

Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis

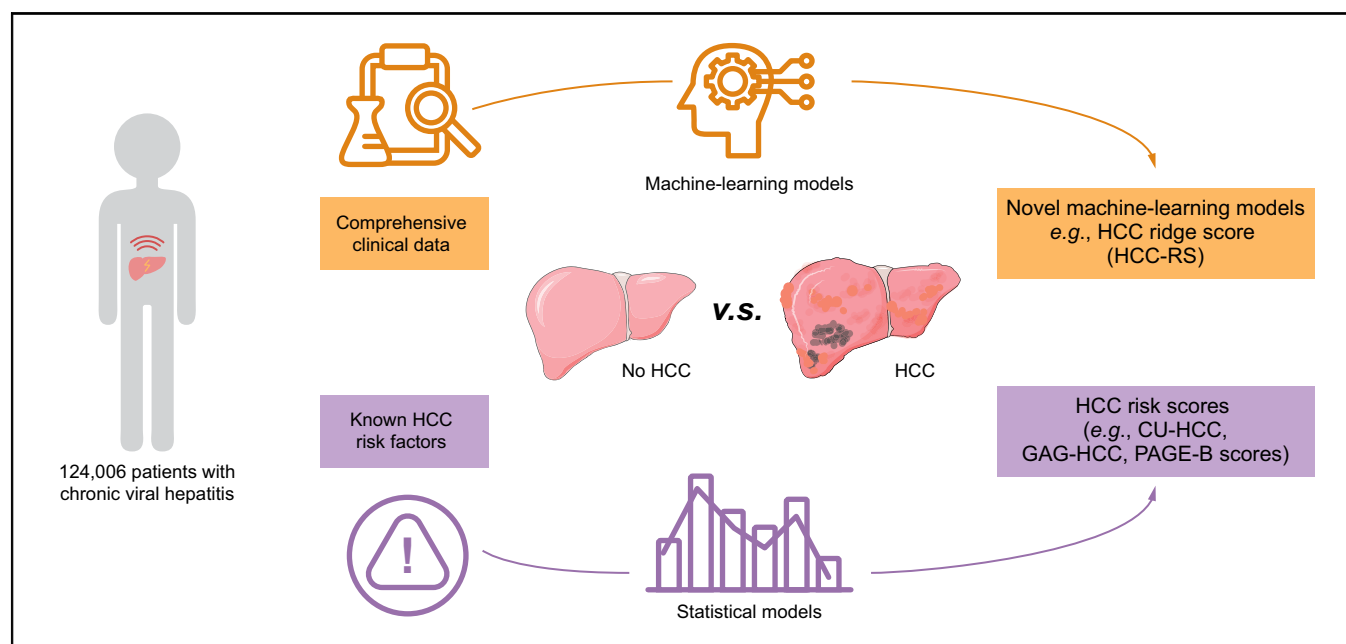
Authors

Grace Lai-Hung Wong, Vicki Wing-Ki Hui, Qingxiong Tan, Jingwen Xu, Hye Won Lee, Terry Cheuk-Fung Yip, Baoyao Yang, Yee-Kit Tse, Chong Yin, Fei Lyu, Jimmy Che-To Lai, Grace Chung-Yan Lui, Henry Lik-Yuen Chan, Pong-Chi Yuen, Vincent Wai-Sun Wong

Correspondence

pcyuen@comp.hkbu.edu.hk (P.-C. Yuen), wongv@cuhk.edu.hk (V.W.-S. Wong).

Graphical abstract



Highlights

- Accurate hepatocellular carcinoma (HCC) risk prediction is helpful in reducing mortality.
- Existing HCC risk scores usually include a few known risk factors and preselected parameters.
- Machine learning allows for direct selection of predictive parameters without subjective preselection.
- HCC ridge score (HCC-RS) built from machine learning modelling has higher accuracy than existing HCC risk scores.
- HCC-RS may be incorporated into electronic medical health systems to facilitate real-time update of HCC risk.

Lay summary

Novel machine learning models generated accurate risk scores for hepatocellular carcinoma (HCC) in patients with chronic viral hepatitis. HCC ridge score was consistently more accurate than existing HCC risk scores. These models may be incorporated into electronic medical health systems to develop appropriate cancer surveillance strategies and reduce cancer death.



Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis

Grace Lai-Hung Wong,^{1,2,3,†} Vicki Wing-Ki Hui,^{1,2,†} Qingxiong Tan,^{4,†} Jingwen Xu,⁴ Hye Won Lee,⁵ Terry Cheuk-Fung Yip,^{1,2,3} Baoyao Yang,⁴ Yee-Kit Tse,^{1,2} Chong Yin,⁴ Fei Lyu,⁴ Jimmy Che-To Lai,^{1,2,3} Grace Chung-Yan Lui,² Henry Lik-Yuen Chan,^{1,6} Pong-Chi Yuen,^{4,*} Vincent Wai-Sun Wong^{1,2,3,*}

¹Medical Data Analytic Centre, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; ²Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; ³Institute of Digestive Disease, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China; ⁴Department of Computer Science, Hong Kong Baptist University, Hong Kong Special Administrative Region, China; ⁵Department of Internal Medicine, Yonsei University College of Medicine, Seoul, South Korea; ⁶Union Hospital, Hong Kong Special Administrative Region, China

JHEP Reports 2022. <https://doi.org/10.1016/j.jhepr.2022.100441>

Background & Aims: Accurate hepatocellular carcinoma (HCC) risk prediction facilitates appropriate surveillance strategy and reduces cancer mortality. We aimed to derive and validate novel machine learning models to predict HCC in a territory-wide cohort of patients with chronic viral hepatitis (CVH) using data from the Hospital Authority Data Collaboration Lab (HADCL).

Methods: This was a territory-wide, retrospective, observational, cohort study of patients with CVH in Hong Kong in 2000–2018 identified from HADCL based on viral markers, diagnosis codes, and antiviral treatment for chronic hepatitis B and/or C. The cohort was randomly split into training and validation cohorts in a 7:3 ratio. Five popular machine learning methods, namely, logistic regression, ridge regression, AdaBoost, decision tree, and random forest, were performed and compared to find the best prediction model.

Results: A total of 124,006 patients with CVH with complete data were included to build the models. In the training cohort (n = 86,804; 6,821 HCC), ridge regression (area under the receiver operating characteristic curve [AUROC] 0.842), decision tree (0.952), and random forest (0.992) performed the best. In the validation cohort (n = 37,202; 2,875 HCC), ridge regression (AUROC 0.844) and random forest (0.837) maintained their accuracy, which was significantly higher than those of HCC risk scores: CU-HCC (0.672), GAG-HCC (0.745), REACH-B (0.671), PAGE-B (0.748), and REAL-B (0.712) scores. The low cut-off (0.07) of HCC ridge score (HCC-RS) achieved 90.0% sensitivity and 98.6% negative predictive value (NPV) in the validation cohort. The high cut-off (0.15) of HCC-RS achieved high specificity (90.0%) and NPV (95.6%); 31.1% of patients remained indeterminate.

Conclusions: HCC-RS from the ridge regression machine learning model accurately predicted HCC in patients with CVH. These machine learning models may be developed as built-in functional keys or calculators in electronic health systems to reduce cancer mortality.

Lay summary: Novel machine learning models generated accurate risk scores for hepatocellular carcinoma (HCC) in patients with chronic viral hepatitis. HCC ridge score was consistently more accurate than existing HCC risk scores. These models may be incorporated into electronic medical health systems to develop appropriate cancer surveillance strategies and reduce cancer death.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of European Association for the Study of the Liver (EASL). This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Keywords: Antiviral treatment; Cirrhosis; Liver cancer; Mortality; World Health Organization.

Received 25 August 2021; received in revised form 20 December 2021; accepted 28 December 2021; available online 22 January 2022

[†] These authors have equal contribution.

* Corresponding authors. Addresses: Department of Medicine and Therapeutics, 9/F Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, Hong Kong Special Administrative Region, China. Tel.: +852 3505 1205; Fax: +852 2637 3852; (V.W.-S. Wong); Department of Computer Science, Room 634, David C Lam Building (DLB634), Hong Kong Baptist University, Kowloon Tong, Hong Kong Special Administrative Region, China. Tel. +852 3411 7091; Fax: +852 3411 7892 (P.-C. Yuen).

E-mail addresses: wongv@cuhk.edu.hk (V.W.-S. Wong), pcyuen@comp.hkbu.edu.hk (P.-C. Yuen).



