Peer Review File

Article information: https://dx.doi.org/10.21037/atm-22-6469

Reviewer Comments

Comment 1: Line 106 – Define UCSF

Reply 1: *We have followed your instructions and defined UCSF criteria. (See page 7, line 113)*

Changes in the text: All transplanted patients were within UCSF criteria (single tumor ≤6.5 cm or ≤3 tumors with the largest tumor diameter ≤4.5 cm with a total tumor diameter ≤8 cm) (1,3).

Comment 2: Line 115 – Define TACE

Reply 2: *We have followed your instructions and defined TACE. We have also included additional information regarding TACE. (See page 7, line 123)*

Changes in the text: Transarterial chemoembolization (TACE) was used as a bridging treatment for patients with an expected waiting time for transplant of more than three months. Data concerning the administration and number of TACE sessions was collected, and no other bridging modalities, such as ablations or resections, were employed.

Comment 3: Lines 123 / 124: RFS is a standard, does not need a definition. However, if the authors would like to include a definition, then please correct it to the appropriate one (i.e. death or tumor recurrence, whichever comes first). Further to this, how was HCC recurrence evaluated as a separate event? Please define (i.e. HCC recurrence yes / no, or within a specific time-span?) Overall and graft survival are also reported, these should be added as secondary endpoints.

Reply 3: *We have revised the definition of RFS as suggested, and we have clarified how HCC recurrence was considered as a separate event. Additionally, we have included overall and graft survival as secondary endpoints. (See page 8, line 134)*

Changes in the text: The primary outcome of this study was recurrence-free survival which includes death or tumour recurrence, whichever occurred first. Secondary endpoints evaluated overall and graft survival, as well as whether patients would experience HCC recurrence within the study period.

Comment 4: Line 192 “researches” – correct to “studies” or alternative

Reply 4: *We have corrected “researches” to “studies” as suggested. (See page 13, line 248)*

Changes in the text: 1.3 Comparison with similar studies

Comment 5: Line 260-261: “Cox proportional hazard with coefficients is more descriptive about data” – this sentence does not make sense, please clarify

Reply 5: *We have modified the text to contain additional information regarding differences between Cox proportional hazards models and machine learning models. We believe that this clarifies the above statement. (See page 17, line 331)*

Changes in the text: The strength of the Cox proportional hazard model lies in its interpretable parameter estimates, which have a straightforward meaning in terms

expected hazard rate. Although this model works well with small datasets, the feature selection process requires careful consideration and expert medical knowledge, and should be based on the confidence interval of individual variables or statistical significance ($p<0.05$). ML methods, such as we used in this study, offer more flexible alternatives for analyzing large, complex, heterogeneous data with a nonlinear relationship. They can intrinsically perform feature selection providing a list of more influential variables without straightforward interpretation. We must be cognizant of the different objectives of both ML and classical statistical methods. Classical statistics is better suited for interpretation and describing the relationships between variables and the outcome of survival, while ML techniques (e.g. using a holdout test set and cross-validating hyperparameter search) are focused on creating the best possible model for accurate predictions of time of event, without intending to explicitly state the relationships between variables.

Comment 6: Tables describe 170 patients, but methods say 167 were included?
Reply 6: *Data from a total of 170 adult patients who underwent transplantation for HCC were collected over the period from March 2013 to December 2019 at the University Hospital Merkur, Zagreb. Unfortunately, due to missing data, three patients had to be excluded from the study. In the "Methods" section describing the data collected for the study, the authors mistakenly stated that data from 167 patients were collected, which was corrected during the revision process. It was also clarified in this section that three patients with incomplete datasets were excluded during the model development. We have modified our text to better clarify exclusion of the mentioned patients. (See page 6, line 101; page 7, line 110)*
Changes in the text: Data consisting of pretransplant parameters from 170 adult patients transplanted for HCC were collected in the period from March 2013 to December 2019 at the University Hospital Merkur, Zagreb.
We also excluded three patients with missing tumour burden scores. This score is composed of several variables which were missing for those patients, and imputation was not considered feasible in these cases.

