DR. VINCENT LINGZHI CHEN (Orcid ID : 0000-0002-0157-6066)
DR. MINDIE H. NGUYEN (Orcid ID : 0000-0002-6275-4989)

Author Manuscript

Article type : Original

Title: Optimizing Hepatitis B Virus Screening in the United States Using a Simple Demographics-Based Model

Authors: Nathan S. Ramrakhiani*[1], Vincent L. Chen*[2], MD, MS, Michael Le[1], MS, Yee Hui Yeo[1,3], MD MSc, Scott D. Barnett, PhD[1], Akbar K. Waljee[2,4], MD, Ji Zhu, PhD[5], Mindie H. Nguyen[1,6], MD, MAS

*Denotes co-first-authorship; Nathan S. Ramrakhiani and Vincent L. Chen contributed equally to this paper.

Authors' institutions:

[1]Division of Gastroenterology and Hepatology, Stanford University Medical Center, Palo Alto, CA, United States

[2]Division of Gastroenterology and Hepatology, University of Michigan, Ann Arbor, MI, United States

[3]Division of General Internal Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, United States

[4]Division of Gastroenterology and Hepatology, Veterans Affairs Ann Arbor Health System, Ann Arbor, MI, United States

[5]Department of Statistics, University of Michigan, Ann Arbor, MI, United States

[6]Department of Epidemiology and Population Health, Stanford University Medical Center, Palo Alto, CA, United States

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/HEP.32142

This article is protected by copyright. All rights reserved

Author emails (in authorship order): nathanrsc1@gmail.com, vichen@med.umich.edu, mihule@stanford.edu, yeehuiy@stanford.edu, scottdb65@gmail.com, awaljee@med.umich.edu, jizhu@umich.edu, mindiehn@stanford.edu

Keywords: targeted screening, viral hepatitis, random forest model, NHANES, vaccination

Author Manuscript

Corresponding author contact information:
Mindie H. Nguyen, MD, MAS
Professor of Medicine (GI, Hepatology, & Liver Transplant) and by courtesy of Epidemiology and Population Health
750 Welch Road, #215
408-431-7567 (research coordinator telephone number)
Palo Alto, CA 94304
mindiehn@stanford.edu

Abbreviations: Chronic hepatitis B (CHB), hepatitis B virus (HBV), hepatitis B surface antigen (HBsAg), National Health and Nutrition Examination Survey (NHANES), adjusted odds ratio (aOR), area under the receiver operating characteristic (AUROC), hepatocellular carcinoma (HCC), electronic health records (EHR), positive predictive value (PPV), negative predictive value (NPV)

Declaration of interest:
There was no external funding to disclose in our research.
Personal disclosures:
Mindie H. Nguyen: Research support: Glycotest, Gilead, Enanta, Pfizer, Vir, B.K. Kee Foundation, National Cancer Institute. Advisory board/consulting: Intercept, Novartis, Spring Bank, Gilead, Janssen, Eisai, Bayer, Laboratory of Advanced Medicine, Helio Health, Eli Lilly.

This article is protected by copyright. All rights reserved

All other authors have nothing to disclose.

Specific author contributions:

Study design: Nathan Ramrakhiani, Vincent Chen, Yee Hui Yeo, Mindie H. Nguyen
Data collection: Nathan Ramrakhiani, Michael Le, Mindie Nguyen
Data analysis: Vincent Chen, Nathan Ramrakhiani, Mindie H. Nguyen
Manuscript drafting: Nathan Ramrakhiani, Vincent Chen, Mindie H. Nguyen
Data interpretation and review and revision of the manuscript: All authors
Study concept and study supervision: Mindie H. Nguyen

Author Manuscript

ABSTRACT:

Background & Aims: Chronic hepatitis B (CHB) affects over 290 million people globally and only 10% have been diagnosed, presenting a severe gap that must be addressed. We developed logistic regression and machine learning (random forest) models to accurately identify patients with HBV, using only easily-obtained demographic data from a population-based data set.

