Supplementary Table 1. Sample sizes used for building the models (total N=377,065).

	Training set			Validation set		
Model	**All samples**	**HCC cases**	**Controls**	**All samples**	**HCC cases**	**Controls**
Base Model	226239	388	225851	150826	259	150567
Overall	226239	388	225851	150826	259	150567
< 65 years	150020	228	149792	100059	147	99912
≥ 65 years	76219	160	76059	50767	112	50655
Male	129602	320	129282	86256	210	86046
Female	96637	68	96569	64570	49	64521
White	205472	325	205147	137096	210	136886
Other	17658	53	17605	11705	46	11659
BMI <25 kg/m^2	79354	86	79268	52694	56	52638
BMI 25-29.9 kg/m^2	92653	162	92491	61859	111	61748
BMI ≥30 kg/m^2	48779	126	48653	32541	88	32453
Diabetes-Yes	20708	91	20617	13917	81	13836
Diabetes-No	205531	297	205234	136909	178	136731
No alcohol intake	17888	40	17848	11933	31	11902
Alcohol ≤ 20 g/d	175481	265	175216	117228	180	117048
Alcohol > 20 g/d	32870	83	32787	21665	48	21617
Excellent/very good Health	118764	126	118638	79223	91	79132
Good Health	76447	174	76273	50617	112	50505
Fair/poor Health	27606	79	27527	18774	53	18721
Independent Models						
< 65 years	150047	225	149822	100032	150	99882
≥ 65 years	76191	163	76028	50795	109	50686
Male	129515	318	129197	86343	212	86131
Female	96724	70	96654	64483	47	64436
Diabetes-Yes	20775	103	20672	13850	69	13781
Diabetes-No	205464	285	205179	136976	190	136786
No alcohol intake	17893	43	17850	11928	28	11900
Alcohol ≤ 20 g/d	175625	267	175358	117084	178	116906
Alcohol > 20 g/d	32721	79	32642	21814	52	21762

Abbreviations: BMI, body mass index

Supplementary Table 2: Performance metrics of the 15 models evaluated in MATLAB for HCC risk prediction

	Elaborate model with 14 features and 5-fold cross-validation					
				Confusion Matrix		
Algorithm[a]	**AUC**	**Sensitivity**	**Specificity**	**Predictions**	**True HCC (n=388)**	**True Non-HCC (n=225,851)**
Fine Tree	0.620	0.000	1.000	Predicted HCC	0	23
				Predicted Non-HCC	388	225828
Medium Tree	0.600	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Coarse Tree	0.580	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Logistic Regression	0.750	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Gaussian Naive Bayes	0.660	0.049	0.984	Predicted HCC	19	3553
				Predicted Non-HCC	369	222298
Kernel Naive Bayes	0.640	0.003	1.000	Predicted HCC	1	69
				Predicted Non-HCC	387	225782
Linear Support Vector Machines	0.530	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Quadratic Support Vector Machines	0.460	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Cubic Support Vector Machines	0.510	0.000	1.000	Predicted HCC	0	16
				Predicted Non-HCC	388	225835
Fine Gaussian Support Vector Machines	0.590	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Medium Gaussian Support Vector Machines	0.490	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Coarse Gaussian Support Vector Machines	0.480	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Ensemble Boosted Trees	0.630	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Ensemble Bagged Trees (Random Forest)	0.590	0.000	1.000	Predicted HCC	0	0
				Predicted Non-HCC	388	225851
Ensemble RUSBoosted Trees[b]	0.720	0.680	0.632	Predicted HCC	264	83108
				Predicted Non-HCC	124	142743

[a]Fifteen algorithms that can accommodate both categorical and numerical variables were evaluated in MATLAB 2019b (MathWorks, Natick, MA) for HCC risk prediction.

[b]The Ensemble RUSBoosted Trees model was chosen for risk prediction due to its overall superior performance in terms of AUC, sensitivity, and specificity collectively compared to the other 14 models.

