Research Article
Liver Transplantation

JOURNAL OF HEPATOLOGY

Multicenter validation of the liver graft assessment following transplantation (L-GrAFT) score for assessment of early allograft dysfunction

Vatche G. Agopian[1,*], Daniela Markovic[2], Goran B. Klintmalm[3], Giovanna Saracino[3], William C. Chapman[4], Neeta Vachharajani[4], Sander S. Florman[5], Parissa Tabrizian[5], Brandy Haydel[5], David Nasralla[6], Peter J. Friend[7], Yuri L. Boteon[8], Rutger Ploeg[7], Michael P. Harlander-Locke[1], Victor Xia[9], Joseph DiNorcia[1], Fady M. Kaldas[1], Hasan Yersiz[1], Douglas G. Farmer[1], Ronald W. Busuttil[1]

[1]Dumont-UCLA Transplant and Liver Cancer Centers, Department of Surgery, David Geffen School of Medicine at UCLA; [2]Department of Biomathematics, David Geffen School of Medicine at UCLA; [3]Annette C. and Harold C. Simmons Transplant Institute, Baylor University Medical Center, Dallas, TX; [4]Section of Transplantation, Department of Surgery, Washington University in St. Louis, St. Louis, MO; [5]Recanati/Miller Transplantation Institute, Mount Sinai Medical Center, New York, NY; [6]Department of Hepatopancreaticobiliary and Liver Transplant Surgery, Royal Free Hospital, London, UK; [7]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK; [8]Queen Elizabeth Hospital Birmingham, Birmingham, UK; [9]Department of Anesthesia, David Geffen School of Medicine at UCLA

Background & Aims: Early allograft dysfunction (EAD) following liver transplantation (LT) negatively impacts graft and patient outcomes. Previously we reported that the liver graft assessment following transplantation (L-GrAFT$_7$) risk score was superior to binary EAD or the model for early allograft function (MEAF) score for estimating 3-month graft failure-free survival in a single-center derivation cohort. Herein, we sought to externally validate L-GrAFT$_7$, and compare its prognostic performance to EAD and MEAF.

Methods: Accuracies of L-GrAFT$_7$, EAD, and MEAF were compared in a 3-center US validation cohort (n = 3,201), and a Consortium for Organ Preservation in Europe (COPE) normothermic machine perfusion (NMP) trial cohort (n = 222); characteristics were compared to assess generalizability.

Results: Compared to the derivation cohort, patients in the validation and NMP trial cohort had lower recipient median MELD scores; were less likely to require pretransplant hospitalization, renal replacement therapy or mechanical ventilation; and had superior 1-year overall (90% and 95% *vs.* 84%) and graft failure-free (88% and 93% *vs.* 81%) survival, with a lower incidence of 3-month graft failure (7.4% and 4.0% *vs.* 11.1%; p <0.001 for all comparisons). Despite significant differences in cohort characteristics, L-GrAFT$_7$ maintained an excellent validation AUROC of 0.78, significantly superior to binary EAD (AUROC 0.68, p = 0.001) and MEAF scores (AUROC 0.72, p <0.001). In *post hoc* analysis of the COPE NMP trial, the highest tertile of L-GrAFT$_7$ was significantly associated with time to liver allograft (hazard ratio [HR] 2.17, p = 0.016), Clavien ≥IIIB (HR 2.60, p = 0.034) and ≥IVa (HR 4.99, p = 0.011) complications; post-LT length of hospitalization (p = 0.002); and renal replacement therapy (odds ratio 3.62, p = 0.016).

Conclusions: We have validated the L-GrAFT$_7$ risk score as a generalizable, highly accurate, individualized risk assessment of 3-month liver allograft failure that is superior to existing scores. L-GrAFT$_7$ may standardize grading of early hepatic allograft function and serve as a clinical endpoint in translational studies (www.lgraft.com).

Lay summary: Early allograft dysfunction negatively affects outcomes following liver transplantation. In independent multicenter US and European cohorts totaling 3,423 patients undergoing liver transplantation, the liver graft assessment following transplantation (L-GrAFT) risk score is validated as a superior measure of early allograft function that accurately discriminates 3-month graft failure-free survival and post-liver transplantation complications.

Introduction

The function of the donor liver allograft is a critical determinant of patient outcomes following liver transplantation (LT),[1–3] and is directly impacted by the degree of ischemia-reperfusion injury that is an unavoidable consequence of the preservation process.[4] The durability of LT as a life-saving treatment for end-stage liver disease[5] underlies the growing indications for LT, including alcoholic hepatitis[6] and the emergence of "transplant oncology",[7] with further expansion limited only by the pool of transplantable organs. This realization has fueled the development of numerous strategies to mitigate ischemia-reperfusion injury and to allow for utilization of extended criteria organs, perhaps none more promising than the emerging machine perfusion technologies.[8,9] Consequently, having an accurate, quantifiable measure of early allograft function is critical to allow

Keywords: Liver transplantation; Early allograft dysfunction; Risk prediction model; Ischemia-reperfusion injury.
Received 21 March 2020; received in revised form 29 August 2020; accepted 15 September 2020; available online 23 September 2020
* Corresponding author. Address: Division of Liver and Pancreas Transplantation, Department of Surgery, David Geffen School of Medicine at UCLA, Ronald Reagan UCLA Medical Center, 757 Westwood Plaza, Suite 8501-B, Los Angeles, CA 90095. Tel.: (310) 267-9610; fax: (310) 267-3590.
E-mail address: vagopian@mednet.ucla.edu (V.G. Agopian).
https://doi.org/10.1016/j.jhep.2020.09.015

for comparative assessments of the benefits of a given intervention.

The concept of early allograft dysfunction (EAD), a term used to identify allografts with poor or marginal function following LT, was first formally defined by Deschenes and colleagues.[10] The most contemporary and widely validated definition in the model for end-stage liver disease (MELD) score era[11] was proposed by Olthoff *et al.*[12]; namely, a serum bilirubin ≥10 mg/dl or international normalized ratio (INR) ≥1.6 on postoperative day 7, or aspartate aminotransferase (AST) or alanine aminotransferase (ALT) >2,000 U/L within first 7 postoperative days. However, the binary nature of this EAD definition lacks granularity, and fails to capture the continuum along which graft dysfunction occurs. The model for early allograft function (MEAF) score[13] was proposed as a tool to enable continuous grading of allograft function, but its accuracy in determining graft failure was not reported.

We recently reported on the liver graft assessment following transplantation (L-GrAFT) risk score,[14] a quantitative, continuous prognostic score measuring early allograft function based on 7 ($L\text{-}GrAFT_7$) or 10 ($L\text{-}GrAFT_{10}$) days of measures of post-LT AST, bilirubin, INR, and platelets. In the original derivation cohort of 2,008 patients undergoing LT, both $L\text{-}GrAFT_7$ (area under the receiver operator characteristics [AUROC] curve 0.83) and $L\text{-}GrAFT_{10}$ (AUROC 0.85) had significantly greater accuracy in predicting 3-month graft failure-free survival than either the binary EAD definition (EAD, AUROC 0.68) or the MEAF score (AUROC 0.70).

In the current study, we sought to externally validate the L-GrAFT score in a large, multicenter US validation cohort (US VC) of 3,201 patients from 3 large US transplant centers from 3 different United Network for Organ Sharing (UNOS) regions, as well as an international European multicenter cohort comprised of 222 patients that participated in the recently reported first randomized controlled trial of normothermic machine perfusion (NMP) in human liver transplantation from the Consortium for Organ Preservation in Europe (COPE) study group.[15] Our primary aim was to evaluate the accuracy of the L-GrAFT score in these validation cohorts, compare its prognostic performance to EAD and MEAF, and evaluate the generalizability of the model by comparing clinical characteristics in the derivation and validation cohorts. In a secondary analysis, we compared the L-GrAFT score in the 2 COPE trial arms and evaluated the ability of L-GrAFT to predict clinically meaningful outcomes and adverse post-LT events in this prospectively maintained trial database.

Patients and methods

The L-GrAFT risk score was developed using a derivation cohort of 2,008 patients undergoing LT at the University of California, Los Angeles utilizing 7 ($L\text{-}GrAFT_7$) or 10 ($L\text{-}GrAFT_{10}$) days of post-LT laboratory measures.[14] We sought to validate L-GrAFT in an external, independent US multicenter cohort and international European cohort. The US VC consisted of consecutive adult patients undergoing primary LT from the Baylor Simmons Transplant Institute (BSTI) in United Network for Organ Sharing (UNOS) region 4, comprised of Baylor University Medical Center, Dallas (BUMC, n = 1,437 2/1/2002–06/30/2015) and Baylor Scott & White All Saints Medical Center, Fort Worth (BAS, n = 262, 1/7/2006–06/30/2015); Mount Sinai Medical Center (MS) in UNOS region 9 (n = 437, 01/01/2012–06/1/2016); and Washington University (WU) in Saint Louis in UNOS region 8 (n = 1,065, 2002–2015). The international European cohort consisted of a *post hoc* analysis of the 222 patients from the NMP randomized controlled-trial (NMP, n = 121 *vs.* static cold storage [SCS], n = 101) that was conducted and published by COPE.[15] Each participating site received Institutional Review Board approval, and analysis of the COPE dataset was approved by the Trial Management Committee and COPE Management Board.

