Elevate Series

ANNALS OF MEDICINE
2024, VOL. 56, NO. 1, 2314237
https://doi.org/10.1080/07853890.2024.2314237

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Development and validation of HBV surveillance models using big data and machine learning

Weinan Dong[a], Cecilia Clara Da Roza[a], Dandan Cheng[b], Dahao Zhang[b], Yuling Xiang[b], Wai Kay Seto[c,d] and William C. W. Wong[a,b]

[a]Department of Family Medicine and Primary Care, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; [b]Department of Family Medicine and Primary Care, The University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong, China; [c]Department of Medicine and State Key Laboratory of Liver Research, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China; [d]Department of Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen, Guangdong, China

ABSTRACT

Background: The construction of a robust healthcare information system is fundamental to enhancing countries' capabilities in the surveillance and control of hepatitis B virus (HBV). Making use of China's rapidly expanding primary healthcare system, this innovative approach using big data and machine learning (ML) could help towards the World Health Organization's (WHO) HBV infection elimination goals of reaching 90% diagnosis and treatment rates by 2030. We aimed to develop and validate HBV detection models using routine clinical data to improve the detection of HBV and support the development of effective interventions to mitigate the impact of this disease in China.

Methods: Relevant data records extracted from the Family Medicine Clinic of the University of Hong Kong-Shenzhen Hospital's Hospital Information System were structuralized using state-of-the-art Natural Language Processing techniques. Several ML models have been used to develop HBV risk assessment models. The performance of the ML model was then interpreted using the Shapley value (SHAP) and validated using cohort data randomly divided at a ratio of 2:1 using a five-fold cross-validation framework.

Results: The patterns of physical complaints of patients with and without HBV infection were identified by processing 158,988 clinic attendance records. After removing cases without any clinical parameters from the derivation sample (n=105,992), 27,392 cases were analysed using six modelling methods. A simplified model for HBV using patients' physical complaints and parameters was developed with good discrimination (AUC = 0.78) and calibration (goodness of fit test p-value >0.05).

Conclusions: Suspected case detection models of HBV, showing potential for clinical deployment, have been developed to improve HBV surveillance in primary care setting in China. (Word count: 264)

KEY MESSAGES

- This study has developed a suspected case detection model for HBV, which can facilitate early identification and treatment of HBV in the primary care setting in China, contributing towards the achievement of WHO's elimination goals of HBV infections.
- We utilized the state-of-art natural language processing techniques to structure the data records, leading to the development of a robust healthcare information system which enhances the surveillance and control of HBV in China.

ARTICLE HISTORY
Received 25 September 2023
Revised 8 January 2024
Accepted 30 January 2024

KEYWORDS
Big data management; big data analytics; infectious disease surveillance; machine learning; China

1. Introduction

Viral hepatitis is a significant global health concern, with high mortality rates resulting from liver cancer and cirrhosis. In 2019, chronic hepatitis B virus (CHB) infection affected 296 million people worldwide, with 0.8 million deaths primarily attributed to chronic liver disease and cancer associated with Hepatitis B Virus (HBV) infection [1]. China has the highest disease burden, accounting for more than one-third of the world's HBV infection cases. More than 80% of infected adults

CONTACT Cecilia Clara Da Roza, claracc@hku.hk, Department of Family Medicine & Primary Care, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pok Fu Lam, Hong Kong, China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07853890.2024.2314237.

in China are unaware that they are carrying the virus, and only approximately 11% are currently receiving treatment [2].

Since the healthcare reforms in 2009, China has made remarkable progress in strengthening its primary healthcare system [3]. By early 2021, a documented network of more than 35,000 community health clinics (CHCs) employing over 434,000 health professionals is established in primary care, providing fundamental health services to 706 million people through a combination of Western medicine and traditional Chinese medicine [4]. The 'Healthy China 2030 Plan' highlights the critical role and commitment of strengthening the primary healthcare system [5]. A nationally representative survey of 149 CHCs and 3,580 frontline primary care practitioners from 20 cities found that 80% of CHCs had the facility to offer HBV testing and 85% of doctors saw the benefits of providing HBV testing; only 19% had been diagnosed with HBV and 15% managed HBV in the previous month. To achieve WHO's targets on 90% diagnosis and treatment of HBV infections by 2030 [6], a clinical decision support system derived from primary care data in China would be helpful for facilitating early identification and treatment of HBV infections [7].

A systematic review and meta-analysis conducted by Liu et al. shed light on the more recent landscape. Their findings, encompassing 3,740 studies and a staggering 231 million people, revealed a declining trend in hepatitis B surface antigen (HBsAg) seroprevalence among the general population in China from 1973–1984 (9.6%) to 2021 (3.0%) [8]. This downward trend coincided with the implementation of several national policies and initiatives aimed at combating the spread of HBV, e.g. the Expanded Program on Immunization (EPI) launched in 2002 and the comprehensive vaccination program in 2009. However, the review estimated that approximately 43.3 million individuals in China still live with CHB in 2021, indicating the persisting challenges in achieving the global goal of HBV elimination by 2030. Furthermore, the review demonstrated that only 18.9% of HBV-infected cases were diagnosed and a mere 12.6% received appropriate treatment over their lifetime [8]. Meanwhile, despite the cost-effectiveness of universal HBV screening for eliminating viral hepatitis in China, as shown in a recent study conducted by Su et al. [9], individuals might still be hesitant to undergo screening because of the discrimination and stigmatization faced by those who have been diagnosed with CHB in the workplace and society [10]. It is imperative to develop a suspected case detection system that can be incorporated into routine health checks to expand current screening and linkage-to-care strategies, and promote early treatment of HBV infections.

Medical records and data are widely available in healthcare institutions at both hospitals and primary care levels because of the informatization requirements in medical systems [11]. However, data utilization remains low owing to the pre-processing requirement of the raw data, with this low-quality limiting data utilization [11]. By virtue of the unique algorithm theory, machine learning (ML) techniques have been more frequently used to discover values from medical data and to build advanced models to assist clinical practice [12]. Bordering Hong Kong in the southern region of China, Shenzhen is a fast-growing municipality that attracts many migrants across other regions of China. Some migrants may have a higher burden of hepatitis [13–16]. As a result, Shenzhen is an important location for implementing public health interventions for hepatitis. According to previous studies, Shenzhen had a high HBV prevalence of 9.7% from 2015 to 2018 and an increase in Hepatitis C Virus (HCV) infections [15]. In this study, we worked primarily with the Department of Family Medicine and Primary Care (FMPC) at the University of Hong Kong-Shenzhen Hospital (HKU-SZH). Currently, approximately 100,000 patients visit the FMPC annually from the Shenzhen municipality, Guangdong, and nearby provinces [16].

Specialised test for HBV including enzyme immunoassay and nucleic test have not been widely applied to the large population for screening because of the cost-effectiveness. Hence, some studies have been conducted to develop more practical digital tools for early identification of HBV patients. Wang et al. developed a model using demographics, routine blood indicator and liver function to identify high-risk population for further HBV test in China [17]. Ajuwon et al. also developed a diagnostic model of HBV infection using blood test parameters based on the routine clinical data in Nigeria [18]. Both of these two studies included liver function parameters as the most powerful predictors. However, liver function parameters are not routinely tested in clinical practice, limiting the practicality of using models based solely on clinical data. Some researchers attempted to use multimodal data to predict or detect diseases, such as images [19] and wearable devices-collected data [20]. It has been proved that multimodal data can provide predictive information in different dimensions hence can yield better predictive performance [21]. In the current study, we sought to additionally extract information from patients' self-reported complaints in text using advanced natural language processing (NLP)

techniques to enhance the predictive performance and improve the practicability of the detection model.

This study aimed to develop and validate HBV case detection models in China using big data from patients' electronic health records (eHR) in family medicine clinics at the HKU-SZH and ML techniques. By leveraging advanced analytical methods to analyse large-scale health data, this study sought to improve the detection of CHB cases and support the development of effective interventions to mitigate the impact of this disease in China.

