

RESEARCH

Open Access



Liver cancer risk stratification using deep learning on nationwide longitudinal health screening data: a retrospective cohort study

Yewon Choi^{1,2} , Sungmin Cho¹ , Changdai Gu^{1,4} , Chungho Kim⁵ , Bomi Park^{5*} and Hwiyoung Kim^{1,3,4*}

Abstract

Background Current liver cancer screening in Korea focuses on viral hepatitis or cirrhosis, despite rising risks from metabolic and alcohol-related liver disease. We aimed to develop a deep learning model that leverages routinely collected national screening and claims data to predict liver cancer risk without requiring additional diagnostic tests.

Methods We conducted a retrospective cohort study of 3,962,209 adults aged 50–69 years who participated in the Korean National Health Screening program between 2010 and 2015, with follow-up until December 31, 2021. A total of 12,401 liver cancer cases were identified. Using data from three biennial screenings over 6 years, we developed a one-dimensional convolutional neural network model to predict 5-year liver cancer risk. The cohort was randomly divided at the patient level into training (80%) and testing (20%) sets. Predictors included demographic, clinical, behavioral, anthropometric, and laboratory features. Model performance was compared with logistic regression, extreme gradient boosting, multilayer perceptron, and current national surveillance criteria, assessed by the area under the receiver operating characteristic curve (AUROC), sensitivity, and specificity. Interpretability was examined using SHapley values and Cox regression, and sensitivity analyses evaluated the impact of screening timing.

Results Our model achieved an AUROC of 0.810 (95% CI, 0.802–0.818) and an AUPRC of 0.029 (95% CI, 0.026–0.034), with a sensitivity of 0.736 (95% CI, 0.720–0.753), clearly outperforming the current national criteria which showed an AUROC of 0.552 (95% CI, 0.546–0.558), an AUPRC of 0.007 (95% CI, 0.006–0.008), and a sensitivity of only 0.112 (95% CI, 0.100–0.125). The top-risk quintile accounted for 65% of incident liver cancer cases and had a 27-fold higher hazard compared to the lowest-risk group. Major predictors included age, viral hepatitis, family history of liver cancer, cholesterol levels, alcohol consumption, and metabolic factors. Sensitivity analyses demonstrated that incorporating all three screening time points yielded the highest overall performance.

Conclusions Applying a deep learning model to routinely collected national screening data improved liver cancer risk stratification and enabled early identification of high-risk individuals, including those without prior liver disease. This approach supports scalable, policy-relevant screening strategies within existing public health infrastructure.

Trial registration Not applicable.

*Correspondence:

Bomi Park
bpark@cau.ac.kr
Hwiyoung Kim
hykim82@yuhs.ac

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Keywords HCC, Machine learning, Liver neoplasms, Lifestyle factor, CNN, Prediction

Introduction

Primary liver cancer is the sixth most common cancer and the third leading cause of cancer-related mortality worldwide [1, 2], with hepatocellular carcinoma (HCC) accounting for more than 85% of cases [3]. In South Korea, liver cancer is the second leading cause of cancer-related death [4]. The 2022 KLCA–NCC Korea Practice Guidelines (KNKPC) recommend biannual HCC surveillance with ultrasonography and serum alpha-fetoprotein (AFP) testing for high-risk individuals, including those with hepatitis B virus (HBV), hepatitis C virus (HCV), or cirrhosis [5]. However, these guidelines may not fully reflect current epidemiologic trends. Widespread HBV vaccination and effective HCV treatment have reduced virally associated HCC, and nearly half of new cases now occur in individuals without viral hepatitis [6, 7]. Moreover, up to 39% of HCC cases are diagnosed in noncirrhotic individuals [8]. Conversely, metabolic dysfunction–associated steatotic liver disease (MASLD) and alcohol-associated liver disease (ALD) are emerging as leading causes [9], closely linked to modifiable lifestyle factors such as obesity, diabetes, alcohol consumption, and smoking [10–12]. These changes underscore the need to expand liver cancer risk stratification beyond conventional high-risk groups [13, 14].

Previous liver cancer risk prediction models have largely targeted individuals with chronic liver disease or viral hepatitis. Although models such as aMAP (age, gender, bilirubin, albumin, platelet count) and GALAD (gender, age, AFP, AFP-L3, des-gamma-carboxy prothrombin) have shown strong predictive performance in high-risk populations [12, 13], they depend on biomarkers (e.g., AFP, albumin) that are not routinely measured in population-level screenings [15]. Furthermore, most previous studies have used static variables and conventional machine learning methods, such as Cox regression or XGBoost [16–18], which often do not account for the temporal progression of disease.

The risk of liver cancer evolves gradually with age, comorbidities, and lifestyle factors [19, 20], suggesting that temporal deep learning (DL) models may be appropriate for prediction. Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformer-based models have shown strong performance in dynamic cancer prediction using longitudinal data [21]. Among these, CNNs are particularly efficient at detecting local patterns in structured, fixed-interval health screening data, such as those collected through Korea's national health examination system, whereas RNNs and Transformer models are better suited for irregular or long-interval data. Therefore, the architecture should

be guided by the structure and frequency of the input variables.

To address these considerations, we developed a 5-year liver cancer risk prediction model based on CNNs, using routinely collected demographic, clinical, and behavioral data from the National Health Insurance Service (NHIS). The performance of the model was compared against logistic regression (LR), XGBoost, multilayer perceptron (MLP), and the KNKPC surveillance criteria. To enhance interpretability, we applied SHapley Additive Explanations (SHAP) to analyze feature contributions and used Cox regression to estimate hazard ratios and assess statistical significance, offering a scalable and interpretable framework for population-level risk prediction [22].

Materials and methods

Data source

NHIS in the Republic of Korea (hereafter, “Korea”) is a nonprofit organization administered by the Korean government that provides mandatory health insurance coverage to the entire Korean population [23]. Individuals aged 20 years and older are eligible for a comprehensive health screening conducted biennially, which includes a self-administered questionnaire assessing lifestyle behaviors, laboratory tests, and anthropometric measurements [24]. The NHIS database contains extensive sociodemographic information, health examination results, inpatient and outpatient medical records, and prescription data. It has been extensively used in epidemiological studies [25–27], and its validity has been established in previous research [23, 28, 29]. The study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis with Artificial Intelligence guidelines [30].

Study population and design

The initial cohort comprised individuals aged 50–69 years in 2012 who had undergone national health screening between 2010 and 2011 ($n = 7,397,609$). Participants were eligible if they completed a health screening at least once every 2 years from 2010 to 2015. Of the initial cohort, 4,861,632 individuals met this criterion. A total of 899,423 individuals were excluded for the following reasons: a prior cancer diagnosis before the index date (January 1, 2016) ($n = 149,619$), death before the index date ($n = 7,076$), diagnosis of another cancer during the follow-up period ($n = 212,077$), or missing data on key variables ($n = 530,651$). We compared basic demographic and clinical characteristics between excluded and included individuals to assess potential selection bias (Table S1). Patients diagnosed with any new cancer other than liver

cancer during follow-up were excluded to reduce confounding, as another malignancy may influence medical surveillance and subsequent liver cancer risk, consistent with large population-based cancer cohort studies [31, 32]. After applying these exclusion criteria, the final analytic cohort consisted of 3,962,209 participants (Fig. S1). Participants were followed up from January 1, 2016, to December 31, 2021, and were classified according to whether they underwent liver cancer diagnosis during follow-up: those diagnosed with liver cancer ($n = 12,401$) and those without a diagnosis ($n = 3,949,808$). For model development and validation, the cohort was stratified by age, sex, and liver cancer diagnosis, and then randomly divided into training (80%) and test (20%) sets. The training set included 9,921 participants diagnosed with liver cancer and 3,159,846 without a liver cancer diagnosis, while the test set comprised 2,480 and 789,962 participants, respectively.

