Research Title: Construction and Clinical Validation of a Predictive Model for Postoperative Adjuvant Therapy in Hepatocellular Carcinoma Based on Whole-Slide Digital Pathological Images and Deep Learning

NCT Number: NCT07417800

Document date: 2025/10/24

Research Protocol

Study Background

Hepatocellular Carcinoma (HCC) is one of the most common malignant tumors worldwide, ranking sixth in incidence and third in mortality globally, causing approximately 480,000 deaths each year. China accounts for more than 45% of global HCC cases, resulting in an extremely heavy disease burden. Curative surgical resection is the primary approach to achieve long-term survival for patients with early-stage HCC. However, the postoperative recurrence rate reaches 50%–70% within five years after surgery, which seriously impairs patient prognosis. Postoperative adjuvant therapy has become a key strategy to delay tumor recurrence and improve survival outcomes. Transarterial Chemoembolization (TACE) and Tyrosine Kinase Inhibitors (TKIs), such as sorafenib and lenvatinib, are widely applied for patients at high recurrence risk after surgery.

Nevertheless, both treatment modalities have notable limitations. The objective response rate of TACE is merely 50%–60%; a considerable number of patients fail to benefit from this therapy and may suffer from treatment-induced liver function injury. Although TKIs can prolong recurrence-free survival (RFS) by 3 to 5 months in high-risk postoperative HCC patients, the treatment response rate is less than 20% in unselected populations. Additionally, the incidence of Grade 3–4 adverse events including hypertension, hand-foot skin reaction and proteinuria exceeds 50%, leading to treatment discontinuation in around 20% of patients due to intolerable toxicities. Currently, there is a lack of efficient and reliable biomarker systems to identify patients who may benefit from treatment. Clinical treatment decisions still rely on empirical judgment based on conventional pathological features such as tumor size and vascular invasion, resulting in insufficient individualized treatment, wasted medical resources and additional treatment burdens on patients.

Recent studies have demonstrated that the Tumor Immune Microenvironment (TIME) is a fundamental biological factor affecting the sensitivity to TACE and TKIs in HCC patients. Preoperative contrast-enhanced CT/MRI can directly reflect tumor vascularity, boundary definition, peritumoral infiltration and satellite lesions. These imaging characteristics are closely correlated with postoperative pathological findings including Microvascular Invasion (MVI) and tumor differentiation, and further determine responses to adjuvant therapy. For example, preoperative MRI radiomic features such as textual heterogeneity and irregular tumor margins have been proven to be significantly associated with postoperative HCC recurrence risk. Moreover, CT perfusion parameters can predict the degree of tumor necrosis after TACE treatment. Furthermore, the combination of preoperative imaging and postoperative pathological images enables the construction of tumor spatiotemporal evolution models to reveal dynamic biological changes before and after therapeutic intervention. Therefore, multimodal data integrating preoperative imaging, postoperative pathological images, pathological reports and clinical indicators can break the limitations of single-source data and substantially improve the accuracy of adjuvant therapy response prediction.

Breakthroughs in artificial intelligence, especially deep learning technology, have opened up a new avenue to explore in-depth pathological information from routine Hematoxylin and Eosin (H&E) stained Whole Slide Images (WSIs). As conventional postoperative pathological materials, H&E WSIs are generated for all HCC patients after surgery without additional sampling or testing. The cellular and histological details contained in WSIs can effectively reflect core characteristics of the TIME. Specifically, Convolutional Neural Networks (CNN) and Vision Transformer (ViT)

architectures can automatically identify morphological patterns related to $CD8^+$ T cell infiltration density and PD-L1 expression, realizing cross-modal prediction of immune status based on H&E morphological features. Latest studies have verified that deep learning models based on H&E WSIs achieve high accuracy in predicting Microsatellite Instability (MSI) in colorectal cancer (AUC = 0.88), PD-L1 expression in non-small cell lung cancer (AUC = 0.80) and Tumor Mutation Burden (TMB) (AUC = 0.91). In the field of HCC, WSI-based deep learning models have also achieved favorable performance in predicting postoperative recurrence risk (AUC = 0.82) and immune cell infiltration (AUC = 0.78). However, no studies have focused on the critical clinical issue of predicting responses to postoperative adjuvant therapy (TACE/TKIs) by integrating preoperative imaging and postoperative multimodal pathological data, and large-scale multicenter prospective clinical validation is still lacking.

Accordingly, this study aims to integrate multicenter clinicopathological data with artificial intelligence algorithms to establish a multimodal predictive model based on preoperative imaging, postoperative pathological WSIs, pathological reports and clinical indicators. It intends to clarify the correlation between preoperative tumor characteristics and responses to postoperative adjuvant therapy, develop clinically applicable digital decision-making tools, and promote the multi-dimensional and full-cycle precise management of HCC.

Study Objectives

This study aims to construct and validate a multimodal data platform by integrating preoperative contrast-enhanced CT/MRI images, postoperative H&E-stained Whole Slide Images (WSIs), pathological reports and clinical indicators collected from HCC patients across multiple centers. Radiomic feature extraction and deep learning algorithms will be adopted to develop a joint predictive model, so as to accurately identify HCC subtypes sensitive to TACE and TKIs preoperatively and quantitatively evaluate treatment efficacy. Meanwhile, this study intends to elucidate the spatiotemporal biological evolution patterns of tumors during postoperative adjuvant therapy, including vascular remodeling and dynamic changes of the tumor immune microenvironment. The predictive performance of the model will be validated in prospective cohorts, and a visual decision support system will be developed to optimize individualized postoperative adjuvant therapy strategies, reduce the rate of ineffective treatment, and advance the multi-dimensional and full-cycle precise treatment and management of hepatocellular carcinoma.

Study Design and Methods

3.1 Trial Design and Study Population

This is a comprehensive study combining diagnostic trials and interventional studies, divided into two major parts and multiple phases, aiming to develop and validate an artificial intelligence-based prediction system for postoperative adjuvant therapy of hepatocellular carcinoma (HCC). The overall study complies with the development specifications of Software as a Medical Device (SaMD) based on artificial intelligence, including retrospective data collection, model development and training, retrospective validation, prospective observational validation and final prospective interventional studies.