ELSEVIER

Introduction

Chronic viral hepatitis (CVH) is the seventh leading cause of mortality globally, responsible for 1.45 million deaths in 2013. The consequences of chronic hepatitis B and C infection—cirrhosis and liver cancer—account for 94% of deaths associated with hepatitis.^{1,2} Hepatocellular carcinoma (HCC) is the second most common cause of cancer death in the Asia-Pacific region.³ Approximately 78% of HCC cases are caused by CVH.^{4,5} The World Health Organization (WHO) published the first global health sector strategy on viral hepatitis in June 2016, setting the goals to reduce CVH incidence and mortality by 90% and 65%,

respectively, by 2030.⁶ Health ministries around the world have planned various strategies and action plans working towards such targets.

In Hong Kong, the Chief Executive's 2017 Policy Address instructed to set up the Steering Committee on Prevention and Control of Viral Hepatitis to formulate strategies that effectively prevent and control viral hepatitis.⁷ The Steering Committee advises the Government on policies and cost-effective, targeted strategies for prevention and control of viral hepatitis.⁸ The Hong Kong Viral Hepatitis Action Plan 2020–2024 was published in October 2020 to set out the strategic plan for reducing the burden of CVH through effective prevention, treatment, and control of viral hepatitis.⁹ Therefore, a comprehensive review of the disease burden of CVH and accurate prediction of HCC would provide pivotal data to the Government and the Steering Committee to guide strategies and action plans, ultimately to achieve the goals set by WHO.

Although most HCC risk prediction models have been developed using traditional regression analysis,¹⁰ machine learning is fast becoming a competitive alternative.¹¹ Machine learning is a comprehensive tool that has arisen in recent years for model development, which allows direct selection of predicting parameters among all available parameters without subjective preselection and maximises data use while minimising bias. In this study, we aimed to develop novel clinical and laboratory parameter-based prediction models using machine learning algorithms to define the risk levels of HCC in patients with CVH. These models can potentially be incorporated into computer-based management systems to facilitate clinical assessment and risk stratification of HCC in patients with CVH.

Patients and methods

Study design and data source

We performed a territory-wide, registry cohort study using data from the Hospital Authority Data Collaboration Lab (HADCL), Hong Kong. As announced in the Hong Kong Chief Executive's 2017 Policy Address, the Hospital Authority (HA) is establishing a Big Data Analytics Platform to support the formulation of healthcare policies, facilitate biotechnological research, and help improve clinical and healthcare services.¹² HADCL was set up as a new alternative channel for more flexible and interactive data sharing in HA, providing a secured collaboration platform between HA and external parties for deeper data analysis within a controlled environment and for conducting health data collaboration projects.¹² HADCL has been open for applications by academic institutions since October 2018 and was formally launched in December 2019. The current study was 1 of the first pilot projects. HADCL provides comprehensive yet anonymised and de-identified data of clinical parameters, namely, demographics, inpatient admissions, transfers and discharges, outpatient appointments, diagnosis, procedures, medications, laboratory tests and results, radiology examinations, clinical notes and summaries, radiology reports, and radiology images from all public hospitals and clinics in Hong Kong.¹³

Patients

We included all patients with CVH, that is, chronic hepatitis B (CHB) and chronic hepatitis C (CHC), from 1 January 2000 to 31 December 2018. CHB was defined as positive HBsAg for at least 6 months; and/or by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis

codes; and/or by use of antiviral treatment for CHB. CHC was defined as positive antibody to HCV (anti-HCV), and/or as detectable HCV RNA and/or HCV genotype, and/or by ICD-9-CM diagnosis codes, and/or by use of antiviral treatment for CHC (Tables S1 and S2).

Patients with missing date of birth or baseline date; coinfecting with HDV based on ICD-9-CM diagnosis codes, viral and/or serological markers; and/or coinfecting with HIV based on ICD-9-CM diagnosis codes were excluded (Table S1). Patients were followed up until death, diagnosis of HCC or hepatic events, last follow-up date (31 December 2018), or 15 years of follow-up, whichever came first. The study protocol was approved by the Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee.

Data collection

Data were retrieved from HADCL from May 2019 to August 2020. Baseline date was defined as the date of first diagnosis of CHB or CHC by viral markers, ICD-9-CM codes, or antiviral drugs, whichever came first. Demographic data including sex and date of birth were captured. At baseline, liver biochemistries, and haematological and virological parameters were collected. Thereafter, serial liver biochemistries as well as viral markers (HBsAg, HBeAg, HBV DNA, and HCV RNA) were collected until the last follow-up date (Table S3). Data on other relevant diagnoses, procedures, concomitant drugs, and laboratory parameters were also retrieved.

Antiviral treatment for CHB included oral nucleos(t)ide analogues (NAs), such as entecavir, tenofovir disoproxil fumarate, tenofovir alafenamide, adefovir dipivoxil, lamivudine and telbivudine, and (pegylated)-interferon for any duration. Antiviral treatment for CHC included (pegylated)-interferon with or without ribavirin, direct-acting antivirals (DAAs) such as asunaprevir, daclatasvir, dasabuvir/ombitasvir/paritaprevir, elbasvir/grazoprevir, glecaprevir/pibrentasvir, sofosbuvir, sofosbuvir/ledipasvir, sofosbuvir/velpatasvir, and sofosbuvir/velpatasvir/voxilaprevir for any duration (Table S2). The medication use was defined as those prescribed and dispensed for at least 4 weeks during the study period, identified by drug codes used in HA internally. The severity of liver fibrosis was assessed with serum formulae, namely, aspartate aminotransferase (AST)-to-platelet ratio index (APRI), Forns index, and Fibrosis-4 (FIB-4) in subgroups of patients with complete data for these formulae (Table S4).¹⁴ Advanced liver fibrosis was defined as APRI ≥ 2 , FIB-4 ≥ 3.25 , or Forns index ≥ 8.4 .¹⁵

Definitions of events

The primary event was HCC, identified based on diagnosis codes (155.0 [hepatocellular carcinoma] and 155.2 [carcinoma of the liver]) or procedure codes for HCC treatment according to ICD-9-CM codes retrieved from the Clinical Data Analysis and Reporting System (CDARS). Secondary events were hepatic events, defined based on ICD-9-CM codes of ascites, spontaneous bacterial peritonitis, variceal bleeding, hepatorenal syndrome, hepatic encephalopathy, liver transplantation, and/or liver-related mortality (Table S1). Liver cirrhosis was identified using ICD-9-CM diagnosis codes for cirrhosis and hepatic events at or before baseline (Table S1). The use of single ICD-9-CM codes for diagnosis was found to be 99% accurate when referenced to clinical, laboratory, imaging, and endoscopy results from electronic medical records.¹⁶

Statistical analysis

Data were analysed using SPSS version 25.0 (SPSS, Inc., Chicago, IL, USA), SAS (9.4; SAS Institute Inc., Cary, NC, USA), and R software (3.5.1; R Foundation for Statistical Computing, Vienna, Austria). Continuous variables were expressed in mean ± SD or median (IQR), as appropriate, whereas categorical variables were presented as n (%). Qualitative and quantitative differences between subgroups were analysed using Chi-square or Fisher’s exact tests for categorical parameters, and Student’s *t* test or the Mann–Whitney *U* test for continuous parameters, as appropriate. Qualitative and quantitative differences between ordinal subgroups were analysed using the Chi-square test for linear trend or Fisher’s exact tests for categorical parameters and one-way ANOVA or the Kruskal–Wallis test for continuous parameters, as appropriate. Cumulative incidences of primary and secondary endpoints with adjustment of competing death were estimated with 95% CI. Hazard ratios and adjusted hazard ratios (aHRs) with 95% CI were estimated with Fine–Gray proportional sub-distribution hazards regression with adjustment of competing death.¹⁷

The cohort was randomly split into training and validation cohorts in a 7:3 ratio. An additional external validation was also performed in an independent cohort of Korean patients. Five popular machine learning methods, namely, logistic regression, ridge regression, AdaBoost, decision tree, and random forest, were performed and compared to find the best prediction model.¹⁸ These machine learning algorithms were chosen as these supervised machine learning models are desirable for ensembling, that is, combining the predictions of multiple machine learning models to produce an accurate prediction.¹⁸ Logistic regression is 1 of the binary classifiers widely used in various medical applications. Ridge regression is applicable in scenarios where independent variables are highly correlated. AdaBoost is a meta-algorithm used in conjunction with many other types of learning algorithms to improve performance.