Predictor and outcome definitions

The primary outcome of this study was the incidence of liver cancer during the follow-up period, as identified in the NHIS database. Liver cancer was defined using the International Classification of Diseases, 10th Revision (ICD-10) codes C22.0 or C22.9 for men and C22.0 for women, with these codes recorded as the primary diagnosis [33]. This operational definition was adopted based on a previous systematic review using the NHIS database, which compared various ICD-10 code combinations with the Korea Central Cancer Registry (KCCR) and identified this sex-specific definition as the most consistent with the registry-based incidence rates [33].

Predictor variables obtained from the NHIS dataset prior to the index date included demographic factors (age, sex, household income); health-related behaviors (smoking status, alcohol consumption, physical activity); clinical measurements (body mass index [BMI], hemoglobin, fasting serum glucose, total cholesterol [TC], waist circumference, high-density lipoprotein [HDL] cholesterol); self-reported medical history (stroke, heart disease, hypertension, diabetes, dyslipidemia); and ICD-10 code-based medical history, including HBV, HCV, ALD, MASLD, cryptogenic liver disease, and cirrhosis, defined according to previously validated criteria (Table S2) [11]. Family history of liver cancer and non-liver cancers was also recorded. Detailed definitions and measurement methods for each predictor variable are provided in Table S3. For each participant, data were extracted from three time points corresponding to the most recent health examination within each interval: 2010–2011 (t1), 2012–2013 (t2), and 2014–2015 (t3). Variables collected at each health examination were treated as temporal variables, whereas baseline characteristics such as age and sex were classified as static variables.

Model development and validation

We developed and validated four models to predict liver cancer incidence: LR, XGBoost, MLP, and a fusion model (Fig. 1). Additionally, we evaluated a separate rule-based model derived from the KNKPC surveillance criteria using the same test set. The fusion model processed temporal features with one-dimensional (1D) convolutional layers and static features with an MLP, then integrated the outputs and passed them to a final classifier. To address class imbalance, we used weighted cross-entropy loss for LR, applied the `scale_pos_weight` parameter in XGBoost [34], and implemented focal loss functions in both the MLP and fusion models [35]. Hyperparameters were optimized through greedy search with five-fold cross-validation within the training set (Table S4). Further details regarding data preprocessing, model architecture, and training procedures are provided in Methods S1 and S2.

Statistical analysis

Continuous variables were compared using Welch's *t*-test [36], and categorical variables were compared using the chi-square test [37]. Model performance was assessed using the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve, sensitivity, and specificity. Differences in AUROC values between models were assessed using the DeLong method [38]. The optimal cutoff values for binary classification were determined using the Youden index [39]. To examine the effect of observation timing on model performance, we conducted a sensitivity analysis across six different observation windows: t1, t2, t3, t1 + t2, t2 + t3, and t1 + t2 + t3.

Kaplan–Meier curves were plotted according to risk quintiles [40], with hazard ratios (HRs) calculated using the lowest quintile (Q1) as the reference. Associations between predictors and liver cancer risk were further analyzed using Cox proportional hazards regression with three model specifications: (1) unadjusted models including only the exposure variable; (2) multicollinearity-adjusted models excluding one variable from each clinically correlated pair; and (3) fully adjusted models including all predefined covariates, regardless of potential multicollinearity [41]. Clinically interchangeable variable pairs were defined a priori (Table S5). For temporal variables measured across the three screening periods, mean values were used as covariates. Using the scaled Schoenfeld residuals, the proportional hazards assumption was tested graphically, and no violation of proportionality was detected. In addition, to account for the competing risk of all-cause mortality, subdistribution hazard ratios were estimated using the Fine-Gray subdistribution hazards model.

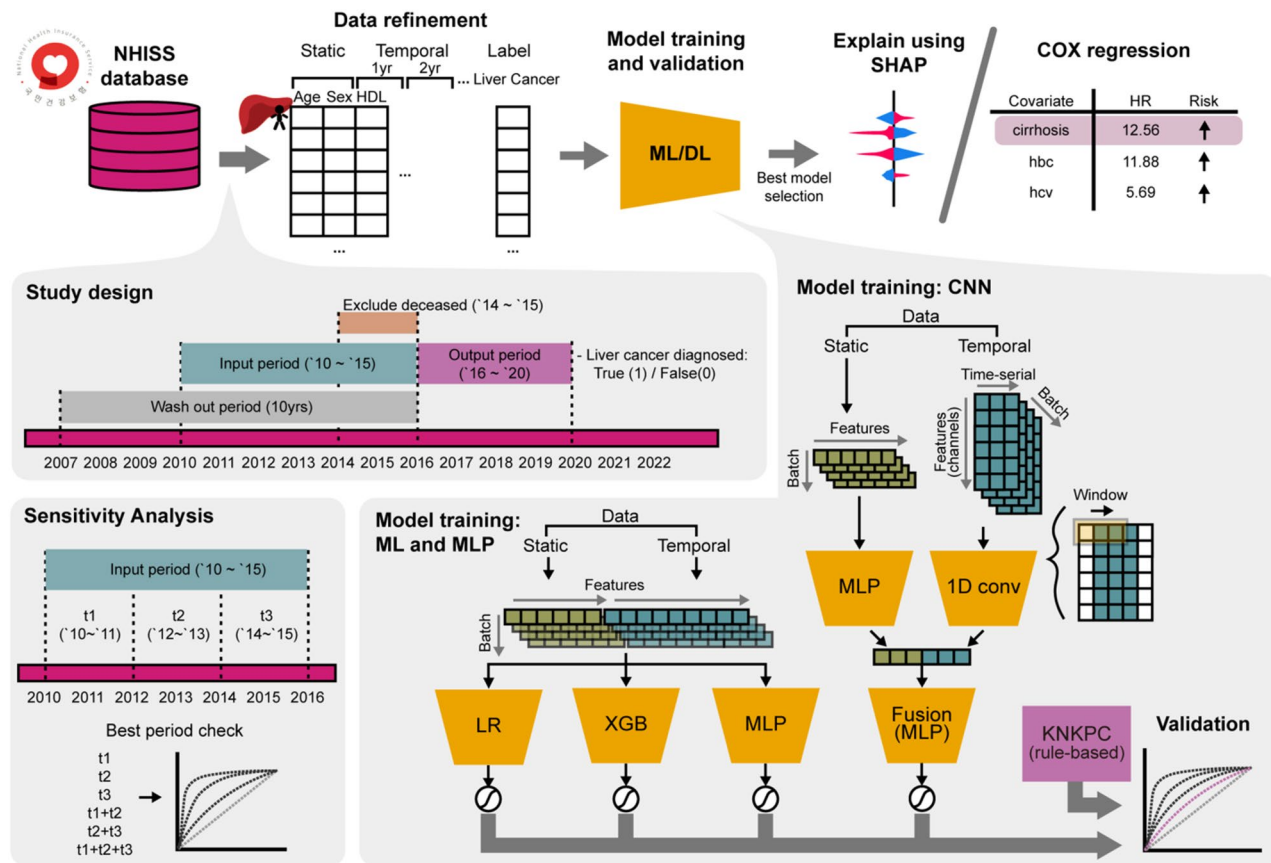


Fig. 1 Workflow of the Study. A liver cancer risk prediction model was developed using Korean national health screening data collected from 2010 to 2015, incorporating a 10-year washout period to exclude individuals with prior liver cancer. Incident cases of liver cancer were tracked over a 5-year follow-up period from 2016 to 2021. The performance of logistic regression, XGBoost, MLP, a CNN-based fusion model, and the KNKPC criteria was evaluated. The fusion model integrated temporal features using 1D CNNs and static variables through the MLP. Model predictions were interpreted using SHAP and Cox regressions. Sensitivity analyses were conducted across different combinations of screening intervals. Abbreviations: LR, linear regression; MLP, multilayer perceptron; CNN, Convolutional neural networks; KNKPC, 2022 KLCA–NCC Korea Practice Guidelines

Statistical significance was defined as a two-tailed $P \leq 0.05$. All analyses and model development were conducted in Python version 3.8, using key packages including PyTorch (v1.9.1) and Scikit-learn (v0.24.2), and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Results

Characteristics of the study population

Table 1 presents the baseline characteristics of 3,962,209 participants at the third health screening, which served as the final examination before the index date. During follow-up, 12,401 individuals (0.3%) were newly diagnosed with liver cancer. The mean age of all participants was 62.1 years (SD 5.5), and 45.0% were male. In the liver cancer group, the mean age was 63.6 years, and 75.4% were male. Compared to cancer-free participants, those with liver cancer were more likely to be in the lower income quartiles (25.5% vs. 28.5% in the highest quartile; 16.3% vs. 16.9% in the lowest quartile), report heavy alcohol consumption (12.2% vs. 5.4%), and be current smokers

(25.4% vs. 12.6%). Participants with liver cancer also had higher mean values for BMI, hemoglobin, fasting glucose, and waist circumference, and lower mean levels of total and HDL cholesterol. Histories of hypertension, diabetes, cirrhosis, hepatitis, and other liver diseases were more prevalent in the liver cancer group, as was a family history of liver cancer. Detailed statistics from all three health examinations are provided in Table S6. The distributions of temporal variables were largely consistent across the three examinations.

