Title

Predicting post—liver transplant outcomes in patients with acute-on-chronic liver failure using Expert-Augmented Machine Learning

Permalink
https://escholarship.org/uc/item/10q1z73c

Journal
American Journal of Transplantation, 23(12)

ISSN
1600-6135

Authors
Ge, Jin
Digitale, Jean C
Fenton, Cynthia
et al.

Publication Date
2023-12-01

DOI
10.1016/j.ajt.2023.08.022

Peer reviewed

Published in final edited form as:

Am J Transplant. 2023 December ; 23(12): 1908–1921. doi:10.1016/j.ajt.2023.08.022.

Predicting post–liver transplant outcomes in patients with acute-on-chronic liver failure using Expert-Augmented Machine Learning

Jin Ge[1,*], **Jean C. Digitale**[2], **Cynthia Fenton**[3], **Charles E. McCulloch**[2], **Jennifer C. Lai**[1], **Mark J. Pletcher**[2], **Efstathios D. Gennatas**[2]

[1]Division of Gastroenterology and Hepatology, Department of Medicine, University of California–San Francisco, San Francisco, California, USA

[2]Department of Epidemiology and Biostatistics, University of California–San Francisco, San Francisco, California, USA

[3]Division of Hospital Medicine, Department of Medicine, University of California–San Francisco, San Francisco, California, USA

Abstract

Liver transplantation (LT) is a treatment for acute-on-chronic liver failure (ACLF), but high post-LT mortality has been reported. Existing post-LT models in ACLF have been limited. We developed an Expert-Augmented Machine Learning (EAML) model to predict post-LT outcomes. We identified ACLF patients who underwent LT in the University of California Health Data Warehouse. We applied the RuleFit machine learning (ML) algorithm to extract rules from

[*]Corresponding author. Jin Ge, 513 Parnassus Avenue, S-357 San Francisco, CA 94143, USA., jin.ge@ucsf.edu (J. Ge).

Author contributions

Authorship was determined using ICMJE recommendations.

J. G.: Study concept and design; data extraction; analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content; statistical analysis; obtained funding; study supervision

J.C.D.: Analysis and interpretation of data; drafting of manuscript; critical revision of the manuscript for important intellectual content

C. F.: Data extraction; critical revision of the manuscript for important intellectual content

C.E.M.: Interpretation of data; critical revision of the manuscript for important intellectual content

J.C.L.: Study concept and design; interpretation of data; critical revision of the manuscript for important intellectual content; study supervision

M.J.P.: Interpretation of data; critical revision of the manuscript for important intellectual content

E.D.G.: Study concept and design; analysis and interpretation of data; critical revision of the manuscript for important intellectual content

Declaration of competing interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jin Ge reports financial support was provided by National Center for Advancing Translational Sciences. Jin Ge reports financial support was provided by American Association for the Study of Liver Diseases. Jin Ge reports financial support was provided by American Society of Transplantation. Jin Ge reports financial support was provided by UCSF Liver Center. Charles E. McCulloch reports financial support was provided by National Center for Advancing Translational Sciences. Mark J. Pletcher reports financial support was provided by National Center for Advancing Translational Sciences. Jennifer C. Lai reports financial support was provided by UCSF Liver Center. Jean C. Digitale reports financial support was provided by National Heart Lung and Blood Institute. Jennifer C. Lai reports financial support was provided by National Institute on Aging. Jin Ge reports a relationship with Merck & Co Inc that includes: funding grants.

Disclosure

The authors of this manuscript have conflicts of interest to disclose as described by the American Journal of Transplantation. J. Ge receives research support from Merck and Co, and consults for Astellas Pharmaceuticals.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ajt.2023.08.022.

decision trees and create intermediate models. We asked human experts to rate the rules generated by RuleFit and incorporated these ratings to generate final EAML models. We identified 1384 ACLF patients. For death at 1 year, areas under the receiver-operating characteristic curve were 0.707 (confidence interval [CI] 0.625–0.793) for EAML and 0.719 (CI 0.640–0.800) for RuleFit. For death at 90 days, areas under the receiver-operating characteristic curve were 0.678 (CI 0.581–0.776) for EAML and 0.707 (CI 0.615–0.800) for RuleFit. In pairwise comparisons, both EAML and RuleFit models outperformed cross-sectional models. Significant discrepancies between experts and ML occurred in rankings of biomarkers used in clinical practice. EAML may serve as a method for ML-guided hypothesis generation in further ACLF research.

Keywords

ACLF; big data; UCHDW; machine learning; posttransplant outcomes

Introduction

Acute-on-chronic liver failure (ACLF) is commonly defined as acute decompensation of end-stage liver disease (ESLD) with extrahepatic organ failure and is associated with high short-term mortality.[1–6] Liver transplantation (LT) is a well-established treatment for patients with ACLF who are refractory to supportive care and treatment for the underlying precipitant. Due to critical illness, however, LT is estimated to be feasible in only 25% of ACLF patients.[7] Moreover, there have been conflicting post-LT outcomes reported for ACLF patients, with some subpopulations having up to 40% 3-month mortality.[8,9] Of note, analyses of the United Network for Organ Sharing (UNOS) database showed that among patients with severe ACLF, mechanical ventilation at transplant and receipt of an organ with an elevated donor risk index were associated with increased post-LT mortality. The derivation of this UNOS-based model, however, required the inclusion of post-LT variables.[10] There is, therefore, still an unmet need for tools to predict post-LT outcomes for ACLF patients in the pre-LT setting (and without the benefit of donor, intraoperative, or post-LT data) to aid with clinical decision-making regarding utility of proceeding to transplantation.[11,12]

Multiple international research consortia, such as the North American Consortium for the Study of End-Stage Liver Disease (NACSELD),[2] the European Association for the Study of the Liver-Chronic Liver Failure Consortium (EF-CLIF),[3] and the Asian Pacific Association for the Study of the Liver ACLF Research Consortium (APASL ACLF)[13]; have developed scoring systems to predict pre-LT outcomes. None of these models, however, evaluates for post-LT outcomes. Currently, there are 2 models that utilize pre-LT data to predict post-LT outcomes. The first is the transplantation for ACLF-3 model (TAM) score, which was trained on 76 patients with severe ACLF at a single French center and validated in 76 patients at 4 other centers.[14] Despite its potential utility, the TAM model has not been studied in non-European settings. More recently, the Sundaram ACLF-LT-Mortality (SALT-M) score, derived from data from 15 liver transplant centers in the United States and validated in 2 French centers, was shown to have an area under the receiver-operating characteristic curve (AUROC) of 0.72 and outperformed the Model for End-Stage

Liver Disease (MELDNa) and its derivatives in assessing post-LT outcomes.[15] The dearth of models predicting post-LT outcomes, however, illustrates the inherent difficulties of modeling this heterogeneous and dynamic clinical syndrome with divergent definitions in different geographies.2,3,5,7,13 Many prediction models do not utilize vast numbers of data features available in electronic health records (EHR) to better define dynamic clinical trajectories seen in patients with ACLF. Our group had previously demonstrated an informatics approach to extract EHR data that yielded a median of 454 features per admission to more accurately represent ACLF patients' clinical courses.[16] Machine learning (ML) is well-suited for analyzing such data but can be misleading when taken out of context of biological or clinical mechanisms.[17,18]

Expert-Augmented Machine Learning (EAML) is an emerging technique that overcomes this limitation of ML by extracting rules from decision-tree ML models for human expert feedback. EAML has 2 potential benefits: (1) to create combined models that incorporate the best of human and ML knowledge, and (2) to evaluate for differences between humans and ML. These differences could represent human biases (eg, experts ignoring important variables identified by ML) or artifacts in the underlying data (eg, experts are identifying the important variables but there is overrepresentation of characteristics or variables in this population not seen by typically seen by human experts, such as differences in etiologies of ACLF or underlying liver disease).

In this study, we utilized a novel multicenter EHR database, the University of California Health Data Warehouse (UCHDW), to construct an EAML model to predict post-LT outcomes in patients with ACLF.

Methods

A flowchart showing the study design is featured in Figure 1.

