HHS Public Access

Author manuscript

Dig Dis Sci. Author manuscript; available in PMC 2017 March 01.

Published in final edited form as:

Dig Dis Sci. 2016 March ; 61(3): 913–919. doi:10.1007/s10620-015-3952-x.

≠Development and Validation of An Algorithm to Identify Nonalcoholic Fatty Liver Disease in the Electronic Medical Record

Kathleen E Corey, MD MPH MMSc[1,2], **Uri Kartoun, PhD**[2,3], **Hui Zheng, PhD**[4], and **Stanley Y Shaw, MD, PhD**[2,3]

[1]Gastrointestinal Unit, Massachusetts General Hospital, Boston, MA, USA

[2]Harvard Medical School, Boston, MA, USA

[3]Center for Systems Biology, Massachusetts General Hospital, Boston, MA, USA

[4]Biostatistics Center, Massachusetts General Hospital, Boston, MA, USA

Abstract

Background and Aims—NAFLD is the most common cause of chronic liver disease worldwide. Risk factors for NAFLD disease progression and liver-related outcomes remain incompletely understood due to the lack of computational identification methods. The present study sought to design a classification algorithm for NAFLD within the Electronic Medical Record (EMR) for the development of large-scale longitudinal cohorts.

Methods—We implemented feature selection using logistic regression with adaptive LASSO. A training set of 620 patients was randomly selected from the Research Patient Data Registry at Partners Healthcare. To assess a true diagnosis for NAFLD we performed chart reviews and considered either a documentation of a biopsy or a clinical diagnosis of NAFLD. We included in our model variables including laboratory measurements, diagnosis codes, and concepts extracted from medical notes. Variables with $P<0.05$ were included in the multivariable analysis.

Results—The NAFLD classification algorithm included number of natural language mentions of NAFLD in the EMR, lifetime number of ICD-9 codes for NAFLD and triglyceride level. This classification algorithm was superior to an algorithm using ICD-9 data alone with AUC of 0.85 vs. 0.75 ($P<0.0001$) and lead to the creation of a new independent cohort of 8,458 individuals with a high probability for NAFLD.

Conclusions—The NAFLD classification algorithm is superior to ICD-9 billing data alone. This approach is simple to develop, deploy and can be applied across different institutions to create EMR based cohorts of individuals with NAFLD.

Corresponding author: Kathleen E Corey, MD, 55 Fruit Street, Blake 4 Boston, MA 02114, (tel) (617) 724-0274, (fax) 617-724-5997, kcorey@partners.org.

Conflict of Interest: None

Keywords

Nonalcoholic fatty liver disease; nonalcoholic steatohepatitis; Electronic Medical Records; triglycerides

INTRODUCTION

Nonalcoholic fatty liver disease (NAFLD) is the most common cause of liver disease in the United States affecting up to 80 million adults.[1] NAFLD is associated with increased all-cause and cardiovascular-related mortality.[2] Six million American adults are estimated to have nonalcoholic steatohepatitis (NASH), the progressive form of NAFLD[3]. NASH can result in cirrhosis and hepatocellular carcinoma (HCC) and by 2020 may be the leading indication for transplant in the United States.[4,5] Despite the high prevalence of both NAFLD and NASH, the current understanding of the risk factors for progressive liver disease, HCC development and cardiovascular-related mortality in NAFLD remains incomplete. Models to accurately predict which patients will develop end stage liver disease, HCC and cardiovascular disease (CVD) are lacking.

The study of risk factors for progressive disease and co-morbidity development in NAFLD is limited by the long time course of liver disease. The development of cirrhosis and HCC occurs over the course of decades and in only a proportion of those with NAFLD.[6] Robust studies of NAFLD progression and co-morbidity development, thus, require long durations and considerable size to detect sufficient outcomes. As a result of these hurdles, the studies of risk factors for NAFLD progression have been limited to either small groups followed over long durations, large populations followed for short periods or cohorts using NAFLD defined only by aminotransferase levels.[6–9] Thus, there is a pressing need to develop large cohorts of patients with NAFLD that can be assessed over sufficient duration. [7]