Model performance

As shown in Fig. 2A; Table 2, the fusion model achieved the highest discriminatory performance in the test set, with an AUROC of 0.810 (95% CI, 0.802–0.818) and an AUPRC of 0.029 (95% CI, 0.026–0.034). This performance exceeded that of LR (AUROC 0.793; 95% CI, 0.784–0.803; AUPRC 0.022; 95% CI, 0.019–0.024), XGBoost (AUROC 0.797; 95% CI, 0.788–0.806; AUPRC 0.025; 95% CI, 0.022–0.029), MLP (AUROC 0.803; 95% CI, 0.793–0.811;

Table 1 Baseline characteristics of the study population at the third health screening

	Total	Non-LiC	LiC	P-value
Number of participants (%)	3,962,209	3,949,808 (99.7)	12,401 (0.3)	-
Age, years, mean (SD)	62.1 (5.5)	62.1 (5.5)	63.6 (5.5)	< 0.001
Sex, N (%)				< 0.001
Men	1,784,333 (45.0)	1,774,978 (44.9)	9,355 (75.4)	-
Women	2,177,876 (55.0)	2,174,830 (55.1)	3,046 (24.6)	-
Household income^a, N (%) 3rd health examination period				< 0.001
1st (highest)	1,127,906 (28.5)	1,124,739 (28.5)	3,167 (25.5)	-
2nd	1,225,479 (30.9)	1,221,375 (30.9)	4,104 (33.1)	-
3rd	940,311 (23.7)	937,201 (23.7)	3,110 (25.1)	-
4th (lowest)	668,513 (16.9)	666,493 (16.9)	2,020 (16.3)	-
3rd health examination period Alcohol consumption status, N (%)				< 0.001
Non-drinking	2,546,510 (64.3)	2,539,995 (64.3)	6,515 (52.5)	-
Light drinking	586,130 (14.8)	584,432 (14.8)	1,698 (13.7)	-
Light-to-moderate drinking	295,306 (7.5)	294,130 (7.4)	1,176 (9.5)	-
Moderate-to-heavy drinking	320,004 (8.1)	318,505 (8.1)	1,499 (12.1)	-
Heavy drinking	214,259 (5.4)	212,746 (5.4)	1,513 (12.2)	-
3rd health examination period Smoking, N (%)				< 0.001
Never smoker	2,715,110 (68.5)	2,709,346 (68.6)	5,764 (46.5)	-
Past smoker	744,699 (18.8)	741,206 (18.8)	3,493 (28.2)	-
Current smoker	502,400 (12.7)	499,256 (12.6)	3,144 (25.4)	-
3rd health examination period Physical activity, times per week, N (%)				< 0.001
0	1,826,577 (46.1)	1,820,495 (46.1)	6,082 (49.0)	-
1–2	848,818 (21.4)	846,362 (21.4)	2,456 (19.8)	-
3–4	893,040 (22.5)	890,410 (22.5)	2,630 (21.2)	-
≥ 5	393,774 (9.9)	392,541 (9.9)	1,233 (9.9)	-
BMI^b, kg/m², mean (SD)	24.2 (3.0)	24.2 (3.0)	24.6 (3.2)	< 0.001
Hemoglobin, g/dL, mean (SD)	13.9 (1.4)	13.9 (1.4)	14.3 (1.5)	< 0.001
Fasting serum glucose, mg/dL, mean (SD)	102.9 (24.8)	102.9 (24.8)	112.1 (34.1)	< 0.001
Total cholesterol, mg/dL mean (SD)	198.0 (40.2)	198.1 (40.2)	175.3 (36.7)	< 0.001
HDL-cholesterol, mg/dL, mean (SD)	54.7 (18.8)	54.7 (18.8)	52.2 (15.2)	< 0.001
Waist circumference, cm, mean (SD)	81.9 (8.5)	81.9 (8.5)	85.2 (8.7)	< 0.001
History of stroke, N (%)				0.009
No	3,868,735 (97.6)	3,856,671 (97.6)	12,064 (97.3)	-
Yes	93,474 (2.4)	93,137 (2.4)	337 (2.7)	-
History of heart disease N (%)				0.018
No	3,742,364 (94.5)	3,730,712 (94.5)	11,652 (94.0)	-
Yes	219,845 (5.5)	219,096 (5.5)	749 (6.0)	-
History of hypertension, N (%)				< 0.001
No	2,513,371 (63.4)	2,506,601 (63.5)	6,770 (54.6)	-
Yes	1,448,838 (36.6)	1,443,207 (36.5)	5,631 (45.4)	-
History of diabetes, N (%)				< 0.001
No	3,428,104 (86.5)	3,418,949 (86.6)	9,155 (73.8)	-
Yes	534,105 (13.5)	530,859 (13.4)	3,246 (26.2)	-
History of dyslipidemia, N (%)				< 0.001
No	3,345,330 (84.4)	3,334,148 (84.4)	11,182 (90.2)	-
Yes	616,879 (15.6)	615,660 (15.6)	1,219 (9.8)	-
History of HCV, N (%)				< 0.001
No	3,955,229 (99.8)	3,943,050 (99.8)	12,179 (98.2)	-
Yes	6,980 (0.2)	6,758 (0.2)	222 (1.8)	-
History of HBV, N (%)				< 0.001
No	3,932,194 (99.2)	3,920,896 (99.3)	11,298 (91.1)	-
Yes	30,015 (0.8)	28,912 (0.7)	1,103 (8.9)	-
History of ALD, N (%)				< 0.001

Table 1 (continued)

	Total	Non-LiC	LiC	P-value
No	3,945,962 (99.6)	3,933,910 (99.6)	12,052 (97.2)	-
Yes	16,247 (0.4)	15,898 (0.4)	349 (2.8)	-
History of MASLD, N (%)				< 0.001
No	3,585,948 (90.5)	3,575,085 (90.5)	10,863 (87.6)	-
Yes	376,261 (9.5)	374,723 (9.5)	1,538 (12.4)	-
History of cryptogenic liver disease, N (%)				< 0.001
No	3,961,156 (100.0)	3,948,778 (100.0)	12,378 (99.8)	-
Yes	1,053 (0.0)	1,030 (0.0)	23 (0.2)	-
History of cirrhosis, N (%)				< 0.001
No	3,960,807 (100.0)	3,948,552 (100.0)	12,255 (98.8)	-
Yes	1,402 (0.0)	1,256 (0.0)	146 (1.2)	-
Family history of liver cancer, N (%)				< 0.001
No	3,872,313 (97.7)	3,860,604 (97.7)	11,709 (94.4)	-
Yes	89,896 (2.3)	89,204 (2.3)	692 (5.6)	-
Family history of non-liver cancer, N (%)				0.002
No	3,472,355 (87.6)	3,461,374 (87.6)	10,981 (88.5)	-
Yes	489,854 (12.4)	488,434 (12.4)	1,420 (11.5)	-

Abbreviations: SD, standard deviation; BMI, body mass index; HDL, high-density lipoprotein; HCV, hepatitis C virus; HBV, hepatitis B virus; ALD, alcohol-related liver disease; MASLD, metabolic dysfunction-associated steatotic liver disease