The University of California Health Data Warehouse (UCHDW)

The UCHDW is a unique data asset created from the EHRs and claims data from the 5 major University of California Health (UCH) Medical Centers (Davis, Irvine, Los Angeles, San Diego, and San Francisco) and managed by the Center for Data-Driven Insights and Innovation (CDI2).[19] UCHDW holds data on 6.2+ million patients seen at UCH since 2012. All data in UCHDW are harmonized in the Observational Medical Outcomes Partnership (OMOP) common data model, version 5.3.1.[20] All data elements in UCHDW are deidentified prior to the receipt by end-users with no clinical notes or imaging. UCHDW has previously been utilized to analyze treatment utilization patterns between UCH Medical Centers.[21] For all analyses, we utilized UCHDW, versioned as of September 22, 2022, and accessed on October 20, 2022.

Study population

We isolated all adults (≥18 years) who underwent an orthotopic liver transplantation procedure, as defined by the OMOP concept identifiers 2109321 or 4067458, based on the ATHENA OMOP vocabulary dictionary,[22] in UCHDW between January 1, 2013, through December 31, 2021. We included patients who had evidence of ACLF prior to the time of LT through a previously published informatics-driven approach validated by manual chart review that showed 88% and 98% positive predictive value for identifying patients with ACLF based on NACSELD and EF-CLIF definitions, respectively. Consistent with this methodology, we excluded all patients who underwent transplant within 48 hours of admission as they were likely admitted electively.[16] We included patients who underwent multiorgan (such as simultaneous liver-kidney transplant) and retransplant procedures. We did not use the APASL ACLF diagnostic criteria due to bacterial infection being the most common precipitant of ACLF in patients in the United States.[23,24]

Measurements

We extracted all structured clinical information associated with the admission of interest. Baseline characteristics included age, sex, race/ethnicity, height, weight, body mass index, and censored identity of the UCH facility (defined as “UC-1,” “UC-2,” and “UC-3”). Laboratory measurements, liver disease etiologies, complications of cirrhosis, comorbid medical conditions, dialysis state, ventilation parameters, and vasopressor administration were extracted based on previously defined OMOP concept identifiers.[22,25]

As patients may have had different lengths of stay before LT, we focused only on data values from the day of admission and the day before LT. We dropped measurements from other time points from consideration to normalize the data and minimize unintentional overfitting. Continuous data features were averaged by hospitalization day. We defined changes between admission and transplant based on the differences between data features between admission and the day before LT.[26–28] All intraoperative data values and values after transplant were excluded from our analyses as our intent was to develop a predictive model utilizing only pretransplant data. Missing data features and variables underwent single imputation with chained RFs, which has been shown to produce low errors and good performances in previous studies utilizing EHR data.[29–31]

Outcomes

The primary outcome was all-cause mortality at 1 year after LT defined based on the date of transplant. The secondary outcomes included the following: (1) all-cause readmissions within 90 days defined as hospitalizations taking place within 90 days from the date of discharge of the index transplantation hospitalization, and (2) all-cause mortality within 90 days after LT defined based on the date of transplant. Death was ascertained based on synchronized data with the California Death Registry and updated monthly.[19]

Model development and EAML

The sample of ACLF patients isolated from UCHDW was split by random sampling into training, validation, and test sets in a 60:20:20 ratio.[32–34] The training set was used to fit the model, the validation set was utilized to tune hyperparameters through a grid-search strategy, and the test set was held-out for independent testing. Learning curves, which show changes in model performance with addition of incremental training data, and calibration curves (both raw and calibrated) were generated (and reported in the Supplementary Material in Supplemental Figures 4–9). We then utilized EAML, as implemented in the rtemis R package, version 0.91, to train 1 ML model for each of our primary and secondary outcomes of interest (total of 3 models).[35] rtemis is a platform for advanced ML research and applications, which incorporates several algorithms, including EAML.[36]

As described above, EAML is an ensemble ML algorithm that incorporates human knowledge by converting high-dimensional training data into Likert scale questions.[35] EAML first trains a predictive model using the RuleFit algorithm,[37] which is a combination of a Gradient Boosting Machine (GBM) decision-tree model (trained on the data to generate rules), and a least absolute shrinkage and selection operator (LASSO) model (used to select rules generated by the GBM model).[37] The RuleFit model training outputs include the detailed rules, model coefficients (represents the change in response associated with the rule), and empirical risk (rating of the rule importance by the machine).

Utilizing the rules selected by RuleFit, we then created an online survey on the Qualtrics platform (example question in Fig. 2) that was sent to 15 hepatologists throughout the world who conduct clinical care and research in ACLF recruited from a convenience sample. These experts were asked to rate rules on a 5-point Likert scale based on perceived associations with the outcomes of interest. We calculated expert rankings based on the averages of these ratings. We then took the differences in rankings between the experts and those generated by the RuleFit model to calculate penalties. These penalties were then incorporated into the RuleFit models by eliminating the top quartile of the most discrepant rules (highest fourths of absolute rank differences between RuleFit and expert rankings) to create the EAML models for each of the 3 outcomes.[35]

Statistical analyses and model performance evaluation

Clinical characteristics and laboratory data were summarized by medians and interquartile ranges (IQR) for continuous variables or numbers and percentages (%) for categorical variables. Comparisons between the training, validation, and test sets were performed using chi-square and Kruskal-Wallis tests where appropriate.

We evaluated the performances of EAML (with expert input) and RuleFit (without expert input) models through the AUROC, which has been used previously to evaluate ML models in transplant hepatology.[38–41] In addition to AUROC, we also calculated (and reported in the Supplementary Material) the area under the precision-recall curve (AUPRC), which may be more informative in models for imbalanced data.[42] To compare the performances of the EAML and RuleFit models versus cross-sectional models (MELDNa, NACELD-ACLF, CLIF--C-ACLF, TAM) and other ML algorithms (Random Forest [RF], GBM, and Elastic-Net Regularized Generalized Linear Model [GLMNET]), we calculated AUROC and AUPRC differences between each pair of models (eg, AUROC differences between EAML and NACSELD) and their confidence intervals (CIs) using bootstrapping with 2000 iterations per pairwise comparison.[43,44] We calculated MELDNa, NACSELD-ACLF, CLIF-C-ACLF, and TAM scores per previously published literature.[2,3,14,45] We used rtemis implementations of RF, GBM, and GLMNET to generate comparison ML models.

All data queries, extractions, and transformations of OMOP concept identifiers in UCHDW were conducted using the Microsoft Azure implementations of Spark, version 2.12. All statistical analyses were performed utilizing Spark-R, version 4.1.3 “One Push-Up” (R Core

Team), and R packages previously noted and documented in Supplementary Material.[46] Two-sided *P* values <.05 were considered statistically significant in all analyses. The use of UCHDW data for this study was authorized by the Institutional Review Board at the University of California, San Francisco under #20-32717 for model generation and #22-37555 for human expert input.

Results

A total of 1384 patients with ACLF were identified from UCHDW from January 1, 2013, through December 31, 2021. Of the 1384 patients, 611 (44.1%) were women, 576 (41.6%) Hispanic, 472 (34.1%) non-Hispanic White, 138 (10.0%) Asian, 60 (4.3%) Black, and 122 (8.8%) of unknown/other race/ethnicity. Retransplant patients accounted for 1.8% (25) of the cohort. Distribution of patients by University of California sites was 410 (29.6%) at UC-1, 173 (12.5%) at UC-2, and 801 (57.9%) at UC-3. The patients were randomly divided based on a 60:20:20 ratio with 841 patients in the training set, 255 in the validation set, and 288 in the test set. The 3 sets were broadly similar across multiple characteristics (eg, age, race/ethnicity, liver disease etiologies, comorbid conditions, and distribution between UCH facilities). Of note, the median MELDNa scores at admission were 34 (IQR 29–39), 34 (IQR 30–38), and 34 (IQR 30–38) for the training, validation, and test sets, respectively. Detailed patient characteristics at time of admission are reported in Table 1.

Primary and secondary outcomes

In the total sample of 1384 patients, 149 (10.8%) met the primary outcome of death at 1 year, 97 (7%) met the secondary outcome of death at 90 days, and 621 (44.9%) met the secondary outcome of readmission within 90 days. Distributions and prevalence of the primary and secondary outcomes were similar between the training, validation, and test sets and are reported in Table 2.