Electronic medical records (EMR) can potentially serve as a rich source of data for the development of such cohorts. EMR data can encompass millions of patients seen in healthcare systems over the past several decades. Creation of patient cohorts from EMR databases can provide both the number of patients and duration of follow-up needed to robustly study the natural history of NAFLD and identify modifiable risk factors for NAFLD progression.[10] However, EMRs are limited by the completeness and accuracy of the entered data. The use of billing data to phenotype patients and outcomes can be inaccurate for many diseases and is not sufficiently reliable to allow for the creation and longitudinal evaluation of patient cohorts. However, recent developments in bioinformatic EMR approaches, including an improved ability to extract concepts from narrative text such as physician notes (using natural language processing), have led to the development of EMR algorithms that allow for reliable patient phenotyping and outcomes assessment for both inflammatory bowel disease and rheumatoid arthritis.[11]

The aim of the present study was to develop a comprehensive classification algorithm to identify patients with NAFLD and to create a well-phenotyped NAFLD patient cohort using EMR data. We hypothesize that a classification algorithm for NAFLD using narrative and

codified EMR data will allow for the accurate classification of the presence of NAFLD with a superior accuracy than through billing data alone.

METHODS

Patients and data for the present study were drawn from the Partners HealthCare EMR utilizing the Partners Research Patient Data Registry (RPDR). This centralized clinical data registry contains data from all institutions in the Partners HealthCare System. We utilized data from the Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), both in Boston which serve the greater Northeast United States. MGH began using an EMR in October 1994 and BWH began in October 1993 although limited data from billing and scheduling programs are available starting in 1979.

Narrative EMR data was defined as data entered into narrative notes including clinician notes, pathology and radiographic reports and included co-morbid conditions, medications and laboratory values. Billing data utilized the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) codes. In addition to billing data, other codified data included laboratory values, medications and patient demographics.

We created a dataset from the RPDR to identify adults, aged ≥18 years of age with either a diagnosis of NAFLD *or* those at high risk of NAFLD. This data query included ICD-9 codes for NAFLD 571.8 and 571.9 and problem list notations of NAFLD, NASH, steatosis or fatty liver. Patients at high-risk of NAFLD were defined as those with elevated aminotransferase levels (ALT ≥30 U/L and/or AST ≥30 U/L), body mass index (BMI) ≥25, a procedure code for liver biopsy, a procedure code for bariatric surgery, or a diagnosis of diabetes mellitus (DM). DM was defined by ICD-9 code, inclusion on patient problem list, HbA1C ≥6.5%, glucose ≥200 mg/dL or use of diabetes medications or supplies (e.g., glucometer or test strips). An ALT and AST level 30 U/L for both genders, rather than a lower level for women, were chosen to optimize the identification of cases of NAFLD. Individuals with the most common causes of liver disease including hepatitis C defined as a with a positive HCV RNA level, hepatitis B defined as positive hepatitis B surface antigen, detectable hepatitis B DNA level or diagnosis of alcohol abuse (alcohol use disorder 29.1×, 303.0×, 303.9×, 305.0×), chronic hepatitis C or chronic hepatitis B were excluded. This query returned 647,392 individuals and 21,432 patients were randomly selected to compose the NAFLD Datamart. From the NAFLD datamart 620 patients were randomly selected using SAS for medical records review as a training set.

Defining NAFLD

On medical record review a diagnosis of NAFLD was made by a trained hepatologist using the following criteria: 1) histology on liver biopsy or 2) clinical diagnosis of NAFLD. Histologic NAFLD required the finding of >5% steatosis in the absence of other chronic liver disease. A clinical diagnosis of NAFLD required 1) the presence of fatty infiltration of the liver on imaging (CT scan, MRI or ultrasound), 2) exclusion of hepatitis C infection, 3) absence of documented alcohol abuse, and 4) at least one risk factor for the development of NAFLD which included obesity, dyslipidemia, hypertension or diabetes mellitus. Alcohol abuse and hepatitis C were excluded as they are known to cause hepatic steatosis that can

misclassified as NAFLD. Patients without any radiographic studies of the liver were excluded. Non-NAFLD was defined by negative imaging for fatty infiltration and/or the presence of other chronic liver disease by histology or serologic evaluation. Individuals with NAFLD and a second liver condition (i.e. NAFLD and hepatitis C infection) were excluded. Variables on each individual were gathered and include age at time of imaging, gender, and ethnicity. Gender and ethnicity were determined by patient report. If a patient had multiple imaging studies, the study with the first report of fatty infiltration of the liver was chosen, or for individuals without NAFLD the most recent imaging study was chosen. Variables including aminotransferase levels, lipid levels, hemoglobin A1C (HbA1C), albumin, platelets, INR and BMI documented within 12 months of imaging were collected. When more than one value was present in this time period the average value was determined.

