8-1-2022

Evaluating the performance and external validity of machine learning-based prediction models in liver transplantation: an international study

Tommy Ivanics

D. So

M. Claasen

D. Wallace

M. Patel

See next page for additional authors

Authors

Tommy Ivanics, D. So, M. Claasen, D. Wallace, M. Patel, A. Gravely, K. Walker, T. Cowling, L. Erdman, and G. Sapisochin

Plenary Abstract Session I

O-005

Evaluating the performance and external validity of machine learning-based prediction models in liver transplantation: an international study

T. Ivanics[1,2,3], D. So[4], M. Claasen[1,5], D. Wallace[6], M. Patel[7], A. Gravely[1], K. Walker[6], T. Cowling[6], L. Erdman[4], G. Sapisochin[1]

[1]University Health Network, Multi-Organ Transplant Program, Toronto, Canada, [2]Henry Ford Hospital, Department of Surgery, Detroit, United States, [3]Uppsala University, Department of Surgical Sciences, Uppsala, Sweden, [4]The Hospital for Sick Children, The Centre of Computational Medicine, Toronto, Canada, [5]Erasmus MC, Department of Surgery, Rotterdam, Netherlands, [6]London School of Hygiene and Tropical Medicine, Department of Health Services Research and Policy, London, United Kingdom, [7]University of Texas Southwestern Medical Center, Division of Surgical Transplantation, Department of Surgery, Dallas, United States

Background: National liver transplant (LT) registry data are curated in many countries. We compared data from three national registries and developed machine learning algorithm (MLA)-based models to predict post-LT 90-day mortality within and across different countries. Predictive performance and external validity of each model was assessed to contextualize the applicability of MLA in LT.
Methods: We studied adult (≥18-years) primary LTs between Jan-2008 and Dec-2018 from United Network for Organ Sharing (UNOS-US), National Health Service Blood and Transplantation (NHSBT-UK), and the Canadian Organ Replacement Registry (CORR-Canada). MLA models for 90-day post-LT mortality were built firstly on each individual registry (based on variables inherent to the individual database) and then using all 3 registries (based on harmonized variables). The predictive abilities of the models were evaluated across countries using area-under-the-receiver-operator-curve (AUROC) and area-under-the-precision-recall-curve (AUPRC).
Results: Patients included were as follows: Canada n=1,214, UK n=5,287, and US n=59,558. ElasticNet had the best performance across both individual registries and harmonized datasets. Model performance diminished from the individualized registries to the harmonized registry (only using variables in common between the three registries), especially in the UK (individualized ElasticNet:AUROC:0.54;Range 0.52-0.56 to harmonized AUROC:0.48; Range:0.48-0.50) and the US (individualized ElasticNet:AUROC:0.70; Range:0.70-0.71 to harmonized AUROC:0.65;Range:0.64-0.65). Model performance after external validation across countries was overall poor.
Conclusions: MLA-based models can be constructed using international LT registries, with independent ElasticNet models demonstrating optimal predictive performance. While MLA-based models yield fair discriminatory potential when used within individual databases, the external validity is poor when applied to different registries across countries. This is likely due to inherent limitations and variability within each dataset. It is conceivable that these limitations may be overcome by increasing the granularity of datasets (e.g., with linkages to other administrative datasets) and placing an increased emphasis on consensus for variable standardization.

Table 1. Harmonized - Cross Country Test Set Performance, Mean (range) across 5 imputations

Model (Country model was trained on)	Registry (Country predictions made on)	AUROC	AUPRC
CA	CA	0.58 (0.57 to 0.60)	0.16 (0.16 to 0.19)
	UK	0.68 (0.67 to 0.70)	0.25 (0.24 to 0.26)
	US	0.57 (0.55 to 0.58)	0.21 (0.18 to 0.24)
UK	CA	0.49 (0.49 to 0.50)	0.04 (0.04 to 0.04)
	UK	0.63 (0.62 to 0.64)	0.05 (0.05 to 0.05)
	US	0.65 (0.65 to 0.65)	0.08 (0.08 to 0.08)
US	CA	0.57 (0.56 to 0.58)	0.05 (0.05 to 0.05)
	UK	0.60 (0.60 to 0.60)	0.05 (0.05 to 0.05)
	US	0.63 (0.63 to 0.63)	0.06 (0.06 to 0.07)

Abbreviations: AUPRC: Area under precision-recall curve, AUROC: Area under receiver-operator characteristic, CA: Canada, UK: United Kingdom, US: United States

O-006

A population-based analysis of long-term outcomes following pediatric acute liver failure highlights high risk populations in access to transplant

J. Ascher Bartlett[1], M. Lo[2], K. Etesami[3], L. Sher[3], R. Kohli[1], J. Emamaullee[3]

[1]Children's Hospital Los Angeles, Gastroenterology, Hepatology and Nutrition, Los Angeles, United States, [2]University of Southern California, Los Angeles, United States, [3]Keck Medicine of University of Southern California, Surgery, Los Angeles, United States

Background: Pediatric acute liver failure (PALF) is a devastating illness that affects otherwise healthy children. Nearly 70% of children with PALF recover, while ≥25% require emergent liver transplant (LT). PALF has an estimated overall mortality of 15%. Beyond short-term outcomes, the long-term clinical course of PALF has not been established. This study aims to characterize long-term outcomes and identify prognostic factors of PALF using registry data.
Methods: Children (<18 years) admitted with PALF were identified by ICD codes in the California Office of Statewide Health Planning and Development Patient Discharge Dataset (1/2005-12/2018). Multivariable Cox proportional hazards models were used to identify risk factors for liver transplant and death.
Results: Among 2162 inpatients with PALF, the mean age at presentation was 9.7±6.2 years, with a median follow-up of 4.6 (IQR 0.06 -10.4) years. 50.1% were female and 44.2% Hispanic. Most deaths occurred within 4 months of presentation. Children <2 years (18.2% of patients) were more likely to undergo LT (HR 2.74 [95% CI 1.96-3.83], p<0.001) and more likely to die (HR 1.66 [95% CI 1.33-2.07], p<0.001) when compared to children >10 years old (52.3% of patients) (Table 1). Mortality was higher among females (HR 1.47 [95% CI 1.17-1.85], p<0.001). No statistically significant difference was observed based on race/ethnicity.