RuleFit and Expert Augmentation

After identification and division of the ACLF patient population as above, we then applied the RuleFit algorithm. RuleFit generated 20 rules for the primary outcome of death at 1 year (Table 3), 18 rules for the secondary outcome of death within 90 days (Table 4), and 6 rules for the secondary outcome of readmission within 90 days (Table 5). The rules generated by RuleFit for each of the outcomes were then distributed to 15 hepatologists throughout the world who conduct clinical care and research in ACLF and rated rule importance. The aggregated physician rankings along with rank differences between RuleFit and experts are also reported in Tables 3–5 for each of the 3 outcomes. Of note, the greatest discrepancies between RuleFit and human experts occurred in the rankings of biomarkers more commonly utilized in clinical practice, such as age and MELDNa score.

EAML model performance versus cross-sectional and other ML models

For the primary outcome of death at 1 year, AUROC were 0.707 (CI 0.625–0.793) for the EAML and 0.719 (CI 0.640–0.800) for the RuleFit models. For the secondary outcome of death at 90 days, AUROC were 0.678 (CI 0.581–0.776) for the EAML and 0.707 (CI 0.615–0.800) for the RuleFit models (Table 6). Pairwise AUROC differences and CIs for the primary outcome of death at 1 year and the secondary outcome of death at 90 days are reported in detail in Table 7 and graphically represented in the Supplementary Material. In general, for the outcomes of death at 1 year and at 90 days, AUROC differences between EAML and RuleFit models showed that RuleFit outperformed EAML but this was not significant: Δ(RuleFit–EAML) was 0.013 (CI −0.027 to 0.052) for death at 1 year and Δ(RuleFit–EAML) was 0.030 (CI −0.100 to 0.071) for death at 90 days. Moreover, AUROC differences between the EAML/RuleFit models and GBM, and those between the EAML/RuleFit models and GLMNET were also not significant. In contrast, for the outcomes of death at 1 year and death at 90 days, the EAML/RuleFit models consistently outperformed cross-sectional models (MELDNa, NACSELD, CLIF-ACLF, and TAM).

For the secondary outcome of readmission at 90 days, AUROC were 0.557 (CI 0.493–0.623) for the EAML and 0.564 (CI 0.498–0.629) for the RuleFit models (Table 6). Pairwise AUROC differences and CIs for the secondary outcome of readmission at 90 days are reported in detail in Table 7 and graphically represented in the Supplementary Material. In general, the EAML and RuleFit models did not show significant differences in predictive abilities versus each other and versus other ML models. Moreover, while EAML/RuleFit showed significant differences in AUROC versus some of the cross-sectional models (MELDNa, NACSELD, and CLIF-ACLF), the overall predictive abilities of all models evaluated were poor.

Discussion

This study is one of the first to explicitly combine human expert knowledge with ML to create an interpretable ML model for a clinical problem within transplantation. In this study, we generated 2 models (EAML, which incorporates human expert content, and RuleFit, which does not incorporate human input) for each of the 3 outcomes (posttransplant mortality at 1 year, posttransplant mortality at 90 days, and readmission at 90 days). Our ML models (EAML and RuleFit) significantly outperformed existing cross-sectional models with mean AUROC clustering around 0.700 for the outcomes of posttransplant mortality at 1 year and mortality at 90 days. In contrast, our ML models did not show good predictive ability for readmission at 90 days—this finding is largely consistent with previous literature showing difficulties with predicting this outcome.[18] This implies that operative, donor, and post-LT variables may be more important for modeling this outcome as opposed to pre-LT variables.

In our pairwise comparisons of models utilizing AUROC differences, we found that there were no significant differences between EAML and RuleFit, and between EAML/RuleFit and other popular ML algorithms, such as GBM and GLMNET. Moreover, while these were not statistically significant, the EAML models consistently had lower AUROC versus the RuleFit models. The most likely explanation in this situation is due to similarities in the training, validation, and test sets, eg, being all derived from the same database. In this circumstance, the process of incorporating expert input with EAML did not improve the performance of the model since the test sets have similar distributions of demographic and clinical characteristics as the training sets.

The purpose of EAML, therefore, in this situation is to reveal key insights from the discrepancies between human experts and ML rankings of rules. These reveal residual biases, artifacts, and areas for future research. For instance, in the EAML model for posttransplant mortality at 1 year, rule #18 (MELDNa at the time of transplant being >32.47) was ranked as the most important by experts, but only tenth most important by RuleFit. This difference in rank by 9 positions indicated that experts may have biases favoring a well-known and established clinical scoring system. In general, across the 3 outcomes, ACLF experts were more likely to overrank the importance of commonly used physiologic and clinical makers, such as MELDNa, age, and white blood cell count. In contrast, RuleFit was more likely to elevate the importance of electrolytes and hematological parameters, such as ionized calcium, sodium, and lactate dehydrogenase as important data features, in comparison to experts.

While the differences in feature importance may be due to human biases, a second reason for this may be due to data artifacts. In our case, all 3 transplant centers are located in California, which is a high MELDNa area of the United States. We invited human experts from around the world, including Europe and Asia, with differing etiologies for ACLF. The human expert rankings may be “correct” for their respective patient populations but may be deemed “incorrect” or discordant compared with ML rankings for our population. This is likely one of the reasons why the TAM model based on French ACLF patients performed poorly in our populations. Unfortunately, due to survey restrictions, it was neither possible nor ethical to identify individual respondents. Finally, one last reason why there are disparities between human experts and the RuleFit model is that the model coefficients may not be a true reflection of feature importance.

These results imply additional avenues for further research in the clinical care of patientswith ACLF (Fig. 3). Moreover, this study demonstrates that EAML’s use may not be limited to predictive modeling, but also as an artificial intelligence-guided method for hypothesis generation. Interestingly, our data indicated a 1-year posttransplant mortality rate of 10.8%, which is higher than the more contemporary estimates (6.4%) of 1-year posttransplant mortality rates for the general posttransplant population but is still largely in line with population-level analyses of the UNOS database.[10,47,48] This 1-year mortality rate is noted to be lower than certain single-center studies in other geographies.[8,9]

Finally, this was the third study to fully utilize UCHDW, a novel big data multicenter EHR database, and the first to derive insights on transplant patients. UCHDW is based on the OMOP common data model, which is also utilized in several other big data multicenter EHR databases, such as the National COVID Cohort Collaborative (N3C),[49] All of Us,[50] and the Veterans Health Administration Corporate Data Warehouse (VHACDW).[51,52] While patients with ACLF and LT patients have been extensively studied in the VHACDW, the VHACDW is not broadly representative of the general population. While patients with cirrhosis have been studied in N3C, the current purviews of N3C limit research topics to those related to the coronavirus pandemic. It is our hope that our analytical approach of utilizing OMOP will become more common as increasing numbers of institutions have or are in the process of harmonizing their EHR data to the OMOP common data model.

Our approach, which relied upon the wholesale automated extraction of EHR data and case-finding, is therefore fundamentally different than the one utilized in the recently published SALT-M score.[15] In their derivation of the SALT-M score, the Multiorgan Dysfunction and Evaluation for Liver Transplant (MODEL) consortium utilized data from manual retrospective review and identified ACLF patients based only on the EF-CLIF diagnostic criteria. In their model development, the MODEL consortium selected features a priori whereas we began model development with all structured data features extractable in the UCHDW database. Despite these very different methodological approaches and final features selected, the SALT-M and EAML/RuleFit models approached AUROC of ≈0.7 in the primary outcome of post-LT death at 1 year.[15] This implies that there is likely a theoretical limit to which pre-LT variables could predict post-LT outcomes and that donor-derived and intraoperative data features are necessary to improve model performance.

There are several limitations to this study due to its retrospective nature, its use of a novel database, and its analytical processes. First, there is selection bias—we had only included patients with ACLF who had undergone LT and not those who were listed but who then subsequently died or recovered and not those who were never listed. This means that the patients with ACLF suffered from a survivorship bias and are unlikely to be representative of the entire population. While it is feasible to pull data for all patients with ACLF who did not undergo LT, we have no visibility into whether these patients were listed for LT and we would not be able to evaluate for the post-LT outcomes. In addition, approximately 2% of the cohort were retransplant patients—these patients are likely distinct and may have a different biological pathway to ACLF versus other patients.

