SEROLOGICAL DETECTION OF HEPATOCELLULAR CARCINOMA - APPLICATION OF MACHINE LEARNING AND IMPLICATIONS FOR DIAGNOSTIC MODELS

Authors: Philip J Johnson, Ehsan Bhatti, Hidenori Toyoda & Shan He

Authors Affiliations:

Philip J Johnson – Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK

Ehsan Bhatti – School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK

Hidenori Toyoda – Department of Gastroenterology, Ogaki Municipal Hospital, Ogaki, Japan

Shan He – School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK

Acknowledgements of Research Support for This Study: PJJ was supported by an educational grant from FUJIFILM Healthcare Americas Corporation.

Corresponding Author: Philip J Johnson, Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool L69 3GE, UK. Philip.Johnson@liverpool.ac.uk ORCID No. 0000-0003-1404-0209
Funding for This Research: PJJ was supported by an educational grant from FUJIFILM Healthcare Americas Corporation

Ethics approval and consent to participate

* This study was approved by the Institutional Review Board of Ogaki Municipal Hospital and was performed in accordance with Helsinki declaration. Patients’ consent was waived due to the retrospective review of the clinical data.
* The study was performed in accordance with the Declaration of Helsinki.

Consent for publication

* In the Ogaki hospital, patients on their first admission were asked to provide written informed consent for the future anonymous use of their clinical data for publication.

Data availability

All reasonable requests for data-sharing will be considered by the authors

**Context summary**

***Key Objective:*** We aimed to apply machine learning (random forest (RF)) to improve on current approaches to serologic diagnosis of hepatocellular carcinoma (HCC), such as the GALAD score, that have been developed and validated on artificially balanced datasets (similar numbers of cases and controls) using regression analysis.

***Knowledge* generated:** We show how GALAD does not generalise well in a prospectively accrued dataset that better reflects HCC prevalence within the population. Additionally, we develop and assess two RF diagnostic models, highlighting their improved performance over GALAD via multiple metrics including AUROC and F1.

**Abstract**

***Background and Aims***

The GALAD score is a biomarker-based statistical model for the serological diagnosis of hepatocellular carcinoma (HCC) that has been developed and validated using the case-control approach with a view to early detection. Performance has, however, been sub-optimal in the first prospective studies which better reflect the ‘real-world’ situation. Here we report the application of machine learning to a large, prospectively accrued, HCC surveillance dataset.

***Patients and methods***

Models were built on a cohort of 3,473 patients with chronic liver disease within a rigorous surveillance programme between 1998 and 2014 during which 459 patients with HCC were detected. Two RF models were trained. The first RF model uses the same variables as the original GALAD model (GALAD-RF); the second is based on routinely available clinical and laboratory features (RF-practical). For comparison we evaluated a logistic regression (LGR) GALAD model trained on this longitudinal prospective dataset (termed GALAD-Ogaki).

***Results***

Models were evaluated using a repetitive cross validation approach with the metrics averaged over 100 independent runs. As judged by AUROC and F1 score, the GALAD RF model significantly outperformed the original GALAD model. The RF-practical model also outperformed the original GALAD model in terms of both AUROC and F1 score and both models outperformed the individual biomarkers. An online web application that implemented the GALAD-RF and RF-practical models is presented.

***Conclusion***

RF-based models improve on the diagnostic performance of original GALAD model in the setting of a standard HCC surveillance program. Further prospective validation studies are warranted using these models and could be expanded to offer prediction of risk of HCC development over defined periods of time.