Approach & Results: We identified participants with data on hepatitis B surface antigen (HBsAg), birth year, sex, race/ethnicity, and birthplace from 10 cycles of the National Health and Nutrition Examination Survey (NHANES, 1999-2018) and divided them into two cohorts: training (cycles 2, 3, 5, 6, 8, 10; n = 39,119) and validation (cycles 1, 4, 7, 9; n = 21,569). We then developed and tested our two models. The overall cohort was 49.2% male, 39.7% White, 23.2% Black, 29.6% Hispanic, and 7.5% Asian/Other, with a median birth year of 1973. In multivariable logistic regression, the following factors were associated with HBV infection: birth year 1991 or after (adjusted OR [aOR] of 0.28, $P < 0.001$), male sex (aOR 1.49, $P = 0.0080$), Black and Asian/Other vs. White (aOR 5.23 and 9.13, $P < 0.001$ for both), and being United States-born (vs. foreign-born) (aOR 0.14, $P < 0.001$). We found that the machine learning model consistently outperformed the logistic regression model, with higher AUROC values (0.83 vs. 0.75 in validation cohort, $P < 0.001$) and better differentiation of high and low risk individuals.

This article is protected by copyright. All rights reserved

Conclusions: Our machine learning model provides a simple, targeted approach to HBV screening, using only easily-obtained demographic data.

Author Manuscript

INTRODUCTION:

Chronic hepatitis B (CHB) is a major global public health concern affecting 290 million people, but only 10% have been diagnosed worldwide.[1] In the United States, CHB affects an estimated 840,000 to 1.59 million people[2, 3] with population-based studies reporting a patient disease awareness and diagnosis rate of only 15-19%.[4, 5, 6] While CHB can progress to cirrhosis, hepatic failure, and hepatocellular carcinoma (HCC), many patients remain asymptomatic until onset of end-stage liver disease secondary to cirrhosis and/or HCC,[7, 8, 9]further contributing to the observed low diagnosis and awareness rates. Delayed diagnosis consequently leads to delayed initiation of antiviral therapies that have been shown to be well tolerated and effective in preventing the development of cirrhosis, HCC, and premature death.[10, 11]

This severe underdiagnosis of CHB has persisted, despite guidelines recommending screening for high-risk individuals since the early 2000s (**Supplementary Table 1**),[12, 13, 14, 15] and this affirms the need for a simpler, more practical approach to screening and diagnosis of hepatitis B virus (HBV) infection. In low-prevalence areas such as the United States or Western Europe, a universal approach to HBV screening is unlikely to be cost-effective.[12] Meanwhile, due to advances in hepatitis B vaccination policy worldwide, the large majority of the CHB burden in the United States occurs in immigrants and older individuals, giving rise to an opportunity for a "semi-universal" screening approach that focuses on specific demographic groups. A "semi-targeted" approach based on a small number of demographic characteristics that are easily obtained from electronic health records (EHR), such as age, sex, race/ethnicity, and birthplace, may enhance CHB screening and diagnosis, due to greater simplicity and data availability, as well as less reliance on culturally sensitive and/or stigmatizing risk assessment questions (e.g. injection drug use, men having sex with men, etc.) (**Supplemental Table 1**).[12, 13, 14, 15]

This article is protected by copyright. All rights reserved

Therefore, using a nationally-representative sample of the non-institutionalized United States civilian population, we sought to develop a data-driven, population-based screening algorithm to accurately identify HBV infection, using only routinely-collected and easily-obtained demographic data.

METHODS:

Data source and study population:

We used data obtained from the National Health and Nutrition Examination Survey (NHANES) database, which consists of a series of nationally representative cross-sectional studies performed by the Centers for Disease Control and Prevention National Center for Health Statistics (NCHS, 2001-2018) in 2-year cycles. NHANES collects data from a complex multistage, stratified, clustered probability sample that is representative of the noninstitutionalized, civilian population of the United States. Use of the data from NHANES allows for the assessment of various health and nutritional complications of adults and children in the United States. NHANES collects data through comprehensive written questionnaires, physical examinations, and biological samples. NHANES data can be downloaded from the NCHS website (https://www.cdc.gov/nchs/nhanes.htm). All participants gave written informed consent, and the NHANES survey is administered by the Centers for Disease Control.

Our study participants were from NHANES 1999-2018 (10 cycles). We excluded patients with missing hepatitis B surface antigen test (HBsAg) data and those with incomplete demographic data (birth year, sex, race/ethnicity, and birthplace) (**Figure 1**). We further divided the study cohort into a training cohort (NHANES cycles 2, 3, 5, 6, 8, 10; n = 39,119) and a validation cohort (NHANES cycles 1, 4, 7, 9; n = 21,569) to develop and compare two potential algorithms, as detailed below.