2. Methods

2.1. Data overview

The study data were extracted from the HKU-SZH Hospital Information System (HIS), which contains all information generated from the clinical practice of Department of Family Medicine in HKU-SZH from June 1, 2018 to May 31, 2021. It was possible for patients to have multiple visits during the study period. With regard to missing data, only laboratory parameters were imputed and not radiological data. It is important to note that missing data can be imputed when the model is applied in real-world practice, so it will not influence the application of the model in the clinical setting. The following procedures were administered to develop and validate an HBV surveillance system based on data from HKU-SZH: (1) data cleaning of the data extracted from HKU-SZH; (2) structuralizing the eHR data in text into quantitative data; (3) completing the data pre-processing, including outlier detection and deletion, missing data imputation, data match, and the building of experimental datasets for further analysis; (4) developing risk assessment models of HBV infection using multiple ML modelling algorithms; (5) validating the risk assessment models using standard measures in order to ensure its accuracy; (6) attempting novel interpretation algorithms to interpret the ML models, so as to ensure its reliability and discover any novel clinical patterns; and (7) deploying the models onto software to strengthen its usability in real-world practice.

A total of four types of data, which were recorded in Simplified Chinese, were extracted, including out-patient records in text, data on laboratory tests, data on examination, and data on prescription and payment items, which are referred to out-patient records and laboratory, imaging, and prescription data in this report. A total of 1,357,177 items with more than 150 clinical parameters from laboratory tests data (e.g. blood test and urinalysis) were identified. The model building did not involve any results directly related to HBV, such as HBsAg, hepatitis B surface antibody (anti-HBs), hepatitis B e antigen (HBeAg), hepatitis B e antibody (anti-HBe), hepatitis B core antibody (anti-HBc), and HBV deoxyribonucleic acid (DNA). 1,104 items of pathological examination, 34,250 items of ultrasound, 15,664 items of radiography, and 398,170 items of electrocardiography reports were identified from examination data. A total of 731,203 items and 232,366 items were identified from the prescription data and records, respectively.

The current study was a retrospective analysis of previously captured data, and all data involved were anonymized. This study was approved by the Research Ethics Board of HKU-SZH (reference number: hkuszh2020145).

2.2. Data cleaning, structuralizing and pre-processing

For laboratory data, the values of the same clinical parameters with different units or measured by different devices were adjusted using the latest clinical guidelines. Imaging data that were not relevant to the HBV diagnosis were excluded. The prescription data on payment items, date, item types, disease coding, diagnosis and department of services were sorted. The records included textual data on patients' complaints, past history, allergies, comorbidities, family history, vital signs (height, weight, temperature, and blood pressure), diagnosis, and treatment morbidities.

The recorded data were then structured using state-of-the-art NLP techniques [22]. Specifically, text segmentation was firstly applied to the records to divide the text into meaningful parts, variable names, physical complaints, adjunct words, negative words, clinical parameters, values, and units. Text segmentation was conducted using bidirectional encoder representations from transformers (BERT). BERT is a pre-trained deep-learning model developed by Google in 2018, with the transformer encoder in the algorithm capable of learning the context of a word based on its bidirectional surroundings [23]. Conditional random fields (CRF), a statistical modelling method based on Bayesian theory, was used to perform part-of-speech (POS) tagging [24]. Based on segmentation results, features of the physical complaints were extracted using named-entity recognition (NER) techniques, together with relationship extraction and pre-defined rules [25]. An example of record structuralization is shown in Figure 1. Subsequently, 200 records were randomly selected for manual checking to ensure consistency between manually extracted physical complaints and ML extracted physical complaints.

The missing values for physical complaints were imputed as zero (absence of physical complaints), and

the missing values for the laboratory data were imputed using generative adversarial imputation nets (GAIN) [26,27]. GAIN consists of two deep neural networks, the generator and a discriminator, to impute missing data and identify the imputed data and return feedback, respectively. In such an antagonistic process, generator and discriminator are trained alternately until the discriminator cannot identify the imputed data (end of the training). The algorithm with the implementation of GAIN can be found in the Supplemental Materials [26]. The data from the same patient were matched using a unique desensitized number. After removing the records without any physical complaints or any valid extracted information, a total of 158,988 records remained. Of these, 27,392 records were matched for valid clinical parameters. We reviewed the records that mismatched the clinical parameters, and found that all of them were for health-related consultation services without the need to conduct laboratory tests or examination.

2.3. Cohort of HBV patients

The diagnosis of CHB was defined using prescription data according to ICPC code D72 [28] and doctors' written diagnosis, and laboratory data according to HBsAg in the HBV 5-item test (consisting of HBsAg, antibody to HBsAg, HBeAg, antibody to HBeAg, and antibody to hepatitis B core antigen) and a serum HBV DNA test [29]. A total of 1,458 unique HBV patients and their corresponding 2,879 clinic attendance records were identified. The physical complaints that could be associated with active CHB reported by patients in the family medicine clinics of the HKU-SZH, are listed in Supplementary Tables 1–3, respectively. As listed in Supplementary Table 4, clinical parameters with observed values of more than 1/3 of the 27,392 complete records were considered in the study. Correlations between pairs of physical complaints were analyzed using Spearman statistics and visualized using a heatmap. Hierarchical clustering was applied to illustrate the patterns of physical complaints in the different disease groups.

2.4. Model development and validation

The cohort data were randomly divided into derivation and test samples in a ratio of 2:1 to develop and validate the HBV detection model. A five-fold cross-validation framework was adopted. Different ML modelling methods were used to develop the model in parallel, including the logistic regression model, Naïve Bayes model, neural network (NN), support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGB).

Figure 1. Examples of structuralization of health record in text: (A) shows the text segmentation of the text records of a real case, while (B) shows the physical complaints extracted from this record data.

The Naïve Bayes model is a classification technique based on Bayes' theorem, with the assumption of independence among predictors [12,24]. The regression and Naïve Bayes models might over-simplify a complex real-world task, but such a strategy usually results in high generalizability and good performance [30]. An NN is the basis of modern artificial intelligence, which encompasses multiple (or single) layers of neurons connected with activation functions [31]. In this study, an NN with a feed-forward solving method was adopted. SVM is one of the most robust classification methods that uses nonlinear kernels [32]. RF is an ensemble learning algorithm based on voting of multiple decision trees [33]. Technically, XGB is also an ensemble learning method in which each sub-model is trained to correct the errors made by the previous model [34]. XGB has been reported to outperform other ML algorithms in real-world tasks [35].

Model development was carried out in three stages: (1) only physical complaints were considered as possible predictors to build the base model; (2) all clinical parameters except those used to determine HBV diagnosis were additionally added to the base model to build model 1; and (3) clinical parameters that were only available in hospitals (not available in general out-patient clinics in community healthcare service institutions) were removed to build model 2 with better generalizability. The loss function of the study task was set as the area under the curve (AUC). The hyper-parameters of the ML models were determined using a grid search [32]. Early stopping to halt the training process was adopted to prevent over-fitting [36]. The developed models were applied to the validation sample to assess the performance of the models, including discrimination measured using AUC and calibration measured using the Hosmer-Lemeshow (H-L) test of goodness of fit. The sensitivity and specificity at different cutoff values are also listed. Sensitivity is a pivotal indicator in HBV monitoring, which reveals the proportion of actual HBV cases being correctly identified [37], and specificity reflects the cost-effectiveness of a model [36]. Therefore, the best cut-off of the models was determined by considering the balance between sensitivity and specificity in a clinical scenario [38].

The Shapley value (SHAP) was used to interpret the model with the best performance to avoid distrust caused by the low interpretability of the machine-learning model. Because ML models are commonly intrinsically uninterpretable, SHAP (a model-agnostic method) operates to analyze input and output pairs to explain an ML model without access to any model internals. Unlike other model-agnostic methods (e.g. local explanation), SHAP has a solid theoretical basis [39]. The nonlinear effects of the physical complaints and clinical parameters on the outcome (HBV) were quantified, and significant interactions between the covariates were also visualized. These quantifications can facilitate the recognition of novel effects or patterns of clinical significance.

The ML models with the best performance were deployed into the suspected case detection software for HBV, which has the potential to complement the HKU-SZH HIS. All significance tests were two-tailed, with a significance level of $p<0.05$. Data analyses were performed using R 3.5.1, STATA 13, and Python 3.5. Our research team has obtained a copyright license of STATA 13.

3. Results

3.1. Pattern of physical complaints

The frequencies of physical complaints reported by patients and non-CHB people in the entire dataset are listed in Table 1. Cough was the most common complaint with a frequency of 9.52%, followed by sleep disorders (7.72%), and throat pain (5.04%). The average number of relevant physical complaints for each patient was 0.83, indicating an extremely sparse feature space. The correlations between each pair of physical complaints are shown in Figure 2. Most complaints were related to upper respiratory illnesses, but the internal correlations were low. The patterns of physical complaints demonstrated by hierarchical clustering are shown in Figure 3. Considering the low number of reported physical complaints in individual patients, this pattern may not be informative for suspected case detection, yet could be indicative of HBV surveillance at the population level.