**Introduction**

Hepatocellular carcinoma (HCC) is one of the most frequently occurring cancers the second most common cause of cancer related mortality 1. Potentially curative therapies for HCC, a major global healthcare problem, are predicated on early diagnosis and detection at a time when the tumour is small. At this early stage, potentially curative therapies can be implemented. To achieve this outcome regular 6-monthly ultrasound (US) examinations of patients with hepatic cirrhosis or advanced fibrosis (the primary risk factor) is recommended, with or without the tumour marker alpha-fetoprotein (AFP) 2-4. However, uptake into surveillance programmes has been poor particularly across the western world where only 20–40% of patients developing HCC are detected within a surveillance programme and early diagnosis remains uncommon 5. To overcome these practical problems many blood-based tests for HCC have been proposed and are currently the source of intensive investigation. Prime amongst these has been alpha-fetoprotein (AFP) but its specificity has been questioned and, recognizing that a single biomarker is unlikely to have adequate sensitivity or specificity, we have used a LGR statistical model involving several biomarkers and clinical features 6-10 to develop a diagnostic statistical known as the GALAD score. This score combines AFP with age and gender with two further serological markers ‘lens culinaris agglutinin-reactive fraction of AFP’ (AFP-L3, an isoform of AFP) and des-gamma-carboxy prothrombin (DCP; also called protein induced by vitamin K absence or antagonist-II (PIVKA-II)).

In line with the current standard biomarker pathway 11, initial investigations (‘phase-II’) are usually based on case-control studies and it was in such a setting that the GALAD score was developed and validated  6, 7, 9, 12. Again, in line with the biomarker pathway strategy, such case-control studies need to be followed by prospective studies as case-control studies can over-estimate biomarker performance. In the case of the GALAD score the first phase III follow-up studies have not fulfilled initial promise 13, 14. Specifically, a significant number of false positive results have been detected and the results of the GALAD model may not be better than the performance of its individual constituent biomarkers 14.

In the course of our own studies involving a large prospectively-accrued dataset, the performance of the GALAD model has also proved less than optimal and we have previously cautioned ‘*despite these encouraging results, the reported performance measures are all based on case/control studies and these might not be the optimal way of testing potential biomarkers’* 15.

In this study, we have considered the reasons behind this falloff in performance and have tested the explicatory hypothesis that case-control studies do not reflect the real-world situation. Specifically, all GALAD studies to date are based on case-control studies in which the numbers of ‘cases’ and ‘controls’ are artificially ‘balanced’ and the resulting models are based on logistic regression. In the practical, real-world situation this is not the case. In any surveillance program or practical diagnostic setting, control subjects will greatly outnumber HCC cases. Prospective studies mirror this real-world situation much more closely as well as ensuring that cases and controls are drawn from the same population. Assessment of diagnostic power for clinical use should therefore be based on prospective studies and use outcome metrics such as the F1 score which is recognised as the appropriate test for unbalanced populations.

We report here an improved approach to the assessment of the diagnostic role of serum biomarkers based on the application of a new machine learning model trained on a large prospectively accrued dataset. We stress that it is not our aim to predict HCC development. Rather, we aim to improve on the current diagnostic serologic tests.

**Patients and Methods**

The model was built on a cohort of 3,473 patients who were part of a surveillance programme at Ogaki Municipal Hospital in Japan. Ogaki Municipal Hospital is a general hospital serving a well-defined and stable local population of approximately 400,000. Of all HCC patients seen in this municipality, more than 70% were detected under surveillance 16. Patients entered the study between March 1998 and April 2014 with the last follow-up occurring in 2021. To be included in the current analysis, patients were required to have detailed documentation of clinical and laboratory features including relevant serological biomarkers. In the event the mean number of serological and clinical records per patient was 26 [range 6 to 137). This dataset henceforth referred to as the ‘*Ogaki dataset’*.

Patients with HCC before or at entry into the programme were excluded if complete data was not available (no disregards). In all sub-group analyses data was > 95% complete. A positive ultrasound screening examination or levels of AFP >20 ng/mL prompted further definitive tests for HCC, Computerised Tomography (CT) or Magnetic Resonance Imaging (MRI) scans. Overall, 445 (12.8%) patients were diagnosed with HCC according to the European Association for the Study of the Liver guidelines 17. As a consequence of the high rates of liver resection, the majority of cases of HCC were also confirmed histologically. A more detailed characterisation of the Ogaki surveillance programme is given in Johnson *et al*. (2022) 18. The patient demographics and clinical characteristics are summarised in Table 1.