Logistic regression model:

We created logistic regression models[16] with HBV infection (defined as positive HBsAg) as the primary outcome and demographic variables (birth year, sex, race/ethnicity, and birthplace [United States vs. foreign-born]) as the primary predictors, with both univariable and multivariable logistic regression, in the training set. We then created a logit score to estimate risk

This article is protected by copyright. All rights reserved

Author Manuscript

for HBsAg seropositivity that included all variables that were significantly associated with positive HBsAg at *P*-value < 0.05 on multivariable regression and was weighted by the beta coefficients corresponding to those variables. This score was created in the training cohort, and we then assessed this model's performance in the validation set.

Random forest model:

We used random forest models[17] to determine demographic factors (birth year, sex, race/ethnicity, and birthplace [United States vs. foreign-born]) that were associated with the primary outcome, HBV infection. Individuals with missing relevant demographic or HBsAg data were excluded. We used the *party* package version 1.3.3 in R with tune length 5 and a fixed seed. We generated a model using the training cohort with down-sampling of the controls (given how unbalanced the set was for HBsAg status) and ten-fold cross-validation to determine test characteristics of the model in the training set. Because we conducted down-sampling, the initial model was poorly-calibrated, so we calibrated the model with a Platt scaling (logistic regression of the risk predicted by the random forest model to the outcome of positive HBsAg) in the training set. We then validated the model in the independent validation cohort without down-sampling.[18]

Comparison of models:

We compared the random forest and logistic regression models in two ways. First, we compared the area under the receiver operating characteristic curve (AUROC) values using the De Long test. Second, we divided participants into deciles of predicted risk based on the logistic regression vs. random forest model, and compared the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the top 20% vs. the bottom 80% of predicted risk. These cutoffs were obtained by observing that actual HBV infection prevalence was far higher in the top 20% of predicted risk in both the logistic regression and random forest models. Confidence intervals for sensitivity and specificity were generated using the Clopper-Pearson method,[19] while confidence intervals for PPV and NPV were based on logit confidence intervals.[20]

Other statistical analysis:

This article is protected by copyright. All rights reserved

Author Manuscript

Descriptive statistics were reported as median (interquartile range) or %. Continuous variables were compared using a Wilcoxon rank-sum test, and categorical variables with a chi-squared test. All analyses were performed using R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria). Two-tailed *P*-values < 0.05 were considered significant.

RESULTS:

Study population:

As shown in **Figure 1**, we obtained the entire 1999-2018 NHANES cohort (N = 101,316), and then excluded patients with missing HBsAg data (n = 40,601) or data on birth year, sex, race/ethnicity, or birthplace (n = 27). In total, we included 60,688 patients with all required data in study analysis. The study cohort was then divided into a training cohort (NHANES cycles 2, 3, 5, 6, 8, 10; n = 39,119), from which we derived both the logistic regression and random forest models, and a validation cohort (NHANES cycles 1, 4, 7, 9; n = 21,569), in which the two models were tested.

As shown in **Table 1**, the overall cohort was 49.2% male, 39.7% White, 23.2% Black, 29.6% Hispanic, and 7.5% Asian/Other, and with a median birth year of 1973. HBsAg-positive participants were more often male (58.1% vs. 49.2%, $P = 0.0034$) and older (median birth year 1960 vs. 1973, $P < 0.001$). In addition, the racial distribution differed significantly between the two groups: the HBsAg-positive group were less likely to be White (11.1% vs. 39.8%, $P < 0.001$) or Hispanic (6.6% vs. 29.7%, $P < 0.001$) patients, and more likely to be Black (34.1% vs. 23.1%, $P < 0.001$) or Asian/Other (48.1% vs. 7.3%, $P < 0.001$), compared to HBsAg-negative participants. In addition, the HBsAg positive group were less likely to be born in the United States than the HBsAg-negative group (35.9% vs. 78.9%, $P < 0.001$).

Development of algorithms to identify HBV

First, we generated logistic regression and random forest models for a semi-targeted screening approach using only the above significant demographic factors, namely sex, year of birth, race/ethnicity, and birthplace.

