Received: 19 August 2019 | Revised: 20 December 2019 | Accepted: 8 January 2020

DOI: 10.1111/liv.14380

ORIGINAL ARTICLE

Liver INTERNATIONAL WILEY

Predicting survival after transarterial chemoembolization for hepatocellular carcinoma using a neural network: A Pilot Study

Aline Mähringer-Kunz[1] | Franziska Wagner[1] | Felix Hahn[1] | Arndt Weinmann[2,3] | Sebastian Brodehl[4] | Sebastian Schotten[1] | Jan B. Hinrichs[5] | Christoph Düber[1] | Peter R. Galle[2] | Daniel Pinto dos Santos[6] | Roman Kloeckner[1]

[1]Department of Diagnostic and Interventional Radiology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

[2]Department of Internal Medicine, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

[3]Clinical Registry Unit (CRU), University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

[4]Institute for Informatics, Johannes Gutenberg-University Mainz, Mainz, Germany

[5]Department of Interventional and Diagnostic Radiology, Hannover Medical School, Hanover, Germany

[6]Department of Radiology, University Hospital Cologne, Cologne, Germany

Correspondence
Roman Kloeckner, Department of Diagnostic and Interventional Radiology, Johannes Gutenberg-University Medical Center, 55131 Mainz, Germany.
Email: roman.kloeckner@unimedizin-mainz.de

Handling Editor: Alejando Forner

Abstract

Background and aims: Deciding when to repeat and when to stop transarterial chemoembolization (TACE) in patients with hepatocellular carcinoma (HCC) can be difficult even for experienced investigators. Our aim was to develop a survival prediction model for such patients undergoing TACE using novel machine learning algorithms and to compare it to conventional prediction scores, ART, ABCR and SNACOR.

Methods: For this retrospective analysis, 282 patients who underwent TACE for HCC at our tertiary referral centre between January 2005 and December 2017 were included in the final analysis. We built an artificial neural network (ANN) including all parameters used by the aforementioned risk scores and other clinically meaningful parameters. Following an 80:20 split, the first 225 patients were used for training; the more recently treated 20% were used for validation.

Results: The ANN had a promising performance at predicting 1-year survival, with an area under the ROC curve (AUC) of 0.77 ± 0.13. Internal validation yielded an AUC of 0.83 ± 0.06, a positive predictive value of 87.5% and a negative predictive value of 68.0%. The sensitivity was 77.8% and specificity 81.0%. In a head-to-head comparison, the ANN outperformed the aforementioned scoring systems, which yielded lower AUCs (SNACOR 0.73 ± 0.07, ABCR 0.70 ± 0.07 and ART 0.54 ± 0.08). This difference reached significance for ART ($P < .001$); for ABCR and SNACOR significance was not reached ($P = .143$ and $P = .201$).

Conclusions: Artificial neural networks could be better at predicting patient survival after TACE for HCC than traditional scoring systems. Once established, such prediction models could easily be deployed in clinical routine and help determine optimal patient care.

KEYWORDS

chemoembolization, diagnostic accuracy study, hepatocellular carcinoma, neural network

Abbreviations: AFP, alpha-fetoprotein; ANN, artificial neural network; AUC, area under the curve; BCLC, Barcelona Clinic Liver Cancer; CT, computed tomography; cTACE, conventional TACE; DEB-TACE, drug eluting bead TACE; EASL, European Association for the Study of the Liver; HCC, hepatocellular carcinoma; MELD, model for end-stage liver disease; MRI, magnetic resonance imaging; OS, overall survival; ROC, receiver operating characteristics; TACE, transarterial chemoembolization.

1 | INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide and the fourth leading cause of cancer death.[1,2] According to the Barcelona Clinic Liver Cancer (BCLC) classification, transarterial chemoembolization (TACE) is the recommended treatment for patients in intermediate-stage HCC (BCLC-B).[3] However, deciding when to repeat and when to cease TACE treatment and possibly change to systemic treatment, or even to best supportive care, can be difficult for even experienced investigators. Several conventional scoring systems have been developed to provide decision support regarding retreatment with TACE in patients with HCC, including the **A**ssessment for **R**etreatment with **T**ACE **(ART)** score,[4] the **ABCR** score[5] and the **SNACOR** score.[6] The ART (**ART**) score comprises the following parameters: increase of aspartate aminotransferase, an increase of Child-Pugh score from baseline and radiologic tumour response. The ABCR score consists of the following parameters: **a**lpha-fetoprotein **(A**FP) and **B**CLC at baseline, increase in **C**hild-Pugh score from baseline and radiological tumour **R**esponse. The SNACOR comprises the parameters tumour **S**ize, tumour **N**umber, baseline **A**FP, **C**hild-Pugh class and **O**bjective radiological **R**esponse. However, none of these scoring systems are widely used in clinical practice. Several attempts of external validation by us and other working groups have failed, with only poor to moderate predictive ability.[7-10]

In recent years, machine learning techniques—in particular artificial neural networks (ANNs)—have been increasingly used for prediction purposes. ANNs are complex and flexible nonlinear computing systems. They were devised in an attempt to build artificial systems based on the characteristics of neurones of the brain, both structurally and functionally. Such networks are trained in a supervised manner by exposure to paired input-output data; once trained they are able to make predictions based on new input data.[11] ANNs have shown promising results compared to conventional statistical approaches. In the field of hepatology, ANNs were superior in predicting mortality of patients with end-stage liver disease compared to model for end-stage liver disease (MELD) as well as in predicting HCC tumour grade and microvascular invasion compared to a conventional linear model.[12,13] Recently, Meek et al suggested stronger implementation of such techniques in interventional oncology.[14] Yet, only very few studies have used similar approaches in patients with HCC in the setting of TACE. Peng et al[15] applied a convolutional neural network to predict tumour response after first TACE based on pattern recognition in computed tomography (CT)-images. Abajian et al[16] analyzed a small number of clinical baseline parameters in combination with magnetic resonance imaging (MRI) parameters and used random forest models to predict tumour response.

