Cover page

Trial Title:

A New Conception About Individualized Treatment Allocation for HCC—Using Machine Learning

Unit of Study: The Second Affiliated Hospital of Air Force Military Medical University

Principal Investigator：Lei Liu

Date: 2023-8-20

NCT number and document date have not been assigned

Study protocols

1. Study population

Between January 2010 and October 2021, a total of 5816 consecutive HCC patients receiving LR or TACE from 15 Chinese tertiary hospitals were retrospectively screened. HCC was diagnosed based on either imaging or pathological results as previously described [1]. To better fit LR and TACE therapies into AI models, liver-confined HCC was incorporated with LR or TACE candidates according to the latest criteria of the BCLC guidelines [2]. Patients with well-preserved liver function, good PS (Eastern Cooperative Oncology Group [ECOG] PS0), and no vascular invasion or extrahepatic spread were divided into BCLC 0, A, and B stages, where they could undergo LR and TACE therapy. In addition, mild tumor-related symptoms (PS1) have not been an absolute contraindication for LR as per recent research [3], and the guidelines of the American Association for the Study of Liver Diseases (AASLD) expanded the PS 0–1 in the BCLC 0–B stage, considering the conceptual overlap between them [4, 5]. Here, the inclusion criteria were as follows: (a) diagnosed with HCC; (b) receiving LR or TACE therapy; (c) complete clinical information; (d) preserved liver function (Child-Pugh Score [CPS] ≤ 7); and (e) PS 0–1. The exclusion criteria included: (a) vascular invasion or extrahepatic spread; (b) decompensated liver cirrhosis (gastrointestinal bleeding, ascites, or hepatic encephalopathy); (c) younger than 18 years; and (d) duration of follow-up of fewer than 30 days. Image assessment, including the diameter of the largest nodule (tumor size, measured in centimeters) and tumor number, was performed by two independent investigators via CT or MRI. The flow chart of the study is depicted in Figure 1. Eligible patients were randomly stratified into train and validation cohorts (ratio=0.85: 0.15). The study was approved by the committee on human research of participating centers and approved by the Medical Ethics Committee of the Tangdu hospital (TDLL-202302-06) and Daping Hospital of the Army Medical University (2022(186)). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

2. Treatment procedures

During the LR procedure, patients were placed in the supine position, often under general anesthesia of the trachea. Operators fully exposed the surgical field and explored the location of the tumor and vascular anatomy. After blocking the first liver portal, operators ligated the liver section blood vessels and bile ducts, assessed the distance between lesions, and ensured residual liver volume. Multiple tumors were removed by en bloc resection or multiple LR with good liver function, including indocyanine green (ICG) clearance tests [6].

During the TACE procedure, an emulsion of mixtures of lipiodol (3–30 ml) and chemotherapeutic drugs was injected, including doxorubicin (10–50 mg), cisplatin (10–110 mg), epirubicin (10–50 mg), or oxaliplatin (100–200 mg). Afterward, either a gelatin sponge or polyvinyl alcohol foam (PVA) particles were introduced to embolize tumor-feeding vessels either selectively or super-selectively [7].

3. Model development

According to survival algorithms in the existing literature, 10 kinds of available AI models were developed in the training set. The details of data preprocessing and model development are shown in Supplement 1. To reduce the overfitting of survival models, the AI algorithms in the training cohort were developed by adopting K-fold cross-validation. In each fold, hyperparameter optimization was tuned by an exhaustive cross-validation grid search, and the models were refitted in the training set with the optimal score parameters. The selection criterion of the AI models was defined as a C-index ≥ 0.7. Based on the AI models, the predictive performance (censored C-index and mean cumulative/dynamic area under the receiver operating characteristic curve [auROC]), calibration curves, and Brier score (BS) were evaluated. The BS represents the mean square difference between observed patient status and predicted survival probabilities; it ranges from 0 to 1 and is negatively correlated with the discrimination and calibration of algorithms. IBS is defined as an overall estimation of BS; a value lower than 0.25 is deemed applicable in the clinic [8].

