ORIGINAL RESEARCH

Machine Learning to Improve Prognosis Prediction of Early Hepatocellular Carcinoma After Surgical Resection

Gu-Wei Ji[1–3],*
Ye Fan[1–3],*
Dong-Wei Sun[1–3],*
Ming-Yu Wu[4]
Ke Wang[1–3]
Xiang-Cheng Li[1–3]
Xue-Hao Wang[1–3]

[1]Hepatobiliary Center, The First Affiliated Hospital of Nanjing Medical University, Nanjing, People's Republic of China; [2]Key Laboratory of Liver Transplantation, Chinese Academy of Medical Sciences, Nanjing, People's Republic of China; [3]NHC Key Laboratory of Living Donor Liver Transplantation (Nanjing Medical University), Nanjing, People's Republic of China; [4]Department of Hepatobiliary Surgery, Wuxi People's Hospital, Wuxi, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xue-Hao Wang; Ke Wang
Hepatobiliary Center, The First Affiliated Hospital of Nanjing Medical University, 300 Guangzhou Road, Nanjing, 210029, People's Republic of China
Tel +86 13305178713; +86 18061675088
Fax +86 25 68136450; +86 25 68136450
Email wangxh@njmu.edu.cn; lancetwk@163.com

Background: Improved prognostic prediction is needed to stratify patients with early hepatocellular carcinoma (EHCC) to refine selection of adjuvant therapy. We aimed to develop a machine learning (ML)-based model to predict survival after liver resection for EHCC based on readily available clinical data.

Methods: We analyzed data of surgically resected EHCC (tumor≤5 cm without evidence of extrahepatic disease or major vascular invasion) patients from the Surveillance, Epidemiology, and End Results (SEER) Program to train and internally validate a gradient-boosting ML model to predict disease-specific survival (DSS). We externally tested the ML model using data from 2 Chinese institutions. Patients treated with resection were matched by propensity score to those treated with transplantation in the SEER-Medicare database.

Results: A total of 2778 EHCC patients treated with resection were enrolled, divided into 1899 for training/validation (SEER) and 879 for test (Chinese). The ML model consisted of 8 covariates (age, race, alpha-fetoprotein, tumor size, multifocality, vascular invasion, histological grade and fibrosis score) and predicted DSS with C-Statistics >0.72, better than proposed staging systems across study cohorts. The ML model could stratify 10-year DSS ranging from 70% in low-risk subset to 5% in high-risk subset. Compared with low-risk subset, no remarkable survival benefits were observed in EHCC patients receiving transplantation before and after propensity score matching.

Conclusion: An ML model trained on a large-scale dataset has good predictive performance at individual scale. Such a model is readily integrated into clinical practice and will be valuable in discussing treatment strategies.

Keywords: liver cancer, artificial intelligence, prognosis, modelling, surgery

Introduction

Hepatocellular carcinoma (HCC), the fourth leading cause of cancer-related death worldwide, typically occurs in patients with chronic liver disease and is an aggressive disease with dismal prognosis.[1] Over the past decades, improved surveillance programs and imaging techniques have led to early HCC (EHCC) diagnosis in 40–50% of patients, at a stage amenable to potentially curative therapies—resection, transplantation or ablation.[2,3] Generally, EHCC is expected to have an excellent outcome after radical therapies. Since total hepatectomy eliminates both the diseased liver and the tumor, liver transplantation (LT) offers the highest chance of cure, with a survival up to 70% at 10 years in selected cases, and remains the best

913

Graphical Abstract
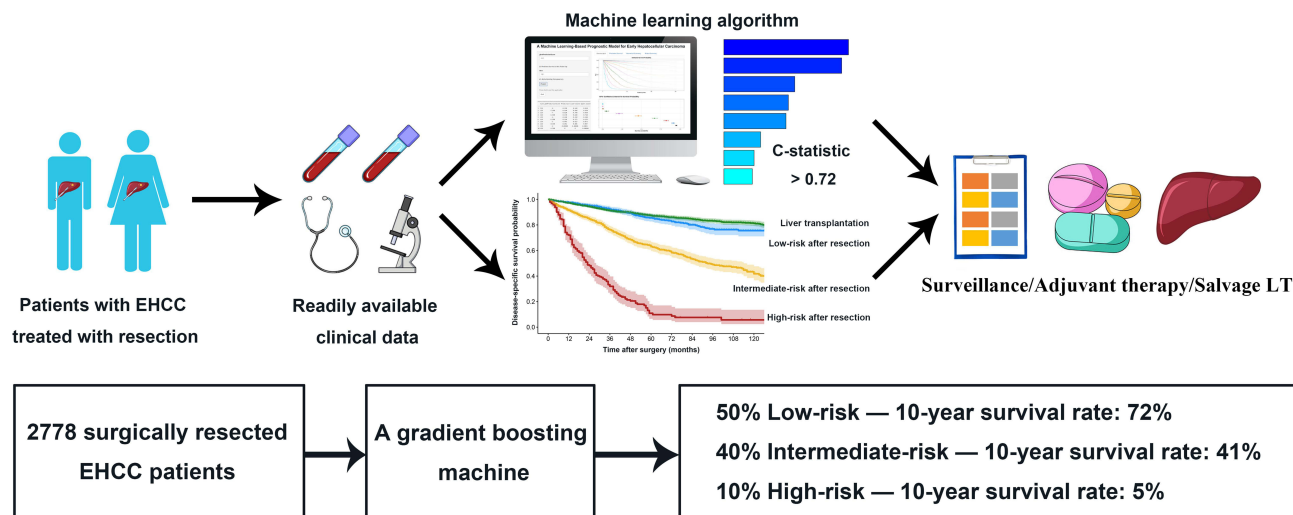
treatment for EHCC.[4] Unfortunately, the critical shortage of donor organs represents the main limitation of LT and results in long waiting times.