Decision tree trains a tree-like classifier in which each node depends on a variable as an easily interpretable classification model. Random forest ensembles multiple decision trees, which increases the generalisation accuracy.¹⁸ These machine learning methods have been used to identify patients with NAFLD in the general population¹¹ and patients with peptic ulcer bleeding.¹⁹ The machine learning models were built first by including all 46 available parameters, followed by a group of most predictive parameters *via* supervised feature selection with filter methods. These techniques look at the intrinsic properties of features and use statistical techniques to evaluate the relationship between a predictor and the target variable. The subset of best ranked parameters was then used for model training. We employed the default hyperparameters in the Scikit-learn Python package (Table S5) and did not perform any fine tuning. Machine learning models were trained using training dataset and tested using evaluation (test) dataset. The evaluation dataset was the test dataset. A detailed description of what was exactly used, and how to formulate these trees, is provided in Table S5. For the implementation, we employed the popular Scikit-learn machine learning package. The default parameters of each model were used, as listed in Table S5. The accuracy of the models was assessed by the area under the receiver operating characteristic curve (AUROC). Dual cut-offs were selected to achieve 90% sensitivity and 90% specificity to rule out and rule in patients with HCC, while maximising the corresponding specificity and sensitivity, respectively. The model with the highest AUROC in the validation cohort was treated as the most predictive model. This model was also compared with common HCC risk scores, namely, CU-HCC score, GAG-HCC score, REACH-B score, PAGE-B score, and REAL-B score.¹⁰ All statistical tests were 2-sided. Statistical significance was taken as a 2-sided *p* value of <0.05. Values of *p* for pairwise comparisons were adjusted using Bonferroni correction.

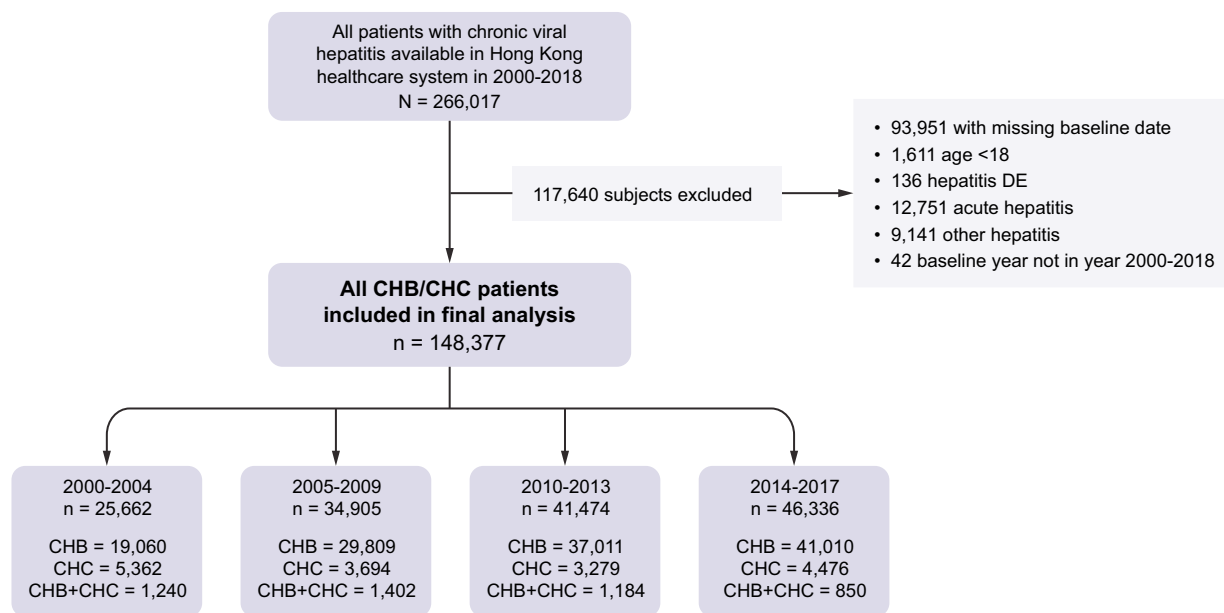


Fig 1. Selection of patients with CHB in the final analysis. CHB, chronic hepatitis B; CHC, chronic hepatitis C.

Table 1. Baseline clinical characteristics of patients with CHV first diagnosed at different periods from 2000 to 2018.

Period	Chronic hepatitis B (N = 126,890)				Chronic hepatitis C (N = 16,811)					
	2000–2004	2005–2009	2010–2013	2014–2018	2000–2004	2005–2009	2010–2013	2014–2018		
No. of patients	n = 19,060	n = 29,809	n = 37,011	n = 41,010	p value	n = 5,362	n = 3,694	n = 3,279	n = 4,476	p value
Male sex (n, %)	12,175 (63.88)	18,746 (62.89)	22,425 (60.59)	24,521 (59.79)	<0.001	3,343 (62.3)	2,616 (70.8)	2,268 (69.2)	3,089 (69.0)	<0.001
Age (years)	48.33 (15.49)	51.31 (14.43)	54.00 (14.17)	58.12 (14.24)	<0.001	51.4 (17.4)	54.7 (15.6)	56.5 (15.5)	57.0 (14.7)	<0.001
Platelet ($\times 10^9/L$)*	200.38 (99.19)	211.88 (90.63)	211.09 (83.18)	214.67 (96.33)	<0.001	209.0 (105.9)	211.2 (104.3)	210.9 (98.3)	218.7 (100.0)	<0.001
Prothrombin time (s)*	12.68 (3.67)	12.01 (3.20)	12.06 (7.49)	12.47 (3.36)	<0.001	12.4 (3.7)	12.3 (5.4)	12.6 (7.7)	12.5 (3.6)	0.341
Albumin (g/L)*	38.26 (6.57)	40.21 (6.09)	40.27 (5.88)	39.40 (6.20)	0.034	36.2 (6.6)	36.9 (6.7)	37.7 (6.5)	38.1 (6.2)	<0.001
Total bilirubin ($\mu\text{mol/L}$)*	11.30 (8.00–17.46)	12.00 (8.00–17.00)	11.60 (8.00–16.00)	11.00 (8.00–16.00)	<0.001	10.0 (7.00–16.0)	11.4 (8.00–17.0)	11.0 (8.00–16.0)	11.0 (8.00–16.0)	<0.001
ALT (IU/L)*	38.00 (23.00–71.00)	33.00 (21.00–59.00)	31.00 (20.00–51.00)	28.00 (18.00–48.00)	<0.001	33.00 (18.00–67.00)	38.00 (21.00–71.00)	36.00 (21.00–65.00)	36.90 (22.00–65.90)	<0.001
AST (IU/L)*	40.00 (26.00–71.00)	32.00 (23.00–50.00)	30.00 (22.00–46.00)	29.00 (21.00–46.00)	<0.001	42.00 (25.00–72.00)	42.00 (27.00–71.00)	40.00 (26.00–67.20)	39.00 (26.00–65.00)	0.187
APRI	1.90 (6.28)	1.02 (3.45)	0.93 (2.64)	1.10 (6.12)	<0.001	1.5 (5.3)	1.3 (2.7)	1.2 (2.0)	1.2 (4.0)	0.057
Forns index	6.05 (2.70)	5.83 (2.34)	5.95 (2.21)	6.32 (2.34)	<0.001	7.3 (2.4)	7.1 (2.4)	7.0 (2.5)	6.6 (2.3)	<0.001
FIB-4	0.75 (1.98)	0.55 (1.93)	0.62 (2.05)	0.84 (4.53)	<0.001	0.6 (1.3)	0.7 (1.7)	0.8 (1.7)	0.8 (4.3)	0.085
AFP ($\mu\text{mol/L}$)*	4.40 (3.–10.00)	3.59 (2.–6.69)	3.02 (2.–5.35)	3.01 (2.–5.15)	<0.001	6.2 (3.–15.1)	5.3 (3.–13.0)	4.5 (3.–9.0)	4.1 (3.–7.6)	<0.001
Positive HBeAg (n, %) [†]	647 (35.63)	1,949 (29.19)	2,953 (22.75)	2,236 (18.25)	<0.001					
Missing (%)	17,244 (90.47)	23,131 (77.60)	24,031 (64.93)	28,756 (70.12)						
HBV DNA (IU/L)*	5.26 (1.13)	1.60 (2.49)	0.50 (1.20)	0.15 (0.49)	<0.001					
Missing (%)	18,970 (99.53)	29,751 (99.81)	36,605 (98.90)	39,910 (97.32)						
Advanced liver disease										
APRI ≥ 2	552 (19.59)	1,017 (9.87)	948 (7.92)	1,307 (8.99)	<0.001	174 (16.2)	172 (13.7)	168 (14.6)	131 (9.6)	<0.001
FIB-4 ≥ 3.25	98 (3.48)	176 (1.71)	246 (2.05)	447 (3.08)	<0.001	21 (2.0)	40 (3.2)	44 (3.8)	39 (2.9)	0.071
Forns index ≥ 8.4	163 (22.42)	609 (14.82)	725 (13.63)	1,211 (16.58)	<0.001	50 (35.0)	101 (27.9)	114 (29.1)	99 (18.6)	<0.001
Comorbidities [‡] (n, %)										
Diabetes mellitus	2,856 (14.98)	5,326 (17.87)	7,985 (21.57)	11,244 (27.42)	<0.001	926 (17.3)	739 (20.0)	687 (21.0)	983 (22.0)	<0.001
Hypertension	4,180 (21.93)	9,566 (32.09)	14,000 (37.83)	18,744 (45.71)	<0.001	1,240 (23.1)	1,320 (35.7)	1,248 (38.1)	1,776 (39.7)	<0.001
Cardiovascular disease	2,667 (13.99)	4,600 (15.43)	7,468 (20.18)	10,020 (24.43)	<0.001	922 (17.2)	811 (22.0)	795 (24.2)	1,081 (24.2)	<0.001
Malignancy					<0.001					
Colorectal cancer	147 (0.77)	473 (1.59)	661 (1.79)	1,082 (2.64)	<0.001	26 (0.5)	28 (0.8)	22 (0.7)	35 (0.8)	0.24
Lung cancers	149 (0.78)	469 (1.57)	635 (1.72)	1,101 (2.68)	<0.001	28 (0.5)	45 (1.2)	58 (1.8)	52 (1.2)	<0.001
Urinary/renal malignancies	36 (0.19)	99 (0.33)	141 (0.38)	209 (0.51)	<0.001	10 (0.2)	16 (0.4)	12 (0.4)	16 (0.4)	0.18
Cervical cancer (female only)	12 (0.06)	66 (0.22)	66 (0.18)	134 (0.33)	<0.001	3 (0.1)	5 (0.1)	0 (0.0)	8 (0.2)	n.a.
Breast cancer	141 (0.74)	504 (1.69)	595 (1.61)	897 (2.19)	<0.001	9 (0.2)	10 (0.3)	12 (0.4)	19 (0.4)	0.101
Lymphoma	199 (1.04)	310 (1.04)	618 (1.67)	1,583 (3.86)	<0.001	8 (0.1)	19 (0.5)	12 (0.4)	12 (0.3)	0.017
Chronic kidney disease	475 (2.49)	550 (1.85)	812 (2.19)	1,149 (2.80)	<0.001	199 (3.7)	139 (3.8)	84 (2.6)	113 (2.5)	0.001