4. Model estimation

An ensemble model was developed and named the precise treatment allocation (PTA) model in our study. It could accurately predict the survival probability of each patient by alternating the different treatments. The inputs of the PTA model mainly comprise four modules: (a) liver function, such as serum albumin, serum bilirubin, and coagulative function indexes; (b) tumor burden, including serum alpha-fetoprotein (AFP), tumor size, and tumor number; (c) demographic data, such as age and gender; and (d) routine laboratory indications, such as routine blood examination and kidney function indexes. To further estimate the predictive performance of the PTA model, it was compared with other prognostic models, including widespread clinical guidelines (BCLC, BCLC B sub-classification [9], HKLC, and the most recent Italian Liver Cancer [ITA.LI.CA] staging system [10]), other tumor-burden-based models (up-to-seven model [11], four-and-seven model [12], and six-and-twelve model [1]), and models incorporating other parameters (hepatoma arterialembolization prognostic [HAP] score, modified HAP [mHAP] II score [13], mHAP III score [14], albumin-bilirubin [ALBI] score [15], and the model to estimate survival in ambulatory HCC patients [MESIAH] [16]), by estimating the auROC and C-index.

Here, HCC patients could achieve two potential survival rates and predicted median survival times (MSTs) by choosing TACE or LR. The ideal LR candidate is defined as a patient whose MST could be extended by more than 50% compared with the MST of primary treatment if LR is adopted in the model. Similarly, ideal TACE candidates are those with an additional survival benefit of more than 50% compared to LR allocation in TACE prediction time. Considering that BCLC and HKLC [17] have clearly defined LR and TACE candidates, the ideal LR and TACE candidates of the PTA model were compared with candidates recommended by these two widely used guidelines.

5. Web implementation

After train and validation, the optimized PTA model was implemented as a web interface, which generates the MST and depict integrated survival curves composed of LR and TACE according to the clinic parameters of specific patients provided by users. Relative algorithms of web development were described in the supplement 1.

6. Statistical analyses

Continuous and normally distributed variables were expressed as mean with a standard difference (SD) and compared using Student's t-test, while non-normally distributed variables were presented as median with interquartile range (IQR) and compared using the non-parametric Mann–Whitney U test. Categorical variables were expressed as numbers and percentages and tested by Fisher's exact test or the χ^2 test.

Overall survival (OS) was defined as the time interval between initial TACE or LR and all-cause death. Patients who survived up to the last follow-up date (October 15th, 2021) or were lost to follow-up were censored. Survival curves were displayed using the Kaplan–Meier (K-M) method and compared with the log-rank test.

Python software was applied for train AI models and detailed codes were uploaded as supplements (https://github.com/ApocalypsezZ/Survival). R software and related packages (survival, rms, reportROC, survminer, and ggplot2) were adopted to conduct statistical analysis and plot figures. $P < 0.05$ was considered statistically significant.

Results information for studies

Basic clinical characteristics

In this multicenter study, 4991 patients met the selection criteria and were included. Of these, 1916 patients (38.39%) were amenable to LR while 3075 patients (61.61%) submitted to TACE (Figure 1). 4192 (84%) and 799 (16%) patients were divided into train and validation sets, respectively. Among train set, 1622 (38.7%) were underwent LR while 2570 (61.3%) adopted TACE treatment. In validation set, 294 (36.8%) and 505 (63.2%) were allocated to LR and TACE treatment, respectively. The demographic and clinicopathological characteristics of the cohort are shown in Table 1. The variables associated to tumor burden were lower in the LR group than those of the TACE group in overall cohort: (a) The largest tumor diameters were smaller in the