According to clinical practice guidelines, liver resection (LR) is the recommended first-line option for patients with EHCC and preserved liver function, although ablation is an alternative treatment modality.[3,5,6] The prognosis following LR may vary even among patients with EHCC and two competing causes of death (tumor recurrence and liver dysfunction) both influence survival.[7] Several HCC staging systems have been proposed to pair prognostic prediction with treatment allocation; however, these proposals—such as Barcelona Clinic Liver Cancer (BCLC) staging, China Liver Cancer (CNLC) staging, Hong Kong Liver Cancer (HKLC) staging and Cancer of the Liver Italian Program (CLIP) score—are not derived from surgically managed patients, except for the American Joint Committee on Cancer (AJCC) system and Japan Integrated Staging (JIS) score, and therefore exhibit modest prognostic accuracy for resected cases.[6–9] A few prognostic models have been developed based on readily available patient and tumor characteristics; however, they are by nature outmoded and rigid tools because all determinants were examined by conventional statistical methods (ie, Cox proportional hazard regression) and assigned fixed weights.[8,10] Hence, new strategies to improve outcome prediction and treatment selection are warranted for EHCC patients.

Machine learning (ML), a subfield of artificial intelligence, leverages algorithmic methods that enable computers to learn from on large-scale, heterogeneous datasets and execute a specific task without predefined rules.[11] ML solutions such as gradient boosting machine (GBM) have outperformed regression modelling in a variety of clinical situations (eg, diagnosis and prognosis).[11–13] Nevertheless, the benefit of ML in predicting prognosis of patients with resected EHCC has yet to be fully explored. Accordingly, we assembled a large, international cohort of EHCC patients to design and evaluate a ML-based model for survival prediction, and compare its performance with existing prognostic systems.

Materials and Methods
Study Population

Patients with EHCC, defined as tumor ≤5 cm and without evidence of extrahepatic disease or major vascular invasion,[14] were retrospectively screened from two sources: (1) Medicare patients treated with surgical therapy (LR+LT) in the Surveillance, Epidemiology, and End Results (SEER) Program, a population-based database in the United States, between 2004 and 2015; (2) consecutive patients treated with LR at two high-volume hepatobiliary centers in China (First Affiliated Hospital of Nanjing Medical University and Wuxi People's Hospital) between 2006 and 2016. The inclusion criteria were (1) adult patients aged ≥20 years; (2) histology-confirmed HCC (International Classification of Diseases for Oncology, Third Edition, histology codes 8170 to 8175 for HCC and site code C22.0 for liver);[15] (3) complete survival data and a survival of ≥1 month. The exclusion criteria were (a) missing information on the type

of surgical procedure; (b) another malignant primary tumor prior to HCC diagnosis; (c) unknown cause of death. Patient selection process is summarized in the flow chart of Figure 1. This study protocol was approved by the Institution Review Board of First Affiliated Hospital of Nanjing Medical University and Wuxi People's Hospital. Written informed consent was waived because retrospective anonymous data were analyzed. Non-identified information was used in order to protect patient data confidentiality. This study was conducted in accordance with the Declaration of Helsinki.

Outcome and Data Collection

The endpoint selected to develop ML-based model was disease-specific survival (DSS), defined as the time from the date of surgery to the date of death from disease (tumor relapse or liver dysfunction). All deaths from any other cause were counted as non-disease-specific and censored at the date of the last follow-up. Follow-up protocol for Chinese cohort included physical examination, laboratory evaluation

and dynamic CT or MRI of the chest and abdomen every 3 months during the first 2 years and every 6 months thereafter. The follow-up was terminated on August 15, 2020.

Electronic and paper medical records were reviewed in detail; all pertinent demographic and clinicopathologic data were abstracted on a standardized template. The following characteristics of interest were ascertained at the time of enrollment: age, gender, race, year of diagnosis, alpha-fetoprotein level, use of neoadjuvant therapy, tumor size, tumor number, vascular invasion, histological grade, liver fibrosis score, and type of surgery.

Machine Learning and Model Performance

We deployed GBM, a decision tree-based ML algorithm that has gained popularity because of its performance and interpretability, to aggregate baseline risk factors and predict the likelihood of survival using the R package "gbm". GBM algorithm[16] assembles multiple base learners, in
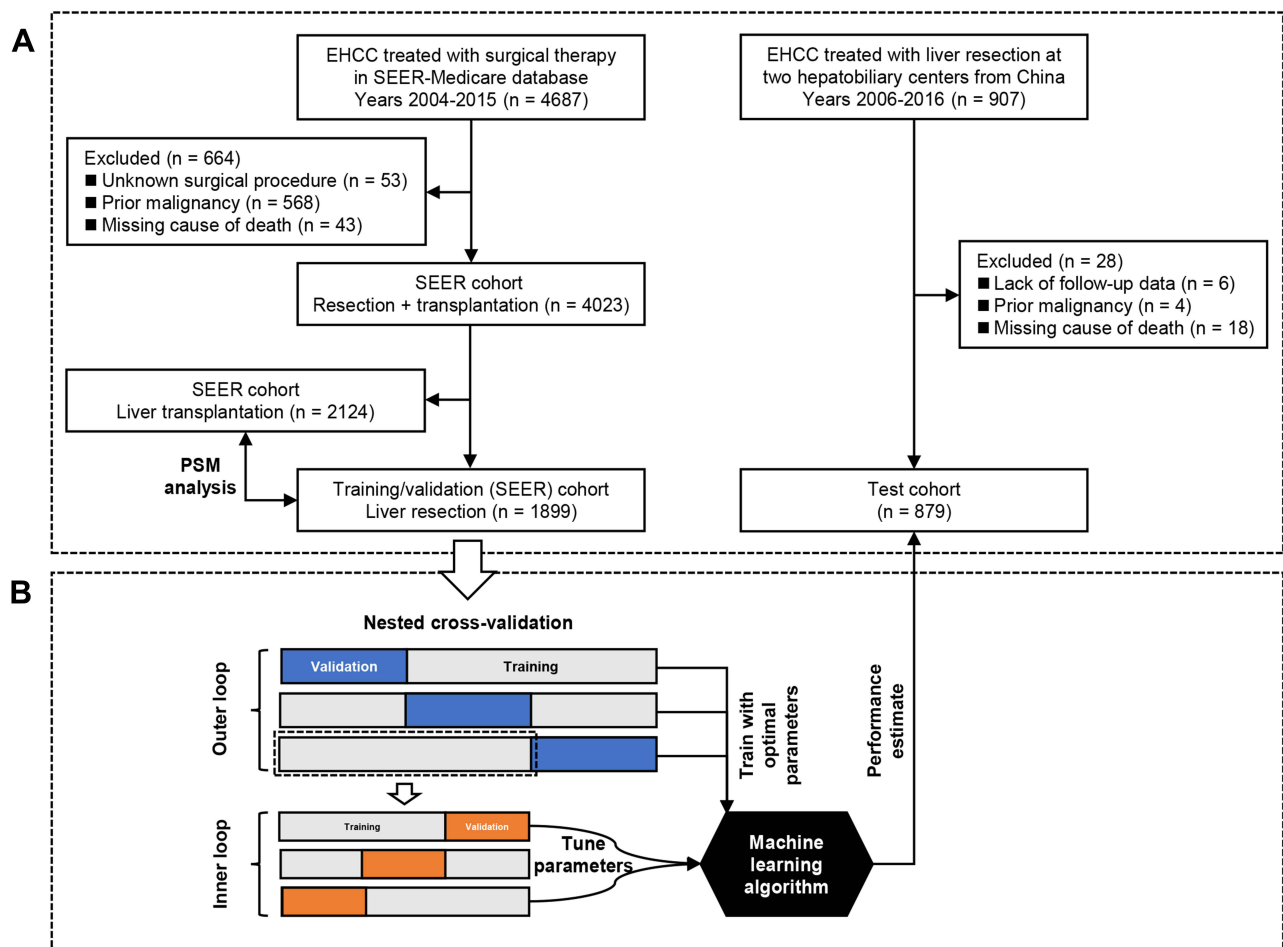