Comment 7: In the Methods section, very little information is provided on the machine-learning methods used (for example hyperparameters such as learning rate and loss function, number of trees in the tree-based models etc.), cross-validation (size of groups? Number of cross validations? Randomized or stratified?), preprocessing etc.
Reply 7: *We have revised the "Methods" section to provide more thorough descriptions of the statistical and machine learning methods employed. This section is now divided into several subsections, each detailing the preprocessing, imputation, classical statistical methods, regularization, and machine learning methods used. (See page 8-10, lines 138-189)*
Changes in the text: Data analysis and ML modelling were performed using the Python version 3.9 programming language with open source libraries for statistics and ML (pandas, numpy, scikit-survival, scipy, statsmodels, seaborn and lifelines) (11,12). Descriptive statistical analysis, including visualization and outlier detection, was conducted on the entire dataset. Preparation of data for this complex analysis is

the first step, and while collecting and classifying data is not an issue, missing data can eliminate patients from a study. Imputation of missing data with substituted values is a recognized statistical method. In this analysis, imputation was performed via chained regression method and our dataset had <1% of missing data (24). Chained regression starts with replacing all missing values with the mean values of each column. The column to be imputed first is the one containing the least amount of missing data, and it is considered a dependent variable, while all other columns are independent variables (including the columns containing missing values that were replaced with mean values). A regression model is then used to impute the missing values of the current column, and then the process the process is moved onto the next column. This was repeated for all columns with missing values up to 10 times, to ensure that the values obtained by regression converged, i.e., stopped changing between iterations. (24) In order to estimate the impact of each variable on survival, a multivariate Cox proportional hazards model was developed on the whole dataset, after a non-parametric univariate Kaplan-Meier approach. We used a forward selection process based on predictive performance CI of each individual variable that is fit into the Cox model. The best final model, with highest CI, is described by the top 7 variables, which was selected in a 3-fold cross-validated grid search and implemented using the scikit-survival library. The final model is built on the whole dataset. (11) The data were also processed through the Coxnet: the elastic net regularized Cox regression model. Regularization is a ML procedure that reduces overfitting by controlling the growth of coefficients. This is mathematically expressed by adding a penalty term to the Cox log partial likelihood loss function. This penalty term is multiplied by a hyperparameter (alpha) that defines the weight of the penalty. When alpha is set to zero, we have a standard Cox model. As alpha increases, the coefficients are eventually shrunk to zero. Finding the ideal set of features is part of the optimization procedure of the hyperparameter alpha. After choosing a specific alpha value, we can perform prediction, either in terms of a risk score using the corresponding scikit-survival function or in terms of the survival or cumulative hazard function (*Figure 1*) (11,12). ML modelling was comprised from methods adapted for survival analysis on censored data. We used RSF, Survival SVM and Survival Gradient Boosting models. Random forest is an effective ML approach for both classification and regression. This method is constructing each tree with a different bootstrap sample and selecting diverse features for split criteria at each node. The final prediction is determined by aggregating the results of individual trees. The depth of trees influences the model's ability to control overfitting. For prediction, a sample is run through each tree until reaching a terminal node. At each terminal node, data is used to non-parametrically estimate survival and cumulative hazard functions with the Kaplan-Meier and Nelson-Aalen estimators respectively (*Figure 2*). The data were split into 75% for training and 25% for testing, with the model based on 1000 trees (7,8,11,13). The standard ML technique SVM can also be extended for survival data, as it has the advantage of being able to handle complex, non-linear relationships and survival using kernel functions. We used both Linear Survival SVM and Kernel SVM (with custom kernels) and in both cases, we used a cross-validated grid search to determine the optimal hyperparameter alpha. The best model was obtained by Linear Survival SVM with regression objective (11,14).

Similar to random forests, gradient boosting is an ensemble-based ML method based on multiple learners, however, the way they are combined is different. The final model is obtained by combining multiple "base learners" (predictors) that should be simple and slightly better than random guessing. The final prediction is the result of an additive model, whereby each of the base models sequentially improves and "boosts" the overall model. In this study, we employed 250 base learners, using the Cox partial likelihood as a loss function, with regression trees and compontent-wise least-squares as base learners and with 75% - 25% train - test split (11,15). The evaluation and selection of the best model were performed based on the 5-fold cross-validated CI. Although our final evaluation is performed on the test set and the selection of the best model is based on results in terms of the CI, it is important to note that a comparison between ML methods and classical statistical methods using CI alone is a limited view of the results, power, and objectives of these methods. (11,12).

Comment 8: Furthermore, some information should be provided on the imputation method used.
Reply 8: *The revised "Methods "section explains the imputation methods in detail, as previously stated in our reply to Comment 7, and corresponding reference is added. (See page 8, line 143)*
Changes in the text: Imputation of missing data with substituted values is a recognized statistical method. In this analysis, imputation was performed via chained regression method and our dataset had <1% of missing data (24). Chained regression starts with replacing all missing values with the mean values of each column. The column to be imputed first is the one containing the least amount of missing data, and it is considered a dependent variable, while all other columns are independent variables (including the columns containing missing values that were replaced with mean values). A regression model is then used to impute the missing values of the current column, and then the process the process is moved onto the next column. This was repeated for all columns with missing values up to 10 times, to ensure that the values obtained by regression converged, i.e., stopped changing between iterations (24). Reference added:
24. Little R, Rubin D. Statistical analysis with missing data.3rd ed. Wiley; 2019.

