Received: 25 September 2024 | Accepted: 6 January 2025

DOI: 10.1097/LVT.0000000000000575

PERSPECTIVE

Evaluating allograft risk models in organ transplantation: Understanding and balancing model discrimination and calibration

David Goldberg[1] | Hemant Ishwaran[2] | Vishnu Potluri[3,4] | Michael Harhay[4] | Emily Vail[5] | Peter Abt[6] | Sarah J. Ratcliffe[7] | Peter P. Reese[4,8]

[1]Department of Medicine, Division of Digestive Health and Liver Diseases, University of Miami Miller School of Medicine, Miami, Florida, USA

[2]Department of Public Health Sciences, Division of Biostatistics, University of Miami Miller School of Medicine, Miami, Florida, USA

[3]Department of Medicine, Renal, Electrolyte, and Hypertension Division, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[4]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[5]Department of Anesthesiology and Critical Care, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[6]Department of Surgery, Division of Transplant Surgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[7]Department of Public Health Sciences, Division of Biostatistics, University of Virginia, Charlottesville, Virginia, USA

[8]Department of Surgery, Vanderbilt Center for Transplant Science, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Correspondence
David Goldberg, Department of Medicine, Division of Digestive Health and Liver Diseases, University of Miami Miller School of Medicine, Don Soffer Clinical Research Building, 1120 NW 14th Street, Room 807, Miami, FL 33136, USA.
Email: dsgoldberg@miami.edu

Abstract

In the field of organ transplantation, the accurate assessment of donor organ quality is necessary for efficient organ allocation and informed consent for recipients. A common approach to organ quality assessment is the development of statistical models that accurately predict posttransplant survival by integrating multiple characteristics of the donor and allograft. Despite the proliferation of predictive models across many domains of medicine, many physicians may have limited familiarity with how these models are built, the assessment of how well models function in their population, and the risks of a poorly performing model. Our goal in this perspective is to offer advice to transplant professionals about how to evaluate a prediction model, focusing on the key aspects of discrimination and calibration. We use liver allograft assessment as a paradigm example, but the lessons pertain to other scenarios too.

Keywords: model assessment, risk score, performance

INTRODUCTION

In the field of organ transplantation, like other areas of medicine, there has been an interest in developing and implementing predictive models. In kidney transplantation, the kidney donor risk index (DRI), which is meant to summarize allograft quality, is integrated directly into organ allocation and other real-time measures of organ

Abbreviations: C-index, concordance index; DRI, donor risk index; OPTN, Organ Procurement and Transplantation Network.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.ltxjournal.com.

quality. Although scores have been developed in the fields of lung and liver transplantation, they have not been implemented into clinical practice.[1–5]

In liver transplantation, the liver DRI was developed in 2006 but was never implemented into clinical practice.[2–5] There have been efforts to develop a better liver donor risk score, including the ID^{2}EAL score.[6] While ID^{2}EAL had superior discrimination compared to the DRI, it too had limitations preventing it from being implemented in clinical practice.[6–10] Given the limitations of available liver donor risk scores, liver transplant physicians often rely on heuristics and past experiences to make decisions that influence whether a donor's liver is used or discarded and a given patient is transplanted or waits for another offer while facing the risk of death or clinical deterioration.[3,11–13] The potential advantages of an accurate liver quality score would include a greater ability to match the projected longevity of organs to the patient's projected survival and the potential to enhance the processes of informed consent.[3,11–13] An accurate liver score could improve outcomes by supplementing the judgment of human clinicians, which can be flawed in the setting of fatigue, distractions, or well-described cognitive biases.

The goal of developing risk models is to create risk scores with excellent discrimination and calibration. However, this goal is often not achieved. Journals and clinicians often focus primarily on measures of discrimination (eg, concordance index [C-index]) when determining the decision to accept a manuscript or implement a model in clinical practices.[1] Therefore, our goal is to present a review of the topics of discrimination and calibration, using the risk score we developed to contextualize these important concepts in the specific context of organ transplantation.

