Received: 12 May 2021 | Revised: 20 July 2021 | Accepted: 27 August 2021

DOI: 10.1002/hep.32142

ORIGINAL ARTICLE

# Optimizing hepatitis B virus screening in the United States using a simple demographics-based model

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

[1]Division of Gastroenterology and Hepatology, Stanford University Medical Center, Palo Alto, California, USA

[2]Division of Gastroenterology and Hepatology, University of Michigan, Ann Arbor, Michigan, USA

[3]Division of General Internal Medicine, Cedars-Sinai Medical Center, Los Angeles, California, USA

[4]Division of Gastroenterology and Hepatology, Veterans Affairs Ann Arbor Health System, Ann Arbor, Michigan, USA

[5]Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

[6]Department of Epidemiology and Population Health, Stanford University Medical Center, Palo Alto, California, USA

**Correspondence**
Mindie H. Nguyen, Division of Gastroenterology and Hepatology, Stanford University Medical Center, 750 Welch Road, #215, Palo Alto, CA 94304, USA.
Email: mindiehn@stanford.edu

## Abstract

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

**Approach and Results:** We identified participants with data on HBsAg, birth year, sex, race/ethnicity, and birthplace from 10 cycles of the National Health and Nutrition Examination Survey (1999–2018) and divided them into two cohorts: training (cycles 2, 3, 5, 6, 8, and 10; $n$ = 39,119) and validation (cycles 1, 4, 7, and 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], 0.28; $p < 0.001$); male sex (aOR, 1.49; $p = 0.0080$); Black and Asian/other versus White (aOR, 5.23 and 9.13; $p < 0.001$ for both); and being USA-born (vs. foreign-born; aOR, 0.14; $p < 0.001$). We found that the ML model consistently outperformed the LR model, with higher area under the receiver operating characteristic values (0.83 vs. 0.75 in validation cohort; $p < 0.001$) and better differentiation of high- and low-risk persons.

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

**Abbreviations:** aOR, adjusted odds ratio; AUROC, area under the receiver operating characteristic; CDC, Centers for Disease Control and Prevention; CHB, chronic hepatitis B; EHR, electronic health records; LR, logistic regression; ML, machine learning; NHANES, National Health and Nutrition Examination Survey; NPV, negative predictive value; PPV, positive predictive value; RF, random forest.

Nathan S. Ramrakhiani and Vincent L. Chen hold co-first authorship and contributed equally to this article.

## INTRODUCTION

Chronic hepatitis B (CHB) is a major global public health concern affecting 290 million persons, but only 10% have been diagnosed worldwide.[1] In the USA, CHB affects an estimated 840,000 to 1.59 million persons,[2,3] with population-based studies reporting a patient disease awareness and diagnosis rate of only 15%–19%.[4–6] Whereas CHB can progress to cirrhosis, hepatic failure, and HCC, many patients remain asymptomatic until onset of end-stage liver disease (ESLD) secondary to cirrhosis and/or HCC,[7–9] 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 persons since the early 2000s (Table S1),[12–15] and this affirms the need for a simpler, more practical approach to the screening and diagnosis of HBV infection. In low-prevalence areas, such as the USA or Western Europe, a universal approach to HBV screening is unlikely to be cost-effective.[12] Meanwhile, because of advances in hepatitis B vaccination policy worldwide, the large majority of the CHB burden in the USA occurs in immigrants and older persons, giving rise to an opportunity for a "semiuniversal" screening approach that focuses on specific demographic groups. A "semitargeted" approach based on a small number of demographic characteristics that are easily obtained from electronic health records (EHRs), such as age, sex, race/ethnicity, and birthplace, may enhance CHB screening and diagnosis, because pf 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.; Table S1).[12–15]

Therefore, using a nationally representative sample of the noninstitutionalized USA 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.

## MATERIALS AND 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 (CDC) 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 USA. Use of the data from NHANES allows for the assessment of various health and nutritional complications of adults and children in the USA. 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 CDC.