Part One: Retrospective Study on the Construction of Tumor Immune Microenvironment Identification and Adjuvant Therapy Prediction Model

(1) Construction and Validation of Tumor Immune Microenvironment Identification Model

Development and Training Phase

Public datasets (e.g., TCGA) and retrospective datasets from our center (n=1500, 400 cases collected so far) will be used as the training set. The core task of this phase is to develop a deep learning model (architectures such as Vision Transformer or ResNet) to automatically identify, segment and quantify key features of the tumor immune microenvironment from H&E-stained digital pathological Whole Slide Images (WSIs), including but not limited to the degree of immune cell infiltration, stroma ratio, vascular invasion and necrotic areas. Meanwhile, preoperative contrast-enhanced CT/MRI images will be integrated. Radiomic feature extraction techniques (such as texture analysis and morphological feature analysis) and deep learning algorithms will be applied to explore tumor vascularity, morphological features of tumor margins and peritumoral infiltration information, and conduct multimodal correlation analysis with pathological features. All images will be independently annotated by at least two senior physicians in a blinded manner. Any discrepancies will be resolved through consensus negotiation or arbitration by a third senior physician to form the gold standard.

External Validation Phase

Retrospective WSI and imaging datasets from 8 to 10 collaborative medical centers across China will be used to preliminarily validate the generalization ability of the model. This phase aims to evaluate the robustness of the model under different scanning devices, pathological slide preparation and staining procedures, and interpretation habits of pathologists.

(2) Construction and Multi-level Clinical Validation of Treatment Response Prediction Model

This is the core part of the study. Based on the quantified tumor immune microenvironment features output by the above model, combined with imaging data and essential clinical variables (such as TNM stage, Child-Pugh liver function classification and AFP level), a classification model will be built to predict patients' responses to four types of postoperative adjuvant therapy regimens: surgery alone, surgery combined with TACE, surgery combined with TACE plus systemic therapy, and surgery combined with systemic therapy alone. A strict hierarchical validation strategy will be adopted in this part.

Model Training Phase

Retrospective data from our center (n=1500) will be used. Inclusion criteria for patients in this phase: ① Histopathologically confirmed HCC; ② Underwent curative surgical resection; ③ Complete preoperative contrast-enhanced CT/MRI images, records of postoperative adjuvant therapy and follow-up data are available. Definition of treatment response: The primary endpoint is Relapse-Free Survival (RFS), and secondary endpoints include Overall Survival (OS) and Objective Response Rate (ORR). Multi-category machine learning algorithms (such as Gradient Boosting Decision Tree and Deep Neural Network) will be used for model training to output the most beneficial treatment regimen for each patient.

Internal Validation Phase

An independent validation set (n≈500, no overlap with the training set) will be randomly divided from the data of our center for preliminary evaluation of the model's discriminative performance and hyperparameter tuning to prevent overfitting.

External Validation Phase 1 (Retrospective Validation and Model Calibration)

Retrospective data (n=3000) will be collected from 8 to 10 medical centers nationwide. The main purposes of this phase are: ① To validate the model performance on larger and more heterogeneous external datasets; ② To calibrate or slightly optimize the model according to

validation results and finalize the predictive model.

External Validation Phase 2 (Large-scale Retrospective Validation)

The finalized model will be tested on a brand-new large-scale retrospective dataset (n=10,000) from 3 to 5 medical centers to further enhance the level of evidence and provide a solid basis for prospective studies.

(3) Study Population

The study population consists of patients who underwent curative surgical resection for HCC. Multiple datasets will be established for model development, training and validation, and all datasets will be constructed in accordance with strict inclusion and exclusion criteria.

(1) Patient Inclusion and Exclusion Criteria

Inclusion Criteria

Histopathologically confirmed diagnosis of hepatocellular carcinoma;

Aged between 18 and 75 years old;

Underwent curative resection (R0 resection) for primary hepatocellular carcinoma;

Complete postoperative formalin-fixed paraffin-embedded (FFPE) H&E-stained tissue slides eligible for digital scanning;

Complete and accessible clinicopathological data and follow-up data, including but not limited to age, gender, HBsAg status, liver cirrhosis background, AFP and PIVKA-II levels, tumor size and number, CNLC stage, AJCC 8th TNM stage, BCLC stage, vascular invasion status, tumor differentiation grade, detailed postoperative adjuvant therapy regimens and medication records, as well as postoperative tumor recurrence and survival data;

Preoperative contrast-enhanced CT or MRI images with standardized scanning parameters and no severe artifacts, meeting the requirements for analytical use.

Exclusion Criteria

Received anti-tumor treatments such as TACE, targeted therapy, immunotherapy and radiotherapy at non-collaborative medical centers before surgery, with unavailable original imaging and clinical baseline data;

Combined with primary malignant tumors of other organs;

Positive surgical margins after surgery (R1 or R2 resection);

Poor quality of pathological tissue slides (such as severe fading, folding, breakage or insufficient tissue volume), which cannot be used for digital scanning or reliable analysis;

Severe missing of clinical or follow-up data;

Missing preoperative imaging data or unqualified imaging quality (such as severe artifacts or incomplete scanning sequences) that cannot be used for analysis.

(2) Definition and Source of Datasets

A multi-center and multi-stage data collection strategy will be adopted for all datasets, which are strictly divided according to purposes to ensure mutual independence of the training set, validation set and test set and avoid data leakage.

A total of 1,500 eligible surgical patients from our center will be enrolled and randomly divided into the model training set (n=1000) and internal validation set (n=500) at a ratio of 2:1, for model training, preliminary validation and optimization respectively. A total of 400 eligible patients have been enrolled so far, and the remaining 1,100 patients will be continuously recruited from the historical medical records of our center according to the inclusion and exclusion criteria.

A multi-center collaboration model will be adopted for external validation. ① External Validation

Set 1: A total of 3,000 retrospective cases from 8 to 10 domestic medical centers will be included for multi-center retrospective validation and model calibration, so as to verify the model's generalization ability across different institutions and conduct calibration or limited iterative optimization on the initial model. ② External Validation Set 2: A total of 10,000 retrospective cases from 3 to 5 large domestic medical centers will be included for the final test of the finalized model to fully evaluate its effectiveness and stability. Only forward validation will be performed on this dataset using the locked model without any parameter adjustment.