Descriptive statistics were calculated after subtraction of missing data from denominator. Total bilirubin, ALT, AST, alpha-fetoprotein, APRI, and FIB-4 are expressed in median (IQR), whereas other continuous variables are expressed in mean \pm SD. Statistical tests involved: chi-square test, Fisher's exact test, Student's t test, Mann-Whitney test, one-way ANOVA, Kruskal-Wallis.

AFP, alpha-fetoprotein; ALT, alanine aminotransferase; APRI, aspartate aminotransferase-to-platelet ratio index; AST, aspartate aminotransferase; CHV, chronic viral hepatitis; FIB-4, Fibrosis-4; ICD-9, International Classification of Diseases, Ninth Revision.

* Data were log-transformed before missing value imputation was performed. Values of p were also calculated based on log-transformed values.

[†] Percentages were computed based on non-missing values.

[‡] Comorbidities were all defined based on ICD-9 diagnosis codes.

Table 2. Cox regression model for the factors associated with various clinical outcomes in patients with complete data for at least 1 of the serum fibrosis formulae.

Parameters	Chronic hepatitis B						Chronic hepatitis C					
	Univariate			Multivariable			Univariate			Multivariable		
	HR	95% CI	p value	aHR	95% CI	p value	HR	95% CI	p value	aHR	95% CI	p value
HCC												
Male sex	2.68	(2.49–2.89)	<0.001	2.23	(1.98–2.50)	<0.001	0.98	(0.82–1.17)	0.819	1.70	(1.54–1.87)	<0.001
Age (years)	1.02	(1.01–1.02)	<0.001	1.02	(1.02–1.03)	<0.001	1.03	(1.03–1.04)	<0.001	1.04	(1.04–1.05)	<0.001
Albumin (g/L)	0.95	(0.95–0.96)	<0.001	0.98	(0.97–0.98)	<0.001	0.99	(0.98–1.00)	0.117	0.95	(0.95–0.96)	<0.001
ALT (>ULN)	2.59	(2.44–2.75)	<0.001	1.77	(1.61–1.94)	<0.001	2.83	(2.37–3.38)	<0.001	0.89	(0.82–0.98)	0.015
Positive HBeAg	1.42	(1.28–1.57)	<0.001	1.46	(1.31–1.62)	<0.001	–	–	–	–	–	–
Antiviral treatment	1.79	(1.69–1.90)	<0.001	1.68	(1.53–1.84)	<0.001	0.05	(0.01–0.18)	<0.001	1.79	(1.51–2.11)	<0.001
Advanced liver fibrosis	3.22	(3.02–3.42)	<0.001	1.41	(1.25–1.58)	<0.001	2.90	(2.45–3.43)	<0.001	1.34	(1.19–1.50)	<0.001
Hepatic events (non-HCC)												
Male sex	1.57	(1.48–1.67)	<0.001	1.32	(1.20–1.46)	<0.001	0.76	(0.66–0.87)	<0.001	0.87	(0.76–1.01)	0.067
Age (years)	1.02	(1.02–1.02)	<0.001	1.01	(1.01–1.02)	<0.001	1.02	(1.01–1.02)	<0.001	1.00	(1.00–1.01)	0.066
Albumin (g/L)	0.92	(0.92–0.92)	<0.001	0.95	(0.94–0.95)	<0.001	0.95	(0.94–0.96)	<0.001	0.96	(0.95–0.97)	<0.001
ALT (>ULN)	1.94	(1.83–2.04)	<0.001	1.28	(1.18–1.40)	<0.001	2.12	(1.85–2.44)	<0.001	1.50	(1.30–1.74)	<0.001
Positive HBeAg	1.17	(1.06–1.29)	0.002	1.17	(1.06–1.30)	0.002	–	–	–	–	–	–
Antiviral treatment	1.18	(1.11–1.25)	<0.001	1.01	(0.93–1.11)	0.758	0.04	(0.01–0.13)	<0.001	0.06	(0.02–0.19)	<0.001
Advanced liver fibrosis	4.34	(4.10–4.59)	<0.001	1.25	(1.13–1.40)	<0.001	4.01	(3.51–4.59)	<0.001	1.56	(1.32–1.85)	<0.001
Death												
Male sex	1.47	(1.42–1.53)	<0.001	1.54	(1.43–1.66)	<0.001	1.22	(1.11–1.33)	<0.001	1.70	(1.54–1.87)	<0.001
Age (years)	1.05	(1.05–1.05)	<0.001	1.04	(1.04–1.04)	<0.001	1.04	(1.04–1.05)	<0.001	1.04	(1.04–1.05)	<0.001
Albumin (g/L)	0.91	(0.91–0.91)	<0.001	0.93	(0.93–0.94)	<0.001	0.94	(0.94–0.95)	<0.001	0.95	(0.95–0.96)	<0.001
ALT (>ULN)	1.01	(0.97–1.05)	0.693	0.93	(0.87–1.00)	0.037	0.90	(0.82–0.97)	0.011	0.89	(0.82–0.98)	0.015
Positive HBeAg	0.76	(0.70–0.83)	<0.001	0.92	(0.85–1.01)	0.069	–	–	–	–	–	–
Antiviral treatment	3.00	(2.89–3.12)	<0.001	2.32	(2.17–2.48)	<0.001	1.22	(1.04–1.43)	0.017	1.79	(1.51–2.11)	<0.001
Advanced liver fibrosis	2.45	(2.35–2.56)	<0.001	1.41	(1.29–1.53)	<0.001	1.80	(1.63–1.98)	<0.001	1.34	(1.19–1.50)	<0.001

aHR, adjusted hazard ratio; ALT, alanine aminotransferase; HCC, hepatocellular carcinoma; HR, hazard ratio; ULN, upper limit of normal.

Table 3. Parameters used to develop the machine learning models.

