Prediction of the Risk of Developing Hepatocellular Carcinoma in Health Screening Examinees: a Korean Cohort Study

Chansik An
National Health Insurance Service Ilsan Hospital
Jong Won Choi
National Health Insurance Service Ilsan Hospital
Hyung Soon Lee
National Health Insurance Service Ilsan Hospital
Hyunsun Lim
National Health Insurance Service Ilsan Hospital
Seok Jong Ryu
National Health Insurance Service Ilsan Hospital
Jung Hyun Chang (jhchang@nhimc.or.kr)
National Health Insurance Service Ilsan Hospital
Hyun Cheol Oh
National Health Insurance Service Ilsan Hospital

Research Article

Keywords: Big Data, Machine Learning, Liver Neoplasms, Precision Medicine

DOI: https://doi.org/10.21203/rs.3.rs-343547/v1

Abstract

Background

Almost all Koreans are covered by mandatory national health insurance and are required to undergo health screening at least once every 2 years. We aimed to develop a machine learning model to predict the risk of developing hepatocellular carcinoma (HCC) based on the screening results and insurance claim data.

Methods

The National Health Insurance Service-National Health Screening database was used for this study (NHIS-2020-2-146). Our study cohort consisted of health screening examinees in 2004 or 2005 without cancer history, which was randomly split into training and test cohorts. Robust predictors were selected using Cox proportional hazard regression with 1,000 different bootstrapped datasets. Random forest and extreme gradient boosting algorithms were used to develop a prediction model for the 12-year risk of HCC development after screening. After optimizing a prediction model via cross validation in the training cohort, the model was validated in the test cohort.

Results

Of 331,694 examinees, 0.8% were diagnosed with HCC during the follow-up period (median, 11.2 years), respectively. Of the selected predictors, older age, male sex, abnormal liver function tests, the family history of chronic liver disease, and underlying chronic liver disease, chronic hepatitis virus or human immunodeficiency virus infection, and diabetes mellitus were associated with increased risk, whereas elevated total cholesterol and underlying dyslipidemia or schizophrenic/delusional disorders were associated with decreased risk of HCC development ($p<0.001$). In the test, our model showed good discrimination and calibration. The *C*-index, AUC, and Brier skill score were 0.868, 0.872, and 0.08, respectively.

Conclusions

Machine learning-based model could be used to predict the risk of HCC development based on the health screening examination results and claim data.

Background

Hepatocellular carcinoma (HCC) is the third most common cause of cancer death worldwide, with over half a million new cases diagnosed annually worldwide [1, 2]. In South Korea (hereafter Korea), HCC and other primary liver cancer are the fourth most common cancer in men and the sixth in women, and the second largest cause of cancer mortality [3].

Almost all Koreans are covered by mandatory national health insurance or Medical Care (a governmental program corresponding to the US Medicaid), and all insured adults aged 40 years or older are required to undergo a national general health screening examination at least once every 2 years. All the claim and

health screening data produced are accumulated in the database of the national health insurance system and can be used for a research purpose with permission. The national health screening examination is intended for screening general health risk factors. However, we postulated that new values could be derived that can be used to predict the risk of development of a certain disease if the examination results are used in combination with the claim data.

As the healthcare insurance claim and screening data contain information related to the risk of developing HCC such as demographic characteristics, family medical history, laboratory results including liver enzymes, and various underlying medical conditions including chronic liver disease and viral infection [4], we hypothesized that a machine learning algorithm may be utilized to predict the risk of HCC for each participant of the national health screening examination.

Several models have been proposed to predict the risk of HCC development [5–10]. However, to the best of our knowledge, they were for patients who are already at high risk for HCC. If a prediction model targets all screening examinees that include not only people who are already aware of their risks for HCC but also those who are not, it could play an additional important role in identifying undiagnosed high-risk patients.

Therefore, the purpose of this study was to identify risk factors and develop a machine learning model to predict the risk of HCC development for an individual examinee within 12 years after the national health screening examination, with a large cohort of Koreans.

Methods

Study population

The National Health Insurance Service-National Health Screening (NHIS-HEALS) database is a sample cohort of 514,795 people, accounting for 10% of all health screening examinees aged 40–80 years in 2002 or 2003 in South Korea, and contains the information on their claim data and the results of their health screening examinations between 2002 and 2015. Detailed information on the NHIS-HEALS database has been outlined elsewhere [11, 12]. This retrospective cohort study was approved by the Institutional Review Board of National Health Insurance Service Ilsan Hospital (NHIMC 2020-06-033), and the informed consent from the participants was waived.

Of the total 514,795 people, 334,966 were included in this study who also underwent the health screening in 2004 or 2005, in order to use 2002 and 2003 as a washout period. People who died (n = 2) or were diagnosed as having cancer (n = 2,487) during the washout period were excluded. Furthermore, people covered by Medical Care (n = 783) were excluded because their healthcare service claim is significantly different than the general population. The final study cohort was randomly split into train and test cohort with a ratio of 7:3, while preserving the same proportions of HCC occurrence (Fig. 1). The test cohort was only used for the final test of the machine learning models developed in the training cohort.

Variables

Variables for input features or outcomes were extracted from the NHIS-HEALS cohort following processing and cleaning the dataset. The full description of variables included can be found in Supplementary Table 1.

Input features

Variables retrieved from the healthcare claim data included sociodemographic variables, underlying medical conditions, and prescription records. In the NHIS-HEALS cohort, diagnoses were coded according to the Korean Standard Classification of Diseases (KCD) version 6, which is based on the International Classification of Diseases 10th revision (ICD-10). However, the diagnosis claimed by the healthcare providers and the actual diagnosis may differ because the dataset was established for recording claims and reimbursements. Therefore, for major diseases such as hypertension, diabetes mellitus, dyslipidemia, heart diseases, and stroke, operational definition was used as previously reported [13]. For example, the diagnosis of hypertension was determined to occur when a patient on antihypertensive medication was admitted for the first time or visited outpatient clinic for a second time with ICD-10 codes for hypertensive disease. (See Supplementary Table 2 for the definitions of all the underlying medical conditions used in this study). In the NHIS-HEALS data, some diagnostic codes were masked as sensitive personal information; for example, human immunodeficiency viruses (HIVs) (B20–B24) were grouped under the B_ code, and mental and behavioral disorders due to psychoactive substance use (F10–F19) and schizophrenia, schizotypal and delusional disorders (F20–F29) were coded as F_ altogether.