A NOVEL RISK MODEL IN LIVER TRANSPLANTATION

As part of a set of projects to improve the assessment of organ quality, we sought to develop a new liver allograft risk model using state-of-the-art machine learning methods. We conducted a retrospective cohort study using data from the Organ Procurement and Transplantation Network (OPTN) of deceased liver donors from May 1, 2007, to March 31, 2022, and recipients of those transplanted organs.[14] We included a number of donor variables that have previously been shown to be predictive of graft outcome and/or biologically plausibly related to the outcome of graft failure.[2–6] We included "baseline" variables (ie, defined only once for a donor and available at the time of organ procurement), in addition to longitudinal laboratory variables from the donor hospitalization at any time point available in DonorNet, 48 hours prior to cross-clamp (ie, terminal), 24 hours prior to cross-clamp, and terminal values (Supplemental Table S1, http://links.lww.com/LVT/A896). The primary study outcome was time-to-graft failure, a continuous outcome defined as death or retransplantation. Secondary outcomes included graft failure assessed at 1, 3, 5, and 10 years. The models were built using random survival forest functions using the "rfsrc" command in R.[15] We used "VarPro" for dimension reduction (a model-free variable selection strategy that determines which variables are related to the outcomes based on the restricted mean survival time.[16] The restricted mean survival time is a measure of the average time-to-graft failure from the time of transplantation to a specific time point.

Despite the inclusion of longitudinal variables not previously explored in liver allograft models, our models had only modest discrimination, with a time-dependent AUC that varied from 0.59 to 0.62, yet excellent calibration (based on calibration plots and time-dependent Brier scores). This AUC is similar to the magnitude to that of the DRI and the ID^{2}EAL risk scores,[17,18] a recently published lung risk score,[1] and the kidney DRI, which is implemented directly into kidney transplant allocation.[19–21] Later in this paper, we will contextualize how a model with similar performance could be implemented in clinical practice.

UNDERSTANDING DISCRIMINATION AND CALIBRATION

When assessing a risk prediction model, the 2 primary performance measures are discrimination and calibration. Simply stated, discrimination refers to the ability to differentiate between those who will versus will not develop the outcome of interest. Calibration refers to the agreement between the frequency of observed events and the predicted probabilities.[22] Notably, prediction models differ from explanatory models, in which the goal is often to gain insight into the relationships between variables in a model and the outcome.

Discrimination

In organ transplantation, measures of discrimination might reflect how often a risk score correctly ranks orders of any 2 donor allografts (or recipients) with respect to their observed outcome. In liver transplantation, the relevant outcome is often graft failure, and the C-statistic reflects whether a model correctly predicts whether allograft A or B reaches that outcome first.[23] We can use the example of all livers donated by deceased donors in the year 2018. Put simply, if any 2 liver allografts are identified from the pool of livers donated in 2018, a C-index of 0.8 shows that the risk score correctly predicts 80% of the time which donated liver will first reach the outcome. Discrimination does not assess a model's accuracy in predicting the rate of the event of a graft failure overall or in important subgroups; the C-statistic for 2 models could be identical with widely varying predictions of the rate. That is why measures of calibration are also needed.

The time horizon for an outcome is a critical determination when deriving a risk model because it impacts the type of model used to derive the score (eg, logistic regression versus Cox regression) and the measure of discrimination. Two very common approaches to discrimination are the receiver operating curve, used for a binary outcome such as death, and the time-dependent AUC (or C-statistic) for a time-to-event outcome. For example, had we decided to focus our posttransplant model only on short-term posttransplant mortality (eg, in-hospital mortality, 30-d mortality), we would have used logistic regression models as the outcome (death) at any point during follow-up would be considered a bad outcome by patients, regardless of when during the short-term postoperative period death occurred. However, in the model we presented, the focus was on long-term outcomes (1, 3, 5, and 10 y); therefore, we fit time-dependent models that accounted for the outcome (death) and time because a death in a patient with cirrhosis that occurred 3 months after diagnosis has more important clinical implications than one occurring 4 years after diagnosis.[10]