SI Conversion Factors: To convert fasting serum glucose to millimoles per liter, multiply by 0.0555; HDL cholesterol to millimoles per liter, multiply by 0.0259; TC to millimoles per liter, multiply by 0.0259; hemoglobin to grams per liter, multiply by 10

^aProxy for socioeconomic status based on the National Health Insurance Service premium

^bBMI is calculated as weight in kilograms divided by height in meters squared

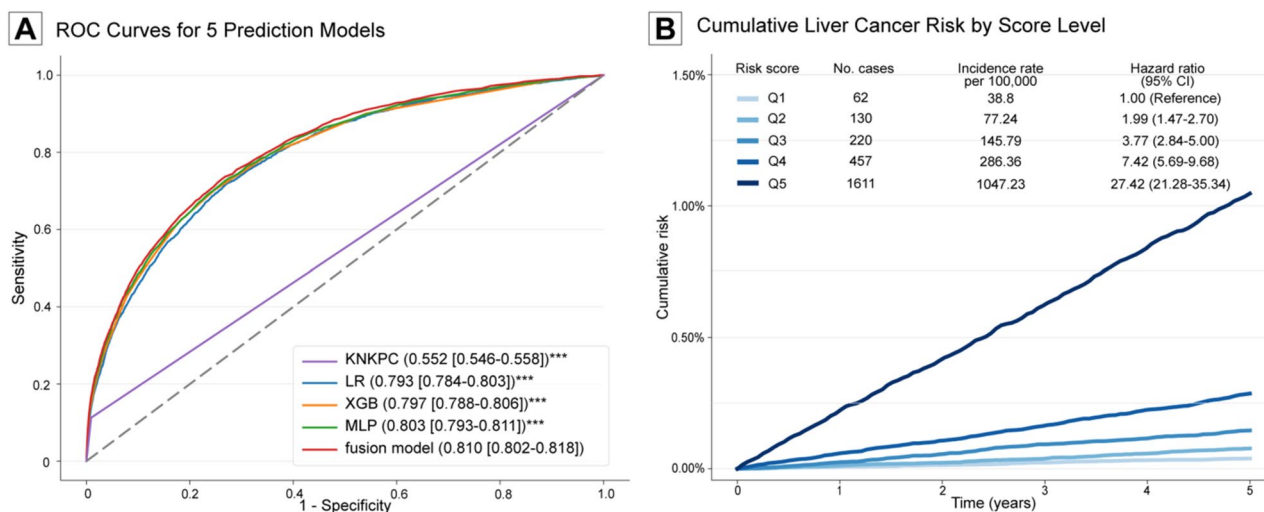


Fig. 2 Model Performance and Risk Stratification. **(A)** Five models were evaluated for 5-year liver cancer prediction: LR, XGBoost, MLP, fusion model, and KNKPC criteria. All pairwise comparisons with the fusion model were statistically significant ($P < 0.001$, DeLong test). Bars represent the AUROC with 95% CIs; *** indicates $P < 0.0001$. **(B)** The 5-year cumulative risk of liver cancer is shown across predicted risk quintiles from the fusion model. Abbreviations: LR, linear regression; MLP, multilayer perceptron; KNKPC, 2022 KLCG–NCC Korea Practice Guidelines; AUROC, area under the receiver operating characteristic curve; CI, confidence interval

AUPRC 0.028; 95% CI, 0.024–0.032) and KNKPC (AUROC 0.552; 95% CI, 0.545–0.558; AUPRC 0.007; 95% CI, 0.006–0.008). All differences were statistically significant ($P < 0.001$, DeLong test). The fusion model attained a sensitivity of 0.736 (95% CI, 0.720–0.753), approximately 6.5 times higher than that of KNKPC (0.112; 95% CI, 0.100–0.125). However, its specificity was lower at 0.732

(95% CI, 0.731–0.733), compared with KNKPC at 0.991 (95% CI, 0.991–0.991).

Figure 2B illustrates cumulative incidence curves stratified by risk quintiles from the fusion model. Of the 2,480 incident liver cancer cases, 1,611 (65.0%) occurred in Q5, corresponding to an incidence rate of 1,047.2 per 100,000 individuals and a hazard ratio of 27.42 (95% CI,

Table 2 Selected risk factors for lic identified by SHAP and multivariable Cox regression

Variables	Events	Person-years	Unadjusted aHR (95% CI)	Collinearity-based adjusted aHR (95% CI)
Age, y				
< 60	3,483	7,600,536	1.00 (Reference)	1.00 (Reference)
≥ 60	8,918	12,078,649	1.61 (1.55–1.68)	1.61 (1.54–1.67)
Sex				
Men	9,355	8,828,336	1.00 (Reference)	1.00 (Reference)
Women	3,046	10,850,849	0.27 (0.25–2.28)	0.29 (0.28–0.31)
Alcohol consumption				
2nd period				
Non-drinker	6,288	12,457,627	1.00 (Reference)	1.00 (Reference)
Light drinker	1,715	2,938,563	1.16 (1.10–1.22)	0.85 (0.80–0.90)
Light-to-moderate	1,179	1,501,231	1.56 (1.46–1.66)	0.89 (0.83–0.95)
Moderate-to-heavy	1,618	1,651,153	1.94 (1.84–2.05)	1.01 (0.96–1.08)
Heavy	1,601	1,130,612	2.81 (2.66–2.97)	1.33 (1.25–1.41)
Total cholesterol				
1st period				
< 180	6,057	55,282,219	1.00 (Reference)	1.00 (Reference)
180–240	5,533	11,197,335	0.45 (0.44–0.47)	0.54 (0.52–0.56)
≥ 240	811	2,953,631	0.25 (0.23–0.27)	0.33 (0.31–0.36)
2nd period				
< 180	6,559	5,829,938	1.00 (Reference)	1.00 (Reference)
180–240	5,133	10,981,950	0.42 (0.40–0.43)	0.51 (0.49–0.53)
≥ 240	709	2,867,298	0.22 (0.20–0.24)	0.32 (0.29–0.34)
3rd period				
< 180	7,114	6,389,494	1.00 (Reference)	1.00 (Reference)
180–240	4,683	10,570,805	0.40 (0.38–0.41)	0.50 (0.48–0.52)
≥ 240	604	2,718,886	0.20 (0.18–0.22)	0.30 (0.28–0.33)
History of dyslipidemia				
No	11,182	16,611,757	1.00 (Reference)	1.00 (Reference)
Yes	1,219	3,067,428	0.59 (0.56–0.63)	0.61 (0.58–0.65)
History of HCV				
No	12,179	19,645,196	1.00 (Reference)	1.00 (Reference)
Yes	222	33,989	10.54 (9.23–12.03)	5.69 (4.96–6.52)
History of HBV				
No	11,298	19,532,881	1.00 (Reference)	1.00 (Reference)
Yes	1,103	146,305	13.04 (12.26–13.87)	11.88 (11.10–12.71)
History of ALD				
No	12,052	19,600,399	1.00 (Reference)	1.00 (Reference)
Yes	349	78,786	7.21 (6.48–8.02)	2.20 (1.94–2.48)
History of cryptogenic liver disease				
No	12,378	19,674,057	1.00 (Reference)	1.00 (Reference)
Yes	23	5,128	7.14 (4.75–10.75)	4.10 (2.72–6.19)
History of cirrhosis				
No	12,255	19,673,029	1.00 (Reference)	1.00 (Reference)
Yes	146	6,157	38.10 (32.36–44.85)	12.56 (10.64–14.83)
Family history of liver cancer				
No	11,709	19,233,089	1.00 (Reference)	1.00 (Reference)
Yes	692	446,096	2.55 (2.36–2.75)	2.49 (2.30–2.69)

Abbreviations: aHR, adjusted hazard ratio; CI, confidence interval; HCV, hepatitis C virus; HBV, hepatitis B virus; ALD, alcohol-related liver disease

Notes: aHRs were calculated using multivariable Cox proportional hazards regression, adjusting for collinearity and relevant covariates