The health screening data included physical examination results (height, weight, and blood pressure), laboratory results (fasting glucose, total cholesterol, hemoglobin, urine stick test, liver enzymes), information obtained from history taking or questionnaires (family medical history, smoking history, alcohol consumption, and exercise habit).

Outcome variables

For probability prediction (i.e., classification task), the outcome was whether HCC was diagnosed within 12 years from the health screening examination. For time-to-event prediction (i.e., survival analysis), the outcome was the time interval between the examination and the diagnosis. The diagnoses of other cancers were considered competing risks. Participants who were not diagnosed with HCC or who died from other causes during the observation period were right-censored. The NHIS-HEALS data contains the date and cause of death statistics extracted from the national database produced by Statistics Korea. The last follow-up date was December 31, 2015.

Statistical analysis and machine learning

All analyses were performed using R 3.3.3. Main packages used include ‘survival (v2.41-3)’, ‘cmprsk (v2.2-7)’, ‘randomForestSRC (v2.5.1)’, ‘caret (v6.0-78)’, ‘survminer (0.4.2)’, and ‘xgboost (0.6.4.1)’. In Table 1, continuous and categorical variables were compared using Mann-Whitney or t-test and chi-square test, respectively. Two-sided probability values of < 0.05 were considered statistically significant.

Table 1
Baseline characteristics

Variable		Training cohort	Test cohort	*p*-value	Total
Sociodemographic characteristics					
Age (years)		54.3 (9.28)	94.27 (9.28)	0.43	54.29 (9.28)
Sex	Female	42.2% (98048/232186)	42.2% (41987/99508)	0.86	42.2% (140035/331694)
	Male	57.8% (134138/232186)	57.8% (57521/99508)		57.8% (191659/331694)
Residence	Metropolitan	16.7% (38808/232186)	16.5% (16371/995038)	0.15	16.6% (55179/331694)
	Urban or suburban	28.2% (65516/232186)	28.2% (28066/99508)		28.2% (93582/331694)
	Rural	55.1% (127862/232186)	55.3% (55071/99508)		55.2% (182933/331694)
Income level	< 30%	28.6% (66305/232186)	28.4% (28309/99508)	0.77	28.5% (94614/331694)
	30–80%	34.9% (80925/232186)	35% (34788/99508)		34.9% (115713/331694)
	> 80%	36.6% (84956/232186)	36.6% (36411/99508)		36.6% (121367/331694)
Physical examination					
Obesity	Normal (BMI < 25 kg/m^2)	66.2% (153721/232070)	66.1% (65756/99459)	0.9	66.2% (219447/331529)
	Overweight (25–30 kg/m^2)	31.3% (72675/323070)	31.4% (31247/99459)		31.3% (103922/331529)
	Obese (> 30 kg/m^2)	2.4% (5674/232070)	2.5% (2456/99459)		2.5% (8130/331529)
Systolic BP (mmHg)		126.57 (17.17)	126.63 (17.23)	0.39	126.59 (17.19)
Diastolic BP (mmHg)		79.18 (11.15)	79.18 (11.16)	0.96	79.18 (11.15)

For continuous variables, numbers in parentheses are standard deviation. BMI = body mass index, BP = blood pressure, AST = aspartate aminotransferase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase.

*This is not a complete table of underlying medical conditions. The full table can be found in Supplementary Table 1.

Variable		Training cohort	Test cohort	*p*-value	Total
Blood test					
AST (IU/L)		26.56 (16)	26.52 (15.72)	0.42	26.55 (15.92)
ALT (IU/L)		25.52 (19.56)	25.50 (19.18)	0.77	25.52 (19.45)
GGT (IU/L)		37.73 (52.2)	37.64 (51.03)	0.66	37.7 (51.85)
Total cholesterol (mg/dL)		198.37 (36.86)	198.2 (36.8)	0.2	198.32 (36.84)
Fasting blood glucose (mg/dL)		97.89 (29.18)	97.81 (28.24)	0.42	97.87 (28.9)
Hemoglobin (g/dL)		13.95 (1.49)	13.94 (1.49)	0.73	13.94 (1.49)
Urine stick test					
Urine pH level (pH)		6.09 (0.64)	6.1 (0.64)	0.39	6.09 (0.64)
Urine Glucose	Negative or weak	97.4% (225319/231397)	97.4% (96616/99173)	0.8	97.4% (321935/330570)
	Medium	0.9% (2196/231397)	0.9% (941/99173)		0.9% (3137/330570)
	Strong	1.7% (3882/231397)	1.6% (1616/99173)		1.7% (5498/330570)
Urine Occult blood	Negative or weak	93.2% (215604/231385)	93.1% (92285/99169)	0.64	93.1% (307889/330554)
	Medium	3.8% (8776/231385)	3.9% (3838/99169)		3.8% (12614/330554)
	Strong	3% (7005/231385)	3.1% (3046/99169)		3% (10051/330554)
Urine protein	Negative or weak	98.2% (227156/231402)	98.1% (97326/99177)	0.24	98.2% (324482/330579)
	Medium	1.3% (2903/231402)	1.2% (1219/99177)		1.2% (4122/330579)
	Strong	0.6% (1343/231402)	0.6% (632/99177)		0.6% (1975/3305/79)
Habit					

For continuous variables, numbers in parentheses are standard deviation. BMI = body mass index, BP = blood pressure, AST = aspartate aminotransferase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase.

*This is not a complete table of underlying medical conditions. The full table can be found in Supplementary Table 1.