To the best of our knowledge, no attempt has yet been made to develop a survival prediction model for patients with HCC undergoing TACE using neural networks. Therefore, the purpose of this study was to implement such a novel approach for treatment stratification and to compare it to conventional prediction scores.

Key points

- Predicting survival in patients with primary liver cancer is essential for deciding further treatment.
- Conventional scoring systems have remained behind expectations; thus, we used artificial intelligence to improve prediction.
- The artificial neural network we developed led to good prediction and outperformed the three most widely known conventional scoring systems.
- The difference reached significance in case of the ART score ($P < .001$); for ABCR and SNACOR significance was not reached ($P = .143$ and $P = .201$).

2 | PATIENTS AND METHODS

We used the TRIPOD guidelines when writing our manuscript.[17] The study was approved by the responsible ethics committee (permit number 2018-13619). Patient records and clinical information were de-identified before analysis.

2.1 | Patients

We performed a database search and identified a total of 860 patients who underwent TACE for HCC at our tertiary referral centre between January 2005 and December 2017. To ensure at least 1 year of follow-up, the final evaluation date was December 31, 2018. The study included only TACE-naïve patients with HCC confined to the liver who then underwent at least two TACE treatments. The study excluded patients who underwent liver transplantation within the follow-up period and patients who developed a portal venous tumour thrombus before the second TACE treatment. After applying these inclusion and exclusion criteria, 282 patients were included in the final analysis (Figure 1).

2.2 | Diagnosis, treatment and follow-up

Hepatocellular carcinoma was diagnosed by histological or radiological evaluation according to the guidelines of the American Association for the Study of Liver Diseases (AASLD) or the European Association for the Study of the Liver (EASL).[18,19] Treatment was performed in a standardized manner described in detail elsewhere.[20,21] All patients underwent CT or MRI prior to their first and second TACE treatment. These examinations were the basis for the radiological assessment of tumour response, which was evaluated by applying the Modified Response Evaluation Criteria in Solid Tumours.[22] Objective tumour response was defined as a partial response or complete response. Stable disease and progressive disease were assessed as a lack of radiological response. Overall survival (OS) was the primary endpoint

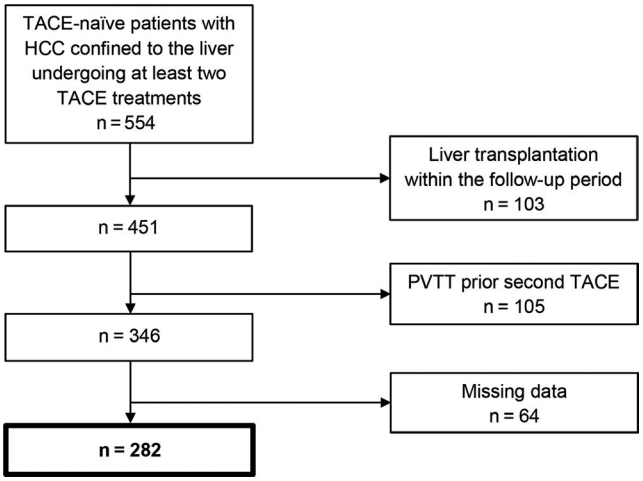

FIGURE 1 Flow diagram showing the reasons for dropout. The initial number of patients fulfilling the inclusion criteria was 554. After applying the exclusion criteria, the final number of patients included in the analysis was 282. Abbreviations: HCC, Hepatocellular carcinoma; PVTT, Portal vein tumour thrombosis; TACE, Transarterial chemoembolization

of this study. To ensure comparability with the conventional scoring systems, the interval was defined as the day prior to the second TACE treatment until death or last follow-up.

2.3 | Data acquisition

Data were acquired from the laboratory database and clinical registry software specially developed for the characterization of patients with HCC.[23] Baseline characteristics including demographic data, aetiology of liver disease, liver function parameters, TACE-related parameters and relevant comorbidities were documented. Tumour load was represented by the total number of lesions, the lesions' sizes and the tumour growth pattern.

2.4 | Design of the neural network

For the machine-learning algorithm, we decided on an ANN, more specifically a multilayer perceptron. The structure of a multilayer perceptron consists of an interconnected group of nodes arranged in multiple layers: an input layer, one or more hidden layers and an output layer. The input layer comprises of all input parameters; the output layer comprises of possible outcomes. The hidden layer(s) comprise(s) of hidden nodes, which connect input and output nodes and allow nonlinear interactions among the input variables. Hidden nodes do not have a real clinical correlate. The nodes are connected by links, each of which carries an associated weight. The network is trained by exposure to paired input-output data. The ANN learns through modification of these weights according to feedback. The final network applies these previously determined weights to new input data and thus makes predictions.[13,24,25]

Our ANN was built using Python 3.7.3 with scikit-learn (https://scikit-learn.org/stable/) (0.19.2). It consisted of 46 input nodes. During fine-tuning, we found that a network architecture with three fully connected hidden layers with sizes 20, 12 and 4 performed best. We used ReLU as activation function on all hidden layers and softmax classification for the final fully connected layer. We used stringent L2-regularization to prevent overfitting. The ANN was used to predict the OS after 1 year, starting from the day prior to the second TACE treatment. Therefore, the final two output nodes represented survival (=1) and death (=2).