Parameters	All (N = 46)	Selected mode 1 (n = 36)	Selected mode 2 (n = 20)
Male sex	✓	✓	✓
Age	✓	✓	✓
Platelet	✓	✓	✓
Albumin	✓	✓	✓
Total bilirubin	✓	✓	✓
ALT	✓	✓	✓
AST	✓		
Alpha-foetoprotein	✓		
International normalized ratio	✓		
Creatinine	✓		
Gamma glutamyl transferase	✓		
Total cholesterol	✓		
HbA _{1c}	✓		
Fasting glucose	✓		
HBV DNA	✓		
Positive HBeAg	✓		
Cirrhosis	✓	✓	✓
Cardiovascular disease	✓	✓	
Colorectal cancer	✓	✓	
Lung cancers	✓	✓	
Urinary/renal malignancies	✓	✓	
Cervical cancer	✓	✓	
Breast cancer	✓	✓	
Lymphoma	✓	✓	
Chronic kidney disease	✓	✓	✓
Osteopenia	✓	✓	
Osteoporosis	✓	✓	
Diabetes mellitus	✓	✓	✓
Hypertension	✓	✓	✓
Anticoagulants	✓	✓	
ACEI/ARB	✓	✓	✓
Antiplatelet agents	✓	✓	✓
Beta blockers	✓	✓	✓
Histamine-2 receptor antagonist	✓	✓	
Insulin	✓	✓	✓
Immunosuppressant	✓	✓	
Loop diuretics	✓	✓	
Metformin	✓	✓	✓
NSAID	✓	✓	
Other lipid-lowering agents	✓	✓	✓
Other oral hypoglycaemic agents	✓	✓	✓
Proton pump inhibitor	✓	✓	✓
Potassium sparing diuretics	✓	✓	
Statins	✓	✓	✓
Sulphonylurea	✓	✓	✓
Thiazides	✓	✓	

ACEI, angiotensin-converting-enzyme inhibitor; ALT, alanine aminotransferase; ARB, angiotensin receptor blocker; AST, aspartate aminotransferase; CTP, Child–Turcotte–Pugh; HbA_{1c}, haemoglobin A_{1c}.

Table 4. AUROC and the 95% CI of the machine learning models in training and validation cohorts to HCC.

Machine learning model	Training cohort* (N = 86,804, HCC = 6,821)			Validation cohort† (N = 37,202, HCC = 2,875)		
	20 selected parameters	36 selected parameters	All parameters	20 selected parameters	36 selected parameters	All parameters
Logistic regression	0.814 ± 0.006	0.829 ± 0.006	0.825 ± 0.006	0.818 ± 0.009	0.832 ± 0.009	0.829 ± 0.009
Ridge regression‡	0.817 ± 0.005	0.839 ± 0.005	0.842 ± 0.005	0.821 ± 0.009	0.840 ± 0.009	0.844 ± 0.009
AdaBoost	0.822 ± 0.006	0.828 ± 0.006	0.828 ± 0.006	0.824 ± 0.009	0.833 ± 0.009	0.832 ± 0.009
Decision tree§	0.877 ± 0.005	0.884 ± 0.005	0.800 ± 0.005	0.802 ± 0.010	0.819 ± 0.010	0.818 ± 0.010
Random forest‡,§	0.987 ± 0.003	0.991 ± 0.003	0.992 ± 0.003	0.807 ± 0.010	0.821 ± 0.010	0.821 ± 0.010

AUROC, area under the receiver operating characteristic curve; HCC, hepatocellular carcinoma.

* AUROC of the 5 machine learning algorithms were overall difference in the training cohort, *p* < 0.05.

† AUROC of the 5 machine learning algorithms were overall difference in the validation cohort, *p* < 0.05.

‡ AUROC higher than decision tree in the validation cohort, *p* < 0.05.

§ AUROC higher than logistic regression and AdaBoost in both cohorts, *p* < 0.05.

Table 5. Accuracy of the machine learning models using selected parameters in diagnosing HCC in the training and validation cohorts.

Machine learning algorithm	Dual cut-offs	n (%) (<lower cut-off / ≥ upper cut-off)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)
Training cohort (n = 86,804)						
Logistic regression	0.18	43,951 (50.6)	0.90 (0.89–0.91)	0.54 (0.542–0.545)	0.143 (0.141–0.146)	0.985 (0.983–0.985)
	0.29	11,527 (13.3)	0.52 (0.51–0.53)	0.90 (0.898–0.902)	0.307 (0.300–0.315)	0.956 (0.955–0.958)
Ridge regression	0.07	48,341 (55.7)	0.90 (0.89–0.91)	0.596 (0.593–0.599)	0.160 (0.156–0.164)	0.986 (0.985–0.987)
	0.15	11,506 (13.3)	0.52(0.51–0.53)	0.900 (0.898–0.902)	0.307 (0.300–0.315)	0.956 (0.955–0.958)
AdaBoost	0.42	43,298 (49.9)	0.91 (0.90–0.92)	0.533 (0.529–0.536)	0.142 (0.140–0.145)	0.985 (0.984–0.986)
	0.45	10,363 (11.9)	0.48 (0.46–0.49)	0.911 (0.909–0.913)	0.313 (0.308–0.320)	0.953 (0.952–0.954)
Decision tree	0.04	48,765 (56.2)	0.92 (0.92–0.93)	0.603 (0.599–0.606)	0.166 (0.163–0.170)	0.990 (0.989–0.991)
	0.17	12,029 (13.9)	0.63 (0.62–0.64)	0.903 (0.902–0.905)	0.356 (0.349–0.363)	0.966 (0.965–0.967)
Random forest	0.45	71,804 (82.7)	0.90 (0.90–0.91)	0.998 (0.997–0.998)	0.976 (0.973–0.979)	0.992 (0.991–0.992)
	0.10	12,074 (13.9)	0.97 (0.96–0.97)	0.932 (0.930–0.933)	0.547 (0.539–0.557)	0.997 (0.997–0.998)
Validation cohort (n = 37,202)						
Logistic regression	0.18	19,448 (52.3)	0.90 (0.89–0.91)	0.568 (0.553–0.565)	0.146 (0.142–0.151)	0.985 (0.984–0.987)
	0.29	4,930 (13.3)	0.52 (0.50–0.54)	0.900 (0.896–0.902)	0.304 (0.293–0.317)	0.957 (0.955–0.959)
Ridge regression	0.07	20,816 (56.0)	0.90 (0.89–0.91)	0.598 (0.593–0.603)	0.158 (0.152–0.164)	0.986 (0.985–0.988)
	0.15	4,932 (13.3)	0.52 (0.50–0.54)	0.900 (0.897–0.903)	0.304 (0.291–0.317)	0.957 (0.955–0.960)
AdaBoost	0.42	18,725 (50.3)	0.91 (0.90–0.92)	0.538 (0.532–0.543)	0.142 (0.137–0.146)	0.987 (0.985–0.988)
	0.45	4,377 (11.8)	0.47 (0.45–0.49)	0.912 (0.909–0.914)	0.310 (0.297–0.323)	0.954 (0.952–0.956)
Decision tree	0.02	17,689 (47.6)	0.90 (0.89–0.91)	0.507 (0.501–0.511)	0.133 (0.129–0.137)	0.983 (0.982–0.985)
	0.17	4,987 (13.4)	0.54 (0.52–0.56)	0.900 (0.897–0.904)	0.312 (0.302–0.330)	0.959 (0.957–0.961)
Random forest	0.01	17,561 (47.2)	0.90 (0.89–0.91)	0.503 (0.496–0.508)	0.132 (0.127–0.137)	0.984 (0.982–0.986)
	0.20	4,561 (12.3)	0.52 (0.50–0.53)	0.910 (0.907–0.913)	0.326 (0.312–0.341)	0.957 (0.955–0.959)

In the training cohort, dual cut-offs were selected to achieve >90% sensitivity and specificity. HCC, hepatocellular carcinoma; NPV, negative predictive value; PPV, positive predictive value.

Results

Demographic characteristics

We identified 266,017 patients with viral hepatitis; 117,640 patients were excluded according to inclusion and exclusion criteria (Fig. 1). In total, 148,377 patients with CVH (126,890 patients with CHB, 16,811 patients with CHC, and 4,676 patients with both CHB and CHC; Fig. 1) were included in the final analysis. The cohorts were predominantly male, and most patients had compensated liver disease. The prevalence of key comorbidities generally increased over time (Table 1).