Part Two: Clinical Validation of Hepatocellular Carcinoma Adjuvant Therapy Prediction Model - Prospective Study

(1) Prospective Observational Study

A total of 1,000 eligible postoperative HCC patients will be consecutively enrolled from 10 to 15 medical centers. Preoperative imaging data, postoperative WSIs and clinical data will be collected and imported into the model to obtain predicted treatment regimens. No intervention will be implemented on actual clinical decisions in this phase. Researchers will record model prediction results and actual clinical decisions. Patients will be divided into the prediction-consistent cohort and prediction-inconsistent cohort according to whether the actual clinical decisions are consistent with model predictions. Long-term follow-up will be conducted to observe patients' actual prognosis, so as to evaluate the consistency between model predictions and actual outcomes, as well as the model's performance degradation in real-world settings.

(2) Study Population

The study population consists of patients who underwent curative surgical resection for HCC. Multiple datasets will be established in the prospective study for model optimization and validation, and all datasets will be constructed in accordance with strict inclusion and exclusion criteria.

(1) Patient Inclusion and Exclusion Criteria

Inclusion Criteria

Histopathologically confirmed diagnosis of hepatocellular carcinoma;

Aged between 18 and 75 years old;

Underwent curative resection (R0 resection) for primary hepatocellular carcinoma;

Complete postoperative formalin-fixed paraffin-embedded (FFPE) H&E-stained tissue slides eligible for digital scanning;

Complete and accessible preoperative CT/MRI images, clinicopathological data and follow-up data, including but not limited to age, gender, HBsAg status, liver cirrhosis background, AFP and PIVKA-II levels, tumor size and number, CNLC stage, AJCC 8th TNM stage, BCLC stage, vascular invasion status, tumor differentiation grade, detailed postoperative adjuvant therapy regimens and medication records, as well as postoperative tumor recurrence and survival data;

Signed the informed consent form for the prospective study.

Exclusion Criteria

Received anti-tumor treatments such as TACE, targeted therapy, immunotherapy and radiotherapy at non-collaborative medical centers before surgery, with unavailable original imaging and clinical baseline data;

Combined with primary malignant tumors of other organs;

Positive surgical margins after surgery (R1 or R2 resection);

Poor quality of pathological tissue slides (such as severe fading, folding, breakage or insufficient

tissue volume), which cannot be used for digital scanning or reliable analysis;
Severe missing of clinical or follow-up data.

(2) Definition and Source of Datasets

A multi-center and multi-stage data collection strategy will be adopted for all datasets, which are strictly divided according to purposes to ensure mutual independence of the training set, validation set and test set and avoid data leakage.

A multi-center collaboration model will be adopted for the prospective cohort. ① Prospective Observational Cohort: A total of 1,000 patients will be consecutively enrolled from 10 to 15 medical centers. Pathological slides will be collected for digital scanning, and preoperative imaging data will be imported into the model. No clinical intervention will be conducted; only model predictions and actual clinical outcomes will be recorded to evaluate the model's performance in real clinical settings. ② Prospective Interventional Cohort: A randomized controlled trial will be carried out in 3 to 5 medical centers with a planned enrollment of 600 patients. After obtaining patients' informed consent, patients will be randomly assigned to the intervention group (AI model-assisted decision-making) and the control group (standard clinical decision-making). Physicians in the intervention group will formulate treatment plans with reference to the optimal treatment regimens predicted by the AI model; the control group will receive routine treatment. Clinical outcomes will be compared between the two groups to provide high-level evidence for the clinical utility of the model.

3.2 Deployment and Intervention of AI System

Aiming to solve the clinical problems of high heterogeneity in postoperative treatment response and difficulties in efficacy prediction for hepatocellular carcinoma, this study intends to construct the Hepatocellular Mixture-of-Experts (HepMoE) multimodal intelligent decision-making system. By fusing multimodal data including pathological imaging and clinical data, this system realizes individualized efficacy prediction and intelligent support for precise treatment decision-making. Centered on digital whole-slide pathology images (multiple stains of H&E and IHC: CK7, CK19, Ki-67), combined with structured clinical information (such as TNM stage, Child-Pugh classification and AFP level), this system innovatively incorporates medical imaging data (contrast-enhanced CT/MRI). Relying on the Mixture-of-Experts framework driven by pathological large models, a multimodal deep learning model is constructed to provide an intelligent solution for postoperative adjuvant therapy of hepatocellular carcinoma.

Equipped with the latest multimodal model capabilities, the HepMoE system fully leverages the achievements of self-supervised representation learning on large-scale whole-slide image corpora. Through transfer learning and fine-tuning mechanisms, HepMoE can efficiently extract cellular and histological phenotypic features from H&E and IHC stained images, and capture tumor morphology, tumor immune microenvironment and molecular phenotypic heterogeneity.

The model adopts Stain-specific Expert Networks to decouple features and conduct fine-grained modeling for different staining channels (H&E, CK7, CK19, Ki-67). A Gated Fusion Module dynamically integrates multimodal inputs to balance the contribution of morphological and immune features. CT/MRI images will be standardized (e.g., pixel normalization), tumor lesions will be automatically segmented using 3D U-Net or ViT models, and radiomic features (such as texture, shape and intensity features) and deep learning features will be extracted. On the basis of clinical variables (such as TNM stage, Child-Pugh classification and AFP level) and textual

pathological reports, individualized treatment benefit assessment and prediction of adjuvant therapy response are realized. Finally, HepMoE will form an integrated precise treatment auxiliary system combining pathological large model representation, interpretable decision-making of imaging models and clinical integration, providing a promotable intelligent decision-making platform for postoperative adjuvant therapy of hepatocellular carcinoma.

3.3 Evaluation Index System (Algorithm Performance, Clinical Utility, System Stability)

This study comprehensively evaluates the overall value of the AI model from three dimensions: algorithm performance, clinical utility and system stability.

Evaluation of Algorithm Performance

Conducted mainly during model development and validation, quantitative indicators are used to objectively assess the accuracy and robustness of the tumor immune microenvironment identification model and treatment response prediction model.