Variable		Training cohort	Test cohort	*p*-value	Total
Smoking (pack-year)		5.98 (11.56)	5.92 (11.46)	0.18	5.96 (11.53)
Alcohol consumption	Rarely	56.3% (128639/2283638)	56.2% (590593/9/8/35)	0.48	56.3% (183692/326241)
	2–3 per month	15.6% (35665/228368)	15.5% (15180/97873)		15.6% (50845/326241)
	1–2 per week	17.4% (39629/228368)	17.5% (17135/97873)		17.4% (56764/326241)
	3–4 per week	6.8% (15453/228368)	6.9% (6744/97873)		6.8% (22197/326241)
	Almost everyday	3.9% (8982/228368)	3.8% (3761/97873)		3.9% (12743/326241)
Exercise	Rarely	50.1% (113877/227/095)	50.1% (48791/97411)	0.59	50.1% (162668/324506)
	1–2 per week	26.9% (61202/227095)	26.9% (26241/9/7411)		26.9% (87443/324506)
	3–4 per week	12.1% (27560/22/095)	12.1% (11781/97411)		12.1% (39341/324506)
	5–6 per week	3.2% (/2/0/227095)	3.2% (3114/97411)		3.2% (10384/324506)
	Almost everyday	7.6% (17186/227095)	7.7% (7484/97411)		7.6% (24670/324506)
Family history					
Liver disease		2.8% (5989/213528)	2.81% (2574/91558)	0.9	2.81% (8563/305086)
Hypertension		9.16% (19648/214527/)	9.16% (8424/92012)	0.77	9.16% (28072/306539)
Stroke		5.49% (11742/214048)	5.44% (4996/91769)	0.85	5.47% (16738/305817)
Heart disease		2.39% (5102/213537)	2.38% (21/76/91569)	0.85	2.39% (7278/305106)
Diabetes mellitus		6.4% (13718/214195)	6.54% (6011/91853)	0.3	6.45% (19729/306048)

For continuous variables, numbers in parentheses are standard deviation. BMI = body mass index, BP = blood pressure, AST = aspartate aminotransferase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase.

*This is not a complete table of underlying medical conditions. The full table can be found in Supplementary Table 1.

Variable	Training cohort	Test cohort	*p*-value	Total
Cancer (all types)	13.1% (28189/215156)	13.22% (12200/9228/)	0.49	13.14% (40389/307443)
Underlying medical condition*				
Diabetes mellitus	6.08% (14114/232186)	6.05% (6025/99508)	0.8	6.07% (20139/331694)
Dyslipidemia	5.95% (13811/232186)	5.94% (5914/99508)	0.96	5.95% (19725/331694)
Hypertension	19.08% (44298/232186)	19.31% (19213/99508)	0.13	19.15% (63511/331694)
Chronic hepatitis virus infection	2.55% (5915/232186)	2.63% (2615/99508)	0.16	2.57% (8530/331694)
Human Immunodeficiency virus	6.69% (15527/232186)	6.82% (6787/99508)	0.16	6.73% (22314/331694)
Schizophrenic or delusional disorders, or mental disorders due to psychoactive substance use	15.55% (36110/232186)	15.68% (15604/99508)	0.35	15.59% (51714/331694)
Chronic liver disease	5.77% (13397/232186)	5.17% (5149/99508)	0.15	5.59% (18546/331694)

For continuous variables, numbers in parentheses are standard deviation. BMI = body mass index, BP = blood pressure, AST = aspartate aminotransferase, ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase.

*This is not a complete table of underlying medical conditions. The full table can be found in Supplementary Table 1.

Feature selection

Including irrelevant features in a machine learning model likely results in overfitting and can undermines the generalizability of a prediction model [14]. Thus, feature selection was performed using Cox proportional hazard (CoxPH) regression in the training cohort. First, multicollinearity among the features was examined by calculating variance inflation factors (VIFs). Systolic/diastolic blood pressure and aspartate transaminase (AST)/alanine transaminase (ALT) were determined to have strong correlation as they showed VIFs > 2.5. Thus, mean average was calculated and used instead of systolic or diastolic blood pressure, and AST was discarded as ALT is more specific to liver disease. Next, with the features that showed statistically significant ($p < 0.05$) associations with HCC in the univariable analysis as input variables, the multivariable analysis was performed to identify independent predictors. In order to select stable features, this selection process was repeated 1,000 times with different datasets resampled by

bootstrapping the training dataset, and only features that were chosen as independent predictors for HCC in > 95% of the 1,000 datasets were selected.

Hazard ratio of predictors for HCC

In the multivariable CoxPH regression, the hazard ratios (HRs) of the selected features were estimated with and without other cancers included as the competing risk. Subdistribution hazard with the competing risk was estimated using the methodology by Fine and Gray [15].

Training machine learning models in the training cohort

Random survival forest (RSF) algorithm was used for predicting the probability of and the time to HCC occurrence, with non-HCC cancers included as competing risks [16]. In addition, we tested whether an ensemble of RSF and multivariate extreme gradient boosting (XGBoost) algorithm could improve the accuracy of probability prediction. Hyperparameters were optimized using grid search by assessing out-of-bag errors for RSF and by 10-fold cross validation with area under receiver operating characteristics curve (AUC) as an evaluation metric for XGBoost. Optimal hyperparameters found were ntree = 100, mtry = 1, and nodesize = 6 for RSF, and max.depth = 4, eta = 0.1, min_child_weight = 1, gamma = 0, lambda = 0, and nrounds = 101 for XGBoost, with other parameters set to default. With the selected features and the optimal hyperparameters, the models were fit to the training dataset. In prediction of the probability of the 12-year development of HCC, the performances of RSF, XGBoost, and both were compared in terms of Brier skill score, AUC, and calibration plot, and the best model was chosen. Although the Brier score is a proper score function that measures the accuracy of probabilistic predictions, it does not tell us how accurate the predictions are compared with anything else, which may result in misleading results especially when a target outcome is rare as in this study. Thus, we used Brier skill score that assess the accuracy of predictions compared to a reference prediction of always predicting 'no HCC development': *Brier skill score=1-(Brier score/Reference Brier Score)*.