The primary objective was to evaluate the discrimination (accuracy) and calibration of the L-GrAFT model in predicting 3-month graft failure (liver retransplantation or patient death), and compare its performance to the existing binary EAD definition[12] and MEAF score[13] in both the multicenter US VC and the international COPE cohort. The risk score models and formulas utilized to calculate them are summarized in Table 1. Variables for analysis included recipient demographics (age, gender, primary end-stage liver disease diagnosis, BMI, diabetes, hypertension); pretransplant laboratory variables (platelets, bilirubin, INR, creatinine, physiologic MELD); pretransplant acuity (requirement for hospitalization, renal replacement therapy [RRT], mechanical ventilation, and vasopressors); donor and operative characteristics (donor age and sex, type of allograft [donation after brain death – DBD, donation after cardiac death – DCD, living donor – LD], donor risk index [DRI],[16] cold ischemia time [CIT], implantation warm ischemia time [WIT], and transfusion of packed red blood cells [uPRBCs]); and post-transplant AST, ALT, bilirubin, INR, and platelets for 7 (COPE) or 10 (US Validation) days following LT. A summary of the missing variables for each cohort is summarized in Table S1.

In a *post hoc* analysis of the COPE dataset, our secondary objective was to compare the L-GrAFT score between the 2 treatment arms (NMP *vs.* SCS), and evaluate the association of the L-GrAFT score with the development of clinically relevant post-transplant outcomes and complications. These included the development of post-LT renal failure requiring RRT, post-LT length of hospitalization, and time to development of adverse events including infectious, bleeding, genitourinary, hepatic (liver allograft), and Clavien grade ≥IIIB or Clavien grade ≥IVA complications, which were recorded prospectively and reported in Extended Data Tables 5 and 6 in the COPE NMP Trial.[15] Of note, "hepatic" or "liver allograft" complications as defined in the original COPE NMP Trial (Extended Data Table 6) include biliary leaks or strictures, ischemic cholangiopathy, drainage of ascites, vascular complications of the hepatic artery, hepatic vein, or portal vein, graft dysfunction, and rejection.

Statistical analysis

Clinical characteristics were compared between the derivation cohort and validation cohort utilizing Chi-Square/Fisher's test for categorical variables and Wilcoxon rank sum test for continuous variables, which were reported as median (IQR). Graft failure-free and overall patient survival curves were computed using Kaplan-Meier methods and compared using log-rank tests. The L-GrAFT scores were computed and their distribution compared between the derivation cohort and each validation cohort using the Wilcoxon rank sum test. Discrimination (ability of the model to accurately predict 3-month graft failure-free survival) was assessed by computing the AUROC and its 95% CIs, whereas calibration (agreement between observed and predicted probabilities) was assessed using methods of Hosmer and Lemeshow. We compared the discriminatory ability of the L-GrAFT with established measures for evaluating early allograft function

Table 1. Summary of risk score models evaluated in the US and COPE validation cohorts.

Multivariate predictors in the L-GrAFT risk score in original derivation cohort			
	Odds ratio*	**95% CI**	***p* value**
L-GrAFT$_7$**			
AUC$_7$ log$_e$¶ AST	1.44	1.19–1.76	<0.001
Slope$_7$ log$_e$ AST (rate of change/day) decrease	0.46	0.33–0.65	<0.001
Log$_e$ AUC$_7$ INR	1.20	1.00–1.42	0.045
AUC$_7$ log$_e$ bilirubin	1.82	1.50–2.21	<0.001
Slope$_7$ log$_e$ bilirubin (rate of change/day) decrease	0.64	0.55–0.75	<0.001
AUC$_7$ log$_e$ platelets	0.95	0.89–1.01	0.075
L-GrAFT$_{10}$***			
AUC$_{10}$ log$_e$ AST	1.30	1.05–1.60	0.015
Slope$_7$ log$_e$ AST (rate of change/day) decrease	0.52	0.36–0.75	0.001
Log$_e$ maximum INR$_{10}$	1.22	1.02–1.45	0.028
AUC$_{10}$ log$_e$ bilirubin	1.90	1.54–2.34	<0.001
Slope$_{10}$ log$_e$ bilirubin (rate of change/day) decrease	0.62	0.51–0.75	<0.001
AUC$_{10}$ log$_e$ platelets	0.80	0.65–0.99	0.040
Slope$_{10}$ log$_e$ platelets (rate of change/day) increase	0.73	0.60–0.88	0.001

ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUC7, area under curve over 7 post-LT days; AUC10, area under curve over 10 post-LT days; EAD, early allograft dysfunction; Log maximum INR10, maximum international normalized ratio observed in 10 post-LT days; L-GrAFT$_{7/10}$, liver graft assessment 7/10 days following transplantation; MEAF, model for early allograft function; PLT, platelets; Slope, indicates rate of change over 7 or 10 days in each respective formula; Slope7 indicates rate of change over 7 days in L-GrAFT10; TBIL, total bilirubin.

Risk score formulas.

L-GrAFT$_7$ = 6.965 − 0.58 × (AUC log AST) + 0.008 × (AUC log AST squared) + 5.254 × (slope log AST) + 4.651 × (slope log AST squared) + 1.141 × (log AUC INR) −0.035 × (AUC log TBIL) + 0.006 × (AUC log TBIL squared) + 4.311 × (slope log TBIL) + 5.847 × (slope log TBIL squared) −0.051 × (AUC log PLT).

L-GrAFT$_{10}$ = 9.77 − 0.429 × (AUC log$_e$ AST) + 0.005 × (AUC log$_e$ AST squared) + 4.607 × (Slope7 log$_e$ AST) + 4.413 × (Slope7 log$_e$ AST squared) + 0.890 × (log$_e$ max INR) − 0.049 × (AUC log$_e$ TBIL) + 0.004 × (AUC log$_e$ TBIL squared) + 5.336 × (slope log$_e$ TBIL) − 0.046 × (AUC log$_e$ PLT) − 5.249 × (slope log$_e$ PLT) + 13.086 × (slope log$_e$ PLT squared).

EAD = One or more of the following 1) AST/ALT >2,000 IU/ml in 1st 7 post-LT days, 2) Serum bilirubin ≥10 mg/dl on postoperative day 7, and 3) INR ≥1.6 on postoperative day 7.

MEAF = (score ALTmax.3pod + score INRmax.3pod + score bilirubin3pod); score ALTmax.3pod = 3.29/(1 + e$^{-1.9132}$(log$_e$ (ALTmax.3pod) − 6.1723); score INRmax.3pod = 3.29/(1 + e$^{-6.8204}$(log$_e$ (INRmax.3pod) − 0.6658); score bilirubin3pod = 3.4 / (1 + e$^{-1.8005}$(log$_e$ (bilirubin3pod) − 1.0607).

*Odds ratio (OR) per standard deviation change relative to the median.

**C-statistic of L-GrAFT7 in original derivation cohort = 0.83.

***C-statistic of L-GrAFT$_{10}$ in original derivation cohort = 0.85.

¶Log$_e$ refers to natural log (LN) for all risk score formulas.

including the binary EAD definition and MEAF score for all patients utilizing DeLong formulas.[17] For patients surviving more than 7 days, missing values for any given laboratory variable were multiply imputed using the subject specific trajectory (intercept and slope) under the mixed effects linear regression model over 200 bootstrapped samples of the data, utilizing the boostrap with re-imputation method by Shao and Sitter[18] in order to account for the uncertainty of the estimates due to imputing missing values. Briefly, the procedure is to re-impute repeatedly using the observed data in the bootstrapped sample to fit the imputation model, and then impute the missing data of the bootstrap sample. Instead of only utilizing complete cases without missing data, which reduces the sample size and may lead to inefficient and possibly biased estimates, multiple imputation methods make use of all the data and are the preferred analytic strategy for missing data.[19] The AUROC and the differences in AUROCs were estimated across the bootstrapped imputed datasets and averaged. The 95% CIs for the point estimates were computed using the percentile method. Sensitivity analyses were performed using complete cases where not a single data point was missing, and the results compared to the imputed analyses. Further sensitivity analyses were performed in subsets of patients stratified by MELD score, hospitalization status, DRI, and individual validation centers representing different UNOS regions using the same methods as above. For patients with an early graft failure before 7 days, L-GrAFT scores were estimated using partially available data prior to graft failure, with additional sensitivity analyses utilizing a landmark analysis approach excluding these early graft failures.

For the secondary analysis in the COPE dataset, L-GrAFT$_7$ scores were calculated prior to and after adjustment for total preservation time, DRI, MELD, DCD, and WIT, and compared between the randomized groups using the Wilcoxon rank sum test. The adjusted analysis was performed using a multivariable regression model to quantify the effect of the above variables on the L-GrAFT$_7$ in the US VC. The corresponding parameter estimates from this multivariable regression were used to adjust the L-GrAFT$_7$ score in the COPE dataset to the overall mean values for each covariate in order to estimate the impact of NMP on early allograft function had these 5 variables been the same in both randomized arms. The association of L-GrAFT$_7$ with length of hospitalization and need for post-LT RRT were evaluated using logistic regression, while time to adverse events were evaluated using the proportional hazards Cox regression model.

Analyses were performed using SAS 9.4 (Copyright (c) 2002–2012 by SAS Institute Inc., Cary, NC, USA) and R version 3.0.2 (Copyright (C) 2013 The R Foundation for Statistical Computing).

Results

The US validation cohort was comprised of 3,201 adult patients undergoing primary LT at 3 independent transplant centers from different UNOS regions, including Baylor Simmons Transplant Institute (comprised of BUMC, n = 1,437 and BAS, n = 262; UNOS region 4), Mount Sinai Medical Center (n = 437; UNOS region 9), and Washington University in Saint Louis (n = 1,065; UNOS region 8). The international COPE cohort was comprised of 222 adult patients enrolled in a prospective randomized controlled

Table 2. Comparison of recipient, donor, and operative characteristics in derivation and 3 US validation cohort centers.