For the tumor immune microenvironment identification model: At the pixel level, Dice Similarity Coefficient (DSC) is used to evaluate the coincidence degree between segmentation results and expert annotations, and Intersection over Union (IoU) is adopted as a supplementary indicator. At the regional level, Precision, Recall and F1-Score are used to evaluate the recognition performance of specific components in the immune microenvironment.

For the treatment response prediction model: Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used to evaluate the binary classification ability of the model, and macro-average and micro-average AUC values are calculated. Accuracy, Precision, Recall and F1-Score calculated based on the confusion matrix are used to evaluate the performance of multi-classification tasks. Calibration Curve and Brier Score are used to assess the accuracy of predicted probabilities.

Evaluation of Clinical Utility

Conducted in the prospective study phase, focusing on evaluating the actual impact of the model on clinical decision-making and patient prognosis.

Primary clinical endpoints: Relapse-Free Survival (RFS), defined as the time from randomization to tumor recurrence or death from any cause. The Kaplan-Meier method is used for survival analysis, the Log-rank test is adopted for inter-group comparison, and the Hazard Ratio (HR) and its 95% confidence interval are calculated. Overall Survival (OS), defined as the time from randomization to death from any cause, is analyzed using the same statistical methods.

Secondary clinical endpoints: Treatment decision change rate, calculated as the proportion of cases where AI recommendations are inconsistent with physicians' initial treatment plans. Physician adherence, defined as the proportion of physicians in the intervention group who finally follow AI recommendations. Decision Curve Analysis (DCA) is used to evaluate the clinical net benefit of the model, and standardized net benefit values under different threshold probabilities are calculated. A 1–5 point Likert scale is used to assess changes in physicians' decision confidence.

Evaluation of System Stability

Focusing on the operational performance and reliability of the AI system in real clinical settings.

Technical performance indicators: Inference speed, recording the average time and standard deviation for processing a single whole-slide image. System availability rate, calculated as: Availability Rate = (1 - Downtime / Total Operating Time) × 100%. Concurrent processing capability, testing the system performance when processing multiple requests simultaneously. API response time, monitoring the communication delay of the interface connected to the hospital

information system.

Usability evaluation: The System Usability Scale (SUS) is adopted, which contains 10 items with a 5-point Likert scale. The total score is converted into a percentage for evaluation. Task completion rate test is conducted to record the success rate of users completing specific operations within the specified time.

Reliability evaluation: Long-term monitoring of relevant indicators including Mean Time Between Failures (MTBF) (the average time for continuous normal operation of the system) and Mean Time To Repair (MTTR) (the average time for system recovery after failure). Error rate statistics are performed to classify and record various system errors and their occurrence frequencies.

All evaluation results will be summarized into a comprehensive evaluation report to support model optimization and clinical application. Data collected during the evaluation will be stored in a dedicated database to facilitate subsequent in-depth analysis. A systematic evaluation system ensures that the artificial intelligence model meets clinical application standards in terms of accuracy, practicality and reliability.

3.4 Sample Size Calculation (Description of Parameter Sources and Calculation Tools for Sample Size)

This study includes multiple phases of retrospective model development and prospective clinical validation with different research objectives, so the basis and methods for sample size estimation vary in each phase. This section elaborates the basis, parameter sources and methods for sample size calculation of each phase, and clearly presents key calculation formulas.

3.4.1 Sample Size for Retrospective Study Phase

(1) Sample Size for Model Training and Validation: Training Set (n = 1,000) and Internal Validation Set (n = 500)

Estimation Basis: The sample size of this phase follows empirical rules of machine learning and feasibility principles. Complex deep learning models for WSIs require a large amount of annotated data to avoid overfitting. Empirical experience shows that 1,000 to 5,000 samples are generally required for each category. This study targets four categories of treatment response classification. A sample size of 1,000 is set to balance model performance and data collection feasibility, which is consistent with the sample size of existing top-tier multimodal AI studies. The size of the internal validation set generally accounts for 20%–30% of the training set. A total of 500 samples can estimate model performance indicators (such as AUC) with a small standard error and meet the requirements of hyperparameter tuning.

(2) External Validation Set 1 (n = 3,000)

Estimation Basis: This phase is used for preliminary evaluation of model calibration and generalization ability, and a sufficient sample size is required to detect clinically significant declines in model performance.

a sample size of 3,000 provides sufficient statistical power to support subgroup analysis and subsequent fine-tuning of the model.

(3) External Validation Set 2 (n = 10,000)

Estimation Basis: This phase is for the final test of the locked model, and a high-precision estimation of performance indicators (such as AUC) is required. The sample size is calculated reversely according to the width of the AUC confidence interval.

Given AUC = 0.80 and the target 95% confidence interval width $\leq$ 0.04 (i.e., $\pm$0.02),

approximately 8,600 cases are required according to calculations using PASS software. A total of 10,000 cases are set to reserve sufficient margin and ensure high credibility of the results.

3.4.2 Sample Size for Prospective Study Phase

(1) Sample Size for Prospective Observational Validation (n = 1,000)

Estimation Basis: This phase evaluates the correlation strength between model predictions and real clinical prognosis (C-index) using sample size formulas for survival analysis. Relapse-Free Survival (RFS) is taken as the primary endpoint.

Assuming the 12-month RFS rate of the control group is 60% and the corresponding HR of the model-consistent group is 0.7, a total of 256 events are expected. Simulated by PASS 2021 software, more than 300 events can be observed in 1,000 cases within a 24-month study period (12 months for enrollment + 12 months for follow-up), which meets the requirements of statistical power and supports stratified analysis.

(2) Sample Size for Prospective Interventional Study

Estimation Basis: This phase is a Randomized Controlled Trial (RCT) with strictly calculated sample size, taking RFS as the primary endpoint.

Sample size formula for Log-Rank test is adopted with the following parameter settings: Baseline event rate: The 12-month RFS rate of the standard treatment group (Arm B) is 65%. Effect size: The 12-month RFS rate of the AI-assisted decision-making group (Arm A) is increased by an absolute value of 10% (up to 75%), corresponding to HR = 0.67. Significance level (α): 0.05. Statistical power (1-β): 80% (conventional standard for RCT). Accrual time: 18 months. Follow-up time: 12 months. Dropout rate: 10%.