3.2. Suspected case detection model of HBV using predictors of physical complaints

We solely used the physical complaints as predictors to develop the model since these data were available for the majority of patients attending primary care clinics. Using the derivation sample ($n=105,992$) with only the physical complaints extracted from the clinical records, different modelling methods were used to build the model. As all the predictors were binary variables, the Bayesian Network (BN) was proposed as the most appropriate method. BN training was performed in two steps: structure learning and parameter learning [40]. We attempted a full BN with scored-based, constraint-based, and backward sequential elimination and joined the algorithm to determine the structures of the BN, as shown in Supplementary Figure 1. All candidate modelling methods described in the Methods section were attempted. All the models

Table 1. List of the frequencies of the present symptoms ($N = 158,988$).

Symptoms	Frequency	%
Abdominal pain	7146	4.49%
Head pain	6060	3.81%
Throat pain	8007	5.04%
Chest pain	4045	2.54%
Waist pain	3567	2.24%
Neck pain	1804	1.13%
Joint pain	3121	1.96%
Chest distress	4526	2.85%
Flatulence	2665	1.68%
Nausea	6282	3.95%
Burping	2197	1.38%
Diarrhea	1875	1.18%
Loss of appetite	1444	0.91%
Abdominal mass	195	0.12%
Ulcer	92	0.06%
Urine yellow	132	0.08%
Gas dysfunction	124	0.08%
Nose running	6463	4.07%
Cough	15128	9.52%
Fever	6719	4.23%
Nasal congestion	5615	3.53%
Dizziness	6885	4.33%
Palpitation	3407	2.14%
Fatigue	5945	3.74%
Weight change	3156	1.99%
Sleep disorders	12275	7.72%
Rash	3664	2.30%
Itchiness	7929	4.99%
Urine problem	1373	0.86%
Urine color	220	0.14%

showed AUC less than 0.63 (0.58-0.63), which was not acceptable for a clinical model.

The distribution of the predicted risk yielded by the best model (NB) is shown in Supplementary Figure 2. The predicted risks of HBV-positive and HBV-negative cases showed slightly different risk distributions. Within the best model, if 90% of HBV-positive cases need to be successfully detected by the risk model (sensitivity = 90%), 75.2% to 78·9% of the whole population will be predicted as positive, and in which, 97.7% will be false positive. According to the above analysis and theory of information [41], we can infer that the physical complaints presented in primary care clinics are not sufficiently informative for the detection of HBV.

3.3. Case detection model of HBV using predictors of physical complaints and clinical parameters

Subsequently, we considered the clinical parameters as additional predictors to see whether extra predictive information can be provided. Using the same derivation sample, we added the laboratory tests indexes (Supplementary Table 4) to build a detection model for HBV. After removing cases without any clinical parameters, 27,392 cases (18,261 in the derivation sample and 9,131 in the test sample) remained and were used to develop and validate the model. Six modelling methods were used to develop the model in parallel; logistic regression, NB, NN, SVM, RF, and XGB. The receiver operating characteristic (ROC) curves of the different models for the test sample are shown in Figure 4. The XGB models showed the best performance with an AUC of 0.823, which was significantly higher ($p = 0.023$ by DeLong's test) than those of the RF (AUC = 0.801), logistic regression (AUC = 0.7683), NB (AUC = 0.709), NN (AUC = 0.726), and SVM (AUC = 0.731) models. The XGB model also showed good calibration, with an H-L test p-value of more than 0.05, indicating good goodness of fit.

The hyperparameters of the best XGB model were: max depth = 7, subsample = 0.6, minimum child weight = 0.6, column sample by level = 0.6, column sample by tree = 0.6, gamma = 3, alpha = 5, iteration number = 122, and learning rate = 0.1. The importance of the predictors evaluated using the SHAP method is illustrated in Figure 5. A summary of the predictors is presented in Supplementary Figure 3, where X-axis represents the SHAP value, indicating the impact of different predictors on the model output. Platelet count was the most important predictor, followed by alkaline phosphatase (ALP), aspartate aminotransferase (AST), mean corpuscular hemoglobin concentration (MCHC), albumin, low-density lipoprotein, alanine transaminase, percentage of neutrophils, lymphocyte percentage, and total cholesterol. The effects of the top ten important predictors are shown in Figure 6, illustrating the quantisation and visualisation of nonlinear effects of the predictors and their interactions. The normal range of platelet is 150-450 * 10^9/L [42], and we observed that a decrease in platelet count to <200 * 10^9/L without fever could be an important predictor of HBV infection. ALP showed a U-shaped effect on the risk of HBV, where ALP levels lower than 50 IU/L together with low high-density lipoprotein cholesterol (HDL-C), or more than 100 IU/L are associated with a higher risk of HBV infection. The linear effect of AST discovered by ML was in line with clinical experience, with a value >40 IU/L indicating impaired liver function.

The sensitivity and specificity of the model at different cutoff values are listed in Table 2. If the model achieved a sensitivity of 80%, the specificity was 70%, which means that approximately 30% of the negative cases were predicted to be false positives. Overall, this full model showed outstanding performance in detecting suspected CHB cases in a secondary care setting, where most laboratory tests are accessible.

3.4. Simplified case detection model of HBV

Among the ten important predictors of the previous model, three were indicators of liver function. Changes

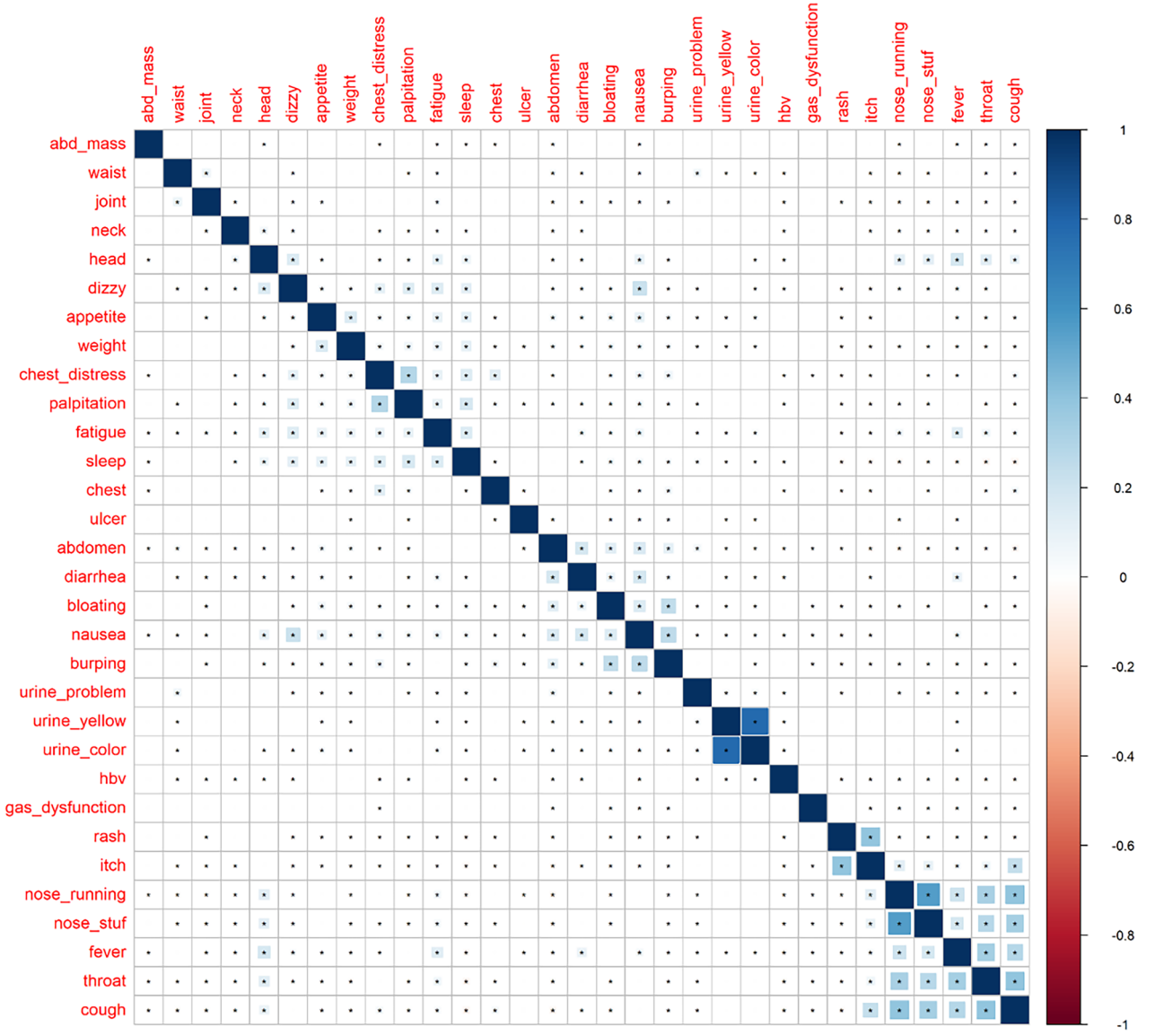

Figure 2. Correlations between each pair of physical complaints in the study sample ($N = 158,988$). (Cells with dot indicate statistically significant correlation. Blue and red colour represent positive and negative correlation, respectively. The correlations were measured using Spearman's rank correlation co-efficient, and the statistical significance was determined using Z-test.)