Abbreviations: AUC, area under the curve; HCC, hepatocellular carcinoma

Supplementary Table 3. Variables used for model training and validation

Variable	Format used for model training and validation
Age at baseline[a]	Continuous
Sex[a]	Male, Female
Body mass index (BMI, kg/m^2)[a]	Continuous
Height (meters)	Continuous
Diabetes[a]	Yes, no
Self-reported general health condition[a]	Excellent/very good, Good, Fair/poor
Alcohol use, g/d[a]	Continuous
Moderate to vigorous physical activity[a]	< 3 times/week vs. ≥3 times/week
Cholesterol, mg/d	Continuous
Saturated fat, g/d	Continuous
Trans-fatty acids, g/d	Continuous
Ounce equivalents of lean meat from eggs per day[a]	Continuous
Vitamin B_6, mg/d	Continuous
Healthy Eating Index score[a]	Continuous

[a]These variables were found to be statistically significant in a multivariable logistic regression model and were used to build a separate 9-variable model

Supplementary Table 4. Results for the base model[a] and stratified by patient characteristics in the validation sample

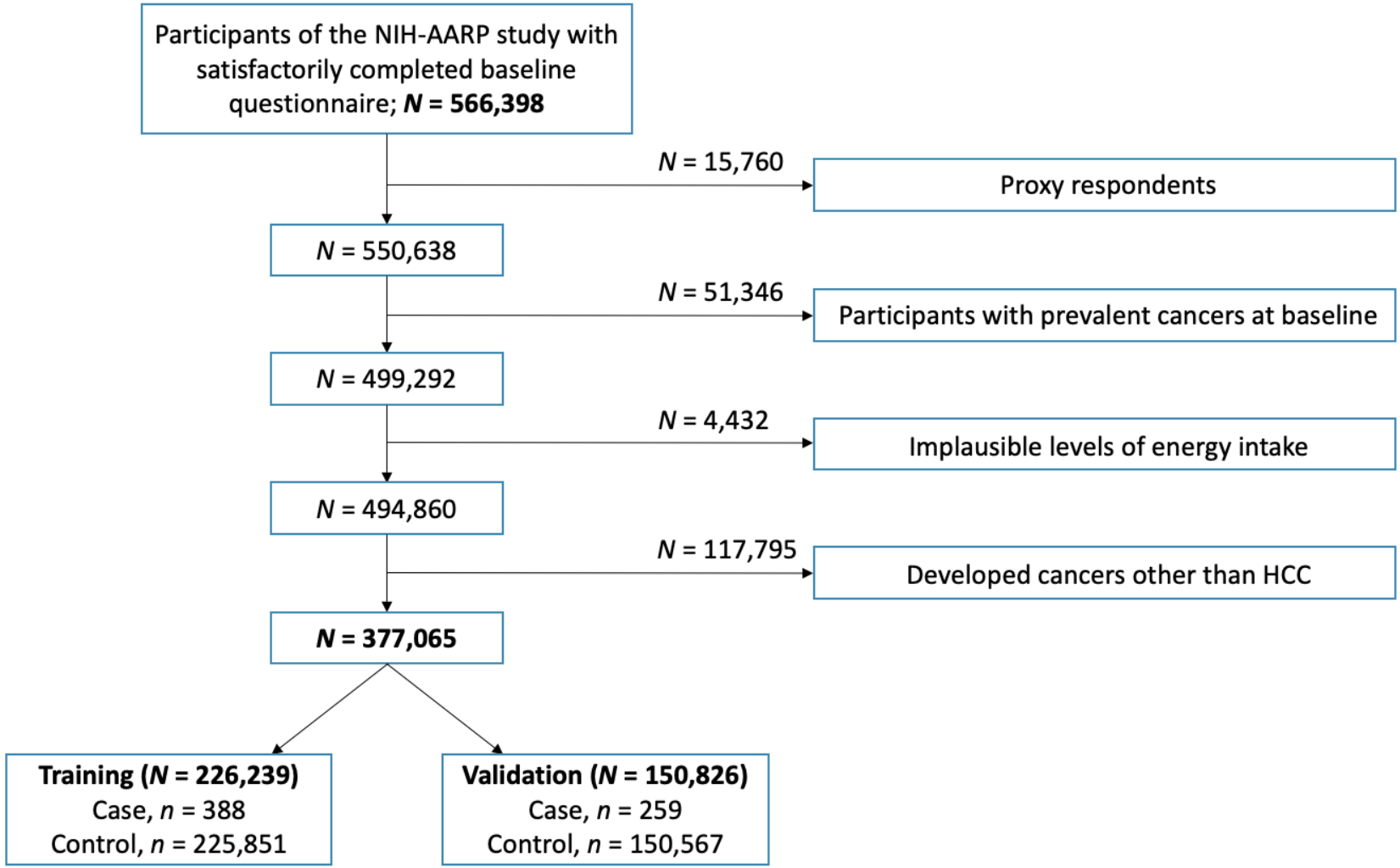
Participant Characteristics	Sensitivity	Specificity	AUC
< 65 years	0.633	0.658	0.645
≥ 65 years	0.786	0.518	0.652
Male	0.848	0.344	0.596
Female	0.061	0.967	0.514
White	0.729	0.607	0.668
Other	0.565	0.649	0.607
BMI <25 kg/m^2	0.643	0.744	0.694
BMI 25-29.9 kg/m^2	0.748	0.559	0.653
BMI ≥30 kg/m^2	0.682	0.489	0.585
Diabetes-Yes	0.790	0.329	0.559
Diabetes-No	0.657	0.639	0.648
No alcohol intake	0.452	0.605	0.529
Alcohol >0 to20 g/d	0.700	0.638	0.669
Alcohol ≥ 20 g/d	0.854	0.466	0.660
Excellent/Very Good Health	0.560	0.710	0.635
Good Health	0.759	0.519	0.639
Fair/Poor Health	0.792	0.436	0.614