LR group (4.60 [3.00;7.00] cm) than in the TACE group (6.30 [4.00;9.90] cm); (b) The tumor lesions of HCC were relatively few in the LR group (1.0 [1.0-1.0] lesion) than in the TACE group (1.0 [1.0-2.0] lesions). In addition, the liver function of HCC patients (CPS and Child-Pugh stage) varied remarkably between the LR and TACE groups. A higher proportion of Child stage B patients were included in the TACE group (17.6%) than in the LR group (11.1%). The median follow-up was 54.0 months (IQR 38.7–69.2 months) for LR therapy and 36.1 months (IQR 24.7–54.7 months) for TACE. The 1-year, 3-year, and 5-year survival probability was 91.7%, 79.6%, and 70.8% for LR allocation, respectively, while 74.8%, 41.5%, and 26.4% for TACE treatment, respectively. Although there were significant differences between the LR and TACE groups, we perceived therapy type as a pivotal factor in the AI models to simulate a real-world clinical setting, considering the precise prediction of AI technologies.

Selection of the optimal AI model

We applied 10 kinds of machine learning survival algorithms in the train cohort, of which seven AI models performed well, with the C-index > 0.7, and were retained: (i) random survival forest (RSF), (ii) conditional survival forest, (iii) Deepsurv algorithm , (iv) Extra Survival Trees, (v) Gompertz, (vi) standard Cox proportional hazards model (CoxPH), and (vii) PTA model based on the above six models. For model estimation, the C-index and IBS of these seven AI models were displayed in the Table 2. The C-index of the PTA model was 0.854 and 0.801 in the train and validation cohorts, respectively, which were higher than other six AI models except for the RFS model in train cohort. Even though the RSF model achieved the highest C-index (0.926) in the train set, while in the validation cohort, the C-index was only 0.765, indicating the potential presence of over-fitting in the RSF model. Given the C-index of the train and validation cohorts, the PTA model derived from the above six algorithms outperformed other single-algorithm models. As expected, the IBS of the six single algorithms and PTA algorithm was far smaller than 0.25, demonstrating their capability for clinical application (Supplement 2-3).

Predictive performance of the PTA model

To further validate the PTA model, we drew the receiver operating characteristic curves (ROC) and calculated the AUC. The AUC values for 1 year, 3 years, and 5 years were 0.912 (95% confidence interval [CI], 0.887–0.938), 0.914 (95% CI, 0.891–0.937), and 0.928 (95% CI, 0.903–0.953), respectively (Supplement 4A-C and Supplement 5). In addition, the AUCs for TACE treatment were 0.893 (95% CI, 0.858–0.928), 0.857 (95% CI, 0.810–0.904), and 0.858 (95% CI, 0.787–0.929), whereas the AUCs for LR were 0.873 (95% CI, 0.805–0.941), 0.850 (95% CI, 0.797–0.903), and 0.839 (95% CI, 0.779–0.899) at 1, 3, and 5 years, respectively (Supplement 5). The full summary of all indicators (accuracy, sensitivity, specificity, positive predictive value [PPV], negative predictive value [NPV], positive likelihood ratio [PLR], and negative likelihood ratio [NLR]) were compiled in Supplement 5. These results suggested that the PTA model could accurately predict HCC survival at the observed follow-up time.

Additionally, the performance and discriminative ability of clinical guidelines (BCLC, BCLC B sub-classification, HKLC, and ITA.LI.CA stage), tumor-burden-based models (up-to-seven model, four-and-seven model, and six-and-twelve model), and models adopting other parameters (HAP score, mHAP II score, mHAP III score, ALBI score, and MESIAH score) were compared with the PTA model (Table 2). As expected, the 1-, 3- and 5-year auROC and C-index of the PTA model were higher than those of the above mentioned models and prognostic scores. Moreover, three calibration curves were plotted to show the predicted probabilities of death caused by HCC in 1, 3, and 5 years (Supplement 4D-F). The observed survival was within a negligible margin of error of the predicted survival, showing that the performance of the PTA model was overwhelmingly good.