Comment 9: It is also unclear what exactly the models were predicting and there is no graphical representation of results. For example, if the models were predicting median RFS, then the underlying function used to predict RFS should be described and, if possible, a survival function graph (similar to Kaplan-Meier) should be produced. If predicting recurrence within a certain time-period, then a confusion matrix should be produced, with sensitivity / specificity etc.
Reply 9: *In this work we took predictions approach (e.g. probability of survival through time, Figure 1) instead of classification and this is why our results are not presented in terms of the confusion matrix or ROC curve. We have modified the "Data analysis" section of the "Methods" as stated in the reply to Comment 7 and we added survival function graphs showing survival prediction for six randomly selected patients (Figure 1, Figure 2), one showing survival using the regularized Cox proportional*

hazards model and the other showing survival using the Survival Random Forest model. Additionally we have added vertical lines to Kaplan Meier graphs showing median survival. (See page 8-10, lines 138-189)
Changes in the text: Same as mentioned in reply to Comment 7.

Comment 10: Generally, the reader is left to believe that the models somehow predict RFS (but the actual RFS predictions are not given) and that some are better than others, according to the c-index, but the authors do not go deeper than that.
Reply 10: *We have taken into account this comment by extending our "Methods" section and generating probability of survival graphs for selected patients from the test set. (See page 8-10, lines 138-189)*
Changes in the text: Same as mentioned in reply to Comment 7.

Comment 11: Recurrence-free survival is described as the primary endpoint of the study, but is not described in the results, nor in the Kaplan-Meier figures. Median and 1-,3-,5- year RFS should be added to results and median RFS should be plotted in a Kaplan-Meier graph. Why is survival described in days? I believe months is the commonly used scale.
Reply 11: *We have added a Kaplan-Meier figure representing recurrence-free survival to the "Results" section, including a comment and the median value. As per your suggestion, the survival graphs have been changed to present survival in months instead of days. (See page 11, line 192)*
Changes in the text: The 1-, 3- and 5-year post-transplant recurrence-free survival rates, as reported in *Figure 3*, were 78%, 70%, and 65%, respectively. Kaplan-Meier curves for 1-, 3- and 5-year post-transplant recipient survival time showed survival probability of 84%, 81% and 80% respectively (*Figure 4*). Graft 1-, 3- and 5-year survival time was 82%, 78% and 76% respectively (*Figure 5*).

Comment 12: Discussion, Lines 175 – 183: this belongs in the methods section
Reply 12: *The specified lines have been revised and incorporated in the "Methods" section of the manuscript. (See page 8-10, lines 138-189)*
Changes in the text: Same as mentioned in reply to Comment 7.

Comment 13: Overfitting as an issue and the advantages / disadvantages of cross-validation compared to external dataset testing should be described in limitations
Reply 13: *We have added the discussion about overfitting and our approach to addressing it in the "Strengths and limitations" part of the "Discussion". (See page 13, line 239)*
Changes in the text: The major problem in ML is overfitting, which means that the model is well-suited to the existing training data, but performs poorly when given new, unseen data. To avoid biased conclusions, we took several steps to address that problem, which is commonly encountered with small datasets. First, we inspected the dataset for potential outliers which can be influential when there is a low number of an observation. We also selected relevant features, performed regularization and controlled the depth of tree-based models. To make the most of our dataset, we executed cross-validation (3- and 5-fold) to identify the best

hyperparameters for all our ML methods. An example of finding the best hyperparameter alpha is given in *Figure 8*. Lastly, we tested our final models on a holdout test set, except for all Cox models, which was built on the entire dataset.