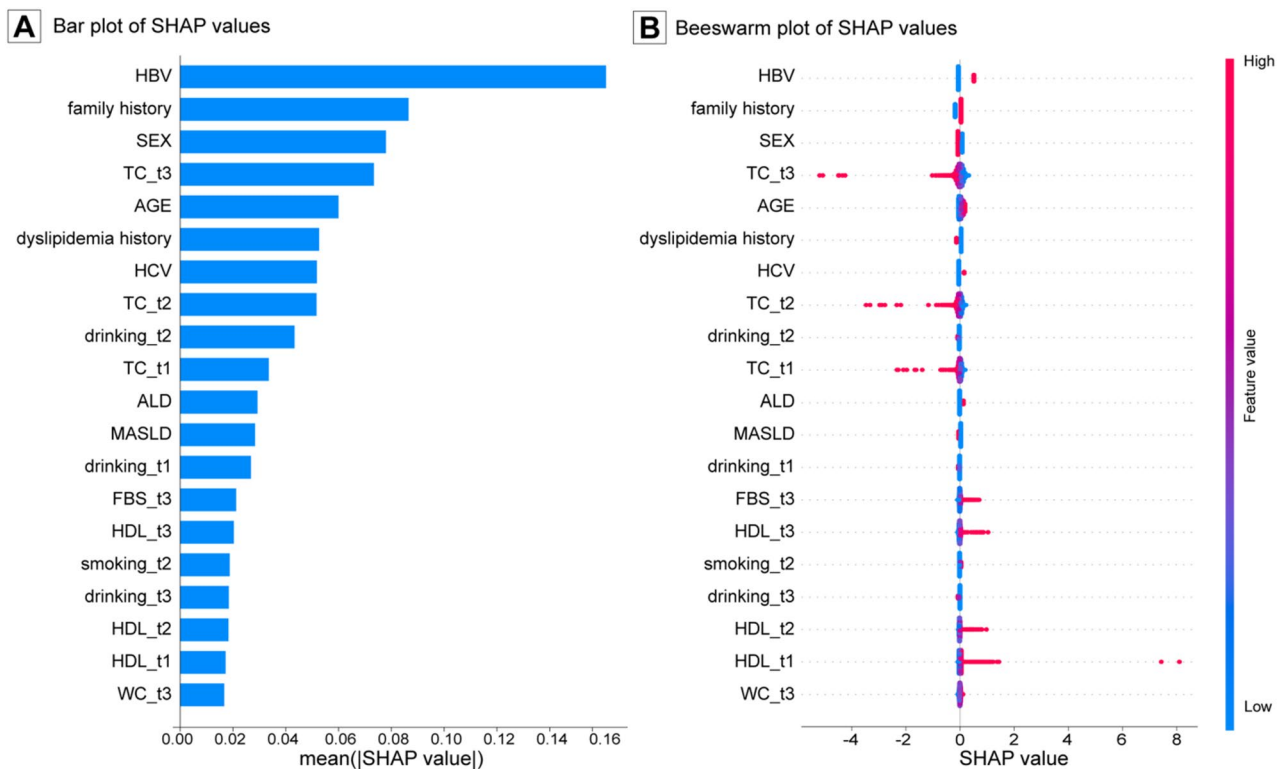


Fig. 3 Global Interpretability Analysis of the Fusion Model Using SHAP Values. **(A)** The bar plot presents the mean absolute SHAP values for each variable, reflecting their overall contribution to the prediction performance of the model. **(B)** The beeswarm plot shows the distribution and impact of each variable on model prediction performance, with color intensity indicating variable values. Abbreviations: SHAP, SHapley Additive exPlanations; HBV, hepatitis B virus; family_history, family history of liver cancer; HDL, high-density lipoprotein; TC, total cholesterol; HCV, hepatitis C virus; ALD, alcohol-related liver disease; MASLD, metabolic dysfunction-associated steatotic liver disease; FBS, fasting serum glucose; WC, waist circumference; t₁, t₂, and t₃ indicate health examination intervals 2010–2011, 2012–2013, and 2014–2015, respectively

21.28–35.34) compared with Q1. Sensitivity analyses using different combinations of screening intervals (Table S7) indicated that incorporating all three time points (t₁, t₂, and t₃) produced the highest overall performance. Among single-interval models, those using the most recent examination (t₃) consistently outperformed those based on earlier time points (t₁ or t₂).

Associations between the predictor and liver cancer

SHAP analysis

Figure 3 illustrates the top 20 features ranked by their mean absolute SHAP values, highlighting the most influential predictors in the liver cancer risk fusion model. HBV infection had the most significant contribution to model predictions, followed by a family history of liver cancer, sex, and TC levels measured at t₃. Age and a history of dyslipidemia also exerted a substantial influence on the predicted risk. Additional influence features included HCV infection, TC levels measured at earlier screenings (t₁ and t₂), and alcohol consumption at t₂. The beeswarm plot (Fig. 3B) illustrates the distribution of SHAP values across individual predictions. Lower TC levels at t₃ were associated with increased predicted liver

cancer risk, as indicated by predominantly positive SHAP values for low feature values. Similarly, higher alcohol consumption at t₂ corresponded with elevated risk, indicated by positive SHAP values for high feature values. Collectively, these results highlight the reliance of the model on a combination of clinical and behavioral variables measured longitudinally to generate individualized predictions of liver cancer risk.

Hazard ratios from Cox proportional hazards models

Table 2 presents a selection of the adjusted hazard ratios (aHRs) for liver cancer incidence estimated using Cox proportional hazards models. Variables were chosen based on either their prominence among the top 10 featured in the SHAP analysis or strong associations observed in the regression models. Older age (≥ 60 years; aHR, 1.61; 95% CI, 1.54–1.67) was associated with higher risk, whereas female sex was associated with lower risk (aHR, 0.29; 95% CI, 0.28–0.31). Heavy alcohol consumption at the second screening was associated with a higher risk (aHR, 1.33; 95% CI, 1.25–1.41), whereas light-to-moderate intake showed an inverse association (aHR, 0.89; 95% CI, 0.83–0.95).

TC exhibited a consistent inverse association with liver cancer risk across all three screening periods. Compared with levels < 180 mg/dL, levels $\geq$ 240 mg/dL were associated with substantially lower risk (aHR, 0.30; 95% CI, 0.28–0.33). A history of dyslipidemia also correlated with reduced risk (aHR, 0.61; 95% CI, 0.58–0.65). Conversely, chronic liver disease showed strong positive associations with liver cancer risk. Cirrhosis was the strongest predictor (aHR, 12.56; 95% CI, 10.64–14.83), followed by HBV (aHR, 11.88; 95% CI, 11.10–12.71), HCV (aHR, 5.69; 95% CI, 4.96–6.52), cryptogenic liver disease (aHR, 4.10; 95% CI, 2.72–6.19), and ALD (aHR, 2.20; 95% CI, 1.94–2.48). A family history of liver cancer was also associated with a substantially increased risk (aHR, 2.49; 95% CI, 2.30–2.69).

The complete set of Cox regression results, including unadjusted, multicollinearity-based adjusted, and fully adjusted models, is presented in Table S8. Adjusted hazard ratios were highly consistent between the two adjusted models, indicating that multicollinearity was effectively managed without compromising robustness or interpretability. Moreover, subdistribution hazard ratios from the Fine-Gray model were consistent with the Cox estimates (Table S9), indicating that the overall patterns of association remained consistent.

Discussion

In this nationwide cohort study involving 3.96 million Korean adults aged 50–69 years, we developed a fusion model to predict 5-year liver cancer incidence using longitudinal health screening data collected over 6 years. The model outperformed LR, XGBoost, MLP, and the KNKPC surveillance criteria, suggesting that population-level health screening data can effectively support liver cancer risk stratification and enable earlier, more sensitive detection than conventional criteria.

The fusion model achieved an AUROC of 0.810 and a sensitivity of 0.736, which was more than six times higher than the KNKPC criteria of 0.112, although specificity was lower. Two-thirds of liver cancer cases were observed in Q5, corresponding to a 27-fold increase in hazard compared with Q1. These findings underscore the potential of the model to identify high-risk individuals and to inform stratified surveillance strategies in both clinical and public health settings.