.

**Machine Learning And Statistical Methods**

Random Forest (RF) is an ensemble learning approach in which a collection of decision trees work in parallel for prediction 19. An RF model consists of 100 decision trees each of which were constructed using the bootstrapping principle where the available training data is sampled with replacement N times until a new dataset of the same size is created 20. Hence each decision tree is trained on a dataset the same size as the training data while having some variation between trees to reduce overfitting. To further reduce overfitting, each tree uses a randomised subset of the available features for its splitting criterion, meaning that no two decision trees are the same. When the trained RF model is presented with new unseen data, each tree classifies the new sample. These classification results are aggregated and outputted as the final prediction by the RF model 21.

We trained RF models on the prospective Ogaki dataset and compared their performance to the original GALAD model and a new GALAD model retrained on the Ogaki dataset (henceforth referred to as GALAD-RF). Both the original and the retrained GALAD models, as well as the RF models were then assessed on the prospective Ogaki dataset.

We trained two RF models using different sets of parameters:

GALAD-RF: An RF model trained using the same parameters as the GALAD model Gender, Age, the percentage of alpha-fetoprotein (AFP) binding to Lens culinaris agglutinin (AFP-L3%), alpha- fetoprotein (AFP) and Des-gamma-carboxy-prothrombin (DCP). This model serves as a direct comparison between Random Forests and logistic regression approaches. See supplementary material for more information about model formation.

"RF-Practical: An RF model trained on a section selection of routine biomarkers that are easy to obtain, which include in full 'AFP', 'Plt', 'ALB', 'Age', 'ALP', 'Tbil', 'Gender', 'Cirrhosis’. This model demonstrates the flexibility of the Random Forest methodology, showcasing its performance using a variety of common serological markers. Other putative biomarkers could be readily incorporated."

**Data Pre-processing**

Following the example of Johnson *et al* 6 in creating GALAD, categorical variables were converted into distinct numeric values. The gender of the patient is set to 1 if the patient is male and 0 if they are female. Similarly, cirrhotic patients were labelled as 1, while non-cirrhotic patients were labelled as 0. Patients were classified as cirrhotic or non-cirrhotic at entry into the study but over the long follow-up several patients progressed to cirrhosis. This transition was established principally by serial estimations of fibrosis index, Fib-4. A consistent Fib-4 figure over >3.25 led to a classification as ‘cirrhosis’.

The next step of pre-processing the Ogaki dataset was the extraction of relevant data for the construction and testing of the machine learning models. Firstly, any data point for an HCC patient recorded after their diagnosis was omitted as treatments may affect the values of their prognostic biological markers. Similarly, we removed the data of any patient who took the medication Warfarin as this has been shown to alter DCP protein concentrations 22. Secondly, for HCC patients we removed any data point that occurred more than 6 months prior to their diagnosis. We chose 6 months because the values of the variables considered were those closest to the date of the definitive HCC diagnosis and, in all cases, less than 6 months prior to that date.

Thirdly, we addressed the imbalance of the HCC and non-HCC cases, which is common in cancer clinical datasets and could affect the training and evaluation of the machine learning models 23. To reduce the imbalance, we followed the up-sampling approach. We first created a non-HCC data pool comprising of the final 5 data points available for each non-HCC patient. From this non-HCC data pool, we initially selected the final data point of each non-HCC patient (without replacement) and then randomly sampled from the pool without replacement until achieving a 20:80 ratio of HCC: non-HCC. i.e., the dataset was reduced to 4785 data points, of which 957 were labelled HCC (3828 non-HCC).

This pre-processing methodology ensured that all eligible patients were included in the study while maintaining a realistic HCC signal within the population. Emphasis was placed on the selection of HCC data points using those that are most likely to capture the change in serological markers as cancer develops.

