ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Application of supervised and semi-supervised learning prediction models to predict progression to cirrhosis in chronic hepatitis C

Yueying Hu[1†], Weijing Tang[2†], Lauren A. Beste[3,4†], Grace L. Su[5,6], George N. Ioannou[7,8], Tony Van[9], Ji Zhu[10†], and Akbar K. Waljee[9,11,12†*]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States of America

[2]Department of Statistics and Data Science, Dietrich College of Humanities and Social Sciences, Carnegie Melon University, Pittsburgh, Pennsylvania, United States of America

[4]General Medicine Service, Veterans Affairs Puget Sound Healthcare System, Seattle, Washington, United States of America

[4]Department of Medicine, Veterans Affairs Puget Sound Healthcare System, Seattle, Washington, United States of America

[5]Gastroenterology Service, VA Ann Arbor Healthcare System, Ann Arbor, Michigan, United States of America

[6]Department of Internal Medicine, Michigan Medicine, Ann Arbor, Michigan, United States of America

[7]Gastroenterology Service, Veterans Affairs Puget Sound Healthcare System, Seattle, Washington, United States of America

[8]Division of Gastroenterology, School of Medicine, University of Washington, Seattle, Washington, United States of America

[†]These authors contributed equally to this work.

***Corresponding author:**
Akbar Waljee
(awaljee@med.umich.edu)

Citation: Hu Y, Tang W, Beste LA, *et al.* Application of supervised and semi-supervised learning prediction models to predict progression to cirrhosis in chronic hepatitis C. *Artif Intell Health.* 2025;2(2):87-99. doi: 10.36922/aih.4671

Received: August 27, 2024

Revised: October 31, 2024

Accepted: December 19, 2024

Published online: January 2, 2025

Abstract

In this study, we aim to examine the efficacy of deep learning methods in predicting the 1-year risk of developing cirrhosis in patients with chronic hepatitis C (CHC), as defined by transient elastography (TE), in comparison with conventional models, as well as to assess whether semi-supervised learning can improve performance relative to supervised learning when the labels are limited. We used the electronic health records of the 169,317 valid patients in the Veterans Health Administration system from 2000 to 2016. Predictor variables contained baseline characteristics, such as age, gender, race, hepatitis C virus genotype, and 26 liver-related longitudinal variables such as sustained virologic response and laboratory data. The response variable, developing cirrhosis, is defined as liver stiffness >12.5 kPa on TE within a 1-year window. Using baseline and longitudinal variables, we fitted four prediction models, including logistic regression (LR), random forest (RF), supervised recurrent neural network (RNN), and semi-supervised RNN (semi-RNN) and evaluated their performances. Both RNN (area under the receiver operating characteristic curve [AuROC] 0.744) and semi-RNN (AuROC 0.785) accurately predicted the risk of cirrhosis within 1 year and significantly outperformed RF (AuROC 0.731) and LR (AuROC 0.724). By enabling early identification of high-risk patients, these models hold promise for targeted interventions in clinical CHC treatment.

Keywords: Semi-supervised learning; Electronic health records; Longitudinal predictors

[9]Center for Clinical Management Research, VA Ann Arbor Healthcare System, Ann Arbor, Michigan, United States of America
[10]Department of Statistics, College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, Michigan, United States of America
[11]Center for Global Health and Equity, University of Michigan, Ann Arbor, Michigan, United States of America
[12]Department of Learning Health Sciences, Michigan Medicine, University of Michigan, Ann Arbor, Michigan, United States of America

1. Introduction

Chronic hepatitis C (CHC) is a leading cause of cirrhosis, liver transplantation, and hepatocellular carcinoma. The development of cirrhosis in patients with CHC is highly variable and non-linear.[1] Many cofactors influence the risk of cirrhosis in patients with CHC, including patient characteristics (e.g., alcohol use, obesity, and age), viral factors (e.g., genotype), successful antiviral treatment, and others.[2] Reliable models to predict cirrhosis risk are needed to facilitate population-level screening and guide treatment decision-making. Machine learning methods to predict cirrhosis development offer greater flexibility than traditional predictive methods because they can accommodate large numbers of predictor variables and are able to handle data with complex inter-variable relationships and irregular collection intervals.

Traditional methods for predicting cirrhosis are subject to several limitations. Liver biopsy, which is considered the gold standard for diagnosis, is invasive and poorly scalable for large populations. Non-invasive markers of liver disease, such as the AST-to-platelet ratio index (APRI), the fibrosis-4 index (FIB-4), transient elastography (TE), and others, offer only single snapshots in time and do not account for longitudinal changes. Fibrosis assessment is not performed across large populations at consistent intervals or in all patients. Moreover, few previous models examining patients with CHC evaluate the risk of continued liver disease progression after antiviral therapy (e.g., due to comorbid liver disease that persists after CHC eradication). In addition, the variable rate of fibrosis progression over time complicates the development of reliable risk prediction models using conventional methods.

Machine learning is a form of artificial intelligence that uses computer algorithms to identify patterns in large datasets and allows computers to assimilate new information without being explicitly programmed.[3] It has demonstrated success in many real-world healthcare applications, including computer-aided interpretation of liver imaging, prediction of hepatocellular carcinoma risk, and prediction of cirrhosis development.[4-6] For example, conventional machine learning models such as logistic regression (LR) and random forest (RF) have been shown to effectively predict the disease progression in CHC by incorporating both baseline predictors and summary statistics of longitudinal predictors, with RF being able to capture non-linear trends that are limitedly represented by LR.[7] Recent advances in deep learning, a subtype of machine learning, see the emergence of the deep recurrent neural network (RNN) as a powerful tool to process sequential data collected at various times.[8] The structure of RNN has shown superior performance for applications such as machine translation, and it is flexible to be applied in both supervised learning and semi-supervised learning. In supervised learning, we use training data with known outcomes ("labeled data") to learn an algorithm that can make accurate predictions for new unseen data. In contrast, semi-supervised learning uses both labeled data and data with missing outcomes ("unlabeled data"), where the unlabeled data can help identify relevant patterns. Semi-supervised learning can help improve prediction performance especially in the case where labeled data is scarce. Both supervised RNN and semi-supervised RNN (semi-RNN) offer advantages over conventional methods, such as LR because they can handle varying time durations and irregular time gaps between two consecutive visits, and they can automatically learn predictive patterns from raw data, rather than requiring pre-specified feature extraction.

Machine learning methods have previously demonstrated superior performance compared to linear Cox proportional hazards in predicting the risk of cirrhosis in patients with CHC, as defined based on APRI score thresholds.[4] However, it remains unknown whether machine learning methods perform well in predicting progression to cirrhosis, as defined by TE, a far more sensitive modality for assessing liver fibrosis. Although TE has become more widely used in the last decade, it is still used in only a minority of patients with CHC. Therefore, we hypothesized that outcomes from patients who underwent TE could be used to train a model to accurately predict the 1-year risk of developing cirrhosis in CHC patients, as defined by TE. The Veterans Health Administration (VHA) serves the largest single cohort of CHC patients in the United States. Our analysis aimed to evaluate the predictive performance of deep learning methods and compare it to conventional models. Moreover, we aimed to assess whether semi-RNN can obtain better performance than a supervised RNN when the number of patients with TE outcomes is limited.