This article is protected by copyright. All rights reserved

Author Manuscript

In multivariable logistic regression analysis (**Table 2)**, all four demographic factors considered were significantly associated with HBV infection. A birth year of 1991 and after (the first year of the universal HBV vaccination recommendation in the United States[21]) corresponded to 72% lower odds of HBV infection (adjusted OR [aOR] of 0.28, 95% CI 0.14-0.55, $P < 0.001$). Male sex was associated with 49% higher odds of HBV infection (aOR 1.49, 95% CI 1.11-2.01, $P = 0.0080$). Compared to White as reference, Black and Asian/Other were associated with more than 5-9 times the odds of having HBV infection, respectively (Black: aOR 5.23, 95% CI 3.10-8.83, $P < 0.001$; Asian/Other: aOR of 9.13, 95% CI 5.23-15.96, $P < 0.001$). Meanwhile, Hispanic ethnicity was associated with 66% lower odds of HBV infection (aOR of 0.34, 95% CI 0.16-0.71, $P = 0.0044$). Being born in the United States (vs. foreign-born) also corresponded to 86% lower odds of HBV infection (aOR 0.14, 95% CI 0.10-0.21, $P < 0.001$). The equation for the logistic regression-based score was -5.17 + 0.40 (if male) + 2.21 (if Asian/other race) + 1.65 (if Black race) - 1.27 (if born 1991 or later) - 1.94 (if born in the United States). It is not possible to display the corresponding score for a random forest model in a readily-interpretable closed form.

Comparison of logistic and machine learning models

Next, we compared the accuracy of the two models in predicting HBsAg status. **Figure 2** displays the ROC curves for the logistic regression and machine learning models in both the training and validation cohorts. In the training cohort, the AUROC was significantly higher for the machine learning model at 0.90 (95% CI: 0.88-0.92) vs. 0.81 (95% CI 0.79-0.84) for the logistic regression model ($P < 0.001$ by De Long test). In the validation cohort, the AUROC was also higher with the machine learning model as compared to the logistic regression model (0.83, 95% CI: 0.78-0.88 vs. 0.75, 95% CI: 0.70-0.80, $P < 0.001$). While the initial machine learning model was poorly-calibrated due to the down-sampling used initially (log loss 1.37), after applying Platt scaling the model became reasonably well-calibrated (log loss 0.02) (**Supp. Figure 1**).

Furthermore, when we evaluated the efficacy of the two models for differentiating participants with a high likelihood of HBV infection from those with a low likelihood by grouping participants into deciles based on their risk of HBV infection (as estimated by either the logistic

This article is protected by copyright. All rights reserved

Author Manuscript

regression or the machine learning model) and analyzing the percentage of participants with HBV infection within each estimated risk decile (**Figure 3),** the machine learning model was also more effective at differentiating the high risk from low risk participants. In both the training (top panes) as well as the validation cohorts (bottom panes), the prevalence of HBV infection was higher in the top 20% of risk with machine learning model (training: 2.1% in machine learning vs. 1.7% in logistic regression model; validation: 1.4% vs. 1.1%) and the percentage of HBV infection was also lower in the bottom 80% with the machine learning model (training: 0.08% in machine learning vs. 0.18% in logistic regression model; validation: 0.12% vs. 0.20%). In addition, with both the training and validation sets, the machine learning model had higher sensitivity, PPV, NPV, and AUROC than the logistic regression model ($P < 0.005$ for all), with no difference in specificity (**Table 3**).

DISCUSSION:

To our knowledge, this is the first study to develop an algorithm to prioritize patients for hepatitis B screening using data from a population-based cohort in the United States and relying only on demographic data that are routinely available in typical healthcare delivery settings. We found that the machine learning model consistently outperformed the logistic regression model, with higher AUROC values (0.83 vs. 0.75 in validation cohort) and more effective identification of high-risk patients (1.4% vs. 1.1% seroprevalence in the top 20% in the machine learning and logistic models, respectively).

To reiterate, CHB is a major public health concern and a significant cause of morbidity and mortality, but it is estimated that worldwide 90% and in the United States 80% of people with CHB have not been diagnosed.[1, 3, 5, 6] As a result, opportunities for HCC surveillance and antiviral therapy to prevent HCC and end-stage liver disease are lost. Identifying patients with HBV also allows for targeted vaccination of family members, partners, and other close contacts, an inexpensive and effective way to prevent HBV transmission.

While universal screening of adults for hepatitis C virus (HCV) has been recommended by the Centers for Disease Control[22] and is likely cost-effective,[23, 24] universal screening for HBV in the United States is unlikely to be cost-effective given its lower prevalence (0.3-0.5%) among

This article is protected by copyright. All rights reserved

Author Manuscript

the general population,[3, 25] below the 2% USPSTF threshold.[12] However, prior studies have found that the prevalence threshold for cost-effective screening programs can be as low as 1% if screening is followed by treatment and vaccination of close contacts.[26] Thus, a semi-targeted approach to identify a higher risk group with an HBV seroprevalence of >1% may allow for a cost-effective semi-universal approach to HBV screening. As such, since our algorithm only requires four simple non-stigmatizing demographic factors, it is likely to be well-received by both patients and care providers. Screening should also be easily implemented because three of these four factors are already part of any medical records (birth year, sex, and race). Birthplace (foreign-born vs. United States-born) can often be gleaned from most EHR systems which usually record a patient's preferred language if different from English, or can be relatively easily added to EHR during patient registration.