Calculated by PASS 2021 software, a total of 200 events are required. Considering the dropout rate, the total required sample size is 600 cases (300 cases in each group). This sample size can detect the clinical difference corresponding to HR = 0.67 and meet the power requirements under follow-up and dropout conditions.

3.5 Bias Control and Fairness

A multi-level and systematic strategy is adopted to strictly control various potential biases in the study and ensure the scientificity and reliability of research results.

Bias control in study design phase: Standardized inclusion and exclusion criteria are formulated to control selection bias. In multi-center studies, all centers are required to adopt a unified patient screening process. Selection bias is assessed by comparing the basic characteristics of enrolled and unenrolled patients. For the prospective interventional study, central randomization (block randomization stratified by center) is used to assign patients to the AI-assisted decision-making group or standard treatment group to ensure balanced baseline characteristics between groups.

Bias control in data collection and processing phase: Strict blinding principles are implemented to reduce information bias and measurement bias. For pathological image annotation, at least two senior pathologists conduct independent annotations without access to any clinical outcomes of patients, and discrepancies are arbitrated by a third expert to establish high-quality gold standards. During model development, algorithm developers are prohibited from accessing data of subsequent validation sets to avoid overfitting caused by manual tuning. An independent clinical endpoint adjudication committee is established in the endpoint evaluation phase. Committee members are blinded to patient grouping information and AI prediction results, and judge tumor recurrence and survival events objectively according to established criteria to minimize assessment bias.

Bias control in algorithm development and validation phase for AI-specific biases: For spectrum bias, slide images scanned by different devices and stained with different procedures from multiple medical centers are collected. Color normalization techniques (such as the Macenko method) are adopted in the preprocessing phase to reduce the impact of staining differences. Multi-center and large-sample external validation is conducted to evaluate and correct the performance differences of the model across different populations and devices so as to control measurement bias. In addition to dividing independent validation sets, regularization and Dropout technologies are applied to prevent overfitting. Cross-validation and Bootstrap sampling results are used when evaluating model performance.

Fairness assessment: Comprehensive fairness evaluation is carried out to analyze whether there are significant differences in model performance across subgroups divided by demographic characteristics (gender, age) and clinical characteristics (tumor stage, etiology, liver function classification). Core evaluation indicators include inter-group AUC difference, balanced odds and fairness metrics for sensitive attributes. If performance discrimination of the model is found in specific subgroups, algorithm correction will be performed to ensure the fairness of algorithm decisions. All fairness evaluation results will be fully disclosed in the research report.

Bias control in statistical analysis phase: Multivariate Cox regression model and Logistic regression model are used to adjust for known prognostic factors (such as TNM stage, vascular invasion and AFP level) to accurately estimate the independent predictive efficacy of the AI model and control potential residual confounding bias.

3.6 Data Governance System (Data Collection, Transmission, Storage Security Control; Full-chain Data Recording)

A sound data governance system is established to ensure the quality, security and compliance of data throughout the research process, covering all links including data collection, transmission, storage, processing, analysis and archiving, forming a complete data management closed loop.

Standardized data collection: Detailed Standard Operating Procedures (SOPs) for data collection are formulated to clarify the definition, format and collection specifications of all types of data. Structured electronic Case Report Forms (eCRFs) are used for clinical data collection with built-in logic verification and range check functions to ensure the accuracy and completeness of data collection. Pathological image data collection follows digital scanning standards with a unified scanning resolution of 0.25μm/pixel and file format of SVS. Equipment calibration and quality control tests are conducted before each batch of scanning. Laboratory test data are directly collected by docking with the Laboratory Information System (LIS) to avoid manual transcription errors.

Security control for data transmission: End-to-end encryption technology and TLS 1.3 protocol are adopted for all data transmission to secure the transmission channel. A complete data transmission log system is established to record the transmission time, content, sender and receiver. Desensitization is performed on data before cross-institutional transmission: direct identifiers are removed and indirect identifiers are generalized. Data integrity verification and digital signature technology are applied during transmission to prevent data tampering.

Data storage management: A hierarchical storage strategy is implemented. Raw data are stored on internal servers of each institution with physical isolation and access control. Analytical data are stored on a dedicated research platform with encrypted storage technology. A complete data backup and recovery mechanism is established: incremental backup is performed daily and full

backup is performed weekly, with backup data stored off-site. The data storage system has passed ISO 27001 certification, and regular security vulnerability scanning and penetration testing are carried out.

Quality control in data processing: A comprehensive data quality evaluation indicator system covering integrity, accuracy, consistency and timeliness is established. Automated data quality inspection tools are used for regular data quality assessment. Once data quality problems are identified, root cause analysis is conducted and corrective and preventive actions are implemented. A complete audit log is recorded for all data processing operations to ensure traceability.

Data archiving and destruction: All research data are archived in accordance with standard specifications after the completion of the study, including raw data, processed data and final analysis data. A strict data access permission management system is established to control the scope of data use. Physical destruction or multiple overwriting technologies are adopted for data destruction to ensure data is irrecoverable. Detailed records are kept throughout the archiving and destruction process.

Data governance organizational structure: A data governance committee is set up to formulate data management policies. Dedicated data administrators are assigned to be responsible for daily data management. A data quality supervision team conducts regular data quality inspections. All researchers are required to receive and pass data management training to ensure the implementation of standardized data management.

3.7 Adverse Events (Specific Definition, Classification, Reporting Methods and Treatment Measures for Adverse Events/Serious Adverse Events)

Adverse events in this study refer to any negative medical outcomes related to AI-assisted decision-making, which are divided into three grades:

Grade 1 (Mild): AI recommendations lead to minor inconvenience or reversible additional examinations without long-term impacts;

Grade 2 (Moderate): AI recommendations result in inappropriate treatment regimens, which require medical intervention for correction without permanent damage;

Grade 3 (Severe): AI recommendations lead to misdiagnosis and mistreatment, causing permanent functional damage or significant shortening of survival time.

Once an adverse event occurs, the clinical center shall report it to the research safety monitoring committee through the online system within 24 hours. The committee consists of experts in hepatology, oncology, pathology and medical ethics. Treatment measures include immediately suspending the use of the AI system, organizing experts to assess the cause of the event, determining the correlation with the AI system, and formulating corrective measures. All adverse events will be fully recorded and included in the final research report.