We employed a repeated K-Fold paradigm consisting of 5 folds and 20 repeats. An RF model was developed using the entirety of the training folds, the testing fold acted as unseen data to evaluate the model. To prevent any data leakage, all of a patient’s data points were kept in the same fold, thereby ensuring the model was generalised and not recognising the same patient across folds. For a fair comparison, the Standard-GALAD model was evaluated on the same testing fold. The results of each of the 100 runs of the repeated K-Fold were recorded and averaged and shown below. This approach guarantees that the methodology is not biased based on the training test split and reduces the risk of overfitting.

**GALAD-Ogaki**

The next comparison was the evaluation of logistic regression as trained on a longitudinal prospective dataset, which was termed GALAD-Ogaki. The Ogaki dataset was pre-processed and repetitively portioned into training and testing splits using 5-Fold Cross Validations and repeated for 20 independent runs. A new GALAD model was trained using the training split and tested using the testing split. As above the entirety of a patient’s data was kept within the same fold ensuring generalisation. The final model was created by averaging the parameter coefficients and intercepts of the model developed during each run.

**Results**

Firstly, we confirmed that the original GALAD model performed poorly in this prospectively-accrued unbalanced dataset. When the true positive rate (sensitivity) was set at 0.9 (Table2) the RF approaches scored higher than GALAD across all metrics, in particular GALAD had an F1 score 0.5 which is inferior to the RF F1 score of 0.63. Similarly, the RF model achieves a higher AUC of .91 compared to GALAD's .86 as seen in Figure 1. Strikingly, at this high sensitivity level, the specificity fell to 0.58 with the original GALAD model whereas with the RF approach the specificity was much higher at 0.83. Whilst the GALAD-Ogaki outperformed the original-GALAD, scoring higher in all metrics, it too was unable to match the performance of the RF models and achieved an F1 score of 0.56.

Secondly, we developed a new, RF-based predictive model solely based on the routinely available clinical/demographic features of RF-Practical. This model gave an F1 of 0.69, a sensitivity of 0.62 and a specificity of 0.96 when the standard probability threshold of 50% (if greater than equal to 50% classifies patient as HCC Yes, otherwise No) was used (Table 3). Additionally, we compared RF-Practical to three RF models each trained using one of the protein biomarkers present in GALAD. RF-Practical outperforms the RF-protein models in all metrics apart from specificity where all models achieve a score of ~0.96. Furthermore, comparison of the protein RFs shows RF-AFP achieving an F1 score of 0.56 whilst RF-AFP-L3 and RF-DCP score 0.37 and 0.38 respectively. This notion of AFP being a more useful predictive biomarker in regards to RF models is further validated when examining the feature importance in RF-GALAD (Figure 2). Interestingly, DCP and AFP-L3 show similar levels of importance reflecting their near identical F1 scores.

**Discussion**

Our analyses suggest that the application of RF methodology significantly improves serological diagnosis of HCC within a prospectively accrued data. This methodology should be of value in the HCC surveillance setting permitting early and more accurate diagnosis at a time when potentially curative therapy can be offered. Such surveillance programs are already widely, albeit not efficiently, applied within public health systems. The importance of AFP is clearly shown in Figures 2 and 3. The RF models consistently outperformed the original GALAD model when the prospective study was used as a case control series i.e., all HCCs at the time of diagnosis acted as the cases and those not developing HCC acted as the control subjects. Similarly using the biomarkers for the original GALAD model but applying RF methodology also significantly improved performance. The first published application of the GALAD model to a prospectively accrued (‘phase III’) dataset 13 reported that the output for the original GALAD model did not improve on the performance of the individual constituent biomarkers particularly AFP L3 and resulted in a high percentage of false positive results. Data presented here (Table 3) shows convincingly that, in terms of the F1 score, the RF approach significantly outperformed any of the individual biomarkers. Our dataset is particularly ‘data-rich’ permitting us to examine combinations of several clinical variables all of which are within the RF-Practical model routinely available. Again, when assessed according to F1 score and AUROC our RF-Practical model, based on clinical parameters, both overall and in the various aetiological categories results in a greatly improved performance. For each of the models we have provided a ‘calculator’ {<https://hcc-rfs.azurewebsites.net/models>} which makes practical application, in clinical practice very straight forward.

