efficacy of adding or removing criteria across the algorithm. It allowed us to adjust NAFLD risk indicator criteria and keyword terminology indicative of hepatic steatosis in radiology and pathology reports. Over 150 MRNs were reviewed at CUIMC during algorithmic development using both inpatient and outpatient EHR records. This extensive review allowed us to fine-tune criteria for

selection of patients within the cohort. Chart review at CUIMC was conducted by two clinical research coordinators, and subsequently verified by a board-certified transplant hepatologist. Chart review at VUMC and UPHS was similarly performed by board-certified transplant hepatologists.

Chart review was also necessary to assess algorithm accuracy. Results of chart review were used to calculate the positive predictive value (PPV) of the algorithm at each of the assessed medical systems. PPV is defined as the proportion of patients identified by the phenotyping algorithm as having the condition, determined by expert chart review. We reviewed the charts of 200 patients, independent of those assessed for diagnostic code selection (for a total of 350 clinical charts reviewed), to calculate PPV at CUIMC. Hepatologists at VUMC and UPHS reviewed 20 charts for PPV calculation.

To determine the true positive rate of our algorithm, sensitivity was assessed at CUIMC and UPHS. At CUIMC, 147 physician-diagnosed NAFLD patients within a registry maintained by the transplant hepatology team were used. All patients were diagnosed between 2006 and 2018. Sensitivity was calculated as the number of patients within this registry that were correctly identified by our algorithm. At UPHS, sensitivity was assessed using 146 physician-diagnosed NAFLD patients with visits to the Gastroenterology department within the validation period. Patients with cirrhosis codes were excluded if NAFLD as an etiology could not be robustly confirmed. Sensitivity assessments were not performed at VUMC as de-identified patient data prohibited this interaction between the hepatologist and biomedical informatics team under the current IRB.

**Results**

Our algorithm identified 16,006 NAFLD patients with verified hepatic steatosis at CUIMC, 20,779 patients at UPHS and 19,575 patients at VUMC. Patient data at CUIMC, the primary algorithm development site, were further interrogated to determine the number of patients with a histologic diagnosis of NASH, determined by algorithmically querying the pathology reports. Of the 16,006 NAFLD patients at CUIMC, 170 were identified with a biopsy-proven NASH diagnosis. Patients meeting advanced fibrosis metrics were also identified. Fibrosis calculations were performed on the 15,890 patients lacking histology reports and 943 patients with scores suggestive of advanced fibrosis, as indicated by an elevated score in at least 2 of the metrics, were identified. Figure 2 shows all steps of the algorithm along with the number of patients with NAFLD identified at each of the 3 healthcare centers. The majority of patients initially meeting criteria for NAFLD risk indicators and NAFLD diagnosis codes were dropped from the algorithm during the verification of hepatic steatosis stage. Of the total potential NAFLD patients (identified after considering inclusion and exclusion criteria), 3.2% at CUIMC, 3.1% at UPHS, and 7.5% of patients at VUMC have algorithmic verified steatosis indicative of NAFLD. *This large drop of patients was primarily due to a lack of available imaging or biopsy data within each system's CDW*. Demographics and summary statistics for the NAFLD patients identified at each healthcare system can be seen in Table 1.

***PPV and Sensitivity:***

Chart review performed by clinical experts at CUIMC of 200 randomly selected patients showed 182 individuals correctly identified by the algorithm as having NAFLD, a positive predictive value (PPV) of 91%. For the validation sites, chart review was performed on 20 randomly selected patients. At UPHS, our algorithm correctly identified 15 of these patients NAFLD, a PPV of 75%. At VUMC, 17 of the 20 patients were correctly identified, for an algorithmic PPV of 85%.