We also recognize that our algorithm can miss 13% of HBV infection and some patients not meeting screening criteria by our semi-targeted approach may have significant risk for HBV such as a young person with a history of injection drug use. Therefore, we advocate for implementation of a semi-targeted, semi-universal screening approach such as ours to remedy the current state of CHB diagnosis and linkage to care in the United States while also applying the existing risk-based approach for those not meeting our semi-targeted screening criteria.

Limitations of this study include a relatively small number of patients with HBV infection due to the low HBsAg prevalence (0.44%) in our study population. The lack of an extremely high-risk (>5%) subpopulation may affect calibration so that risk estimates may not generalize to higher risk populations; however, these high-risk persons should already be captured by existing risk-based guidelines. Also, the model is derived from a United States population-based cohort and may not be applicable to other countries, especially more racially-homogenous countries. Strengths include the fact that our study cohort is a nationally representative sample of the non-institutionalized civilian population of the United States, and each cycle of NHANES is independent of the other cycles. Screening results of our random forest machine learning model may be even more efficient at identifying HBV infection in areas with a higher prevalence of infection such as the various metropolitan areas of the West and Northeast of the United States.

This article is protected by copyright. All rights reserved

Author Manuscript

Our models can be readily modified to apply to populations with higher HBsAg seroprevalence by changing the cutoff value for a "positive" screen.

In summary, we developed a data-driven, population-based screening algorithm for HBV infection in the United States, using only demographic data that is routinely collected by healthcare providers and EHR systems. Our machine learning model consistently outperformed the logistic regression model, laying the groundwork for what could eventually be a practical and cost-effective HBV screening strategy for low prevalence regions with more "imported" HBV infection such as the United States or Western Europe. We also advocate for additional risk-based screening for populations with specific exposure risks as per professional society and CDC guidelines. Lastly, we encourage validation of our algorithm in other populations as well as future studies to evaluate the cost-effectiveness of this semi-targeted and semi-universal HBV screening approach.

Author Manuscript

REFERENCES:

1 Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: a modelling study. Lancet Gastroenterol Hepatol 2018;**3**:383-403.

2 Wong RJ, Brosgart CL, Welch S, Block T, Chen M, Cohen C*, et al.* An Updated Assessment of Chronic Hepatitis B Prevalence Among Foreign-Born Persons Living in the United States. Hepatology 2021.

3 Le MH, Yeo YH, Cheung R, Henry L, Lok AS, Nguyen MH. Chronic Hepatitis B Prevalence Among Foreign-Born and U.S.-Born Adults in the United States, 1999-2016. Hepatology 2020;**71**:431-43.

4 Lim JK, Nguyen MH, Kim WR, Gish R, Perumalswami P, Jacobson IM. Prevalence of Chronic Hepatitis B Virus Infection in the United States. Am J Gastroenterol 2020;**115**:1429-38.

5 Ogawa E, Yeo YH, Dang N, Le MH, Jeong D, Tran S*, et al.* Diagnosis Rates of Chronic Hepatitis B in Privately Insured Patients in the United States. JAMA Netw Open 2020;**3**:e201844.

This article is protected by copyright. All rights reserved

6 Yeo YH, Nguyen MH. Review article: current gaps and opportunities in HBV prevention, testing and linkage to care in the United States-a call for action. Aliment Pharmacol Ther 2021;**53**:63-78.

7 Abara WE, Qaseem A, Schillie S, McMahon BJ, Harris AM. Hepatitis B Vaccination, Screening, and Linkage to Care: Best Practice Advice From the American College of Physicians and the Centers for Disease Control and Prevention. Ann Intern Med 2017;**167**:794-804.

8 McMahon BJ. Natural history of chronic hepatitis B. Clin Liver Dis 2010;**14**:381-96.

9 Nguyen MH, Wong G, Gane E, Kao JH, Dusheiko G. Hepatitis B Virus: Advances in Prevention, Diagnosis, and Therapy. Clin Microbiol Rev 2020;**33**.

10 Tan M, Bhadoria AS, Cui F, Tan A, Van Holten J, Easterbrook P*, et al.* Estimating the proportion of people with chronic hepatitis B virus infection eligible for hepatitis B antiviral treatment worldwide: a systematic review and meta-analysis. Lancet Gastroenterol Hepatol 2021;**6**:106-19.