To improve interpretability, we applied both SHAP and Cox proportional hazards regression. Most major predictors—including HBV and HCV infections, family history of liver cancer, male sex, older age, ALD, elevated fasting glucose, smoking, and waist circumference—aligned with the findings of previous studies and current clinical guidelines [42–44]. Lower TC measured at t1, t2, and t3, as well as a history of dyslipidemia, showed a strong inverse association with liver cancer, consistent

with previous meta-analyses [45]. This inverse association should be interpreted cautiously. Statins, which are widely prescribed for dyslipidemia, have been linked to reduced HCC risk [46], and the absence of medication data prevented adjustment for this potential confounder and may have inflated the apparent protective association. Beyond confounding, biological mechanisms may also contribute, as cholesterol is a precursor for bile acids that protect against hepatic injury and may slow HCC progression [47, 48]. Experimental evidence from Qin et al. [49] revealed that elevated serum cholesterol levels enhanced natural killer cell-mediated antitumor activity and suppressed liver tumor growth in mice. Some predictors exhibited discrepancies between SHAP and Cox models. For instance, alcohol consumption appeared slightly protective in SHAP, showing a weak inverse trend, whereas Cox regression revealed a J-shaped association: decreased risk in light drinkers and increased risk in heavy drinkers relative to nondrinkers. Several mechanisms have been proposed to explain the association between alcohol consumption and liver cancer. Light to moderate alcohol consumption has been reported to improve insulin sensitivity and increase adiponectin levels, potentially reducing metabolic dysfunction associated with liver cancer [50–52]. In contrast, excessive alcohol consumption induces carcinogenesis through accumulated oxidative stress and the hepatotoxic metabolite acetaldehyde, leading to hepatocyte injury, fibrosis, and ultimately cirrhosis [53, 54]. However, the seemingly protective effect among light drinkers should be interpreted with caution. This finding may be influenced by the “sick quitter” effect, in which former drinkers who ceased alcohol consumption due to illness are categorized as non-drinkers [55, 56]. Similarly, HDL cholesterol contributed positively to predicted risk in SHAP but was inversely associated in the Cox regression, potentially reflecting nonlinear threshold effects reported in previous studies [57]. Although baseline characteristics indicated a higher crude prevalence of MASLD among liver cancer cases, MASLD was not a significant predictor in adjusted Cox models and contributed minimally in SHAP analysis. Although previous studies have linked MASLD to increased liver cancer risk [58, 59], HCC development in patients with MASLD typically occurs over 10–20 years [60], suggesting that our follow-up period may have been insufficient to capture this long-term progression. It is also well recognized that MASLD is substantially under-recorded in population-based and administrative databases [61, 62], which may have limited the accuracy of case identification and led to an underestimation of its true predictive contribution. Overall, SHAP and Cox regression showed substantial concordance in identifying major predictors, although some differences were observed. Cirrhosis and cryptogenic liver disease were

strongly associated with liver cancer in Cox models but did not rank among the top SHAP features, likely reflecting SHAP's emphasis on complex patterns and interactions versus the Cox model's focus on adjusted associations, a contrast also observed in UK Biobank analyses [63]. Combining both approaches improves interpretability and supports population-level liver cancer risk assessment.

Previous general population-based liver cancer risk models employing statistical or machine learning approaches have shown moderate performance (AUC: 0.712–0.873) but frequently relied on cross-sectional data or variables not routinely collected, and they did not fully account for nonlinear associations commonly observed in clinical data [16, 18, 64]. To address these limitations, we developed a model that simultaneously integrates temporal and static predictors, achieving superior accuracy in predicting liver cancer risk within large-scale cohorts relative to previous studies of comparable size.

The model captures a shift in liver cancer etiology toward metabolic risk factors and facilitates risk-based prescreening using routinely collected health screening data. Its robust discrimination of the highest-risk group (Q5) highlights its potential as a population-level screening tool. Although separation among intermediate-risk groups (Q2–Q3) was modest, further refinement could improve continuous stratification and support more nuanced clinical decision-making. The higher sensitivity of the model compared to the KNKPC algorithm indicates improved potential for early detection. However, with a specificity of 0.732, approximately 26.8% of non-cancer individuals would be classified as false positives, potentially triggering a large volume of unnecessary follow-up procedures that impose considerable financial costs, strain healthcare resources, and induce psychological stress [65]. To mitigate these potential harms while maintaining high sensitivity, a stepped, risk-stratified strategy may be considered, whereby the model assigns individuals to risk tiers and surveillance intervals are tailored accordingly. Further research will be needed to determine the optimal surveillance interval for each risk group. Although incorporating additional biomarkers may further enhance predictive performance, the principal strength of this study lies in revealing that routinely collected, policy-supported data can be effectively leveraged for risk stratification. Beyond clinical applicability, the model holds policy-level significance by showing how existing screening data can be repurposed to proactively identify high-risk individuals, providing a foundation for integrating predictive modeling into national cancer control programs to enable targeted prevention, optimize resource allocation, and support cost-effective, data-driven policy decisions. However, this study

has several limitations. First, key predictors, including alcohol consumption, physical activity, and specific self-reported medical or family histories, may be subject to reporting bias. Second, the operational definition of liver cancer differed by sex (C22.0 or C22.9 for men and C22.0 for women). However, it was derived from a validated comparison of NHIS-based definitions with the KCCR, which showed that including C22.9 for women tended to overestimate incidence rates. Thus, this sex-specific definition provided the closest alignment with national registry data, and the likelihood of underestimating liver cancer incidence among women in this study is minimal. Third, the exclusively Korean cohort may limit the generalizability of our findings to populations with different ethnic, environmental, or healthcare contexts, highlighting the need for validation in more diverse cohorts. In addition, medication data were unavailable, preventing the assessment of their protective or modifying effects. Furthermore, the retrospective design constrains causal inference, and residual confounding may persist despite statistical adjustments. In particular, regarding the observed J-shaped association between alcohol intake and liver cancer risk, unmeasured confounding may exist due to the inclusion of former drinkers among non-drinkers [66], potential underreporting or misclassification of alcohol intake, and unmeasured differences in socioeconomic status, dietary habits, or other lifestyle characteristics. Fourth, a cost-effectiveness analysis evaluating the trade-off between high sensitivity and lower specificity was not conducted in this study. Future work should address this gap by comparing the KNKPC criteria with a stepped, risk-stratified screening strategy using appropriate health economic modeling, such as Markov or other state-transition models. Fifth, about 10% of individuals were excluded due to missing key variables. Although some variables such as sex, income, and smoking showed moderate SMD differences, they were included as model predictors, reducing concerns about bias. Most clinical variables had small SMDs, suggesting limited impact on model performance (Table S1). Future studies should adopt more robust methods for handling missing data. Finally, deploying DL models requires substantial computational resources and technical infrastructure, which may pose challenges in some clinical settings, even as institutional investment in AI-driven healthcare systems continues to grow.

Conclusions

This nationwide study developed a DL-based fusion model leveraging longitudinal health screening data to predict 5-year liver cancer risk in adults aged over 50 years, with superior discrimination compared to conventional models and current surveillance criteria. Its principal strength lies in showing that AI applied to routinely

collected data can enhance identification of high-risk individuals and support more efficient, scalable, and data-driven strategies for liver cancer prevention within existing national screening infrastructure.

Abbreviations

HCC	Hepatocellular Carcinoma
KNKPC	2022 KLCA–NCC Korea Practice Guidelines
AFP	Alpha-Fetoprotein
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
MASLD	Metabolic Dysfunction–Associated Steatotic Liver Disease
ALD	Alcohol-Associated Liver Disease
DL	Deep Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
NHIS	National Health Insurance Service
LR	Logistic Regression
MLP	Multilayer Perceptron
SHAP	Shapley Additive Explanations
ICD-10	International Classification of Diseases, 10th Revision
BMI	Body Mass Index
TC	Total Cholesterol
HDL	High-Density Lipoprotein
AUROC	Area Under the Receiver Operating Characteristic Curve
HR	Hazard Ratio
aHR	Adjusted Hazard Ratio
CI	Confidence Interval
FBS	Fasting Serum Glucose
WC	Waist Circumference
t1, t2, t3	Health examination intervals (2010–2011, 2012–2013, and 2014–2015, respectively)
SD	Standard Deviation

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-025-03323-x>.