Table 2. The sensitivity and specificity of the model at different cutoff values.

Sensitivity	Specificity	Cutoff
90.20%	47.51%	0.0120
85.30%	58.99%	0.0170
80.12%	70.28%	0.0247
70.03%	79.25%	0.0347
66.28%	81.81%	0.0383
60.23%	84.92%	0.0445
50.14%	89.98%	0.0586

Note: The values of sensitivities and specificities were selected by sensitivity level of 90%, 85%, 80%, 70%, 60% and 50%, and cut-off of HBV prevalence in the cohort data.

in liver function are not usually apparent in the early stages of HBV infection, and these indicators are not widely available in primary care settings, particularly in community-based healthcare institutions. In this section, only common blood test parameters and physical complaints were considered to improve the usability of the model in primary care clinics.

Using the same dataset ($n = 27,392$ [18,261 in the derivation sample and 9,131 in the test sample]), models were developed using the same modelling methods. As shown in Figure 7, the XGB model showed outstanding performance at an AUC of 0.780, which was significantly better than the other models, hence was selected as the final simplified case-detection model. This model showed good calibration with an H-L test p-value of more than 0.05, indicating no significant difference between the predicted and observed risks. The hyperparameters of the XGB model were

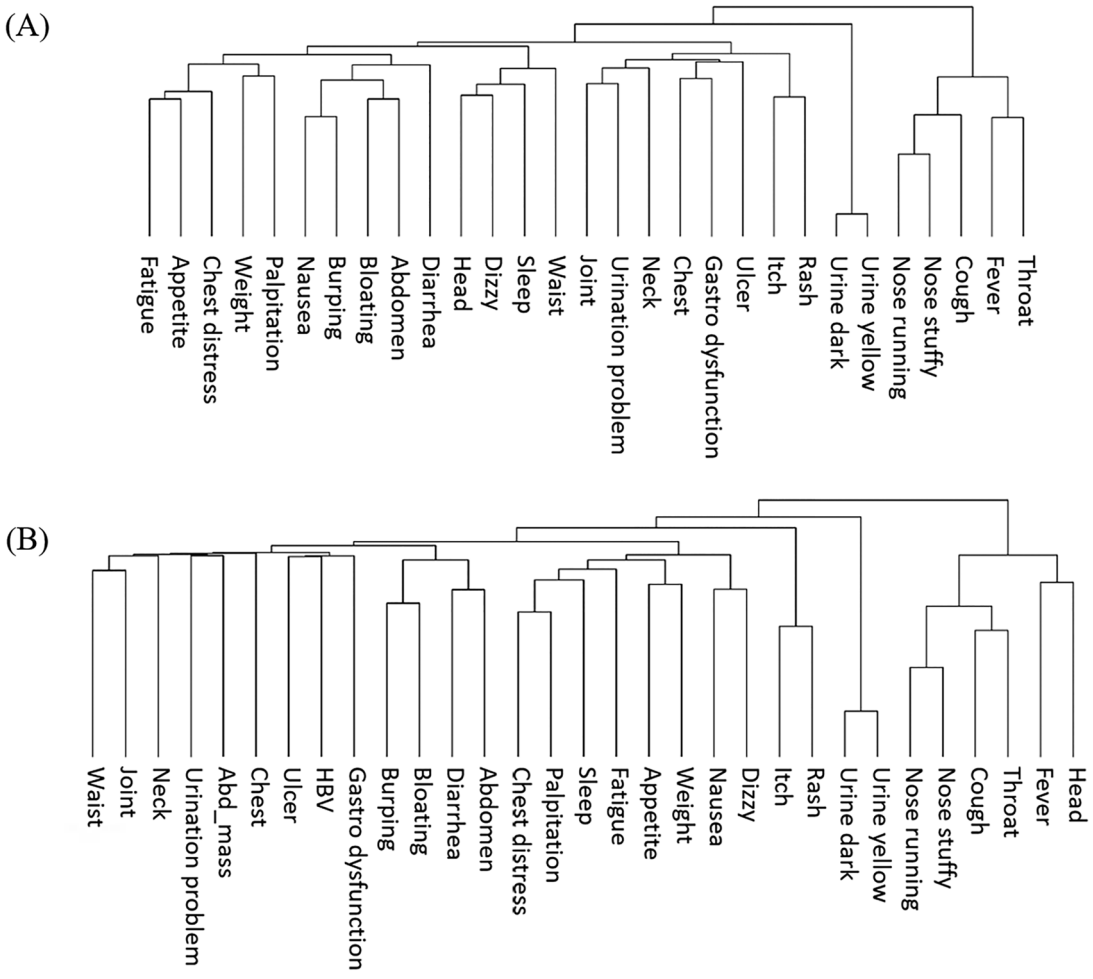

Figure 3. Patterns of physical complaints: (A) is the pattern of CHB patients; and (B) is the pattern of non-CHB people (symptoms in the same cluster usually present at the same time, and vice versa).

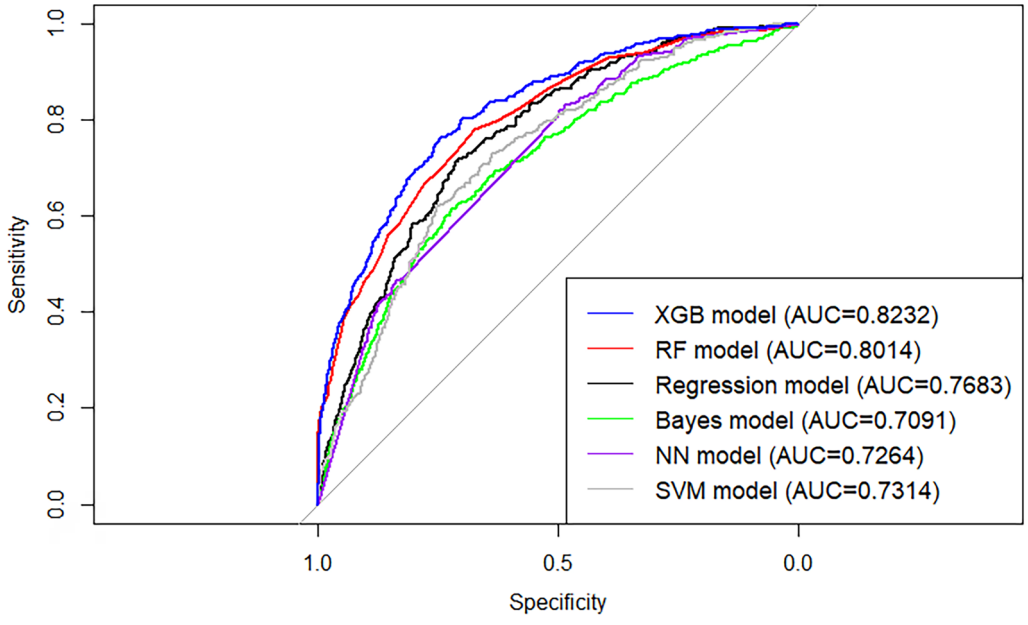

Figure 4. ROC curves of different HBV detection models on test sample ($n = 9{,}131$).