Comment 14: The authors use a large part of the discussion to describe the significance of the variables chosen by the models. This should be shortened, especially since no discussion takes place of other studies using machine-learning methods in liver transplantation. There are a couple of references in this paper, but, as far as I know, there are more papers in the literature regarding this topic. These should be sought out, referenced and discussed (merits / limitations of the studies, especially compared to this study) in 1-2 paragraphs.
Reply 14: The discussion was restructured to provide a brief overview of the use of machine learning in liver transplantation and transplant oncology, and to compare this study to more recent, relevant works. The list of references was updated to include recent papers concerning machine learning, while some older references were omitted. Significant variables were presented and explained in less detail due to the shortening of that part of the discussion. (See page 13-17, lines 249-325)
Changes in the text: Even though ML has existed for decades, it is a relatively new concept in medical data analysis and its popularity is currently increasing. Comparing the ML approaches to statistical methods such as Cox proportional hazards regression is a topic often discussed in literature (25,26,27). Each approach has its strengths and weaknesses. ML is algorithmic in its nature and it can identify patterns in data through numerous iterations to learn the relationships between parameters. As opposed to classical statistical modelling no assumptions about underlying distributions are made. Statistical analysis relies on hypothesis testing, data analysis and explanation of the relation between variables, while ML focuses more heavily on prediction of unseen data. Learning a model can take into account a large number of variables with their complex, nonlinear relations, while statistical analysis usually focuses on a relatively small number of parameters (25). ML is increasingly becoming an invaluable tool for evaluation of pre- and post-transplant aspects of cadaveric and living-donor liver transplantation. High-performing imaging assessments enabled by ML algorithms can provide more accurate pathohistological evaluation of graft quality and streamline the process of liver segmentation. ML can also be used to facilitate timely detection of liver tumours in the setting of HCC, as well as predicting post-transplant morbidity and mortality with greater accuracy (8,28,29). Models predicting waitlist dropout are also reported, bearing particular relevance for patients with HCC (30). Transplant oncology is a rapidly evolving field, and optimizing organ allocation in tumor patients and predicting tumour recurrence after transplantation is of great significance (8,10,31). Traditional models for patient selection and prediction of recurrence in liver transplantation for HCC rely on classical statistical methods and may be limited in a complex multifactorial setting. Nam et al. developed a model (MoRAL-AI) that uses deep neural networks to predict the HCC recurrence, taking into account tumour biology, as indicated by biomarkers, and imaging-assessed tumour size (32). Ivanics et al. conducted a comparative evaluation of multiple ML models to develop the Toronto postliver transplantation HCC recurrence calculator.

Their research showed that the Coxnet model had the best metrics and outperformed RSF, survival SVM, and neural networks (DeepSurv), and demonstrated the importance of exploring various ML models on data analysis (7). The application of a ML technique has the potential to make use of retrospective data to create more precise prognostic models. However, this can also be seen as a disadvantage, depending on the quality of the dataset. In comparison to similar studies, our approach was to investigate several different ML methods and the Cox proportional hazards regression on our dataset. Creating a predictive calculator or a score based on a limited number of patients from a single center can lead to imprecise outcomes due to potential biases. However, identifying predictors of survival is important for comparison with experience from other centers. External validation of similar studies is needed to obtain clinically relevant results.

4.4 Explanations of findings

Several variables stood out as having an impact on survival and tumour recurrence and some appeared in more than one model. The most detrimental variable in all the models was donor CRP. CRP is a non-specific acute phase reactant associated with various inflammatory diseases, sepsis and malignant tumours, that is widely available, inexpensive and has been in clinical use for many years. It is synthetized in hepatocytes, both in normal and HCC cells, and can reflect the degree of local inflammation since it encourages proliferation of hepatocytes and promotes HCC growth (33-36). Inflammation creates a microenvironment that favours DNA damage and neoangiogenesis, thus facilitating tumour growth and creating a vicious circle in which tumour creates inflammation that helps it develop (37). In literature, CRP is described as a prognostic factor of several types of cancer –oesophageal squamous cell carcinoma, cervical cancer and non-small cell lung cancer (38,39,40). Elevated CRP levels were also found to be predictive of overall survival and tumour recurrence in non-transplanted HCC patients after liver resection or treatment with locoregional therapies (41). Albeit the evidence for influence of cadaveric donor CRP on LT is scarce in literature, previous studies have reported that an elevated CRP in HCC patients undergoing living donor LT is predictive of a poor outcome (42,43). NLR as an index has demonstrated its value in infections, cardiovascular and inflammatory diseases, and in several types of primary and metastatic cancer. The role of NLR in HCC has emerged after the observation that sorafenib treatment in HCC patients had significantly better survival benefit in those with low NLR (19,20,44). It is considered that NLR represents the balance between the protumour inflammatory status and the antitumour adaptive immunity. Increase in NLR is suggestive of an increase in overall inflammatory status or a decrease in adaptive immunity. Hence inflammation, a stimulating factor in tumour microenvironment and a well-known indicator of tumour progression, was found important in survival and tumour recurrence in our group through two separate variables in our analysis. We considered NLR as a continuous variable with a reference range according to Forget et al., even though the optimal “cut-off point” in clinical settings has been debated in literature (45,46,47).