Algorithmic sensitivity was assessed at CUIMC using 147 clinically-verified NAFLD patients. Our NAFLD algorithm attained a sensitivity of 58.5%, identifying 86 of the patients. Figure 3 shows sensitivity across the 3 stages of the algorithm at CUIMC. 141 of the patients were identified following step 1 (95.9% sensitivity), 112 after step 2 (76.2% sensitivity), and 86 following step 3 (58.5% sensitivity). The majority of patients not identified by the algorithm do not have imaging data within the CDW, suggesting that imaging was performed outside of the healthcare center, thus they are missed by the algorithm. 146 clinically-diagnosed NAFLD patients at UPHS were identified by our approach, for a full algorithm sensitivity of 40%. All patients were identified following Step 1 criteria (100% sensitivity), 101 after step 2 (69% sensitivity), and 59 following step 3 (60% sensitivity).

## Discussion

The identification of patients with under-recognized disease is critical towards addressing the public health crisis of NAFLD. Delayed recognition by frontline health care providers requires time, effort, and significant support. While education and resources for provider and patient recognition are necessary, they are not sufficient. Patients with end-stage disease are encountered within the EHR and collaborative efforts between data scientists and physicians can help detect them for targeted intervention, including cancer screening and transplant evaluation referral. Healthcare systems with available interrogatable data serve as a meaningful starting point for development of techniques to better identify groups at risk groups for NAFLD as well as sub-groups at risk for other manifestations of disease[21]. This work highlights an algorithm to identify NAFLD and NASH patients within a healthcare system. Institutions supporting OMOP CDM can implement the algorithm to identify at-risk patients. Institutions without CDM support can still use the detailed workflow provided within the supplementary tables with diagnostic codes (ICD and LOINC) and search terms for pathology and radiology modalities to assist in patient identification.

Our algorithm aggregates large amounts of clinical data and uses imaging or histologic components to verify hepatic steatosis. The algorithm exhibits a high PPV of 91% at CUIMC, 85% at VUMC, and 75% at UPHS showing very good generalizability of the approach. The algorithm also incorporates QC measures to reduce the rate of false positives. During chart review at CUIMC, we found that patients with only 1 diagnostic code of NAFLD or a risk indicator were predominantly not true NAFLD patients. QC steps requiring a minimum number of unique diagnoses were employed to remove these patients. The algorithm was designed to prioritize PPV so that patients who truly have NAFLD are selected, and the overall false positive rate is reduced. *This comes with the limitation that not all NAFLD patients will be included, as is reflected by the algorithm's reduced sensitivity of 59% at CUIMC and 40% at UPHS. The sensitivity values are largely affected by missing radiology and histology data for potential NAFLD patients. The majority of potential NAFLD patients identified by the algorithm did not have imaging data within the CDW, and could therefore not by queried in Step 3, verification of hepatic steatosis. Further investigation identified that many of these patients had imaging performed outside of the given healthcare system and are therefore missed by the algorithm*. *At CUIMC, this is particularly stark in the referral center where patients often arrive with outside imaging and reports. Outside images are not integrated within the CUIMC clinical data warehouse and cannot be parsed by the algorithm*. EHRs accessible to national health systems with centralized, shared data are not predicted to have this limitation. Another limitation of the study is small sample sizes for manual chart review at the external validation sites.

Future directions include prospective contact of patients identified through the algorithm particularly those with fibrosis scores indicative of advanced fibrosis. Identified patients can undergo screening and surveillance after informed consent. Additional applications include further validation in national cohorts such as the *All of Us* Research Program, and integration of genomic sequencing of groups both protected from and at-risk for disease progression. Also,

iterative processing of available non-invasive scoring systems (FIB-4, NAFLD-FS, APRI) to monitor longitudinal progression in fibrosis will help identify groups who rapidly progress through stages of disease. These individuals may serve as an ideal population for genomic discovery.

Important sub-phenotypes that may represent other disease processes (e.g., familial hyperlipidemia with hepatic steatosis) may emerge for earlier recognition, avoidance of misdiagnosis, and intervention. As important as the predictive capacity for this algorithm are specifics regarding diverse individuals who fail to be captured with significant disease, potentially limiting the algorithm's diagnostic potential. As telemedicine scales with local testing (imaging and laboratory), point of care measurements accessible to the patient at home can serve as other sources of usable data and patient-facing identification. Data aggregation may look different in the next century following a de-centralized model with the potential of home-based technologies that can be uploaded to a central data warehouse. Clinical utilization of this technology should be subjected to robust clinical validation and longitudinal cohort assessment prior to rapid deployment and scaling for large populations.