Figure 1 Analytical framework for survival prediction. (**A**) Flow diagram of the study cohort details. (**B**) A machine learning pipeline to train, validate and test the model.

a step-wise fashion, with each successive learner fitting the residuals left over from previous learners to improve model performance: (1) $F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m)$, where $h(x; a_m)$ is a base learner, typically a decision tree; (2) $F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$, where a_m is optimized parameters in each base learner and β_m is the weight of each base learner in the model. Each base learner may have different variables; variables with higher relative importance are utilized in more decision trees and earlier in the boosting algorithm. The model was trained using stratified 3×3-fold nested cross-validation (3 outer iterations and 3 inner iterations) on the training/validation cohort; a grid search of optimal hyper-parameter settings was run using the R package "mlr". Figure 1 shows the ML workflow schematically.

Model discrimination was quantified using Harrell's C-statistic and 95% confidence intervals [CIs] were assessed by bootstrapping. Calibration plots were used to assess the model fit. Decision curve analysis was used to determine the clinical net benefit associated with the adoption of the model.[17]

Statistical Analysis

Differences between groups were tested using χ2 test for categorical variables and Mann–Whitney U-test for continuous variables. Survival probabilities were assessed using the Kaplan–Meier method and compared by the Log rank test. The optimal cutoffs of GBM predictions were determined to stratify patients at low, intermediate, or high risk for disease-specific death by using X-tile software version 3.6.1 (Yale University School of Medicine, New Haven, CT).[18] Propensity score matching (PSM) was used to balance the LR versus LT for EHCC in SEER cohort using 1:1 nearest neighbor matching with a fixed caliper width of 0.02. Cases (LR) and controls (LT) were matched on all baseline characteristics other than type of surgery using the R package "MatchIt". All analyses were conducted using R software version 3.4.4 (www.r-project.org). Statistical significance was set at $P<0.05$; all tests were two-sided.

Results
Patient Data for Machine Learning

A total of 2778 EHCC patients (2082 males and 696 females; median age, 60 years; interquartile range [IQR], 54–67 years) treated with LR were identified and divided into 1899 for the training/validation (SEER) cohort and 879 for the test (Chinese) cohort. Patient characteristics of the training/

validation and test cohorts are summarized in Table 1. There were 625 disease-related deaths recorded (censored, 67.1%) during a median (IQR) follow-up time of 44.0 (26.0–74.0) months in the SEER cohort, and 258 deaths were recorded (censored, 70.6%) during a median (IQR) follow-up of 52.5 (35.8–76.0) months in the Chinese cohort. Baseline characteristics and post-resection survival differed between the cohorts.

Machine Learning Model and Prognostic Performance

We investigated 12 potential model covariates using GBM algorithm. According to the results of nested cross-validation, we utilized 2000 decision trees sequentially, with at least 5 observations in the terminal nodes of the trees; the decision tree depth was optimized at 3, corresponding to 3-way interactions, and the learning rate was optimized at 0.01. Covariates with a relative influence greater than 5 (age, race, alpha-fetoprotein level, tumor size, multifocality, vascular invasion, histological grade and fibrosis score) were integrated into the final model developed to predict DSS (Figure 2A and B).

The final GBM model demonstrated good discriminatory ability in predicting post-resection survival specific for EHCC, with a C-statistic of 0.738 (95% CI 0.717–0.758), and outperformed the 7th and 8th edition of AJCC staging systems ($P<0.001$) in the training/validation cohort (Table 2). The internal validation group was the 3×3-fold nested cross-validation of the final model of the training cohort with 211 patients in each fold. For the composite outcome, the GBM model yielded a median C-statistic of 0.727 (95% CI 0.706–0.761) and performed better than AJCC staging systems ($P<0.05$) in the internal validation group (Figure 2C). In the test cohort, the GBM model provided a C-statistic of 0.721 (95% CI, 0.689–0.752) in predicting DSS after resection of EHCC and was clearly superior to AJCC, BCLC, CNLC, HKLC, CLIP and JIS systems ($P<0.05$). Note that prediction scores differed between training/validation and test sets ($P<0.001$) (Figure S1). The discriminatory performance of ML-based model exceeded those of AJCC staging systems even in sub-cohorts stratified by covariate integrity (complete/missing) (Table S1). Furthermore, the GBM model exhibited greater ability to discriminate survival probabilities than simple prognostic strategies, such as multifocal EHCC with vascular invasion indicating a dismal prognosis following LR, in sub-cohorts with complete strategy-related information ($P<0.001$) (Table S2).