Each of the 46 parameters formed one input node. The selected input data comprised the general demographic parameters age and gender, type of TACE (conventional TACE [cTACE], drug eluting bead TACE [DEB-TACE]), type of imaging before the second TACE (CT, MRI) and all parameters used by the above-mentioned risk scoring systems ART, ABCR and SNACOR. These scoring systems comprise the parameters BCLC stage, alpha-fetoprotein level, tumour size and number, Child-Pugh score, radiological tumour response and aspartate aminotransferase level. We also included other potentially clinically meaningful parameters regarding aetiology (alcohol abuse, hepatitis B and C, non-alcoholic steatohepatitis[26]), comorbidities (nicotine abuse,[27] obesity,[28] diabetes[26]) and tumour growth pattern.[29] In addition, we included the following potentially meaningful parameters indicating liver function: MELD,[30] bilirubin,[30,31] albumin[31] and the international normalized ratio.[30] TACE-related parameters (alanine aminotransferase[32]), other laboratory values (thrombocyte count,[33] sodium level[34]) and sarcopenia[28] were also evaluated. Sarcopenia was measured by means of the skeletal muscle index, which was calculated at the level of L3 as described elsewhere.[28] Measurements were performed using a dedicated Picture Archiving and Communication System (Sectra®, Linköping, Sweden). All parameters were captured before the first and second TACE treatment. All continuous input parameters were standardized by subtracting the mean and dividing by the standard deviation. The design and architecture of our ANN including an input layer, hidden layers and an output layer is provided in the Figure S1.

2.5 | Training and validation of the ANN

Using an 80:20 split, the first 225 of the total 282 patients treated were allocated to the diagnostic training dataset. This dataset was used for training and fine-tuning of the ANN. The more recently treated 20% comprised 57 patients and formed the holdout validation dataset.

In the diagnostic training dataset, a five-fold cross validation approach was used to maximize training capabilities: for each fold, one model was trained with a subset of 180 patients used for training and 45 patients for testing. The complete model is then constructed as the average of the five folds.

The holdout validation dataset was used for final evaluation: the patients in this set were never used for training or fine-tuning of the

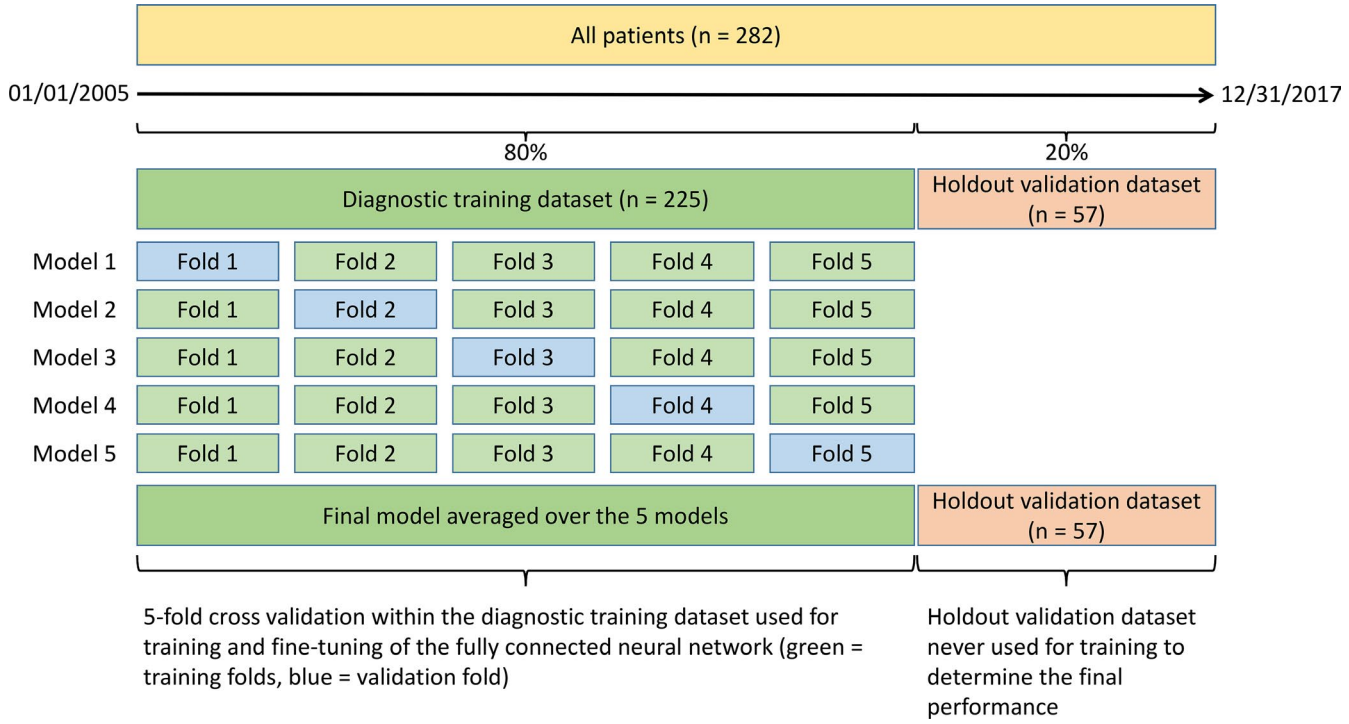

FIGURE 2 Diagram defining the different datasets and visualizing the process of building and validating the model. In the first step, 80% of the patients (n = 225) were allocated to the diagnostic training dataset. This dataset was used for training following a five-fold cross validation approach. In the second step, the more recently treated 20% of the patients (n = 57) were used for validation

model, but fed into the model once training was completed to enable head-to-head comparisons with existing scores (Figure 2).

Both groups were compared regarding the main baseline characteristics. Detailed data including the pertinent *P*-values are provided in Table 1.

Both groups were similar regarding baseline characteristics except for age ($P = .020$), albumin ($P = .019$), type of TACE ($P < .001$) and type of imaging prior to second TACE ($P < .001$).