Second, we do not have intraoperative or donor-derived data for the patients in our cohort. While the inclusion of donor variables may have improved predictive ability, the inclusion of donor-derived variables was not consistent with our intention, which was to utilize only pre-LT candidate variables to predict post-LT outcomes. The ultimate clinical decision with which this model would help is whether to proceed to LT without the benefits of knowing donor characteristics. In addition, UCHDW, our data source, did not contain donor-derived variables. Third, EAML and RuleFit ensemble algorithms are ultimately built upon decisiontree algorithms, which have several limitations. These include overfitting due to overly deep trees, instability due to sensitivity to changes in the training data, difficulty with handling continuous variables due to information loss from dichotomization, and the preference for local over global optimality.[53,54]

Fourth, there are also several limitations related to UCHDW. We only sourced data from 3 transplant centers within UCH; this means that we did not have any visibility or access to the clinical data of these patients if they were admitted at other institutions prior to their admission at a UCH facility. UCHDW, being deidentified, does not have provisions for the reidentification of records at this time; we could not conduct a manual chart review to validate ACLF identification. In addition, all 3 UCH facilities included are in the state of California, which has some of the highest MELDNa scores at the time of transplant. External validation should be undertaken for these models prior to their potential deployment in clinical practice.

Finally, the analysis codes utilized to derive the data from UCHDW were written for this specific (UCHDW) implementation of the OMOP common data model. While OMOP is a common data model that allows for generalization of analyses across different datasets, there may be minor variations and differences in data structures, semantics, and coding. The OMOP-based extraction methods and algorithms for these analyses have not been tested on other OMOP-based data sources—further research is required to evaluate for true "out-of-the-box" interoperability.

Despite these limitations, this study represents "proof of concept" for several key conceptual developments for health services research in transplantation: (1) use of human expert augmentation in ML modeling, (2) generation of multiple ML models that outperform traditional cross-sectional models for predicting posttransplant outcomes in ACLF, and (3) utilizing of a novel data source and common data model in transplant hepatology. With external validation and refinement, the EAML models generated in this study could be refined and evaluated in an iterative manner in clinical decision support systems to actively guide clinical decision-making. In such a clinical decision support-based implementation, prospective surveillance of outcomes would then allow for active feedback to further improve these models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The authors of this study were supported in part by the KL2TR001870 (National Center for Advancing Translational Sciences, Ge and McCulloch), AASLD Anna S. Lok Advanced/Transplant Hepatology Award AHL21-104606 (AASLD Foundation, Ge), American Society of Transplantation LICOP Grant Award CA-0182782 (American Society of Transplantation, Ge), P30DK026743 (UCSF Liver Center Grant, Ge and Lai), F31HL156498 (National Heart, Lung, and Blood Institute, Digitale), UL1TR001872 (National Center for Advancing Translational Sciences, Pletcher), and R01AG059183/K24AG080021 (National Institute on Aging, Lai). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or any other funding agencies. The funding agencies played no role in the analysis of the data or the preparation of this manuscript.

Data availability

The authors thank the Center for Data-Driven Insights and Innovation (CDI2) at the University of California Health (UCH),[19] for its analytical and technical support related to the use of the UC Health Data Warehouse (UCHDW) and related data assets, including the UC COVID Research Data Set (CORDS).

Abbreviations:

ACLF	acute-on-chronic liver failure
APASL ACLF	Asian Pacific Association for the Study of the Liver ACLF Research Consortium
AUPRC	area under the precision-recall curve

AUROC	area under the receiver-operating characteristic curve
CDI2	Center for Data-Driven Insights and Innovation
CI	confidence interval
CORDS	UC COVID Research Data Set
EAML	Expert-Augmented Machine Learning
EF-CLIF	European Association for the Study of the Liver-Chronic Liver Failure Consortium
EHR	electronic health record
ESLD	end-stage liver disease
FiO2	fraction of inspired oxygen
GBM	Gradient Boosting Machine
GLMNET	Elastic-Net Regularized Generalized Linear Model
ICD-10-CM	International Classification of Diseases, Tenth Revision, Clinical Modification
LASSO	least absolute shrinkage and selection operator
LT	liver transplantation
MELDNa	Model for End-Stage Liver Disease
ML	machine learning
MODEL	Multiorgan Dysfunction and Evaluation for Liver transplantation
N3C	National COVID Cohort Collaborative
NACSELD	North American Consortium for the Study of End-Stage Liver Disease
OMOP	Observational Medical Outcomes Partnership
PaO2	arterial partial pressure of oxygen
RF	Random Forest
SALT-M	Sundaram ACLF-LT-Mortality
SpO2	partial oxygen saturation
TAM	transplantation for ACLF-3 model
UCH	University of California Health

UCHDW	University of California Health Data Warehouse
UNOS	United Network for Organ Sharing
VHACDW	Veterans Health Administration Corporate Data Warehouse

References

1. Bajaj JS, O'Leary JG, Reddy KR, et al. Survival in infection-related acute-on-chronic liver failure is defined by extrahepatic organ failures. Hepatology. 2014;60(1):250–256. 10.1002/hep.27077. [PubMed: 24677131]
2. O'Leary JG, Reddy KR, Garcia-Tsao G, et al. NACSELD acute-on-chronic liver failure (NACSELD-ACLF) score predicts 30-day survival in hospitalized patients with cirrhosis. Hepatology. 2018;67(6):2367–2374. 10.1002/hep.29773. [PubMed: 29315693]
3. Jalan R, Saliba F, Pavesi M, et al. Development and validation of a-prognostic score to predict mortality in patients with acute-on-chronic liver failure. J Hepatol. 2014;61(5):1038–1047. 10.1016/j.jhep.2014.06.012.
4. Gustot T, Moreau R. Acute-on-chronic liver failure vs. traditional acute-decompensation of cirrhosis. J Hepatol. 2018;69(6):1384–1393. 10.1016/j.jhep.2018.08.024. [PubMed: 30195459]
5. Hernaez R, Solà E, Moreau R, Ginès P. Acute-on-chronic liver failure: an update. Gut. 2017;66(3):541–553. 10.1136/gutjnl-2016-312670. [PubMed: 28053053]
6. Moreau R, Jalan R, Gines P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology. 1437.e1;144(7):1426–1437. doi:10.1053/j.gastro.2013.02.042.
7. Sarin SK, Choudhury A. Acute-on-chronic liver failure: terminology, mechanisms and management. Nat Rev Gastroenterol Hepatol. 2016;13(3):131–149. 10.1038/nrgastro.2015.219. [PubMed: 26837712]
8. Levesque E, Winter A, Noorah Z, et al. Impact of acute-on-chronic liver failure on 90-day mortality following a first liver transplantation. Liver Int. 2017;37(5):684–693. 10.1111/liv.13355. [PubMed: 28052486]
9. Umgelter A, Lange K, Kornberg A, Büchler P, Friess H, Schmid RM. Orthotopic liver transplantation in critically ill cirrhotic patients with multiorgan failure: a single-center experience. Transplant Proc. 2011;43(10):3762–3768. 10.1016/j.transproceed.2011.08.110. [PubMed: 22172843]
10. Sundaram V, Jalan R, Wu T, et al. Factors associated with survival of patients with severe acute-on-chronic liver failure before and after liver transplantation. Gastroenterology. 2019;156(5):1381–1391.e3. 10.1053/j.gastro.2018.12.007. [PubMed: 30576643]
11. Bajaj JS, Verna EC. What role should ACLF play in liver transplant prioritization? survey of us-based transplant providers. Liver Transpl. August 2020;26(12):1658–1661. 10.1002/lt.25861.
12. Wu T, Sundaram V. Transplantation for acute-on-chronic liver failure. Clin Liver Dis (Hoboken). 2019;14(4):152–155. 10.1002/cld.852. [PubMed: 31709045]
13. Sarin SK, Kedarisetty CK, Abbas Z, et al. Acute-on-chronic liver failure: consensus recommendations of the Asian Pacific Association for the Study of the Liver (APASL) 2014. Hepatol Int. 2014;8(4):453–471. 10.1007/s12072-014-9580-2.
14. Artzner T, Michard B, Weiss E, et al. Liver transplantation for critically ill cirrhotic patients: stratifying utility based on pretransplant factors. Am J Transplant. 2020;20(9):2437–2448. 10.1111/ajt.15852.
15. Hernaez R, Karvellas CJ, Liu Y, et al. The novel SALT-M score predicts1-year post-transplant mortality in patients with severe acute-on-chronic liver failure. J Hepatol. June 2023;79(3):717–727. 10.1016/j.jhep.2023.05.028. [PubMed: 37315809]
16. Ge J, Najafi N, Zhao W, Somsouk M, Fang M, Lai JC. A methodology to generate longitudinally updated acute-on-chronic liver failure prognostication scores from electronic health record data. Hepatol Commun. 2021;5(6):1069–1080. 10.1002/hep4.1690. [PubMed: 34141990]

17. Danziger J, Zimolzak AJ. Residual confounding lurking in big data: A source of error. In: Secondary Analysis of Electronic Health Records. Springer International Publishing; 2016:71–78. 10.1007/978-3-319-43742-2_8.