In the setting of binary, rather than time-to-event outcomes, the AUROC or AUC is equivalent mathematically to Harrell C-index. However, the measures have slightly different interpretations. The AUC measures separability (ie, discrimination) based on the ROC, which plots sensitivity versus specificity. In contrast, the C-index measures the rank correlation between predicted risk scores and observed outcomes (ie, how often does the subject who experiences the outcome have a higher predicted risk probability than the person who did not have the outcome).[23–25] Both metrics have important limitations in the assessment of time-to-event outcomes, including (1) difficulty accounting for censored outcomes and (2) handling of time-dependence because 2 subjects could have the same risk score because they both achieve the outcomes despite vast differences in survival duration (ie, time-to-the outcome). With respect to censored outcomes, this is an issue when there is potential for survival beyond the time horizon that is analyzed. In our model, we focused on outcomes out to 10 years, even though median posttransplant survival is 12 years. As a result, our discrimination at 10 years does not account for events that occur after 10 years, and deaths at 11 years and 15 years would both be considered censored outcomes. With respect to time-dependence, even when evaluating survival under a time-to-event framework, the AUC in its simplest form still must account for the outcome in a binary fashion at various time intervals (eg, 3 y); therefore, deaths at year 1 versus year 2 are both considered an outcome. Approaches to account for the time-dependent issue are calculating and graphically depicting time-dependent AUCs at multiple time intervals, or more sophisticated methods that involve integrated AUCs that account for the AUC and its change over time for the duration of the analytic time period.[26]

Calibration

Calibration assesses how close a risk score's prediction of the expected outcome is to the observed outcome. For example, a prediction model with poor calibration might predict that survival probability at 3 years was 75% for livers assessed in the lowest quartile of quality when the observed (actual) survival is only 50%. Some liver transplant clinicians and patients might be willing to consider transplantation in the setting of a 75% probability of surviving, but not 50%. In practice, a model is fit for the outcome of interest, which is then often converted to a risk score based on the beta coefficients of the model serving as multipliers for the model variables. For example, if a subgroup of the recipients in the population that a given model has assigned a risk score of 2 reach the outcome 30% of the time (ie, observed), then a perfectly calibrated model will correctly predict that the outcome rate will be 30% for this subgroup (ie, expected). Calibration needs to be assessed in a few different ways. Oftentimes, authors present a quantitative assessment of calibration, most commonly using the Brier score. This provides an overarching measure of the agreement between predicted versus observed outcomes but can give a false sense of security because it is a summary metric, that may overlook subgroups for which a model is poorly calibrated. As a result, it is critical to also be provided with a graphical assessment of calibration using a calibration plot, which compares values of predicted probability of outcome versus true probability (Figure 1). Graphically, a well-calibrated model/risk score will have predicted and true probabilities that nearly overlap (ie, the model/score predicts the outcome that is close to reality, as shown in Figure 1), while a model with a lower Brier score has better calibration.[27]

The techniques to measure calibration are broadly similar for binary and time-to-event outcomes, in that they both use Brier scores and calibration plots. However, for the Brier score, one can calculate time-to-event Brier scores, which account for the time to an event, but assess the calibration at different time points. In this way, the calibration can vary as a function of follow-up times. Similarly, for calibration plots, one evaluates the observed versus predicted survival probability (ie, time-to-event) and can evaluate it for all the follow-ups, as well as at discrete time points (eg, calibration plots at 1, 3, and 5 y).

