10-9-2021

The Toronto Postliver Transplantation Hepatocellular Carcinoma Recurrence Calculator: A Machine Learning Approach

Tommy Ivanics

Walter Nelson

Madhukar S. Patel

Marco P.A.W. Claasen

Lawrence Lau

See next page for additional authors

Authors

Tommy Ivanics, Walter Nelson, Madhukar S. Patel, Marco P.A.W. Claasen, Lawrence Lau, Andre Gorgen, Phillipe Abreu, Anna Goldenberg, Lauren Erdman, and Gonzalo Sapisochin

The Toronto Postliver Transplantation Hepatocellular Carcinoma Recurrence Calculator: A Machine Learning Approach

Tommy Ivanics [1,2,3,*] Walter Nelson,[4,5,*] Madhukar S. Patel,[6] Marco P.A.W. Claasen,[1,7] Lawrence Lau,[1] Andre Gorgen,[1] Phillipe Abreu,[1] Anna Goldenberg,[8] Lauren Erdman,[8,9,**] and Gonzalo Sapisochin[1,10,**]

[1]Multi-Organ Transplant Program, Division of General Surgery, Toronto General Hospital, University Health Network, University of Toronto, Toronto, ON, Canada; [2]Department of Surgery, Henry Ford Hospital, Detroit, MI; [3]Department of Surgical Sciences, Uppsala University, Akademiska Sjukhuset, Uppsala, Sweden; [4]Centre for Data Science and Digital Health, Hamilton Health Sciences, Hamilton, ON, Canada; [5]Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada; [6]Division of Surgical Transplantation, Department of Surgery, University of Texas Southwestern Medical Center, Dallas, TX; [7]Department of Surgery, Erasmus MC, University Medical Center Rotterdam, the Netherlands; [8]Centre for Computational Medicine, SickKids Research Institute, University of Toronto, Toronto, ON, Canada; [9]Center for Computational Medicine, SickKids Research Institute, Toronto, ON, Canada; and [10]Abdominal Transplant & HPB Surgical Oncology, Toronto General Hospital, University of Toronto, Toronto, ON, Canada

Liver transplantation (LT) listing criteria for hepatocellular carcinoma (HCC) remain controversial. To optimize the utility of limited donor organs, this study aims to leverage machine learning to develop an accurate posttransplantation HCC recurrence prediction calculator. Patients with HCC listed for LT from 2000 to 2016 were identified, with 739 patients who underwent LT used for modeling. Data included serial imaging, alpha-fetoprotein (AFP), locoregional therapies, treatment response, and posttransplantation outcomes. We compared the CoxNet (regularized Cox regression), survival random forest, survival support vector machine, and DeepSurv machine learning algorithms via the mean cross-validated concordance index. We validated the selected CoxNet model by comparing it with other currently available recurrence risk algorithms on a held-out test set (AFP, Model of Recurrence After Liver Transplant [MORAL], and Hazard Associated with liver Transplantation for Hepatocellular Carcinoma [HALT-HCC score]). The developed CoxNet-based recurrence prediction model showed a satisfying overall concordance score of 0.75 (95% confidence interval [CI], 0.64-0.84). In comparison, the recalibrated risk algorithms' concordance scores were as follows: AFP score 0.64 (outperformed by the CoxNet model, 1-sided 95% CI, >0.01; $P = 0.04$) and MORAL score 0.64 (outperformed by the CoxNet model 1-sided 95% CI, >0.02; $P = 0.03$). The recalibrated HALT-HCC score performed well with a concordance of 0.72 (95% CI, 0.63-0.81) and was not significantly outperformed (1-sided 95% CI, ≥0.05; $P = 0.29$). Developing a comprehensive posttransplantation HCC recurrence risk calculator using machine learning is feasible and can yield higher accuracy than other available risk scores. Further research is needed to confirm the utility of machine learning in this setting.

Liver Transplantation 0 1–10 2021 AASLD.
Received May 26, 2021; accepted September 23, 2021.

Liver transplantation (LT) is the best treatment option for patients with early stages of hepatocellular carcinoma (HCC).[1-4] However, the use of LT depends on maintaining a balance between patient-specific survival benefit, the availability of alternative treatment modalities,[5,6] and the equitable distribution of donor organs.[5,7-12]

Current selection criteria aim to avoid transplantation futility by excluding patients at a high risk of tumor recurrence.[10,11] Selecting patients with HCC within Milan criteria has been shown to provide excellent patient outcomes.[13-15] However, the Milan

Abbreviations: AFP, alpha-fetoprotein; BMI, body mass index; CI, confidence interval; ETOH, Alcohol; HALT-HCC, Hazard Associated with Liver Transplantation for Hepatoceullar Carcinoma; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; HCV, hepatitis

criteria has been challenged by other series reporting equivalent outcomes for transplanted patients with larger and more numerous tumors. Furthermore, while the use of parameters such as tumor size and number simplify the criteria,[14,16,17] recent studies have shown that the sole reliance on morphologic features does not adequately reflect tumor biology.[18,19] Thus, the need to incorporate additional prognostic factors, such as serum alpha-fetoprotein (AFP score)[20-22] and neutrophil-to-lymphocyte ratio (Model of Recurrence After Liver Transplant [MORAL] score),[23] became apparent and was further explored. In addition, radiologic and AFP responses to downstaging or bridging treatment have also been suggested to be important in predicting outcomes after LT.[24-26]

Given the opportunity cost of suboptimal organ allocation, creating a more precise and quantitative posttransplantation outcome calculator remains paramount. However, one of the main hurdles in developing such a calculator has been the limitation of standard statistical methods to account for many variables and their potential for various interactions. Looking forward, the amount of clinical data is only likely to increase. Machine learning represents a tool that can be used to derive meaning from such data.[27] Traditionally, humans have analyzed data and adapted systems to the changes in data patterns. However, as the volume of data surpasses the ability of humans to interpret and write rules, there will likely be a natural inclination to increasingly turn to automated systems that can actively learn from the data and adapt to shifting landscapes. With progress in applying machine learning techniques in medicine, we propose that these methods can be utilized to identify complex nonlinear relationships between a comprehensive set of factors and recipient outcomes in transplantation oncology.[28-31] Thus, we hypothesize that an accurate posttransplantation HCC recurrence calculator can be developed using a machine learning algorithm mapped on preoperative patient and tumor characteristics and have designed this study as a proof of concept.