Table 1 Baseline Characteristics in the Training/Validation and Test Cohorts

Variables	Training/Validation (SEER) (n = 1899)	Test (China) (n = 879)	*P*-value
Age, years	62.0 (56.0–69.0)	56.0 (47.0–63.0)	<0.001
Gender			<0.001
Male	1380 (72.7)	702 (79.9)	
Female	519 (27.3)	177 (20.1)	
Race			<0.001
White	1020 (53.7)	0 (0.0)	
Asian/Pacifific Islander	598 (31.5)	879 (100.0)	
Black/American Indian/Alaskan	275 (14.5)	0 (0.0)	
Unknown	6 (0.3)	0 (0.0)	
Year of diagnosis			<0.001
Year 2010 and before	881 (46.4)	171 (19.5)	
Year 2011 and after	1018 (53.6)	708 (80.5)	
Neoadjuvant therapy			<0.001
No	1802 (94.9)	834 (94.9)	
Yes	70 (3.7)	45 (5.1)	
Unknown	27 (1.4)	0 (0.0)	
Multifocality			0.048
No	1562 (82.3)	737 (83.8)	
Yes	325 (17.1)	142 (16.2)	
Unknown	12 (0.6)	0 (0.0)	
Vascular invasion			<0.001
No	1516 (79.8)	765 (87.0)	
Yes	290 (15.3)	114 (13.0)	
Unknown	93 (4.9)	0 (0.0)	
Histological grade			<0.001
Well-differentiated	407 (21.4)	156 (17.7)	
Moderately differentiated	920 (48.4)	384 (43.7)	
Poorly differentiated or undifferentiated	360 (19.0)	296 (33.7)	
Unknown	212 (11.2)	43 (4.9)	
Tumor size, cm	3.0 (2.3–4.0)	3.0 (2.2–4.0)	0.784
Alpha-fetoprotein level			0.063
Normal	566 (29.8)	283 (32.2)	
Elevated	920 (48.4)	384 (43.7)	
Unknown or undetermined	413 (21.8)	212 (24.1)	
Fibrosis score			<0.001
None to moderate fibrosis	337 (17.7)	506 (57.6)	
Severe fibrosis or cirrhosis	441 (23.3)	371 (42.2)	
Unknown	1121 (59.0)	2 (0.2)	
Type of surgery			<0.001
Wedge or segmental resection	1265 (66.6)	745 (84.8)	
Lobectomy	440 (23.2)	94 (10.7)	
Extended lobectomy or other hepatectomy	194 (10.2)	40 (4.5)	
Median DSS time, months*	138.0 (114.0-undefined)	Undefined	0.003

Notes: Continuous variables reported as median (interquartile range) and categorical variables reported as number (percentage). *Numbers in parentheses are 95% confidence interval.
Abbreviations: SEER, Surveillance, Epidemiology, and End Results; DSS, disease-specific survival.
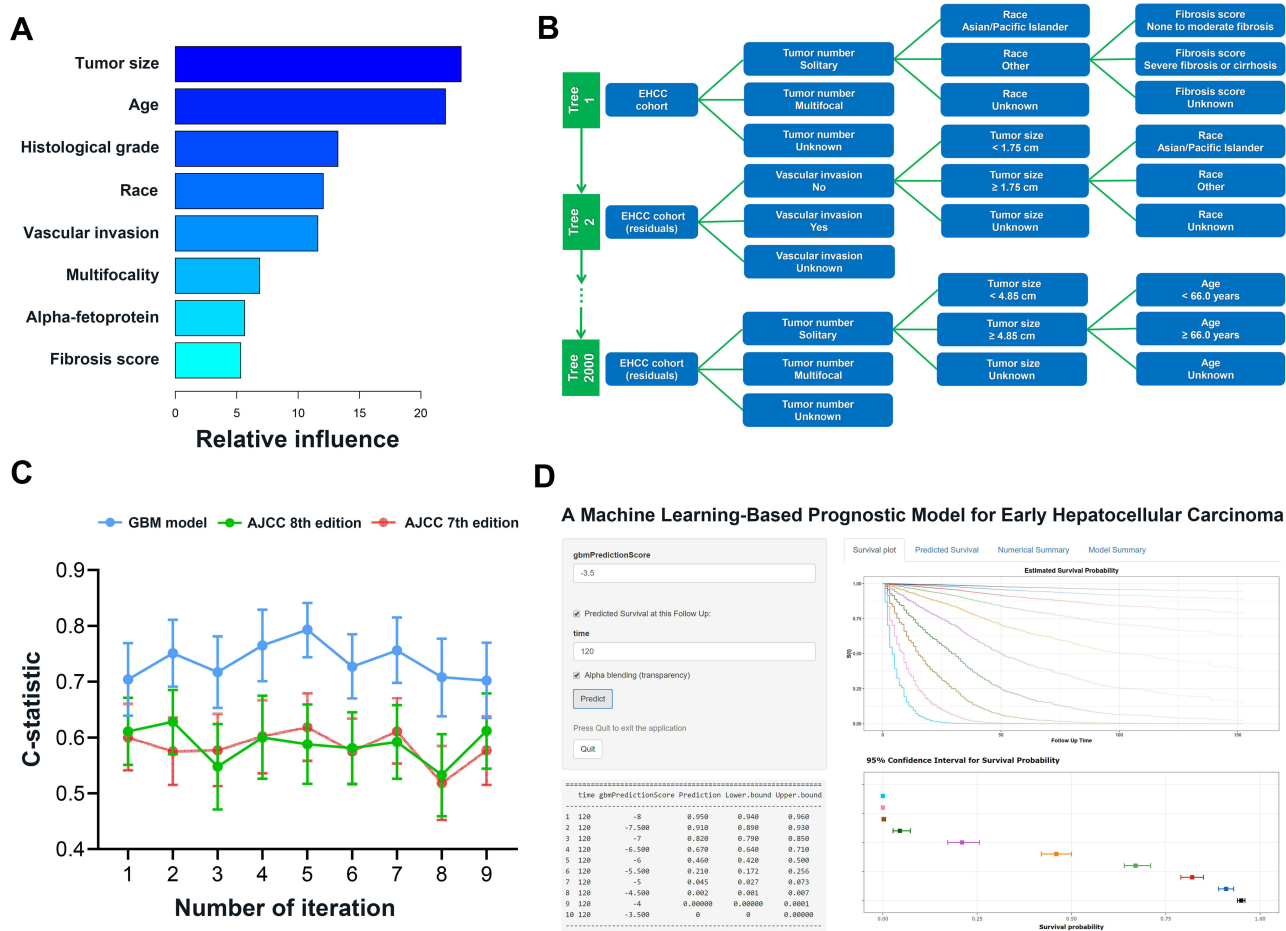
Figure 2 Overview of the machine-learning-based model. (**A**) Relative importance of the variables included in the model. (**B**) Illustrative example of the gradient boosting machine (GBM). GBM builds the model by combining predictions from stumps of massive decision-tree-base-learners in a step-wise fashion. GBM output is calculated by adding up the predictions attached to the terminal nodes of all 2000 decision trees where the patient traverses. (**C**) Performance of GBM model as compared with that of American Joint Committee on Cancer (AJCC) staging in the internal validation group. (**D**) Online model deployment based on GBM output.