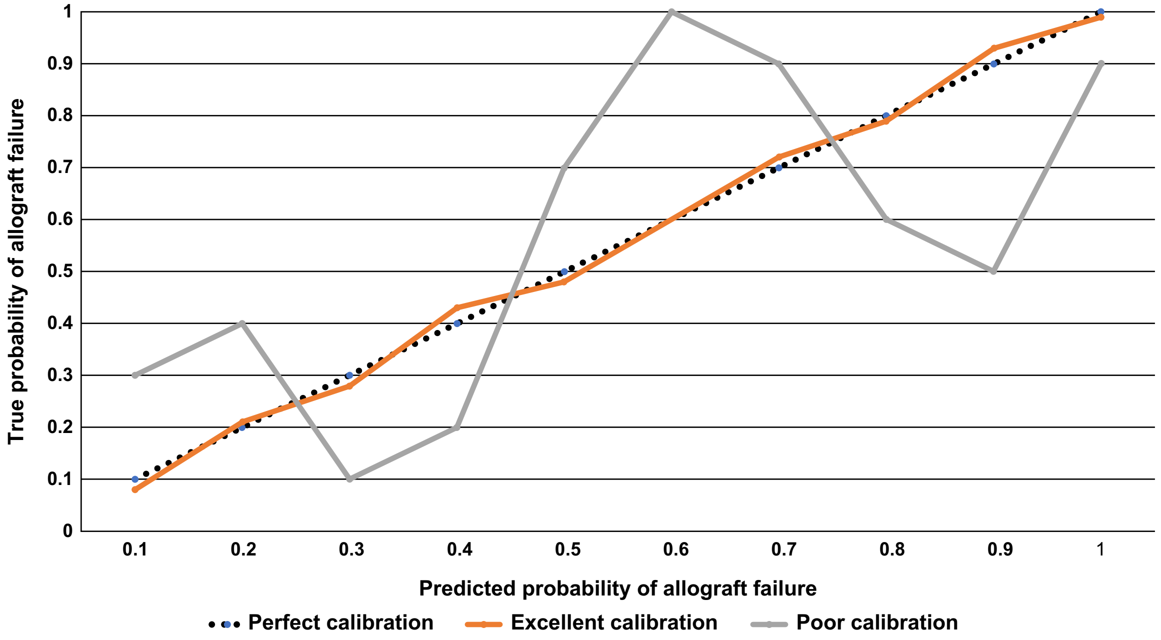

FIGURE 1 Example of a calibration plot of predicted versus observed survival.

APPLYING RISK MODELS TO CLINICAL CASES IN LIVER TRANSPLANTATION

The ideal scenario is a risk model that has high discrimination and calibration. However, in organ transplantation, many risk scores do not have excellent performance metrics. In the setting of our risk model (or prior risk models such as the liver DRI or the kidney DRI), clinicians often have to weigh the pros and cons of models with different discrimination and calibrations and determine whether applying a given model will improve clinical decision-making and/or outcomes.

To help contextualize the impact of discrimination and calibration on the clinical utility of a risk score, we present a simple hypothetical example of how a clinician can apply such models to deceased donor liver offers. We acknowledge that one must not always sacrifice discrimination for calibration (and vice versa), but for the sake of simplicity, we developed an example where we only varied one of these model performance parameters (Figures 2A–F).

We present scenarios in which a transplant surgeon named Dr. Brown has 2 deceased donor liver offers for a specific patient. We chose extremes for the sake of simplicity (ie, high discrimination/low calibration vs. low discrimination calibration/high calibration) using extremes of liver "quality" (ie, livers with good outcomes vs. bad outcomes) to contextualize how risk models could be used in clinical practice.

In example 1 (Figures 2A–C), both donor livers have good observed outcomes (ie, the actual survival). However, if Dr. Brown had to rely on a high discrimination/low calibration model, she would expect both livers to have bad outcomes (an incorrect prediction due to the low calibration) and, therefore, decline both livers. Conversely, if Dr. Brown relied on a high calibration/low discrimination model (ie, similar to the risk score we developed), she would likely be willing to accept either liver but would not be able to accurately determine which liver would have the better outcomes. In example 1, the C-statistic of 0.60, means that the liver with lower observed survival would be (incorrectly) estimated to be "better" (and accepted) 40% of the time. In this scenario, the C-statistic reflects the proportion of times that the model would correctly identify the donor that will have the better outcome (and 1 minus that equals the proportion of time it would incorrectly categorize the donor with the worse outcome as the better donor). The recipient would still have a good outcome because the surgeon correctly accepted a liver that was accurately predicted to have excellent long-term survival. In a sense, the surgeon had a low-risk choice to make, because both options were good.

In example 2 (Figures 2D–F), we present an alternative scenario where both liver grafts would have poor observed outcomes. If the surgeon relied on a high discrimination/low calibration model, they would be able to accurately assess which organ would have better survival and would transplant it, not recognizing that they were overestimating the predicted longevity of the organ. Conversely, if the surgeon employed a model with high calibration, they would decline both offers, confident in the prediction that both livers would not have good long-term outcomes.