C virus; IQR, interquartile range; LASSO, least absolute shrinkage and selection operator; LT, liver transplantation; MELD, Model for End-Stage Liver Disease; MELD–Na, Model for End-Stage Liver Disease–sodium; MLA, Machine learning algorithm; MORAL, Model of Recurrence After Liver Transplant; mRECIST, modified Response Evaluation Criteria In Solid Tumors; NASH, nonalcoholic steatohepatitis; NLR, Neutrophil-lymphocyte ratio; SRTR, Scientific Registry of Transplant Recipients.

Address reprint requests to Gonzalo Sapisochin, M.D., Ph.D., M.Sc., HBP & Multi-Organ Transplant Program, Division of General Surgery, Toronto General Hospital, University Health Network, University of Toronto, 585 University Avenue, 11PMB184, Toronto, M5G 2N2, ON, Canada. Telephone: +1 416 340 4800 ext. 5169; FAX: +1 416 340 3237; E-mail: gonzalo.sapisochin@uhn.ca

Gonzalo Sapisochin consults for, advises, and received grants from Roche. He consults for Novartis and also advises AstraZeneca.

Tommy Ivanics, Madhukar S. Patel, Marco P.A.W. Claasen, Andre Gorgen, Phillipe Abreu, Anna Goldenberg, and Lauren Erdman were responsible for literature review, interpretation of results, and write-up of the manuscript. Walter Nelson was responsible for the conception of the project, statistical analysis, literature review, interpretation of results, and write-up of the manuscript. Lawrence Lau, Lauren Erdman, and Gonzalo Sapisochin were responsible for the conception of the project, literature review, interpretation of results, and write-up of the manuscript.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**These authors share the first authorship.*

***These authors share the last authorship.*

Additional supporting information may be found in the online version of this article.

View this article online at wileyonlinelibrary.com.

DOI 10.1002/lt.26332

Methods

This study was approved by our institutional review board (REB#15-9989), and a waiver of informed consent was obtained.

STUDY POPULATION

Patients who underwent LT for HCC from 2000 to 2016 were identified from the prospectively maintained Toronto General Hospital LT database. A detailed description of listing criteria has been outlined elsewhere.[19] Moreover, in contrast to the United States, there is no mandatory 6-month waiting time for patients with HCC exception points. In Ontario specifically, patients with HCC that meet selection criteria for listing start at 22 points (Model for End-Stage Liver Disease [MELD]) and increase by 3 points every 3 months.[32] Recipients with incidentally discovered HCC in the explanted liver were excluded. Data on age, sex, body mass index, comorbidities, etiology of liver disease, and MELD score were collected. HCC-specific variables were tumor size, volume, number, and AFP levels, all at transplantation, listing, and delisting (dropout). Bridging therapy (administered or not), the timing of bridging therapy, and number of sessions were also included.

DEVELOPING A MACHINE LEARNING MODEL

A machine learning approach was used to create a model to determine the risk of posttransplantation HCC recurrence. Nearly 15% of randomly selected patients were held out as a test set. The remaining 85% of patients constituted the development set. This splitting ratio was chosen based on a classic rule of prescribing an 80%:20% split between the development and test sets, with an increased training set size of 85% to account for the rate of right censorship in the recurrence outcome.

The candidate machine learning models were selected on the basis of being representative of the main paradigms of machine learning: regularized regression, ensembled decision trees, support vector machines, and deep neural networks. CoxNet refers to the usual Cox proportional hazards model, with an added penalty term that regularizes the coefficients during model fitting. This penalty term has the effect of driving coefficients with little or no independent predictive value to 0 and shrinking other coefficients to prevent overfitting. Random survival forests are a generalization of decision trees. Whereas the output of a traditional decision tree is a probability or binary decision, the output of a survival tree is a valid cumulative hazard function. A survival random forest consists of a large number of these trees with the final prediction of the forest being the mean prediction from each individual tree. Survival support vector machines project data into high-dimensional space using a patient-patient similarity function defined over the predictors. The model then learns to rank the training samples; new samples are then ranked relative to the training samples. Lastly, DeepSurv is an alternative to the Cox proportional hazards model where the relative risk term is parameterized by an artificial neural network instead of linear regression, enabling the application of deep learning. Within the development set, the 4 machine learning algorithms were compared by 5-fold cross-validation to identify the best performing algorithm to develop the final model (Supporting Fig. 1 and Supporting Table 1). The number of folds was chosen to ensure sufficient data in each fold for model selection in each iteration of cross-validation. Of note, all the models other than CoxNet investigated for the presence of nonlinear interactions in the set of available pre-LT variables. The mean concordance for each algorithm for the optimal set of hyperparameters across the held-out folds during the cross-validation step (only on the development set) was reported to assess each model's performance. The best performing model, according to the concordance index, was trained on the full development set with the optimal set of hyperparameters, of which all non-zero coefficients are reported in Table 1. This was referred to as the final model.