	Derivation *vs.* US validation cohort			Comparison of 3 US validation centers			
	DC (n = 2,008)	US VC (n = 3,201)	*p* value	BSTI (n = 1,699)	MS (n = 437)	WU (n = 1,065)	*p* value
Recipient characteristics							
Age, years*	56 (49–62)	55 (49–61)	0.004	53 (48–59)	59 (52–64)	56 (50–62)	<0.001
Male, %	64.4	67.5	0.023	67.5	67.5	67.5	0.999
Diagnosis, %			<0.001				<0.001
HCV	38.5	38.7		35.0	45.5	41.7	
Alcohol	13.7	19.8		25.0	12.4	14.7	
NASH	6.8	6.1		3.1	8.0	10.1	
Cryptogenic	13.8	9.5		11.0	2.5	10.1	
Cholestatic	5.6	9.2		8.9	11.0	8.9	
HBV	6.8	4.4		4.2	7.8	3.2	
Autoimmune	2.1	2.5		3.1	2.8	1.4	
Other	12.8	9.8		9.7	10.0	10.0	
BMI*	27 (24–31)	28 (25–32)	<0.001	28 (25–32)	27 (24–31)	28 (25–32)	0.003
Diabetes, %	25.9	24.2	0.166	20.1	32.0	27.4	<0.001
Hypertension, %	29.6	32.0	0.078	29.9	41.7	31.4	<0.001
Pre-LT laboratory and acuity							
Platelets ($\times 10^3$/μl)*	59 (41–87)	77 (52–113)	<0.001	74 (51–108)	78 (50–119)	82 (56–120)	<0.001
Bilirubin (mg/dl)*	6.8 (2–24.7)	3.4 (1.5–9.2)	<0.001	3.3 (1.5–8.1)	3.4 (1.3–13.2)	3.7 (1.6–9.6)	0.248
INR*	1.6 (1.3–2.1)	1.6 (1.3–2.0)	0.010	1.5 (1.2–1.8)	1.6 (1.2–2.5)	1.7 (1.3–2.3)	<0.001
Creatinine (mg/dl)*	1.4 (0.8–3.9)	1.0 (0.8–1.5)	<0.001	1.0 (0.8–1.5)	1.1 (0.8–2.2)	1.0 (0.7–1.6)	<0.001
Lab MELD*	31 (25–38)	18 (13–26)	<0.001	18 (12–23)	19 (12–34)	20 (14–29)	<0.001
Hospitalized, %	46.1	23.3	<0.001	18.8	36.4	25.2	<0.001
Pre-LT RRT, %	31.7	8.8	<0.001	6.4	14.1	10.4	<0.001
Ventilator, %	19.8	3.7	<0.001	3.4	4.6	3.5	0.491
Donor and operative							
Donor age, years*	41 (25–53)	43 (26–56)	0.001	42 (25–55)	42 (28–53)	45 (27–57)	0.019
Donor male, %	62.6	56.9	<0.001	57.7	58.3	55.2	0.351
Graft type, %			<0.001				<0.001
DBD	94.4	93.9		96.8	80.1	94.9	
DCD	4.2	3.2		1.0	6.9	5.1	
Living donor	1.3	3.0		2.2	13.0	0	
DRI	1.4 (1.2–1.7)	1.6 (1.3–1.9)	<0.001	1.5 (1.2–1.8)	n.a.	1.7 (1.5–2.1)	<0.001
uPRBCs*	11 (7–18)	4 (2–8)	<0.001	5 (2–8)	4 (1–10)	4 (1–8)	<0.001
CIT, hours*	6.7 (5.1–8.5)	6.2 (4.5–8.4)	<0.001	7.6 (5.8–9.5)	4.8 (3.7–5.9)	5.4 (3.9–7.1)	<0.001
WIT, minutes*	40 (35–47)	39 (31–50)	<0.001	49 (41–58)	32 (27–37)	32 (26–38)	<0.001
Follow-up, months	36 (13–77)	52 (24–95)	<0.001	60 (24–97)	35 (12–49)	59 (27–101)	<0.001

CIT, cold ischemia time; DBD, donation after brain death; DCD, donation after cardiac death; DRI, donor risk index; INR, international normalized ratio; MELD, model for end-stage liver disease; n.a., not available; NASH, non-alcoholic steatohepatitis; RRT, renal replacement therapy; uPRBCs, units of packed red blood cells; WIT, warm ischemia time.
*Continuous variables expressed as median (IQR).

trial comparing NMP (n = 121) to SCS (n = 101). The L-GrAFT$_{10}$, L-GrAFT$_7$, EAD, and MEAF scores were calculated for each patient in both validation cohorts based on the published risk scores and formulas (Table 1), with the exception of the COPE cohort where L-GrAFT$_{10}$ was not evaluated because only 7 days of post-LT laboratory studies were prospectively recorded. Consequently, L-GrAFT$_7$ is the primary focus of the analyses in both validation cohorts, with evaluation of L-GrAFT$_{10}$ limited to the US validation cohort and presented as supplementary data. A risk score calculator for both L-GrAFT$_7$ and L-GrAFT$_{10}$ is available at www.lgraft.com and is provided as a supplementary file.

US validation cohort *vs.* L-GrAFT model derivation cohort

Recipient, donor and operative characteristics of the UCLA derivation cohort and all 3 US VC centers are shown in Table 2. Compared to the derivation cohort, US VC recipients were significantly younger (55 *vs.* 56 years, $p = 0.004$); more likely to be male (67.5% *vs.* 64.4%, $p = 0.023$), have alcohol-related (19.8% *vs.* 13.7%) or cholestatic (9.2% *vs.* 5.6%) liver disease but less likely to have cryptogenic (9.5% *vs.* 13.8%) or HBV-related cirrhosis (4.4% *vs.* 6.8%, overall diagnosis $p < 0.001$); had higher median BMI (28 *vs.* 27, $p < 0.001$) and platelet counts (77 *vs.* 59 × 10^3/μL, $p < 0.001$), but significantly lower median pre-LT total bilirubin (3.4 *vs.* 6.8 mg/dl, $p < 0.001$), INR (1.6 [IQR 1.3–2.0] *vs.* 1.6 [IQR 1.3–2.1], $p = 0.010$), serum creatinine (1.0 *vs.* 1.4 mg/dl, $p < 0.001$), and laboratory MELD scores (18 *vs.* 31, $p < 0.001$); and were significantly less likely to require pre-LT hospitalization (23.3% *vs.* 46.1%, $p < 0.001$), RRT (8.8% *vs.* 31.7%, $p < 0.001$), and mechanical ventilation (3.7% *vs.* 19.8%, $p < 0.001$). Regarding donor and operative characteristics, US VC had significantly older donors (43 *vs.* 41 years, $p < 0.001$), fewer male donors (56.9% *vs.* 62.6%, $p < 0.001$), more living donors (3.0% *vs.* 1.3%, $p < 0.001$), higher DRI organs (1.6 *vs.* 1.4, $p < 0.001$), fewer operative blood transfusions (4 *vs.* 11 uPRBCs, $p < 0.001$), and shorter median CIT (6.2 *vs.* 6.7 h, $p < 0.001$) and implantation WIT (39 *vs.* 40 min, $p < 0.001$) compared to the derivation cohort. At a median follow-up of 52 months (IQR 24–95), the overall patient and graft failure-free survival rates in the US VC were 90%, 83%, 77% and 88%, 81%, 74% at 1, 3, and 5-years, with a 7% incidence of 3-month graft failure, compared to 84%, 74%, 68% and 81%, 71%, 65%, respectively, and an 11% incidence of 3-month graft failure in the derivation cohort ($p < 0.001$; Fig. 1A,B).

Comparison of the 3 individual US validation cohorts also showed significant differences in age, diagnosis, BMI, frequency of diabetes and hypertension, pretransplant platelets, INR, creatinine, laboratory MELD, need for hospitalization, RRT,

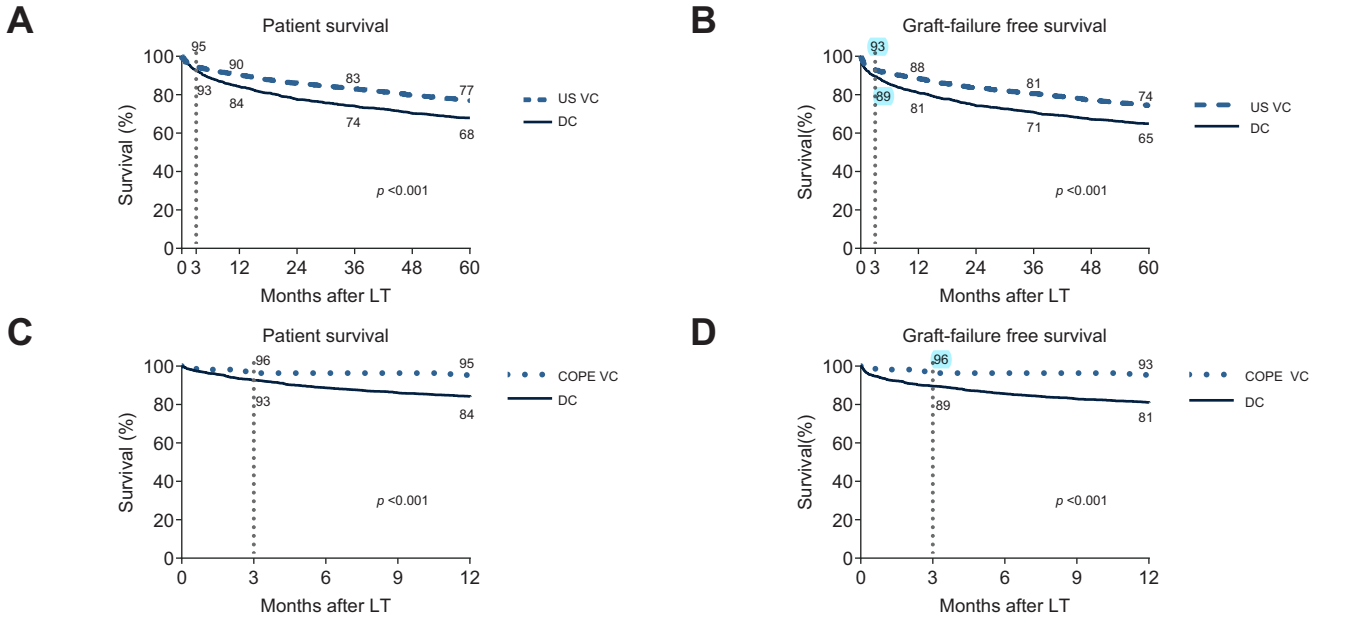


Fig. 1. Kaplan-Meier overall patient and graft failure free survival. (A) Comparison of patient survival in the original DC and US VC. (B) Comparison of graft survival in the original DC and US VC. (C) Comparison of patient survival in the original DC and COPE NMP cohort. (D) Comparison of graft survival in the original DC and COPE NMP cohort. p <0.05 significant (log-rank test). COPE, Consortium for Organ Preservation in Europe; DC, derivation cohort; LT, liver transplantation; NMP, normothermic machine perfusion; VC, validation cohort.