Comparison with clinical guidelines

Next, we compared the discriminative performance of the PTA model in selecting potential candidates with that of the BCLC and HKLC guidelines. As stated above, the LR and TACE candidates were determined following the BCLC and HKCL guidelines. Among the patients who underwent LR therapy (Supplement 6A-C), the PTA model identified 1386 potential LR candidates, BCLC guidelines selected 1196 cases, and HKLC chose 1765 candidate patients. K-M cures clearly showed that the survival difference between recommended and non-recommended populations was the largest in the PTA model compared with other guidelines. Meanwhile, the hazard ratio (HR) calculated by Cox regression analysis was 10.6, 2.02, and 1.93 respectively in the PTA model, BCLC, and HKLC guidelines (non-recommended as reference). Likewise, among the patients who accepted TACE, the PTA model, BCLC, and HKLC guidelines discriminated 323, 964, and 521 cases, respectively. The HR for the PTA model, BCLC, and HKLC guidelines was 25.4, 0.901, and 0.503, respectively (Supplement 6D-F).

Moreover, the LR candidates recommended by the PTA model had a prolonged OS compared to those suggested by BCLC and HKLC (Figure 2A); a similar conclusion was reached for the TACE group (Figure 2B). Intriguingly, the PTA model also found potential allocation “error” populations in the TACE and LR groups, depicted as hypothetical K-M curves (Figure 3A-B). The MST predicted by the PTA model was significantly higher compared with that of the original therapy allocation, suggesting that alternative therapies can enable patients to achieve a better prognosis and that mistaken allocation either to LR or TACE is still abundant. The patients with actual treatment and LR or TACE candidates are clearly shown in Figure 3C. Regarding the original LR allocation, 1458 cases had the right allocation, while 458 “model-found” cases were misled and should have been given TACE therapy. Similarly, there were abundant erroneous allocations among the original TACE allocations (2752, 89.5%).

Given that BCLC, the HKLC staging system, and PS could affect the prognosis of the potential benefit-gained population, we stratified patients into different BCLC subgroups (BCLC 0–A, BCLC B, and BCLC C stages; Supplement 7), HKLC subgroups (HKLC I–II and HKLC III; Supplement 8), and different PS (PS0 and PS1; Supplement 9). The PTA model could accurately select benefit-gained patients for LR and TACE across different BCLC and HKLC stages and PS. Collectively, the K-M curves and HR consistently showed that the PTA algorithm could more effectively select TACE and LR candidates than the two authoritative therapeutic guidelines.

Web development and clinical application

To make an easy-to-apply tool and visualize the individual predictive results, we implemented the PTA models as a web interface, which could be freely accessed from www.pta4hcc.com. Users may input four groups of parameters: (a) demographic information, such as age and gender; (b) liver function; (c) tumor burden files; and (d) routine laboratory results like kidney function. The PTA model calculates the survival rate for each month when adopting different treatments (now including LR and TACE) and generates MST, all of which can be displayed on the user-friendly interface. For example, a 65-year-old man with hepatitis B virus infection and no prior treatment would possess a better prognosis if LR therapy was adopted instead of TACE after the diagnosis of HCC. His tumor burden characteristics are as follows: the number of tumors is four, the maximum tumor size is 8 cm, the AFP is 300 μg/L, and the ECOG or PS is 0. In addition, the information on the other two groups is also input into the website in the complete version. The output is his predicted survival curves derived from the PTA model, showing that LR's MST is 78 months, while TACE's MST is only 23 months (Supplement 10).

In parallel with the complete PTA algorithm, a concise version including the nine most important features was also developed on this website. This simple AI version could also calculate the integrated survival model immediately upon the input of these important features, especially for cases where some above-stated module features were unavailable. For example, a patient with two tumor lesions, a tumor size of 12 cm, 100 μg/L AFP, good PS, 120 g/L HGB, 35 U/L AST, 40 g/L ALB, and normal INR would benefit from TACE therapy rather than LR. The simulative survival curve indicates that the MST was 15 and 10 months for TACE and LR allocation, respectively (Supplement 11).