Our study participants were from NHANES 1999–2018 (10 cycles). We excluded patients with missing HBsAg test 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, and 10; $n$ = 39,119) and a validation cohort (NHANES cycles 1, 4, 7, and 9; $n$ = 21,569) to develop and compare two potential algorithms, as detailed below.

### Logistic regression model

We created logistic regression (LR) models[16] with HBV infection (defined as positive HBsAg) as the primary outcome and demographic variables (birth year, sex, race/ethnicity, and birthplace [USA- vs. foreign-born]) as the primary predictors, with both uni- and multivariable logistic regression, in the training set. We then created a logit score to estimate risk for HBsAg seropositivity that included all variables that were significantly associated with positive HBsAg at $p < 0.05$ 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 (RF) models[17] to determine demographic factors (birth year, sex, race/ethnicity, and birthplace [USA- vs. foreign-born]) that were associated with the primary outcome, HBV infection. Persons 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 downsampling of the controls (given how unbalanced the set was for HBsAg status) and 10-fold cross-validation to determine test characteristics of the model in the training set. Because we conducted downsampling, the initial model was poorly calibrated, so we calibrated the model with a Platt scaling (LR of the risk predicted by the RF model to the outcome of

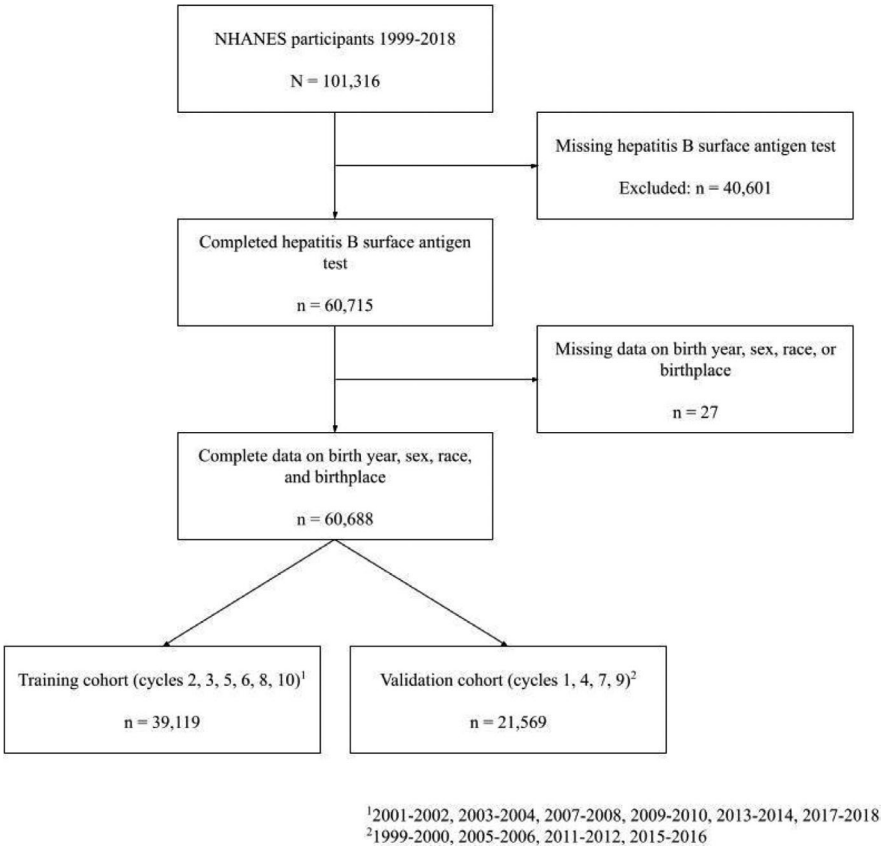


**FIGURE 1** Study design

positive HBsAg) in the training set. We then validated the model in the independent validation cohort without downsampling.[18]

## Comparison of models

We compared the RF and LR models in two ways. Initially, 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 LR vs. RF model and compared the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the top 20% versus 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 LR and RF models. CIs for sensitivity and specificity were generated using the Clopper-Pearson method,[19] whereas CIs for PPV and NPV were based on logit CIs.[20]

## Other statistical analysis

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

# RESULTS

## Study population

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; Figure 1). 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, and 10; $n$ = 39,119), from which we derived both the LR and RF models, and a validation cohort (NHANES cycles 1, 4, 7, and 9; $n$ = 21,569), in which the two models were tested.