**Figures:**

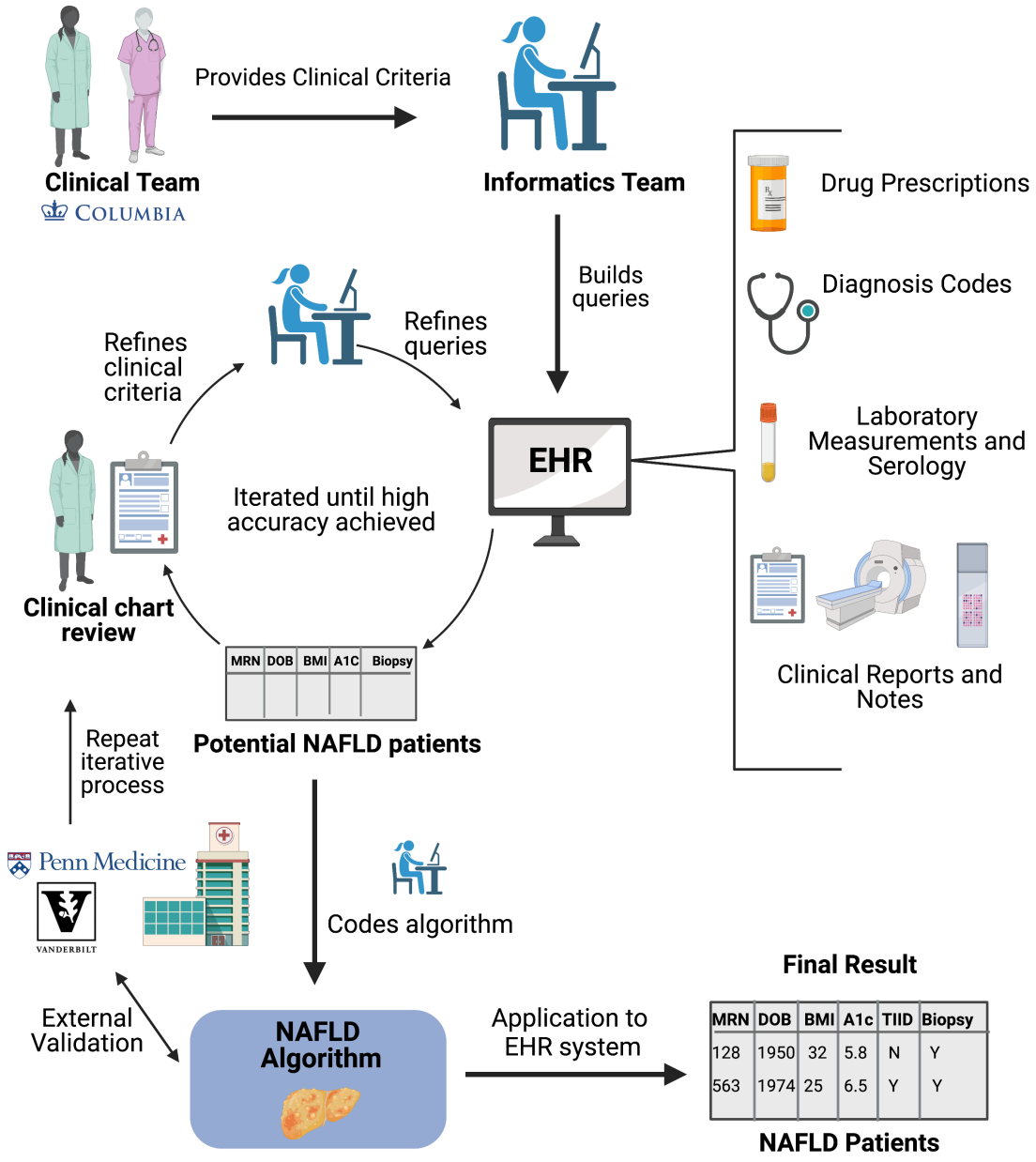


**Figure 1**: Illustration of the NAFLD algorithm development and validation process. The algorithm was developed at Columbia University Irving Medical Center (CUIMC) by clinical and informatics teams. Clinical criteria for the algorithm, provided by medical experts, was used by the bioinformatics team to design queries which produced a set of potential NAFLD patients. The charts of these patients were reviewed by the clinical team to determine true NAFLD status.

Clinical criteria, as represented in the EHR system, was adjusted based on chart review results and the queries were refined. This process was repeated across each step of the algorithm until high accuracy was achieved. Once achieved, the queries were used to code the algorithm. Algorithmic validation was performed at the University of Pennsylvania Healthcare System (UPHS) and Vanderbilt Medical Center (VUMC) where the iterative process described above was repeated by clinical and informatic experts at each site. The final output of the algorithm is a list of NAFLD patients along with clinical characteristics (subset depicted above). EHR = Electronic Health Record; A1c=Glycated hemoglobin; TIID=Type II Diabetes; Y= Yes; N= No; DOB=Date of Birth