3.8 Statistical Analysis Plan (Methods for Missing Data Processing, Definition and Selection of Statistical Analysis Datasets, Statistical Methods for Comparing Primary and Secondary Outcome Indicators, Algorithm Stability Evaluation)

R (Version 4.2.0), Python (Version 3.9) and SPSS (Version 26.0) software will be used for all statistical analyses. All statistical tests are two-sided with a significance level of $\alpha=0.05$.

Missing data processing: Multiple imputation method is adopted. For continuous variables, the Multiple Imputation by Chained Equations (MICE) is used to create 5 imputed datasets for combined analysis. Mode imputation is applied for categorical variables. The missing data mechanism will be tested: Complete Case Analysis is adopted for data Missing Completely at

Random (MCAR); multiple imputation is used for data Missing at Random (MAR); selection models are applied for data Missing Not at Random (MNAR). Sensitivity analysis is conducted to compare results obtained by different missing data processing methods.

Definition of statistical analysis datasets:

Full Analysis Set (FAS): Includes all randomized patients who received at least one efficacy assessment, analyzed according to the Intention-To-Treat (ITT) principle.

Per-Protocol Set (PPS): Includes patients who fully complied with the research protocol, excluding patients with major protocol violations.

Safety Set (SS): Includes all patients who received at least one treatment, used for safety evaluation.

Primary analysis is based on the FAS, and the PPS is used for sensitivity analysis to verify the robustness of results.

Statistical methods for primary outcome indicators: For categorical variables (such as treatment response grouping), weighted F1-score and macro-average AUC are taken as main evaluation indicators, and Cohen's Kappa coefficient is used to assess the consistency between model predictions and actual clinical conditions. For survival data (RFS and OS), the Kaplan-Meier method is used to draw survival curves, the Log-rank test is used for inter-group comparison, and the Cox proportional hazards regression model is established to calculate the Hazard Ratio (HR) and its 95% confidence interval while adjusting for potential confounding factors such as age, gender and TNM stage.

Statistical methods for secondary outcome indicators: Decision Curve Analysis (DCA) is used to evaluate the clinical net benefit of the model. Calibration Curve and Brier Score are used to assess the accuracy of predicted probabilities. Repeated-measures Analysis of Variance is used for inter-group comparison of continuous variables. The Mann-Whitney U test is adopted for ordinal variables such as physician adherence and decision confidence scale scores.

Algorithm stability evaluation: Bootstrap sampling (1,000 repetitions) is used to calculate the 95% confidence interval of model performance indicators for stability assessment. SHapley Additive exPlanations (SHAP) values are used to analyze feature importance and improve model interpretability. Intraclass Correlation Coefficient (ICC) is calculated to evaluate the consistency of the model across different centers; ICC > 0.75 is regarded as good consistency.

Subgroup analysis and sensitivity analysis: Subgroup analysis is performed according to key clinical characteristics including age (≤60 years vs >60 years), gender, TNM stage (Stage I-II vs Stage III-IV) and degree of liver cirrhosis. Interaction test is used to evaluate subgroup differences. Sensitivity analysis is conducted by changing statistical model parameters and adopting different missing data processing methods to verify the robustness of research results.

Sample size re-estimation plan: If the deviation between the event rate of primary outcome indicators observed in interim analysis and the initial assumption exceeds 20%, the Lan-DeMets method combined with O'Brien-Fleming boundaries will be used for sample size re-estimation. Re-estimation is limited to once and must be approved by the Data Safety Monitoring Board (DSMB).

All statistical analyses are conducted in accordance with the pre-specified statistical analysis plan. All post-hoc analyses are clearly marked as exploratory. Point estimates and 95% confidence intervals are reported for all statistical results, and p-values are retained to three decimal places. Conclusions will not be drawn based solely on p-values.

3.9 Full-process Quality Control of Project Data (Data Entry, Coding and Storage Methods, Measures to Improve Data Quality; Measures to Ensure Subject Compliance in Clinical Observation and Follow-up; Measures to Ensure Consistency Across Multi-center Studies)

A comprehensive and systematic data quality management system is established covering the whole process of data collection, transmission, storage, processing and analysis to ensure the integrity, accuracy, consistency and reliability of research data.

Quality control for data collection: Unified Standard Operating Procedures (SOPs) for data collection are formulated to clarify the definition, collection time and recording standards of all observation indicators. An Electronic Data Capture (EDC) system is used for data entry with built-in logic verification and range check functions to identify abnormal values and inconsistent data in real time. Dual independent data entry is adopted for clinical data, and discrepancies are verified against original documents by a third party. Pathological image data collection follows standardized procedures. Calibrated scanning devices are used, and quality control tests are conducted before each batch of scanning to ensure consistent image resolution and color.

Security control for data transmission and storage: Encrypted protocols (SSL/TLS) are adopted for data transmission. A hierarchical management strategy is implemented for data storage: raw data are stored on internal servers of hospitals, and desensitized data for analysis are stored on a secure cloud platform. A complete data backup mechanism is established: incremental backup is performed daily and full backup is performed weekly, with backup data stored off-site. A complete log is recorded for all data operations to ensure traceability.

Data governance and multi-level quality audit mechanism: A data quality management team is set up to conduct regular random inspections on data quality and generate monthly data quality reports. A data quality evaluation indicator system including data completeness rate, consistency rate and timeliness rate is established. Root cause analysis and Corrective and Preventive Actions (CAPA) are implemented once data quality problems are identified. The entire data cleaning process is fully recorded with all data modification traces retained.

Quality control for clinical observation and follow-up: Detailed SOPs for clinical observation and follow-up are formulated to unify observation indicators, measurement methods and recording standards. An electronic follow-up system is applied to automatically remind follow-up time points and reduce the loss to follow-up rate. Regular training and quality assessment are conducted for researchers to ensure standardized operations. An endpoint event adjudication committee is established to judge tumor recurrence and survival status in a blinded manner uniformly and ensure the objectivity of endpoint data.