Supplementary Material 1

Acknowledgements

Administrative support was also provided by the National Health Insurance Service of Korea (NHIS-2023-1-795). We would like to thank Enago (www.enago.co.kr) for English language editing.

Author contributions

Drs. Park and H. Kim had access to all the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: Choi, Cho, Gu, Park, H. Kim. Acquisition, analysis, or interpretation of data: Choi, Cho, Gu, C. Kim, Park, H. Kim. Drafting of the manuscript: Choi, Cho, Gu. Critical review of the manuscript for important intellectual content: Choi, Cho, Gu, C. Kim, Park, H. Kim. Statistical analysis: Choi, Cho. Obtaining funding: Park, H. Kim. Administrative, technical, or material support: Choi, C. Kim. Supervision: Park and H. Kim. All authors read and approved the final manuscript.

Funding

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (Grant No. HA23C0083). This work was also supported by the Technology Innovation Program (RS-2025-02220286, Development of large language AI model-based techniques and platforms for nursery record generation and task automation) funded By the Ministry of Trade, Industry & Resources (MOTIR, Korea).

Data availability

Data cannot be shared publicly because the data are from the Korean National Health Insurance Service (NHIS) health screening cohort and are subject to access restrictions. Data are available from the NHIS Institutional Data Access/

Ethics Committee for qualified researchers who meet the criteria for access to confidential data. Contact: <https://nhiss.nhis.or.kr/>; Address: 32 Gungang-ro, Wonju-si, Gangwon-do 26464, Republic of Korea.

Declarations

Ethics approval and consent to participate

The Institutional Review Board of Chung-Ang University waived the requirement for both ethical approval and informed consent for this study, as it is a retrospective study that used anonymized data in accordance with the Bioethics and Safety Act (1041078–202112-HR-336–01). The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2000.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Institute for Innovation in Digital Healthcare, Yonsei University Health System, Seoul 03722, South Korea

²Department of Integrative Medicine, Yonsei University College of Medicine, Seoul 06229, South Korea

³Department of Neurosurgery, Yonsei University College of Medicine, 50-1, Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea

⁴Department of Artificial Intelligence, Yonsei University, Seoul 03722, South Korea

⁵Department of Preventive Medicine, College of Medicine, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea

Received: 23 September 2025 / Accepted: 15 December 2025

Published online: 17 January 2026

References

- Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74:229–63.
- Qiu S, Cai J, Yang Z, et al. Trends in hepatocellular carcinoma mortality rates in the US and projections through 2040. *JAMA Netw Open*. 2024;7:e2445525–2445525.
- Kulik L, El-Serag HB. Epidemiology and management of hepatocellular carcinoma. *Gastroenterology*. 2019;156:477–e4911.
- Jung K-W, Won Y-J, Kong H-J, Lee ES. Cancer statistics in Korea: Incidence, Mortality, Survival, and prevalence in 2016. *Cancer Res Treat*. 2019;51:417–30.
- Korean Liver Cancer Association (KLCA) and National Cancer Center (NCC) Korea. 2022 KLCA–NCC Korea practice guidelines for the management of hepatocellular carcinoma. *Korean J Radiol*. 2022;23:1126–1240.
- Kanwal F, Khaderi S, Singal AG, et al. Risk factors for HCC in contemporary cohorts of patients with cirrhosis. *Hepatol Baltim Md*. 2023;77:997–1005.
- McGlynn KA, Petrick JL, El-Serag HB. Epidemiology of hepatocellular carcinoma. *Hepatol Baltim Md*. 2021;73:4–13.
- Vaz J, Strömberg U, Midlöv P, Eriksson B, Buchebner D, Hagström H. Unrecognized liver cirrhosis is common and associated with worse survival in hepatocellular carcinoma: A nationwide cohort study of 3473 patients. *J Intern Med*. 2023;293:184–99.
- Le P, Tatar M, Dasarthy S, et al. Estimated burden of metabolic Dysfunction–Associated steatotic liver disease in US Adults, 2020 to 2050. *JAMA Netw Open*. 2025;8:e2454707.
- Kanwal F, Neuschwander-Tetri BA, Loomba R, Rinella ME. Metabolic dysfunction-associated steatotic liver disease: update and impact of new nomenclature on the American association for the study of liver diseases practice guidance on nonalcoholic fatty liver disease. *Hepatol Baltim Md*. 2024;79:1212–9.