mechanical ventilation, as well as donor age, graft type, DRI, operative transfusions, and cold and implantation WIT (all p < 0.05; Table 2). However, the degree of pre-LT acuity of any individual US VC as assessed by recipient MELD, and need for hospitalization, RRT, or mechanical ventilation was significantly lower compared to the derivation cohort (pairwise p values not reported).

COPE validation cohort *vs.* L-GrAFT model derivation cohort

Recipient, donor and operative characteristics of the derivation cohort and international COPE validation cohort (COPE VC) are compared in Table 3. COPE VC recipients were significantly more likely to be male (72.1% *vs.* 64.4%, p = 0.024), have underlying alcohol-related (29.3% *vs.* 10.1%), non-alcoholic (9.9% *vs.* 5.4%) or cholestatic (19.8% *vs.* 5.2%) liver disease, but less likely HCV-related (3.6% *vs.* 18.8%) or cryptogenic (0% *vs.* 8.1%) cirrhosis, or HCC (16.7% *vs.* 38.1%; overall diagnosis p value <0.001); and had significantly lower median pre-LT bilirubin (2.0 *vs.* 6.8 mg/dl, p <0.001), INR (1.3 *vs.* 1.6, p <0.001), serum creatinine (0.9 *vs.* 1.4 mg/dl, p <0.001), laboratory MELD score (14 *vs.* 31, p <0.001), and need for pretransplant hospitalization (0% *vs.* 46.1%, p <0.001), RRT (1.8% *vs.* 31.7%, p <0.001), and mechanical ventilation (0% *vs.* 19.8%, p <0.001). Regarding donor and operative characteristics, COPE VC patients received allografts from significantly older donors (56 *vs.* 41 years, p <0.001), received more DCD allografts (24.8% *vs.* 4.2%, p <0.001) with higher median DRI (1.6 *vs.* 1.4, p <0.001) and longer preservation times (9.7 *vs.* 6.7 h, p <0.001), but had fewer operative transfusions (2 *vs.* 11 uPRBCs, p <0.001) and shorter implantation WIT (36 *vs.* 40 min, p <0.001). At 12 months follow-up, the overall patient and graft failure-free survivals of the COPE VC were 95% and 93%, respectively, with a 4% incidence of 3-month graft failure, compared to 84% and 81%, respectively, with an 11% incidence of 3-month graft failure in the derivation cohort (p <0.001; Fig. 1C,D).

Distribution of L-GrAFT$_7$ scores and association with 3-month graft failure in derivation and validation cohorts

Despite significant differences in recipient demographics, pre-transplant acuity, donor and operative characteristics between the UCLA derivation cohort and both the US and international COPE validation cohorts, the distribution of L-GrAFT$_7$ scores across intervals of risk score was very similar (Fig. 2A–C, solid dark blue bars), with a median L-GrAFT$_7$ of −2.94 (IQR −3.65 to −2.07), −3.15 (IQR −3.59 to −2.47), and −2.96 (−3.49 to −2.11) in the derivation cohort, US VC, and COPE VC, respectively (p = 0.217). The conditional risk of graft failure with a given risk score was negligible for L-GrAFT$_7$ <−1.5, but continually increased for intervals of risk score ≥−1.5 in derivation cohort, US VC, and COPE VC groups (Fig. 2A–C, solid black bars). L-GrAFT$_7$ scores >−1.5 were observed in only 14.4%, 10.1%, and 14.9% of the entire cohort in the derivation cohort, US VC, and COPE VC (Fig. 2A–C, solid dark blue bars), but accounted for 51.2%, 56.2%, and 41.7% of all graft failures in each respective cohort (Fig. 2A–C, light blue bars).

External validation of L-GrAFT and comparison with existing models of early allograft dysfunction

US validation cohort

External validation of the L-GrAFT$_7$ model in the 3,201 patients comprising the US VC is shown in Table 4. The L-GrAFT$_7$ had an excellent overall C-statistic (AUROC) of 0.78 (IQR 0.75–0.82), significantly superior to both the EAD (0.68 [IQR 0.65–0.71], p <0.001) and MEAF score (0.72 [IQR 0.68–0.76], p <0.001) in predicting 3-month graft failure-free survival (Fig. 3A–C). Furthermore, in extensive sensitivity analyses looking at subsets of

Table 3. Comparison of recipient, donor, and operative characteristics in derivation and COPE validation cohorts.

	Derivation cohort (n = 2,008)	COPE cohort (n = 222)	*p* value
Recipient characteristics			
Age, years*	56 (49–62)	55 (48–63)	0.989
Male, %	64.4	72.1	0.024
Diagnosis¶, %			<0.001
HCV	18.8	3.6	
Alcohol	10.1	29.3	
NASH	5.4	9.9	
Cryptogenic	8.1	0	
Cholestatic	5.2	19.8	
HBV	2.3	2.3	
Autoimmune	1.8	3.2	
Other	10.2	15.2	
HCC	38.1	16.7	
BMI*	27 (24–31)	26 (24–32)	0.619
Pre-LT laboratory and acuity			
Bilirubin (mg/dl)*	6.8 (2–24.7)	2.0 (1.0–4.0)	<0.001
INR*	1.6 (1.3–2.1)	1.3 (1.2–1.5)	<0.001
Creatinine (mg/dl)*	1.4 (0.8–3.9)	0.9 (0.7–1.1)	<0.001
Lab MELD*	31 (25–38)	14 (10–18)	<0.001
Hospitalized, %	46.1	0	<0.001
Pre-LT RRT, %	31.7	1.8	<0.001
Ventilator, %	19.8	0	<0.001
Donor and operative			
Donor age, years*	41 (25–53)	56 (46–65)	<0.001
Donor male, %	62.6	56.8	0.090
Graft type, %			<0.001
DBD	94.4	75.2	
DCD	4.2	24.8	
Living donor	1.3	0	
DRI	1.4 (1.2–1.7)	1.6 (1.4–1.9)**	<0.001
Preservation, %			<0.001
Static cold storage	100	45.5	
Normothermic perfusion	0	54.5	
uPRBCs*	11 (7–18)	2 (0–4)	<0.001
Total preservation time, hours*+	6.7 (5.1–8.5)	9.7 (7.5–12.5)	<0.001
WIT, minutes*	40 (35–47)	36 (28–47)	<0.001

CIT, cold ischemia time; DBD, donation after brain death; DCD, donation after cardiac death; HCC, hepatocellular carcinoma otherwise unspecified; INR, international normalized ratio; MELD, model for end-stage liver disease; NASH, non-alcoholic steatohepatitis; RRT, renal replacement therapy; uPRBCs, units of packed red blood cells; WIT, warm ischemia time.

*Continuous variables expressed as median IQR.

¶percentages of given underlying liver disease reported patients without liver cancer, all primary liver cancer patients reported as HCC irrespective of underlying liver disease.

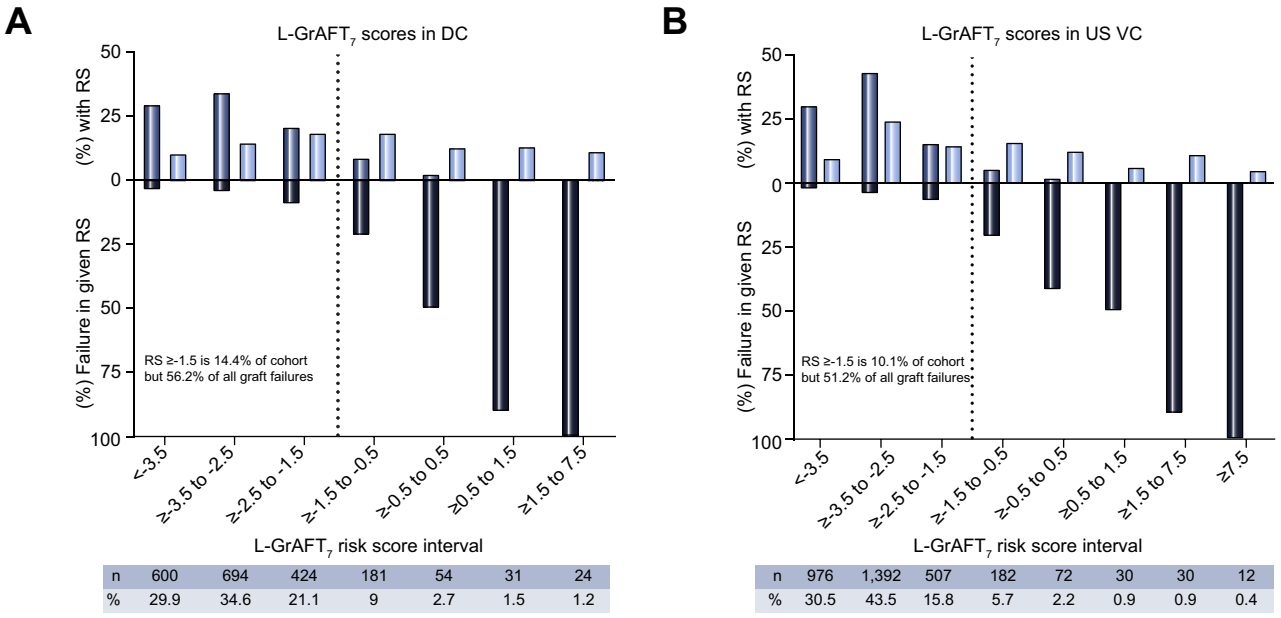
**for international COPE DRI calculation, assumed all donors are regional.