Antiviral treatment

More than 40% of patients with CHB had received antiviral treatment by 2018. The increase in the cumulative treatment uptake first became obvious from 2005–2009 to 2010–2013 (from 12.05% to 17.76%), and then it dramatically increased in 2014–2018 (from 17.76% to 40.64%) (Table S6). The majority (51,191/51,572; 99.3%) of these patients received NAs as antiviral treatment; whereas only 981/51,572 patients (1.9%) received conventional or pegylated interferon as antiviral treatment. More than 30% of patients with CHC had received antiviral treatment by 2018. The majority (5,219/5,660; 92.2%) of these patients received conventional or pegylated interferon and ribavirin as antiviral treatment, whereas only 441/5,660 patients (7.8%) received DAAs as antiviral treatment as these only became available in Hong Kong in late 2013.

Among 44,193 patients with CHB with complete data for APRI, Forns, and/or FIB-4 indices, 5,849 patients (13.2%) had advanced liver fibrosis (Table S7). There appeared to be a trend for decreasing prevalence of advanced liver fibrosis, as evidenced by any of these 3 indices (from 21.8% in 2000–2004 to 13.6% in 2014–2018), by APRI ≥2 over the years (from 19.6% in 2000–2004 to 9.0% in 2014–2018), and by Forns index ≥8.4 (from 22.4% in 2000–2004 to 16.6% in 2014–2018). Only small proportions of patients (1.7–3.5%) were defined to have advanced liver fibrosis by FIB-4 ≥3.25; hence, an obvious trend was demonstrated. Treatment indication according to alanine aminotransferase (ALT) above 2 times the upper limit of normal (ULN) also decreased over time, from 25.0% in 2000–2004 to 12.6% in 2014–2018.

Among 5,249 patients with CHC with complete data for APRI, Forns, and/or FIB-4 indices, 903 (17.2%) patients had advanced liver fibrosis (Table S7). There was also a trend for decreasing prevalence of advanced liver fibrosis as evidenced by any of these 3 indices (from 18.6% in 2000–2004 to 10.2% in 2014–2018), by APRI ≥2 over the years (from 16.3% in 2000–2004 to 9.6% in 2014–2018), and by Forns index ≥8.4 (from 35.0% in 2000–2004 to 18.6% in 2014–2018). Again, only small proportions of patients were defined to have advanced liver fibrosis by FIB-4 ≥3.25 (2.0–3.8%).

Predictors of HCC, hepatic events, and death

By univariate and multivariable analyses, all well-established risk factors for HCC (namely, male sex, advanced age, hypoalbuminaemia, raised ALT, positive HBeAg, advanced liver fibrosis) were identified in patients with CHB (Table 2). The aHR was 1.41 (95% CI 1.25–1.58, *p* <0.001) for advanced liver fibrosis. Antiviral treatment was found to be a risk factor, which was likely explained by prescription bias whereby patients at higher risk of HCC (*viz.* cirrhosis and more active hepatitis) would have received antiviral therapy. Similar risk factors for hepatic events

Table 6. Accuracy of HCC-RS compared with existing HCC risk scores in diagnosing HCC in the validation (n = 37,202) cohorts.

Risk scores	AUROC	Dual cut-offs	n (%)	<lower cut-off	≥upper cut-off	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%) (95% CI)	NPV (%) (95% CI)
HCC-RS	0.840	0.07	20,816 (56.0)	4,932 (13.3)	15,884 (42.7)	90.0 (89.0–0.91)	59.8 (59.3–60.3)	15.8 (15.2–16.4)	98.6 (98.5–98.8)
CU-HCC score	0.672	<5	27,083 (72.8)	7,812 (21.0)	19,271 (52.8)	52.2 (50.3–54.0)	90.0 (89.7–90.3)	30.4 (29.1–31.7)	95.7 (95.5–96.0)
GAG-HCC score	0.745	≥20	25,781 (69.3)	2,939 (7.9)	22,842 (61.4)	46.4 (28.6–64.3)	74.0 (69.9–78.4)	10.3 (6.4–14.3)	95.6 (94.2–97.1)
REACH-B score	0.671	<8	18,601 (50.0)	558 (1.5)	18,043 (48.5)	32.1 (14.3–50.0)	71.5 (67.2–75.6)	9.1 (4.5–14.0)	94.8 (93.6–96.2)
PAGE-B score	0.748	<10	10,193 (27.4)	17,969 (48.3)	8,176 (21.9)	64.3 (46.4–82.1)	93.4 (91.1–95.6)	12.3 (8.8–15.9)	97.0 (95.5–98.4)
REAL-B score	0.712	<4	6,548 (17.6)	4,278 (11.5)	2,270 (6.1)	28.6 (14.3–46.4)	93.4 (91.1–95.6)	21.1 (10.5–33.3)	95.5 (94.5–96.6)
		≥8				72.7 (54.6–90.9)	52.8 (45.4–59.7)	16.2 (12.1–20.0)	94.1 (89.8–97.9)
						4.5 (0–13.5)	98.9 (97.4–100)	33.3 (0–100)	89.2 (88.7–90.2)
						95.7 (94.9–96.5)	29.4 (28.9–30.0)	10.7 (10.6–10.9)	98.7 (98.5–99.0)
						81.1 (79.4–82.7)	54.6 (54.0–55.3)	13.7 (13.4–14.0)	97.0 (96.8–97.3)
						96.0 (95.2–96.9)	19.2 (18.5–19.8)	12.0 (11.9–12.2)	97.7 (97.2–98.2)
						27.0 (25.0–29.1)	90.3 (89.8–90.7)	24.2 (22.6–25.8)	91.5 (91.3–91.7)

AUROC, area under the receiver operating characteristic curve; HCC, hepatocellular carcinoma; HCC-RS, HCC ridge score; NPV, negative predictive value; PPV, positive predictive value.

and death were identified, with aHRs for advanced liver fibrosis of 1.25 (95% CI 1.25–1.58, $p < 0.001$) and 1.41 (95% CI 1.29–1.53, $p < 0.001$), respectively (Table 2).

Using univariate and multivariable analyses, we identified all well-established risk factors for HCC (viz, male sex, advanced age, hypoalbuminaemia, and advanced liver fibrosis) in patients with CHC (Table 2). The aHR was 1.34 (95% CI 1.19–1.50, $p < 0.001$) for advanced liver fibrosis. Antiviral treatment was again found to be a risk factor, which was likely because this had been preferentially offered to patients with more advanced liver disease. Elevated ALT was an independent risk factor for hepatic events but not for HCC and death. Similar risk factors for hepatic events and death were identified, with aHRs for advanced liver fibrosis of 1.56 (95% CI 1.32–1.85, $p < 0.001$) and 1.34 (95% CI 1.19–1.50, $p < 0.001$), respectively (Table 2).

Machine learning models to predict HCC

A cohort of 124,006 patients was included to build the models, by including all 46 available parameters, with 36 or 20 selected parameters with best predictive power (Table 3). Baseline data of these parameters were used in these models. In the training cohort (n = 86,804; 6,821 HCC), random forest, decision tree, and ridge regression performed the best with all parameters (AUROC = 0.992 ± 0.001, 0.800 ± 0.004, and 0.842 ± 0.006, respectively), with 36 selected parameters (AUROC = 0.991 ± 0.002, 0.884 ± 0.004, and 0.839 ± 0.006, respectively) or with 20 selected parameters (AUROC = 0.987 ± 0.003, 0.877 ± 0.005, and 0.817 ± 0.005, respectively). In the validation cohort (n = 37,202; 2,875 HCC), ridge regression had consistently high accuracy with all parameters (0.844 ± 0.009), with 36 selected parameters (0.840 ± 0.009) and with 20 selected parameters (0.821 ± 0.009) (Table 4). Table 5 summarises the sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) of these models in the training and validation cohorts. The dual cut-off approach was applicable in more than 60% of patients in most models; the applicability was particularly high with random forest (96.6%) in the training cohort, but not in the validation cohort (59.5%). We have developed Windows software applications for these models, which are available via <https://drive.google.com/drive/folders/1Nb0rKpYakiHzcYnRUNmUfNPhDwd1rHHz?usp=sharing>.

HCC ridge score

As the ridge regression model achieved consistently good performances in the training and validation cohorts with all or selected parameters, HCC ridge score (HCC-RS) was formed for further comparisons. Different cut-offs of HCC-RS were tested, and these are summarised in Table S8. In order to achieve high sensitivity (≥90%), the low cut-off was set below 0.1, and that for high specificity (≥90%) was set between 0.1 and 0.2.