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 (Table 1). HBsAg-positive participants were more often male (58.1%

**TABLE 1** Baseline characteristics of the overall cohort ($n$ = 60,688)

| Variable | Overall cohort ($n$ = 60,688) | HBsAg negative ($n$ = 60,418) | HBsAg 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: USA | 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 |
| 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 |
| HDL, 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 |
| LDL, 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, ×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, mm Hg ($n$ = 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, mm Hg ($n$ = 52,209) | 68.0 (58.0–76.0) | 68.0 (58.0–76.0) | 72.0 (64.0–80.0) | <0.001 |

*Note:* All data presented as either median (interquartile range) or percentage. All variables presented as variable name, units ($n$ = number of patients with complete data).

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 USA than the HBsAg-negative group (35.9% vs. 78.9%; $p$ < 0.001).

## Development of algorithms to identify HBV

Initially, we generated LR and RF models for a semi-targeted screening approach using only the above significant demographic factors, namely sex, year of birth, race/ethnicity, and birthplace.

In multivariable LR 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

**TABLE 2** LR of predictors of hepatitis b infection, training cohort ($n$ = 39,119)

| | Unadjusted | | Adjusted | |
|---|---|---|---|---|
| **Variable** | **OR (95% CI)** | ***p* value** | **OR (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: USA | 0.11 (0.08–0.16) | <0.001 | 0.14 (0.10–0.21) | <0.001 |

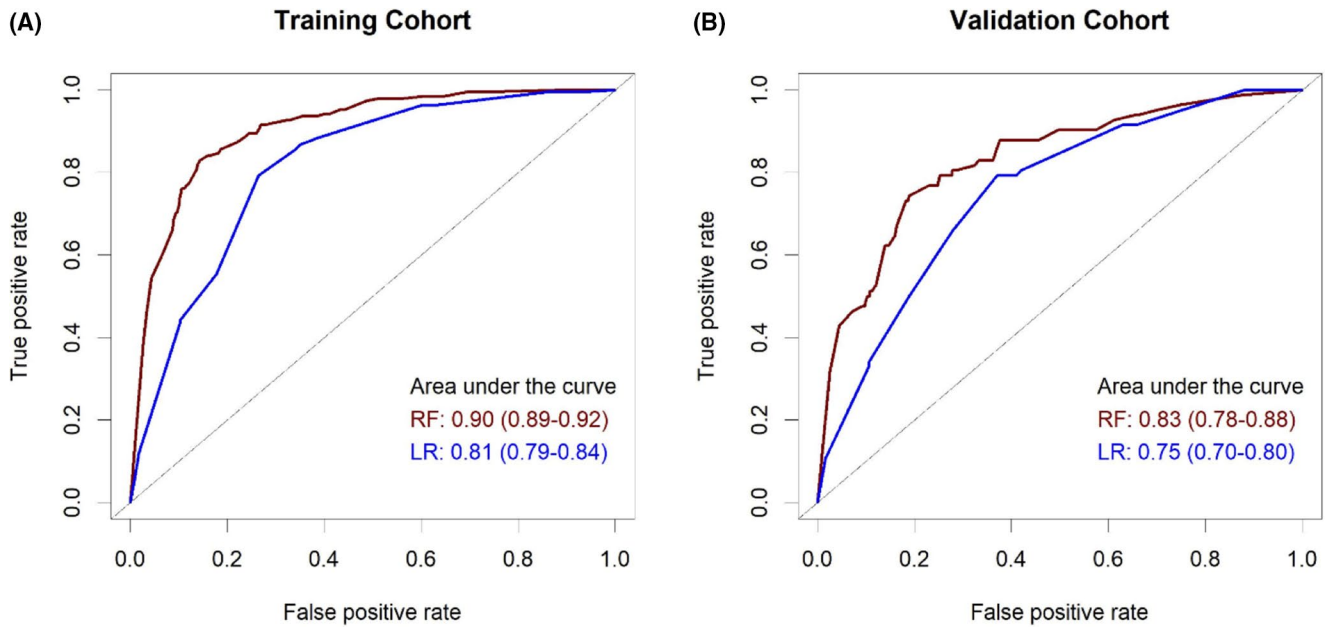


**FIGURE 2** ROC curves for LR and RF ML models in the training (A) and validation cohorts (B) [Color figure can be viewed at wileyonlinelibrary.com]