We believe that the prospective setting more accurately reflects the real clinical situation in which such models would be applied. We suggest that the limitations of the original GALAD model, as recently reported by Tayob et al., 13, reflect the inadequacy of logistic regression analysis in which cases and controls are artificially balanced. We specifically tested this hypothesis by applying LGR to this dataset and comparing the results to our RF model (Table 2). Our analysis indicated that, the generalized linear model, logistic regression, on which both Standard-GALAD and GALAD-Ogaki are based, cannot handle the non-linear relationship between the biomarkers and HCC to generate accurate HCC diagnostic results.

In this study we examined the possibility of using serology to diagnose HCC i.e. the chance that, in a patient with chronic liver disease HCC is present. We achieved by focussing on the sample taken at the time of, or very shortly before, HCC diagnosis assessed by definitive radiological methods 17. Equally however, we will be able to extend the role of serology to predict (as opposed to diagnose) HCC by extending the window of tests from 6 months prior to conventional diagnosis to 3 or even 5 years. Furthermore, in this extended surveillance time-period we can explore models that incorporate repeated measures of the same patient i.e., whether the change in serological markers over time is a good predictive indicator of HCC.

We anticipate that this new model will find a role as part of ongoing HCC surveillance programs. Although such programs usually include AFP nowadays, this parameter is only included as present (above reference range) or absent whereas the new GALAD model offers increased accuracy using quantitated measures of this and other diagnostic factors.

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline characteristics | Overall | HCC | Non-HCC |
| Patients, n | 3473 | 445 | 3028 |
| Age (years) | 61.0 (51.7, 68.4) | 65.0 (58.3, 70.1) | 60.3 (50.5, 68.1) |
| Males, n (%) | 1715 (49.4) | 269 (60.4) | 1446 (47.8) |
| Albumin (g/L) | 42 (39, 44)  | 39 (36, 41)  | 42 (39, 44)  |
| Bilirubin (μmol/L) | 10.0 (8.3, 15.0)  | 11.7 (10.0, 16.7)  | 10.0 (8.3, 13.3)  |
| ALBI score | -2.87 (-3.08, -2.61) | -2.57 (-2.85, -2.32) | -2.91 (-3.11, -2.66) |
| Platelets (103/mm3) | 185 (138, 234)  | 123 (89, 164)  | 193 (151, 239) |
| Log(AFP) (ng/mL) | 0.4 (0.1, 0.7)  | 0.9 (0.6, 1.2)  | 0.4 (0.1, 0.7)  |
| Cirrhosis at enrolment, n (%) | 886 (25.5) | 274 (61.6) | 612 (20.2) |
| Follow-up (years) | 11.4 (7.7, 15.7) | 6.6 (4.5, 9.7) | 12.2 (8.5, 16.2) |
| Aetiology, n (%) | Hepatitis B | 821 (23.6) | 59 (13.3) | 762 (25.2) |
| Hepatitis C | 1896 (54.6) | 344 (77.3) | 1552 (51.3) |
| Hepatitis B + C | 16 (0.5) | 5 (1.1) | 11 (0.4) |
| Non-viral | 740 (21.3) | 37 (8.3) | 703 (23.2) |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Name** | **Threshold** | **F1** | **Sensitivity/TPR** | **Specificity/TNR** | **PCC** | **AUC** | **Precision/PPV** | **NPV** |
| Original-GALAD | 5.426 | 0.507 | 0.903 | 0.581 | 0.742 | 0.856 | 0.355 | 0.959 |
| RF-GALAD | 10.26 | 0.623 | 0.907 | 0.745 | 0.826 | 0.907 | 0.477 | 0.969 |
| RF-Practical | 11.32 | 0.614 | 0.906 | 0.734 | 0.82 | 0.911 | 0.467 | 0.969 |
| GALAD-Ogaki | 10.319 | 0.555 | 0.903 | 0.658 | 0.781 | 0.888 | 0.403 | 0.964 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Name** |  | **F1** | **Sensitivity/TPR** | **Specificity/TNR** | **PCC** | **AUC** | **Precision/PPV** | **NPV** |
| RF-Practical |  | 0.686 | 0.618 | 0.955 | 0.787 | 0.911 | 0.774 | 0.909 |
| RF-AFP |  | 0.557 | 0.464 | 0.951 | 0.708 | 0.830 | 0.703 | 0.876 |
| RF-AFP-L3 |  | 0.374 | 0.269 | 0.960 | 0.614 | 0.747 | 0.624 | 0.840 |
| RF-DCP |  | 0.380 | 0.278 | 0.955 | 0.616 | 0.709 | 0.605 | 0.841 |