2. Data and methods

2.1. Data source

The national VHA system is the largest integrated healthcare system in the United States. It includes 172 medical centers and 1,069 outpatient sites of care, serving 9 million enrollees.[9] All data were obtained from the VA Corporate Data Warehouse, which is a comprehensive repository of data from the VA's universal electronic medical record system including laboratory data, biometric data, diagnoses, and pharmacy data.[10]

All study procedures were approved by the VA Ann Arbor Institutional Review Board. All procedures conform to the ethical guidelines of the 1975 Declaration of Helsinki. A waiver of informed patient consent was obtained before project initiation.

2.2. Study population

We identified 182,747 VHA users with a history of positive HCV RNA tests seen in the VHA at least once between January 2000 and January 2016. Patients were followed from the date of the first APRI (enrollment) to their last visit recorded in the VA system through January 2019. To ensure that patients did not have cirrhosis at enrollment, we included only patients with APRI results <2.0 (72% negative predictive value for cirrhosis in CHC) at enrollment.[11] Because antiviral treatment outcome is a key predictor of cirrhosis development, we excluded an additional 13,430 patients who received antiviral treatment regimens but lacked RNA tests in VHA electronic records to document whether sustained virologic response (SVR) was achieved. After these exclusions, the cohort contained 169,317 patients, among which 10,575 patients had undergone TE after enrollment. Finally, since we aimed to develop longitudinal models predicting the development of cirrhosis over a 1-year period, we excluded TE results for 297 patients who had less than 1 year of available follow-up time between enrollment and their last available TE. This resulted in a final analytic cohort of 10,278 patients with valid TE results (the "labeled cohort") for 1-year prediction and a cohort of 159,039 patients without TE results (the "unlabeled cohort").

2.3. Progression to cirrhosis defined by TE

TE was introduced into the VHA system in 2013 for the non-invasive assessment of fibrosis and can be considered a reliable measure for cirrhosis outcome. Our primary outcome, the development of cirrhosis, was defined based on liver stiffness >12.5 kPa on TE measured at least once in the VHA data. The earliest date of liver stiffness >12.5 kPa on available TEs is defined as the date of cirrhosis.

2.4. Predictor variables

Predictor variables for predicting cirrhosis development were selected based on our previous research and biological plausibility. We employed both baseline and longitudinal variables for our analysis. The baseline predictors consisted of age at the enrollment, gender, race, and HCV genotype. The longitudinal predictors, which may be assessed multiple times, included achievement of SVR, body mass index, and 24 laboratory blood tests. The achievement of SVR was defined as a serum HCV RNA viral load below the lower limit of detection performed at least 12 weeks after the end of HCV treatment, where we identified all antiviral treatment regimens received, including both interferon and direct-acting antiviral-based therapies. The blood tests used in this study included total bilirubin, aspartate aminotransferase (AST), alanine aminotransferase (ALT), alpha-fetoprotein (AFP), alkaline phosphatase (ALP), albumin, AST:ALT ratio, FIB-4, APRI, blood urea nitrogen, creatinine, glucose, international normalized ratio (INR), hemoglobin, leukocyte count, platelet count (PLT), sodium, potassium, chloride, and total protein. FIB-4 and APRI scores were defined using published formulae to assess the degree of liver fibrosis.[12] In addition, the laboratory values of AST, ALT, AFP, and ALP, which were measured through standardized blood tests, were divided by the corresponding upper limits of normal to account for differences in reference ranges across laboratories.

2.5. Cohort building

Labeled patients were followed from enrollment (time 0) to the date of the last available TE or the date of diagnosis of cirrhosis through TE, if applicable. Unlabeled patients (i.e., those without TE outcomes) were followed from enrollment to the last visit documented in the VHA records (Figure 1). The training cohort was created by randomly selecting visit dates from the patient's follow-up records. This approach simulates the scenario in which we aim to predict the risk of cirrhosis within a year of a clinical visit based on a patient's medical history.

2.5.1. Labeled cohort for supervised learning

All patients with known cirrhosis outcomes by TE were included in this cohort (Figure 1). The models used baseline predictors as well as the entire trajectory of the longitudinal predictors from enrollment to their sampled visit time *t*. The outcome measured whether the patient developed cirrhosis within 1 year, starting from time *t*.

2.5.1.1. Cases

There were 2,247 patients in the labeled cohort who developed cirrhosis during follow-up according to their