In addition, to billing, laboratory and demographic data we extracted additional variables from approximately 3.5 million notes associated with the patients. We counted the number of occurrences of medical conditions such as “obesity”, “high blood pressure”, and “alcohol problem” (and synonymous expressions or abbreviations) in all notes associated with the patients, including physician narrative notes, radiology and other diagnostic reports, and discharge summaries. We combined similar expressions for several of the variables, for example, the “alcohol problem” variable included expressions such as “alcohol abuse”, “alcohol dependence”, and “problem drinking”. A complete list of the variables is available in eTable 1.

From the 620-patient training set variables associated with the presence of NAFLD were assessed and included in our classification algorithm (described below). This algorithm was then applied to the 21,432-patient NAFLD datamart to create a testing set. From this potential NAFLD group (excluding patients from the training set) 611 unique patients were randomly selected and were used as a testing set.

For additional validation we deployed our algorithm on an independent database of 314,292 patients from the Partners RPDR. This EMR database contains clinical information for patients with diabetes at Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH) between 1990 – 2010 distinct from our primary cohort.

Statistical analysis

Categorical variables were compared using the Chi squared test. Continuous variables were compared using the T test or Mann Whitney test, as appropriate. To determine odds ratio for the variables associated with the presence of NAFLD, logistic regression was performed. Those variables with a $P<0.05$ were included in the multivariable analysis. The logistic classification model was selected using the adaptive LASSO procedure.[12]. The linear score was calculated using the regression coefficients estimated in the logistic regression model:

$$\text{Linear Score} = -1.0742 + 0.449 * \text{fatty\_liver\_codes\_life} + 0.0792 * \text{Number\_all\_NAFLD} + 0.00765 * \text{Triglycerides}$$

The probability for NAFLD was calculated using the inverse logit function:

Probability(NAFLD)= exp(linear score)/(1+exp(linear score)).

For the training and final algorithm the sensitivity, specificity, negative predictive value (NPV) and positive predictive value (PPV) were calculated. A calculated probability of >0.85 was used as an indication for a positive NAFLD diagnosis while a calculated probability of ≤0.85 was an indication for no NAFLD. Various probability cut-offs ranging from 0.80 to 0.95 were tested and 0.85 was chosen as it provided the optimal combination of sensitivity, specificity, PPV and NPV. Sensitivity was defined by the number of patients with a finding of positive classification for NAFLD and positive chart review for NAFLD divided by the number of patients with positive chart review for NAFLD. Specificity was defined by the number of the patients with both negative classification for NAFLD and negative chart review for NAFLD divided by the number of patients with negative chart review for NAFLD. PPV was defined as the number of patients with both positive classification for NAFLD and positive chart review for NAFLD divided by the number of patients with positive classification for NAFLD. Negative predictive value (NPV) was defined as the number of patients with negative classification for NAFLD and negative chart review for NAFLD divided by the number of patients with negative classification for NAFLD. All statistical analysis was performed on SAS 9.4 (SAS Institute Inc., Cary, NC). This study was approved by the Partners Healthcare Human Research Committee which serves as the institutional review board for both BWH and MGH.

RESULTS

Baseline Characteristics of Training Set

620 patients were randomly selected for the training set from the final NAFLD Datamart (n=21,432). On chart review 444 individuals (71.6%) met criteria for NAFLD, while 176 individuals (28.4%) did not meet criteria. (Table 1) Diabetes mellitus was more prevalent among patients with NAFLD (77.0% vs. 50.0%, P<0.0001). In addition, those with NAFLD were older and had higher mean HbA1C levels, higher triglyceride levels and lower HDL levels.

NAFLD Classification Algorithm

Table 2 contains the variables chosen for the final classification algorithm by using logistic regression with adaptive LASSO and regression coefficients.