Validation in the test cohort

The performance of the final model was evaluated in the test cohort: AUC, Brier skill score, and calibration plot for the probability, and concordance index (*C*-index) for the time to HCC development. The sensitivity, specificity, and accuracy for 12-year HCC development were calculated at the optimal cutoff probability obtained from AUC analysis. Kaplan-Meier curve with log-rank test was used to compare the survival curves between three groups divided according to the predicted probability: low-risk (< 5%), intermediate-risk (5–20%), and high-risk (> 20%) groups.

Results

Study population

The final study population consisted of 331,694 screening examinees, with 232,186 (70.0%) in the training cohort and 99,508 (30.0%) in the test cohort (Fig. 1). The age ranged from 42 to 82 (mean, 54) years at the time of the examination, and the ratio of males to females was 5.8:4.2. There was no significant difference

in variables between the training and test cohorts (Table 1 and Supplementary Table 1). The median follow-up time was 11.2 years (up to 12.0 years). Of the total examinees, 0.8% (1,746 in the training cohort and 724 in the test cohort) were diagnosed with HCC, and 9.0% (19,034 and 8,151, respectively) were diagnosed with other cancers during the follow-up period.

Selected predictors and their hazard ratios for HCC

Stable predictors that showed significant association with the risk of HCC development in > 95 % of 1,000 different resampled datasets were age, sex, family history of chronic liver disease, ALT, gamma-glutamyl transpeptidase (GGT), total blood cholesterol level, and preexisting chronic liver disease, chronic hepatitis virus infection, HIV infection, diabetes mellitus, dyslipidemia, or schizophrenic/delusional disorders or mental disorders due to psychoactive substance use (Supplementary Table 3).

In the multivariable CoxPH regression, older age (HR, 1.616, with non-HCC cancers as competing risks), male sex (HR, 3.154), higher levels of ALT (HR, 1.060) or GGT (HR, 1.024), family history of chronic liver disease (HR, 2.677), and preexisting chronic liver disease (HR, 3.283), chronic hepatitis virus infection (HR, 2.103), HIV infection (HR, 4.020), and diabetes mellitus (HR, 1.583) were associated with increased risk, whereas a higher level of total cholesterol (HR, 0.902) and preexisting dyslipidemia (HR, 0.487) or schizophrenic/delusional disorders or mental disorders due to psychoactive substance use (HR, 0.616) were associated with decreased risk of HCC development ($p < 0.001$ for all variables). HRs were not significantly affected by whether or not the development of non-HCC cancers was considered competing risks (Table 2).

Table 2
Multivariable Cox proportional hazard regression for HCC with and without other cancers included as competing risks

	No competing risk			Competing risk included		
	HR	95% CI	*p*-value	HR	95% CI	*p*-value
Age	1.700	1.617–1.788	< 0.001	1.616	1.538–1.698	< 0.001
Male sex (vs. female)	3.225	2.821–3.690	< 0.001	3.154	2.754–3.610	< 0.001
Family history of chronic liver disease	2.581	2.158–3.087	< 0.001	2.677	2.224–3.222	< 0.001
ALT	1.060	1.052–1.068	< 0.001	1.060	1.047–1.070	< 0.001
GGT	1.023	1.021–1.027	< 0.001	1.024	1.020–1.028	< 0.001
Total cholesterol	0.898	0.885–0.911	< 0.001	0.902	0.886–0.914	< 0.001
Chronic liver disease	3.228	2.857–3.646	< 0.001	3.283	2.878–3.743	< 0.001
Chronic hepatitis virus infection	2.111	1.790–2.488	< 0.001	2.103	1.749–2.528	< 0.001
HIV infection	4.096	3.619–4.636	< 0.001	4.020	3.530–4.578	< 0.001
Diabetes mellitus	1.586	1.368–1.839	< 0.001	1.583	1.349–1.857	< 0.001
Dyslipidemia	0.509	0.391–0.662	< 0.001	0.487	0.362–0.651	< 0.001
Schizophrenic or delusional disorders, or mental disorders due to psychoactive substance use	0.623	0.532–0.730	< 0.001	0.616	0.522–0.727	< 0.001

HCC = hepatocellular carcinoma, HR = hazard ratio, CI = confidence interval, ALT = alanine aminotransferase, GGT = gamma-glutamyl transferase, HIV = human immunodeficiency

Probability prediction

In the training cohort, the XGBoost showed better performance than the RSF model in predicting the risk of HCC development. For discriminating whether HCC will develop or not, the AUCs (± standard deviation) of the XGBoost and RSF models were 0.879 (± 0.014) and 0.862 (± 0.021) in the cross validation and out-of-

bag validation, respectively. In terms of calibration, the Brier skill scores were 0.113 and 0.064, which can be interpreted as 11.3% and 6.4% improvement in Brier score compared to the baseline model, respectively. An ensemble of XGBoost and RSF showed comparable AUC (0.880 [± 0.013]) and Brier skill score (0.116) to XGBoost alone, but it was determined to show the best calibration curve (Fig. 2). Therefore, the ensemble model was chosen as our final model (Table 3).

Table 3
Performances of machine learning models in prediction of the probability of and the time to the development of HCC

Model	Evaluation metric	CV or OOB error in the training cohort	Validation in the test cohort
		Value (± SD)	AUC (95% CI)
Probability of developing HCC within 12 years			
Random survival forest	AUC	0.862 (± 0.021)	
	BSS	0.064	
Extreme gradient boosting	AUC	0.879 (± 0.014)	
	BSS	0.113	
Ensemble of two models	AUC	0.880 (± 0.013)	0.872 (0.858–0.887)
	BSS	0.116	0.080
Time to cancer occurrence if HCC develops			
Cox proportional hazard	C-index	0.837 (± 0.007)	0.848 (0.837–0.859)
Random survival forest	C-index	0.869 (± 0.011)	0.868 (0.861–0.875)

HCC = hepatocellular carcinoma, CV = cross validation, OOB = out-of-bag, SD = standard deviation, AUC = area under receiver operating characteristics curve, CI = confidence interval, BSS = Brier skill score, C-index = concordance index.