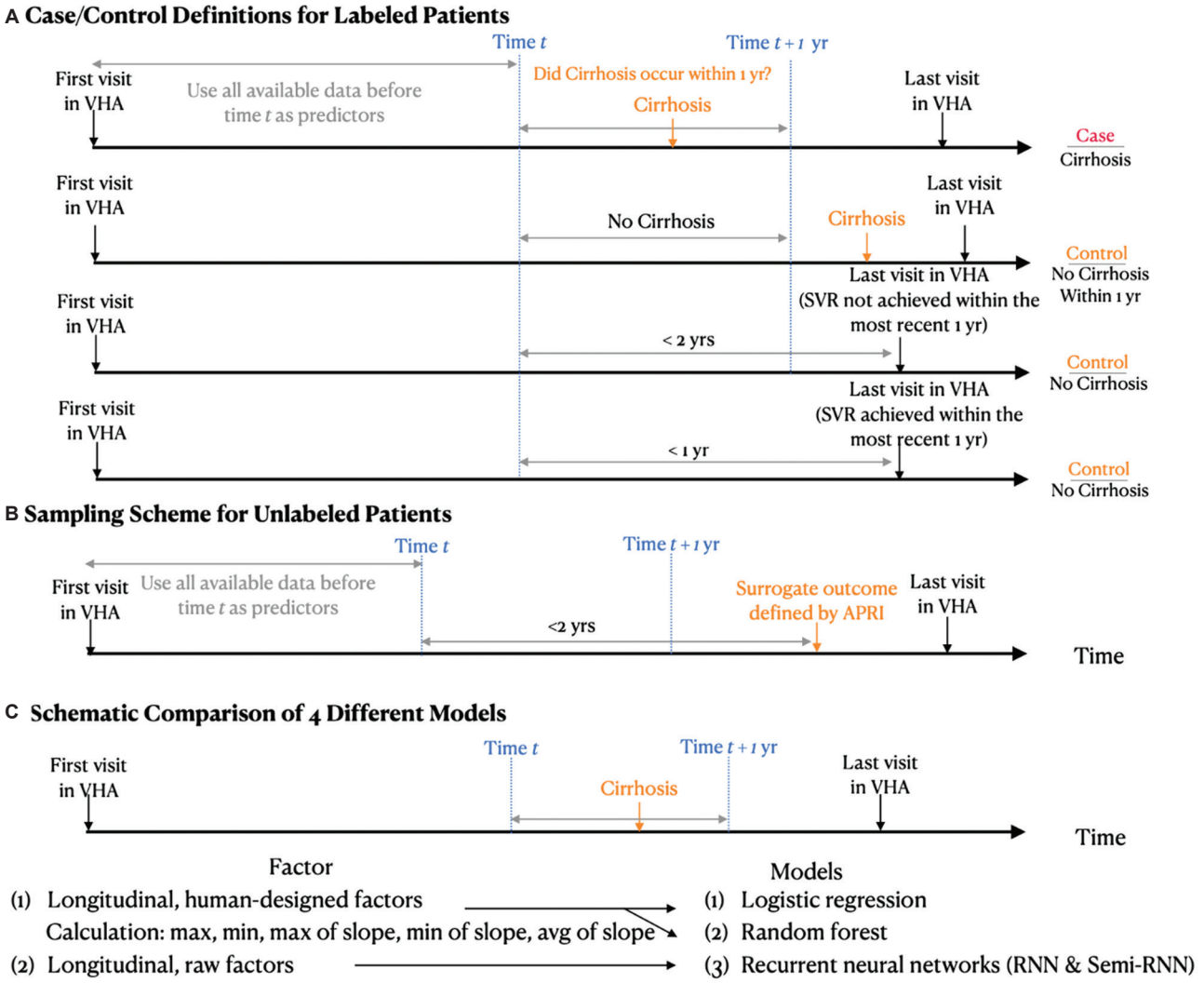

Figure 1. Sampling scheme for population and comparison of models. (A) We identified patients who were infected with HCV and had a valid TE outcome within 1 year of their sampled visit (time *t*) to the VHA. The patients who developed cirrhosis according to TE within 1 year of time *t* were considered cases, while those who did not were assigned as controls. We used all available data at and before time *t* to predict the probability of developing cirrhosis within 1 year. The first two examples highlighted patients who developed cirrhosis during follow-up; the last two examples showed patients who did not develop cirrhosis during the same period. (B) We were then left with patients who were infected with HCV but did not have a valid TE outcome within 1 year of time *t*. We randomly sampled one visit 1 – 2 years before the surrogate outcome (time *t*) to ensure that the distribution of predictors would be similar to the labeled cohort. (C) Schematic comparison of the four different models we developed to predict cirrhosis (LR and RF used human-designed longitudinal features before time *t*, while RNN and semi-RNN used raw longitudinal factors before time *t*).
Abbreviations: APRI: AST-to-platelet ratio index; HCV: Hepatitis C virus; LR: Logistic regression; RF: Random forest; RNN: Supervised recurrent neural network; Semi-RNN: Semi-supervised recurrent neural network; SVR: Sustained virologic response; TE: Transient elastography; VHA: Veterans Health Administration.

TEs (liver stiffness >12.5 kPa). Among 48 patients who developed cirrhosis but had less than 1 year of follow-up from enrollment, we randomly sampled 1 visit (time *t*) before their first positive outcome. Out of the remaining 2,199 patients, 2,134 had laboratory visits within 1 year before their first positive outcome. For each patient, we randomly sampled 1 visit (time *t*) from all the visits they had during that 1-year period. We thereby obtained 2,182 cases in which the TE outcome became positive within 1 year of the sampled visit (time *t*) (Figure 1).

2.5.1.2. Controls

Of the 2,199 patients who developed cirrhosis more than 1 year after enrollment, their outcome was negative from enrollment to 1 year before their first diagnosis. By randomly sampling visits within that window, we obtained

2,199 control samples. Of the 8031 patients who did not develop cirrhosis and maintained a follow-up of more than 1 year, 976 of them achieved SVR within 1 year of their most recent TE. We assumed that they would not develop cirrhosis in the subsequent year, and thus randomly sampled one visit as time *t* within 1 year before the last available TE. For the rest of the patients who neither achieved SVR in the most recent 1 year nor developed cirrhosis, 6,345 patients had documented records within the 1 – 2-year window before the last available TE, for whom we randomly sampled 1 visit as time *t* in that window. Finally, for the remaining 710 patients who did not have any records within 2 years before the last available TE, we selected the most recent documented visit before the last available TE as time *t*. We ended up with 10,230 control samples in which TE outcome was assumed to remain negative within 1 year of the sampled visit (time *t*) (Figure 1).

2.5.2. Additional unlabeled cohort for semi-supervised learning

In addition to the labeled cohort described above, we included patients without known TE outcomes (unlabeled patients) to improve the feature representation of longitudinal predictors. We defined a surrogate outcome as the achievement of two consecutive APRIs >2[4] to ensure a similar sampling scheme to the labeled cohort and to avoid potential bias. After removal of visits later than the 1[st] date of developing the surrogate outcome, we randomly sampled one visit (time *t*) within the 1 – 2-year windows before the surrogate outcome for the 159,039 unlabeled patients (Figure 1).

2.6. Models for supervised and semi-supervised learning

We developed four different models to predict the probability of developing cirrhosis within 1 year after time *t* using baseline predictors and longitudinal predictors from enrollment to time *t*. To utilize longitudinal information, we employed two approaches in our analysis. The first approach was to compute summary statistics for each longitudinal predictor, including maximum, minimum, maximum of slope, minimum of slope, and total variation. These summary statistics were combined with baseline predictors and used to train conventional machine learning models.[13] We opted for LR (a classic linear method) and RF (a highly non-linear method based on decision trees) to evaluate the effectiveness of conventional machine learning methods based on human-designed factors.

The second approach to handle raw longitudinal predictors from enrollment to time *t* was using an RNN, which excels in processing sequential data with irregular time gaps and eliminates the need for feature extraction. The adaptable structure of RNNs also enables them to support both supervised and semi-supervised learning, a capability that is not easy to attain with LR or RF. We developed a supervised RNN utilizing labeled data only, and a semi-RNN that employed the abundant unlabeled data to improve classification performance.

Specifically, for the supervised RNN model, we used gated recurrent units (GRU)[14] to regulate information flow and remember long-term information. The longitudinal information from hidden units of GRU was passed to a max pooling layer, and then was merged with time-invariant information from baseline predictors using a feedforward neural network (FNN). Finally, we built another FNN to process the combined information, and used a sigmoid activation function in the output layer to predict the probability of developing cirrhosis within 1 year (Figure 2). In all FNNs, we used rectified linear unit (ReLU) as the non-linear activation function. To train the model, we minimized the binary cross-entropy loss, which is named the supervised loss, through the Adam stochastic algorithm.[13] Adam's adaptive learning rates naturally deal with noisy gradients, which can be viewed as a form of implicit regularization. By dampening the effect of noisy updates (due to averaging over time), Adam avoids the tendency to overfit to noise in the training data.[15] We also used dropout technique[16] to prevent overfitting and an early stopping mechanism with a patience of 10 epochs to prevent unnecessary training beyond the optimal point.