Variables in the final NAFLD classification algorithm included number of text mentions of NAFLD over a patient's lifetime, number of billing codes for NAFLD over an individual's lifetime and triglyceride level within 12 months of the radiographic report and led to the creation of the final model derived is as follows:

$$\begin{aligned} &logit(p(NAFLD)) \\ &\quad = \beta_0 + \beta_1(Total\ number\ of\ NAFLD\ icd9\ codes) \\ &\quad + \beta_2(Total\ number\ of\ NAFLD\ expressions\ extracted\ from\ notes) \\ &\quad + \beta_3(Most\ recent\ triglycerides\ value\ past\ 12\ months) \end{aligned}$$

Algorithm Performance for NAFLD and non-NAFLD in Testing Set

Six hundred eleven unique patients were randomly selected from the population identified by the training algorithm to serve as the testing set. Three hundred sixty-six patients (59.9%) had NAFLD on chart review, while two hundred forty-five patients (40.1%) did not meet criteria for NAFLD.

Among individuals identified with NAFLD on chart review, only 202 (sensitivity 55.2%) carried an ICD-code diagnosis for NAFLD (571.8). Thus, the use of the ICD-9 code alone would miss 44.8% of NAFLD patients in this cohort. Using both ICD-9 codes 571.8 and 571;9 improves the accuracy of NAFLD identification only slightly, identifying 212 individual with NAFLD (57.9%) and missing 42.1%.

When the classification algorithm was applied to the testing set, using an 85% threshold, 89% of patients classified as having NAFLD truly had NAFLD (PPV) and 91% of patients without NAFLD by chart review did not have NAFLD by the algorithm (specificity). The area under the curve (AUC) for the NAFLD classification model in the testing set was 0.85. (Table 3).

We then applied our classification algorithm to an *independent datamart* of 314,292 patients from the Partners RPDR. The classification algorithm identified 8,458 patients as having NAFLD using the 85% threshold. Medical records review of 100 individuals identified as having NAFLD with the algorithm and randomly selected from this datamart found that 91 were correctly identified giving a PPV of 91%.

Algorithm Performance Compared to Billing Data and Veterans Administration (VA) Model

We compared the performance of our NAFLD classification algorithm to NAFLD billing data. (Table 3). In the ICD-9 coding data model individuals with any lifetime ICD-9 code for NAFLD (571.8 and 571.9) were defined as having NAFLD. The NAFLD classification algorithm had a superior accuracy for the identification of NAFLD () when compared to a model utilizing ICD-9 billing codes for NAFLD alone (AUC 0.85 95% CI 0.81–0.88 vs. 0.75 95% CI 0.72–0.79, P<0.0001) (Figure 1).The NAFLD classification algorithm had identical PPV to billing data (PPV 89% for both) and superior NPV (56% vs. 36%) and specificity (91% vs. 73%). Billing data had a higher sensitivity than the NAFLD classification algorithm (63% vs. 51%).

We also compared our model to the model developed by Husain et al., to classify NAFLD in the Veterans Administration (VA) population, a predominantly male population. The NAFLD classification algorithm was superior to the VA model in PPV (89% vs. 80.8%) and nearly equivalent in specificity (91% vs. 92.4%). However, the VA model carried higher NPV and sensitivity (78.0% vs. 56% and 55% vs. 51%, respectively).

DISCUSSION

The present study identifies a classification algorithm for the identification of individuals with NAFLD using a combination of codified and narrative data in the EMR. In addition, we demonstrate that this algorithm is superior to algorithms using ICD-9 billing data alone. The

final NAFLD algorithm provides an AUC of 0.85 with a PPV of 89% and specificity of 91%. This accuracy was superior to that defined by coding data alone (AUC 0.75) and has a higher PPV than the algorithm used to identify NAFLD in the VA system.