18. Hu C, Anjur V, Saboo K, et al. Low predictability of readmissions and death using machine learning in cirrhosis. Am J Gastroenterol. 2021;116(2):336–346. 10.14309/ajg.0000000000000971. [PubMed: 33038139]

19. Center for Data-Driven Insights and Innovations. UCOP. Accessed February 17, 2021. https://www.ucop.edu/uc-health/functions/center-for-data-driven-insights-and-innovations-cdi2.html:(CDI2).

20. Observational Health Data Sciences and Informatics. Standardized Data: The OMOP Common Data Model. Accessed February 17, 2021. https://www.ohdsi.org/data-standardization/the-common-data-model/.

21. Peterson TA, Fontil V, Koliwad SK, Patel A, Butte AJ. Quantifying variation in treatment utilization for type 2 diabetes across five major university of california health systems. Diabetes Care. 2021;44(4):908–914. 10.2337/dc20-0344. [PubMed: 33531419]

22. Pintus R, Yang Y, Rushmeier H. ATHENA. J Comput Cult Herit. 2015;8(1):1–25. 10.1145/2659020.

23. Singh H, Pai CG. Defining acute-on-chronic liver failure: east, West or Middle ground? World J Hepatol. 2015;7(25):2571–2577. 10.4254/wjh.v7.i25.2571. [PubMed: 26557949]

24. Zaccherini G, Weiss E, Moreau R. Acute-on-chronic liver failure: definitions, pathophysiology and principles of treatment. JHEP Rep. 2021;3(1):100176. 10.1016/j.jhepr.2020.100176. [PubMed: 33205036]

25. N3C Consortium Ge J, Pletcher MJ, Lai JC. Outcomes of SARS-CoV-2 infection in patients with chronic liver disease and cirrhosis: A national COVID cohort collaborative study. Gastroenterology. 2021;161(5):1487–1501. 10.1053/j.gastro.2021.07.010.e5. [PubMed: 34284037]

26. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48(6):S106–S113. 10.1097/MLR.0b013e3181de9e17.suppl. [PubMed: 20473190]

27. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc. 2015;22(4):872–880. 10.1093/jamia/ocv024. [PubMed: 25896647]

28. Singh A, Nadkarni G, Gottesman O, Ellis SB, Bottinger EP, Guttag JV. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. J Biomed Inform. 2015;53:220–228. 10.1016/j.jbi.2014.11.005.

29. Wong KC-Y, Xiang Y, Yin L, So HC. Uncovering clinical risk factors and predicting severe COVID-19 cases using UK Biobank data: machine learning approach. JMIR Public Health Surveill. 2021;7(9):e29544. 10.2196/29544. [PubMed: 34591027]

30. Rios R, Miller RJH, Manral N, et al. Handling missing values in machine learning to predict patient-specific risk of adverse cardiac events: insights from REFINE SPECT registry. Comput Biol Med. 2022;145:105449. 10.1016/j.compbiomed.2022.105449. [PubMed: 35381453]

31. Liu D, Oberman HI, Muñoz J, Hoogland J, Debray TPA. Quality Control, Data Cleaning, Imputation. 2021. 10.48550/arxiv.2110.15877.arXiv.

32. Razavian N, Major VJ, Sudarshan M, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. npj Digit Med. 2020;3:130. 10.1038/s41746-020-00343-x. [PubMed: 33083565]

33. Ayala Solares JR, Diletta Raimondi FE, Zhu Y, et al. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. J Biomed Inform. 2020;101:103337. 10.1016/j.jbi.2019.103337. [PubMed: 31916973]

34. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020;145(2):463–469. 10.1016/j.jaci.2019.12.897. [PubMed: 31883846]

35. Gennatas ED, Friedman JH, Ungar LH, et al. Expert-augmented machine learning. Proc Natl Acad Sci U S A. 2020;117(9):4571–4577. 10.1073/pnas.1906831117. [PubMed: 32071251]

36. Gennatas ED. Rtemis ML. Accessed January 2, 2023. https://rtemis.lambdamd.org/.

37. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2(3):916–954. 10.1214/07-AOAS148.

38. Lau L, Kankanige Y, Rubinstein B, et al. Machine-learning algorithms predict graft failure after liver transplantation. Transplantation. 2017; 101(4):e125–e132. 10.1097/TP.0000000000001600. [PubMed: 27941428]

39. Ferrarese A, Sartori G, Orrù G, et al. Machine learning in liver transplantation: a tool for some unsolved questions? Transpl Int. 2021;34(3):398–411. 10.1111/tri.13818. [PubMed: 33428298]

40. Spann A, Yasodhara A, Kang J, et al. Applying machine learning in liver disease and transplantation: A comprehensive review. Hepatology. 2020;71(3):1093–1105. 10.1002/hep.31103. [PubMed: 31907954]

41. Ivanics T, So D, Claasen MPAW, et al. Machine learning–based mortality prediction models using national liver transplantation registries are feasible but have limited utility across countries. Am J Transplant. 2023;23(1):64–71. 10.1016/j.ajt.2022.12.002. [PubMed: 36695623]

42. Branco P, Torgo L, Ribeiro R. A Survey of Predictive Modelling under Imbalanced Distributions. 2015. 10.48550/arxiv.1505.01658.arXiv.

43. DiCiccio TJ, Efron B. Bootstrap confidence intervals. Stat Sci. 1996;11(3):189–228. 10.1214/ss/1032280214.

44. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med. May 2000; 19(9):1141–1164. 10.1002/(sici)1097-0258(20000515)19:9&lt;1141::aid-sim479&gt;3.0.co;2-f. [PubMed: 10797513]

45. Kim WR, Biggins SW, Kremers WK, et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med. 2008;359(10):1018–1026. 10.1056/NEJMoa0801209. [PubMed: 18768945]

46. R Core Team. R. A Language and Environment for Statistical Computing. 2013.

47. Kwong AJ, Ebel NH, Kim WR, et al. OPTN/SRTR 2020 annual data report: liver. Am J Transplant. 2022;22(suppl 2):204–309. 10.1111/ajt.16978.

48. Sundaram V, Kogachi S, Wong RJ, et al. Effect of the clinical course of acute-on-chronic liver failure prior to liver transplantation on post-transplant survival. J Hepatol. 2020;72(3):481–488. 10.1016/j.jhep.2019.10.013. [PubMed: 31669304]

49. Haendel MA, Chute CG, Bennett TD, et al. The national COVID cohort collaborative (N3C): Rationale, design, infrastructure, and deployment. J Am Med Inform Assoc. 2021;28(3):427–443. 10.1093/jamia/ocaa196. [PubMed: 32805036]

50. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All of US Research Program: transforming i2b2 data into the OMOP common data model. PLOS ONE. 2019;14(2):e0212463. 10.1371/journal.pone.0212463. [PubMed: 30779778]

51. Wang H, Belitskaya-Levy I, Wu F, et al. A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program. BMC Med Inform Decis Mak. 2021;21(1):289. 10.1186/s12911-021-01643-2.

52. Viernes B, Lynch KE, South B, Coronado G, DuVall SL. Characterizing VA users with the OMOP common data model. Stud Health Technol Inform. 2019;264:1614–1615. 10.3233/SHTI190561. [PubMed: 31438258]

53. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: an overview and their use in medicine. J Med Syst. 2002;26(5):445–463. 10.1023/a:1016409317640. [PubMed: 12182209]

54. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–674. 10.1109/21.97458.