2.6 | Statistical analysis

Continuous data were described by medians and ranges and compared using a two-tailed unpaired Student's *t* test or Wilcoxon test where appropriate. Categorical data were described as percentages and compared using the chi-squared test or Fisher's exact test. The predictive performance of the ANN was measured using the area under the receiver operating characteristic curve (AUC); the same approach was used to compare the ANN to ART, ABCR and SNACOR. The AUC ranges from 0 to 1, and values can be interpreted as follows: 0.9-1, 'excellent prediction'; 0.8-0.9, 'good prediction'; 0.7-0.8, 'fair prediction'; 0.6-0.7, 'poor prediction'; and 0.5-0.6, 'very poor prediction'.[35,36] A value <0.5 indicates 'anti-prediction'. Cumulative/dynamic receiver operating characteristic (ROC) curves were obtained using Python 3.7.3 with matplotlib 3.1.0 (http://matplotlib.org/). The prediction of the ANN was further expressed in sensitivity, specificity and positive and negative predictive values. A $P < .05$ was considered significant. As this analysis was intended to be exploratory, the *P*-values should be interpreted in a descriptive manner. R 3.5.2 and R 3.6.0 (A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, https://www.R-project.org; accessed 2018/2019) was used for statistical analyses.

3 | RESULTS

In the diagnostic training cohort, the 1-year OS was 71.5%, and in the holdout validation cohort, the 1-year OS was 63.1% ($P = .283$).

3.1 | Predictive performance of the neural network

For predicting 1-year survival, the ANN had a mean AUC in the diagnostic training cohort of 0.77 ± 0.13 (Figure 3).

These results were further verified in the holdout validation cohort, which had a mean AUC of 0.83 ± 0.06 (Figure 4), a positive predictive value of 87.5%, and a negative predictive value of 68.0%. The sensitivity was 77.8% and specificity 81.0%.

3.2 | Predictive performance of the neural network compared to conventional scoring systems

In the last step, the performance of the ANN was compared to the existing scoring systems ART, ABCR and SNACOR.[4-6] The AUCs were 0.54 ± 0.08, 0.70 ± 0.07 and 0.73 ± 0.07, respectively, and therefore,

TABLE 1 Baseline characteristics of patients in both groups

	Diagnostic training dataset (n = 225)	Holdout validation dataset (n = 57)	*P*-value
Age, years	66 (16-90)	68 (49-85)	.020
Gender			.809
Male	188 (83.6%)	49 (86.0%)	
Female	37 (16.4%)	8 (14.0%)	
BCLC stage			.139
A	37 (16.4%)	5 (8.8%)	
B	173 (76.9%)	51 (89.5%)	
C	11 (4.9%)	0 (0.0%)	
D	4 (1.8%)	1 (1.7%)	
Child-Pugh stage			.397
A	168 (74.7%)	43 (75.4%)	
B	53 (23.5%)	13 (22.8%)	
C	4 (1.8%)	1 (1.8%)	
Tumour size, mm	49 (10-270)	60 (10-180)	.293
Tumour number			.205
1	90 (40.0%)	23 (40.4%)	
2	34 (15.1%)	10 (17.5%)	
3	22 (9.8%)	11 (19.3%)	
4	24 (10.7%)	2 (3.5%)	
5	9 (4.0%)	2 (3.5%)	
6	9 (4.0%)	2 (3.5%)	
7	1 (0.4%)	2 (3.5%)	
8	2 (0.9%)	1 (1.8%)	
9	3 (1.3%)	0 (0.0%)	
≥10	31 (13.8%)	4 (7.0%)	
Diffuse tumour			.316
Yes	29 (12.9%)	4 (7.0%)	
No	196 (87.1%)	53 (93.0%)	
Alcohol[a]			.954
Yes	110 (48.9%)	27 (47.4%)	
No	115 (51.1%)	30 (52.6%)	
HBV[a]			.111
Yes	42 (18.7%)	5 (8.8%)	
No	183 (81.3%)	52 (91.2%)	
HCV[a]			.270
Yes	62 (27.6%)	11 (19.3%)	
No	163 (72.4%)	46 (80.7%)	
NASH[a]			.507
Yes	12 (5.3%)	5 (8.8%)	
No	213 (94.7%)	52 (91.2%)	
Unknown aetiology[a]			1.000
Yes	30 (13.3%)	9 (15.8%)	
No	195 (86.7%)	48 (84.2%)	
Nicotine abuse			.684
Yes	51 (22.7%)	15 (26.3%)	

(Continues)

TABLE 1 (Continued)

	Diagnostic training dataset (n = 225)	Holdout validation dataset (n = 57)	*P*-value
No	174 (77.3%)	42 (73.7%)	
Obesity			.339
Yes	89 (39.6%)	18 (31.6%)	
No	136 (60.4%)	39 (68.4%)	
Diabetes			.307
Yes	99 (44.0%)	30 (52.6%)	
No	126 (56.0%)	27 (47.4%)	
MELD score	10 (5-24)	10 (5-17)	.790
SMI	0.215 (0.125-0.324)	0.212 (0.139-0.341)	.843
AFP, ng/ml	16 (1-164852)	14 (1-26881)	.342
Bilirubin, mg/dl	1.05 (0.3-5.9)	1.04 (0.3-4.5)	.915
Albumin, g/L	35 (20-54)	34 (22-42)	.019
INR	1.1 (0.9-3.2)	1.2 (0.9-1.9)	.364
ALT, U/L	43 (4-477)	36 (12-245)	.283
AST, U/L	56 (16-358)	58 (23-187)	.997
Thrombocyte count, per nl	135 (34-720)	130 (34-458)	.345
Sodium, mmol/L	138 (124-147)	138 (131-142)	.373
TACE-type			<.001
cTACE	144 (64.0%)	8 (14.0%)	
DEB-TACE	81 (36.0%)	49 (86.0%)	
Previous therapy before first TACE			.772
Surgical resection	44 (19.6%)	12 (21.1%)	
Local ablation therapy	4 (1.8%)	0 (0.0%)	
Sorafenib therapy	3 (1.3%)	1 (1.7%)	
none	174 (77.3%)	44 (77.2%)	
Subsequent therapy after second TACE[b]			.493
TACE only	175 (77.8%)	50 (87.7%)	
Surgical resection	9 (4.0%)	0 (0%)	
Local ablation therapy	10 (4.4%)	1 (1.8%)	
SIRT	5 (2.2%)	1 (1.8%)	
Systemic therapy	24 (10.7%)[c]	5 (8.8%)[d]	
SIRT + Sorafenib	2 (0.9%)	0 (0.0%)	
Type of imaging before the second TACE			<.001
CT	189 (84.0%)	17 (29.8%)	
MRI	36 (16.0%)	40 (70.2%)	

Note: The table provides baseline patient characteristics of both groups (diagnostic training dataset and the holdout validation dataset). Data are given as n (%) or median (range) unless otherwise noted.