[1] WANG Q, XIA D, BAI W, et al. Development of a prognostic score for recommended TACE candidates with hepatocellular carcinoma: A multicentre observational study [J]. J Hepatol, 2019, 70(5): 893-903.

[2] REIG M, FORNER A, RIMOLA J, et al. BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update [J]. J Hepatol, 2022, 76(3): 681-93.

[3] WU H, XING H, LIANG L, et al. Real-world role of performance status in surgical resection for hepatocellular carcinoma: A multicenter study [J]. Eur J Surg Oncol, 2019, 45(12): 2360-8.

[4] SCHNADIG I D, FROMME E K, LOPRINZI C L, et al. Patient-physician disagreement regarding performance status is associated with worse survivorship in patients with advanced cancer [J]. Cancer, 2008, 113(8): 2205-14.

[5] MARRERO J A, KULIK L M, SIRLIN C B, et al. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases [J]. Hepatology, 2018, 68(2): 723-50.

[6] MAKUUCHI M, KOKUDO N, ARII S, et al. Development of evidence-based clinical guidelines for the diagnosis and treatment of hepatocellular carcinoma in Japan [J]. Hepatol Res, 2008, 38(1): 37-51.

[7] MATSUI O, KADOYA M, YOSHIKAWA J, et al. Small hepatocellular carcinoma: treatment with subsegmental transcatheter arterial embolization [J]. Radiology, 1993, 188(1): 79-83.

[8] ADEOYE J, HUI L, KOOHI-MOGHADAM M, et al. Comparison of time-to-event machine learning models in predicting oral cavity cancer prognosis [J]. Int J Med Inform, 2022, 157(104635.

[9] BOLONDI L, BURROUGHS A, DUFOUR J F, et al. Heterogeneity of patients with intermediate (BCLC B) Hepatocellular Carcinoma: proposal for a subclassification to facilitate treatment decisions [J]. Semin Liver Dis, 2012, 32(4): 348-59.

[10]FARINATI F, VITALE A, SPOLVERATO G, et al. Development and Validation of a New Prognostic System for Patients with Hepatocellular Carcinoma [J]. PLoS Med, 2016, 13(4): e1002006.

[11]MAZZAFERRO V, LLOVET J M, MICELI R, et al. Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond the Milan criteria: a retrospective, exploratory analysis [J]. Lancet Oncol, 2009, 10(1): 35-43.

[12]YAMAKADO K, MIYAYAMA S, HIROTA S, et al. Subgrouping of intermediate-stage (BCLC stage B) hepatocellular carcinoma based on tumor number and size and Child-Pugh grade correlated with prognosis after transarterial chemoembolization [J]. Jpn J Radiol, 2014, 32(5): 260-5.

[13]PARK Y, KIM S U, KIM B K, et al. Addition of tumor multiplicity improves the prognostic performance of the hepatoma arterial-embolization prognostic score [J]. Liver Int, 2016, 36(1): 100-7.

[14]CAPPELLI A, CUCCHETTI A, CABIBBO G, et al. Refining prognosis after trans-arterial chemo-embolization for hepatocellular carcinoma [J]. Liver Int, 2016, 36(5): 729-36.

[15] PINATO D J, SHARMA R, ALLARA E, et al. The ALBI grade provides objective hepatic reserve estimation across each BCLC stage of hepatocellular carcinoma [J]. J Hepatol, 2017, 66(2): 338-46.

[16] YANG J D, KIM W R, PARK K W, et al. Model to estimate survival in ambulatory patients with hepatocellular carcinoma [J]. Hepatology, 2012, 56(2): 614-21.

[17] YAU T, TANG V Y, YAO T J, et al. Development of Hong Kong Liver Cancer staging system with treatment stratification for patients with hepatocellular carcinoma [J]. Gastroenterology, 2014, 146(7): 1691-700.e3.