+total preservation time is time from donor x-clamp to portal reperfusion, and is the same as CIT for all allografts undergoing static cold storage, for normothermic machine perfusion patients, time on the pump counts towards total preservation time.

patients stratified by hospitalization status prior to LT, recipient high (≥30) or low (<30) MELD, donor allograft quality stratified at the DRI median, and individual validation cohort centers, the L-GrAFT$_7$ had similarly excellent accuracies and consistently outperformed both the EAD and MEAF scores. In a separate sensitivity analysis to verify that the imputation of missing variables did not significantly impact the accuracy evaluation, a complete case analysis of 1,295 patients who did not have a single postoperative laboratory value of AST, INR, bilirubin, and platelets missing from day 1–10 was performed. The L-GrAFT$_7$ had an even higher overall C-statistic of 0.82 (IQR 0.78–0.86), significantly superior to the EAD (0.73 [IQR 0.69–0.76], p <0.001) and MEAF score (0.73 [IQR 0.69–0.78], p <0.001), with subanalyses across the same recipient subsets stratified by hospitalization status, recipient MELD, donor DRI, and transplant center revealing similar superiority for L-GrAFT$_7$. Finally, in a sensitivity landmark analysis excluding 67 patients with early graft failures prior to day 7, L-GrAFT$_7$ had a significantly superior AUROC compared to both EAD and MEAF in both imputed analysis (L-GrAFT$_7$ 0.74 [IQR 0.69–0.78] *vs.* EAD 0.67 [IQR 0.63–0.71; p <0.001] and MEAF 0.66 [IQR 0.61–0.71; p = 0.001]) and complete case analysis (L-GrAFT$_7$ 0.77 [IQR 0.72–0.83] *vs.* EAD 0.69 [IQR 0.64–0.74; p <0.001] and MEAF 0.68 [IQR 0.62–0.74; p = 0.002]).

The L-GrAFT$_7$ model also demonstrated good calibration, with excellent agreement in the observed and predicted probabilities of 3-month graft failure stratified by deciles of risk score in each US VC (Fig. S1; BTSI, Fig. S1A–B, goodness of fit p = 0.303; MS, Fig. S1C–D, goodness of fit p = 0.370; WU, Fig. S1E–F, goodness of fit p <0.001).

Validation results for the L-GrAFT$_{10}$ model are shown in Table S2. In an analysis of all 3,201 patients in the US VC (Fig. S2), L-GrAFT$_{10}$ had a C-statistic of 0.82 (IQR 0.78–0.85), significantly superior to EAD (0.68 [IQR 0.65–0.71], p <0.001) and MEAF (0.72 [IQR 0.68–0.76], p <0.001), with similar superiority across all subset sensitivity analyses and in a complete case analysis of 1,295 patients (L-GrAFT$_{10}$ C-statistic 0.84 [IQR 0.80–0.88], compared to EAD 0.73 [IQR 0.69–0.76] and MEAF 0.73 [IQR 0.69–0.78]; both p <0.001). In a similar landmark analysis excluding patients with early graft failure, L-GrAFT$_{10}$ had a significantly superior AUROC compared to both EAD and MEAF in



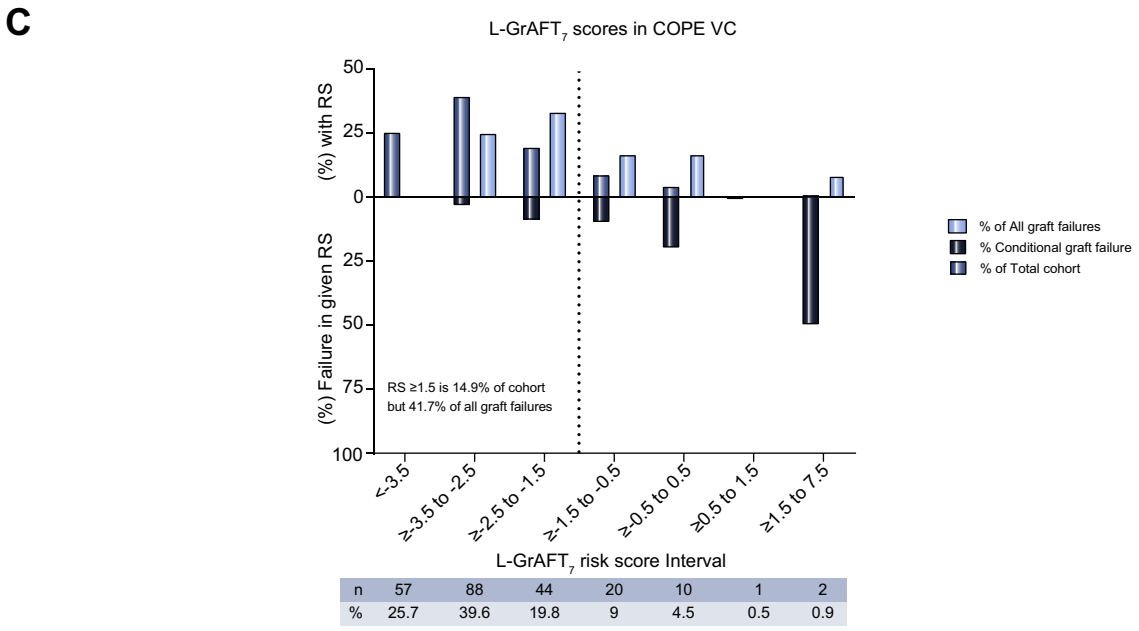


Fig. 2. Histogram comparing the distribution of L-GrAFT$_7$ scores (solid dark blue bars), conditional risk of graft failure (solid black bars) and percentage of all graft failures (light blue bars) within predefined risk score intervals. (A) US DC. (B) US VC. (C) International COPE cohort. Total number (n) and percentage of total (%) for each risk score interval shown in tables below. L-GrAFT$_7$ risk scores ≥ −1.5 contribute only 14.4%, 10.1%, and 14.9% in the DC (A), US VC (B), and COPE (C), but account for 56.2%, 51.2%, and 41.7% of all graft failures, respectively. COPE, Consortium for Organ Preservation in Europe; DC, derivation cohort; VC, validation cohort.

both imputed analysis (L-GrAFT$_{10}$ 0.77 [IQR 0.72–0.81] *vs.* EAD 0.67 [IQR 0.63–0.71; p <0.001] and MEAF 0.66 [IQR 0.61–0.71; p <0.001]) and complete case analysis (L-GrAFT$_{10}$ 0.80 [IQR 0.74–0.85] *vs.* EAD 0.69 [IQR 0.64–0.74; p <0.001] and MEAF 0.68 [IQR 0.62–0.74; p <0.001]).

International COPE validation cohort

External validation of the L-GrAFT$_7$ model in the COPE validation cohort is shown in Fig. 3. L-GrAFT$_7$ had a C-statistic of 0.81 (IQR 0.67–0.94; Fig 3D), significantly superior to EAD 0.64 (IQR 0.47–0.81, p = 0.040; Fig 3E) and MEAF 0.57 (IQR 0.35–0.79, p = 0.025; Fig 3F) in the 171 patients with non-missing laboratory values allowing for calculation of all 3 scores. In a sensitivity analysis of 219 of the 222 COPE patients where L-GrAFT$_7$ and EAD can be calculated (Fig. S3), L-GrAFT$_7$ had a C-statistic of 0.77 (IQR 0.63–0.90), significantly superior to EAD 0.65 (IQR 0.50–0.80, p = 0.007).

Evaluation of L-GrAFT$_7$ in prospective COPE randomized controlled trial arms

L-GrAFT$_7$ scores and a comparison of its individual components in the NMP and SCS groups in the COPE RCT cohort are shown in Table S3. NMP and SCS recipients had similar unadjusted median

Table 4. Validation AUROCs for L-GrAFT$_7$ compared to EAD and MEAF in US validation cohort.