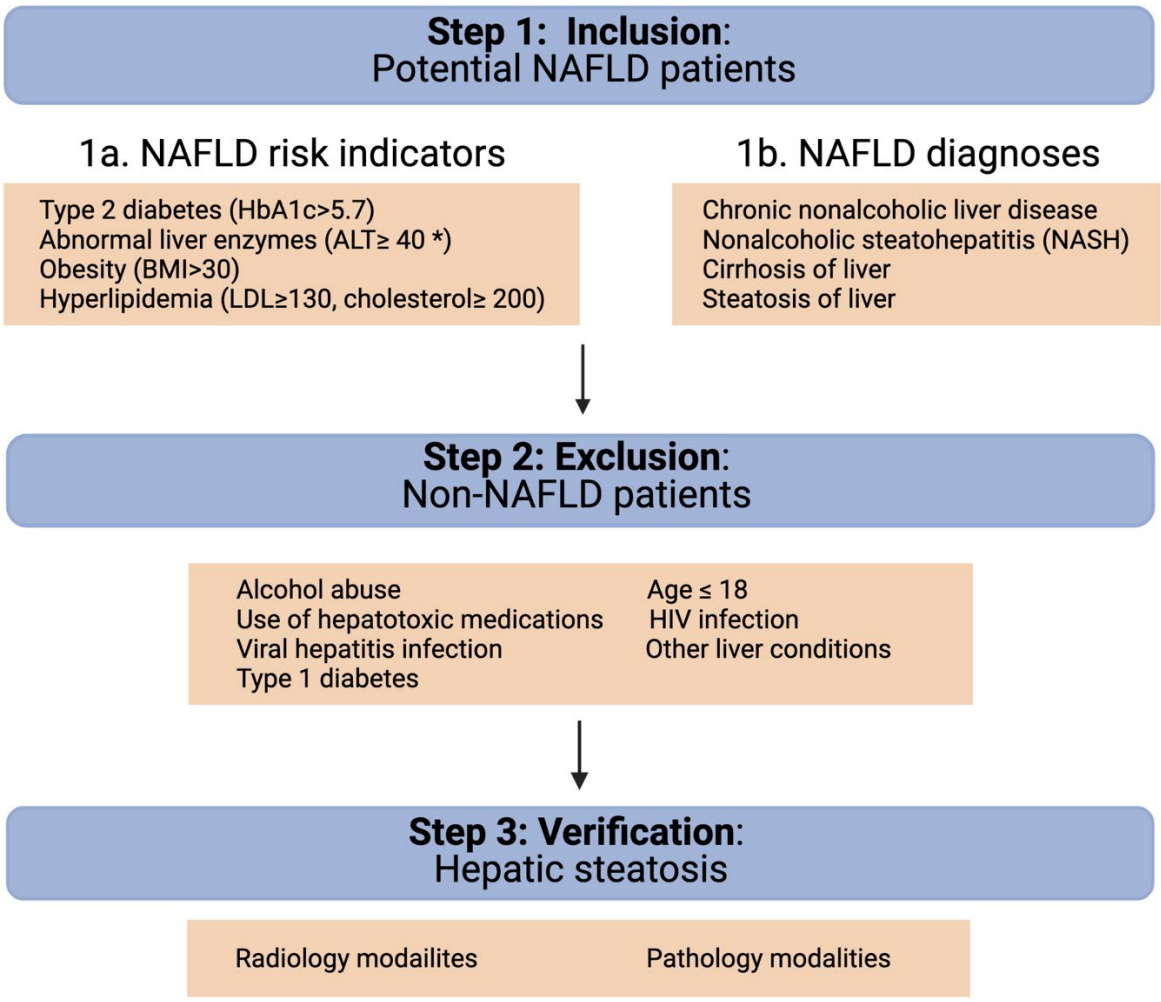


**Figure 2**: The three main steps of the NAFLD algorithm with a *small subset* of the criteria shown. Complete codes for selection or exclusion criteria can be found in the following Supplementary Tables: Step 1a (S Table 1), Step 1b (S Table 2), Step 2 (S Table 3), and Step 3 (S Table 4 and S Table 5). *Patients with an ALT ≥ 40 as seen with at least 2 measurements take 6 months apart.