Quality control for multi-center studies: Unified training is organized for researchers from all centers before the study starts to ensure consistent understanding of the research protocol. A centralized monitoring plan is formulated with regular on-site and remote monitoring. Standardized data collection tools and unified data definitions are used to ensure data consistency across all centers. Regular coordination meetings are held to solve problems encountered in data collection in a timely manner. Centralized random audit is implemented to ensure the consistent implementation of the research protocol in all centers.

Data security management: A data classification and grading management system is established, and encrypted storage and strict access control are implemented for sensitive data. All researchers receive data security and privacy protection training and sign confidentiality agreements. Regular

security vulnerability scanning and risk assessment are carried out to eliminate potential security hazards. A data security emergency response plan is formulated with a corresponding emergency response mechanism.

The above comprehensive and systematic quality control measures ensure the authenticity, accuracy and integrity of research data and provide a solid guarantee for the reliability of research conclusions. All quality control activities are fully recorded to form a complete quality assurance document system.

7.1 Common Risks of Interventional Therapy

Postembolization syndrome: fever, pain, nausea and vomiting;

Liver function injury: transient transaminitis, jaundice and ascites; severe cases may lead to liver failure. Embolization-related complications: cholecystitis, gastroduodenal erosion or ulcer, splenic infarction, etc.;

Puncture-related complications: hematoma at the puncture site, pseudoaneurysm and vascular injury;

Others: renal function injury, allergic reaction, infection, etc.

7.2 Common Risks of Targeted Drug Therapy

Common adverse reactions: hypertension, hand-foot skin reaction (redness, desquamation and pain of hands and feet), diarrhea, decreased appetite, fatigue, weight loss, abnormal thyroid function, proteinuria, etc.;

Hematological toxicity: thrombocytopenia, leukopenia and increased bleeding risk;

Severe but rare adverse reactions: cardiac dysfunction, arterial thromboembolism, gastrointestinal perforation, impaired wound healing, etc.

7.3 Preventive Measures

Standard operating procedures will be strictly followed. The physical condition of subjects will be comprehensively evaluated before treatment. Subjects will be closely monitored during and after treatment. Routine blood tests, liver and kidney function tests, electrocardiograms and other examinations will be performed regularly to detect and manage any possible adverse reactions in a timely manner. Complete emergency plans and rescue measures are prepared, and detailed guidelines for adverse reaction management (such as management of hypertension and skin reactions) are provided for researchers.

In case of severe intolerable adverse events (SAE), the research physician will reduce the drug dose, suspend or permanently terminate the study treatment in accordance with the protocol, and provide necessary medical support for the subjects. All adverse events will be recorded in the Case Report Form. The Ethics Committee, clinical trial institution and sponsor shall be notified immediately. Relevant national and provincial drug regulatory authorities shall be reported within 24 hours. All SAEs must be reported through the internal hospital adverse event and near-miss no-fault reporting system.

Informed Consent for Human Research, The Second Affiliated Hospital of Zhejiang University School of Medicine

Dear Patient:

You are invited to participate in a multi-center, randomized, controlled and prospective clinical study entitled Construction and Clinical Validation of a Postoperative Adjuvant Therapy Prediction Model for Hepatocellular Carcinoma Based on Digital Whole Slide Pathology Images and Deep Learning. Before you decide whether to participate, please read the following content carefully to help you understand this study, its purpose, procedures, duration, potential benefits, risks and inconveniences of participation.

The introduction of this study is as follows:

I. Study Background and Objectives

Liver cancer is a common and severe disease. Surgical resection is the main treatment for early-stage liver cancer, but the postoperative recurrence rate is relatively high. To reduce the risk of recurrence, doctors usually recommend postoperative adjuvant therapy, mainly including local treatment (TACE) and oral targeted drugs (TKIs, such as sorafenib and lenvatinib).

However, these treatments are not equally effective for all patients. Some patients may gain limited therapeutic effects or suffer from side effects which force them to discontinue treatment. At present, doctors mainly rely on clinical experience to judge the most suitable treatment for each patient, lacking more accurate prediction methods.

Recent studies have found that the tumor immune microenvironment affects the efficacy of the above treatments. Traditional methods for detecting the tumor immune microenvironment are complicated and costly. Now we intend to adopt artificial intelligence technology to analyze routine pathological slides (H&E stained slides) made after surgery, so as to find clues that can predict the efficacy of TACE or TKIs. Preoperative imaging (such as contrast-enhanced CT/MRI) can directly reflect tumor vascularity, boundary definition, peritumoral infiltration and satellite lesions. These imaging characteristics are closely correlated with postoperative pathological findings including microvascular invasion and tumor differentiation, and further affect responses to adjuvant therapy.

The objective of this study is to establish an artificial intelligence prediction model. If successful, it may help doctors select more appropriate postoperative adjuvant therapy regimens for patients in the future, making treatment more effective and individualized.

II. Specific Procedures and Processes

This is a comprehensive study combining diagnostic trials and interventional studies, divided into two major parts and multiple phases. The prospective trial is the main part in this stage, aiming to validate the clinical application efficacy of the liver cancer adjuvant therapy prediction model.

Prospective Observational Study: A total of 1,000 eligible postoperative HCC patients will be consecutively enrolled from 10 to 15 medical centers. WSIs and preoperative CT/MRI images will be collected and imported into the model to obtain predicted treatment regimens. No intervention will be implemented on actual clinical decisions in this stage. Researchers will record model prediction results and actual clinical decisions. Long-term follow-up will be conducted to observe patients' actual prognosis, so as to evaluate the consistency between model predictions and actual outcomes, as well as the model's performance degradation in real-world settings.

III. What You Need to Do If You Participate in This Study

Fully understand the potential translational value and significance of this study;

Agree to allow the study team to obtain all your previous medical documents, examination results and pathological slides generated in this hospital;

Cooperate with doctors for follow-up and receive relevant examinations;

Receive postoperative adjuvant therapy recommended by doctors, including interventional therapy, targeted and immunotherapy, or regular follow-up without adjuvant therapy after surgery.

IV. Potential Benefits of Participating in This Study

Standardized long-term follow-up management: Participants will receive unified and standardized postoperative follow-up monitoring, which helps to detect potential postoperative tumor recurrence at an early stage, conduct early intervention in a timely manner and fully guarantee the effect of postoperative rehabilitation.