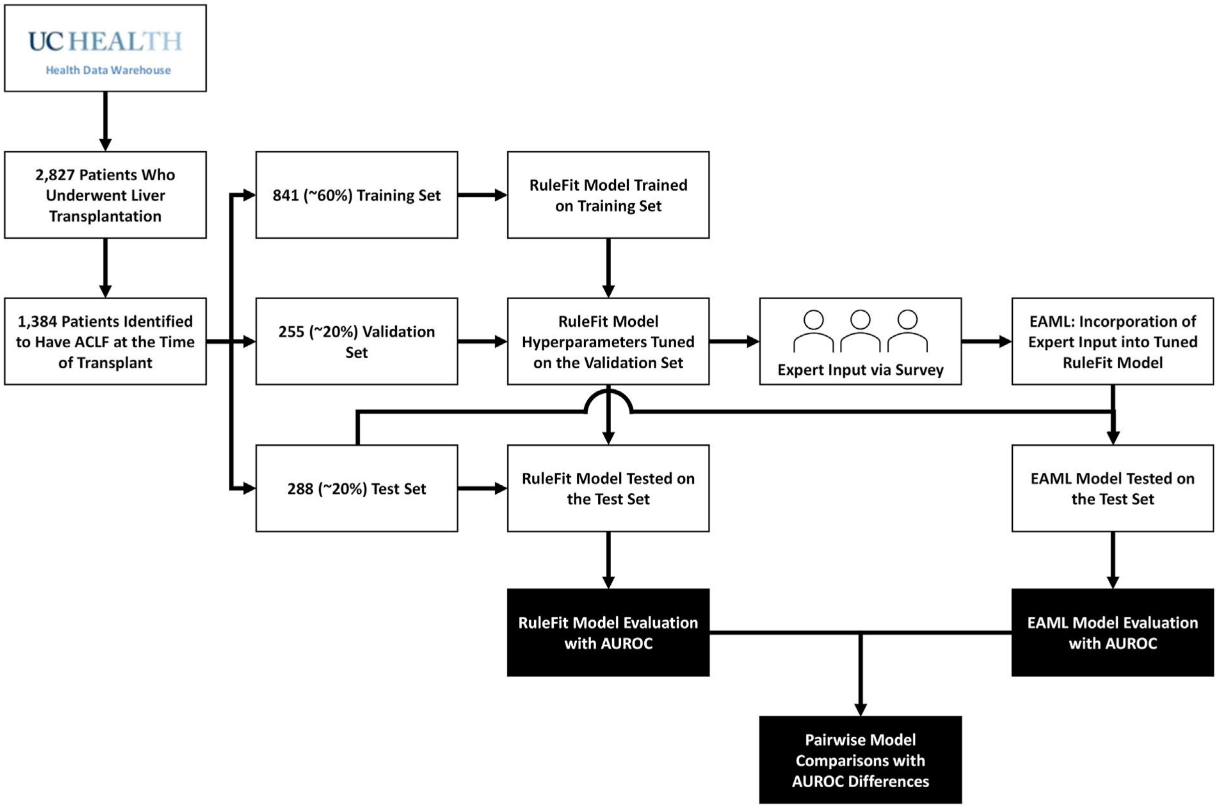
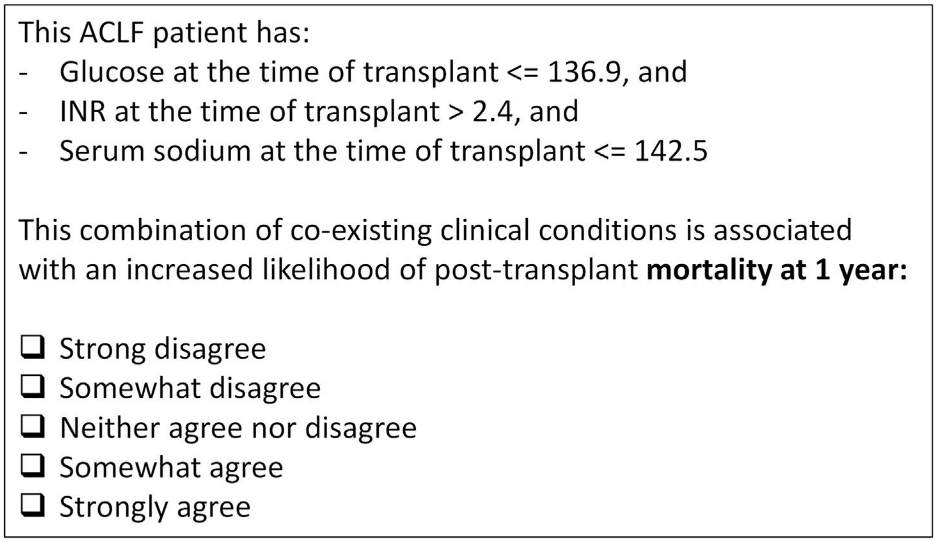

UC HEALTH
Health Data Warehouse
2,827 Patients Who Underwent Liver Transplantation
1,384 Patients Identified to Have ACLF at the Time of Transplant
841 (~60%) Training Set
255 (~20%) Validation Set
288 (~20%) Test Set
RuleFit Model Trained on Training Set
RuleFit Model Hyperparameters Tuned on the Validation Set
Expert Input via Survey
EAML: Incorporation of Expert Input into Tuned RuleFit Model
RuleFit Model Tested on the Test Set
EAML Model Tested on the Test Set
RuleFit Model Evaluation with AUROC
EAML Model Evaluation with AUROC
Pairwise Model Comparisons with AUROC Differences

Figure 1.

Study design flowchart.



This ACLF patient has:

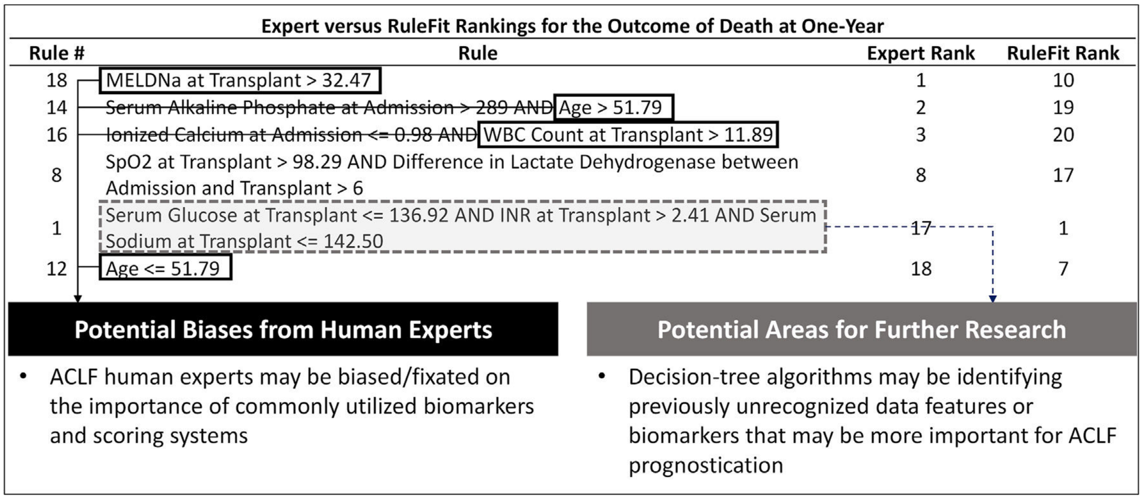
- Glucose at the time of transplant <= 136.9, and
- INR at the time of transplant > 2.4, and
- Serum sodium at the time of transplant <= 142.5

This combination of co-existing clinical conditions is associated with an increased likelihood of post-transplant **mortality at 1 year:**

□ Strong disagree
□ Somewhat disagree
□ Neither agree nor disagree
□ Somewhat agree
□ Strongly agree

Figure 2.

Example survey questions utilized to obtain expert input.



Expert versus RuleFit Rankings for the Outcome of Death at One-Year

Rule #	Rule	Expert Rank	RuleFit Rank
18	MELDNa at Transplant > 32.47	1	10
14	~~Serum Alkaline Phosphate at Admission > 289 AND~~ Age > 51.79	2	19
16	~~Ionized Calcium at Admission <= 0.98 AND~~ WBC Count at Transplant > 11.89	3	20
8	SpO2 at Transplant > 98.29 AND Difference in Lactate Dehydrogenase between Admission and Transplant > 6	8	17
1	Serum Glucose at Transplant <= 136.92 AND INR at Transplant > 2.41 AND Serum Sodium at Transplant <= 142.50	17	1
12	Age <= 51.79	18	7

Potential Biases from Human Experts

- ACLF human experts may be biased/fixated on the importance of commonly utilized biomarkers and scoring systems

Potential Areas for Further Research

- Decision-tree algorithms may be identifying previously unrecognized data features or biomarkers that may be more important for ACLF prognostication

Figure 3.