recommendation in the USA[21]) corresponded to 72% lower odds of HBV infection (adjusted OR [aOR], 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 a 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, 9.13; 95% CI, 5.23–15.96; $p < 0.001$). Meanwhile, Hispanic ethnicity was associated with 66% lower odds of HBV infection (aOR, 0.34; 95% CI, 0.16–0.71; $p = 0.0044$). Being born in the USA (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 LR-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 USA). It is not possible to display the corresponding score for an RF model in a readily interpretable closed form.

## Comparison of LR and ML models

Next, we compared the accuracy of the two models in predicting HBsAg status. Figure 2 displays the receiver operating characteristic (ROC) curves for the LR and ML models in both the training and validation cohorts. In the training cohort, the AUROC was significantly higher for the ML model, at 0.90 (95% CI, 0.88–0.92)

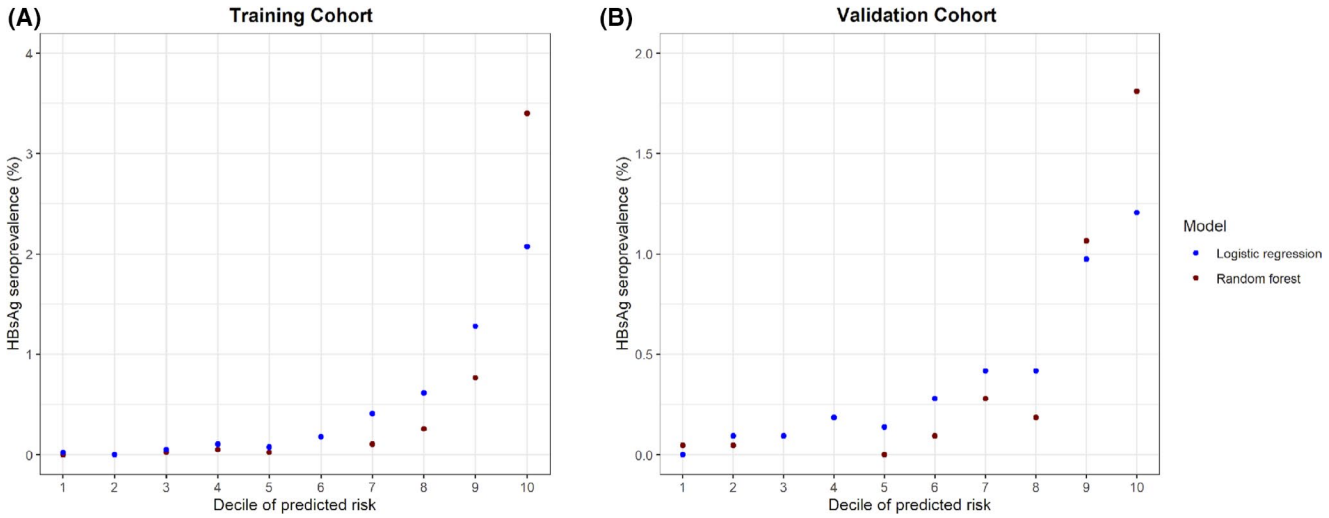


**FIGURE 3** Percentage of patients with hepatitis B viral infection within each calculated risk decile for LR and ML models in both the training cohort (A) and validation cohort (B) [Color figure can be viewed at wileyonlinelibrary.com]

**TABLE 3** Performance characteristics of the ML and LR models (positive model result: top 20% score; $n$ = 60,688)

| Test characteristic | ML | LR | $p$ 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 |

versus 0.81 (95% CI, 0.79–0.84) for the LR model ($p$ < 0.001 by the De Long test). In the validation cohort, the AUROC was also higher with the ML model as compared to the LR model (0.83; 95% CI, 0.78–0.88 vs. 0.75; 95% CI, 0.70–0.80; $p$ < 0.001). Whereas the initial ML model was poorly calibrated because of the downsampling used initially (log loss, 1.37), after applying Platt scaling the model became reasonably well calibrated (log loss, 0.02; Figure S1).