	AUROC (95% CI)						
	L-GrAFT$_7$	EAD	MEAF	L-GrAFT$_7$ *vs.* EAD		L-GrAFT$_7$ *vs.* MEAF	
				AUC Diff	*p* value	AUC Diff	*p* value
Overall model (n = 3,201)*	**0.78 (0.75–0.82)**	**0.68 (0.65–0.71)**	**0.72 (0.68–0.76)**	**0.100**	**<0.001**	**0.060**	**<0.001**
Subgroup analysis							
Pre-LT recipient status							
Hospitalized	0.82 (0.76–0.88)	0.72 (0.66–0.78)	0.73 (0.67–0.79)	0.103	<0.001	0.092	0.010
Home	0.76 (0.72–0.81)	0.67 (0.64–0.71)	0.71 (0.66–0.76)	0.089	<0.001	0.052	0.018
Recipient lab MELD							
≥30	0.80 (0.73–0.87)	0.73 (0.67–0.79)	0.72 (0.65–0.80)	0.071	0.015	0.076	0.066
<30	0.77 (0.73–0.81)	0.67 (0.63–0.71)	0.71 (0.67–0.76)	0.101	<0.001	0.060	0.006
DRI							
DRI above median	0.78 (0.73–0.84)	0.67 (0.62–0.72)	0.70 (0.64–0.75)	0.117	<0.001	0.087	0.002
DRI below median	0.75 (0.68–0.82)	0.69 (0.63–0.75)	0.69 (0.62–0.77)	0.056	0.077	0.054	0.152
Validation center							
BSTI – BAS + BUMC	0.80 (0.75–0.85)	0.72 (0.68–0.76)	0.76 (0.72–0.81)	0.080	<0.001	0.038	0.096
Mount Sinai	0.81 (0.73–0.90)	0.68 (0.59–0.77)	0.74 (0.64–0.84)	0.135	<0.001	0.073	0.065
Washington University	0.73 (0.66–0.81)	0.64 (0.57–0.70)	0.68 (0.60–0.77)	0.097	0.003	0.048	0.213
Complete case analysis**	**0.82 (0.78–0.86)**	**0.73 (0.69–0.76)**	**0.73 (0.69–0.78)**	**0.095**	**<0.001**	**0.0848**	**<0.001**
Subgroup analysis							
Pre-LT recipient status							
Hospitalized	0.83 (0.76–0.90)	0.74 (0.67–0.81)	0.69 (0.61–0.77)	0.091	<0.001	0.141	0.002
Home	0.81 (0.76–0.87)	0.72 (0.67–0.76)	0.75 (0.70–0.81)	0.097	<0.001	0.062	0.015
Recipient lab MELD							
>30	0.81 (0.73–0.89)	0.73 (0.65–0.80)	0.72 (0.63–0.81)	0.084	0.001	0.092	0.073
<30	0.82 (0.77–0.87)	0.72 (0.67–0.76)	0.74 (0.68–0.79)	0.101	<0.001	0.082	0.001
DRI							
DRI above median	0.79 (0.71–0.86)	0.69 (0.63–0.75)	0.69 (0.62–0.77)	0.099	<0.001	0.094	0.020
DRI below median	0.84 (0.77–0.92)	0.77 (0.70–0.84)	0.76 (0.67–0.84)	0.079	0.011	0.089	0.052
Validation center							
BSTI – BAS + BUMC	0.81 (0.75–0.88)	0.75 (0.70–0.80)	0.80 (0.74–0.86)	0.062	0.026	0.017	0.599
Mount Sinai	0.82 (0.73–0.91)	0.68 (0.58–0.77)	0.73 (0.63–0.84)	0.141	<0.001	0.084	0.035
Washington University	0.76 (0.67–0.85)	0.67 (0.59–0.74)	0.72 (0.63–0.81)	0.093	0.003	0.039	0.419

AUROC, area under receiver operator characteristics curve; AUC diff, mean difference between model AUCs; DRI, donor risk index; EAD, early allograft dysfunction; L-GrAFT$_7$, liver graft assessment 7 days following transplantation; MEAF, model for early allograft function; MELD, model for end-stage liver disease.
*Overall model accuracies based on all 3,201 patients based on 200 bootstrapped samples with imputation of missing variables.
**Complete case analysis in subset of 1,295 patients with 0 missing variables in the first 7 post-LT days, performed as a sensitivity analysis.

L-GrAFT$_7$ scores (−2.88 [−3.47 to −2.12] *vs.* −3.05 [−3.56 to −2.09], $p = 0.616$). Evaluating individual L-GrAFT$_7$ lab measures, NMP LT recipients had significantly lower median overall AST (AUC$_7$ logAST, 33.4 *vs.* 35.5, $p = 0.001$) and bilirubin (AUC$_7$ logBilirubin, 4.69 *vs.* 6.73, $p = 0.027$) exposure over 7 days; similar measures of INR (AUC$_7$ INR), platelets (AUC$_7$ logPlatelets), and rate of normalization of bilirubin (Slope$_7$ logBilirubin); but slower rate of normalization of AST over 7 days (Slope$_7$ logAST −0.32 *vs.* −0.47, $p < 0.001$) compared to SCS. However, after adjusting for factors that significantly impacted the L-GrAFT$_7$ score in the larger US VC – total preservation time, DRI, DCD donors, recipient lab MELD, and implantation WIT – NMP recipients had a calculated median L-GrAFT$_7$ of −2.93 (IQR −3.71 to −2.03) and SCS recipients had a median L-GrAFT$_7$ of −2.65 (IQR −3.34 to −1.98, $p = 0.079$).

The association of L-GrAFT$_7$ with time to post-LT adverse events and need for RRT, irrespective of trial arm, is shown in Table 5. Increasing tertiles of L-GrAFT$_7$ score were not significantly associated with the development of a composite measure of any adverse event, or infectious, bleeding, and genitourinary complications. However, increasing tertiles of L-GrAFT$_7$ were associated specifically with liver allograft complications (tertile 3 [HR 2.17, $p = 0.016$] compared to tertile 1), clinically relevant Clavien grade ≥IIIB (tertile 3 [HR 2.60, $p = 0.034$] compared to tertile 1) and grade ≥IVa complications (tertile 3 [HR 4.99, $p = 0.011$] compared to tertile 1), and need for post-LT RRT (odds ratio 3.62 for L-GrAFT$_7$ >median *vs.* <median, $p = 0.016$). Finally, COPE LT recipients with the highest L-GrAFT$_7$ scores (>90th percentile) had significantly longer post-LT length of hospitalization compared to recipients with lower risk scores (17.2 *vs.* 11.0 days, $p = 0.002$), even after adjusting for total preservation time, DRI, DCD donors, recipient lab MELD, and implantation WIT (17.1 *vs.* 11.7 days, $p = 0.006$).

Discussion

Despite an increasing number of liver transplants performed in the United States, there has been a concomitant record number of new liver waitlist registrants totaling 11,844 in 2018.[20] Subsequently, organ allograft shortage is the greatest barrier to realizing the life-saving benefits of LT, incentivizing the development of technological advances[8,21] and translational studies[22] to mitigate ischemia-reperfusion injury and allow for a safe and meaningful expansion of the donor pool. As the transplant community anticipates a growing number of such clinical trials, a tool to accurately measure early allograft function that is associated with relevant clinical outcomes has become an increasingly important unmet need. The L-GrAFT risk score[14] vastly outperformed the most widely used EAD definition[12] and the MEAF score[13] in predicting 3-month graft failure-free survival in the original derivation cohort. Herein, we report a large scale, multicenter and international validation of the L-GrAFT score in more than 3,200 patients from 3 US transplant centers, and 222

Table 5. Association of L-GrAFT$_7$ with time to post-LT complications in the prospective COPE validation dataset.

Complication*	Proportion of patients (%)	Hazard ratio	95% CI	*p* value
Any adverse event	125/222 (56.3)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		1.12	0.73–1.73	0.596
L-GrAFT$_7$ Tertile 3		1.16	0.75–1.80	0.497
Infectious complication	42/222 (18.9)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		0.74	0.34–1.62	0.457
L-GrAFT$_7$ Tertile 3		0.61	0.27–1.38	0.235
Bleeding complication	15/222 (6.8)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		0.99	0.29–3.42	0.988
L-GrAFT$_7$ Tertile 3		0.79	0.21–2.96	0.729
Genitourinary complication	25/222 (11.3)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		1.18	0.36–3.88	0.781
L-GrAFT$_7$ Tertile 3		2.26	0.78–6.50	0.131
Hepatic complication	92/222 (41.4)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		1.83	0.97–3.48	0.063
L-GrAFT$_7$ Tertile 3		2.17	1.16–4.09	0.016
Clavien grade ≥IIIB	57/222 (25.7)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		2.09	0.84–5.17	0.112
L-GrAFT$_7$ Tertile 3		2.60	1.08–6.27	0.034
Clavein grade ≥IVA	40/222 (18.0)			
L-GrAFT$_7$ Tertile 1		1.00	ref	ref
L-GrAFT$_7$ Tertile 2		3.09	0.84–11.4	0.091
L-GrAFT$_7$ Tertile 3		4.99	1.43–17.4	0.011
		Odds ratio		
Renal failure requiring RRT+	21/219 (9.6)			
L-GrAFT$_7$ >median		3.62	1.28–10.3	0.016

COPE, Consortium for Organ Preservation in Europe; L-GrAFT$_7$, liver graft assessment 7 days following transplantation; LT, liver transplantation; RRT, renal replacement therapy.
*Adverse events and complications as defined in the Extended Data Tables 5 and 6 in COPE RCT publication[15]; + requiring RRT >1-week post-LT.

patients in the European COPE cohort who participated in the first human prospective randomized clinical trial of NMP.[15] Our findings show the L-GrAFT to be significantly superior to both the EAD and MEAF scores in both US and European cohorts, and demonstrate for the first time the utility of L-GrAFT in a human clinical trial of NMP; namely, that L-GrAFT not only accurately predicted graft and patient survival following LT, but was highly associated with relevant clinical endpoints such as resource utilization and the development of serious post-LT adverse events.