TABLE 1. Our Model's Coefficients (Development Set)

Variable	Coefficient	Hazard Ratio
Age	–0.004	0.996
Number of bridging therapies	0.228	1.257
Etiology: other	–0.153	0.858
Total tumor diameter (at listing)	0.041	1.042
Largest lesion size (at listing)	0.020	1.021
log-AFP (before transplantation)	0.191	1.210
Largest lesion size (before transplantation)	0.020	1.020
Within Milan criteria (before transplantation)	–0.060	0.942
Neutrophil count (before transplantation)	0.025	1.025
Sodium (before transplantation)	–0.010	0.990
Tumor burden score (before transplantation)	0.038	1.039

NOTE: The model was fit to a standardized predictor matrix, that is, the mean and standard deviation were subtracted prior to model fitting. However, these coefficients have been rescaled according to the standard deviation to ensure that the hazard ratios are interpretable with respect to the original units (with the exception of log-AFP, which must still be interpreted on the natural log-scale).
y = (age × 0.23) + (etiology other × –0.15) + (total tumor diameter × 0.04) + (largest lesion size [at listing] × 0.02) + (log-AFP*0.19) + (largest lesion size lesion [before LT] × 0.02) + (within Milan criteria [before LT] × –0.06) + (neutrophil count [before LT] × 0.02) + (sodium [before LT] × –0.01) + (tumor burden score [before LT] × 0.04)
Age: per 1-year increase
Etiology: other (reference non-other)
Total tumor diameter: per 1-cm increase
Largest lesion size (at listing): per 1-cm increase
AFP: per 1-unit increase (ng/mL)
Largest lesion size (before LT): per 1-cm increase
Within Milan criteria (before LT): reference not within Milan criteria before LT
Neutrophil count (before LT): per 1-unit increase ($\times 10^9$/L)
Sodium (before LT): per 1-unit increase (mmol/L)
Tumor burden score (before LT): per 1-unit increase.

COMPARISON WITH PREVIOUSLY PUBLISHED MODELS

We compared the final performing machine learning model (CoxNet) with the models underlying several other HCC recurrence scores: MORAL,[23] AFP,[33] and Hazard Associated with Liver Transplantation for

TABLE 2. Coefficient Comparison

Variable/Condition	Reported Coefficient	Refitted Coefficient
a. MORAL model		
Maximum AFP ≥200 (before transplantation)	0.318	0.483
Neutrophil-to-lymphocyte ratio ≥5 (before transplantation)	0.417	0.063
Maximum lesion size ≥3 cm (before transplantation)	0.265	0.454
b. AFP model		
Maximum lesion size ≤3 cm (at listing)	—	—
3 cm < maximum lesion size ≤6 cm (at listing)	0.069	0.111
Maximum lesion size >6 cm (at listing)	0.343	0.356
Lesion count ≤3 (at listing)	—	—
Lesion count ≥4 (at listing)	0.177	0.157
AFP ≤100 (at listing)	—	—
100 < AFP ≤ 1000 (at listing)	0.170	0.097
AFP >1000 (at listing)	0.241	0.279
c. HALT-HCC model		
Tumor Burden Score (before transplantation)	0.376	0.363
log-AFP (before transplantation)	0.547	0.609
MELD-Na score (before transplantation)	0.077	–0.028

NOTE: All coefficients for each model were normalized such that their absolute values sum to one, to account for potential scale differences when comparing the between columns. Directional discrepancies are underlined. AFP units is in ng/mL. All coefficients for each model were normalized such that their absolute values sum to 1, to account for potential scale differences when comparing the between columns. Directional discrepancies are underlined.

Hepatoceullar Carcinoma (HALT-HCC).[34] These models were selected because they use pre-LT variables and evaluated a similar outcome (recurrence) which allowed for appropriate comparisons to be performed. The AFP model and the MORAL score are well-known prognostic scores used to predict HCC recurrence following LT.[15,33] The AFP model aims to identify HCC candidates with a low recurrence risk who would otherwise be excluded based on the Milan criteria. It also takes into account the largest tumor diameter, number of nodules, and the AFP level.[22] The pretransplantation MORAL score (pre-MORAL [herein referred to as MORAL score]) uses preoperative neutrophil-to-lymphocyte ratio, AFP, and the maximum tumor size to predict post-LT recurrence. The AFP variables at listing and delisting were log-transformed to satisfy the assumption of normally distributed residuals. The Tumor Burden Score,[35] an input to the HALT-HCC model, was derived and used as an engineered predictor to the machine learning models for fairness. For each comparison algorithm, the corresponding Cox model was recalibrated on our development set and evaluated on the held-out test set. As expected, this improved the resulting concordance statistics on the held-out test set for each of the comparison models. In Table 2, we report the differences between the coefficients reported in each comparison model's original publication and the coefficients found by recalibrating the Cox model on our development set. Lastly, we assessed, both before and after recalibration, whether CoxNet offered advantages over the other models by testing for improvement in model performance according to the concordance index.

FOLLOW-UP AND RECURRENCE DEFINITION

Following transplantation, patients were followed with either contrast-enhanced computed tomography of the chest and abdomen or ultrasound together with AFP measurements in 3-month intervals for the first 2 years. After that, surveillance occurs every 6 months for 2 years and then yearly. Additional imaging studies were performed for any suspected recurrence, including contrast-enhanced computed tomography, contrast-enhanced ultrasound, or magnetic resonance imaging.[36] The time to recurrence was calculated from transplantation to the first imaging study that confirmed tumor recurrence.