No additional medical burden or physical trauma: This is a purely observational study. No additional puncture, blood collection, imaging examination or surgical intervention will be performed. Only your existing surgical pathological data, imaging data and clinical data will be used for research analysis. No extra examination fees, treatment fees or physical trauma will be incurred.

Contribute to the development of precise treatment for liver cancer: Your clinical data will help build an intelligent prediction system for postoperative adjuvant therapy of liver cancer, enabling more individualized precise treatment for more liver cancer patients in the future, reducing ineffective treatment and toxic and side effects, and bringing important clinical value to the progress of liver cancer diagnosis and treatment. The pathological tissues and clinical data you contribute will help us verify this innovative technology, which is expected to promote the progress of postoperative adjuvant therapy for hepatocellular carcinoma and improve the postoperative survival of liver cancer patients.

V. Potential Adverse Reactions, Risks and Preventive Measures

This is a prospective observational study. Potential adverse reactions are mainly related to the corresponding treatments, including interventional therapy (e.g., Transarterial Chemoembolization, TACE) and targeted drugs (e.g., lenvatinib, sorafenib, regorafenib).

Common Risks of Interventional Therapy

Postembolization syndrome: fever, pain, nausea and vomiting;

Liver function injury: transient transaminitis, jaundice and ascites; severe cases may lead to liver failure. Embolization-related complications: cholecystitis, gastroduodenal erosion or ulcer, splenic infarction, etc.;

Puncture-related complications: hematoma at the puncture site, pseudoaneurysm and vascular injury;

Others: renal function injury, allergic reaction, infection, etc.

Common Risks of Targeted Drug Therapy

Common adverse reactions: hypertension, hand-foot skin reaction (redness, desquamation and pain of hands and feet), diarrhea, decreased appetite, fatigue, weight loss, abnormal thyroid function, proteinuria, etc.;

Hematological toxicity: thrombocytopenia, leukopenia and increased bleeding risk;

Severe but rare adverse reactions: cardiac dysfunction, arterial thromboembolism, gastrointestinal perforation, impaired wound healing, etc.

Preventive Measures

Standard operating procedures will be strictly followed. Your physical condition will be

comprehensively evaluated before treatment. You will be closely monitored during and after treatment. Routine blood tests, liver and kidney function tests, electrocardiograms and other examinations will be performed regularly to detect and manage any possible adverse reactions in a timely manner. Complete emergency plans and rescue measures are prepared, and detailed guidelines for adverse reaction management (such as management of hypertension and skin reactions) are available for medical staff.

If you suffer from severe intolerable toxic and side effects, the research physician will reduce the drug dose, suspend or permanently terminate the study treatment in accordance with the protocol and provide you with necessary medical support.

VI. Cost Description

Costs Undertaken by the Study

Analysis fees for the research-related prediction model;

Fees for additional examinations specified in the study protocol beyond routine clinical diagnosis and treatment (such as blood tests or imaging examinations added for research purposes);

Fees for study drugs (if applicable) or study-specific operations.

Costs Undertaken by Yourself

Fees for radical resection of liver cancer (this is your standard treatment and will not be covered by the study);

All medical expenses incurred during adjuvant therapy (interventional or targeted therapy), including but not limited to hospitalization fees, surgical operation fees, targeted drug fees, auxiliary medication fees, routine laboratory tests and routine imaging examinations. These expenses shall be paid jointly by you and your medical insurance in accordance with conventional medical policies. The study will not cover the fees for your standard treatment.

VII. Compensation for Participation, Including Compensation for Research-related Injuries

Medical Treatment and Compensation for Research-related Injuries

If you suffer from adverse events or injuries directly caused by this study rather than your own underlying diseases, the study sponsor will bear the medical expenses and provide corresponding compensation for damages related to the study. You will not waive any legitimate rights and interests. Even if you have signed this informed consent form, you still have the right to seek compensation through legal channels. For detailed terms of injury compensation, application procedures and insurance information, please consult your research physician or nurse for relevant explanatory documents.

VIII. Alternative Treatment Options

If you refuse to participate in this study, there is no alternative research-related arrangement, and your routine clinical treatment will not be affected in any way.

IX. Confidentiality of Your Personal Information

Your medical records (including research medical records and laboratory test reports) will be kept in the hospital in accordance with regulations. Researchers, members of the Ethics Committee, monitors, auditors, drug regulatory authorities and other relevant personnel are allowed to access your medical records. Other personnel unrelated to the study have no right to access your medical records without permission. Your personal identity will not be disclosed in any published results of this study. We will make every effort to protect the privacy of your personal medical data within the scope permitted by laws and regulations.

X. Termination of Participation in the Study

Your participation in this study is completely voluntary. You may refuse to participate, or withdraw from the study at any time without giving any reason. This will not affect your relationship with doctors, your medical treatment or any other legitimate interests. In addition, your participation may be terminated for the following reasons:

You fail to follow the medical advice of the research physician;

You develop severe medical conditions requiring treatment;

The research physician believes that terminating your participation in the study is in your best health interests.

XI. Ethics Committee

This study has been reported to the Human Research Ethics Committee of The Second Affiliated Hospital of Zhejiang University School of Medicine. After comprehensive review including risk assessment for research subjects, the committee has approved this study. For issues related to ethics and your rights during the study, please contact the Human Research Ethics Committee of The Second Affiliated Hospital of Zhejiang University School of Medicine:

Tel: 0571-87783759 (Daytime); 13757118366 (On-duty at night)

Email: HREC2013@126.com

I confirm that I have read and fully understood this informed consent form. I voluntarily accept the treatment involved in this study and agree that my medical data can be used for the publication of this study.

Subject Signature: _____

Contact Information: _____

Date: _____

Authorized Representative Signature: _____

Relationship with the Subject: _____

Contact Information: _____

Date: _____

(Witness, if applicable): _____

Contact Information: _____

Date: _____

I confirm that I have explained the detailed content of this study to the patient, including their rights as well as potential benefits and risks. A copy of the signed informed consent form has been provided to the patient.

Researcher Signature: _____

Contact Information (Mobile Phone): _____

Date: _____