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

## DISCUSSION

This study developed an algorithm to prioritize patients for hepatitis B screening using data from a population-based cohort in the USA and relying only on demographic data that are routinely available in typical health

care delivery settings. We found that the ML model consistently outperformed the LR model, with higher AUROC values (0.83 vs. 0.75 in the validation cohort) and more effective identification of high-risk patients (1.4% vs. 1.1% seroprevalence in the top 20% in the ML and LR 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 USA 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 ESLD 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.

Although universal screening of adults for HCV has been recommended by the CDC[22] and is likely cost-effective,[23,24] universal screening for HBV in the USA is unlikely to be cost-effective given its lower prevalence (0.3%–0.5%) among the general population,[3,25] below the 2% U.S. Preventive Services Task Force threshold.[12] However, past 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 semitargeted approach to identify a higher-risk group with an HBV seroprevalence of >1% may allow for a cost-effective semiuniversal approach to HBV screening. As such, given that our algorithm only requires four simple nonstigmatizing 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. USA-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 EHRs during patient registration.

We also recognize that our algorithm can miss 13% of HBV infection, and some patients not meeting screening criteria by our semitargeted 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 semitargeted, semiuniversal screening approach such as ours to remedy the current state of CHB diagnosis and linkage to care in the USA while also applying the existing risk-based approach for those not meeting our semitargeted screening criteria.

Limitations of this study include a relatively small number of patients with HBV infection because of 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 USA 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 noninstitutionalized civilian population of the USA, and each cycle of NHANES is independent of the other cycles. Screening results of our RF ML 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. 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 USA, using only demographic data that are routinely collected by health care providers and EHR systems. Our ML model consistently outperformed the LR 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 USA or Western Europe. We also advocate for additional risk-based screening for populations with specific exposure risks as per professional society and CDC guidelines. Last, we encourage validation of our algorithm in other populations as well as future studies to evaluate the cost-effectiveness of this semitargeted and -universal HBV screening approach.

## CONFLICT OF INTEREST

Dr. Nguyen advises for and received grants from Gilead. She advises for Intercept, Novartis, Eisai, Bayer, Exact Science, Laboratory of Advanced Medicine, Janssen, and Spring Bank. She received grants from Pfizer, Enanta, Vir, and Glycotest.

## AUTHOR CONTRIBUTIONS

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

## ORCID

*Vincent L. Chen* https://orcid.org/0000-0002-0157-6066
*Mindie H. Nguyen* https://orcid.org/0000-0002-6275-4989

## REFERENCES


1. Polaris Observatory Collaborators. 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;74:607–26.
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.
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:e00046-19.
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 ASF, McMahon BJ, Brown RS, 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.; Centers for Disease Control and Prevention (CDC). 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. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Cham, Switzerland: Springer Nature Switzerland AG; 2009.
17. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
18. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D, editors. Advances in large-margin classifiers. Cambridge, MA: MIT Press; 1999. p. 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. Stat Med. 2007;26:2170–83.
21. Recommendations of the Immunization Practices Advisory Committee (ACIP). Hepatitis B virus: a comprehensive strategy for eliminating transmission in the United States through universal childhood vaccination. 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.


## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.
Supplementary Material

**How to cite this article:** Ramrakhiani NS, Chen VL, Le M, Yeo YH, Barnett SD, Waljee AK, et al. Optimizing hepatitis B virus screening in the United States using a simple demographics-based model. Hepatology. 2022;75:430–437. https://doi.org/10.1002/hep.32142