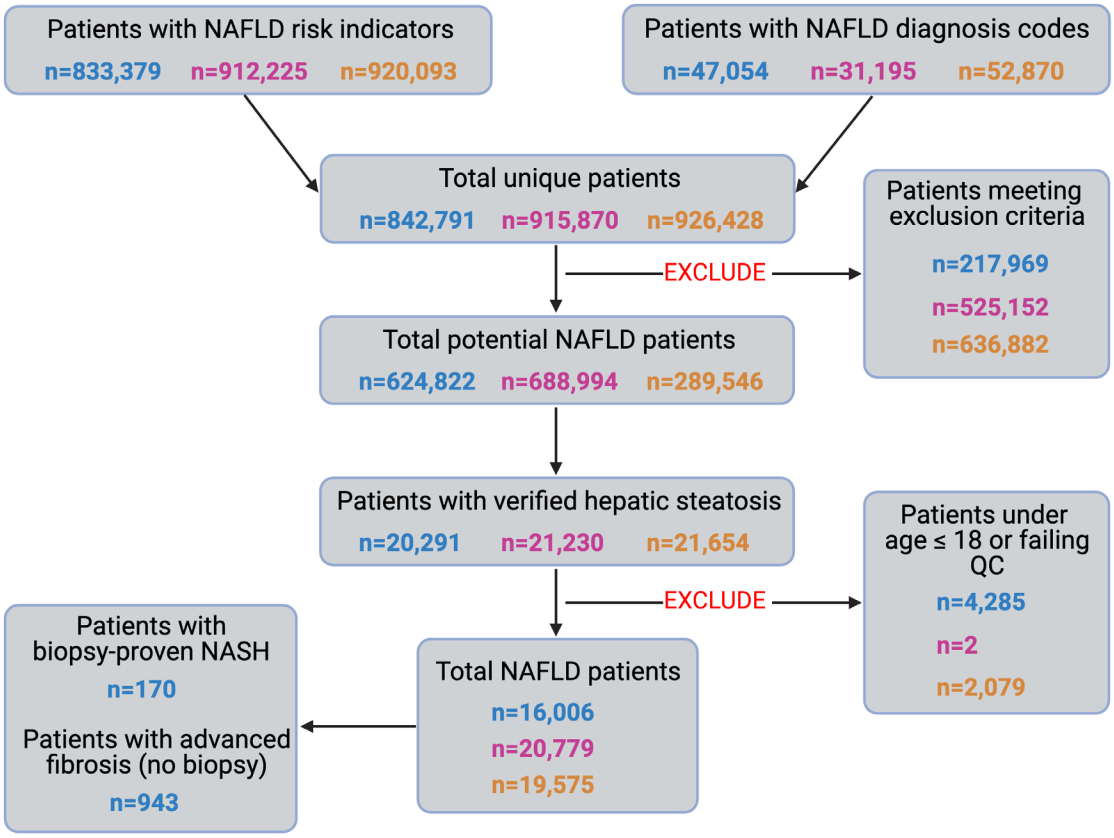


**Figure 3**: Counts of patients at each stage of the algorithm. Data from Columbia University Irving Medical Center (CUIMC) is in blue, that from University of Pennsylvania Healthcare System (UPHS) is in pink/purple, and numbers from Vanderbilt Medical Center (VUMC) are in orange.

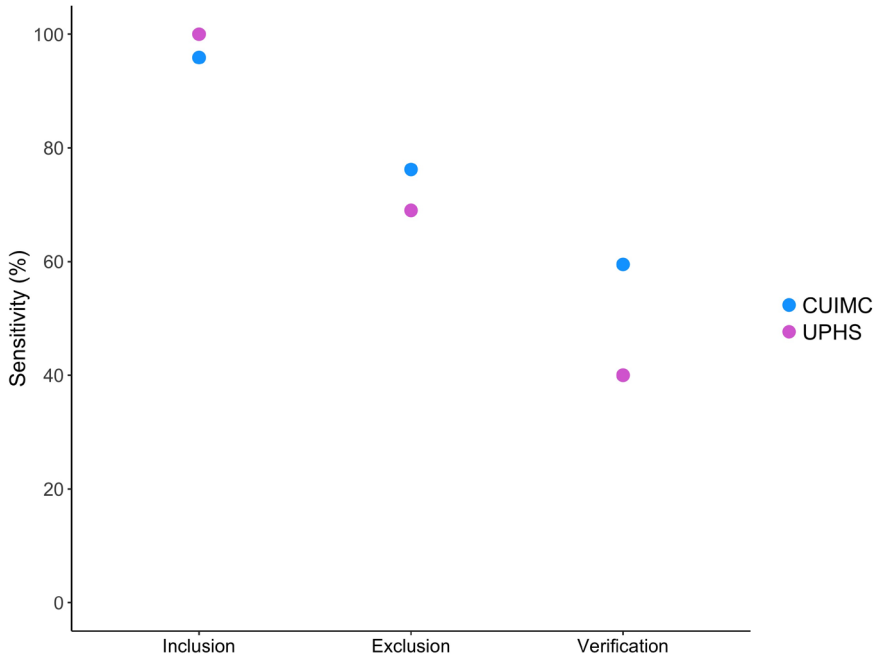


**Figure 4**: Sensitivity at Columbia University Irving Medical Center (CUIMC) (blue) and University of Pennsylvania Healthcare System (UPHS) (pink/purple) after each of the 3 stages of the algorithm. “Inclusion” refers to the identification of NAFLD patients. “Exclusion” is the removal of patients meeting exclusion criteria. “Verification” refers to verification of hepatic steatosis, the final step of our algorithm. 147 known NAFLD patients from CUIMC and 146 from UPHS were used for sensitivity analysis.