DATA ANALYSIS

Continuous variables were described using median and interquartile range (IQR), whereas categorical variables were described using frequency and percentage (%). Disease-free, intention-to-treat, and posttransplantation survival were evaluated using the Kaplan-Meier method in R (R Core Team).[37] All models were fit using the scikit-survival package in Python 3.6.[38] The grid search space for each model is given in the Supporting Material. All *P* values were computed via bootstrapping (the baseline procedure described by Kang et al.[39]), except where expressly noted. Two-sided *z*-tests were used in our initial models (CoxNet, survival random forest, survival support vector machine, and DeepSurv) to test the alternative hypothesis of differences in model performance. One-sided *z*-tests were used to compare our best model with 3 competing models in order to test the alternative hypothesis that our model performs better than previously published algorithms.

Results

The study data set comprised 1013 patients with HCC listed for LT. Of these, 831 (82%) were male, and the most common cause of underlying liver disease was chronic hepatitis C (51.7%). At listing, 304 (30%) patients were beyond Milan criteria. Of the listed patients, 739 (73.0%) underwent LT, of which 142 (19.2%) had grafts from a living donor. While on the waiting list, 625 (61.7%) underwent bridging treatment, with 241 (38.6%) having more than 1 treatment. Of a total of 977 bridging therapy treatments, the majority received either radiofrequency ablation (564 treatments [57.7%]) or transarterial chemoembolization (311 treatments [31.8%]). The median time on the waiting list was 6.1 months (IQR, 3.0-10.3). During the wait time, 269 (26.6%) patients dropped out and 5 had not experienced an event (dropout or transplantation) by the end of the study follow-up. Baseline patient characteristics are summarized in Table 3.

Among the patients who underwent LT, 143 (19.4%) had tumor recurrence after a median follow-up of 4.5 years (IQR, 2.0-8.9). The 1-, 3-, and 5-year disease-free survival rates were 91.5%, 82.4%, and 79.9%, respectively (Fig. 1A). Most recurrences (79%) occurred in the first 3 years after LT, as demonstrated by Fig. 2. The median intention-to-treat overall survival was 3.4 years (IQR, 1.5-7.4). The 1-, 3-, and 5-year intention-to-treat overall survival rates were 83.3%, 63.8%, and 55.5%, respectively (Fig. 1B).

When using these data to fit CoxNet, survival random forest, survival support vector machine, and DeepSurv models, CoxNet performed the best according to cross-validation within the development set (Supporting Table 1), although this was not statistically significant. The characteristics of the derivation and test cohorts are provided in Supporting Table 2. Based on both model performance and parsimony, CoxNet was selected as the final model, and trained on the full development set. The CoxNet model being the best performing model also signified that nonlinear interactions were not included. The selected optimal hyperparameters include an L1-ratio of 0.55 (commonly referred to in the elastic net literature as α), suggesting that the benefit of CoxNet is its ability to both automatically exclude meaningless predictors via L1-regularization and reduce the contributions of less meaningful predictors via L2-regularization. The inclusion of particular variables is always determined by their contribution to predictive performance, rather than their individual statistical significance—a fundamental difference between the regularization-based and forward selection–based approaches, respectively.

TABLE 3. Patients Baseline Characteristics

Variable	Total (n = 1013)
Male sex, n (%)	831 (82.0)
Age (years), median (IQR)	59 (53.6-63.7)
BMI (kg/m^2), median (IQR)	26.9 (24.2-30.4)
MELD score at listing, median (IQR)	10 (8-14)
AFP level at listing (ng/mL), median (IQR)	11 (5-45)
Etiology, n (%)	
HCV	524 (51.7)
HBV	204 (20.1)
ETOH	138 (13.6)
NASH	67 (6.6)
Other	80 (7.9)
Months on waiting list, median (IQR)	6.1 (3.0-10.3)
Median tumor size at listing (cm), median (IQR)	2.8 (1.9-3.9)
Tumor number at listing, median (IQR)	1 (1-2)
Within Milan criteria at listing, n (%)	709 (70.0)
Bridging therapy (yes), n (%)	625 (61.7)
Number of bridging therapies, n (%)	
0	388 (38.3)
1	384 (37.9)
2	156 (15.4)
3	64 (6.3)
4	19 (1.9)
5	2 (0.2)
Dropout rate while on waiting list, n (%)	269 (26.6)
Median tumor size before LT (cm), median (IQR)	1.6 (0.0-3.0)
Tumor number before LT, median (IQR)	1 (0-2)
Within Milan criteria before LT, n (%)	580 (78.5)
Transplanted, n (%)	739 (73.0)
Living donor liver graft, n (%)	142 (19.2)
Milan on pathology, n (%)	368 (49.9)
Median tumor size on pathology (cm), median (IQR)	3.0 (2.0-4.0)
Tumor number on pathology, median (IQR)	5 (3-8)
Tumor differentiation, n (%)	
Well	89 (14.5)
Moderate	457 (74.3)
Poor	69 (11.2)
Unable to be assessed/missing	124
Microvascular invasion (yes), n (%)	207 (28.0)
Follow-up of transplanted patients (years), median (IQR)	4.5 (2.0-8.9)
Recurrence rate, n (%)	143 (19.4)