Disagreements between experts and RuleFit may reflect biases, artifacts, and areas for further research.

Table 1

Baseline clinical and demographic characteristics of the training, validation, and test set populations.

Characteristic	Training (N = 841)	Validation (N = 255)	Test (N = 288)	*P* value
Female	370 (44)	106 (42)	135 (47)	0.46
Age, y (IQR)	57.5 (49.1–63.8)	56.2 (46.2–62.9)	58.0 (47.4–64.5)	0.19
UC health site				0.64
UC-1	251 (30)	70 (27)	89 (31)	
UC-2	110 (13)	34 (13)	29 (10)	
UC-3	480 (57)	151 (59)	170 (59)	
Race/ethnicity				0.21
Hispanic	357 (42)	107 (42)	112 (39)	
White	284 (34)	80 (31)	108 (38)	
Asian	89 (11)	19 (7)	30 (10)	
Black	34 (4)	13 (5)	13 (5)	
Unknown/other	68 (8)	33 (13)	21 (7)	
Etiology of liver disease				
Alcohol associated	293 (35)	90 (35)	94 (33)	0.76
Nonalcoholic fatty	207 (25)	63 (25)	73 (25)	0.97
Hepatitis C	207 (25)	61 (24)	66 (23)	0.84
Hepatitis B	87 (10)	17 (7)	24 (8)	0.17
Autoimmune	67 (8)	24 (9)	23 (8)	0.75
Previous complications of cirrhosis				
Ascites	746 (89)	231 (91)	258 (90)	0.68
Hepatic encephalopathy	636 (76)	191 (75)	214 (74)	0.90
Esophageal varices	468 (56)	139 (55)	170 (59)	0.51
Spontaneous bacterial peritonitis	179 (21)	60 (24)	70 (24)	0.50
Hepatocellular carcinoma	113 (13)	35 (14)	46 (16)	0.56
Comorbidities				
Chronic renal failure	476 (57)	131 (51)	173 (60)	0.12
Diabetes	362 (43)	107 (42)	120 (42)	0.90
Coronary artery disease	231 (27)	60 (24)	84 (29)	0.31

Characteristic	Training (N = 841)	Validation (N = 255)	Test (N = 288)	*P* value
Congestive heart failure	130 (15)	32 (13)	41 (14)	0.50
Laboratory tests				
MELDNa	34.1 (29.0–39.1)	33.8 (30.1–38.1)	33.7 (29.5–37.6)	0.90
Sodium	134.0 (129.0–138.0)	135.0 (130.0–139.0)	134.0 (129.0–138.0)	0.24
Creatinine	2.0 (1.2–3.4)	2.0 (1.2–3.1)	2.0 (1.3–3.2)	0.74
Albumin	3.1 (2.6–3.7)	3.1 (2.6–3.6)	3.2 (2.6–3.6)	0.50
Aspartate transferase	70.0 (43.5–125.0)	37.0 (23.0–72.0)	36.0 (21.0–62.0)	0.30
Alanine transferase	36.0 (21.0–66.5)	37.0 (23.0–72.0)	36.0 (21.0–62.0)	0.76
Alkaline phosphatase	111.0 (78.8–163.0)	117.5 (82.0–177.8)	110.0 (79.0–156.0)	0.17
Total bilirubin	12.0 (4.6–24.5)	11.6 (5.2–19.8)	11.3 (4.3–22.0)	0.51
White blood cell count	7.2 (5.1–11.3)	7.7 (5.0–12.1)	7.5 (5.0–11.3)	0.75
Hemoglobin	8.7 (7.8–10.2)	8.8 (7.9–10.0)	8.8 (7.8–10.1)	0.87
Platelet	53.0 (37.0–88.5)	54.0 (37.0–87.3)	53.0 (35.0–79.0)	0.41
International normalized ratio	2.3 (1.8–2.9)	2.3 (1.9–3.1)	2.3 (1.8–3.0)	0.35
Infection	110 (13)	36 (14)	46 (16)	0.47
Hemodialysis	60 (7)	17 (7)	29 (10)	0.22
NACSELD-ACLF	559 (66)	160 (63)	183 (64)	0.44
CLIF-ACLF				0.34
Grades 1–2	178 (21)	65 (26)	64 (22)	
Grade 3	663 (79)	190 (75)	222 (77)	
TAM (CLIF-ACLF 3 only)				0.69
0–1	113 (17)	31 (16)	39 (18)	
2	120 (18)	27 (14)	28 (13)	
3	37 (6)	10 (5)	12 (5)	

Abbreviations: CLIF-ACLF, Chronic Liver Failure–acute-on-chronic liver failure; MELDNa, Model for End-Stage Liver Disease; NACSELD, North American Consortium for the Study of End-Stage Liver Disease; TAM, transplantation for ACLF-3 model.

Table 2

Outcomes of the training, validation, and test set populations.

Outcomes	**Training (N = 841)**	**Validation (N = 255)**	**Test (N = 288)**	***P* value**
Outcomes				
Death at 1 y	87 (10)	28 (11)	34 (13)	.78
Death at 90 d	55 (7)	19 (7)	23 (9)	.68
Readmission at 90 d	367 (44)	127 (50)	127 (50)	.21

Table 3

RuleFit and expert rankings for the primary outcome of mortality at 1 year.

Rule description	# Cases	Model coefficient	Empirical risk (importance)	RuleFit importance	Expert importance	Rank difference
Serum glucose[b] ≤ 136.92 AND INR[b] > 2.41 AND serum sodium[b] ≤ 142.50	229	0.8	0.97	1	17	−16
Age ≤ 51.79	263	0.45	0.93	7	18	−11
Serum glucose[b] > 168 AND differences in serum glucose ≤ 68.32 AND no new bacterial infection (transplant versus admission)	83	0.41	0.95	6	15	−9
Serum sodium[a] ≤ 138.50 AND oxygen saturation[b] ≤ 98.29	232	0.62	0.96	4	12	−8
Serum glucose[b] > 136.92	410	−0.03	0.86	11	19	−8
Differences in serum albumin ≤ −0.55	84	0.72	0.96	2	9	−7
Serum calcium[b] ≤ 8.45 AND ionized calcium[b] ≤ 1.36 AND differences in temperature ≤ 0.15	81	0.44	0.96	3	7	−4
Total bilirubin[b] ≤ 26.41 AND differences in serum albumin > −0.55 AND differences in serum phosphorus ≤ 0.98	435	0.36	0.92	8	10	−2
Ionized calcium[b] ≤ 1.36 AND differences in temperature > 0.15	388	0.02	0.92	9	11	−2
Differences in serum albumin > −0.55 AND differences in serum phosphorus > 0.98	140	−0.15	0.83	14	16	−2
Serum glucose[b] ≤ 136.92 AND serum sodium[b] > 142.50	37	−0.38	0.76	18	20	−2
Alkaline phosphatase[a] ≤ 289 AND differences in serum potassium > 1 AND age > 51.79 y	25	0.63	0.96	5	5	0
Differences in serum glucose > 68.32	138	−0.16	0.83	15	13	2
Ionized calcium[b] > 1.36	77	−0.62	0.79	16	14	2
Alkaline phosphatase[a] > 63 AND serum bicarbonate[a] ≤ 26.90 AND differences in hemoglobin > −0.10	283	−0.35	0.85	13	6	7
Temperature[a] > 98.05 AND MELDNa[b] ≤ 32.47 AND differences in WBC ≤ 3.90	199	−0.21	0.85	12	4	8
MELDNa[b] > 32.47	436	0	0.92	10	1	9
Oxygen saturation[b] > 98.29 AND differences in lactate dehydrogenase > 6	135	−0.61	0.79	17	8	9
Alkaline phosphatase[a] > 289 AND age > 51.79 y	16	−0.44	0.75	19	2	17
Ionized calcium[a] ≤ 0.98 AND WBC[b] > 11.89	12	−1.02	0.67	20	3	17

Differences are defined between values on the day prior to transplant versus those on the day of admission.