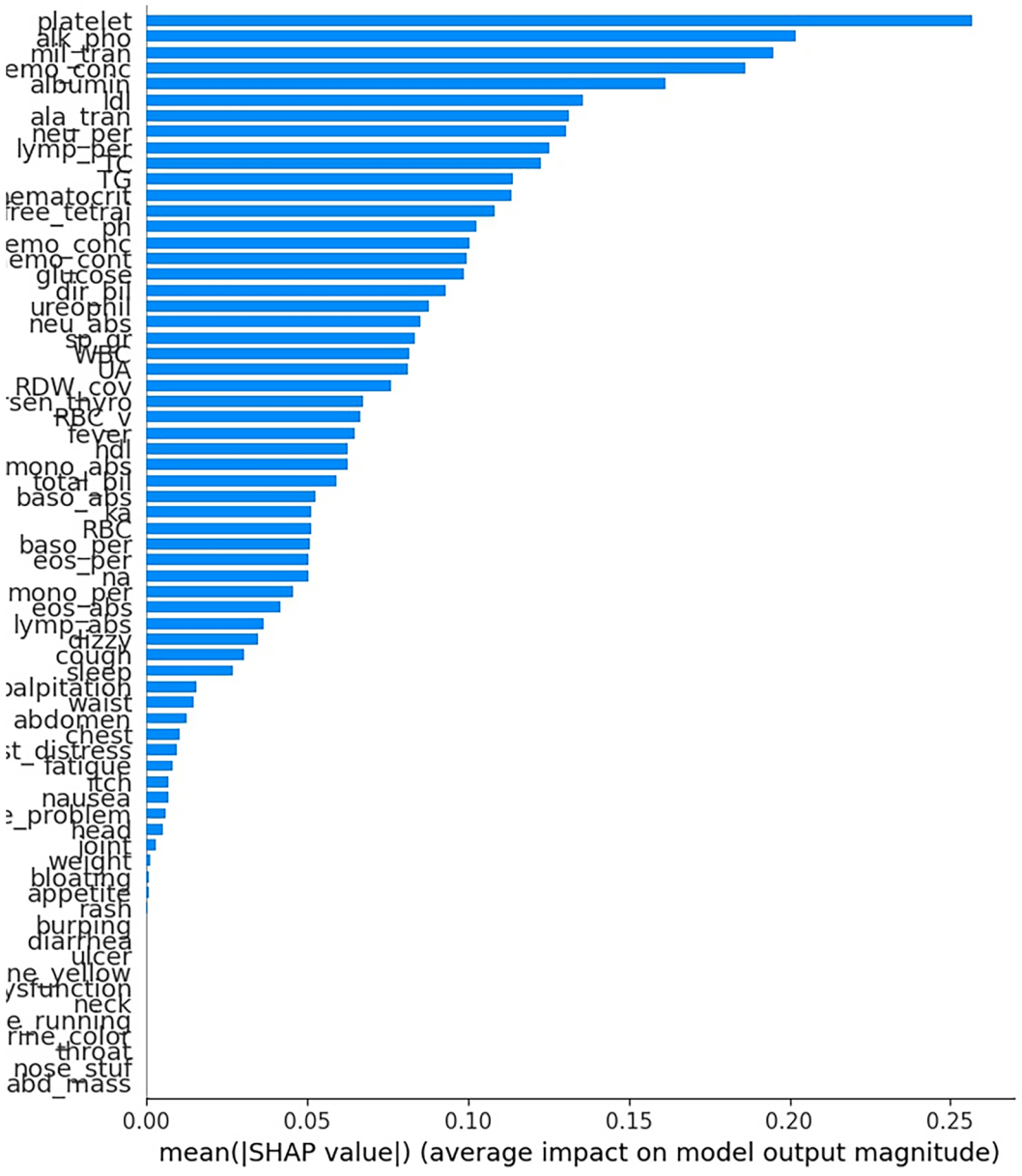

Figure 5. Importance of the predictors in XGB model based on physical complaints and clinical parameters (X-axis is the SHAP value of different predictors. Predictors with higher SHAP value can provide more predictive power and are more important).

max depth = 7, subsample = 0.7, minimum child weight = 0.7, column sample by level = 0.6, column sample by tree = 0.8, gamma = 3, alpha = 0, iteration number = 99, and learning rate = 0.1. This XGB model was not as good as that developed using all clinical parameters, but its discrimination power was acceptable for clinical use. The importance ranking of the predictors in the model evaluated using the SHAP method is shown in Figure 8. The platelet count was the most important predictor, followed by the percentage of lymphocytes, hemoglobin concentration, white cell count, mean corpuscular hemoglobin (MCH), MCHC, and hematocrit. The sensitivities and specificities at different cutoff values are listed in Table 2.

Although the cutoff is commonly determined using the Youden Index to determine the highest sum of sensitivity and specificity, it is more appropriate to determine the cutoff according to the cost of false positives and false negatives, as they are generally unequal in practice [43]. Using the prevalence of CHB in the cohort (3.8%) as the cutoff, the confusion matrix of the detection model is presented in Figure 9. More than 70% of CHB cases could be successfully detected, while 30% of HBV-negative cases would be false negatives. This simplified model has the potential to be used in primary care settings to detect suspected CHB cases with good discrimination (AUC = 0.78) and calibration (goodness-of-fit test p-value >0.05).

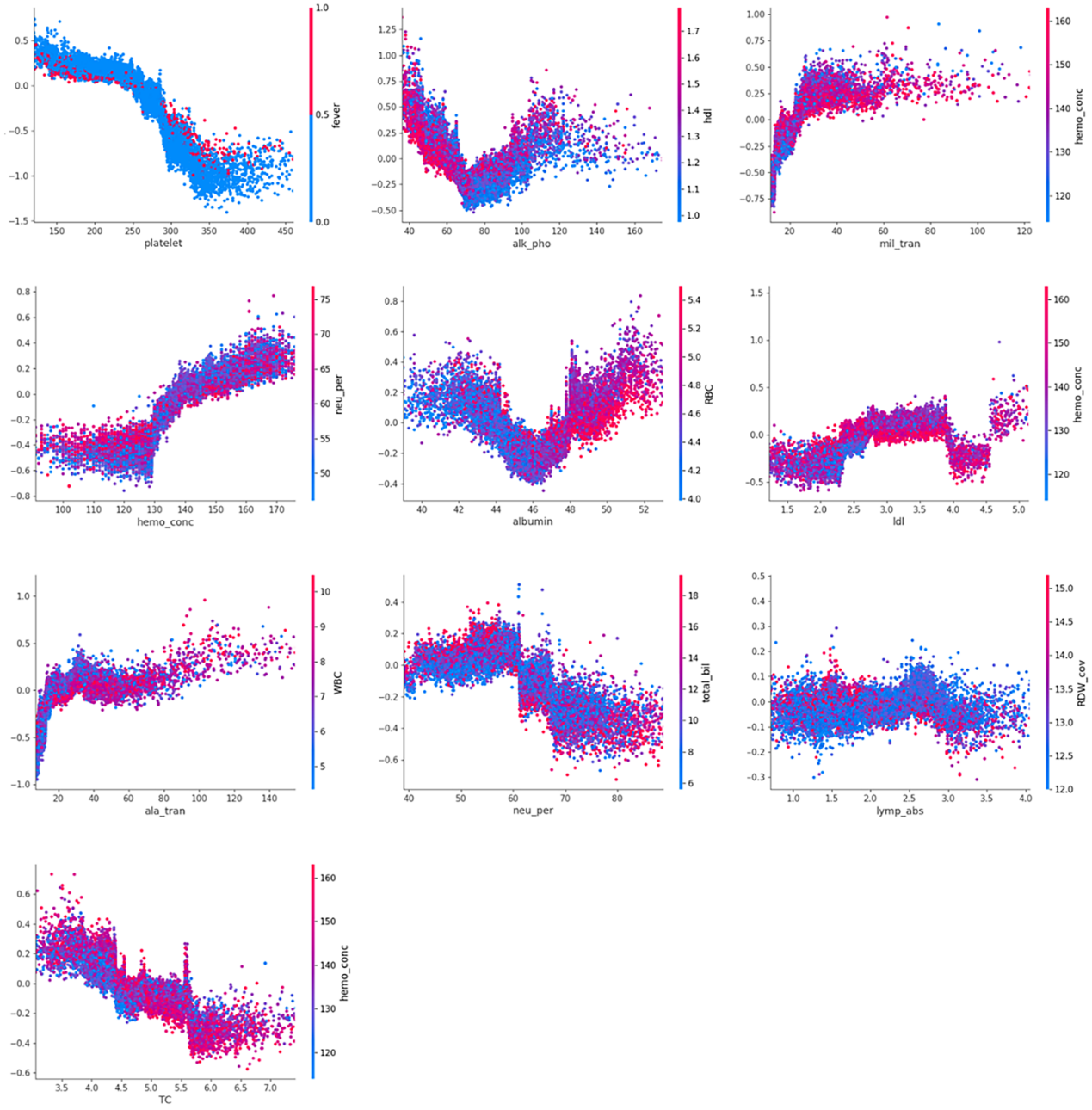

Figure 6. Nonlinear effects of the top ten important predictors in the XGB model based on symptoms and clinical parameters.