For the semi-RNN model, we incorporated an auxiliary task in addition to the primary prediction task of the supervised RNN. The auxiliary task was to predict the values of longitudinal predictors at the next visit, which can be trained using unlabeled data. We shared the layers of GRU between the auxiliary task and the prediction task. Jointly learning both tasks could help improve feature representation by leveraging unlabeled data. We defined the negative log-likelihood for longitudinal predictors as the unsupervised loss, and we minimized the weighted sum of the supervised and unsupervised losses to train the model end-to-end.[17] The weight, which controls the trade-off between supervised learning and unsupervised learning, was selected by hyperparameter tuning.

2.7. Statistical analysis

To conduct the analysis, we randomly split the labeled data into a training set (40%), a validation set (30%), and a testing set (30%). The unlabeled data used in semi-RNN belonged to the training set. We learned each model using the training and validation set, and then evaluated their performance on the same testing set. This procedure was

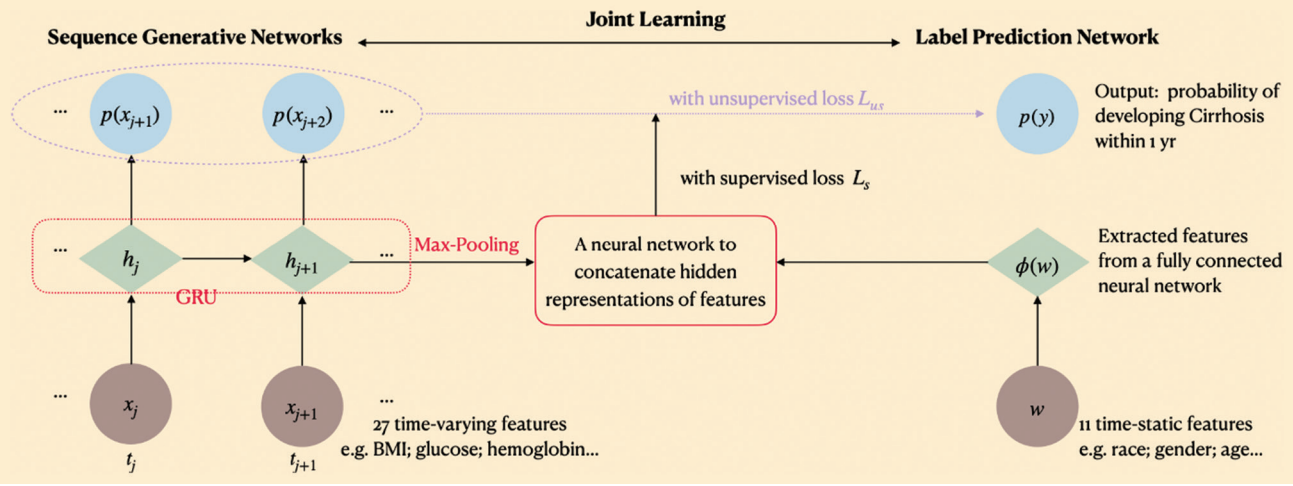

Figure 2. Model architecture for RNN and semi-RNN
Abbreviations: BMI: Body mass index; GRU: Gated recurrent units; RNN: Supervised recurrent neural network; Semi-RNN: Semi-supervised recurrent neural network.

carried out for 10 times, and the resulting performance characteristics on the testing sets were averaged over the 10 splits to examine each method.

Due to the irregularity in the number and timing of clinic visits across patients, and the incomplete availability of certain predictors at each visit, we employed imputation techniques to address missing data. For the LR and RF models, we first calculated summary statistics for the longitudinal predictors. We then applied the missForest algorithm,[18] which efficiently handles multivariate data comprising both continuous and categorical variables without relying on distributional assumptions or requiring extensive parameter tuning. For the RNN and semi-RNN models, we first imputed missing entries of longitudinal predictors at time 0 by the mean of non-missing entries at enrollment in the training set. We then filled in the remaining missing entries with the latest non-missing value before the respective time points. Once missing data were imputed, we standardized all the data using the mean and standard deviation of the corresponding training set.

We tuned hyperparameters for each method as follows: For two conventional machine learning methods, we combined the training and validation sets to conduct 10-fold cross-validation and selected the optimal hyperparameters as those achieving the highest area under the receiver operating characteristic curve (AuROC).[19] We adopted the penalized LR given the collinearity among longitudinal summary statistics and tuned the regularization strength for the LR model. For the RF model, we searched the optimal number of trees and the number of features used for each split. Both models were implemented using the Scikit-learn library in Python version 3.9.7.[20] For the RNN and semi-RNN models, given the computation cost of training deep learning models, we selected the optimal hyperparameters as those achieving the highest AuROC on the hold-out validation set. We randomly generated 100 combinations of hyperparameters, searching the optimal hidden sizes of model structures within the range of [16, 128], batch size from the set {64, 128, 256, 512}, dropout rates in the range of [0.2, 0.5], learning rates of Adam optimizer within the range of [0.0001, 0.01], and weights for semi-supervised learning within the range of [0.001, 1], where applicable. We implemented both RNN models using PyTorch 1.12.1[21] on a high-performance computing cluster.

We evaluated the models' ability to distinguish if the patient developed cirrhosis within 1 year by measuring their performance characteristics using AuROC and the area under the precision-recall curve (AuPRC).[19] To compare the overall accuracy of the models, we used the Brier score[22] where a score of 0 indicates perfect accuracy. Furthermore, we compared the performance characteristics of the conventional models and RNN models using a paired sample t-test. The two-sided p-values were reported to assess the statistical significance.

Given that machine learning algorithms typically require large datasets to achieve optimal performance, we evaluated the robustness of four models by analyzing their performance when only a limited amount of labeled data are available. To achieve this, we reduced the amount of labeled training and validation data to 50%, 20%, 10%, and 5% of the original sets, and repeated the training, validation, and testing procedures as above. We plotted the changes in three performance characteristics for each model at each percentage level.

3. Results

3.1. Cohort population characteristics

Out of the 12,412 labeled samples, most were from men (12,040 [97.0%]). The sample population exhibited a racially diverse distribution, with the majority comprising individuals of White and Black ethnic backgrounds (Table 1). Compared to the control samples, cases who developed cirrhosis within 1 year of the sampled visit were less likely to achieve SVR (147 [6.74%] versus 734 [7.18%]), had higher levels of serum AST, ALT, and bilirubin levels (mean [SD] AST: 66.8 [46.3] U/L versus 46.7 [33.7 U/L; mean [SD] ALT: 75.7 [69.7] U/L versus 56.3 [51.8] U/L; mean [SD] bilirubin: 0.767 [0.614] mg/dL versus 0.641 [0.372] mg/dL), higher FIB-4 and APRI scores (mean [SD] FIB-4 score: 3.3 [2.7] versus 1.9 [1.9]; mean [SD] APRI score: 3.29 [2.63] versus 2.14 [1.72]), and lower platelet count (mean [SD]: 176 [66.7] × 103/μL vs 207 [68.1] × 10^3/μL) at the time of the sampled visit (time *t*). Overall, the unlabeled cohort had similar characteristics to the labeled cohort except for a few variables. The unlabeled cohort had a larger proportion of missing data for race/ethnicity groups and genotypes and were more likely to achieve SVR at time *t* than the labeled cohort (13,286 [8.35%] versus 881 [7.10%]).