The study of the natural history of NAFLD, risk factors for liver-related outcomes and risk factors for cardiovascular disease has been limited by a lack of large, longitudinal cohorts. The cohorts evaluating the natural history of NAFLD and liver-related outcomes are limited by small size, ranging from 109 to 132 total individuals. Matteoni et al., evaluated the relationship between NAFLD and the development of cirrhosis and liver-related death.[6] While this study had sufficient follow-up, with a mean of 8.2 years, only 132 individuals had complete data and could be included. Similar limitations were found in the cohort followed by Dam-Larsen et al and Ekstedt et al who, while able to follow patients over 16.7 and 13.7 years, respectively, had limited NAFLD cohorts of only 109 and 129 patients.[7,8] Cohorts of larger size have been limited by the reliance on abnormal aminotransferase levels to define NAFLD and controls.[13] Recent studies have demonstrated that the majority of patients with NAFLD have normal aminotransferase levels suggesting that studies using elevated levels to detect NAFLD may suffer from misclassification bias.[14] The present study seeks to address the limitations of these cohorts by rigorously defining NAFLD either histotologically or radiographically while excluding other causes of liver disease. In addition, the resulting cohort includes 8,458 individuals followed in this healthcare system over at least one year. Thus, the present algorithm allows for creation of a rigorously defined, sufficiently large cohort to allow for longitudinal evaluation of liver-related and cardiovascular outcomes in NAFLD.

The development of algorithms for the study of liver disease in EMR databases is limited.[15] Husain et al have developed an algorithm for the identification of NAFLD associated with elevated ALT in the VA System.[16] This algorithm serves as a valuable tool for the study of NAFLD within the VA system. This study is limited, however, by its development in a predominantly male population and its strict requirement for inclusion of individuals with multiple elevated ALTs over time. Data suggests that 79–86% of patients with NAFLD, either steatosis or NASH, have normal aminotransferase levels and thus the VA algorithm may exclude a large proportion of patients.[14,17] The present study expands on this work by identifying an algorithm developed from two large urban tertiary medical centers outside the VA system and includes both men and women equally. In addition, our algorithm allows for the identification of individuals with NAFLD whose aminotransferase levels are normal.

Our algorithm is also superior to a model relying of ICD-9 codes for NAFLD. NAFLD lacks a distinct and specific ICD-9 code. Administrative data is frequently inaccurate and errors in billing data are common.[18] NAFLD is traditionally billed using ICD-9 codes 571.8, ‘other chronic nonalcoholic liver disease’ or 571.9, ‘unspecified chronic liver disease without alcohol’. However, these codes may be used for liver conditions other than NAFLD, decreasing the accuracy of these codes in identifying NAFLD patients. Our model, with the inclusion of NLP, significantly increased the accuracy over billing data alone.

Strengths

The present study has several strengths. The algorithm was derived from a large database from two large academic medical centers. The final algorithm utilizes only three variables and does not require additional testing or documentation outside the standard of care. This approach is straightforward to deploy and to extend across different institutions to create multicenter EMR based cohorts.

Limitations

While inclusion in our study did not require the presence of fatty infiltration on radiographic imaging, a clinical diagnosis of NAFLD did require positive imaging for fatty infiltration or NAFLD on biopsy. As a result, patients without imaging or histologic confirmation of NAFLD in our system were excluded from our algorithm development. While this limitation may exclude some patients, this criterion allows for a robust definition of NAFLD and identification of cases. In addition, the goal of the present study was to develop an algorithm to identify cases of NAFLD, rather than unaffected controls. As a result we begin our study with a datamart of high risk patients which could be applied in other EMRs to identify cases. Our study is also limited to two large academic medical centers in the Northeastern United States. Further evaluation of the validity of the presented approach is needed in other geographic areas and in community medical centers. However, recent data suggests that despite differences in EMR structure classification approaches can be successfully used across multiple healthcare systems.[19]

NAFLD is the leading cause of chronic liver disease in the United States and can result in end stage liver disease and HCC.[5] In addition, NAFLD is an independent risk factor for the development of cardiovascular disease, the leading cause of death among individuals with NAFLD.[20] Understanding the pathogenesis of NAFLD progression and the risk factors for CVD in NAFLD is imperative to limiting the morbidity and mortality resulting from NAFLD. Unfortunately, the development and progression of hepatic fibrosis and its associated complications occur over decades which limit the ability to evaluate these outcomes. The creation of a robust EMR-based cohort of individuals with NAFLD inclusive of several decades of patients will allow for the identification of risk factors of NAFLD progression and of liver-related outcomes include HCC development and need for liver transplantation. As EMR penetration increases and institutional databases become increasingly linked they will be increasing opportunities to study disease processes such as NAFLD at a population level. In addition, institutions such as ours are championing efforts to link EMR to biorepository specimens. Such linkage will allow for the study of NAFLD biomarkers and genetics.[21] The present algorithm is the first step to the development of such cohorts by allowing for comprehensive and accurate case identification. It is important to note that EMR based studies will not replace the need for rigorous, long-term prospective studies to evaluate liver-related complications and cardiovascular disease in NAFLD.