Performance of HCC-RS vs. common HCC risk scores

The AUROC was 0.672, 0.745, 0.671, 0.748, and 0.712 for CU-HCC score, GAG-HCC score, REACH-B score, PAGE-B score, and REAL-B score, respectively (Table 6). Using dual cut-offs, the low cut-off of REAL-B score (<4) had the highest sensitivity (96.0%) but was applicable only to a relatively small proportion of patients (17.6%); the high cut-off of REACH-B score (≥14) had the highest specificity (98.9%) but was applicable only to a minor proportion of patients (1.5%). HCC-RS performed better than these common HCC risk scores in terms of larger AUROC (0.840), high

applicability, and the small proportion of patients falling into the grey zone (30.7%).

External validation

External validation was performed with an independent cohort of 4,462 Korean patients, with 1,072 patients developing HCC. The accuracy of 4 of the 5 machine learning algorithms with 20 selected parameters remained satisfactory and comparable with those in the training cohort; the AUROC of logistic regression, ridge regression, AdaBoost, and random forest was above 0.80, whereas that of decision tree (0.799) was less satisfactory (Table S9). Table S10 summarises the sensitivity, specificity, PPV and NPV of these models in the external validation cohort. Dual cut-offs approach was applicable in more than 62% of patients in most models.

Discussion

This was the first territory-wide cohort study using the HADCL platform, which facilitated the inclusion of the majority of patients with CVH under the care of the public healthcare system in Hong Kong. We demonstrated that machine learning models by ridge regression and random forest were accurate to predict HCC in patients with CVH. These models may be developed as built-in functional keys or calculators in electronic health systems to facilitate hepatitis elimination.

Electronic health records (EHRs) are universally adopted in nearly all hospitals around the world; EHRs are rapidly growing in terms of number of patients, and quantity and variety of data. EHRs provide robust and comprehensive demographic and laboratory data in thousands to millions of patients. Unfortunately, clinical observations and anthropometric measurements might be missing in some EHRs, especially in regions where manual entries of data are not available.¹² One recent example of applying machine learning in hepatology was with the non-alcoholic fatty liver disease (NAFLD) ridge score, which was developed based on 5 laboratory parameters and 1 comorbidity.¹¹ The beauty of this NAFLD ridge score was its excellent NPV of 96% to exclude NAFLD.

Machine learning is not a stranger in the field of HCC but is more commonly applied to predicting clinical outcomes and prognosis.²⁰ Machine learning enables the processing of nonlinear data, identification of novel patterns between variables and outcomes, and inclusion of many different variables.²⁰ Some early applications of machine learning in patients with HCC included combinations of salivary metabolites derived from machine learning models for early detection of HCC.²¹ Despite these advantages, machine learning is neither standardised nor available for clinical practice.

Our current study demonstrates a novel application of machine learning in a much wider population of patients with CVH, which affects more than 300 million people worldwide. Similar to the NAFLD ridge score we built,¹¹ we included common clinical and laboratory parameters readily available from patients. This substantially increased the utility and applicability of the machine learning models. We once again found that ridge regression had consistently high accuracy, as was found with the NAFLD ridge score. Ridge regression is a technique for analysing multiple regression data that suffer from multicollinearity; when multicollinearity occurs, least squares estimates are unbiased, but their variances are large and, as such, may deviate far from the true value.¹⁸ Hence, ridge regression is particularly suitable

for machine learning models in clinical medicine, as many parameters included in the models are closely related and multicollinearity commonly occurs.¹⁰

The current machine learning models can be further optimised to account for changes in the epidemiology of patients with CVH, namely, increasing age and comorbidities²² and increasing antiviral treatment uptake, which modifies the natural history and reduces HCC risk.²³ Furthermore, recent studies suggest that tenofovir may further reduce HCC risk in patients with CHB.^{24,25} DAAs are also increasingly used in patients with CHC.²⁶ With increasing and changing use of these antiviral therapies, the machine learning models should be continuously optimised. A common approach is hyperparameter optimisation, which would be tuned by an exhaustive cross-validation grid search, with an independent cohort size of 30% of the entire cohort, as in our study.²⁷ Longitudinal, serial data may further increase the accuracy of risk prediction; however, prediction with irregular medical time series is challenging because the intervals between consecutive records vary significantly over time. Existing methods often handle this problem by generating regular time series from irregular medical records without considering the uncertainty in the generated data induced by the varying intervals. Thus, a novel Uncertainty-Aware Convolutional Recurrent Neural Network is proposed by our group, which introduces the uncertainty information in the generated data to boost risk prediction, potentially increasing the accuracy of these machine-learning methods.²⁸

The strength of our study includes the large sample size, as our study is by far the largest real-life cohort study, with close to 150,000 patients with CVH in total. The HADCL dataset contains robust demographic data, diagnosis coding system and death information, and complete serial laboratory parameters and drug information, which facilitated the analysis of the impact of comorbidities, use of medications, and clinical events. Data from real-life cohorts represent a wider spectrum of patients than do those from randomised controlled trials, in which patients with multiple comorbidities are often excluded. Findings from real-life cohorts are thus more readily applicable to routine clinical practice. Nonetheless, our study also had a few limitations. First, missing data and irregular intervals of laboratory measurement may lead to biases, as in other registry studies, although these biases can be partially compensated by our large cohort size. In particular, some laboratory assays, namely, serum HBV DNA, varied across different institutions, operators, and time periods. Fortunately, all public virology laboratories in Hong Kong adopted very similar assays; for example, Roche COBAS® AmpliPrep/COBAS® TaqMan® HBV Test v2.0 (Roche Diagnostics, Basel, Switzerland) is used for measuring serum HBV DNA. Also, the detection limit and technology of HBV DNA has changed significantly, particularly in a few landmark years: 2003, when the detection limit was lowered to 2,000 IU/ml, and 2010–2011, when the detection limit was further lowered to 10–20 IU/ml. We compared the accuracy of the machine learning methods in 2000–2010 and 2011–2018, and their accuracies were comparable (Tables S11–S13). Second, we might have missed some comorbidities of milder severity because of missed coding, such as hypertension, diabetes mellitus, cardiovascular disease, and early-stage chronic kidney disease. Coinfections with HCV and HDV were not 100% excluded as their antibodies were checked in only 4,359 (3.2%) and 69 (0.1%) of 135,395 patients, respectively, but these coinfections were uncommon (0.5 and <0.1%, respectively) in Hong Kong.²² This negative bias would result in under-

reporting of the prevalence of these comorbidities. Third, ascertainment bias may affect the reliability of the study because of the inaccurate entry of HCC in the HADCL dataset; however, the use of single ICD-9-CM codes in CDARS for the diagnosis of key events such as HCC was previously validated to be 99% accurate when referenced to clinical, laboratory, imaging, and endoscopy results from electronic medical records.¹⁶ Fourth, the exact time of CVH diagnosis could be earlier than we identified as some laboratory results for viral markers came from private laboratories before patients entered the public healthcare system; diagnosis coding was also not mandatory before 2008. Last, other unmeasured or uncaptured factors might have confounded the results. We do not have information on HBV and HCV

genotypes. Previous studies have shown that the majority of patients with CHB in Hong Kong have either genotype B or C HBV, and genotype C HBV is associated with an increased risk of HCC²⁹; genotypes 1 and 6 HCV are found in more than 80% of patients with CHC in Hong Kong.³⁰

In conclusion, this novel HCC-RS from ridge regression machine learning model accurately predicted HCC in patients with CVH. These machine learning models may be developed as built-in functional keys or calculators in electronic health systems to reduce cancer mortality. Prospective studies and randomised trials comparing machine learning model-guided HCC surveillance with routine clinical practice for the early diagnosis of HCC in patients with CVH are warranted.

Abbreviations

aHR, adjusted hazard ratio; ALT, alanine aminotransferase; anti-HCV, antibody to hepatitis C virus; APRI, aspartate transaminase-to-platelet ratio index; AUROC, area under the receiver operating characteristic curve; CDARS, Clinical Data Analysis and Reporting System; CHB, chronic hepatitis B; CHC, chronic hepatitis C; CI, confidence intervals; CVH, chronic viral hepatitis; DM, diabetes mellitus; HADCL, Hospital Authority Data Collaboration Lab; HBeAg, hepatitis B e antigen; HBsAg, hepatitis B surface antigen; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; ICD-9-CM, International Classification of Diseases, Ninth Revision Clinical Modification; NA, nucleos(t)ide analogue; RS, ridge score; WHO, World Health Organization.

Financial support

This work was supported by the Health and Medical Research Fund (HMRF) of the Food and Health Bureau (reference no.: 07180216) awarded to GW.