Abbreviations: AFP, alpha-fetoprotein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BCLC, Barcelona Clinic Liver Cancer; CT, computed tomography; cTACE, conventional transarterial chemoembolization; DEB-TACE, drug eluting bead transarterial chemoembolization; HBV, hepatitis B virus; HCV, hepatitis C virus; INR, international normalized ratio; MELD, Model for End-stage Liver Disease; MRI, magnetic resonance imaging; NASH, non-alcoholic steatohepatitis; SIRT, selective internal radiation therapy; SMI, skeletal muscle index; TACE, transarterial chemoembolization.

[a]The sum of aetiologies could exceed 100% because patients could have more than one aetiology.

[b]We considered only therapies performed within the observation period of 12 months.

[c]All 24 patients received sorafenib; in eight patients sorafenib therapy had been started between the first and second TACE.

[d]Two patients received nivolumab, the remainder received sorafenib.

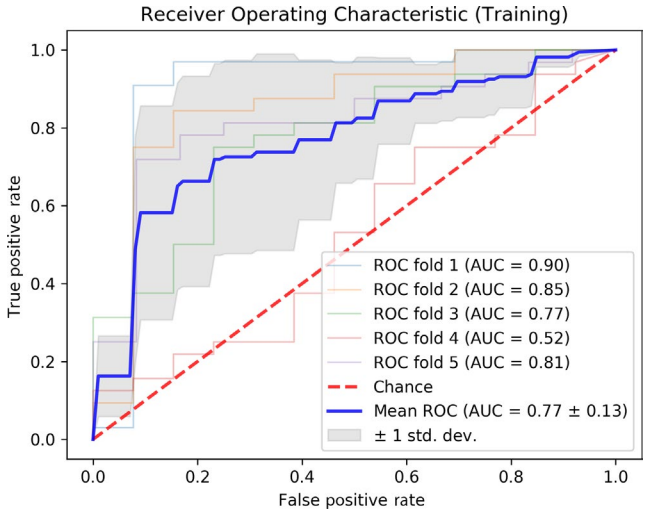

FIGURE 3 Receiver operating characteristic curve of the diagnostic training cohort. The analysis of the predictive ability of the artificial neural network in the diagnostic training cohort yielded an area under the curve of 0.77 ± 0.13

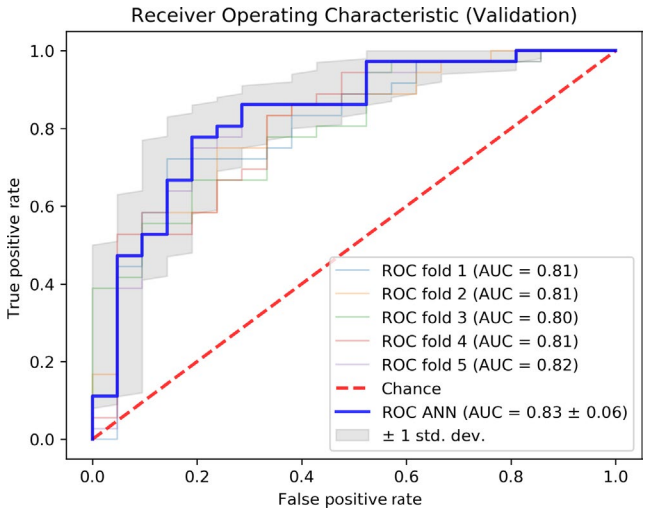

FIGURE 4 Receiver operating characteristic curve of the holdout validation cohort. The analysis of the predictive ability of the artificial neural network in the holdout validation cohort yielded an area under the curve of 0.83 ± 0.06

lower than the AUC of our ANN (Figure 5). This difference reached significance in case of the ART score ($P < .001$); for ABCR and SNACOR significance was not reached ($P = .143$ and $P = .201$ respectively).

The complete ANN is publicly available online in the Mendeley repository.[37] To facilitate easy implementation of the model, the repository comprises the Python script, a sample data file and a detailed manual.

4 | DISCUSSION

This is the first study applying an ANN for survival prediction in patients with HCC undergoing TACE. The ANN achieved a promising performance at predicting 1-year survival in patients with HCC prior

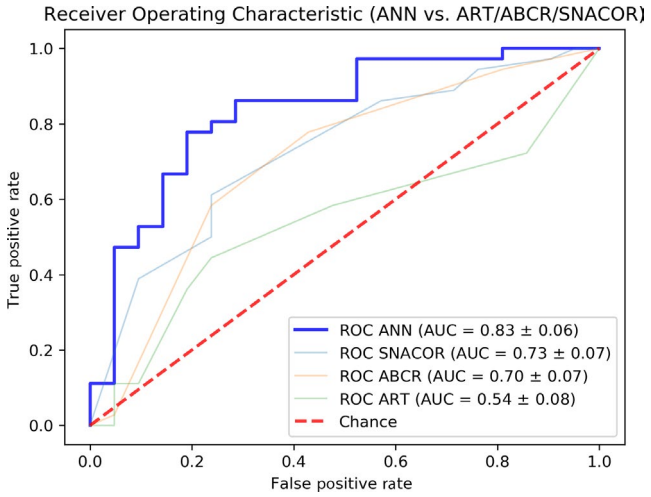

FIGURE 5 Receiver-operating characteristic analysis comparing the predictive ability of the artificial neural network (ANN) to ART, ABCR and SNACOR. The respective areas under the curve (AUCs) were 0.83 for the ANN, 0.73 for SNACOR, 0.70 for ABCR and 0.54 for ART. These AUCs correspond to 'good prediction' for the ANN, 'fair prediction' for SNACOR and ABCR and 'very poor prediction' for ART

to the second TACE treatment. With an AUC of 0.83, the ANN outperformed conventional scoring systems.