What these scenarios also highlight is the dangers of using a well-calibrated (or poorly calibrated) model, regardless of its discrimination in the setting of organ

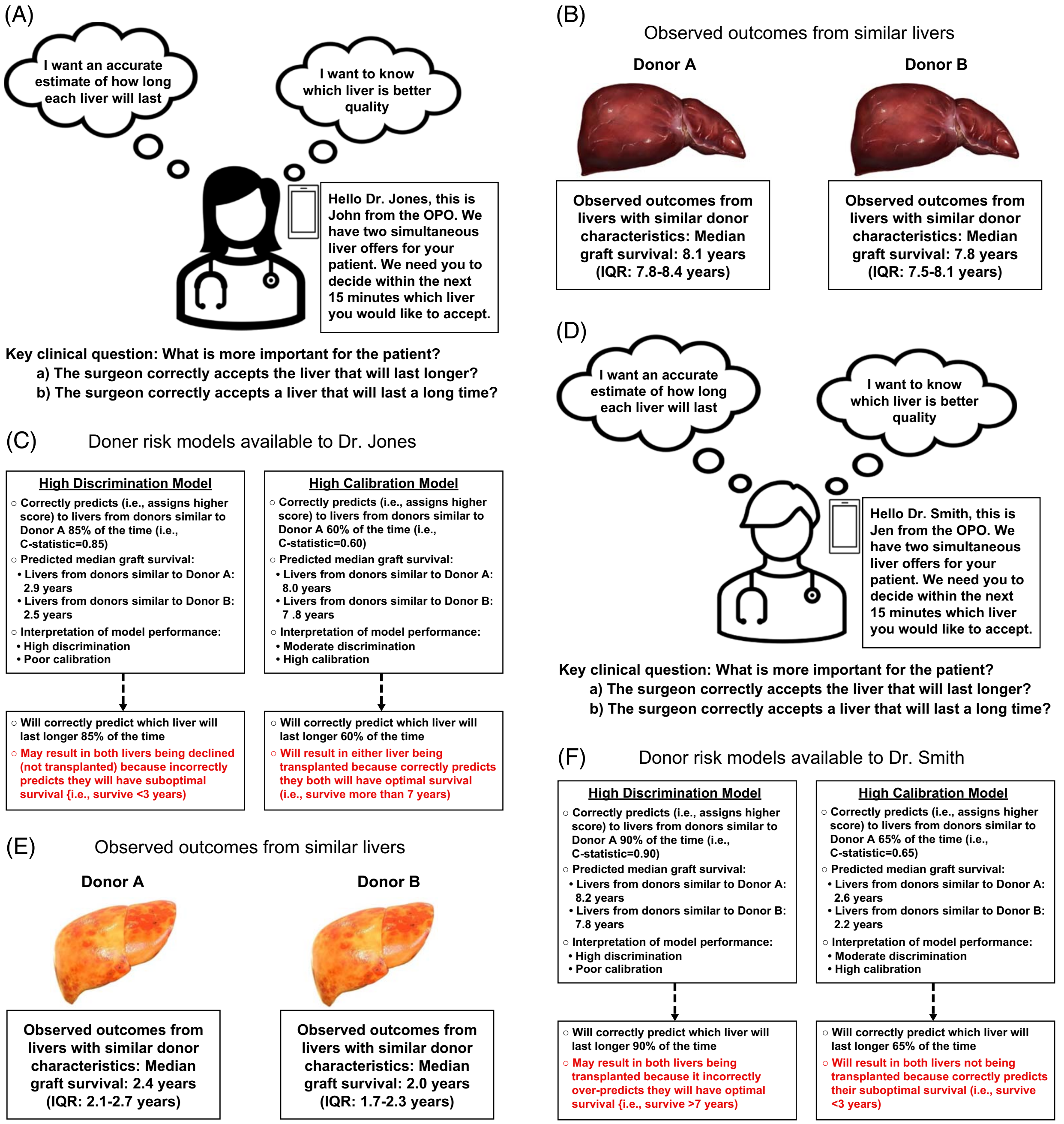
(A)
I want an accurate estimate of how long each liver will last
I want to know which liver is better quality
Hello Dr. Jones, this is John from the OPO. We have two simultaneous liver offers for your patient. We need you to decide within the next 15 minutes which liver you would like to accept.
Key clinical question: What is more important for the patient?
a) The surgeon correctly accepts the liver that will last longer?
b) The surgeon correctly accepts a liver that will last a long time?
(B)
Observed outcomes from similar livers
Donor A
Donor B
Observed outcomes from livers with similar donor characteristics: Median graft survival: 8.1 years (IQR: 7.8-8.4 years)
Observed outcomes from livers with similar donor characteristics: Median graft survival: 7.8 years (IQR: 7.5-8.1 years)
(C)
Doner risk models available to Dr. Jones
High Discrimination Model
○ Correctly predicts (i.e., assigns higher score) to livers from donors similar to Donor A 85% of the time (i.e., C-statistic=0.85)
○ Predicted median graft survival:
• Livers from donors similar to Donor A: 2.9 years
• Livers from donors similar to Donor B: 2.5 years
○ Interpretation of model performance:
• High discrimination
• Poor calibration
High Calibration Model
○ Correctly predicts (i.e., assigns higher score) to livers from donors similar to Donor A 60% of the time (i.e., C-statistic=0.60)
○ Predicted median graft survival:
• Livers from donors similar to Donor A: 8.0 years
• Livers from donors similar to Donor B: 7 .8 years
○ Interpretation of model performance:
• Moderate discrimination
• High calibration
○ Will correctly predict which liver will last longer 85% of the time
○ May result in both livers being declined (not transplanted) because incorrectly predicts they will have suboptimal survival {i.e., survive <3 years)
○ Will correctly predict which liver will last longer 60% of the time