One of the hallmarks of validating a clinically useful model is to establish its generalizability to populations that are different from the original model derivation cohort.[23] In this regard, our L-GrAFT validation study was robust. The US VC included 3,201 patients who had different clinical characteristics compared to the original derivation cohort, including significant differences in recipient diagnosis, and notably lower acuity as measured by the recipient laboratory MELD and proportion of patients requiring pretransplant hospitalization, RRT, and mechanical ventilation. Not surprisingly, this was reflected in the inferior graft and patient survival observed in the derivation cohort that would be anticipated with transplantation of sicker recipients (Fig. 1A,B). Additionally, the 3 individual US VCs also differed from one another across numerous recipient and donor characteristics, and represent 3 different UNOS regions where differences in waiting times, practice patterns, patient selection, and center-specific operative and perioperative management exists – further supporting the generalizability of the model. Finally, inclusion of an international European cohort of 222 patients participating in the first randomized controlled trial of NMP in liver transplantation, where there was significantly greater utilization of DCD organs and with longer preservation times allowed for validation of L-GrAFT$_7$ in practice circumstances not typically observed in the United States.

Despite significant differences in recipient and donor characteristics and graft and patient outcomes among the derivation and validation cohorts, L-GrAFT maintained an excellent accuracy in discriminating 3-month graft failure-free survival. L-GrAFT$_7$, based on 7 days of post-LT laboratory measures, had a validated C-statistic of 0.78 in the US validation cohort of 3,201 patients, significantly superior to the EAD (C-statistic 0.68, p <0.001) and MEAF (C-statistic 0.72, p <0.001). In a sensitivity analysis of 1,295 patients without any missing laboratory variables over 7 post-LT days (Table 4), L-GrAFT$_7$ had even a higher C-statistic of 0.82, outperforming both EAD (C-statistic 0.73, p <0.001) and MEAF (C-statistic 0.73, p <0.001), and confirming that imputation of random missing variables in the overall cohort did not impact the reliability of the analysis. L-GrAFT$_{10}$, based on 10 days of post-LT variables, had even higher accuracy and outperformed EAD and MEAF in similar analyses. Further sensitivity analysis also verified the generalizability of the L-GrAFT to recipients with different characteristics based on hospitalization status, high or low recipient MELD and DRI, and validation center in both the overall population and complete case analysis for both L-GrAFT$_7$ and L-GrAFT$_{10}$. In the prospective COPE VC, only 7 post-LT days of laboratory variables were

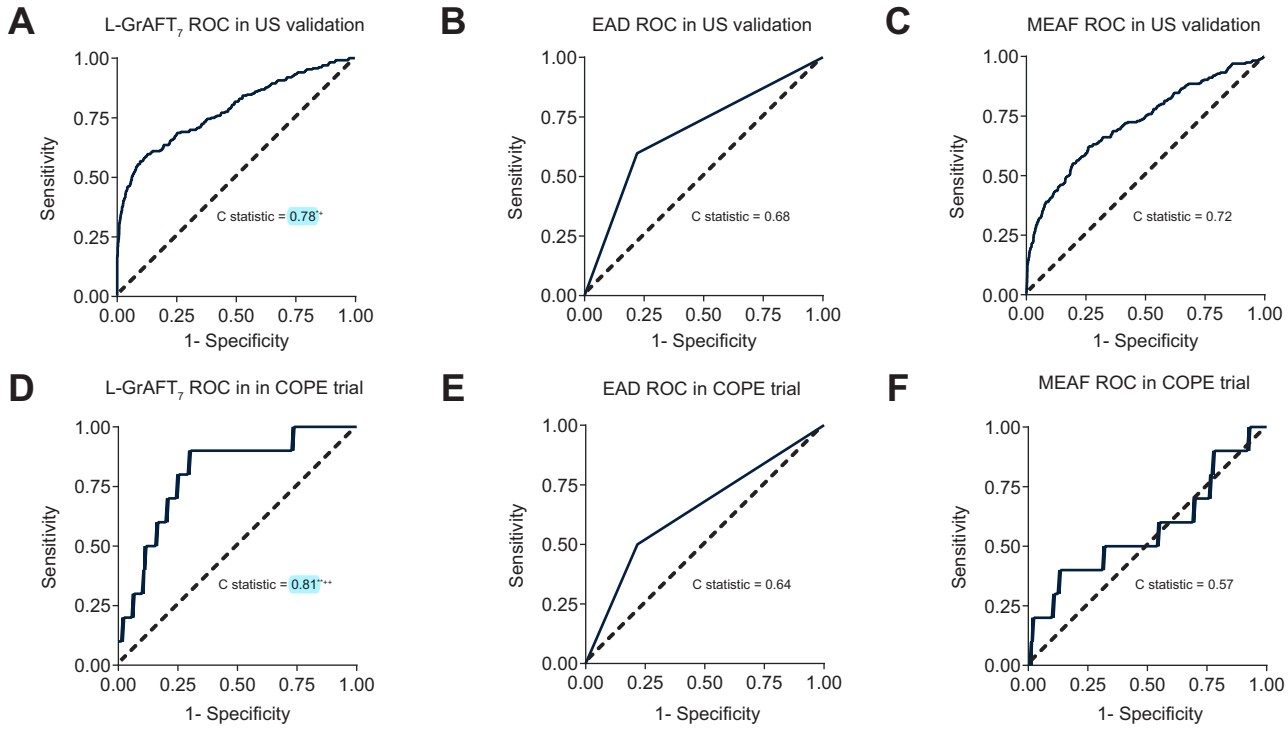


Fig. 3. Comparison of AUROC curves for 3 models in US VC and international COPE cohort. (A) AUROC for L-GrAFT$_7$ in US VC. (B) AUROC for binary EAD in US VC. (C) AUROC for MEAF in US VC. (D) AUROC for L-GrAFT$_7$ in COPE cohort. (E) AUROC for EAD in COPE cohort. (F) AUROC for MEAF in COPE cohort. ** p <0.001 compared to EAD and MEAF in the US validation cohort (log-rank test). $^{**}p$ = 0.040 compared to EAD in COPE validation cohort (log-rank test). ^{++}p = 0.025 compared to MEAF in the COPE validation cohort (log-rank test). COPE, Consortium for Organ Preservation in Europe; EAD, early allograft dysfunction; L-GrAFT$_7$, liver graft assessment 7 days following transplantation; MEAF, model for early allograft function; VC, validation cohort.

recorded, allowing for validation of L-GrAFT$_7$ as a more accurate tool (C-statistic of 0.81) compared to EAD (C-statistic 0.64, p = 0.040; Fig. 3E) and MEAF (C-statistic 0.57, p = 0.024; Fig. 3F) in predicting 3-month graft failure-free survival. Evaluation of the distribution of L-GrAFT$_7$ in the derivation cohort and both validation cohorts consistently showed that L-GrAFT$_7$ scores greater than −1.5 comprise only 10–15% of each cohort but nearly half of all graft failures, with the conditional risk of graft failure increasing significantly with scores above −1.5. Very few previous studies have reported the accuracy of these models in discriminating 3-month survival. The original study defining MEAF by Pareja and colleagues[13] did not report the C-statistic. Recently, Jochmans *et al.*[24] compared the EAD to the MEAF score in 660 patients undergoing LT, and found C-statistics of 0.64 and 0.73 in predicting 3-month survival, respectively, in line with our current findings.

Arguably one of the most exciting advances in the field of liver transplantation has been utilization of machine perfusion technologies (NMP, HOPE, D-HOPE, *etc.*) to mitigate ischemia-reperfusion injury and allow for utilization of marginal allografts.[9] The widespread adoption of such technologies will certainly require more randomized controlled trials; however, high graft and patient survival rates pose a barrier in achieving sufficient sample sizes should they be used as the primary endpoint, necessitating the need for other surrogates of graft function that correlate with clinically meaningful outcomes beyond survival.[25] In fact, the majority of all currently registered liver machine perfusion trials (ClinicalTrials.gov, n = 21) have designated EAD as either the primary or secondary outcome, and the FDA has mandated utilizing EAD as the primary outcome for the US NMP RCT (NCT02775162) utilizing the OrganOx Metra NMP device, the approval of which will depend on the trial meeting its primary endpoint of lower EAD in the experimental group. In this context, our *post hoc* analysis of the COPE NMP study, which revealed far greater accuracy of L-GrAFT in measuring early allograft function compared to EAD, provides a strong rationale for utilizing L-GrAFT as a clinical endpoint in translational trials. Furthermore, in unadjusted analysis, there was no difference in L-GrAFT$_7$ scores between NMP and SCS trial arms; the trial was not powered for this outcome. Evaluation of the individual components of the L-GrAFT$_7$ score did reveal significantly lower exposure (AUC) of AST and bilirubin in NMP; however, rate of normalization of AST was faster in the SCS group, resulting in non-significant differences in the overall L-GrAFT$_7$ score (Table S3). However, these unadjusted results presuppose that trial randomization balanced for factors impacting the L-GrAFT score. After adjusting for total preservation time, recipient MELD, donor risk index, DCD organs, and WIT – factors which had a significant impact on L-GrAFT$_7$ in multivariate regression in the US VC – we found that L-GrAFT$_7$ favored the NMP cohort over SCS, and estimated what the impact of NMP would have been had these 5 factors been equally distributed amongst the groups.

Arguably the most important finding from our *post hoc* COPE analysis was that L-GrAFT$_7$ was highly associated with clinically meaningful outcomes beyond graft and patient survival. The strengths of this analysis were based on the fact that post-LT morbidity and adverse events were meticulously recorded in the setting of a prospective randomized trial, representing

datapoints that are notoriously subject to bias when evaluated in a retrospective fashion. Beyond L-GrAFT$_7$ being highly predictive of 3-month graft failure-free survival, it was significantly associated with the development of liver allograft complications specifically, as well as Clavien grade IIIB and greater complications. The fact that L-GrAFT$_7$ was not associated with genitourinary, infectious, or adverse events in general supports its utility as a specific marker of hepatic allograft function. Furthermore, higher L-GrAFT$_7$ scores were associated with longer post-LT length of hospitalization (marker of cost and resource utilization),[26] as well as the development of post-LT RRT, a clinically important adverse event known to be associated with allograft dysfunction.[3] Taken as a whole, these results corroborate the importance and utility of L-GrAFT$_7$ in identifying the degree of hepatic allograft injury following LT.