4. Discussion

In the current study, suspected HBV case detection models were developed to support HBV surveillance in primary care settings. Using a state-of-the-art NLP algorithm, patient-reported physical complaints can be accurately identified from clinical records in the text. Using the extracted physical complaints and clinical parameters, the models were capable of detecting HBV cases with an accuracy of approximately 80%.

HBV infection poses substantial public health challenges in less developed countries across the globe. Early screening and detection of HBV infection can enable early treatment and mitigate the risk of severe complications such as liver cancer [29,44]. A recent study conducted by Su et al. [9] showed that population-wide universal screening for HBV infection was cost-effective and could potentially save 3.5 million lives. Nonetheless, it is crucial to consider the restrictions on real-world applications. The current hepatitis control strategy in China places significant reliance on secondary care facilities for HBV detection on an *ad hoc* basis [45]. Previous studies have also identified that the dominant barriers to HBV detection in China are patients who report no or

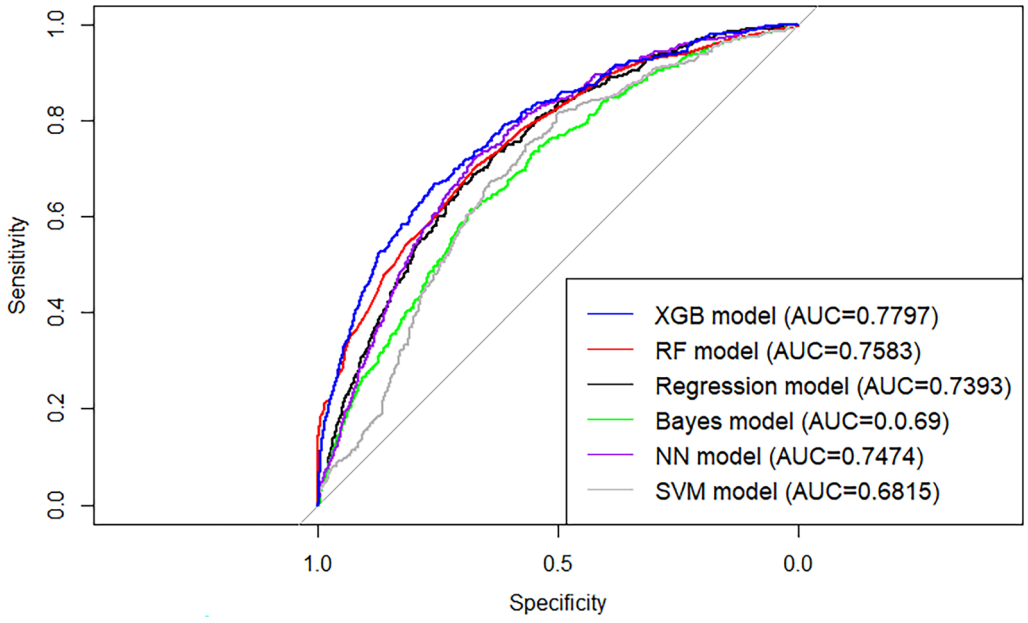

Figure 7. ROC curves of different HBV detection models based on symptoms and common clinical parameters on test sample (n = 9,131).

minimal physical complaints, as well as inadequate training and support for primary care practitioners [9,46]. The current study maximized the use of clinical information from text records and quantitative laboratory test parameters by utilizing the latest ML technologies to achieve precise detection of suspected cases across various clinical scenarios. Despite the implementation of the eHR system in public healthcare institutions in China, there is still a considerable gap in data usage. BERT, a pre-trained deep-learning model [47], along with the widely used generative pre-trained transformers (GPT) [48] developed by Google and OpenAI, respectively, represent two mainstream methods for constructing large language models. The utilization of pre-trained models in this study facilitated the text segmentation of clinical writing records and accurate entity identification of relevant physical complaints and characteristics. This, in turn, led to improved extraction of vital information from complex clinical data in the text and the development of exceptional models.

Existing models in the field of HBV-related research mainly focus on the prediction of long-term risks associated with HBV patients, such as cirrhosis and hepatocellular carcinoma, with the aim of supporting treatment decision-making [49,50]. Currently, there is still a gap in the development of a case detection model. In the present study, a comprehensive model incorporating both physical complaint features and all relevant clinical parameters was built, which could aid clinicians in identifying suspected HBV cases through a review of the historical data. The model developed in this study can help prevent missing diagnoses of HBV, even if a patient visits the clinic for a seemingly unrelated reason such as cough. Moreover, the simplified model is of great value for assisting clinicians with less experience in CHB diagnosis in primary care settings. According to evidence from 2016, over 80% of patients with HBV infection remained undiagnosed in China [2]. The models developed in the current study showed a sensitivity of more than 80% with a false positive rate of 30%, and a sensitivity of more than 90% with a false positive rate of 50%, which might have the potential to be implemented in real-world clinical practice, facilitating early identification of HBV cases. This study discovered that the combination of low ALP and low HDL-C levels is linked to an increased risk of HBV infection, which might be due to malnutrition. However, further investigation is required to examine whether this approach is feasible in primary care settings, and whether this would make a difference to clinical care.

The models were integrated into the workflow of HKU-SZH to aid clinical decision-making. We will report the effectiveness of these models and feedback from clinicians in the near future. While the current study provides valuable insights, it has several limitations. First, the models were developed and validated using data from only one hospital, which might restrict their generalization. Second, some patients had HBV infection, but their status may not have been recorded in the hospital dataset. Third, the absence of certain clinical parameters may reduce the effectiveness of the model in clinical settings.

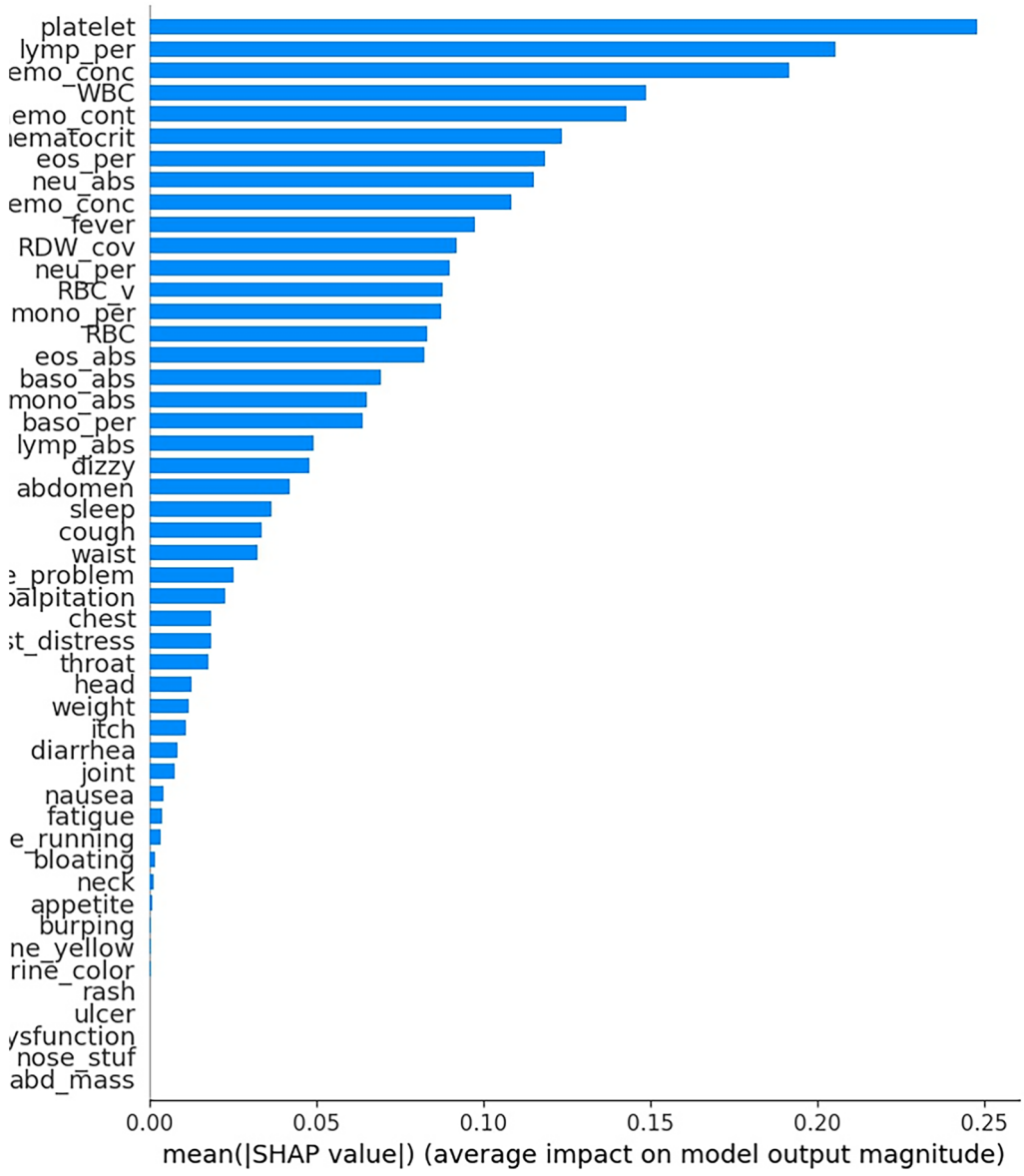

Figure 8. Importance ranking of the predictors in XGB model based on symptoms and common clinical parameters.