Figure 1



Figure 3

Figure 2.

**References**

1. 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(6):394-424.

2. Aghemo A. Update on HCC Management and Review of the New EASL Guidelines. Gastroenterol Hepatol (N Y). 2018;14(6):384-6.

3. Wolf E, Rich NE, Marrero JA, Parikh ND, Singal AG. Use of Hepatocellular Carcinoma Surveillance in Patients With Cirrhosis: A Systematic Review and Meta-Analysis. Hepatology. 2021;73(2):713-25.

4. Tzartzeva K, Singal AG. Testing for AFP in combination with ultrasound improves early liver cancer detection. Expert Rev Gastroenterol Hepatol. 2018;12(10):947-9.

5. Goldberg DS, Taddei TH, Serper M, Mehta R, Dieperink E, Aytaman A, et al. Identifying barriers to hepatocellular carcinoma surveillance in a national sample of patients with cirrhosis. Hepatology. 2017;65(3):864-74.

6. Johnson PJ, Pirrie SJ, Cox TF, Berhane S, Teng M, Palmer D, et al. The detection of hepatocellular carcinoma using a prospectively developed and validated model based on serological biomarkers. Cancer Epidemiol Biomarkers Prev. 2014;23(1):144-53.

7. Berhane S, Toyoda H, Tada T, Kumada T, Kagebayashi C, Satomura S, et al. Role of the GALAD and BALAD-2 Serologic Models in Diagnosis of Hepatocellular Carcinoma and Prediction of Survival in Patients. Clin Gastroenterol Hepatol. 2016;14(6):875-86.e6.

8. Huang C, Fang M, Xiao X, Wang H, Gao Z, Ji J, et al. Validation of the GALAD model for early diagnosis and monitoring of hepatocellular carcinoma in Chinese multicenter study. Liver Int. 2022;42(1):210-23.

9. Yang JD, Addissie BD, Mara KC, Harmsen WS, Dai J, Zhang N, et al. GALAD Score for Hepatocellular Carcinoma Detection in Comparison with Liver Ultrasound and Proposal of GALADUS Score. Cancer Epidemiol Biomarkers Prev. 2019;28(3):531-8.

10. Harris PS, Hansen RM, Gray ME, Massoud OI, McGuire BM, Shoreibah MG. Hepatocellular carcinoma surveillance: An evidence-based approach. World J Gastroenterol. 2019;25(13):1550-9.

11. Parikh ND, Mehta AS, Singal AG, Block T, Marrero JA, Lok AS. Biomarkers for the Early Detection of Hepatocellular Carcinoma. Cancer Epidemiol Biomarkers Prev. 2020;29(12):2495-503.

12. Best J, Bechmann LP, Sowa JP, Sydor S, Dechêne A, Pflanz K, et al. GALAD Score Detects Early Hepatocellular Carcinoma in an International Cohort of Patients With Nonalcoholic Steatohepatitis. Clin Gastroenterol Hepatol. 2020;18(3):728-35.e4.

13. Tayob N, Kanwal F, Alsarraj A, Hernaez R, El-Serag HB. The Performance of AFP, AFP-3, DCP as Biomarkers for Detection of Hepatocellular Carcinoma (HCC): A Phase 3 Biomarker Study in the United States. Clin Gastroenterol Hepatol. 2023;21(2):415-23.e4.