Beyond its utility as a clinical endpoint for translational studies, L-GrAFT may also be used to identify LT recipients at the highest risk of graft loss due to ischemia-reperfusion injury, allowing for possible mitigation of this risk through more proactive monitoring and pharmacologic interventions.[27] With promising new preclinical data on the use of RNA interference[28] and extracellular vesicles[29] to treat liver ischemia-reperfusion injury, L-GrAFT may also allow for allocation of what will undoubtedly be high-cost/high-resource interventions to patients who stand to benefit most. Finally, L-GrAFT can be utilized to more accurately and effectively counsel liver transplant recipients and their families about the risks of adverse outcomes, and potentially guide earlier retransplantation in the highest risk patients.

In conclusion, we have validated the L-GrAFT score in a large, multicenter, international cohort as a highly accurate measure of EAD that is superior to the existing binary EAD definition and MEAF score. Its consistently excellent performance across varying recipient, donor, and operative characteristics as well as heterogeneous practice environments establishes it as a highly generalizable tool allowing for an individualized risk prediction of 3-month graft failure, and for a standardized grading of allograft function across all transplant settings. In a *post hoc* analysis of the first human randomized controlled trial of NMP in LT, L-GrAFT$_7$ accurately predicted 3-month graft failure-free survival, and was highly associated with post-LT resource utilization and time to clinically relevant serious adverse events specific to hepatic allograft dysfunction. L-GrAFT$_7$ should be considered for use as a highly accurate clinical endpoint to evaluate the efficacy of interventions in the increasing number of translational studies that aim to mitigate hepatic ischemia-reperfusion injury, as well as identifying LT recipients at the highest risk of graft failure who may benefit from pharmacologic intervention and early retransplantation.

Abbreviations

ALT, alanine aminotransferase; AST, aspartate aminotransferase; AUC, area under the curve; AUROC, area under the receiver operator characteristic; CIT, cold ischemia time; COPE, Consortium for Organ Preservation in Europe; COPE VC, COPE validation cohort; DBD, donation after brain death; DCD, donation after cardiac death; DRI, donor risk index; EAD, early allograft dysfunction; HCC, hepatocellular carcinoma; INR, international normalized ratio; LD, living donor; L-GrAFT, liver graft assessment following transplantation; LT, liver transplantation; MEAF, model for early allograft function; MELD, model for end-stage liver disease; NASH, non-alcoholic steatohepatitis; NMP, normothermic machine perfusion; RRT, renal replacement therapy; SCS, static cold storage; UNOS, United Network for Organ Sharing; uPRBC, units of packed red blood cells; US VC, US validation cohort; WIT, warm ischemia time.

Financial support

The authors received no financial support to produce this manuscript.

Conflict of interest

The authors declare no conflicts of interest that pertain to this work.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Study concept and design: VA, DM; Acquisition of data: VA, GS, NV, PT, BH, DN, YLB, MPH-L, VX; Analysis and interpretation of data: VA, DM, GBK, WCC, SSF, DN, PJF, RP, JD, FMK, HY, DGF, RWB; Drafting of the manuscript: VA, DM, DN; Critical revision of manuscript: VA, GBK, GS, WCC, NV, SSF, PT, BH, DN, PJF, YLB, RP, MPH-L, VX, JD, FMK, HY, DGF, RWB; Statistical analysis: VA, DM; Study Supervision: VA.

Data availability statement

Data for this study is confidential patient information regulated by the IRB of each participating institution. Requests to access data will have to be in compliance with each institutional IRB.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2020.09.015.

References

[1] Lee DD, Singh A, Burns JM, Perry DK, Nguyen JH, Taner CB. Early allograft dysfunction in liver transplantation with donation after cardiac death donors results in inferior survival. Liver Transpl 2014;20:1447–1453.
[2] Pomposelli JJ, Goodrich NP, Emond JC, Humar A, Baker TB, Grant DR, et al. Patterns of early allograft dysfunction in adult live donor liver transplantation: the A2ALL experience. Transplantation 2016;100:1490–1499.
[3] Wadei HM, Lee DD, Croome KP, Mai ML, Golan E, Brotman R, et al. Early allograft dysfunction after liver transplantation is associated with short- and long-term kidney function impairment. Am J Transplant 2016;16:850–859.
[4] Zhai Y, Petrowsky H, Hong JC, Busuttil RW, Kupiec-Weglinski JW. Ischaemia-reperfusion injury in liver transplantation–from bench to bedside. Nat Rev Gastroenterol Hepatol 2013;10:79–89.
[5] Agopian VG, Petrowsky H, Kaldas FM, Zarrinpar A, Farmer DG, Yersiz H, et al. The evolution of liver transplantation during 3 decades: analysis of 5347 consecutive liver transplants at a single center. Ann Surg 2013;258:409–421.
[6] Im GY, Cameron AM, Lucey MR. Liver transplantation for alcoholic hepatitis. J Hepatol 2019;70:328–334.
[7] Abreu P, Gorgen A, Oldani G, Hibi T, Sapisochin G. Recent advances in liver transplantation for cancer: the future of transplant oncology. JHEP Rep 2019;1:377–391.
[8] Boteon YL, Afford SC. Machine perfusion of the liver: which is the best technique to mitigate ischaemia-reperfusion injury? World J Transpl 2019;9:14–20.
[9] Dutkowski P, Guarrera JV, de Jonge J, Martins PN, Porte RJ, Clavien PA. Evolving trends in machine perfusion for liver transplantation. Gastroenterology 2019;156:1542–1547.
[10] Deschenes M, Belle SH, Krom RA, Zetterman RK, Lake JR. Early allograft dysfunction after liver transplantation: a definition and predictors of outcome. National Institute of Diabetes and Digestive and Kidney Diseases Liver Transplantation Database. Transplantation 1998;66:302–310.

[11] Wiesner R, Edwards E, Freeman R, Harper A, Kim R, Kamath P, et al. Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology 2003;124:91–96.
[12] Olthoff KM, Kulik L, Samstein B, Kaminski M, Abecassis M, Emond J, et al. Validation of a current definition of early allograft dysfunction in liver transplant recipients and analysis of risk factors. Liver Transpl 2010;16:943–949.
[13] Pareja E, Cortes M, Hervas D, Mir J, Valdivieso A, Castell JV, et al. A score model for the continuous grading of early allograft dysfunction severity. Liver Transpl 2015;21:38–46.
[14] Agopian VG, Harlander-Locke MP, Markovic D, Dumronggittigule W, Xia V, Kaldas FM, et al. Evaluation of early allograft function using the liver graft assessment following transplantation risk score model. JAMA Surg 2018;153:436–444.
[15] Nasralla D, Coussios CC, Mergental H, Akhtar MZ, Butler AJ, Ceresa CDL, et al. A randomized trial of normothermic preservation in liver transplantation. Nature 2018;557:50–56.
[16] Feng S, Goodrich NP, Bragg-Gresham JL, Dykstra DM, Punch JD, DebRoy MA, et al. Characteristics associated with liver graft failure: the concept of a donor risk index. Am J Transplant 2006;6:783–790.
[17] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.
[18] Shao J, Sitter RR. Bootstrap for imputed survey data. J Am Stat Assoc 1996;91:1278–1288.
[19] Perkins NJ, Cole SR, Harel O, Tchetgen Tchetgen EJ, Sun B, Mitchell EM, et al. Principled approaches to missing data in epidemiologic studies. Am J Epidemiol 2018;187:568–575.
[20] 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 s1):193–299.
[21] Bramstedt KA, Hoang JB. Technology- and policy-based strategies for increasing supply of deceased donor livers. AMA J Ethics 2016;18:143–152.
[22] Nacif LS, Kim V, Galvao F, Ono SK, Pinheiro RS, Carrilho FJ, et al. Translational medical research and liver transplantation: systematic review. Transl Gastroenterol Hepatol 2018;3:91.
[23] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–524.
[24] Jochmans I, Fieuws S, Monbaliu D, Pirenne J. "Model for early allograft function" Outperforms "early allograft dysfunction" as a predictor of transplant survival. Transplantation 2017;101:e258–e264.
[25] Lee DD, Croome KP, Shalev JA, Musto KR, Sharma M, Keaveny AP, et al. Early allograft dysfunction after liver transplantation: an intermediate outcome measure for targeted improvements. Ann Hepatol 2016;15:53–60.
[26] Croome KP, Hernandez-Alejandro R, Chandok N. Early allograft dysfunction is associated with excess resource utilization after liver transplantation. Transplant Proc 2013;45:259–264.
[27] Cannistra M, Ruggiero M, Zullo A, Gallelli G, Serafini S, Maria M, et al. Hepatic ischemia reperfusion injury: a systematic review of literature and the role of current drugs and biomarkers. Int J Surg 2016;33(Suppl 1):S57–S70.
[28] Bruggenwirth IMA, Martins PN. RNA interference therapeutics in organ transplantation: the dawn of a new era. Am J Transplant 2020;20:931–941.
[29] Ali M, Pham A, Wang X, Wolfram J, Pham S. Extracellular vesicles for treatment of solid organ ischemia-reperfusion injury. Am J Transplant 2020;20:3294–3307.