○ Will result in either liver being transplanted because correctly predicts they both will have optimal survival (i.e., survive more than 7 years)
(D)
I want an accurate estimate of how long each liver will last
I want to know which liver is better quality
Hello Dr. Smith, this is Jen from the OPO. We have two simultaneous liver offers for your patient. We need you to decide within the next 15 minutes which liver you would like to accept.
Key clinical question: What is more important for the patient?
a) The surgeon correctly accepts the liver that will last longer?
b) The surgeon correctly accepts a liver that will last a long time?
(E)
Observed outcomes from similar livers
Donor A
Donor B
Observed outcomes from livers with similar donor characteristics: Median graft survival: 2.4 years (IQR: 2.1-2.7 years)
Observed outcomes from livers with similar donor characteristics: Median graft survival: 2.0 years (IQR: 1.7-2.3 years)
(F)
Donor risk models available to Dr. Smith
High Discrimination Model
○ Correctly predicts (i.e., assigns higher score) to livers from donors similar to Donor A 90% of the time (i.e., C-statistic=0.90)
○ Predicted median graft survival:
• Livers from donors similar to Donor A: 8.2 years
• Livers from donors similar to Donor B: 7.8 years
○ Interpretation of model performance:
• High discrimination
• Poor calibration
High Calibration Model
○ Correctly predicts (i.e., assigns higher score) to livers from donors similar to Donor A 65% of the time (i.e., C-statistic=0.65)
○ Predicted median graft survival:
• Livers from donors similar to Donor A: 2.6 years
• Livers from donors similar to Donor B: 2.2 years
○ Interpretation of model performance:
• Moderate discrimination
• High calibration
○ Will correctly predict which liver will last longer 90% of the time
○ May result in both livers being transplanted because it incorrectly over-predicts they will have optimal survival {i.e., survive >7 years)
○ Will correctly predict which liver will last longer 65% of the time
○ Will result in both livers not being transplanted because correctly predicts their suboptimal survival (i.e., survive <3 years)