[a]The variables used for constructing the base model are reported in Supplementary Table 1, samples sizes are provided in Supplementary Table 2.
Abbreviations: BMI, body mass index

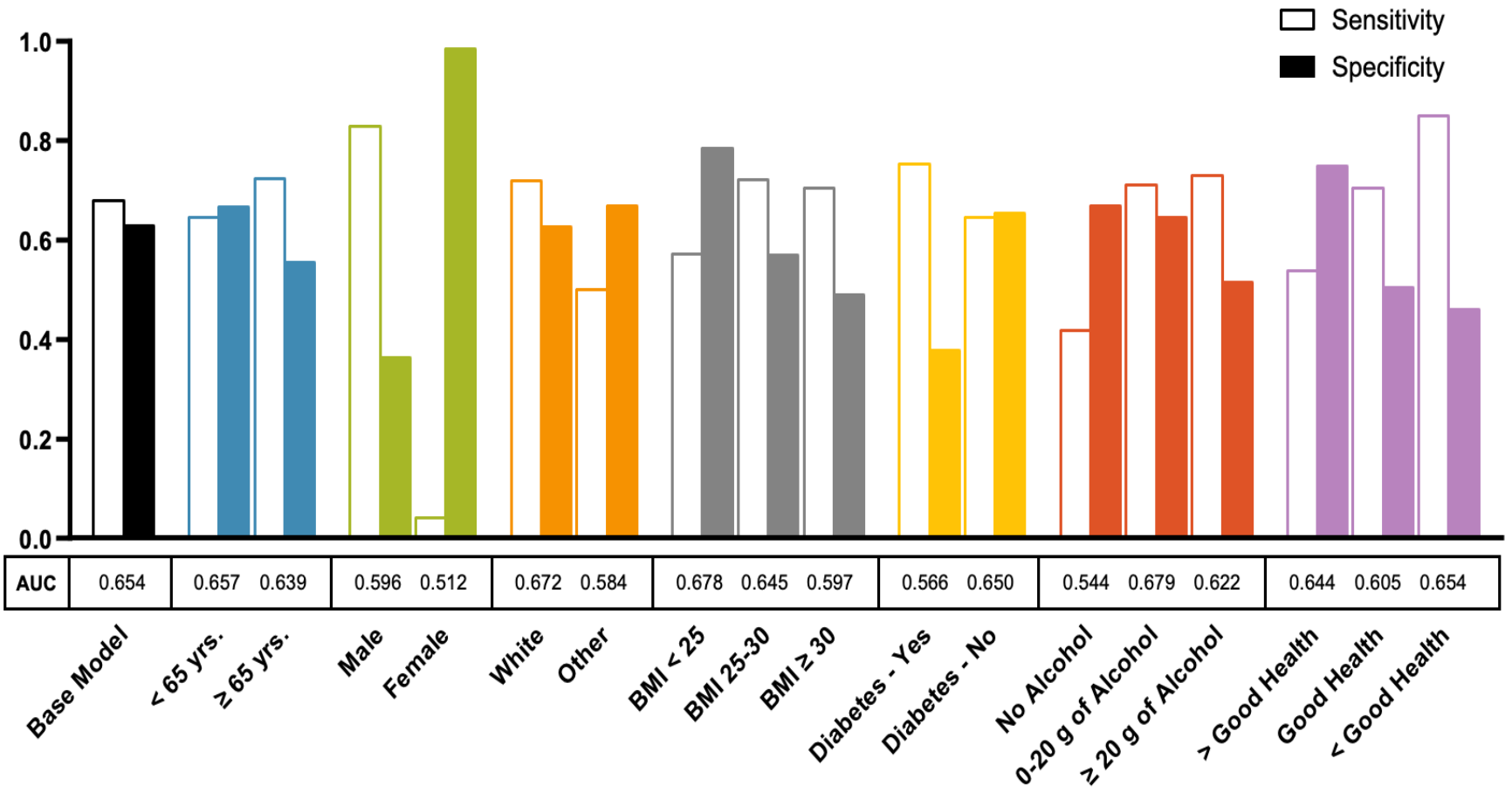
Supplementary Table 5. Results for independent models[a] tested in the validation sample

Patient characteristics	Sensitivity	Specificity	AUC
< 65 years	0.693	0.659	0.676
≥ 65 years	0.624	0.607	0.615
Male	0.547	0.644	0.596
Female	0.553	0.670	0.611
Diabetes-Yes	0.478	0.603	0.540
Diabetes-No	0.674	0.604	0.639
No Alcohol	0.500	0.598	0.549
0-20 g of Alcohol	0.697	0.627	0.662
≥ 20 g of Alcohol	0.596	0.630	0.613