In the test cohort, our prediction model showed good calibration with a trend of mild underestimation with probabilities < 20% and mild overestimation with probabilities > 20% (Fig. 2). The AUC was 0.872 (95% CI, 0.858–0.887). The Brier skill score was 0.08. Using 1% as a cutoff probability, the sensitivity, specificity, and accuracy were 70.5% (95% CI, 70.1–70.9), 87% (95% CI, 86.8–87.2), and 86.9% (95% CI, 86.7–87.1), respectively. In the Kaplan-Meier curve with log-rank test, the curves for the three risk groups (i.e., low, < 5%; intermediate, 5–20%; and high, > 20%) were separated well ($p < 0.001$) in the test cohort (Fig. 3).

Time-to-event prediction

During the observation period, approximately 0.7% developed HCC (1,614/232,186 in the training cohort and 667/99,508 in the test cohort). The median time to cancer development was 302 weeks (5.8 years), ranging from 1 to 609 weeks (up to 11.7 years).In prediction of the time to HCC occurrence, the RSF model showed better discriminative ability than CoxPH in the test cohort with the c-indices of CoxPH and RSF being 0.848 (95% CI, 0.837–0.859) and 0.868 (95% CI, 0.861–0.875), respectively. Representative cases of individual predictions of the time-to-HCC by RSF are shown in Fig. 4.

Discussion

In this study, we developed a machine learning model to predict the risk of developing HCC within the following 12 years in an individual health screening examinee, based on the information available from the examination results and the history of medical service use. The model showed good calibration and discrimination in the test. We believe that one of the greatest strengths of this model is that it extracts information hidden in the big data that otherwise would have been discarded, and creates a new value, that is, providing an individual examinee with the estimated risk of developing a certain disease in the future. We hope that, after further development and validation, prediction models of this kind will be integrated in the national healthcare system and provide people with additional helpful information.

The main goal of machine learning lies on making the most accurate prediction possible, while traditional statistical analysis is mainly focused on generalization from sample statistics to population parameters. Thus, machine learning is often referred to as a ‘black box’; data goes in, predictions come out, but the processes between the input and the output are unclear, which is okay as long as its prediction is accurate [17]. However, we wanted to first identify a set of valid and stable input features, or predictors, and examine their associations with the outcome, instead of putting all data into an algorithm and asking it to try to make a good prediction, for the following reasons. First, even when the main goal is to make accurate predictions, it is still important to understand the relationship of predictors with an outcome, so that we can take appropriate action about the causes of the outcome. Second, complex algorithms can be so flexible that they pick up meaningless or noisy signals from input data to make good predictions only in a certain dataset but fail to generalize to other datasets with different noises. Therefore, by the rigorous feature selection process, we aimed to remove noisy signals, that is, non-significant, unstable input features; in our results many seemingly irrelevant underlying diseases such as hemorrhoid or chronic rhinitis were frequently selected as independent risk factors for HCC in resampled datasets (Supplementary Table 3).

Older age, male sex, chronic liver disease, heavy alcohol consumption, diabetes, and HIV infection are well-known risk factors for HCC.[18, 19] All of these risk factors were independent predictors in our cohort as well. Although drinking habit by questionnaire was not selected as a predictor in our model, we believe that the use of ALT and GGT, which were strong predictors in our model, is a more object approach for assessing the effect of alcohol consumption than the 5-point scale questionnaire used in our health screening examination, as a previous study showed [20].

In contrast to underlying diabetes, underlying dyslipidemia and higher total cholesterol were associated with the lower risk in our cohort. This opposite association between diabetes, dyslipidemia, and HCC is in line with the results of an epidemiologic study of HCC and metabolic risk factors in a nationwide Taiwan cohort [21]. This may be partly explained by that in this study dyslipidemia was diagnosed when both the diagnosis and the use of lipid-lowering drugs were confirmed (Supplementary Table 2), and current evidence suggests that statin use could contribute to a decline in HCC incidence [18, 22]. However, hypercholesterolemia without taking lipid-lowering drugs was also an independent risk factor [21]. More research is warranted on the effect and mechanism of dyslipidemia on the risk of HCC development and prognosis.

Family history of liver cancer is also a known risk factor for HCC [23, 24]. In our study, family history of chronic liver disease, not cancer, was a strong predictor. This is not a surprising result considering that chronic liver disease is one of the strongest risk factors for liver cancer. The presence or absence of family history of cancer was also asked in our health screening questionnaire, but it includes all types of cancer, which is probably the reason that it was not included as a significant risk factor.

Interpretation of the lower risk of HCC in patients with mental disorders due to psychoactive substance use or schizophrenic and delusional disorders is hampered by the fact that those diagnoses were considered sensitive personal information and grouped together under the unidentified code in our dataset. However, as mental disorders due to use of alcohol, which is most commonly used psychoactive substance, probably affected the outcome towards an increased risk, schizophrenic and delusional disorders were likely attributed to the decreased risk of HCC. Especially, schizophrenia has been reported by a meta-analysis study to be protective against HCC development [25]. Some investigators suggested the correlation between tumor suppressor genes and schizophrenia as possible explanation of its potential protective effect against cancer [26].

Our prediction model has limitations. Our model was developed and validated using a single ethnic (i.e., Asian) population from a single country. Thus, the generalizability of the model to other countries or ethnic groups is not guaranteed. However, we believe that our approach (i.e., machine learning predictor based on the claim and health screening data) can be applied to various cohorts similarly and used to produce their own, even multi-national, prediction models. In addition, as mentioned above, some diagnoses were masked and grouped together for the protection of sensitive personal information. We expect that more detailed information from the national health insurance database will be made available for research purposes in the future.

Conclusions

In conclusion, machine learning could be used to develop a prediction model for the 12-year risk of HCC in individual health screening examinees, based on the information retrieved from the examination results and healthcare claim data.