We made an effort to identify cases with HBV infection using various data sources, including HBV tests, diagnosis codes, drug usage, and self-reported records. However, due to concealing, there is a possibility that some cases may have been missed. To address the issue of missing values, we have employed advanced AI methods for imputation. While the accuracy of prediction based on imputed values remains unknown, we have focused on providing the predictive power primarily through the top predictors identified in our study. There remains difficulties in optimizing our models to accommodate all primary care settings in China. Further cohorts and studies from other institutions are essential for validating the performance of the detection models in different patient populations and refining our models before they can be widely applied in the country. To be specific, future efforts should focus on recalibrating the detection model to ensure its applicability and accuracy in diverse settings. The characteristics of patient populations can vary across different regions, and it is crucial to validate and adapt the model accordingly. Furthermore, gathering feedback from clinicians can provide insights into the practical implementation of the model and enhance its integration into the clinical workflow.

5. Conclusions

In conclusion, the current study developed suspected HBV case detection models to support HBV surveillance in China. The models were developed based on clinical records and laboratory data, and thus have the potential to be incorporated into clinical workflows for

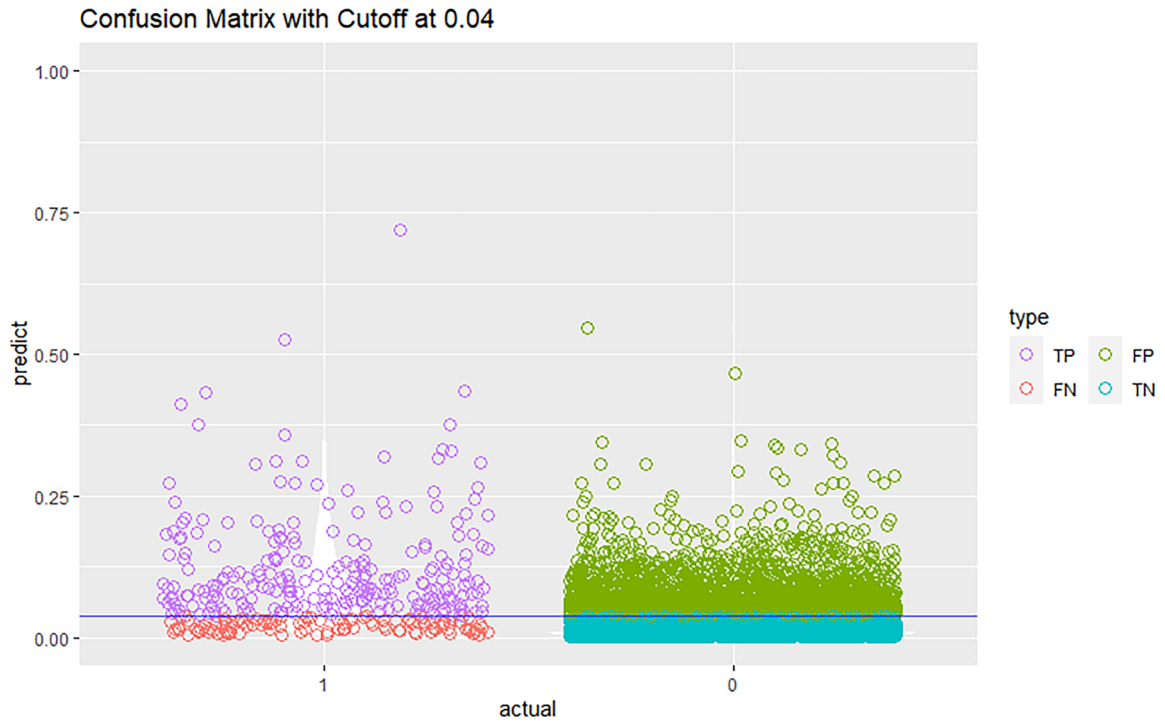

Figure 9. Confusion matrix of the risk XGB model based on physical complaints and clinical parameters at cutoff of HBV prevalence in cohort. (TP = true positive; FP = false positive; FN = false negative; TN = true negative).

decision-making assistance. The developed models demonstrated outstanding accuracy, reaching approximately 80%, which is capable of detecting suspected HBV patients in primary care settings. The utilization of advanced ML technologies allows for the full exploitation of clinical data, delivering genuine value to clinical practice.

Acknowledgements

We would like to express our sincere gratitude to the University of Hong Kong-Shenzhen Hospital for their invaluable support. This enabled us to develop and validate innovative HBV surveillance models using big data and machine learning techniques. Without this support, this study would not have been possible. We extend our heartfelt thanks to the hospital and all those involved in making this collaboration successful.

Author contributions

WD was the main author who designed and implemented the methods, analyzed the data, and developed the HBV surveillance model. WCWW provided supervision and advice throughout the study. CCDR prepared the first draft of the manuscript. DC, DZ, and YX reviewed and modified the manuscript. WKS reviewed and verified the underlying data, and provided professional comments to the modification of the manuscript. All authors had full access to all the data in the study, read and approved the final version of the manuscript, and had final responsibility for the decision to submit for publication.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Weinan Dong http://orcid.org/0000-0002-3541-1649
Cecilia Clara Da Roza http://orcid.org/0009-0005-8900-8790
Dandan Cheng http://orcid.org/0009-0001-4247-9545
Dahao Zhang http://orcid.org/0009-0000-5572-8902
Yuling Xiang http://orcid.org/0009-0001-9462-0723
Wai Kay Seto http://orcid.org/0000-0002-9012-313X
William C. W. Wong http://orcid.org/0000-0003-2540-4055

Data availability statement

Data of this study are available from the corresponding author upon reasonable request.

References

[1] World Health Organization. Global progress report on HIV, viral hepatitis and sexually transmitted infections., 2021. Geneva: WHO; 2021.
[2] World Health Organization. Up to 10 million people in China could die from chronic hepatitis by 2030 – urgent