Calibration plots presented excellent agreement between model predicted and actual observed survival in both the training/validation and test cohorts (Figure S2A and B). Decision curve analysis demonstrated that the GBM model provided better clinical utility for EHCC in designing clinical trials than the "treat all" or "treat none" strategy across the majority of the range of reasonable threshold probabilities (Figure S2C and D). The model is publicly accessible for use on Github (https://github.com/radgrady/EHCC_GBM), with an app (https://mlehcc.shinyapps.io/EHCC_App/) that allows survival estimates at individual scale (Figure 2D).

Risk Stratification

We utilized X-tile analysis to generate two optimal cut-off values (−6.35 and −5.32 in GBM predictions, Figure S3) that separated EHCC patients into 3 strata with a highly different probability of post-resection survival in the training/validation cohort: low risk (760 [40.0%]; 10-year DSS, 75.6%),

intermediate risk (948 [49.9%]; 10-year DSS, 41.8%), and high risk (191 [10.1%]; 10-year DSS, 5.7%) (*P*<0.001). In the test cohort, the aforementioned 3 prognostic strata by using the GBM model were confirmed: low risk (634 [72.1%]; 10-year DSS, 69.0%), intermediate risk (194 [22.1%]; 10-year DSS, 37.9%), and high risk (51 [5.8%]; 10-year DSS, 4.7%) (*P*<0.001) (Table 3). Visual inspection of the survival curves again revealed that, compared with the 8th edition AJCC criteria, the GBM model provided better prognostic stratification in both the training/validation and test cohorts (Figure 3). Differences in the baseline patient characteristics according to risk groups defined by the GBM model are summarized in Table S3.

Resection versus Transplantation in the SEER Cohort

We also gathered data of 2124 EHCC patients (1671 males and 453 females; median age, 58 years; IQR, 53–62 years)

Table 2 Performance of GBM Model and Staging Systems

Prognostic Marker	C-Statistic (95% CI)	P-value
Training/validation cohort (n=1899)		
GBM model	0.738 (0.717–0.758)	Ref
AJCC 8th edition	0.588 (0.566–0.611)	<0.001
AJCC 7th edition	0.585 (0.564–0.605)	<0.001
Test cohort (n=879)		
GBM model	0.721 (0.689–0.752)	Ref
AJCC 8th edition	0.667 (0.634–0.700)	<0.001
AJCC 7th edition	0.675 (0.645–0.704)	<0.001
BCLC stage	0.603 (0.574–0.633)	<0.001
CNLC stage	0.596 (0.567–0.624)	<0.001
HKLC stage	0.637 (0.607–0.667)	<0.001
CLIP classification [a]	0.588 (0.547–0.629)	<0.001
JIS score	0.677 (0.645–0.708)	0.002

Note: [a]Available at baseline (667/879) and compared with GBM model in test cohort.
Abbreviations: GBM, gradient boosting machine; AJCC, American Joint Committee on Cancer; BCLC, Barcelona Clinic Liver Cancer; CNLC, China Liver Cancer; HKLC, Hong Kong Liver Cancer; CLIP, Cancer of the Liver Italian Program; JIS, Japan Integrated Staging.

treated with LT from the SEER-Medicare database. SEER data demonstrated that considerable differences existed between LR (n=1899) and LT (n=2124) cohorts in terms of all listed clinical variables except for alpha-fetoprotein level (Table S4). Upon initial analysis, we found a remarkable survival benefit of LT over LR for patients with EHCC (hazard ratio [HR] 0.342, 95% CI 0.300–0.389, P<0.001), which was further confirmed in a well-matched cohort of 1892 patients produced by PSM (HR 0.342, 95% CI 0.285–0.410, P<0.001). Although a trend for higher survival probability was observed after 5 years in the LT cohort, no statistically significant difference in DSS was

observed when compared with low-risk LR cohort (HR 0.850, 95% CI 0.679–1.064, P=0.138). After PSM, 420 patients in the LT cohort were matched to 420 patients in the low-risk LR cohort; the trend for improved survival remained after 5 years in the matched LT cohort while the matched comparison also yielded no significant survival difference (HR 0.802, 95% CI 0.561–1.145, P=0.226) (Figure 4). By contrast, when compared with intermediate- and high-risk patients treated with LR, remarkable survival benefits were observed in patients treated with LT both before and after PSM (P<0.001) (Table S5).

Discussion

In this study involving over 2700 EHCC patients treated with resection, a gradient-boosting ML model was trained, validated and tested to predict post-resection survival. Our results demonstrated that this ML model utilized readily available clinical information, such as age, race, alpha-fetoprotein level, tumor size and number, vascular invasion, histological grade and fibrosis score, and provided real-time, accurate prognosis prediction (C-statistic >0.72) that outperform traditional staging systems. Among the model covariates, tumor-related characteristics, such as size, multifocality and vascular invasion, as well as liver cirrhosis are known risk factors for poor survival following resection of HCC.[7–10] Besides, multiple population-based studies have shown the racial and age differences in survival of HCC.[19,20] Therefore, our ML model is a valid and reliable tool to estimate prognosis of EHCC patients. This study represents, to our knowledge, the first application of a state-of-the-art ML survival prediction algorithm in EHCC based on large-scale, heterogeneous datasets.