11 Lok AS, McMahon BJ, Brown RS, Jr., Wong JB, Ahmed AT, Farah W*, et al.* Antiviral therapy for chronic hepatitis B viral infection in adults: A systematic review and meta-analysis. Hepatology 2016;**63**:284-306.

12 Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB*, et al.* Screening for Hepatitis B Virus Infection in Adolescents and Adults: US Preventive Services Task Force Recommendation Statement. Jama 2020;**324**:2415-22.

13 Terrault NA, Lok ASF, McMahon BJ, Chang KM, Hwang JP, Jonas MM*, et al.* Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. Hepatology 2018;**67**:1560-99.

14 Weinbaum CM, Williams I, Mast EE, Wang SA, Finelli L, Wasley A*, et al.* Recommendations for identification and public health management of persons with chronic hepatitis B virus infection. MMWR Recomm Rep 2008;**57**:1-20.

15 Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, Caughey AB*, et al.* Screening for Hepatitis B Virus Infection in Pregnant Women: US Preventive Services Task Force Reaffirmation Recommendation Statement. Jama 2019;**322**:349-54.

16 Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009) The Elements of Statistical Learning. Springer.

17 Leo Breiman (2001) Random forests. Machine Learning 45, 5-32.

This article is protected by copyright. All rights reserved

Author Manuscript

18 Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 1999;**10**:61-74.

19 Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;**26**:404-13.

20 Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case–control studies. Statistics in Medicine 2007;**26**:2170-83.

21 Hepatitis B virus: a comprehensive strategy for eliminating transmission in the United States through universal childhood vaccination. Recommendations of the Immunization Practices Advisory Committee (ACIP). MMWR Recomm Rep 1991;**40**:1-25.

22 Schillie S, Wester C, Osborne M, Wesolowski L, Ryerson AB. CDC Recommendations for Hepatitis C Screening Among Adults - United States, 2020. MMWR Recomm Rep 2020;**69**:1-17.

23 Tatar M, Keeshin SW, Mailliard M, Wilson FA. Cost-effectiveness of Universal and Targeted Hepatitis C Virus Screening in the United States. JAMA Netw Open 2020;**3**:e2015756.

24 Chaillon A, Rand EB, Reau N, Martin NK. Cost-effectiveness of Universal Hepatitis C Virus Screening of Pregnant Women in the United States. Clin Infect Dis 2019;**69**:1888-95.

25 Patel EU, Thio CL, Boon D, Thomas DL, Tobian AAR. Prevalence of Hepatitis B and Hepatitis D Virus Infections in the United States, 2011-2016. Clin Infect Dis 2019;**69**:709-12.

26 Hutton DW, Tan D, So SK, Brandeau ML. Cost-effectiveness of screening and vaccinating Asian and Pacific Islander adults for hepatitis B. Ann Intern Med 2007;**147**:460-9.

Author Manuscript

This article is protected by copyright. All rights reserved

Table 1. Baseline characteristics of overall cohort (n = 60,688)

Variable	Overall cohort (n = 60,688)	HBsAg[1] negative (n = 60,418)	HBsAg[1] positive (n = 270)	P-value
Male sex	49.20%	49.20%	58.10%	0.0034
Birth year	1973 (1951-1989)	1973 (1951-1989)	1960 (1949-1974)	<0.001
1911-1930	6.30%	6.30%	5.20%	0.53
1931-1950	18.30%	18.30%	21.10%	0.24
1951-1970	22.50%	22.50%	40.70%	<0.001
1971-1990	32.20%	32.20%	28.90%	0.27
1991-2010	20.60%	20.60%	4.10%	<0.001
Race/ethnicity				
White	39.70%	39.80%	11.10%	<0.001
Black	23.20%	23.10%	34.10%	<0.001
Hispanic	29.60%	29.70%	6.60%	<0.001
Asian or other	7.50%	7.30%	48.10%	<0.001
Birth place: United States	78.70%	78.90%	35.90%	<0.001
Income:poverty line ratio (n = 56,026)	1.9 (1.0-3.7)	1.9 (1.0-3.7)	1.7 (1.0-3.3)	0.26
Body mass index, kg/m^2 (n = 59,783)	25.7 (21.5-30.4)	25.7 (21.5-30.4)	25.0 (22.3-28.6)	0.50
Diabetes (n = 60,648)	7.60%	7.60%	12.20%	0.0078