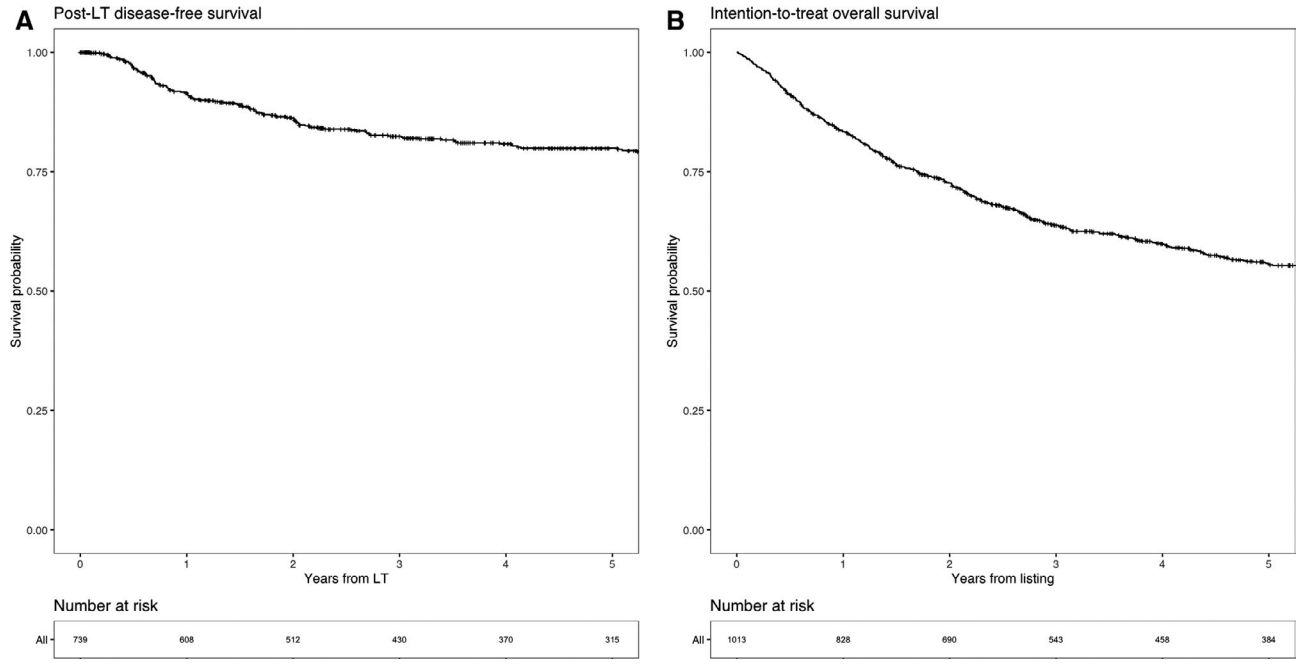

FIG. 1. (A) Posttransplantation disease-free survival. (B) Intention-to-treat overall survival.

Within the held-out test set, we next assessed whether our CoxNet model outperformed the AFP, MORAL, and HALT-HCC scores, using the coefficients given in their original publications. Indeed, our final model, with a concordance of 0.75, outperformed the MORAL score, with a concordance of 0.62 and a difference of 0.13 (1-sided 95% confidence interval [CI], >0.04; $P = 0.01$); the AFP score, with a concordance of 0.63 and a difference of 0.12 (1-sided 95% CI, >0.03; $P = 0.02$); and the HALT-HCC score, with a concordance of 0.64 and a difference of 0.10 (1-sided 95% CI, >0.03; $P = 0.02$). Lastly, the conventional models were recalibrated on the development set used to train our machine learning model, and we subsequently assessed whether the gain in performance of the machine learning approach holds over the recalibrated models. The recalibrated MORAL score, with a concordance of 0.64, continued to be outperformed by the CoxNet model (1-sided 95% CI, >0.02; $P = 0.03$). Similarly, the recalibrated AFP score, with a concordance of 0.64, was outperformed by our model (1-sided 95% CI, >0.01; $P = 0.04$). However, the recalibrated HALT-HCC score performed well (with a concordance of 0.72) and was not significantly outperformed (1-sided 95% CI, ≥0.048; $P = 0.29$).

As the final model is linear, its coefficients (Table 1) can be interpreted as in other Cox models. In particular, the risk score can be computed by multiplying the coefficient for each variable with its value and summing up the products. In addition, the final machine learning model is available as an online calculator at https://hcccalculator.ccm.sickkids.ca.

Discussion

This study aimed to develop an accurate posttransplantation HCC recurrence calculator using available clinicopathologic data. A machine learning approach was used due to the vast number of potentially predictive factors and the possibility of multiple nonlinear interactions. Our results demonstrate the feasibility of applying machine learning in transplantation oncology and suggest that this risk prediction method provides improved accuracy over other currently available risk scores, including AFP and MORAL.

Four well-known supervised learning algorithms were developed in this study, with the CoxNet model selected based on model performance and parsimony. To assess this new Toronto HCC recurrence calculator's validity, its performance was compared with 3 well-known prognostic calculators—the AFP model,[33] the MORAL score,[40] and HALT-HCC

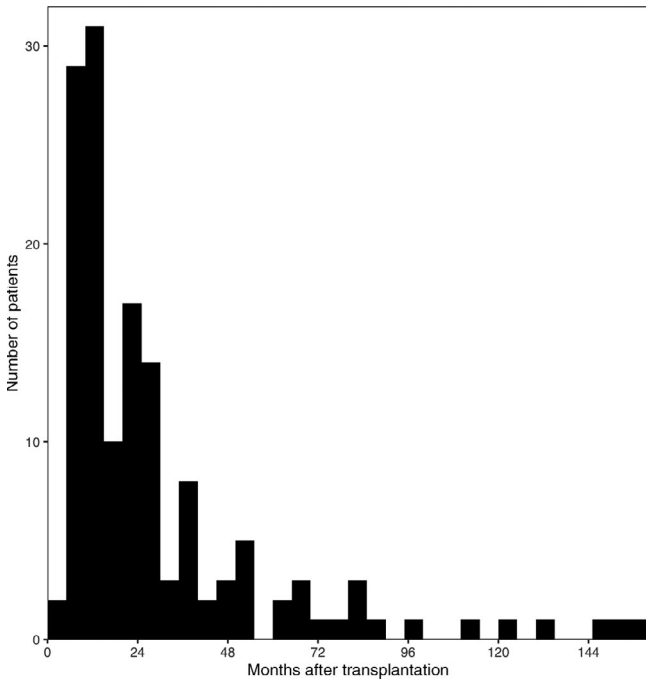

FIG. 2. Distribution of recurrences over time (months).