In SEER cohort, the 10-year survival rate of EHCC after LR was around 50%, which seemed acceptable but

Table 3 Disease-Specific Survival According to Risk Stratification

Risk Group	Median Time, Months (95% CI)	2-Year Rate, % (95% CI)	5-Year Rate, % (95% CI)	10-Year Rate, % (95% CI)	Hazard Ratio (95% CI)	P-value
Training/validation cohort (n=1899)						
Low-risk (n=760)	Undefined	95.2 (93.7–96.8)	85.5 (82.7–88.5)	75.6 (71.1–80.3)	1	
Intermediate-risk (n=948)	93.0 (83.0–114.0)	84.6 (82.3–87.0)	63.0 (59.6–66.7)	41.8 (36.7–47.6)	3.039 (2.536–3.641)	<0.001*
High-risk (n=191)	23.0 (20.0–28.0)	48.2 (41.5–55.9)	10.9 (6.6–18.2)	5.7 (2.4–13.6)	3.876 (2.907–5.169)	<0.001†
Test cohort (n=879)						
Low-risk (n=634)	Undefined	95.5 (93.9–97.2)	81.2 (77.9–84.7)	69.0 (63.8–74.6)	1	
Intermediate-risk (n=194)	68.0 (51.8–80.6)	78.6 (73.0–84.7)	52.6 (45.2–61.3)	37.9 (29.3–49.2)	3.237 (2.289–4.578)	<0.001*
High-risk (n=51)	26.0 (22.0–39.0)	56.3 (44.1–71.8)	18.9 (9.4–38.2)	4.7 (0.8–29.7)	2.607 (1.601–4.243)	<0.001†

Note: *P value versus low-risk; †P value versus intermediate-risk.
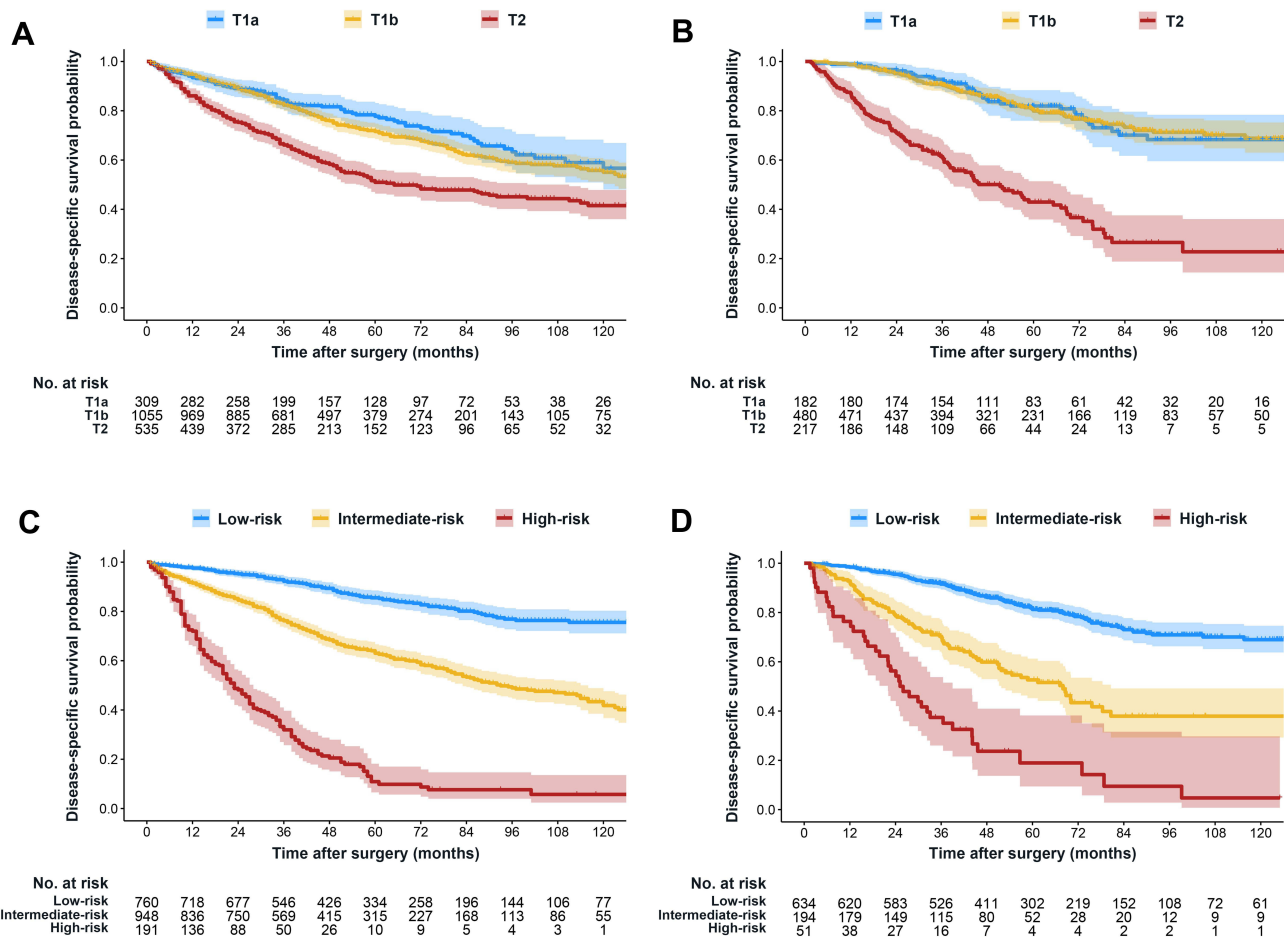**Abbreviation:** CI, confidence interval.