Author Manuscript

This article is protected by copyright. All rights reserved

Coronary artery disease (n = 39,843)	4.10%	4.10%	2.40%	0.20
Hemoglobin A1c, % (n = 52,908)	5.4 (5.1-5.7)	5.4 (5.1-5.7)	5.5 (5.2-5.9)	<0.001
Glucose, mg/dL (n = 25,487)	96.5 (90.0-105.0)	96.5 (90.0-105.0)	96.7 (90.0-106.5)	0.79
High-density lipoprotein, mg/dL (n = 60,319)	51.0 (42.0-61.0)	51.0 (42.0-61.0)	54.0 (43.2-63.8)	0.027
Low-density lipoprotein, mg/dL (n = 25,521)	106.0 (84.0-131.0)	106.0 (84.0-131.0)	105.5 (86.2-133.8)	0.26
Triglycerides, mg/dL (n = 26,523)	99.0 (67.0-149.0)	99.0 (67.0-149.0)	92.0 (69.8-137.8)	0.59
Alanine aminotransferase, U/L (n = 52,504)	21.0 (16.0-34.0)	21.0 (16.0-34.0)	25.0 (19.0-37.0)	<0.001
Aspartate aminotransferase, U/L (n = 52,498)	22.0 (19.0-27.0)	22.0 (19.0-27.0)	26.0 (21.0-34.0)	<0.001
Alkaline phosphatase, U/L (n = 52,598)	67.0 (50.0-88.0)	67.0 (50.0-88.0)	64.0 (52.5-87.0)	0.65
Gamma-glutamyl transferase, U/L (n = 52,596)	18.0 (13.0-28.0)	18.0 (13.0-27.0)	21.0 (14.0-36.0)	<0.001
Total bilirubin, mg/dL (n = 52,575)	0.6 (0.5-0.8)	0.6 (0.5-0.8)	0.7 (0.5-0.9)	0.039
Creatinine, mg/dL (n = 52,603)	0.8 (0.7-1.0)	0.8 (0.7-1.0)	0.9 (0.7-1.0)	0.0026
Platelets, x 10^9/L (n = 60,601)	259.0 (218.0-305.0)	259.0 (218.0-306.0)	215.0 (173.0-259.0)	<0.001
Systolic blood pressure, mmHg (n = 52,029)	116.0 (106.0-130.0)	116.0 (106.0-130.0)	120.0 (110.0-134.0)	<0.001
Diastolic blood pressure, mmHg (n = 52,209)	68.0 (58.0-76.0)	68.0 (58.0-76.0)	72.0 (64.0-80.0)	<0.001

All data presented as either median (IQR) or %. All variables presented as variable name, units (n = number of patients with complete data).

Author Manuscript

This article is protected by copyright. All rights reserved

Table 2. Logistic regression of predictors of hepatitis B infection, training cohort (n = 39,119)

Variable	Unadjusted		Adjusted	
	Odds ratio (95% CI)	**P-value**	**Odds ratio (95% CI)**	**P-value**
Birth year				
1990 and before	Referent	<0.001	Referent	<0.001
1991 and after	0.19 (0.10-0.37)		0.28 (0.14-0.55)	
Male sex	1.52 (1.13-2.03)	0.0050	1.49 (1.11-2.01)	0.0080
Race				
White	Referent		Referent	
Black	5.78 (3.44-9.72)	<0.001	5.23 (3.10-8.83)	<0.001
Hispanic	0.98 (0.48-1.99)	0.96	0.34 (0.16-0.71)	0.0044
Asian or other	30.95 (18.91-50.68)	<0.001	9.13 (5.23-15.96)	<0.001
Birth place: United States	0.11 (0.08-0.16)	<0.001	0.14 (0.10-0.21)	<0.001