action needed to bring an end to the 'silent. epidemic'. Beijing: WHO; 2016.
[3] Feng J, Gong Y, Li H, et al. Development trend of primary healthcare after health reform in China: a longitudinal observational study. BMJ Open. 2022;12(6):1. doi: 10.1136/bmjopen-2021-052239.
[4] National Health Commission of the People's Republic of China. 2020. Statistical bulletin of China's health development. National Health Commission of the People's Republic of China; 2021.
[5] Xinhua News Agency. "Healthy China 2030" plan outline. Beijing (China): Xinhua News Agency; 2016.
[6] World Health Organization. Combating hepatitis B and C to reach elimination by 2030. Geneva: WHO; 2016.
[7] Wong WCW, Lo YR, Jiang S, et al. Improving the hepatitis Cascade: assessing hepatitis testing and its management in primary health care in China. Fam Pract. 2018;35(6):731–16. doi: 10.1093/fampra/cmy032.
[8] Liu Z, Lin C, Mao X, et al. Changing prevalence of chronic hepatitis B virus infection in China between 1973 and 2021: a systematic literature review and meta-analysis of 3740 studies and 231 million people. Gut. 2023;72(12): 2354–2363. doi: 10.1136/gutjnl-2023-330691.
[9] Su S, Wong WC, Zou Z, et al. Cost-effectiveness of universal screening for chronic hepatitis B virus infection in China: an economic evaluation. Lancet Glob Health. 2022; 10(2):e278–e287. doi: 10.1016/S2214-109X(21)00517-9.
[10] Jin D, Treloar C, Brener L. Hepatitis B virus related stigma among Chinese living in mainland China: a scoping review. Psychol Health Med. 2022;27(8):1760–1773. doi: 10.1080/13548506.2021.1944651.
[11] Liang J, Li Y, Zhang Z, et al. Evaluating the applications of health information technologies in China during the past 11 years: consecutive survey data analysis. JMIR Med Inform. 2020;8(2):e17006. doi: 10.2196/17006.
[12] Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x.
[13] Leng A, Li Y, Wangen KR, et al. Hepatitis B discrimination in everyday life by rural migrant workers in hongqi. Hum Vaccin Immunother. 2016;12(5):1164–1171. doi: 10.1080/21645515.2015.1131883.
[14] Zou X, Chow EPF, Zhao P, et al. Rural-to-urban migrants are at high risk of sexually transmitted and viral hepatitis infections in China: a systematic review and meta-analysis. BMC Infect Dis. 2014;14(1):490. doi: 10.1186/1471-2334-14-490.
[15] Tao J, Zhang W, Yue H, et al. Prevalence of hepatitis B virus infection in Shenzhen, China, 2015-2018. Sci Rep. 2019;9(1):13948–13948. doi: 10.1038/s41598-019-50173-5.
[16] Wong WCW, Yang NS, Li J, et al. Crowdsourcing to promote hepatitis C testing and linkage-to-care in China: a randomized controlled trial protocol. BMC Public Health. 2020;20(1):1048. doi: 10.1186/s12889-020-09152-z.
[17] Wang Y, Du Z, Lawrence WR, et al. Predicting hepatitis B virus infection based on health examination data of community population. Int J Environ Res Public Health. 2019;16(23):4842. doi: 10.3390/ijerph16234842.
[18] Ajuwon BI, Richardson A, Roper K, et al. The development of a machine learning algorithm for early detection of viral hepatitis B infection in Nigerian patients. Sci Rep. 2023;13(1):3244. doi: 10.1038/s41598-023-30440-2.
[19] Chakraborty C, Gupta B, Ghosh SK. Mobile metadata assisted community database of chronic wound images. Wound Medicine. 2014;6:34–42. doi: 10.1016/j.wndm.2014.09.002.
[20] Othman SB, Almalki FA, Chakraborty C, et al. Privacy-preserving aware data aggregation for IoT-based healthcare with green computing technologies. Comput Electr Eng. 2022;101:108025. doi: 10.1016/j.compeleceng.2022.108025.
[21] Kishor A, Chakraborty C. Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE. Int J Syst Assur Eng Manag. 2021. doi: 10.1007/s13198-021-01174-z.
[22] Wong A, Plasek JM, Montecalvo SP, et al. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy. 2018;38(8):822–841. doi: 10.1002/phar.2151.
[23] Kades K, Sellner J, Koehler G, et al. Adapting bidirectional encoder representations from transformers (BERT) to assess clinical semantic textual similarity: algorithm development and validation study. JMIR Med Inform. 2021;9(2):e22795-e22795. doi: 10.2196/22795.
[24] Barber D. Bayesian reasoning and machine learning. Cambridge: Cambridge University Press; 2012.
[25] Zhong B, Xing X, Luo H, et al. Deep learning-based extraction of construction procedural constraints from construction regulations. Adv Eng Inf. 2020;43:101003. doi: 10.1016/j.aei.2019.101003.
[26] Dong W, Fong DYT, Yoon J-S, et al. Generative adversarial networks for imputing missing data for big data clinical research. BMC Med Res Methodol. 2021;21(1):78. doi: 10.1186/s12874-021-01272-3.
[27] Yoon J, Jordon J, van der Schaar M. GAIN: missing data imputation using generative adversarial nets. 2018.
[28] OECD [Internet]. Medical Classifications. [cited 15 May 2023]. Available from: https://www.oecd-ilibrary.org/content/component/9789264270985-24-en.
[29] Terrault NA, Lok ASF, McMahon BJ, et al. Update on prevention, diagnosis, and treatment of chronic hepatitis B: AASLD 2018 hepatitis B guidance. Hepatology. 2018;67(4):1560–1599. doi: 10.1002/hep.29800.
[30] Mayo M, Frank E. Improving naive bayes for regression with optimized artificial surrogate data. Appl Artif Intel. 2020;34(6):484–514. doi: 10.1080/08839514.2020.1726615.
[31] Ciaburro G, Venkateswaran B. Neural networks with R: smart models using CNN, RNN, deep learning, and artificial intelligence principl[es]. 1st ed. Birmingham (UK): PACKT Publishing; 2017.
[32] Maity S, Rastogi A, Djeddi C, et al. A novel optimized method for feature selection using non-linear Kernel-Free twin quadratic surface support vector machine. In: Cham: springer International Publishing; 2021:339–353.
[33] Cichosz P. Data mining algorithms: explained using R. Chichester, West Sussex: John Wiley & Sons Inc.; 2015.
[34] Kavzoglu T, Teke A. Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random Forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). Arab J Sci Eng. 2022;47(6):7367–7385. doi: 10.1007/s13369-022-06560-8.
[35] Singh M, Tyagi V, Gupta PK, et al. Accelerating the performance of sequence classification using GPU based ensemble learning with extreme gradient boosting. In:

vol 1613. Switzerland: Springer International Publishing AG; 2022:257–268.
[36] Wang B, Li C, Pavlu V, et al. Regularizing model complexity and label structure for Multi-Label text classification. 2017.
[37] Cameron AR, Meyer A, Faverjon C, et al. Quantification of the sensitivity of early detection surveillance. Transbound Emerg Dis. 2020;67(6):2532–2543. doi: 10.1111/tbed.13598.
[38] Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochem Med (Zagreb). 2016;26(3):297–307. doi: 10.11613/BM.2016.034.
[39] Merrick L, Taly A. The explanation game: explaining machine learning models using Shapley values. In: Holzinger A, Kieseberg P, Tjoa AM, Weippl E, editors. International Cross-Domain Conference for Machine Learning and Knowledge Extraction. Cham: Springer International Publishing; 2020. pp. 17-–38. doi: 10.1007/978-3-030-57321-8_2
[40] Lee N, Kim J-M. Conversion of categorical variables into numerical variables via Bayesian network classifiers for binary classifications. Computational Statistics & Data Analysis. 2010;54(5):1247–1265. doi: 10.1016/j.csda.2009.11.003.
[41] Eiseman NA, Bianchi MT, Westover MB. The information theoretic perspective on medical diagnostic inference. Hosp Pract (1995). 2014;42(2):125–138. doi: 10.3810/hp.2014.04.1110.
[42] NIH. [Internet]. Platelet disorders. Bethesda: National Heart, Lung, and Blood Institute; [cited 15 May 2023]. Available from: //www.nhlbi.nih.gov/health/thrombocytopenia
[43] Yin J, Tian L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with youden index. Computational Statistics & Data Analysis. 2014;77:1–13. doi: 10.1016/j.csda.2014.01.021.
[44] Terrault NA, Bzowej NH, Chang KM, et al. AASLD guidelines for treatment of chronic hepatitis B. Hepatology. 2016;63(1):261–283. doi: 10.1002/hep.28156.
[45] Marley G, Seto WK, Yan W, et al. What facilitates hepatitis B and hepatitis C testing and the role of stigma among primary care patients in China? J Viral Hepat. 2022;29(8):637–645. doi: 10.1111/jvh.13711.
[46] Zhou X, Zhang F, Ao Y, et al. Diagnosis experiences from 50 hepatitis B patients in Chongqing, China: a qualitative study. BMC Public Health. 2021;21(1):2195. doi: 10.1186/s12889-021-11929-9.
[47] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019. arXiv preprint arXiv:181004805.
[48] Ouyang L, Wu J, Xu J, et al. Training language models to follow instructions with human feedback. In. Ithaca: Cornell University Library, arXiv.org; 2022.
[49] Lee MH, Yang HI, Liu J, et al. Prediction models of long-term cirrhosis and hepatocellular carcinoma risk in chronic hepatitis B patients: risk scores integrating host and virus profiles. Hepatology. 2013;58(2):546–554. doi: 10.1002/hep.26385.
[50] Kao JH. Risk stratification of HBV infection in Asia-Pacific region. Clin Mol Hepatol. 2014;20(3):223–227. doi: 10.3350/cmh.2014.20.3.223.