As there is a dire need for more objective decision-making, several prediction tools using a conventional score-based approach have been developed for treatment stratification in patients with HCC. The ART, ABCR and SNACOR scores aim to improve treatment stratification for patients with HCC prior to their second TACE treatment.[4-6] They achieved AUCs/Harrell's C indices between 0.60 and 0.75 for 1-year survival.[5,6] However, in external validations by several study groups, their predictive ability could not be reproduced,[7-10] for example, in our own external validations, we obtained Harrell's C indices between 0.54 and 0.59.[7,8] This difference is probably at least partially due to the so-called 'overfitting' effect, which has been described as 'a phenomenon occurring when a model maximizes its performance on some set of data, but its predictive performance is not confirmed elsewhere due to random fluctuations of patients' characteristics in different clinical and demographical backgrounds'.[38] The phenomenon of non-replicability has been recognized as a common problem, particularly in the life sciences. It describes how different factors including inherent characteristics of the systems, bias in reporting, or problems in study design, execution, or interpretation lead to different results between the original study and the replication attempt.[39]

Only very few studies have used machine learning-approaches in patients with HCC in the setting of TACE.[15,16] Further, these studies follow a different methodology. Peng et al uses a convolutional neural network based on pattern recognition in CT-images. Abajian et al use a total of five features (two MR-imaging-based parameters and three clinical parameters) and perform logistic regression and random forest models. Both studies aimed to predict tumour response prior to the first TACE.[15,16] In contrast, we used

a fully connected ANN based on a multitude of clinical parameters (n = 46) to predict OS prior to the second TACE. Similar approaches using an ANN have already been used following tumour resection.[40-42] Regarding its use in interventional oncology, Wu et al tried to predict disease free survival in patients with HCC after radiofrequency ablation.[43] Until now, it has never been tried in the setting of TACE.

Using this novel approach, we achieved a promising predictive performance with an AUC of 0.83. Once trained, an ANN like ours can easily be implemented in clinical routine and might help to determine further treatment; an explanatory figure can be found in Figure S2. A head-to-head comparison of our ANN with the ART, ABCR and SNACOR scores yielded highest AUCs for the ANN, corresponding to 'good prediction', followed by SNACOR and ABCR ('fair prediction') and the lowest for ART ('very poor prediction'). However, the predictive performance was only significantly better in case of the ART score, for ABCR and SNACOR significance was not reached. As this head-to-head comparison is based solely on the small holdout validation group comprising only 57 patients, the non-significance might be due to underpower.

One of the main advantages of such an ANN is that it can include a broad choice of variables. In our case, we used a total of 46 input variables covering most evidence-based prognostic variables used in daily clinical practice to characterize patients with cirrhosis and/or HCC. Moreover, ANNs are easily scalable when the complexity, the number of patterns and the number of inputs of the dataset increase, and might therefore carry advantages over classical machine learning techniques like random forest classifiers etc.[44]

However, the use of an ANN is associated with several shortcomings. In contrast to traditional statistical approaches, it is somehow a 'black box', which does not allow for easy identification of parameters associated with good predictive ability. Furthermore, an ANN cannot deal with missing values. Therefore, to avoid multiple imputations, the data have to be as complete as possible. Unfortunately, medical patient data is often incomplete or difficult to retrieve from different existing data sources (eg separate radiology information system, hospital information system, laboratory information system etc). Another issue is the lack of digitization as some information is still paper-based. In the future, the broad introduction of novel tools, such as structured reporting, may improve data quality, completeness and availability, facilitating the training and application of neural networks.

Our analysis has several limitations. Firstly, and most importantly, our study lacks an independent external validation cohort. Although we used a holdout patient cohort not used for training as validation, further external validation is mandatory. To encourage independent study groups to verify our results and to address the issue of non-reproducibility, we provide the ANN for download in our Mendeley repository.[37]

Secondly, the study design was retrospective and the final sample size (n = 282) was only moderate. Most likely, our dataset used herein is too small to use the ANN to its full capacity. Therefore, the performance might probably not be superior to classical approaches in this case; however, it is very unlikely that the performance falls behind that of any classical approach. Ideally, training and subsequent validation would be performed with a sufficiently large patient cohort using a multicenter approach. Such a multicenter approach would increase the robustness of the model and also tackle the problem of 'overfitting'. Thirdly, we included all variables used by all three scoring systems, including some handcrafted variables (eg Child-Pugh, BCLC). Using such handcrafted variables still requires additional user input; however, the parameters used herein are commonly applied clinical parameters to characterize patients with liver disease, which should be readily available in most liver centres.