The findings of the present study have several important implications. To our knowledge, this is the first study to develop an algorithm for the identification of NAFLD in a large population of both men and women. In addition, our algorithm provides greater accuracy than that from ICD-9 billing data alone. This approach will allow for the development of

large, longitudinal cohorts to follow outcomes in patients with NAFLD and determine risk factors for both liver-related and cardiovascular-related morbidity and mortality in these patients.

We conclude that the present classification algorithm for NAFLD using codified and narrative data is superior to an algorithm using ICD-9 billing data alone and this approach will allow for the development of large, longitudinal EMR-based cohorts of individuals with NAFLD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to acknowledge Dr. Ashwin N. Ananthakrishnan, MBBS, MPH who provided feedback and critical comments.

Financial Support: This study was funded in part by grants from the NIH K23 DK099422 (KEC) and NIH U54 LM008748 (SYS).

References

1. Williams CD, Stengel J, Asike MI, et al. Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. Gastroenterology. 2011; 140:124–131. [PubMed: 20858492]
2. Byrne CD, Targher G. NAFLD: A multisystem disease. Journal of hepatology. 2015; 62:S47–S64. [PubMed: 25920090] Musso G, Gambino R, Cassader M, Pagano G. Meta-analysis: natural history of non-alcoholic fatty liver disease (NAFLD) and diagnostic accuracy of non-invasive tests for liver disease severity. Ann Med. 2011; 43:617–649. [PubMed: 21039302]
3. Vernon G, Baranova A, Younossi ZM. Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Alimentary pharmacology & therapeutics. 2011; 34:274–285. [PubMed: 21623852]
4. White DL, Kanwal F, El-Serag HB. Association between nonalcoholic fatty liver disease and risk for hepatocellular cancer, based on systematic review. Clin Gastroenterol Hepatol. 2012; 10:1342–1359. e1342. [PubMed: 23041539]
5. Charlton M. Cirrhosis and liver failure in nonalcoholic fatty liver disease: Molehill or mountain? Hepatology. 2008; 47:1431–1433. [PubMed: 18393323]
6. Matteoni CA, Younossi ZM, Gramlich T, Boparai N, Liu YC, McCullough AJ. Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. Gastroenterology. 1999; 116:1413–1419. [PubMed: 10348825]
7. Dam-Larsen S, Franzmann M, Andersen IB, et al. Long term prognosis of fatty liver: risk of chronic liver disease and death. Gut. 2004; 53:750–755. [PubMed: 15082596]
8. Ekstedt M, Franzen LE, Mathiesen UL, et al. Long-term follow-up of patients with NAFLD and elevated liver enzymes. Hepatology. 2006; 44:865–873. [PubMed: 17006923]
9. Soderberg C, Stal P, Askling J, et al. Decreased survival of subjects with elevated liver function tests during a 28-year follow-up. Hepatology. 2010; 51:595–602. [PubMed: 20014114]
10. Sung KC, Kim BS, Cho YK, et al. Predicting incident fatty liver using simple cardio-metabolic risk factors at baseline. BMC gastroenterology. 2012; 12:84. [PubMed: 22770479]
11. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis care & research. 2010; 62:1120–1127. [PubMed: 20235204] Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in

electronic medical records using natural language processing: a novel informatics approach. Inflammatory bowel diseases. 2013; 19:1411–1420. [PubMed: 23567779]

12. Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Assocation. 2006; 101:1418–1429.Friedman, JHT.; Tibshirani, R. The elements of statistical learning. New York: Springer-Verlag; 2001.
13. Dunn W, Xu R, Wingard DL, et al. Suspected nonalcoholic fatty liver disease and mortality risk in a population-based cohort study. Am J Gastroenterol. 2008; 103:2263–2271. [PubMed: 18684196] Ong JP, Pitts A, Younossi ZM. Increased overall mortality and liver-related mortality in non-alcoholic fatty liver disease. J Hepatol. 2008; 49:608–612. [PubMed: 18682312]
14. Targher G, Bertolini L, Rodella S, et al. Nonalcoholic fatty liver disease is independently associated with an increased incidence of cardiovascular events in type 2 diabetic patients. Diabetes Care. 2007; 30:2119–2121. [PubMed: 17519430]
15. Kramer JR, Davila JA, Miller ED, Richardson P, Giordano TP, El-Serag HB. The validity of viral hepatitis and chronic liver disease diagnoses in Veterans Affairs administrative databases. Alimentary pharmacology & therapeutics. 2008; 27:274–282. [PubMed: 17996017]
16. Husain N, Blais P, Kramer J, et al. Nonalcoholic fatty liver disease (NAFLD) in the Veterans Administration population: development and validation of an algorithm for NAFLD using automated data. Aliment Pharmacol Ther. 2014; 40:949–954. [PubMed: 25155259]
17. Browning JD, Szczepaniak LS, Dobbins R, et al. Prevalence of hepatic steatosis in an urban population in the United States: impact of ethnicity. Hepatology. 2004; 40:1387–1395. [PubMed: 15565570]
18. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Medical care. 2004; 42:1066–1072. [PubMed: 15586833]
19. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. Journal of the American Medical Informatics Association : JAMIA. 2013; 20:e147–e154. [PubMed: 23531748]
20. Corey KE, Chalasani N. Management of Dyslipidemia as a Cardiovascular Risk Factor in Individuals With Nonalcoholic Fatty Liver Disease. Clin Gastroenterol Hepatol. 2014
21. Trivedi B. Biomedical science: betting the bank. Nature. 2008; 452:926–929. [PubMed: 18441548] Murphy S, Churchill S, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. Genome research. 2009; 19:1675–1681. [PubMed: 19602638]