Abbreviations

HCC=hepatocellular carcinoma, NHIS-HEALS=National Health Insurance Service-National Health Screening, HIV=human immunodeficiency viruse, CoxPH=Cox proportional hazard, AST=aspartate transaminase, ALT=alanine transaminase, HR=hazard ratio, RSF=random survival forest, XGBoost=extreme gradient boosting, AUC=area under receiver operating characteristics curve, GGT=gamma-glutamyl transpeptidase

Declarations

Ethics approval and consent to participate

The Institutional Review Board of National Health Insurance Service Ilsan Hospital (NHIMC 2020-06-033) approved this Health Insurance Portability and Accountability Act-compliant retrospective study and waived the informed consent. All methods were performed in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to the provisions of the National Health Insurance Service (NHIS), but the codes and other materials are available from the corresponding author on reasonable request.

Competing interests

The authors declare no conflicts of interest.

Funding

This work was supported by National Health Insurance Service Ilsan Hospital grant (NHIMC 2020-06-033).

Authors' contributions

CA, JWC, HSL, SJR, and JHC conceived the study. HL obtained and extracted data. CA and HL cleaned data. CA performed machine learning. CA and HL performed statistical analysis. CA wrote the paper. CA, JWC, HSL, SJR, and JHC interpreted data. All authors have taken due care to ensure the integrity of this work, and all authors read and approved the final manuscript. JHC and HCO were in charge of the overall direction.

Acknowledgement

This research was conducted as part of the Ilsan Machine Intelligence with National health insurance big Data (I-MIND) project. This study used the National Health Insurance Service (NHIS) database (NHIS-2020-

2-146). The authors alone are responsible for the content of this article.

References

1. Liu Z, Jiang Y, Yuan H, Fang Q, Cai N, Suo C, et al. The trends in incidence of primary liver cancer caused by specific etiologies: results from the Global Burden of Disease Study 2016 and implications for liver cancer prevention. J Hepatol. 2018;70:674–83.
2. Mittal S, El-Serag HB. Epidemiology of Hepatocellular Carcinoma. J Clin Gastroenterol. 2013;47;S2–6.
3. Kim BH, Park J-W. Epidemiology of liver cancer in South Korea. Clin Mol Hepatology. 2018;24:1–9.
4. Kim S, Kim M-S, You S-H, Jung S-Y. Conducting and Reporting a Clinical Research Using Korean Healthcare Claims Database. Korean J Fam Medicine. 2020;41:146–52.
5. Hsu Y-C, Yip TC-F, Ho HJ, Wong VW-S, Huang Y-T, El-Serag HB, et al. Development of a scoring system to predict hepatocellular carcinoma in Asians on antivirals for chronic hepatitis B. J Hepatol. 2018;69:278–85.
6. El-Serag HB, Kanwal F, Davila JA, Kramer J, Richardson P. A New Laboratory-Based Algorithm to Predict Development of Hepatocellular Carcinoma in Patients With Hepatitis C and Cirrhosis. Gastroenterology. 2014;146:1249-1255.e1.
7. Kuang S-Y, Jackson PE, Wang J-B, Lu P-X, Muñoz A, Qian G-S, et al. Specific mutations of hepatitis B virus in plasma predict liver cancer development. P Natl Acad Sci Usa. 2004;101:3575–80.
8. Yang H-I, Yuen M-F, Chan HL-Y, Han K-H, Chen P-J, Kim D-Y, et al. Risk estimation for hepatocellular carcinoma in chronic hepatitis B (REACH-B): development and validation of a predictive score. Lancet Oncol. 2011;12:568–74.
9. Ripoll C, Groszmann RJ, Garcia-Tsao G, Bosch J, Grace N, Burroughs A, et al. Hepatic venous pressure gradient predicts development of hepatocellular carcinoma independently of severity of cirrhosis. J Hepatol. 2009;50:923–8.
10. Wong VW, Yu J, Cheng AS, Wong GL, Chan H, Chu ES, et al. High serum interleukin-6 level predicts future hepatocellular carcinoma development in patients with chronic hepatitis B. Int J Cancer. 2009;124:2766–70.
11. Seong SC, Kim Y-Y, Park SK, Khang YH, Kim HC, Park JH, et al. Cohort profile: the National Health Insurance Service-National Health Screening Cohort (NHIS-HEALS) in Korea. Bmj Open. 2017;7:e016640.
12. Ahn E. Introducing big data analysis using data from National Health Insurance Service. Korean J Anesthesiol. 2020;73:205–11.
13. Choi E-K. Cardiovascular Research Using the Korean National Health Information Database. Korean Circ J. 2019;50:754.
14. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019;112:103375.

15. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. J Am Stat Assoc. 2012;94:496–509.
16. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. Biostatistics. 2014;15:757–73.
17. Medicine TLR. Opening the black box of machine learning. Lancet Respir Medicine. 2018;6:801.
18. McGlynn KA, Petrick JL, London WT. Global Epidemiology of Hepatocellular Carcinoma An Emphasis on Demographic and Regional Variability. Clin Liver Dis. 2015;19:223–38.
19. Shiels MS, Cole SR, Kirk GD, Poole C. A Meta-Analysis of the Incidence of Non-AIDS Cancers in HIV-Infected Individuals. Jaids J Acquir Immune Defic Syndromes. 2009;52:611–22.
20. Niemelä O. Biomarker-Based Approaches for Assessing Alcohol Use Disorders. Int J Environ Res Pu. 2016;13:166.
21. Chiang C, Lee L, Hung S, Lin W, Hung H, Yang W, et al. Opposite association between diabetes, dyslipidemia, and hepatocellular carcinoma mortality in the middle-aged and elderly. Hepatology. 2014;59:2207–15.
22. German MN, Lutz MK, Pickhardt PJ, Bruce RJ, Said A. Statin Use is Protective Against Hepatocellular Carcinoma in Patients With Nonalcoholic Fatty Liver Disease. J Clin Gastroenterol. 2020;54:733–40.
23. Yu M-W, Chang H-C, Liaw Y-F, Lin S-M, Lee S-D, Liu C-J, et al. Familial Risk of Hepatocellular Carcinoma Among Chronic Hepatitis B Carriers and Their Relatives. Jnci J National Cancer Inst. 2000;92:1159–64.
24. Hassan MM, Spitz MR, Thomas MB, Curley SA, Patt YZ, Vauthey J-N, et al. The association of family history of liver cancer with hepatocellular carcinoma: A case-control study in the United States. J Hepatol. 2009;50:334–41.
25. Xu D, Chen G, Kong L, Zhang W, Hu L, Chen C, et al. Lower risk of liver cancer in patients with schizophrenia: a systematic review and meta-analysis of cohort studies. Oncotarget. 2017;8:102328–35.
26. Zhuo C, Wang D, Zhou C, Chen C, Li J, Tian H, et al. Double-Edged Sword of Tumour Suppressor Genes in Schizophrenia. Front Mol Neurosci. 2019;12:1.