Figure 3 Kaplan-Meier survival plots demonstrating disparities between groups. Disease-specific survival stratified by the 8th edition of the American Joint Committee on Cancer T stage and the machine-learning model in the training/validation (**A** and **C**) and the test (**B** and **D**) cohort.
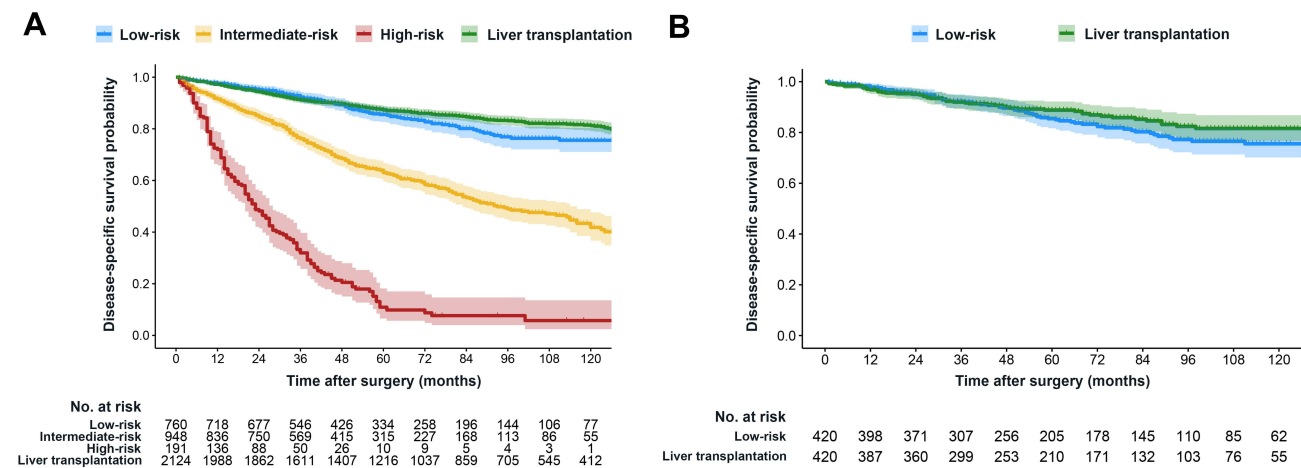

Figure 4 Comparison of survival after resection versus transplantation before and after propensity score matching in SEER-Medicare database. (**A**) Kaplan–Meier curves for different risk groups stratified by the model in the SEER resection cohort (n=1899) and patients in the SEER transplantation cohort (n=2124). (**B**) Kaplan–Meier curves for low-risk patients treated with resection and patients treated with transplantation in propensity score-matched cohort (n=840).

was remarkably lower than that after LT (around 80%). No adjuvant therapies are able to prevent tumor relapse and cirrhosis progression; however, patients with dismal

prognosis should be considered candidates for clinical trials of adjuvant therapy.[7] Salvage LT has also been a highly applicable strategy to alleviate both graft shortage

and waitlist dropout with excellent outcomes that are comparable to upfront LT.[1,5] Priority policy, defined as enlistment of patients at high mortality risk before disease progression, was then implemented to improve the transplantability rate.[21] Promisingly, our ML tool may help clinicians better identify EHCC patients who are at high risk of disease-related death, engage in clinical trials, and meet priority enlistment policy. Specifically, the GBM model identified 10% of EHCC patients who suffered from extremely dismal prognosis following LR in this study. Given its small proportion and survival benefit, we advocate the pre-emptive enlistment of high-risk subset for salvage LT after LR to avoid the later emergence of advanced disease (ie, tumor recurrence and liver decompensation) ultimately leading to death. Moreover, 40% of EHCC patients were at intermediate risk of disease-related death; adjuvant treatments that target HCC and cirrhosis are desirable. In turn, nearly half of EHCC patients were categorized as low risk by using the GBM model. The low-risk subset permits satisfactory long-term survival after LR and may receive no adjuvant therapy. We note that DSS curves are separated after 5 years for low-risk patients treated with LR as compared with patients treated with upfront LT, and thus long-lasting surveillance should be maintained.

Prior efforts to improve prognostic prediction of EHCC have mostly been reliant on tissue-based or imaging-assisted quantification of research biomarkers.[9,22] However, a more accurate, yet more complex, prognosis estimate does not necessarily present a better clinical tool. Parametric regression models are ubiquitous in clinical research because of their simplicity and interpretability; however, regression analysis should be performed in complete cases only.[23] Moreover, regression modeling strategies assume that relationships among input variables are linear and homogeneous but complicated interactions exist between predictors.[24,25] Decision tree-based methods represent a large family of ML algorithms and can reveal complex non-linear relationships between covariates. GBM algorithm has been widely applied in big data analysis and consistently utilized by the top performers of ML predictive modelling competitions.[14,26] GBM algorithm utilizes the boosting procedure to combine stumps of massive decision-tree-base-learners, which is similar to the clinical decision-making process for a patient by aggregating consultations from multiple specialists, each which would that look at the case in a slightly different way. Thus, our GBM model directly integrates interpretability

in order to mitigate this issue. Compared with other tree-based ensemble methods such as random forest, GBM algorithm also has a built-in functionality to handle missing values that permits utilizing data from, and assigning classification to, all observations in the cohort without the need to impute data. We applied nested cross-validation scheme for hyperparameter tuning in GBM as it prevents information leaking between observations used for training and validating the model, and estimates the external test error of the given algorithm on unseen datasets more accurately by averaging its performance metrics across folds.[27] Comparable discriminatory ability in the training/validation cohort, the test cohort as well as sub-cohorts from different clinical scenarios suggested good reproducibility and reliability of the proposed GBM model.

Our study has several limitations that warrant attention. First, all the presented analyses are retrospective; prospective validations of the ML model in different populations are warranted prior to routine use in clinical practice. Second, the study cohort included population-based cancer registries with limited information regarding patient and tumor characteristics; unavailable confounders, such as biochemical parameters, surgical margin status and recurrence treatment modality could not be adjusted for modeling. Third, SEER-Medicare database contains a considerable amount of missing data in several important clinical variables, such as fibrosis score. Indeed, missing data represent an unavoidable feature of all clinical and population-based databases; however, improper management of data resource, such as simply excluding cases with missing data, can introduce considerable bias, as previously noted across numerous cancer types.[28] We therefore contend that integrating missingness into our GBM model indicates good transferability in future clinical practice.

Conclusions

In conclusion, ML approach is both feasible and accurate, and a novel way to consider analysis of survival outcomes in clinical scenarios. Our results suggest that a GBM model trained on readily-available clinical data provides good performance that is better than staging systems in predicting prognosis. Although several issues must be addressed, such as prospective validations and ethical challenges, prior to its widespread use, such an automated tool may complement existing prognostic sources and lead to better personalized treatments for patients with resected EHCC.

Abbreviations

EHCC, early hepatocellular carcinoma; LT, liver transplantation; LR, liver resection; BCLC, Barcelona Clinic Liver Cancer; China Liver Cancer, CNLC; HKLC, Hong Kong Liver Cancer; CLIP, Cancer of the Liver Italian Program; AJCC, American Joint Committee on Cancer; ML, machine learning; GBM, gradient boosting machine; SEER, Surveillance, Epidemiology, and End Results; DSS, disease-specific survival; PSM, propensity score matching; IQR, interquartile range.