Even though we included a broad variety of potentially clinically meaningful parameters, it is possible that we missed variables that would have further improved the prediction. Moreover, some variables were not included because they were not available for all patients, for example, most patients lacked tumour grading and status of small vessel infiltration because they were diagnosed non-invasively. Furthermore, it may be possible that the inclusion of other more advanced parameters could further enhance prediction, for example, radiomic data including texture analysis.[45-47]

Another possible reason of bias could be the large period of data recruitment comprising 13 years. Meanwhile, several technical improvements were made for TACE. Although we introduced DEB-TACE in 2006 at our institution–and therefore both TACE regimes were constantly used throughout the whole recruitment period—the distribution between cTACE and DEB-TACE was significantly different between training and holdout validation group due to a shift towards DEB-TACE in recent years. However, both techniques were equally effective in the two largest multicenter RCTs.[20,48] Consequently, both techniques are equally endorsed by the most recent EASL guideline and the choice is left to the operator.[18] Additionally, several new systemic drugs became available influencing the switch from TACE to systemic therapy.[49]

We used CT and MRI for measuring tumour load as well as for determination of tumour response. In recent years, there was a shift towards MRI, consequently the proportion of patients receiving MRI was greater in the validation group. However, both imaging modalities are accepted for HCC-imaging as well as for determination of tumour response.[18,22] Further, we included primarily Caucasian patients. Due to fundamental differences in patient characteristics, for example, regarding aetiology of liver disease, these results may not be transferable to Asian patients. Lastly, we decided on a neural network with three hidden layers. As there is no commonly accepted design of such networks for similar purposes, it may be possible that the current design is not the best one available. This suggests that a perfectly designed network could allow for even better prediction.

5 | CONCLUSIONS

Neural networks could be better at predicting patient survival after TACE for HCC compared to existing scoring systems using a conventional statistical approach. Once established, such prediction models

could easily be deployed into clinical routine and help determine optimal patient care. Especially less experienced investigators might profit from support mechanisms based on such machine learning algorithms.

Nevertheless, clinical reality is more complex than such a network can capture. Therefore, it may only serve as one of several components in decision-making and cannot replace a clinician's long-lasting practical experience. Inclusion of additional parameters in the prediction model could potentially further increase its performance. This could not only include clinical parameters that have already demonstrated predictive value, such as postembolization syndrome after first TACE,[50] but also novel predictive parameters such as texture analysis.[45-47] Potentially, a combination of our ANN with a convolutional neural network using pattern recognition might further enhance prediction. However, to avoid the problem of overfitting and to enhance generalizability, the network needs to be built on a broader database including clinical data from several institutions.

ACKNOWLEDGEMENTS

The study includes data from the doctoral thesis of one of the authors (FW). DPdS and RK contributed equally to the manuscript. We thank Lukas Müller for his support with data acquisition.

CONFLICT OF INTEREST

PRG has received grants and personal fees from Bayer and personal fees from Bristol-Myers Squibb, MSD Sharp & Dohme, Lilly, Sillajen, SIRTEX and AstraZeneca. AW has received speaker fees and travel grants from Bayer. DPDS has received personal fees from Cook. RK has received speaker fees from BTG, Guerbet and SIRTEX and personal fees from Boston Scientific, Bristol-Myers Squibb, Guerbet and SIRTEX. None of these companies supported this study, and none of the authors report a conflict of interest.

ORCID

Sebastian Schotten https://orcid.org/0000-0003-3359-0732
Peter R. Galle https://orcid.org/0000-0001-8294-0992
Daniel Pinto dos Santos https://orcid.org/0000-0003-4785-6394
Roman Kloeckner https://orcid.org/0000-0001-5492-4792

REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019;144:1941-1953.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68:394-424.
3. Llovet JM, Bru C, Bruix J. Prognosis of hepatocellular carcinoma: the BCLC staging classification. *Semin Liver Dis*. 1999;19:329-338.
4. Sieghart W, Hucke F, Pinter M, et al. The ART of decision making: retreatment with transarterial chemoembolization in patients with hepatocellular carcinoma. *Hepatology*. 2013;57:2261-2273.
5. Adhoute X, Penaranda G, Naude S, et al. Retreatment with TACE: the ABCR SCORE, an aid to the decision-making process. *J Hepatol*. 2015;62:855-862.
6. Kim BK, Shim JH, Kim SU, et al. Risk prediction for patients with hepatocellular carcinoma undergoing chemoembolization: development of a prediction model. *Liver Int*. 2016;36:92-99.
7. Kloeckner R, Pitton MB, Dueber C, et al. Validation of clinical scoring systems ART and ABCR after transarterial chemoembolization of hepatocellular carcinoma. *J Vasc Intervent Radiol*. 2017;28:94-102.
8. Mähringer-Kunz A, Weinmann A, Schmidtmann I, et al. Validation of the SNACOR clinical scoring system after transarterial chemoembolisation in patients with hepatocellular carcinoma. *BMC Cancer*. 2018;18:489.
9. Terzi E, Terenzi L, Venerandi L, et al. The ART score is not effective to select patients for transarterial chemoembolization retreatment in an Italian series. *Dig Dis*. 2014;32:711-716.
10. Arizumi T, Ueshima K, Iwanishi M, et al. Evaluation of ART scores for repeated transarterial chemoembolization in Japanese patients with hepatocellular carcinoma. *Oncology*. 2015;89(Suppl 2):4-10.
11. Zador AM. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nat Commun*. 2019;10:3770.
12. Cucchetti A, Vivarelli M, Heaton ND, et al. Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. *Gut*. 2007;56:253-258.
13. Cucchetti A, Piscaglia F, Grigioni AD, et al. Preoperative prediction of hepatocellular carcinoma tumour grade and micro-vascular invasion by means of artificial neural network: a pilot study. *J Hepatol*. 2010;52:880-888.
14. Meek RD, Lungren MP, Gichoya JW. Machine learning for the interventional radiologist. *AJR Am J Roentgenol*. 2019;231:782-784.
15. Peng J, Kang S, Ning Z, et al. Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging. *Eur Radiol*. 2020;30:413-424.
16. Abajian A, Murali N, Savic LJ, et al. Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning-an artificial intelligence concept. *J Vasc Intervent Radiol*. 2018;29(850–857):e851.
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162:55-63.
18. European Association for the Study of the Liver. EASL clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol*. 2018;69:182-236.
19. 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;68:723-750.
20. Lammer J, Malagari K, Vogl T, et al. Prospective randomized study of doxorubicin-eluting-bead embolization in the treatment of hepatocellular carcinoma: results of the PRECISION V study. *Cardiovasc Intervent Radiol*. 2010;33:41-52.
21. Lencioni R, de Baere T, Burrel M, et al. Transcatheter treatment of hepatocellular carcinoma with Doxorubicin-loaded DC Bead (DEBDOX): technical recommendations. *Cardiovasc Intervent Radiol*. 2012;35:980-985.
22. Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. *Semin Liver Dis*. 2010;30:52-60.
23. Weinmann A, Koch S, Niederle IM, et al. Trends in epidemiology, treatment, and survival of hepatocellular carcinoma patients between 1998 and 2009: an analysis of 1066 cases of a German HCC Registry. *J Clin Gastroenterol*. 2014;48:279-289.
24. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet*. 1995;346:1075-1079.
25. Shi H-Y, Lee K-T, Lee H-H, et al. Comparison of artificial neural network and logistic regression models for predicting