Figures

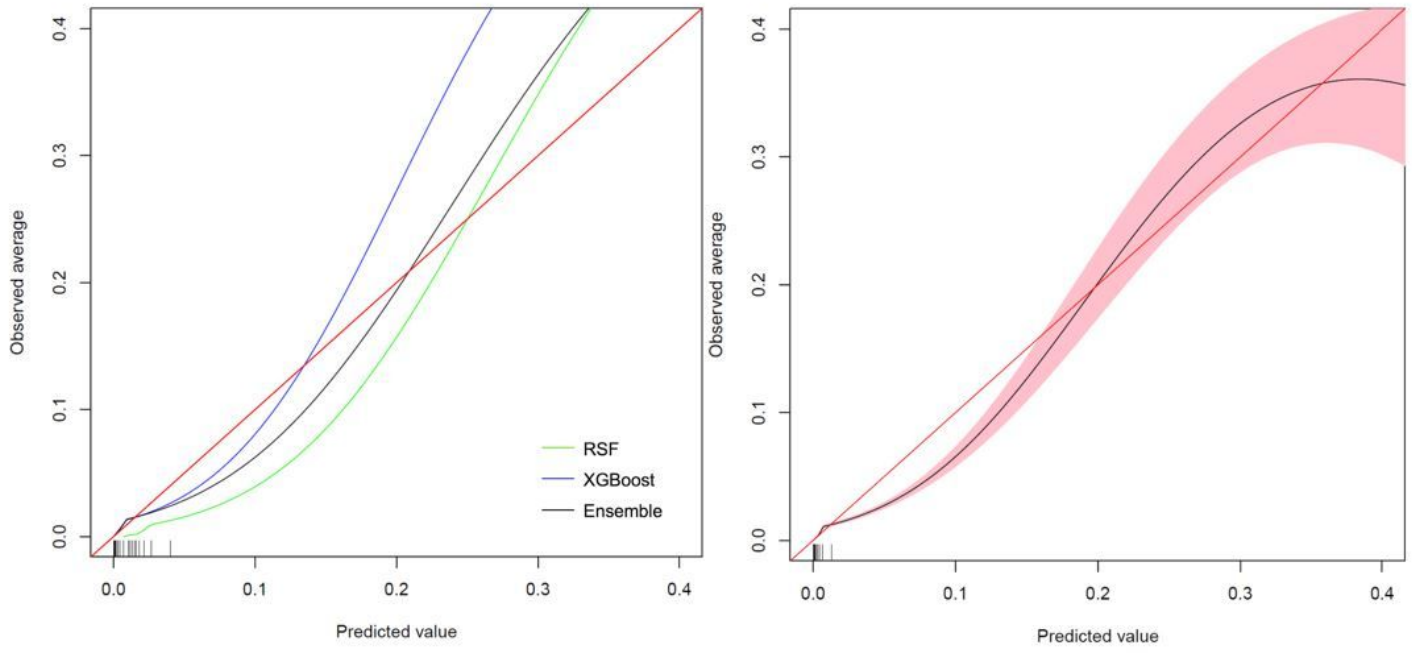

Figure 2

Calibration curves. The left panel shows calibration curves in the training cohort (Red, ideal line; Green, XGBoost; Blue, RandomForest; Black: Ensemble model). The right panel shows the calibration curve of the final model in the test cohort, with the area of pinkish shades indicating 95% confidence interval.