3.2. Prediction of progression to cirrhosis

The semi-RNN model demonstrated the highest mean (SD) AuROC (0.785 [0.062]), followed by the RNN model (0.744, [0.009]), both of which significantly surpass those attained by the RF and LR models. In addition, both the semi-RNN and RNN models demonstrated significantly higher mean (SD) AuPRCs (0.448 [0.119] and 0.371 [0.010], respectively), and significantly lower mean (SD)

Table 1. Patient characteristics in model building

Characteristic	Controls (*n*=10,230)	Cases (*n*=2,182)	Labeled cohort (*n*=12,412)	Unlabeled cohort (*n*=159,039)
Number of visits	55.4	66.6	57.4	36.1
Male, *n* (%)	9,904 (96.8)	2,136 (97.9)	12,040 (97.0)	154,299 (97.0)
Race/Ethnicity, *n* (%)				
Black	4,974 (48.6)	889 (40.7)	5,863 (47.2)	53,254 (33.5)
Hispanic	453 (4.43)	133 (6.10)	586 (4.72)	7,345 (4.62)
White	4,235 (41.4)	1,023 (46.9)	5,258 (42.4)	82,044 (51.6)
Other	192 (1.88)	47 (2.15)	239 (1.93)	2,765 (1.74)
Missing	376 (3.68)	90 (4.12)	466 (3.75)	13,631 (8.57)
Genotype, *n* (%)				
1	8,573 (83.8)	1,817 (83.3)	10,390 (83.7)	101,874 (64.1)
2	924 (9.03)	193 (8.85)	1,117 (9.00)	14,873 (9.33)
3	479 (4.68)	131 (6.00)	610 (4.91)	8,686 (5.47)
≥4	90 (0.88)	19 (0.87)	109 (0.88)	1,152 (0.72)
Missing	164 (1.60)	22 (1.00)	186 (1.50)	32,454 (20.4)
SVR achieved at time *t*, *n* (%)	734 (7.18)	147 (6.74)	881 (7.10)	13,286 (8.35)
Age at enrollment, mean (SD)	52.1 (7.42)	52.5 (6.72)	52.1 (7.30)	52.8 (19.1)
BMI at time t, mean (SD)	27.7 (5.16)	28.7 (5.70)	27.8 (5.27)	27.5 (5.55)
Laboratory test results at time *t*, mean (SD)				
AST, U/L	46.7 (33.7)	66.8 (46.3)	50.3 (37.0)	47.5 (33.3)
ALT, U/L	56.3 (51.8)	75.7 (69.7)	59.7 (55.9)	56.7 (64.3)
PLT, ×10^3/μL	207 (68.1)	176 (66.7)	202 (68.9)	218 (76.8)
Bilirubin, mg/dL	0.641 (0.372)	0.767 (0.614)	0.663 (0.427)	0.716 (0.604)
INR	1.07 (0.299)	1.12 (0.313)	1.08 (0.302)	1.09 (0.333)
Creatinine, mg/dL	1.17 (1.05)	1.23 (1.28)	1.18 (1.09)	1.17 (1.13)
FIB-4 score	2.14 (1.72)	3.29 (2.63)	2.35 (1.96)	1.98 (1.40)
APRI score	0.677 (0.728)	1.16 (1.15)	0.762 (0.839)	0.644 (0.701)

Abbreviation: ALT: Alanine aminotransferase; APRI: AST-to-platelet ratio index; AST: Aspartate aminotransferase; BMI: Body mass index; FIB-4: Fibrosis-4 index; INR: International normalized ratio; PLT: Platelet count; SD: Standard deviation; SVR: Sustained virologic response.

Brier scores (0.120 [0.015] and 0.129 [0.002], respectively), compared to the LR and RF models (Table 2). All p-values of four paired t-tests including LR versus RNN, LR versus semi-RNN, RF versus RNN, and RF versus semi-RNN are below 0.05 for AuROC, Brier score, AuPRC, proportion of samples who test positive at 80% sensitivity, specificity at 80% sensitivity, positive predictive value at 80% sensitivity, and negative predictive value at 80% sensitivity.

3.3. Model robustness

Robustness characterizes a model's capacity to sustain consistent and reliable predictions under varying conditions. In this study, we deliberately reduced the volume of labeled data to assess the models' stability and generalizability. The superior performance of the semi-RNN and RNN models became evident when 50% and 100% of the training and validation data for the labeled cohort was used (Figure 3). This finding underscores the critical role of substantial labeled data in optimizing the predictive power of deep learning models.

3.4. Model calibration

To examine the calibration of the models, we chose a representative split with an AuROC closest to the mean of 10 splits for the RNN model. The calibration plot demonstrates the correspondence between predicted probabilities and observed outcomes. A perfect calibration is denoted by a 45° diagonal line, signifying that the model's predicted probabilities precisely match the actual probabilities of events occurring. Proximity to this ideal calibration line illustrates superior calibration. In Figure 4 (and Figures S1-S3), all four models exhibited good calibration across various proportions of labeled training and validation data (10%, 20%, 50%, and 100%) for predicting 1-year risks, with semi-RNN emerging as the optimal performer. This suggests that the models are reliable in estimating risks and have the potential for use in clinical decision-making.

3.5. Feature attribution of models

To elucidate the decision-making processes within a neural network model, we applied the feature attribution technique for an exemplary patient, who had lower predicted risk scores at first, and then higher predicted risk scores in later visits, for a representative split from the RNN model when 100% of the training and validation set was used for the labeled cohort. The feature attribution technique can quantify the contribution of individual features to a model's prediction[23] through the calculation of the gradient of features against the loss function at two different visits and compare the feature importance based on their centered, adjusted values. The plots (Figure 5) revealed that, at the later visit with a higher predicted risk score, the RNN model relied more heavily on features such as AFP, FIB-4, and albumin, despite the centered values of these features not being extreme. In contrast, compared to an earlier visit, the model placed less emphasis on features, such as time to first visit and glucose but greater emphasis on features, such as SVR and the standardized ratio of AFP. We also provided a similar variable importance analysis for RF and LR shown in Figure S4.