Author Manuscript

This article is protected by copyright. All rights reserved

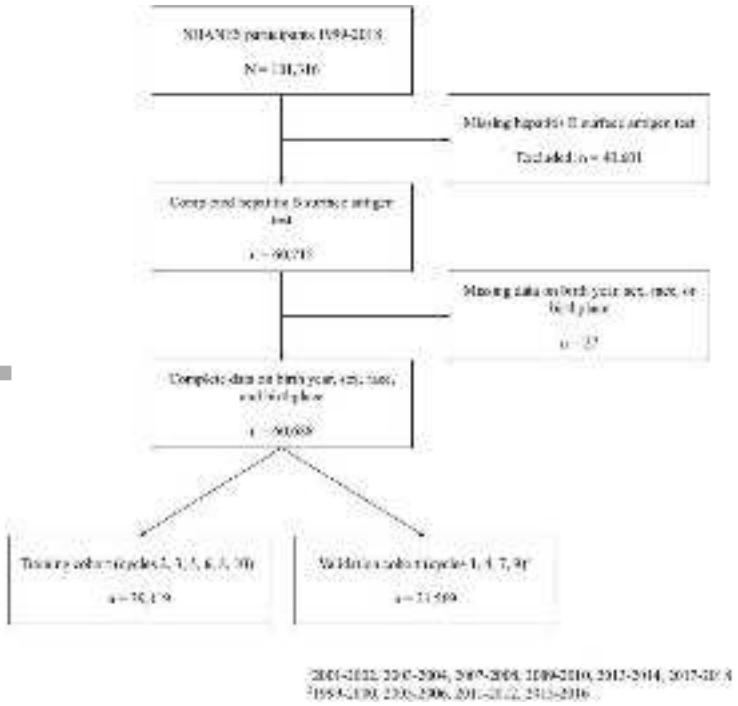
Table 3. Performance characteristics of the machine learning and logistic regression models (positive model result: top 20% score) (n = 60,688)

Test characteristic	Machine learning	Logistic regression	P-value
Training cohort (n = 39,119)			
Sensitivity	86.7% (81.0-91.2%)	69.7% (62.6-76.2%)	<0.001
Specificity	80.3% (79.9-80.7%)	80.2% (79.8-80.6%)	0.35
PPV	2.08% (1.97-2.21%)	1.67% (1.52-1.84%)	<0.001
NPV	99.92% (99.88-99.94%)	99.82% (99.77-99.85%)	<0.001
AUROC (overall)	0.904 (0.885-0.924)	0.814 (0.788-0.839)	<0.001
Validation set (n = 21,569)			
Sensitivity	75.6% (64.9-84.4%)	57.3% (45.9-68.2%)	<0.001
Specificity	80.2% (79.7-80.8%)	80.2% (79.6-80.7%)	0.41
PPV	1.44% (1.27-1.63%)	1.09% (0.90-1.31%)	0.0019
NPV	99.88% (99.83-99.92%)	99.80% (99.74-99.84%)	0.0017
AUROC (overall)	0.828 (0.781-0.876)	0.752 (0.704-0.800)	<0.001

Author Manuscript

This article is protected by copyright. All rights reserved

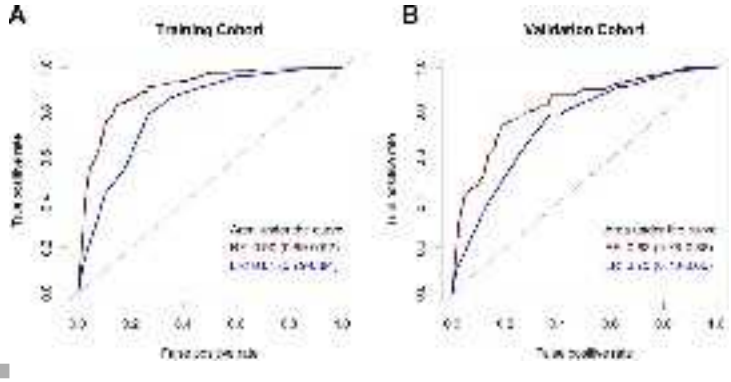
Author Manuscript



hep_32142_f1.tiff

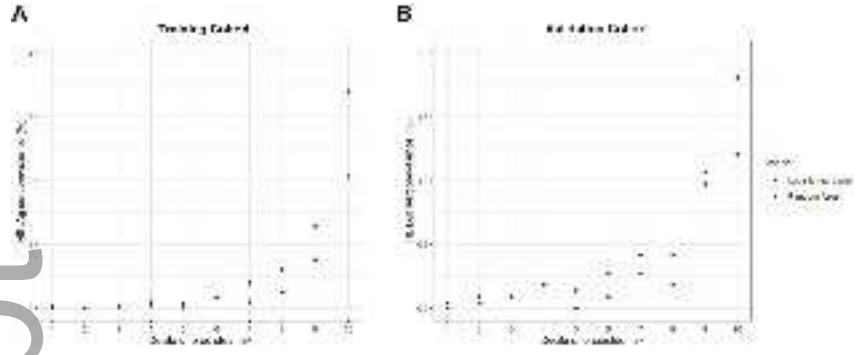
This article is protected by copyright. All rights reserved

Author Manuscript



hep_32142_f2.tiff

This article is protected by copyright. All rights reserved



hep_32142_f3.tiff

Author Manuscript

This article is protected by copyright. All rights reserved