11. Pinheiro PS, Zhang J, Setiawan VW, Cranford HM, Wong RJ, Liu L. Liver cancer etiology in Asian subgroups and American Indian, Black, Latino, and white populations. *JAMA Netw Open*. 2025;8:e252208.
12. Rinella ME, Lazarus JV, Ratzliff V, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *J Hepatol*. 2023;79:1542–56.
13. Serra-Burriel M, Juanola A, Serra-Burriel F, et al. Development, validation, and prognostic evaluation of a risk score for long-term liver-related outcomes in the general population: a multicohort study. *Lancet*. 2023;402:988–96.
14. Qayed E. The evolving landscape of hepatocellular carcinoma mortality in the US. *JAMA Netw Open*. 2024;7:e2445533.
15. Trevisani F, Garuti F, Neri A. Alpha-fetoprotein for Diagnosis, Prognosis, and transplant selection. *Semin Liver Dis*. 2019;39:163–77.
16. An C, Choi JW, Lee HS, Lim H, Ryu SJ, Chang JH, Oh HC. Prediction of the risk of developing hepatocellular carcinoma in health screening examinees: a Korean cohort study. *BMC Cancer*. 2021;21:755.
17. Kim HY, Lampertico P, Nam JY, et al. An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B. *J Hepatol*. 2022;76:311–8.
18. Thomas J, Liao LM, Sinha R, Patel T, Antwi SO. Hepatocellular carcinoma risk prediction in the NIH-AARP diet and health study cohort: A machine learning approach. *J Hepatocell Carcinoma*. 2022;9:69–81.
19. Ioannou GN, Tang W, Beste LA, Tincopa MA, Su GL, Van T, Tapper EB, Singal AG, Zhu J, Waljee AK. Assessment of a deep learning model to predict hepatocellular carcinoma in patients with hepatitis C cirrhosis. *JAMA Netw Open*. 2020;3:e2015626.
20. Khong TMT, Bui TT, Kang H-Y, Park E, Ki M, Choi Y-J, Kim B, Oh J-K. Cancer risk according to lifestyle risk score trajectories: a population-based cohort study. *BJC Rep*. 2025;3:28.
21. Moglia V, Johnson O, Cook G, de Kamps M, Smith L. Artificial intelligence methods applied to longitudinal data from electronic health records for prediction of cancer: a scoping review. *BMC Med Res Methodol*. 2025;25:24.
22. Sundrani S, Lu J. Computing the hazard ratios associated with explanatory variables using machine learning models of survival data. *JCO Clin Cancer Inf*. 2021;5:364–78.
23. Cheol Seong S, Kim Y-Y, Khang Y-H, et al. Data resource profile: the National health information database of the National health insurance service in South Korea. *Int J Epidemiol*. 2017;46:799–800.
24. Shin DW, Cho J, Park JH, Cho B. National general health screening program in Korea: history, current status, and future direction. *Precis Future Med*. 2022;6:9–31.
25. Cho S, Park S, Lee SK, Oh SN, Kim KH, Ko A, Park SM. Associations of changes in alcohol consumption on the risk of depression/suicide among initial nondrinkers. *Depress Anxiety*. 2024;2024:7560390.
26. Park H, Kim D, Jang E, et al. Modifiable lifestyle factors and lifetime risk of atrial fibrillation: longitudinal data from the Korea NHIS-HealS and UK biobank cohorts. *BMC Med*. 2024;22:194.
27. Park B, Kim CH, Jun JK, Suh M, Choi KS, Choi JJ, Oh HJ. A machine learning risk prediction model for gastric cancer with SHapley additive explanations. *Cancer Res Treat*. <https://doi.org/10.4143/crt.2024.843>.
28. Lee J, Lee JS, Park S-H, Shin SA, Kim K. Cohort profile: the National health insurance Service-National sample cohort (NHIS-NSC), South Korea. *Int J Epidemiol*. 2017;46:e15.
29. Seong SC, Kim Y-Y, Park SK, et al. Cohort profile: the National health insurance Service-National health screening cohort (NHIS-HEALS) in Korea. *BMJ Open*. 2017;7:e016640.
30. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster BV, et al. TRIPOD+ AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385.
31. Forjaz G, Howlader N, Scoppa S, Johnson CJ, Mariotto AB. Impact of including second and later cancers in Cause-Specific survival estimates using Population-based registry data. *Cancer*. 2022;128:547–57.
32. Kang MJ, Lim J, Han S-S, Park HM, Cho SC, Park S-J, Kim S-W, Won Y-J. Exocrine pancreatic cancer as a second primary malignancy: A population-based study. *Ann Hepato-Biliary-Pancreat Surg*. 2023;27:415–22.
33. Kim YR, Baek JY, Seo SH, Park H, Cho S, Shin A, et al. Operational definition of liver cancer in studies using data from the National Health Insurance Service: a systematic review. *J Cancer Prev*. 2023;28:47–52.
34. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. 2021. <https://doi.org/10.48550/arXiv.1908.01672>.
35. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. 2018. <https://doi.org/10.48550/arXiv.1708.02002>.
36. WELCH BL. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED. *Biometrika*. 1947;34:28–35.
37. Cochran WG. The χ^2 test of goodness of fit. *Ann Math Stat*. 1952;23:315–45.
38. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–45.
39. Fluss R, Faraggi D, Reiser B. Estimation of the Youden index and its associated cutoff point. *Biom J*. 2005;47:458–72.
40. Stel VS, Dekker FW, Tripepi G, Zoccali C, Jager KJ. Survival analysis I: the Kaplan-Meier method. *Nephron Clin Pract*. 2011;119:c83–8.
41. Cox DR. Regression models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972;34:187–220.
42. Ilagan-Ying YC, Gordon KS, Tate JP, Lim JK, Torgersen J, Lo Re V III, Justice AC, Taddei TH. Risk score for hepatocellular cancer in adults without viral hepatitis or cirrhosis. *JAMA Netw Open*. 2024;7:e2443608.
43. Singh SP, Madke T, Chand P. Global epidemiology of hepatocellular carcinoma. *J Clin Exp Hepatol*. 2025;15:102446.
44. 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 Syndr*. 2009;52:611.
45. Zhang Z, Xu S, Song M, Huang W, Yan M, Li X. Association between blood lipid levels and the risk of liver cancer: a systematic review and meta-analysis. *Cancer Causes Control*. 2024;35:943–53.
46. Wang Y, Wang W, Wang M, Shi J, Jia X, Dang S. A meta-analysis of statin use and risk of hepatocellular carcinoma. *Can J Gastroenterol Hepatol*. 2022;2022:5389044.
47. Luo W, Guo S, Zhou Y, Zhu J, Zhao J, Wang M, Sang L, Wang B, Chang B. Hepatocellular carcinoma: novel Understandings and therapeutic strategies based on bile acids (Review). *Int J Oncol*. 2022;61:117.
48. Liu Y, Chen K, Li F, et al. Probiotic *Lactobacillus rhamnosus* GG prevents liver fibrosis through inhibiting hepatic bile acid synthesis and enhancing bile acid excretion in mice. *Hepatol Baltim Md*. 2020;71:2050–66.
49. Qin W-H, Yang Z-S, Li M, et al. High serum levels of cholesterol increase antitumor functions of nature killer cells and reduce growth of liver tumors in mice. *Gastroenterology*. 2020;158:1713–27.
50. Joosten MM, Beulens JWJ, Kersten S, Hendriks HFJ. Moderate alcohol consumption increases insulin sensitivity and ADIPOQ expression in postmenopausal women: a randomised, crossover trial. *Diabetologia*. 2008;51:1375–81.
51. Davies MJ, Baer DJ, Judd JT, Brown ED, Campbell WS, Taylor PR. Effects of moderate alcohol intake on fasting insulin and glucose concentrations and insulin sensitivity in postmenopausal women: a randomized controlled trial. *JAMA*. 2002;287:2559–62.
52. Beulens JWJ, de Zoete EC, Kok FJ, Schaafsma G, Hendriks HFJ. Effect of moderate alcohol consumption on adipokines and insulin sensitivity in lean and overweight men: a diet intervention study. *Eur J Clin Nutr*. 2008;62:1098–105.
53. Tanihi M. Alcohol and hepatocarcinogenesis. *Clin Mol Hepatol*. 2020;26:736–41.
54. McKillop IH, Schrum LW, Thompson KJ. Role of alcohol in the development and progression of hepatocellular carcinoma. *Hepatic Oncol*. 2016;3:29–43.
55. Shaper AG, Wannamethee G, Walker M. Alcohol and mortality in British men: explaining the U-shaped curve. *Lancet Lond Engl*. 1988;2:1267–73.
56. Stockwell T, Zhao J, Panwar S, Roemer A, Naimi T, Chikritzhs T. Do moderate drinkers have reduced mortality risk? A systematic review and Meta-Analysis of alcohol consumption and All-Cause mortality. *J Stud Alcohol Drugs*. 2016;77:185–98.
57. Liu Z, Yuan H, Suo C, Zhao R, Jin L, Zhang X, Zhang T, Chen X. Point-based risk score for the risk stratification and prediction of hepatocellular carcinoma: a population-based random survival forest modeling study. *eClinicalMedicine*. 2024. <https://doi.org/10.1016/j.eclim.2024.102796>.
58. Rodriguez LA, Schmittiel JA, Liu L, Macdonald BA, Balasubramanian S, Chai KP, Seo SI, Mukhtar N, Levin TR, Saxena V. Hepatocellular carcinoma in metabolic Dysfunction-Associated steatotic liver disease. *JAMA Netw Open*. 2024;7:e2421019.
59. Jeong S, Oh YH, Ahn JC, et al. Evolutionary changes in metabolic dysfunction-associated steatotic liver disease and risk of hepatocellular carcinoma: A nationwide cohort study. *Clin Mol Hepatol*. 2024;30:487–99.
60. Hagström H, Shang Y, Hegmar H, Nasr P. Natural history and progression of metabolic dysfunction-associated steatotic liver disease. *Lancet Gastroenterol Hepatol*. 2024;9:944–56.

61. Hayward KL, Johnson AL, Horsfall LU, Moser C, Valery PC, Powell EE. Detecting non-alcoholic fatty liver disease and risk factors in health databases: accuracy and limitations of the ICD-10-AM. *BMJ Open Gastroenterol.* 2021;8:e000572.
62. Park JW, Yoo J-J, Lee DH, Chang Y, Jo H, Cho YY, Lee S, Kim LY, Jang JY. Evolving epidemiology of non-alcoholic fatty liver disease in South Korea: incidence, prevalence, progression, and healthcare implications from 2010 to 2022. *Korean J Intern Med.* 2024;39:931–44.
63. Liu X, Morelli D, Littlejohns TJ, Clifton DA, Clifton L. Combining machine learning with Cox models to identify predictors for incident post-menopausal breast cancer in the UK biobank. *Sci Rep.* 2023;13:9221.
64. Li X, Wang Y, Li H, Wang L, Zhu J, Yang C, Du L. Development of a prediction model and risk score for Self-Assessment and High-Risk population identification in liver cancer screening: prospective cohort study. *JMIR Public Health Surveill.* 2024;10:e65286.
65. Geh D, Rana FA, Reeves HL. Weighing the benefits of hepatocellular carcinoma surveillance against potential harms. *J Hepatocell Carcinoma.* 2019;6:23–30.
66. WANNAMETHEE G, SHAPER AG. Men who do not drink: A report from the British regional heart study. *Int J Epidemiol.* 1988;17:307–16.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.