Table 2. Comparison of performance metrics of the models predicting cirrhosis development within 1 year in patients at risk when 100% of labeled data were used

Characteristic, mean (SD)	LR	RF	RNN	Semi-RNN	*P*-value*
AuROC	0.724 (0.008)	0.731 (0.008)	0.744 (0.009)	0.785 (0.062)	<0.050
Brier score	0.133 (0.002)	0.131 (0.002)	0.129 (0.002)	0.120 (0.015)	<0.050
AuPRC	0.345 (0.009)	0.358 (0.006)	0.371 (0.010)	0.448 (0.119)	<0.050
Proportion of samples who test positive at 90% sensitivity	0.699 (0.018)	0.685 (0.025)	0.674 (0.021)	0.597 (0.109)	>0.050
Specificity at 90% sensitivity	0.344 (0.022)	0.361 (0.030)	0.374 (0.026)	0.467 (0.133)	>0.050
Positive predictive value at 90% sensitivity	0.227 (0.006)	0.232 (0.009)	0.236 (0.009)	0.277 (0.061)	>0.050
Negative predictive value at 90% sensitivity	0.940 (0.004)	0.943 (0.004)	0.945 (0.004)	0.953 (0.010)	>0.050
Proportion of samples who test positive at 80% sensitivity	0.538 (0.014)	0.533 (0.015)	0.517 (0.018)	0.455 (0.095)	<0.050
Specificity at 80% sensitivity	0.518 (0.017)	0.524 (0.018)	0.544 (0.022)	0.619 (0.115)	<0.050
Positive predictive value at 80% sensitivity	0.263 (0.007)	0.265 (0.009)	0.274 (0.011)	0.328 (0.084)	<0.050
Negative predictive value at 80% sensitivity	0.923 (0.003)	0.924 (0.003)	0.926 (0.003)	0.933 (0.010)	<0.050

*All *P*-values of four paired t-tests including LR versus RNN, LR versus semi-RNN, RF versus RNN, and RF versus semi-RNN are below 0.05 for AuROC, Brier score, AuPRC, proportion of samples who test positive at 80% sensitivity, specificity at 80% sensitivity, positive predictive value at 80% sensitivity, and negative predictive value at 80% sensitivity.

Abbreviations: AuROC: Area under the receiver operating characteristic curve; AuPRC: Area under the precision-recall curve; LR: Logistic regression; RF: Random forest; RNN: Supervised recurrent neural network; Semi-RNN: Semi-supervised recurrent neural network.

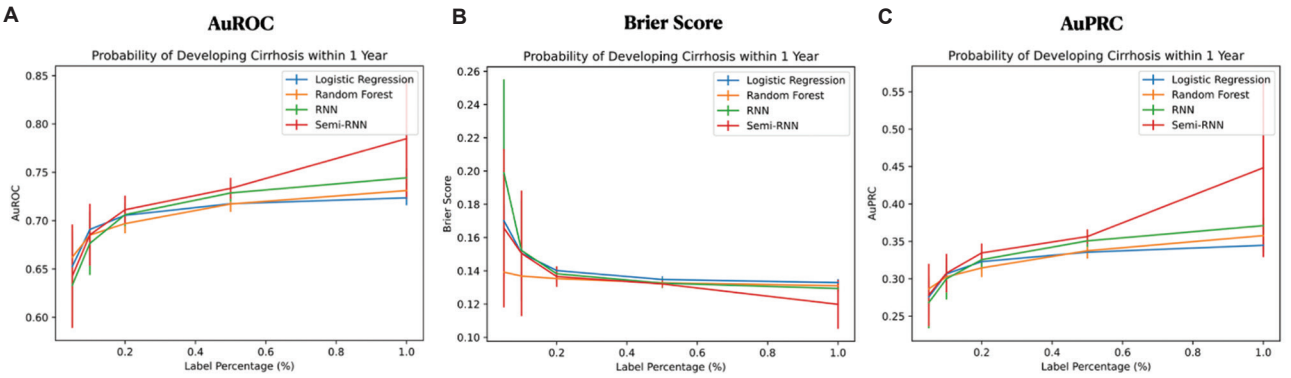

Figure 3. Model performance when various percentages of the training and validation set was used for the labeled cohort. We tested the four developed models using 10 identical testing sets, each with best hyperparameters selected during the validation process. To evaluate performance, we used three different metrics: (A) AuROC; (B) Brier score; and (C) AuPRC. We then calculated the mean of these metrics across all 10 splits to represent the performance of each method using a certain percentage of training and validation data from the labeled cohort.
Abbreviations: AuROC: Area under the receiver operating characteristic curve; AuPRC: Area under the precision-recall curve; RNN: Supervised recurrent neural network; Semi-RNN: Semi-supervised recurrent neural network.

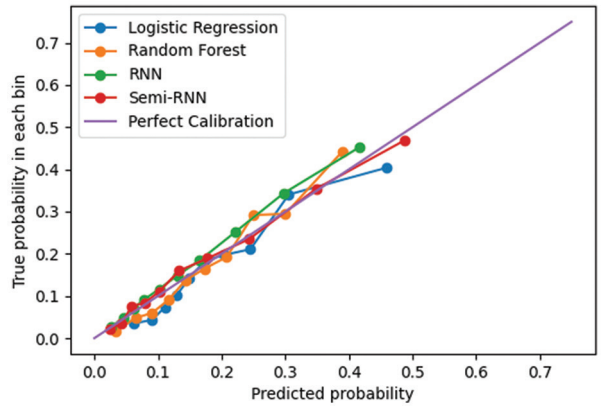

Figure 4. Calibration curve of models when 100% of the training and validation sets was used for the labeled cohort. We selected a representative split that had an AuROC closest to the mean value across 10 splits for the RNN model. Then, we generated the calibration curves for the four models we developed, using 10 quantiles, and compared them to the perfect calibration line.
Abbreviations: AuROC: Area under the receiver operating characteristic curve; RNN: Supervised recurrent neural network; Semi-RNN: Semi-supervised recurrent neural network.

4. Discussion

Our findings suggest that machine learning methods could be useful in identifying patients at high risk for progression to cirrhosis, as defined by TE outcomes. We developed and compared four machine learning models for predicting cirrhosis development, including three supervised learning models which were trained using only patients with known TE outcomes, and a semi-supervised learning model that incorporated both patients with and without TE outcomes for training. We found that RNN models, which are good at processing sequential data, can accurately predict the progression to cirrhosis among patients with CHC under both supervised and semi-supervised learning. In particular, semi-supervised RNN achieved a better predictive performance when we utilized all of our unlabeled data. Unlike prior studies, our cohort includes patients who had received prior HCV treatment, given that the risk of fibrosis progression may persist after antiviral therapy due to other patient cofactors.

Our results demonstrate that machine learning models are feasible options for estimating cirrhosis progression risk in large populations. We anticipate many potential uses, especially in guiding outreach interventions for people with the greatest risk for progression to cirrhosis. For example, using the RNN model, we determined that 90% of all cirrhosis diagnoses occurred in samples with the highest mean (SD) 67% (2.1%) of risk scores, whereas 80% of cirrhosis occurred in samples with the highest mean (SD) 52% (1.8%) of risk scores. Therefore, the RNN model suggests potential benefit in focusing proactive outreach on the top 52% (or 67%) of samples with the highest risk scores, where 80% (or 90%) of cirrhosis cases developed, respectively (Table 2). Given that cirrhosis is the most important risk factor for liver cancer, as well as the cause for multiple disease complications in its own right, clinical decision support for cirrhosis screening could be implemented based on individualized risk.