Data Sharing Statement

Data for model training and validation as well as R codes are available at Github (https://github.com/radgrady/EHCC_GBM). Test data are available from the corresponding author (Xue-Hao Wang) on reasonable request.

Ethics Approval and Informed Consent

This study protocol was approved by the Institution Review Board of First Affiliated Hospital of Nanjing Medical University and Wuxi People's Hospital. Written informed consent was waived because retrospective anonymous data were analyzed. Non-identified information was used in order to protect patient data confidentiality.

Funding

This study was supported by the Key Program of the National Natural Science Foundation of China (31930020) and the National Natural Science Foundation of China (81530048, 81470901, 81670570).

Disclosure

The authors declare no potential conflicts of interest.

References

1. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR. A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat Rev Gastroenterol Hepatol*. 2019;16(10): 589–604. doi:10.1038/s41575-019-0186-y
2. Llovet JM, Montal R, Sia D, Finn RS. Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat Rev Clin Oncol*. 2018;15(10):599–616. doi:10.1038/s41571-018-0073-4
3. European Association for the Study of the Liver. EASL clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol*. 2018;69(1):182–236. doi:10.1016/j.jhep.2018.03.019
4. Pinna AD, Yang T, Mazzaferro V, et al. Liver transplantation and hepatic resection can achieve cure for hepatocellular carcinoma. *Ann Surg*. 2018;268(5):868–875. doi:10.1097/SLA.0000000000002889
5. Marrero JA, Kulik LM, Sirlin CB, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology*. 2018;68(2):723–750. doi:10.1002/hep.29913
6. Zhou J, Sun H, Wang Z, et al. Guidelines for the diagnosis and treatment of hepatocellular carcinoma (2019 Edition). *Liver Cancer*. 2020;9(6):682–720. doi:10.1159/000509424
7. Villanueva A. Hepatocellular carcinoma. *N Engl J Med*. 2019;380 (15):1450–1462. doi:10.1056/NEJMra1713263
8. Chan AWH, Zhong J, Berhane S, et al. Development of pre and post-operative models to predict early recurrence of hepatocellular carcinoma after surgical resection. *J Hepatol*. 2018;69(6):1284–1293. doi:10.1016/j.jhep.2018.08.027
9. Ji GW, Zhu FP, Xu Q, et al. Radiomic features at contrast-enhanced CT predict recurrence in early stage hepatocellular carcinoma: a Multi-Institutional Study. *Radiology*. 2020;294(3):568–579. doi:10.1148/radiol.2020191470
10. Shim JH, Jun MJ, Han S, et al. Prognostic nomograms for prediction of recurrence and survival after curative liver resection for hepatocellular carcinoma. *Ann Surg*. 2015;261(5):939–946. doi:10.1097/SLA.0000000000000747
11. Deo RC. Machine learning in medicine. *Circulation*. 2015;132 (20):1920–1930. doi:10.1161/CIRCULATIONAHA.115.001593
12. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347–1358. doi:10.1056/NEJMra1814259
13. Eaton JE, Vesterhus M, McCauley BM, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. *Hepatology*. 2020;71(1):214–224. doi:10.1002/hep.30085
14. Nathan H, Hyder O, Mayo SC, et al. Surgical therapy for early hepatocellular carcinoma in the modern era: a 10-year SEER-medicare analysis. *Ann Surg*. 2013;258(6):1022–1027. doi:10.1097/SLA.0b013e31827da749
15. Fritz AG. *International Classification of Diseases for Oncology: ICD-O. 3*. Geneva, Switzerland: World Health Organization; 2000.
16. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–1232. doi:10.1214/aos/1013203451
17. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26 (6):565–574. doi:10.1177/0272989X06295361
18. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res*. 2004;10(21):7252–7259. doi:10.1158/1078-0432.CCR-04-0713
19. Altekruse SF, Henley SJ, Cucinelli JE, McGlynn KA. Changing hepatocellular carcinoma incidence and liver cancer mortality rates in the United States. *Am J Gastroenterol*. 2014;109(4):542–553. doi:10.1038/ajg.2014.11
20. Dasari BV, Kamarajah SK, Hodson J, et al. Development and validation of a risk score to predict the overall survival following surgical resection of hepatocellular carcinoma in non-cirrhotic liver. *HPB (Oxford)*. 2020;22(3):383–390. doi:10.1016/j.hpb.2019.07.007
21. Ferrer-Fàbrega J, Forner A, Liccioni A, et al. Prospective validation of ab initio liver transplantation in hepatocellular carcinoma upon detection of risk factors for recurrence after resection. *Hepatology*. 2016;63(3):839–849. doi:10.1002/hep.28339
22. Qiu J, Peng B, Tang Y, et al. CpG methylation signature predicts recurrence in early-stage hepatocellular carcinoma: results from a Multicenter Study. *J Clin Oncol*. 2017;35(7):734–742. doi:10.1200/JCO.2016.68.2153
23. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393. doi:10.1136/bmj.b2393

24. Loftus TJ, Tighe PJ, Filiberto AC, et al. Artificial intelligence and surgical decision-making. *JAMA Surg*. 2020;155(2):148–158. doi:10.1001/jamasurg.2019.4917

25. Shindoh J, Andreou A, Aloia TA, et al. Microvascular invasion does not predict long-term survival in hepatocellular carcinoma up to 2 cm: reappraisal of the staging system for solitary tumors. *Ann Surg Oncol*. 2013;20(4):1223–1229. doi:10.1245/s10434-012-2739-y

26. Bibault JE, Chang DT, Xing L. Development and validation of a model to predict survival in colorectal cancer using a gradient-boosted machine. *Gut*. 2021;70(5):884–889. doi:10.1136/gutjnl-2020-321799

27. Maros ME, Capper D, Jones DTW, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat Protoc*. 2020;15(2):479–512. doi:10.1038/s41596-019-0251-6

28. Jeong CW, Washington SL 3rd, Herlemann A, Gomez SL, Carroll PR, Cooperberg MR. The new surveillance, epidemiology, and end results prostate with watchful waiting database: opportunities and limitations. *Eur Urol*. 2020;78(3):335–344. doi:10.1016/j.eururo.2020.01.009