score.[34] The coefficients derived for the recalibrated AFP and MORAL models largely resemble those provided in the original publications. By contrast, the MELD–sodium (MELD-Na) coefficient derived for the HALT-HCC score in our data set was directionally different from the published coefficient, suggesting a discrepancy. Notably, the MELD-Na in the HALT-HCC was the only directionally discrepant variable across all 3 comparison models. The improved performance of the Toronto HCC recurrence calculator may be due, in part, to the methodologic advantages of the machine learning approach. The HALT-HCC score was generated using a cohort of 420 patients from Cleveland Clinic and subsequently validated in a larger cohort of US liver recipients from the Scientific Registry of Transplant Recipients (SRTR), where Cleveland Clinic patients are included.[34] In the subsequent international validation cohort by Firl et al., of the 4089 patients included, 1851 were from North America (of which the US constituents were part of the SRTR data; 460 [45.1%] in the training and 1391 [45.3%] in the validation cohorts, respectively).[41] Given this overlap in patients in the development and subsequent validation cohort, it is conceivable that this may have resulted in overly optimistic c-indices. Ideally, a fair assessment of the performance of the present machine learning algorithm (MLA) model and the HALT-HCC should thus be done in an external patient cohort. In other words, a cohort that does not contain any patients used to derive the HALT-HCC or patients from Toronto. Notwithstanding this, this model development aimed not necessarily to outperform currently available models per se, but rather to evaluate the feasibility of MLA as a proof-of-concept study and shed light on various techniques and potential pitfalls in the process. By statistically appraising all variables in the data set, potentially informative features were integrated without additional selection bias. Notably, the critical variable selection step sifted out poorly predictive or collinear variables, limiting noise and over-fitting, and in this way worked to optimize the model.

Although our model exhibited strong predictive performance using internal data, external validation is required before the model can be more broadly applied.[42] Within this context, given that many transplantation centers restrict listing to patients that meet the Milan criteria, it is unclear how well the present model would perform in that setting, especially given that the present cohort included a higher proportion of patients that exceeded Milan criteria. In a study by Schrem et al., a prognostic model developed in Germany to predict 90-day post-LT mortality based on pretransplantation donor and recipient variables could not be validated in a cohort of patients from the United Kingdom.[43] This highlighted the challenge of suboptimal translation between 2 transplantation environments with different donor/recipient populations, health care systems, allocation policies, and clinical/surgical practices. Because of the long-standing application of the Extended Toronto Criteria,[19] in our model, patients were not excluded based on HCC size or number alone, resulting in 30% of the patients in the data set falling beyond Milan criteria. This further reflects the ability of the algorithm's predictive performance to be maintained for patients who are beyond conventional macromorphological transplantation criteria. Further, a prediction model validated on internal data alone will likely yield the most optimistic representation of the model. Nonetheless, it remains to be determined whether internally derived prognostic models are required to be generalizable to other settings or whether individual health care systems should optimize models to serve their specific population best.[44,45] Regardless, as this study suggests, incorporating factors reflective of tumor biology, with an emphasis on excellent covariate fidelity and

granularity, remains an essential tenet in developing a calculator with high accuracy.

Over the past 2 decades, machine learning algorithms have been increasingly applied for cancer diagnosis, prognostication, and treatment outcome prediction.[31,42,46,47] Recently, an MLA approach based on a random forest workflow has been developed by a group in Germany to predict disease-free survival after liver resection for HCC.[48] This model had a robust predictive potential for early recurrence with an area under the curve of 0.79 (0.66-0.92).[48] In our model, recurrence was selected rather than recurrence-free survival (which considers both death and recurrence as an event), as it may potentially offer greater clinically relevant insight to refining postoperative management, given that we censored patients who died from non-HCC related reasons. As the standard machine learning workflow involves model performance monitoring and retraining to account for model drift, a multidisciplinary partnership between clinicians and data scientists is required, with a commitment to the curation and iterative maintenance of data sets to allow for the development of meaningful decision-support tools.[45] This process should involve, first and foremost, a robust, consistent, and objective means of collecting data. The data may be clinicopathologic characteristics from electronic medical records, genomics, and imaging studies.[31] Clinicians should strive to establish interdisciplinary partnerships that strive toward a common goal, rather than a "turf and credit" mindset. Leveraging the knowledge and technologies of such partners can achieve synergism. For instance, clinicians help provide a clinically relevant outcome, and data scientists can identify the optimal methodology to make predictions for the outcome based on the available data. The HCC recurrence calculator developed in this study demonstrates the potential for integrating machine learning in transplantation oncology. Increased accuracy in outcome prediction enables clinicians and patients to make better-informed decisions regarding their care. In the case of HCC, where LT is a potentially curative treatment modality, the importance of a quantifiable and accurate recurrence calculator is of particular relevance to ensure fair and equitable access to the limited number of available donor organs.