Conflicts of interest

GW has served as an advisory committee member for Gilead Sciences and Janssen; has served as a speaker for Abbott, Abbvie, Bristol-Myers Squibb, Echosens, Furui, Gilead Sciences, Janssen, and Roche; and has received a research grant from Gilead Sciences. TY has served as a speaker and an advisory committee member for Gilead Sciences. GL has served as an advisory committee member for Gilead, speaker for Merck and Gilead, and received a research grant from Gilead. HC is an advisor for AbbVie, Aligos, Aptorum, Arbutus, Hepion, Janssen, Gilead, GSK, Merck, Roche, Vaccitech, VenatoRx, and Vir Biotechnology; and a speaker for Mylan, Gilead, and Roche. VW has served as an advisory committee member for AbbVie, Allergan, Echosens, Gilead Sciences, Janssen, Perspectum Diagnostics, Pfizer, and Terns and as a speaker for Bristol-Myers Squibb, Echosens, Gilead Sciences, and Merck. The other authors declare that they have no competing interests.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Full access to all of the data in the study and integrity and accuracy of the data analysis: GW, QT, Y-KT, TY, VH. Study concept and design: All authors. Acquisition and analysis of data: GW, QT, Y-KT, TY, VH. Interpretation of data and drafting and critical revision of the manuscript for important intellectual content: All authors.

Data availability statement

Owing to the HADCL policy, the raw data remains confidential and will not be shared.

Acknowledgements

HADCL provided data, tools, platforms, health informatics, and professional support. All data were anonymised and not identifiable. Cathy Chow, PhD, Regional Scientific Director, Healthcare, Hong Kong, provided manuscript editing.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhepr.2022.100441>.

References

Author names in bold designate shared co-first authorship

- [1] World Health Organization. Hepatitis B fact sheet. <http://www.who.int/en/news-room/fact-sheets/detail/hepatitis-b>. Accessed 18 June 2021.
- [2] World Health Organization. Hepatitis C fact sheet. <http://www.who.int/en/news-room/fact-sheets/detail/hepatitis-c>. Accessed 18 June 2021.
- [3] GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2018;392:1736–1788.
- [4] Sarin SK, Kumar M, Eslam M, George J, Al Mahtab M, Akbar SMF, et al. Liver diseases in the Asia-Pacific region: a lancet gastroenterology & hepatology commission. *Lancet Gastroenterol Hepatol* 2020;5:167–228.
- [5] World Health Organization. Global hepatitis report. 2017; <https://www.who.int/hepatitis/publications/global-hepatitis-report2017/en/>. [Accessed 20 April 2021].
- [6] World Health Organization. Global health sector strategy on viral hepatitis 2016–2021. <https://www.who.int/hepatitis/strategy2016-2021/ghss-hep/en/>. Accessed 12 July 2021.
- [7] Government HKSAR. https://www.policyaddress.gov.hk/2017/eng/pdf/Agenda_Ch6.pdf. Accessed 21 June 2021.
- [8] Health Do. Steering committee on prevention and control of viral hepatitis. https://www.hepatitis.gov.hk/english/about_us/about_us.html#:~:text=Contact%20Us,-Steering%20Committee%20on%20Prevention%20and%20Control%20of%20Viral%20Hepatitis,prevent%20and%20control%20viral%20hepatitis. Accessed 12 July 2021.
- [9] Government HKSAR. Hong Kong viral hepatitis action plan 2020–2024. October 2020; https://www.hepatitis.gov.hk/doc/action_plan/Action%20Plan_Full%20Version_PDF_en.pdf. [Accessed 21 June 2021].
- [10] Yip TC, Hui VW, Tse YK, Wong GL. Statistical strategies for HCC risk prediction models in patients with chronic hepatitis B. *Hepatology Res* 2021;7:7.
- [11] **Yip TC, Ma AJ**, Wong VW, Tse YK, Chan HL, Yuen PC, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. *Aliment Pharmacol Ther* 2017;46:447–456.
- [12] Hospital Authority. Hospital Authority Data Collaboration Lab. <https://www3.ha.org.hk/data/DCL/Index/>. Accessed 21 June 2021.
- [13] Hospital Authority. Data Catalogue, Data Collaboration Project, Hospital Authority Data Collaboration Lab. <https://www3.ha.org.hk/data/DCL/ProjectDataCatalogue/>. Accessed 21 June 2021.
- [14] Wong GL, Wong VW, Choi PC, Chan AW, Chan HL. Development of a non-invasive algorithm with transient elastography (Fibroscan) and serum test formula for advanced liver fibrosis in chronic hepatitis B. *Aliment Pharmacol Ther* 2010;31:1095–1103.
- [15] Liang LY, Wong VW, Tse YK, Yip TC, Lui GC, Chan HL, et al. Improvement in enhanced liver fibrosis score and liver stiffness measurement reflects lower risk of hepatocellular carcinoma. *Aliment Pharmacol Ther* 2019;49:1509–1517.

- [16] Wong JC, Chan HL, Tse YK, Yip TC, Wong VW, Wong GL. Statins reduce the risk of liver decompensation and death in chronic viral hepatitis: a propensity score weighted landmark analysis. *Aliment Pharmacol Ther* 2017;46:1001–1010.
- [17] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 1999;94:496–509.
- [18] Wong GL, Yuen PC, Ma AJ, Chan AW, Leung HH, Wong VW. Artificial intelligence in prediction of non-alcoholic fatty liver disease and fibrosis. *J Gastroenterol Hepatol* 2021;36:543–550.
- [19] Wong GL, Ma AJ, Deng H, Ching JY, Wong VW, Tse YK, et al. Machine learning model to predict recurrent ulcer bleeding in patients with history of idiopathic gastroduodenal ulcer bleeding. *Aliment Pharmacol Ther* 2019;49:912–918.
- [20] Zou ZM, Chang DH, Liu H, Xiao YD. Current updates in machine-learning in the prediction of therapeutic outcome of hepatocellular carcinoma: what should we know? *Insights Imaging* 2021;12:31.
- [21] Hershberger CE, Rodarte AI, Siddiqi S, Moro A, Acevedo-Moreno LA, Brown JM, et al. Salivary metabolites are promising non-invasive biomarkers of hepatocellular carcinoma and chronic liver disease. *Liver Cancer Int* 2021;2:33–44.
- [22] Wong GL, Wong VW, Yuen BW, Tse YK, Luk HW, Yip TC, et al. An aging population of chronic hepatitis B with increasing comorbidities: a territory-wide study from 2000 to 2017. *Hepatology* 2020;71:444–455.
- [23] Hui VW, Chan SL, Wong VW, Liang LY, Yip TC, Lai JC, et al. Increasing antiviral treatment uptake improves survival in patients with HBV-related HCC. *JHEP Rep* 2020;2:100152.
- [24] Yip TC, Lai JC, Wong GL. Secondary prevention for hepatocellular carcinoma in patients with chronic hepatitis B: are all the nucleos(t)ide analogues the same? *J Gastroenterol* 2020;55:1023–1036.
- [25] Yip TC, Wong VW, Chan HL, Tse YK, Lui GC, Wong GL. Tenofovir is associated with lower risk of hepatocellular carcinoma than entecavir in patients with chronic HBV infection in China. *Gastroenterology* 2020;158:215–225e6.
- [26] Hui YT, Wong GLH, Fung JYY, Chan HLY, Leung NWW, Liu SD, et al. Territory-wide population-based study of chronic hepatitis C infection and implications for hepatitis elimination in Hong Kong. *Liver Int* 2018;38:1911–1919.
- [27] Liu X, Lu J, Zhang G, Han J, Zhou W, Chen H, et al. A machine learning approach yields a multiparameter prognostic marker in liver cancer. *Cancer Immunol Res* 2021;9:337–347.
- [28] Tan Q, Ye M, Ma AJ, Yang B, Yip TC, Wong GL, et al. Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series. *IEEE Trans Neural Netw Learn Syst* 2021:4665–4679.
- [29] Chan HL, Hui AY, Wong ML, Tse AM, Hung LC, Wong VW, et al. Genotype C hepatitis B virus infection is associated with an increased risk of hepatocellular carcinoma. *Gut* 2004;53:1494–1498.
- [30] Wong GL, Chan HL, Loo CK, Hui YT, Fung JY, Cheung D, et al. Change in treatment paradigm in people who previously injected drugs with chronic hepatitis C in the era of direct-acting antiviral therapy. *J Gastroenterol Hepatol* 2019;34:1641–1647.