**Table 1: Demography and summary information for identified NAFLD patients**

F =Female; M= male; U=unknown/undeclared. Age of diagnosis is based on earliest date of verified hepatic steatosis. Mean age is noted with standard deviation in parentheses. Race/ethnicity values may not aggregate to 100% as some sites code race and ethnicity separately.

| | | | |
|---|---|---|---|
| | CUIMC | UPHS | VUMC |
| **Mean age at diagnosis (+/- Standard Deviation)** | 57.4 (+/- 15.7) | 57.0 (+/- 15.4) | 52.3 (+/- 16.8) |
| **Sex** | F=55.7% M=44.3% | F=44.2% M=55.8% | F=44.4 M=55.6% U=0.01% |
| **Type 2 diabetes** | 53.3% | 31.0% | 24.9% |
| **Obesity** | 45.2% | 60% | 48.8% |
| **Percent Race/Ethnicity** | | | |
| White | 33.9% | 68.4% | 79.3% |
| Black | 9.0% | 16.1% | 12.5% |
| Hispanic | 31.1% | White = 4.8% Black = 1.2% | 3.7% |
| Asian | 1.7% | 2.7% | 0.9% |
| Other | 0.2% | 4.0% | 1.1% |
| Unknown | 24.1% | 2.9% | 6.0% |

**References:**

1. Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol 2018;15:11–20.

2. Alexander M, Loomis AK, Fairburn-Beech J, et al. Real-world data reveal a diagnostic gap in non-alcoholic fatty liver disease. BMC Med 2018;16:130.

3. Estes C, Razavi H, Loomba R, et al. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. Hepatol Baltim Md 2018;67:123–133.

4. Carroll RJ, Eyler AE, Denny JC. Intelligent use and clinical benefits of electronic health records in rheumatoid arthritis. Expert Rev Clin Immunol 2015;11:329–337.

5. Basile AO, Ritchie MD. Informatics and machine learning to define the phenotype. Expert Rev Mol Diagn 2018;18:219–226.

6. Corey KE, Kartoun U, Zheng H, et al. Using an Electronic Medical Records Database to Identify Non-Traditional Cardiovascular Risk Factors in Nonalcoholic Fatty Liver Disease. Am J Gastroenterol 2016;111:671–676.

7. Farmer R, Mathur R, Bhaskaran K, et al. Promises and pitfalls of electronic health record analysis. Diabetologia 2018;61:1241–1248.

8. Van Vleck TT, Chan L, Coca SG, 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 2019;129:334–341.

9. Corey KE, Kartoun U, Zheng H, et al. ≠Development and Validation of An Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record. Dig Dis Sci 2016;61:913–919.

10. Eremić-Kojić N, Đerić M, Govorčin M, et al. Assessment of hepatic steatosis algorithms in non-alcoholic fatty liver disease. Hippokratia 2018;22:10–16.

11. Klann JG, Joss MAH, Embree K, et al. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLOS ONE 2019;14:e0212463.

12. Chelico JD, Wilcox AB, Vawdrey DK, et al. Designing a Clinical Data Warehouse Architecture to Support Quality Improvement Initiatives. AMIA Annu Symp Proc 2017;2016:381–390.

13. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the Science for Active Surveillance: Rationale and Design for the Observational Medical Outcomes Partnership. Ann Intern Med 2010;153:600.

14. Anon. Observational Medical Outcomes Partnership (OMOP) | FNIH. Available at: https://fnih.org/what-we-do/major-completed-programs/omop [Accessed June 20, 2019].

15. Reisinger SJ, Ryan PB, O'Hara DJ, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc JAMIA 2010;17:652–662.

16. 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. Hepatol Baltim Md 2018;67:328–357.

17. Huhdanpaa HT, Tan WK, Rundell SD, et al. Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. J Digit Imaging 2018;31:84–90.

18. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatol Baltim Md 2006;43:1317–1325.

19. Lin Z-H, Xin Y-N, Dong Q-J, et al. Performance of the aspartate aminotransferase-to-platelet ratio index for the staging of hepatitis C-related fibrosis: An updated meta-analysis. Hepatology 2011;53:726–736.

20. Angulo P, Hui JM, Marchesini G, et al. The NAFLD fibrosis score: A noninvasive system that identifies liver fibrosis in patients with NAFLD. Hepatology 2007;45:846–854.

21. Wattacheril J. Extrahepatic Manifestations of Nonalcoholic Fatty Liver Disease. Gastroenterol Clin North Am 2020;49:141–149.