Abbreviations: MELDNa, Model for End-Stage Liver Disease.

[a] Indicates value on the day of admission.

[b] Indicates value on the day prior to transplant.

Table 4

RuleFit and expert rankings for the secondary outcome of mortality at 90 days.

Rule description	# Cases	Model coefficient	Empirical risk (importance)	RuleFit importance	Expert importance	Rank difference
Alkaline phosphatase[b] ≤ 69.50 AND serum ALT[b] ≤ 44 AND age ≤ 69.60 y	110	0.2	0.98	3	18	−16
Serum ALT[b] ≤ 32 AND differences in serum glucose ≤ −0.85	112	0.11	0.98	2	15	−12
Heart rate[b] ≤ 79.30	278	0.17	0.97	4	16	−12
Hemoglobin[b] > 7.35 AND ionized calcium[b] ≤ 1.54 AND differences in lactate dehydrogenase ≤ 24.92	611	0.69	0.96	9	12	−3
Platelet[b] ≤ 37.50 AND temperature[b] ≤ 100.17 AND differences in total bilirubin ≤ 5.73	232	0.72	0.98	1	3	−2
Serum chloride[a] > 103.10	254	0.23	0.96	8	10	−2
Serum chloride[a] ≤ 103.10 AND serum chloride[a] ≤ 101.50 AND ionized calcium[a] > 1.15	210	0.26	0.97	6	7	−1
Temperature[b] > 98.85 AND differences in lactate dehydrogenase ≤ 40.05	192	0.39	0.97	5	5	0
Alkaline phosphatase[a] > 76.50 AND differences in serum glucose > −0.85	465	−0.28	0.92	10	10	0
Serum calcium[b] > 8.60 AND hemoglobin[b] > 8.85 AND differences in MELDNa ≤ 0.20	118	−0.7	0.84	16	16	0
Blood urea nitrogen[b] > 25.50 AND temperature[b] ≤ 98.85 AND differences in lactate dehydrogenase ≤ 40.05	324	0.39	0.97	7	6	1
Serum AST[b] ≤ 34 AND heart rate[b] > 79.30	61	−0.36	0.85	14	13	1
Serum AST[b] > 34 AND heart rate[b] > 79.30 AND differences in serum creatinine ≤ −1.95	94	−0.26	0.84	15	13	2
Serum albumin[b] > 3.15 AND differences in bicarbonate ≤ 3.15 AND differences in oxygen saturation > −1.33	316	−0.3	0.91	11	8	3
Serum Chloride[a] ≤ 103.10 AND Serum Chloride [a] > 101.50	74	−0.03	0.86	13	8	5
Difference in Total Bilirubin > 5.73	218	−0.11	0.89	12	1	11
Differences in Lactate Dehydrogenase > 40.05	86	−0.06	0.79	18	4	14
Temperature[b] > 100.17 AND Differences in Total Bilirubin ≤ 5.73	24	−0.1	0.79	17	2	15

Differences are defined between values on the day prior to transplant versus those on the day of admission.

[a] Indicates value on the day of admission.

[b] Indicates value on the day prior to transplant.

Table 5

RuleFit and expert rankings for the secondary outcome of readmissions at 90 days.

Rule description	# Cases	Model coefficient	Empirical risk (Importance)	RuleFit importance	Expert importance	Rank difference
Differences in phosphorus > 8.50	42	0.21	0.76	1	2	−1
Blood urea nitrogen[a] > 22.50 AND serum glucose[b] ≤ 109.50	101	0.08	0.69	2	3	−1
Blood urea nitrogen[a] > 22.50 AND serum glucose[b] > 109.50 AND respiratory rate[a] ≤ 24.50	416	−0.18	0.5	5	6	−1
Serum phosphorus[a] > 3.25 AND differences in bicarbonate ≤ 2.41 AND differences in mean arterial pressures > −10.45	319	−0.04	0.51	4	4	0
Serum albumin[b] ≤ 4.65 AND serum ALT[a] > 57.50 AND new hemodialysis (transplant versus admission)	117	0.12	0.68	3	1	2
Serum calcium[b] > 9.05 AND serum magnesium[a] ≤ 2.05 AND differences in heart rate ≤ 23.95 AND differences in platelet ≤ −7.50	164	−0.36	0.42	6	4	2

Differences are defined between values on the day prior to transplant versus those on the day of admission

[a] Indicates value on the day of admission

[b] Indicates value on the day prior to transplant

Table 6

AUROC and 95% confidence intervals for EAML/RuleFit.

Outcome	EAML	RuleFit
Death at 1 year	0.707 (0.625 to 0.793)	0.719 (0.640 to 0.800)
Death at 90 days	0.678 (0.581 to 0.776)	0.707 (0.615 to 0.800)
Readmission at 90 days	0.557 (0.493 to 0.623)	0.564 (0.498 to 0.629)

Abbreviations: AUROC, area under the receiver-operating characteristic curve; EAML, Expert-Augmented Machine Learning.

Table 7

AUROC differences and 95% confidence intervals for EAML/RuleFit versus other models.

Differences between models	RuleFit	GBM	GLMNET	RF	NACSELD	MELDNa	CLIF-ACLF	TAM
Δ(Model–EAML) AUROC for death at 1 year	0.013 (CI −0.027 to 0.052)	−0.002 (CI −0.102 to 0.095)	−0.002 (CI −0.116 to 0.108)	−0.248 (CI −0.362 to −0.135)	−0.206 (CI −0.289 to −0.112)	−0.292 (CI −0.349 to −0.212)	−0.365 (CI −0.431 to −0.298)	−0.384 (CI −0.520 to −0.253)
Δ(Model–RuleFit) AUROC for death at 1 year		−0.016 (CI −0.121 to 0.092)	−0.015 (CI −0.125 to 0.096)	−0.262 (CI −0.377 to −0.146)	−0.225 (CI −0.304 to −0.132)	−0.31 (CI −0.366 to −0.230)	−0.379 (CI −0.444 to −0.314)	−0.403 (CI −0.537 to −0.271)
Δ(Model–EAML) AUROC for death at 90 days	0.030 (CI −0.01 to 0.071)	−0.086 (CI −0.241 to 0.071)	−0.060 (CI −0.204 to 0.083)	−0.160 (CI −0.318 to −0.009)	−0.234 (CI −0.333 to −0.110)	−0.307 (CI −0.359 to −0.209)	−0.387 (CI −0.465 to −0.308)	−0.459 (CI −0.604 to −0.315)
Δ(Model–RuleFit) AUROC for death at 90 days		−0.099 (CI −0.253 to 0.052)	−0.091 (CI −0.230 to 0.051)	−0.190 (CI −0.347 to −0.030)	−0.245 (CI −0.343 to −0.122)	−0.319 (CI −0.371 to −0.222)	−0.399 (CI −0.475 to −0.322)	−0.474 (CI −0.621 to −0.325)
Δ(Model–EAML) AUROC for readmission at 90 days	0.006 (CI −0.049 to 0.062)	−0.006 (CI −0.089 to 0.075)	−0.033 (CI −0.132 to 0.056)	0.012 (CI −0.077 to 0.098)	−0.077 (CI −0.131 to −0.034)	−0.073 (CI −0.112 to −0.017)	−0.077 (CI −0.105 to −0.031)	−0.069 (CI −0.109 to 0.027)
Δ(Model–RuleFit) AUROC for readmission at 90 days		−0.018 (CI −0.101 to 0.066)	−0.042 (CI −0.129 to 0.043)	0.007 (CI −0.072 to 0.086)	−0.108 (CI −0.162 to −0.068)	−0.105 (CI −0.143 to −0.050)	−0.108 (CI −0.136 to −0.061)	−0.068 (CI −0.103 to 0.025)

Abbreviations: AUROC, area under the receiver-operating characteristic curve; CLIF-ACLF, Chronic Liver Failure-acute-on-chronic liver failure; EAML, Expert-Augmented Machine Learning; GBM, Gradient Boosting Machine; GLMNET, Elastic-Net Regularized Generalized Linear Model; MELDNa, Model for End-Stage Liver Disease; NACSELD, North American Consortium for the Study of End-Stage Liver Disease; RF, Random Forest; TAM, transplantation for ACLF-3 model.