Development and Validation of a Machine Learning-based screening Algorithm to Predict High Risk of Hepatitis C Infection

Suk-Chan Jang, PharmD, PhD[1]; Pilar Hernández Con, MD[1]; Chanakan Jenjai, PharmD[1]; Ashley Stultz, BS[1]; Shunhua Yan, MEd[1]; Debbie L. Wilson, PhD[1]; Wei-Hsuan Lo-Ciganic, PhD[2,3]; James Huang, PhD [1]; Ashley Norse, MD[4]; Faheem Guirgis, MD[1]; Robert L. Cook, MD, MPH[1]; Christine Gage, DO[4]; Khoa Nguyen, PharmD[1]; David R. Nelson, MD[1]; Haesuk Park, PhD[1*]

[1]University of Florida, Gainesville, FL; [2]University of Pittsburgh, Pittsburg, PA; [3]North Florida/South Georgia Veterans Health System GRECC, Gainesville, FL; [4]University of Florida, Jacksonville, FL

EPH75

BACKGROUND

- Hepatitis C virus (HCV) infections are rising sharply in the United States amid the opioid epidemic.
- Despite this increase, HCV often remains undiagnosed due to its asymptomatic nature, leaving approximately one-third to one-half of individuals with HCV infection unaware of their infections.
- **Objective:** To develop and validate a machine learning (ML)-based algorithm to screen individuals at high risk of HCV infection.

METHODS

- **Design: Prognostic modeling with retrospective cohort data**
- **Source:** 2016-2023 OneFlorida+ database (all-payer electronic health records)
- **Study population:** Individuals aged 18 to 79 years who were tested for HCV (antibody, RNA, or genotype)
  - Exclusion criteria: a prior diagnosis of HCV or treatment for HCV before first HCV test
  - Case: Individuals having a positive RNA or genotype test result
  - Control: Individuals having a negative antibody or RNA test result with no history of positive test results
- **Index Date:** Date of the first positive RNA or genotype test result for the cases and the last negative antibody or RNA test result for controls
- **Study outcome:** HCV infection
- **Predictor candidates:** 275 potential predictors during a 6-month baseline period including sociodemographic and clinical characteristics (e.g., comorbidities, procedures, medications)
- **ML approaches:** Elastic net (EN), random forest (RF), gradient boosting machine (GBM), and deep neural network (DNN)
  - Stratification of individuals into deciles based on predicted risk

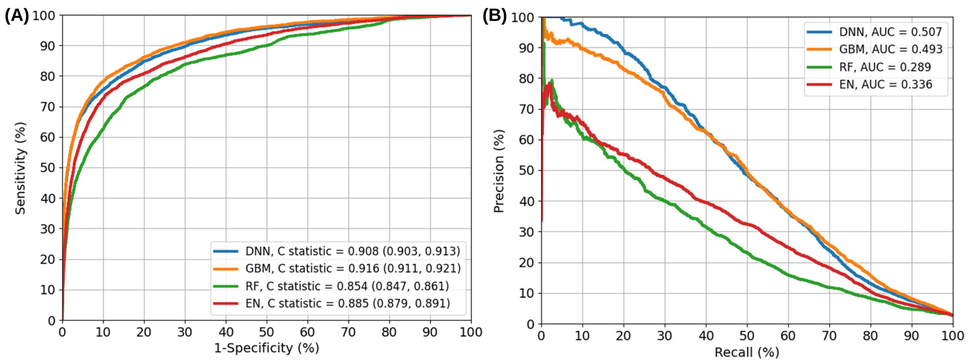
RESULTS

- Among 445,624 individuals, 11,823 (2.65%) tested positive for HCV.
- Training (75%) and validation (25%) samples had similar characteristics **(Table 1).**
- The GBM model (C statistic, 0.916 [95% CI, 0.911-0.921]) outperformed the EN (0.885 [0.879-0.891]), RF (0.854 [0.847-0.861]) and DNN (0.908 [0.903-0.913]) models (P<0.0001) (**Figure 1**). Using the Youden index, GBM achieved 79.39% sensitivity and 89.08% specificity, identifying one positive HCV case per six tests.
- 75.63% and 90.25% of HCV cases were captured within the top first and third risk deciles, respectively (**Figure 2**).
- Key risk predictors included being White, older age, history of HIV, injection drug use, substance use disorder diagnosis and undergoing HIV testing, and more emergency department visits (**Figure 3**).