The VHA system contains the largest cohort of CHC patients in a single U.S. healthcare system and is an ideal environment for developing machine-learning models for cirrhosis prediction.[3] Nevertheless, our results should be interpreted within the context of several limitations. Most importantly, we used TE as a proxy for cirrhosis, as opposed to liver biopsy – the historical gold standard. While TE

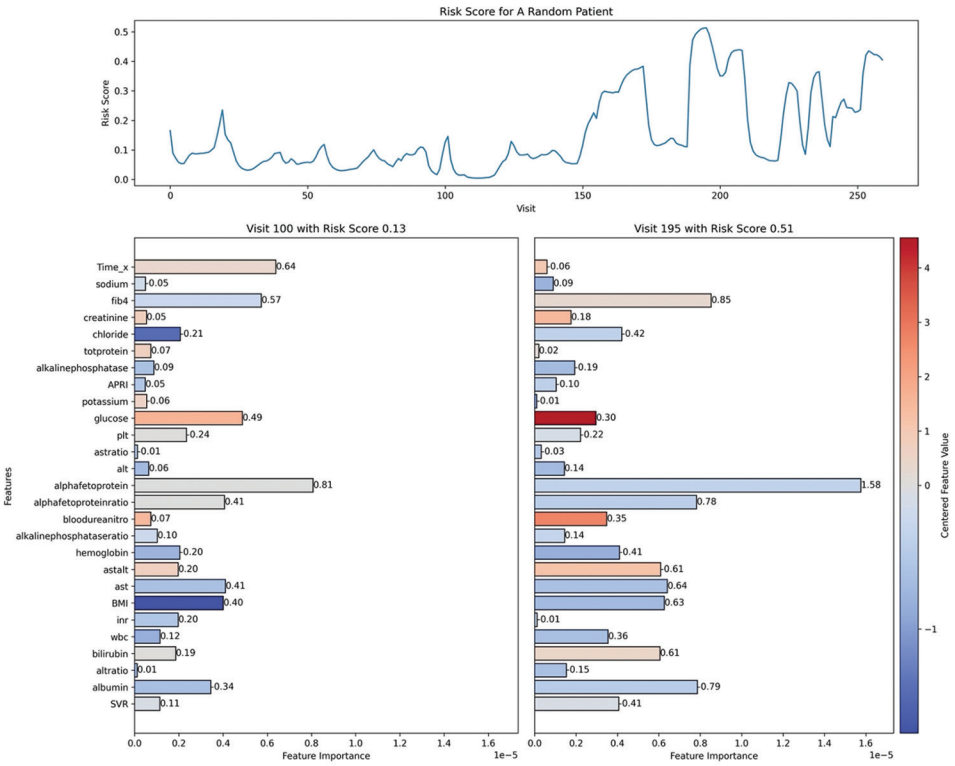

Figure 5. Feature attribution at two visits for an exemplary patient from the representative RNN model when 100% of the training and validation sets was used for the labeled cohort. The length of each bar represents the absolute value of the feature gradient with respect to the loss function at the specific visit, with the number next to each bar indicating the corresponding absolute gradient. The color of the bar reflects the centered feature value: red represents values above the mean, and blue represents values below the mean. The deeper the color, the greater the deviation from the mean, indicating more extreme values of the feature.
Abbreviation: RNN: Supervised recurrent neural network.

has robust performance and has largely supplanted liver biopsy in the assessment of fibrosis in CHC before antiviral treatment, controversy remains in the performance of TE after antiviral therapy.

Second, our data from the VHA population is inherently imbalanced with a majority of male subjects, potentially limiting the generalizability of results to other populations with CHC, as cirrhosis risk factors can vary significantly between genders. For example, women tend to have a different metabolic response to alcohol and certain hepatotoxic medications, leading to a lower threshold for liver damage compared to men.[24] In addition, autoimmune liver diseases, such as primary biliary cirrhosis and autoimmune hepatitis, are more prevalent in women and may influence their liver disease risk profiles.[25] These gender-specific risk factors may not have been fully captured in our analysis. Given the challenges in obtaining electronic health records from external populations, we addressed this concern by stratifying our predictive performance testing by gender, race, and SVR achievement to assess the impact of sampling bias. Specifically, we divided our test data into different subgroups and evaluated the optimal model selected through hyperparameter tuning. Our findings indicate that the RNN model exhibited better predictive ability for female subjects (0.772) compared to male subjects (0.745), for non-White subjects (0.748) versus White subjects (0.737), and better predictive ability on subjects who did not achieve SVR (0.746) at time *t* compared to those who did (0.742), for the RNN model, as shown in Table 3. These results along with Table S1 suggest that the model's performance remains robust across these non-dominant features, mitigating concerns about generalizability. However, future work with more balanced cohorts is needed to validate our machine learning models in external populations.

Third, our predictor variables were limited to those that could be extracted from the VHA's large administrative healthcare database. This prevented the inclusion of alcohol use as a predictor. However, we were able to include several serologic markers suggestive of ongoing alcohol use, including AST, AST:ALT ratio, and platelets. Similarly, diabetes was not specifically included as a predictor,

Table 3. Comparison of performance metrics of the models predicting cirrhosis development within 1 year in patients at risk when 100% of labeled data were used for different subgroups (RNN)

Characteristic	Male	Female	White	Other race	SVR achieved	SVR not achieved
AuROC	0.745	0.772	0.737	0.748	0.742	0.746
Brier score	0.130	0.093	0.140	0.121	0.123	0.129
AuPRC	0.373	0.373	0.392	0.356	0.409	0.370
Proportion of samples who test positive at 90% sensitivity	0.672	0.549	0.678	0.665	0.693	0.659
Specificity at 90% sensitivity	0.377	0.487	0.376	0.380	0.876	0.393
Positive predictive value at 90% sensitivity	0.239	0.201	0.261	0.219	0.220	0.242
Negative predictive value at 90% sensitivity	0.944	0.947	0.937	0.950	0.925	0.947
Proportion of samples who test positive at 80% sensitivity	0.514	0.427	0.531	0.500	0.503	0.510
Specificity at 80% sensitivity	0.548	0.613	0.534	0.562	0.552	0.552
Positive predictive value at 80% sensitivity	0.278	0.225	0.295	0.261	0.267	0.278
Negative predictive value at 80% sensitivity	0.926	0.941	0.914	0.935	0.924	0.927

Abbreviations: AuROC: Area under the receiver operating characteristic curve; AuPRC: Area under the precision-recall curve; RNN: Supervised recurrent neural network; SVR: Sustained virologic response.

although all serum glucose values and BMI values were incorporated into the longitudinal data.

Our analysis was designed to replicate a common clinical scenario in which a provider must estimate the probability that a specific patient will develop cirrhosis within the upcoming year, based on information available before the time of the visit. In the future, models predicting cirrhosis could potentially be deployed in electronic health records to guide in identifying high-risk patients to target for intervention, such as CHC treatment or intensive lifestyle modification. Further, machine learning models need to be developed to predict the development of cirrhosis in chronic liver diseases other than CHC.

5. Conclusion

Deep learning models using RNN resulted in superior predictive performance than conventional machine learning methods, such as LR and RF with substantive use of labeled data. The performance can be further improved by taking advantage of unlabeled data through semi-supervised learning. Our results suggest that these deep learning models are effective tools for identifying patients at high risk for cirrhosis progression. When integrated into clinical decision-making systems, they could support targeted interventions, potentially improving the management and treatment of CHC in large healthcare systems.