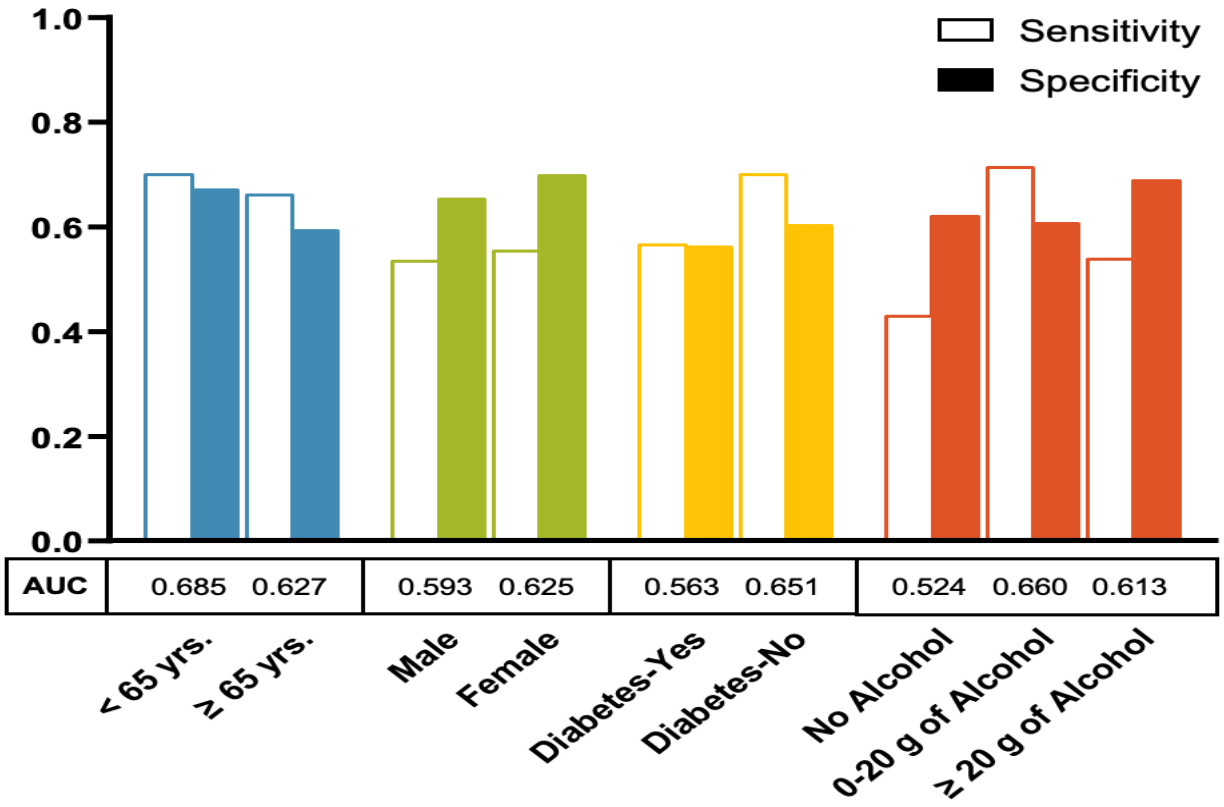
[a]The variables used for constructing the base model are reported in Supplementary Table 1, samples sizes are provided in Supplementary Table 2.
Abbreviations: BMI, body mass index



Supplementary Figure 1. Flowchart showing variable selection process; the NIH-AARP Diet and Health Study prospective cohort (1995-1996, with follow-up to 2011). Abbreviations: HCC, hepatocellular carcinoma.



Supplementary Figure 2. Performance of the elaborate base model for hepatocellular carcinoma risk prediction across population subgroups defined by age, gender, race, body mass index (BMI), diabetes status, alcohol intake, and self-reported general health condition among participants in the NIH-AARP Diet and Health Study prospective cohort (N=377,065; cases: 647, controls=376,418). The elaborate base model was built based on the following variables and then stratified by the patient characteristics: age (continuous), sex, BMI (continuous), height (continuous) diabetes (yes, no), general health status (excellent/very good, good, fair/poor), alcohol (continuous), moderate-to-vigorous physical activity (< 3 times/week, ≥3 times/week), dietary cholesterol (continuous), saturated fat (continuous), trans-fatty acids (continuous), ounce equivalents of lean meat from eggs per day (continuous), dietary vitamin B6 (mg/d, continuous), and healthy eating index scores (continuous).



Supplementary Figure 3. Hepatocellular carcinoma risk prediction models developed separately by age, gender, diabetes status, and alcohol intake among participants in the NIH-AARP Diet and Health Study prospective cohort. Each model was built separately based on the following 14 variables (the elaborate model variables): age (continuous), sex, BMI (continuous), height (continuous) diabetes (yes, no), general health status (excellent/very good, good, fair/poor), alcohol (continuous), moderate-to-vigorous physical activity (< 3 times/week, ≥3 times/week), dietary cholesterol (continuous), saturated fat (continuous), trans-fatty acids (continuous), ounce equivalents of lean meat from eggs per day (continuous), dietary vitamin B6 (mg/d, continuous), and healthy eating index scores (continuous).