FIGURE 2 Simplified schematic of how to apply an allograft risk model for livers with good versus poor observed outcomes. (A) Clinical scenario for offers of livers with good outcomes. (B) Data on observed outcomes of the 2 donor livers with good outcomes. (C) Implications of relying on a model with high discrimination/poor calibration versus moderate discrimination/high calibration for livers with good outcomes. (D) Clinical scenario for offers of livers with poor outcomes. (E) Data on observed outcomes of the 2 donor livers with poor outcomes. (F) Implications of relying on a model with high discrimination/poor calibration versus moderate discrimination/high calibration poor outcomes.

offers, because of the uncertainties about the model's ability to accurately assess the failure rate of a given organ. Our hypothetical scenario is an oversimplification of the complex process of organ offer decision-making, whereby surgeons need to also consider whether a patient is better off accepting a given offer or waiting for another offer. Those decisions need to consider how sick a patient is and their likelihood of surviving to receive another offer, as well as the probability the next offer will be better than the current offer. But our scenario does highlight one aspect of that process, whereby a well-calibrated model would allow a surgeon to feel confident that the given model they are using is accurately predicting the survival of the offered organ, and whether that meets an acceptable threshold from the perspective of the surgeon and the patient.

In a situation where one organ is “good” and one is “bad,” the impact of discrimination and calibration can

TABLE 1 Overview of comparisons of discrimination and calibration

	Discrimination	Calibration
Goal of measure	Ability to differentiate (ie, rank order) those who will versus will not develop an outcome	Measure of agreement between frequency of observed events and predicted probabilities
Question addressed	Is patient A or patient B more likely to reach the outcome?	How accurate is this model in predicting the true outcome of patient A and patient B?
Measure(s) of performance	Binary outcome: ROC Time-to-event outcome: time-dependent AUC or C-statistic	Quantitative assessment: Brier score or GOF test Qualitative assessment: calibration plots
Definition of "good" performance[23]	Moderate/acceptable: 0.60–0.79 Moderate/good: 0.70–0.79 Good/very good: 0.80–0.89 Excellent: 0.90–1.00	Quantitative: 0.0–0.24 Qualitative: Visual inspection calibration plot
Limitations	Time-to-event outcomes: censored outcomes and time dependence Rank orders but does not assess the accuracy of predicted outcomes	Brier score only global measure and is limited to subgroup analyses Calibration plot interpretation subjective

Abbreviations: AUC, area under the curve; GOF, goodness of fit; ROC, receiver operating curve.

be nuanced. If a model had high discrimination but low calibration, the model would correctly indicate the better organ but would not be able to determine the true difference between organ quality (ie, both might be predicted as "good" or "bad," or one "good" and one "bad"). Alternatively, a model with low discrimination and high calibration is rare. In this case, the differences between a "good" or "bad" organ should be clearly discriminated. However, these types of models often have poor calibration and discrimination in a small subgroup of the population, which can lead to biased/poor decision-making and discrimination/inequity in health care utilization within the subgroup.

EVALUATING OUR LIVER ALLOGRAFT RISK SCORE GIVEN THE DISCRIMINATION AND CALIBRATION

The risk score we presented in the introduction is best seen as a teaching example, and not as a clinical tool that we would propose to implement into real-world care without further refinement and validation. However, it helps to contextualize how a risk score with similar performance to ours could be applied in the scenarios described in Figure 2. If a surgeon were offered 2 similar livers that in truth would have "good" outcomes (ie, Figures 2A–C) and used our risk score, they would only correctly pick the liver with a better-observed outcome 60% of the time but would believe that it is reasonable to use either liver because the probability of a good long-term survival was high. Conversely, if the surgeon were offered 2 similar livers that would be expected to have "bad" outcomes (ie, Figures 2D–F) and used our risk score, their ability to determine which liver would last longer would be only slightly better than the flip of a coin (ie, AUC≈0.6), but can decline both offers knowing that the allograft would have lasted around 2 years. Therein lies the importance of calibration at the extremes. As a result, models with performance similar to the one we are developing could help to guide clinician decision-making at the time of an organ offer and to better match donors and recipients based on the graft's predicted outcome and the recipient's expected posttransplant survival (ie, akin to the matching of kidney donors and recipients using the kidney donor profile index [donor] and expected posttransplant survival [recipient] scores) (Table 1).