in-hospital mortality after primary liver cancer surgery. *PLoS ONE*. 2012;7:e35781.
26. Raffetti E, Portolani N, Molfino S, et al. Role of aetiology, diabetes, tobacco smoking and hypertension in hepatocellular carcinoma survival. *Dig Liver Dis*. 2015;47:950-956.
27. Kolly P, Knopfli M, Dufour JF. Effect of smoking on survival of patients with hepatocellular carcinoma. *Liver Int*. 2017;37:1682-1687.
28. Fujiwara N, Nakagawa H, Kudo Y, et al. Sarcopenia, intramuscular fat deposition, and visceral adiposity independently predict the outcomes of hepatocellular carcinoma. *J Hepatol*. 2015;63:131-140.
29. Siriwardana RC, Liyanage CAH, Gunetilleke B, et al. Diffuse-type hepatoma: a grave prognostic marker. *Gastrointest Tum*. 2017;4:20-27.
30. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. *Hepatology*. 2001;33:464-470.
31. Johnson PJ, Berhane S, Kagebayashi C, et al. Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach-the ALBI grade. *J Clin Oncol*. 2015;33:550-558.
32. Lin Z-H, Li X, Hong Y-F, et al. Alanine aminotransferase to hemoglobin ratio is an indicator for disease progression for hepatocellular carcinoma patients receiving transcatheter arterial chemoembolization. *Tumor Biol*. 2016;37:2951-2959.
33. Pang Q, Qu K, Zhang J-Y, et al. The prognostic value of platelet count in patients with hepatocellular carcinoma: a systematic review and meta-analysis. *Medicine*. 2015;94:e1431.
34. Biolato M, Miele L, Vero V, et al. Hepatocellular carcinoma treated by conventional transarterial chemoembolization in field-practice: serum sodium predicts survival. *World J Gastroenterol*. 2014;20:8158-8165.
35. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240:1285-1293.
36. Duncan I. *Healthcare Risk Adjustment & Predictive Modeling*. New Hartford, CT: Actex Learning; 2018.
37. Kloeckner R. Predicting survival after transarterial chemoembolization for hepatocellular carcinoma using a neural network: A pilot study; 2019; Mendeley Data, v2. https://doi.org/10.17632/cdf8tb3pxm.2
38. Facciorusso A, Bhoori S, Sposito C, Mazzaferro V. Repeated transarterial chemoembolization: an overfitting effort? *J Hepatol*. 2015;62:1440-1442.
39. Reproducibility and Replicability in Science. Washington (DC); 2019. ISBN-13: 978-0309486163. Natl Academy Pr.
40. Qiao G, Li J, Huang A, et al. Artificial neural networking model for the prediction of post-hepatectomy survival of patients with early hepatocellular carcinoma. *J Gastroenterol Hepatol*. 2014;29:2014-2020.
41. Chiu H-C, Ho T-W, Lee K-T, Chen H-Y, Ho W-H. Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network. *Sci World J*. 2013;2013:201976.
42. Ho W-H, Lee K-T, Chen H-Y, Ho T-W, Chiu H-C. Disease-free survival after hepatic resection in hepatocellular carcinoma patients: a prediction approach using artificial neural network. *PLoS ONE*. 2012;7:e29179.
43. Wu CF, Wu YJ, Liang PC, et al. Disease-free survival assessment by artificial neural networks for hepatocellular carcinoma patients after radiofrequency ablation. *J Formos Med Assoc*. 2017;116:765-773.
44. Fernandez-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res*. 2014;15:3133-3181.
45. Xu X, Zhang H-L, Liu Q-P, et al. Radiomic analysis of contrast-enhanced CT predicts microvascular invasion and outcome in hepatocellular carcinoma. *J Hepatol*. 2019;70:1133-1144.
46. Ni M, Zhou X, Lv Q, et al. Radiomics models for diagnosing microvascular invasion in hepatocellular carcinoma: which model is the best model? *Cancer Imaging*. 2019;19:60.
47. Yuan C, Wang Z, Gu D, et al. Prediction early recurrence of hepatocellular carcinoma eligible for curative ablation using a radiomics nomogram. *Cancer Imaging*. 2019;19:21.
48. Golfieri R, Giampalma E, Renzulli M, et al. Randomised controlled trial of doxorubicin-eluting beads vs conventional chemoembolisation for hepatocellular carcinoma. *Br J Cancer*. 2014;111:255-264.
49. Kudo M. Systemic therapy for hepatocellular carcinoma: latest advances. *Cancers*. 2018;10(11):412.
50. Mason MC, Massarweh NN, Salami A, Sultenfuss MA, Anaya DA. Post-embolization syndrome as an early predictor of overall survival after transarterial chemoembolization for hepatocellular carcinoma. *HPB (Oxford)*. 2015;17:1137-1144.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Mähringer-Kunz A, Wagner F, Hahn F, et al. Predicting survival after transarterial chemoembolization for hepatocellular carcinoma using a neural network: A Pilot Study. *Liver Int*. 2020;40:694–703. https://doi.org/10.1111/liv.14380