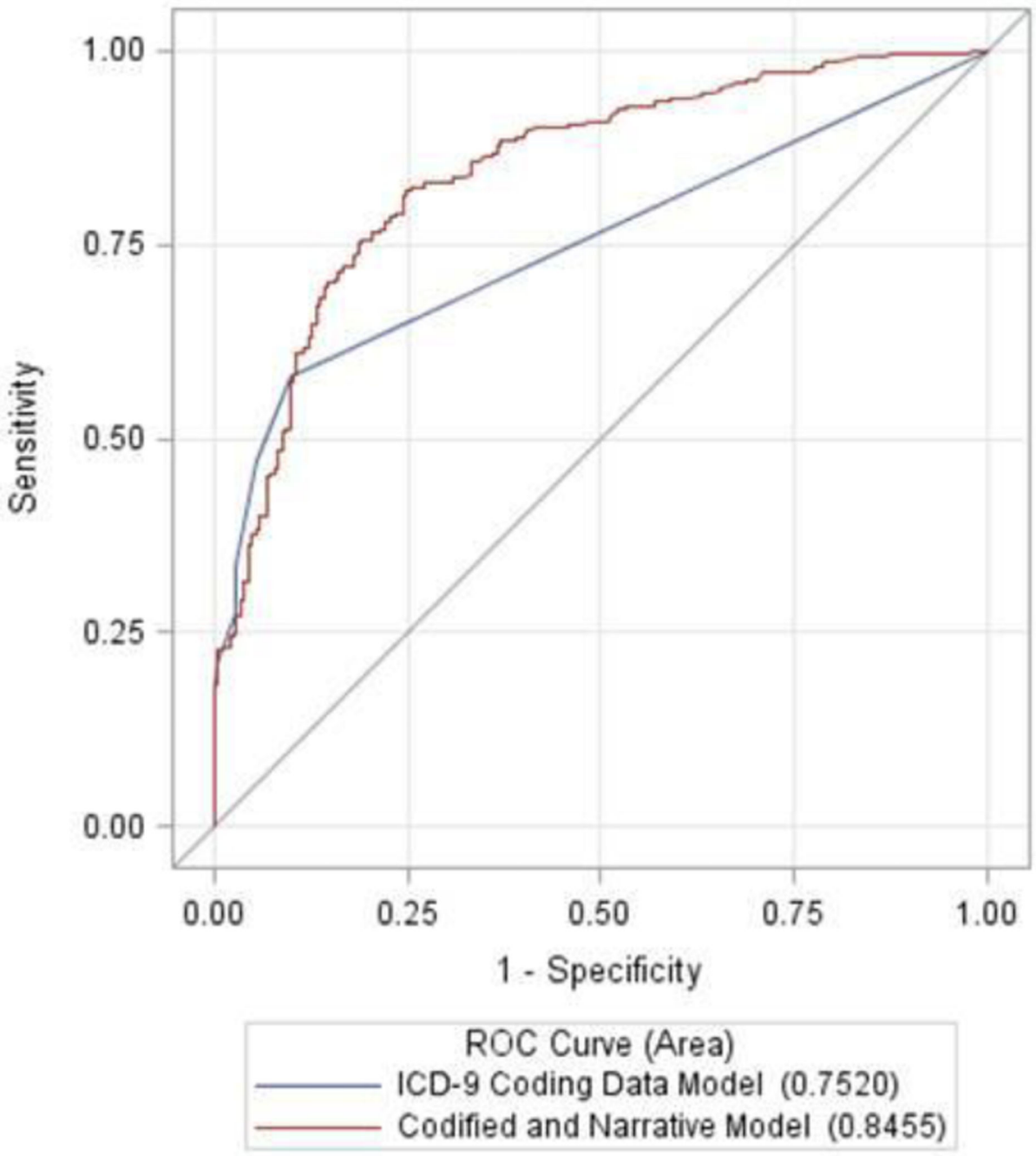

Figure 1.
ROC Curves comparing the Performance of the Codified and Narrative Model for NAFLD to the ICD-9 Coding Data Model for NAFLD

Table 1

Baseline Characteristics of Training Set

	Identified as NAFLD	Identified as Not NAFLD	p-value
Age, years (SD)	61.6 (12.7)	58.7 (15.2)	0.01*
Female (%)	247(56%)	95(54%)	0.71†
Ethnicity (%)			
White	332 (74.8%)	128 (72.7%)	
African American	37 (8.3%)	22 (12.5%)	
Hispanic	43 (9.7%)	11 (6.3%)	0.28†
Asian	7 (1.6%)	5 (2.8%)	
Other	25 (5.6%)	10 (5.7%)	
BMI (SD)	32.5 (7.3)	32.9 (7.2)	0.55*
Diabetes, (%)	342 (77.0%)	88 (50.0%)	<0.0001†
HbA1c, % (SD)	7.4% (1.5%)	7.1% (1.5%)	0.04‡
ALT, U/L (SD)	42.1 (75.8)	44.3 (59.2)	0.97‡
AST, U/L (SD)	34.3 (39.1)	42.6 (78.3)	0.22‡
Low-density lipoprotein, mg/dL (SD)	97.2 (34.8)	95.7 (35.1)	0.77‡
Triglycerides Level, mg/dL (SD)	187.9 (140.8)	129.3 (76.8)	<0.0001‡
High-density lipoprotein level, mg/dL, (SD)	45.8 (12.9)	53.7 (21.0)	0.002‡
Total cholesterol, mg/dL (SD)	179.2 (42.1)	173.7 (44.8)	0.60‡

† Chi-square test

* Student's t-test

‡ Wilcoxon rank sum test

Table 2

Variables selected from the training algorithm (narrative plus codified data) using logistic regression

Variable	Regression Coefficient	Standard Error	p-value
Number of Fatty Liver mentions extracted from notes over a lifetime (E.g.. Steatosis, fatty liver, NAFLD, NASH)*	*0.4490*	*0.1398*	0.0013
Lifetime Number of NAFLD ICD-9 Codes	*0.0792*	*0.0187*	<0.0001
Triglyceride Level within 12 months of NAFLD Radiographic Report	*0.0077*	*0.0019*	<0.0001

* For complete term list see Supplementary table 1

Table 3

Performance of Various Algorithms

	PPV	NPV	Sensitivity	Specificity	AUC
Codified and Narrative Model	89%	56%	51%	91%	0.85
ICD-9 Coding Data Model	89%	36%	63%	73%	0.75
Husain et al., VA model	80.8%	78.0%	55.0%	92.4%	Not done