Table 1. Baseline Sociodemographic and Clinical Characteristics Stratified by Training and Validation Sample

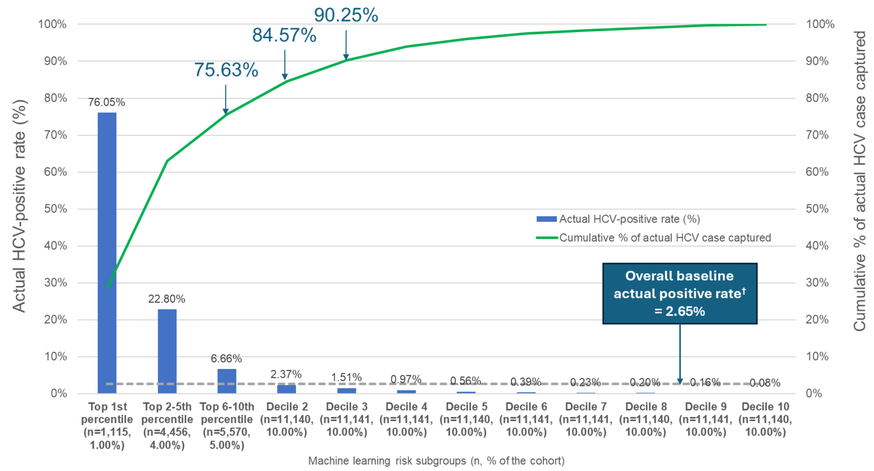
	Training (n=334,218)	Validation (n=111,406)
Age, mean (SD)	45.3 (15.9)	45.4 (15.9)
Male, n (%)	124,118 (37.1)	41,247 (37.0)
Race/ethnicity, Black	88,459 (26.5)	20,604 (26.6)
Payer type, n (%), Medicaid	62,235 (18.6)	20,604 (18.5)
Smoking, n (%)	40,551 (12.1)	13,567 (12.2)
Cirrhosis, n (%)	5,093 (1.5)	1,732 (1.6)
HIV testing, n (%)	60,182 (18.0)	20,245 (18.2)
HIV, n (%)	4,369 (1.3)	1,396 (1.6)
Injection drug use (opioid), n (%)	4,622 (1.4)	1,537 (1.4)
Injection drug use (others), n (%)	5,432 (1.6)	1,868 (1.7)
Substance use disorder (others), n (%)	8,943 (2.7)	3,083 (2.8)
Depression, n (%)	12,466 (3.7)	4,159 (3.7)

Figure 1. Performance of Machine Learning Models for HCV Infection



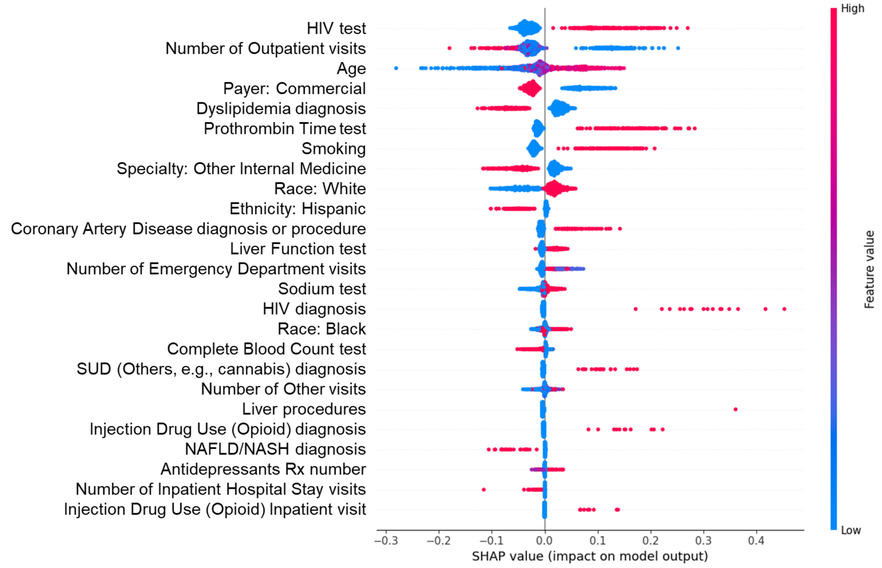
(A) Areas under the receiver operating characteristic curves (AUC), with C statistics and 95% confidence intervals.
(B) Precision-recall curves (precision = positive predictive value, and recall = sensitivity). Precision-recall curves for methods closer to the upper-right corner or above another method have improved performance.

Figure 2. HCV Infection by Risk Subgroup Assessed Using the Gradient Boosting Machine Model With the Validation Sample



† Universal screening for all individuals

Figure 3. Top 25 Important Predictors for HCV Infection Selected by the Gradient Boosting Machine Model



The y-axis lists each feature, and the x-axis indicates the SHapley Additive exPlanations value; the color bar on the right side illustrates the descending order of the average impact on HCV infection for each feature, showing positive (red) or negative (blue) correlations, with overlapping points jittered along the y-axis.

Conclusions

- This study demonstrated the efficacy of ML algorithms, particularly a GBM model, for identifying individuals at high risk of HCV infection.
- The findings of this study underscore the potential for implementing targeted screening in clinical settings to improve HCV screening as part of public health strategies.

Contact Email:
sukchan.jang@ufl.edu