14. Singal AG, Tayob N, Mehta A, Marrero JA, El‐Serag H, Jin Q, et al. GALAD demonstrates high sensitivity for HCC surveillance in a cohort of patients with cirrhosis. Hepatology. 2022;75(3):541-9.

15. Johnson P, Zhou Q, Dao DY, Lo YD. Circulating biomarkers in the diagnosis and management of hepatocellular carcinoma. Nature reviews Gastroenterology & hepatology. 2022;19(10):670-81.

16. Toyoda H, Kumada T, Tada T, Yama T, Mizuno K, Sone Y, et al. Differences in the impact of prognostic factors for hepatocellular carcinoma over time. Cancer Sci. 2017;108(12):2438-44.

17. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. J Hepatol. 2018;69(1):182-236.

18. Johnson PJ, Innes H, Hughes DM, Kalyuzhnyy A, Kumada T, Toyoda H. Evaluation of the aMAP score for hepatocellular carcinoma surveillance: a realistic opportunity to risk stratify. British journal of cancer. 2022;127(7):1263-9.

19. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008;26(9):1011-3.

20. Breiman L. Bagging predictors. Machine learning. 1996;24:123-40.

21. Biau G, Scornet E. A random forest guided tour. Test. 2016;25:197-227.

22. Ertle JM, Heider D, Wichert M, Keller B, Kueper R, Hilgard P, et al. A combination of α-fetoprotein and des-γ-carboxy prothrombin is superior in detection of hepatocellular carcinoma. Digestion. 2013;87(2):121-31.

23. Provost F. Machine learning from imbalanced data sets 101. 2008.

**Legends to Figures**

Figure 1. AUROC plot of the models.

Figure 2. Plot comparing the importance of individual input features in GALAD-RF

Figure 3. Plot comparing the importance of individual input features of RF-Practical

 **Legend to tables**

Table 1. Patient demographics and characteristics in the Ogaki cohort. Figures are given as median(range).

Table 2. Results of models where sensitivity is set to .9. Values equal to or greater than the threshold are classified as HCC. TPR (true positive rate); TNR (true negative rate); PCC (probability of correct classification); AUC (area under receiver operating characteristic curve); PPV (positive predictive value); NPV (negative predictive value).

Table 3. Results of RF-Practical and single biomarker feature RF models. A threshold value of 50 is used for all models. TPR (true positive rate); TNR (true negative rate); PCC (probability of correct classification); AUC (area under receiver operating characteristic curve); PPV (positive predictive value); NPV (negative predictive value).

Supplementary Information

**Standard-GALAD**

The Standard-GALAD logistic regression model proposed by Johnson *et al* (6) was developed using the original case-controlled dataset. The model uses an individual’s Gender, Age, AFP-L3 score and the logarithmic values of their AFP and DCP scores in the following equation.

* + - 1. $Z= -10.08+\left(0.09\*age\right)+\left(1.67\*sex\right)+\left(2.34\*log\left(AFP\right) \right)+\left(0.04\*AFP L3\right)+(1.33\*log\left(DCP\right) ) $

The second equation is then used to convert the GALAD score into a probability of the patient developing HCC.

* + - 1. $Probability HCC= \frac{exp⁡(Z)}{1+exp⁡(Z)}$

**Hyperparameters**

We utilise the default parameters of the random forest classifier model in the SK-Learn Package a full list of the hyperparameters are available here (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>).

If the reviewers feel it would be better, we can list the hyperparameters in full in the supplementary material.

Hyperparameters of note:
Number of trees = 100

Function used to measure quality of split = Gini

Bootstrap when developing trees = True

The citation for the package: Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

**Addendum terminology**

|  |
| --- |
| 1. Sensitivity (Specificity) is calculated for each characteristic as the percentage of cases (no cases) correctly predicted to develop (not develop) HCC out of all HCC (non HCC) patients with the given characteristic. Reported threshold is the point closest to the top-left corner on the ROC curve.
2. PCC=Probability of correct classification.
3. AUC=Area under ROC curve.
4. PPV=Positive Predictive Value.
5. NPV=Negative Predictive Value.
 |