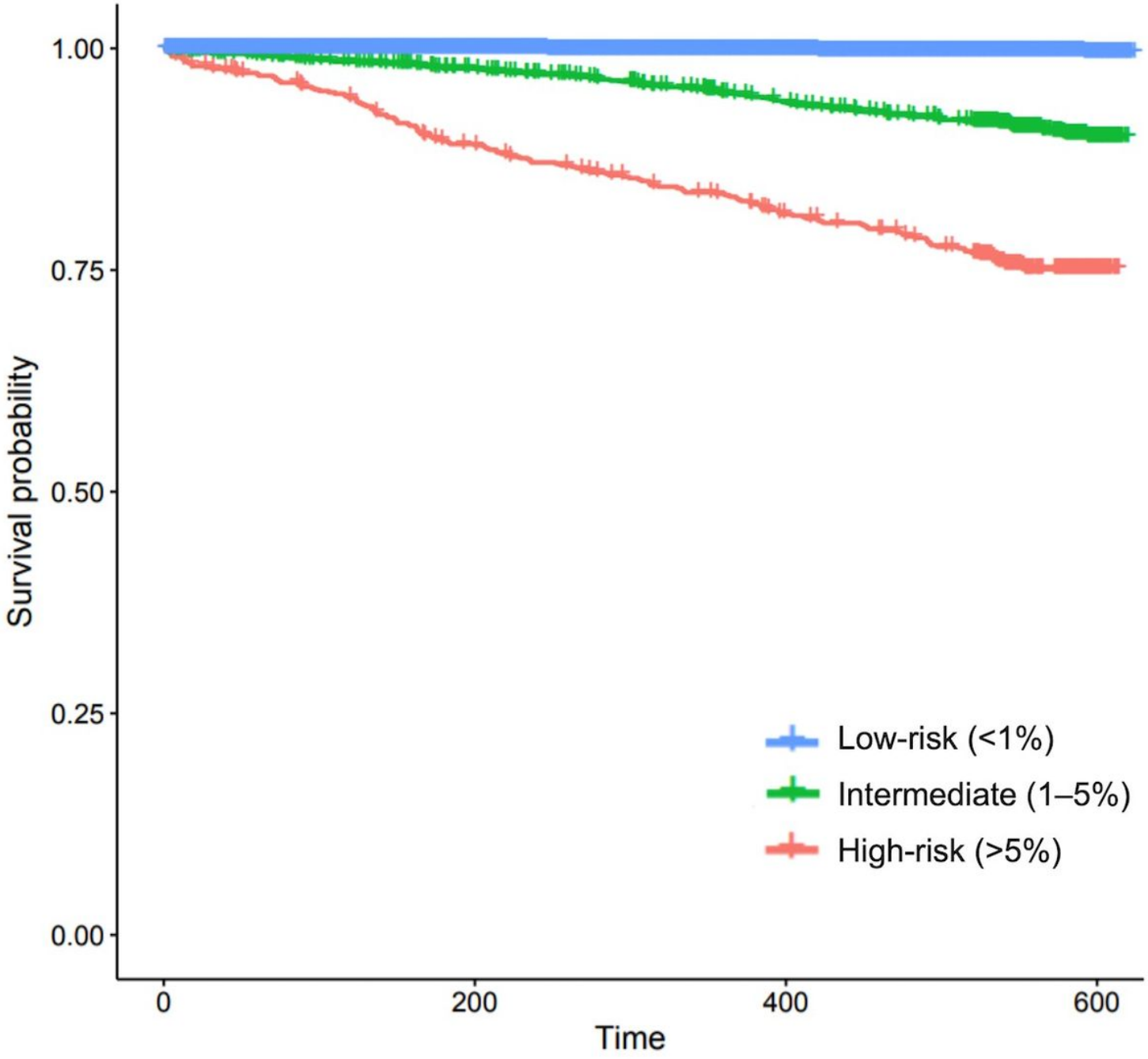

Figure 3

Survival curves of the three risk groups. Low-risk, probability of <5%; intermediate-risk, 5–20%; and high-risk, >20%.

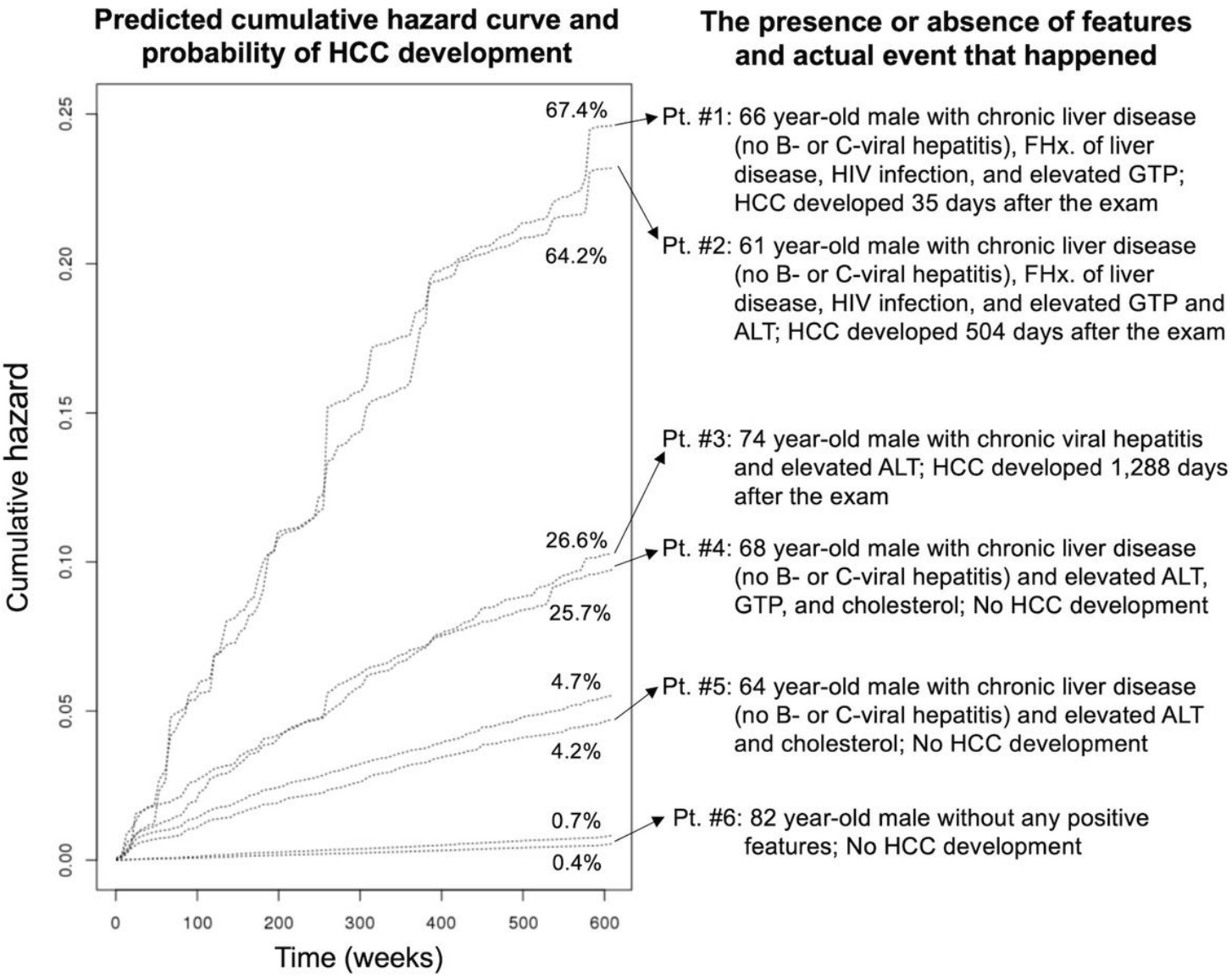

Figure 4

Representative cases with different predicted risks. In the left panel, cumulative hazard curves of eight screening examinees are shown with the predicted risks of developing hepatocellular carcinoma. In the right panel, the risk factors they had and the actual events that happened to them are summarized.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfile.docx