Acknowledgments

None.

Funding

The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Drs. Waljee, Beste, Ioannou, and Su are funded by IIR 16-024 from the United States (U.S.) Department of Veterans Affairs Health Services R&D (HSRD) Service.

Conflict of interest

Grace Su, MD, is an equity owner of Applied Morphomics and Prenovo. Dr. Su has a patent with the University of Michigan regarding image analysis of liver disease. Dr. Su has received funding from the National Institutes of Health, the Department of Veteran Affairs, and the Department of Defense. Dr. Su has received but there is no conflict of interest with the preparation of this manuscript. The remaining authors report they have no competing interests as well.

Author contributions

Conceptualization: Akbar K. Waljee, Weijing Tang, Lauren A. Beste, Ji Zhu, Yueying Hu
Formal analysis: Yueying Hu, Weijing Tang
Investigation: Yueying Hu, Weijing Tang
Methodology: Akbar K. Waljee, Weijing Tang, Ji Zhu, Yueying Hu, Lauren A. Beste
Writing – original draft: Yueying Hu, Lauren A. Beste, Weijing Tang
Writing – review & editing: All authors

Ethics approval and consent to participate

Approval to conduct the study was gained from the Institutional Review Board at the VA Ann Arbor Healthcare System, and informed consent from patients was waived.

Consent for publication

Not applicable.

Availability of data

These analyses were performed using data from the Corporate Warehouse Domains that are available only within the U.S. Department of Veterans Affairs firewall in a secure research environment, the VA Informatics and Computing Infrastructure (VINCI). To comply with VA privacy and data security policies and regulatory constraints, only aggregate summary statistics and results of our analyses are permitted to be removed from the data warehouse for publication. The authors have provided detailed results of the analyses in the paper. These restrictions are in place to maintain veteran privacy and confidentiality. Access to these data can be granted to persons who are not employees of the VA; however, there is an official protocol that must be followed for doing so. The authors also confirm that VA policies are currently being developed that should allow an interested researcher to obtain a de-identified, raw dataset upon request with a data use agreement. Those wishing to access the data that were used for this analysis may contact Jennifer Burns, MHSA, who is a senior data manager at the VA Center for Clinical Management Research, to discuss the details of the VA data access approval process. Her contact information is as follows: Jennifer.Burns@va.gov; UM North Campus Research Complex, Department of Veterans Affairs, 2800 Plymouth Road Bldg 16, Ann Arbor, MI.

References

1. Zeremski M, Dimova RB, Pillardy J, de Jong YP, Jacobson IM, Talal AH. Fibrosis progression in patients with chronic hepatitis C virus infection. *J Infect Dis*. 2016;214(8):1164-1170.

   doi: 10.1093/infdis/jiw332

2. Freeman AJ, Law MG, Kadlor JM, Dore GJ. Predicting progression to cirrhosis in chronic hepatitis C virus infection. *J Infect Dis*. 2003;10(4):285-293.

   doi: 10.1046/j.1365-2893.2003.00436.x

3. Waljee AK, Higgins PD. Machine learning in medicine: A primer for physicians. *Am J Gastroenterol*. 2010;105(6):1224-1226.

   doi: 10.1038/ajg.2010.173

4. Konerman MA, Beste LA, Van T, *et al*. Machine learning models to predict disease progression among veterans with hepatitis C virus. *PLoS One*. 2019;14(1):e0208141.

   doi: 10.1371/journal.pone.0208141

5. Singal AG, Mukherjee A, Elmunzer BJ, *et al*. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. *Am J Gastroenterol*. 2013;108(11):1723-1730.

   doi: 10.1038/ajg.2013.332

6. Chen YW, Luo J, Dong C, *et al*. Computer-aided diagnosis and quantification of cirrhotic livers based on morphological analysis and machine learning. *Comput Math Methods Med*. 2013;2013:264809.

   doi: 10.1155/2013/264809

7. Konerman MA, Zhang Y, Zhu J, *et al*. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology*. 2015;61:1832-1841.

   doi: 10.1002/hep.27750

8. Weerakody PB, Wong KW, Wang G, Ela W. A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*. 2021;441:161-178.

   doi: 10.1016/j.neucom.2021.02.046

9. *Veterans Health Administration*. Available from: https://www.va.gov/health [Last accessed on 2024 Dec 22].

10. *Corporate Data Warenouse (CDW): US Department of Veterans Affairs*. Available from: https://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm [Last accessed on 2014 Mar 28].

11. Oliveira AC, El-Bacha I, Vianna MV, Parise ER. Utility and limitations of APRI and FIB4 to predict staging in a cohort of nonselected outpatients with hepatitis C. *Ann Hepatol*. 2016;15(3):326-332.

    doi: 10.5604/16652681.1198801

12. Lin ZH, Xin YN, Dong QJ, *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(3):726-736.

    doi: 10.1002/hep.24105

13. Ioannou GN, Tang W, Beste LA, *et al*. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. *JAMA Netw Open*. 2020;3(9):e2015626.

    doi: 10.1001/jamanetworkopen.2020.15626

14. Chung, J, Gulcehre, C, Cho KH, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv. Preprint posted online 2014.

    doi: 10.48550/arXiv.1412.3555

15. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv. Preprint posted online 2014.

    doi: 10.48550/arXiv.1412.6980

16. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res*.

2014;15(1):1929-1958.

doi: 10.5555/2627435.2670313

17. Tang W, Ma J, Waljee AK, Zhu J. Semi-supervised joint learning for longitudinal clinical events classification using neural network models. *Stat*. 2020;9:e305.

doi: 10.1002/sta4.305

18. Stekhoven D, Bühlmann P. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2012;28(1):112-118.

doi: 10.1093/bioinformatics/btr597

19. Boyd K, Eng KH, Page CD. Area under the precision-recall curve: Point estimates and confidence intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, editors. *Machine Learning and Knowledge Discovery in Databases*. Springer; 2013. p. 451-466.

doi: 10.1007/978-3-642-40994-3_29

20. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

doi: 10.5555/1953048.2078195

21. Paszke A, Gross S, Chintala S, *et al*. *Automatic Differentiation in Pytorch*. Available from: https://openreview.net/pdf?id=BJJsrmfCZ [Last accessed on 2023 Feb 18].

22. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18(17-18):2529-2545.

doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

23. Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in NLP. arXiv. Preprint posted online 2016.

doi: 10.48550/arXiv.1506.01066

24. Becker U, Deis A, Sørensen TI, *et al*. Prediction of risk of liver disease by alcohol intake, sex, and age: A prospective population study. *Hepatology*. 1996;23(5):1025-1029.

doi: 10.1002/hep.510230513

25. Liberal R, Grant C, Mieli-Vergani G, Vergani D. Autoimmune hepatitis: A comprehensive review. *J Autoimmun*. 2013;41:126-139.

doi: 10.1016/j.jaut.2012.11.002