AFP is a widely accepted marker with prognostic significance in HCC, and also the only tumour marker routinely used for prognostication and treatment selection of patients with HCC. Even though exact tumour staging can only be confirmed after

histologic study of the explanted liver (tumour size and number, vascular invasion, differentiation), AFP is considered to be a representative parameter that correlates with vascular invasion, and thus can be very predictive of HCC recurrence (48,49,50). In some of the recent studies, it was not determined that AFP is a good prognostic marker for patient survival, but the rate of tumour recurrence showed a positive correlation with elevated AFP values (51,52). In our group, the median value of AFP was 11 ng/ml and HCC recurrence was reported in 19% of patients. Since AFP was found to be significant in the RSF model as well as in the Cox proportional hazards multivariate analysis, we can conclude that moderate elevation of AFP may be associated with increased incidence of HCC recurrence.
The share of elderly people in the general population is increasing and so is the age of transplant recipients. The elderly tend to have more comorbid conditions that can affect postoperative complications and survival (53,54). The mean age in our cohort was 62.3 ± 7.1 years which supports the claim that LT is successful and feasible in older population. Although older age has been highlighted as a risk factor in our models, our results suggest that recipients should be carefully selected according to their comorbidities and that old age alone is not an exclusion factor for LT.
Identifying other risk factors in combination with age could help in optimal patient selection.
References added:
28. Veerankutty FH, Jayan G, Yadav MK, et al. Artificial Intelligence in hepatology, liver surgery and transplantation: Emerging applications and frontiers of research. World J Hepatol 2021;13:1977-1990.
29. Tran J, Sharma D, Gotlieb N, Xu W, Bhat M. Application of machine learning in liver transplantation: a review. Hepatol Int 2022;16:495-508.
30. Kwong A, Hameed B, Syed S, et al. Machine learning to predict waitlist dropout among liver transplant candidates with hepatocellular carcinoma. Cancer Med 2022;11:1535-1541.
31. Nitski O, Azhie A, Qazi-Arisar FA, et al. Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. Lancet Digit Health 2021;3:e295-e305.
32. Nam JY, Lee J-H, Bae J, et al. Novel model to predict HCC recurrence after liver transplantation obtained using deep learning: a multicenter study. Cancers 2020;12:2791.

Comment 15: The authors do not describe the clinical / operative process at all, despite mentioning the homogeneity of a single-center study cohort as a potential advantage. The operative and immunosuppressive standards should be briefly mentioned in the methods section (eg. Piggy-back or total cava replacement, Tacrolimus / MMF etc).
Reply 15: *We have modified our text as advised and gave brief information about operative and immunosuppressive standards and later follow-ups. (See page 7, line 103)*
Changes in the text: All patients underwent whole liver cadaveric transplantation using grafts from Donation after Brain Death (DBD) donors. The grafts were procured using University of Wisconsin solution with aortic and portal flush. The

piggy-back technique was used for implantation. Standard immunosuppressive regimen included Tacrolimus, Mycophenolate Mofetil, and steroids with steroid tapering over 3 months. All recipients underwent regular follow-up visits with surveillance for HCC recurrence using cross sectional imaging every 3 months in the first two years post-transplant, and at least every 6 months thereafter.

Comment 16: TACE treatments are mentioned, what about ablations / resections as bridging to transplantation?
Reply 16: *We have followed your instructions in Comment 2 and defined TACE. We have also included additional information regarding TACE. Other bridging modalities, such as ablations or resections, were not used. (See page 7, line 123)*
Changes in the text: Same as mentioned in reply to Comment 2.

Additional changes in manuscript:
The „Key findings" subsection of „Discussion" has been modified due to other rearangements of the manuscript asked by the reviewers. (See page 12, line 226)
Changes in the text: This study has confirmed the predictive superiority of ML methods in comparison to traditional statistical analysis. We have also shown the importance of utilizing and comparing several different ML methods due to the specific ways in which they approach data processing. This study has also selected donor CRP as the most relevant predictor of recurrence free survival, which was not the case in similar studies. However, this needs to be confirmed with analysis of larger datasets and experience from other centers.
Two figures ("Figure 6. Random Survival Forest feature importance with mean feature weight and standard deviation" and "Figure 7. Elastic net regularization - best Cox regression model model features") were added to "Variable selection" part of the "Results" to better clarify our findings.

While reviewing the manuscript, some grammatical errors, typos and minor improvements in style were made in the reviewed version.