This study is limited by its retrospective single-center study design. Further, the generalizability of the results requires external validation. Although machine learning results yield high-performance prediction models, several additional limitations warrant mention. First, the quality of the data output is dependent on the quality of data input. Objective data are thus preferred over subjective data, such as the modified Response Evaluation Criteria in Solid Tumors (mRECIST), which may differ between institutions and radiologists, for model input, the latter of which is prone to bias. As a surrogate for response, the size of the largest tumor size at both listing and before LT were included. The latter variable captured the size of the viable portion of the tumor, which was smaller when there was a successful response to bridging therapy. Although the variables used in this study were objective, misclassification bias may affect the validity of model performance. Second, despite the high predictive capacity of machine learning models, there is potential for prediction "overfitting," which may generate overly optimistic results.[49] This limitation may be overcome by external validation before algorithms are adopted for clinical use. Previous non-MLA–based prognostic scores have included AFP response and found it to be predictive of outcomes.[24] In the present MLA-based model, both AFP at listing and AFP before LT were introduced into the model, but only AFP before LT remained, as it resulted in improved model predictive performance. Because AFP response is a linear combination of AFP before transplantation and AFP at listing, all models, including the selected elastic net model, implicitly consider AFP response. As with other HCC prognostic calculators, it is important to note that the time point at which the values of relevant variables are known dictates the clinical time point when the model predictions are relevant. For instance, variables known just before LT are only relevant for patients who can reach that point and not dropout from the waiting list. As such, our model, similar to many of the other well-known HCC prediction models, would be unable to guide the decision of whether or not to list a patient for transplantation given that it does not represent an intention-to-treat analysis of survival but rather than a per-protocol (from the time of LT) prediction for recurrence. The utility of this model is therefore primarily to provide a prognostic assessment of oncologic outcomes from the time of transplantation, which may potentially help individualize screening for recurrence or lower thresholds for early institution of future adjuvant therapies/clinical trial inclusion based on predicted recurrence risk. Future models should also seek to evaluate the prognostic performance of transplantation outcomes from before or at the time of listing, as this would potentially help refine current patient selection

for LT in the treatment of HCC. Lastly, variables from multiple time points (transplantation listing, bridging therapy, and transplantation) were incorporated into this model. This contrasts with de facto data obtained across multiple institutions in electronic health records to facilitate billing or patient care, which may lack the same standardization, completeness, or granularity.[44] This may, in turn, limit the calculator's performance in those settings.[47] Notwithstanding these limitations, machine learning algorithms represent a powerful statistical platform that can improve clinical decision making and, most importantly, patient outcomes.

Conclusion

The development of a posttransplantation HCC recurrence risk calculator using machine learning is feasible using a comprehensive data set of relevant patient and tumor features before LT. This proof-of-concept study underscores the potential of a machine learning approach to augment individual clinical decision making and help safeguard equitable organ allocation.

REFERENCES

1) Bruix J, Sherman M, American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. Hepatology 2011;53:1020-1022.
2) Forner A, Reig M, Bruix J. Hepatocellular carcinoma. Lancet 2018;391:1301-1314.
3) Bruix J, Reig M, Sherman M. Evidence-based diagnosis, staging, and treatment of patients with hepatocellular carcinoma. Gastroenterology 2016;150:835-853.
4) Llovet JM, Ducreux M, Lencioni R, Di Bisceglie AM, Galle PR, Dufour JF, et al. EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2012;56:908-943.
5) Johnson RJ, Bradbury LL, Martin K, Neuberger J. Organ donation and transplantation in the UK—the last decade: a report from the UK national transplant registry. Transplantation 2014;97(suppl. 1):S1-S27.
6) Neuberger J, James O. Guidelines for selection of patients for liver transplantation in the era of donor-organ shortage. Lancet 1999;354:1636-1639.
7) Kwong A, Kim WR, Lake JR, Smith JM, Schladt DP, Skeans MA, et al. OPTN/SRTR 2018 annual data report: liver. Am J Transplant 2020;20(suppl 1):193-299.
8) Freeman RB, Edwards EB, Harper AM. Waiting list removal rates among patients with chronic and malignant liver diseases. Am J Transplant 2006;6:1416-1421.
9) Northup PG, Intagliata NM, Shah NL, Pelletier SJ, Berg CL, Argo CK. Excess mortality on the liver transplant waiting list: unintended policy consequences and model for End-Stage Liver Disease (MELD) inflation. Hepatology 2015;61:285-291.
10) Mehta N, Dodge JL, Goel A, Roberts JP, Hirose R, Yao FY. Identification of liver transplant candidates with hepatocellular carcinoma and a very low dropout risk: Implications for the current organ allocation policy. Liver Transpl 2013;1343-1353.
11) Mazzaferro V, Sposito C, Coppa J, Miceli R, Bhoori S, Bongini M, et al. The long-term benefit of liver transplantation for hepatic metastases from neuroendocrine tumors. Am J Transplant 2016;16:2892-2902.
12) Yeh H, Smoot E, Schoenfeld DA, Markmann JF. Geographic inequity in access to livers for transplantation. Transplantation 2011;91:479-486.
13) Mazzaferro V. Squaring the circle of selection and allocation in liver transplantation for HCC: an adaptive approach. Hepatology 2016;63:1707-1717.
14) Mazzaferro V, Llovet JM, Miceli R, Bhoori S, Schiavo M, Mariani L, et al. Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond the Milan criteria: a retrospective, exploratory analysis. Lancet Oncol 2009;10:35-43.
15) Mazzaferro V, Sposito C, Zhou J, Pinna AD, De Carlis L, Fan J, et al. Metroticket 2.0 model for analysis of competing risks of death after liver transplantation for hepatocellular carcinoma. Gastroenterology 2018;154:128-139.
16) Bruix J, Fuster J, Llovet JM. Liver transplantation for hepatocellular carcinoma: Foucault pendulum versus evidence-based decision. Liver Transpl 2003;9:700-702.
17) Llovet JM. Expanding HCC criteria for liver transplant: the urgent need for prospective, robust data. Liver Transpl 2006;12: 1741-1743.
18) Mazzaferro V, Battiston C, Sposito C. Pro (with caution): extended oncologic indications in liver transplantation. Liver Transpl 2018;24:98-103.
19) Sapisochin G, Goldaracena N, Laurence JM, Dib M, Barbas A, Ghanekar A, et al. The extended Toronto criteria for liver transplantation in patients with hepatocellular carcinoma: a prospective validation study. Hepatology 2016;64:2077-2088.
20) Carlis L, Giacomoni A, Lauterio A, Slim A, Sammartino C, Pirotta V, et al. Liver transplantation for hepatocellular cancer: should the current indication criteria be changed? Transplant Int 2003;16:115-122.
21) De Carlis L, Giacomoni A, Pirotta V, Lauterio A, Slim AO, Sammartino C, et al. Surgical treatment of hepatocellular cancer in the era of hepatic transplantation. J Am Coll Surg 2003;196: 887-897.
22) Notarpaolo A, Layese R, Magistri P, Gambato M, Colledan M, Magini G, et al. Validation of the AFP model as a predictor of HCC recurrence in patients with viral hepatitis-related cirrhosis who had received a liver transplant for HCC. J Hepatol 2017;66:552-559.
23) Halazun KJ, Najjar M, Abdelmessih RM, Samstein B, Griesemer AD, Guarrera JV, et al. Recurrence after liver transplantation for hepatocellular carcinoma: a new MORAL to the story. Ann Surg 2017;265:557-564.
24) Halazun KJ, Tabrizian P, Najjar M, Florman S, Schwartz M, Michelassi F, et al. Is it time to abandon the Milan criteria? Results of a bicoastal US collaboration to redefine hepatocellular carcinoma liver transplantation selection policies. Ann Surg 2018;268: 690-699.
25) Agopian VG, Harlander-Locke MP, Ruiz RM, Klintmalm GB, Senguttuvan S, Florman SS, et al. Impact of pretransplant bridging locoregional therapy for patients with hepatocellular carcinoma within Milan criteria undergoing liver transplantation: analysis of 3601 patients from the US Multicenter HCC Transplant Consortium. Ann Surg 2017;266:525-535.
26) Lee DD, Samoylova M, Mehta N, Musto KR, Roberts JP, Yao FY, Harnois DM. The mRECIST classification provides insight into tumor biology for patients with hepatocellular carcinoma awaiting liver transplantation. Liver Transpl 2019;25:228-241.

27) Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. JAMA 2019;322:1806-1816.
28) Hibi T, Sapisochin G. What is transplant oncology? Surgery 2019;165:281-285.
29) Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1-22.
30) Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg 2018;268: 70-76.
31) Ivanics T, Patel MS, Erdman L, Sapisochin G. Artificial intelligence in transplantation (machine-learning classifiers and transplant oncology). Curr Opin Organ Transplant 2020;25:426-434.
32) Brahmania M, Marquez V, Kneteman NM, Bhat M, Marleau D, Wong P, et al. Canadian liver transplant allocation for hepatocellular carcinoma. J Hepatol 2019;71:1058-1060.
33) Duvoux C, Roudot–Thoraval F, Decaens T, Pessione F, Badran H, Piardi T, et al. Liver transplantation for hepatocellular carcinoma: a model including α-fetoprotein improves the performance of Milan criteria. Gastroenterology 2012;143:985-986.
34) Sasaki K, Firl DJ, Hashimoto K, Fujiki M, Diago-Uso T, Quintini C, et al. Development and validation of the HALT-HCC score to predict mortality in liver transplant recipients with hepatocellular carcinoma: a retrospective cohort analysis. Lancet Gastroenterol Hepatol 2017;2:595-603.
35) Sasaki K, Morioka D, Conci S, Margonis GA, Sawada YU, Ruzzenente A, et al. The tumor burden score: a new, "metroticket" prognostic tool for colorectal liver metastases based on tumor size and number of tumors. Ann Surg 2018;267:132-141.
36) Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, 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.
37) R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2019.
38) Pölsterl S. Scikit-survival: a library for time-to-event analysis built on top of scikit-learn. J Mach Learn Res 2020;21:1-6.
39) Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. Stat Med 2015;34:685-703.
40) Galle PR, Tovoli F, Foerster F, Wörns MA, Cucchetti A, Bolondi L. The treatment of intermediate stage tumours beyond TACE: from surgery to systemic therapy. J Hepatol 2017;67:173-183.
41) Firl DJ, Sasaki K, Agopian VG, Gorgen A, Kimura S, Dumronggittigule W, et al. Charting the path forward for risk prediction in liver transplant for hepatocellular carcinoma: international validation of HALTHCC among 4089 patients. Hepatology 2020;71:569-582.
42) Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17.
43) Schrem H, Focken M, Gunson B, Reichert B, Mirza D, Kreipe H-H, et al. The new liver allocation score for transplantation is validated and improved transplant survival benefit in Germany but not in the United Kingdom. Liver Transpl 2016;22:743-756.
44) Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Informatics Assoc 2017;24:198-208.
45) Sendak M, Gao M, Nichols M, Lin A, Balu S. Machine learning in health care: a critical appraisal of challenges and opportunities. EGEMS (Wash DC) 2019;7:1.
46) Singal AG, Mukherjee A, Elmunzer JB, Higgins PDR, Lok AS, Zhu JI, et al. Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol 2013;108:1723-1730.
47) Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-1358.
48) Schoenberg MB, Bucher JN, Koch D, Börner N, Hesse S, De Toni EN, et al. A novel machine learning algorithm to predict disease free survival after resection of hepatocellular carcinoma. Ann Transl Med 2020;8:434.
49) Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375: 1216-1219.