CONCLUSIONS

Risk scores that accurately predict allograft survival have the potential to maximize the good from donated organs and bolster informed consent. However, both within and outside organ transplantation, few risk scores are ever implemented in health care to change practice. This lack of implementation is due to several factors, which include the fact that many models (1) have not been externally validated, (2) have only moderate predictive ability, (3) are not built to be easy to implement, and (4) do not address the clinically relevant questions.[28–31] In the setting of donor and recipient graft survival prediction, we argue that calibration of prediction models deserves the same attention as discrimination, and in some scenarios, calibration may be more important. Although the optimal scenario is one where a risk model can determine who will have the best outcome (ie, discrimination) while correctly predicting survival in absolute terms (ie, calibration), this is not always possible. Therefore, in the setting of a model where one has to balance discrimination versus calibration, it is critical to correctly frame the clinical question and the manner in which the risk score would be applied in practice. Only when this is done can one decide the optimal model for the specific clinical question. The Transparent Reporting for Individual Prognosis or Diagnosis initiative includes guidance on how to report multivariable prediction models, including addressing issues related to model calibration.[32] However, because the Transparent Reporting for Individual Prognosis or Diagnosis checklist for model development does not specifically call out calibration, we would urge biomedical journals publishing manuscripts involving predictive models to require all manuscripts to include a discussion of model discrimination and calibration, similar to requirements related to IRB approval.

FUNDING INFORMATION

This work was supported in part by the Health Resources and Services Administration contract HHSH250-2019-00001C. Research support was provided by NIDDK R01 DK123041 (David Goldberg, Hemant Ishwaran, Peter Abt, Michael Harhay, Sarah J. Ratcliffe, Peter P. Reese). The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

CONFLICTS OF INTEREST

Michael Harhay advises the American Thoracic Society. Emily Vail received grants from eGenesis. Peter P. Reese received grants from Gilead and eGenesis. He is employed by the National Kidney Foundation. The remaining authors have no conflicts to report.

REFERENCES

1. Cantu E, Diamond J, Ganjoo N, Nottigham A, Ramon CV, McCurry M, et al. Scoring donor lungs for graft failure risk: The Lung Donor Risk Index (LDRI). Am J Transplant. 2024;24:839–49.
2. Gadsden MM, Goldberg DS. Hepatic dysfunction in deceased donors in the age of the opioid epidemic. Transplantation. 2018; 102:e403.
3. 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; 6:783–90.
4. Flores A, Asrani SK. The donor risk index: A decade of experience. Liver Transpl. 2017;23:1216–25.
5. Mataya L, Aronsohn A, Thistlethwaite JR Jr., Friedman Ross L. Decision making in liver transplantation—Limited application of the liver donor risk index. Liver Transpl. 2014;20:831–7.
6. Asrani SK, Saracino G, Wall A, Trotter JF, Testa G, Hernaez R, et al. Assessment of donor quality and risk of graft failure after liver transplantation: The ID(2) EAL score. Am J Transplant. 2022;22:2921–30.
7. Ye J, Hua M, Zhu F. Machine learning algorithms are superior to conventional regression models in predicting risk stratification of COVID-19 patients. Risk Manag Healthc Policy. 2021;14:3159–66.
8. Sulaiman S, Kawsara A, El Sabbagh A, Mahayni AA, Gulati R, Rihal CS, et al. Machine learning vs. conventional methods for prediction of 30-day readmission following percutaneous mitral edge-to-edge repair. Cardiovasc Revasc Med. 2023;56:18–24.
9. Luo C, Zhu Y, Zhu Z, Li R, Chen G, Wang Z. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure. J Transl Med. 2022;20:136.
10. Goldberg D, Mantero A, Kaplan D, Delgado C, John B, Nuchovich N, et al. Accurate long-term prediction of death for patients with cirrhosis. Hepatology. 2022;76:700–11.
11. Volk ML, Biggins SW, Huang MA, Argo CK, Fontana RJ, Anspach RR. Decision making in liver transplant selection committees: A multicenter study. Ann Intern Med. 2011;155: 503–8.
12. Volk ML, Roney M, Merion RM. Systematic bias in surgeons' predictions of the donor-specific risk of liver transplant graft failure. Liver Transpl. 2013;19:987–90.
13. Schnier KE, Cox JC, McIntyre C, Ruhil R, Sadiraj V, Turgeon N. Transplantation at the nexus of behavioral economics and health care delivery. Am J Transplant. 2013;13:31–5.
14. Ge J, Wood N, Segev DL, Lai JC, Gentry S. Implementing a height-based rule for the allocation of pediatric donor livers to adults: A Liver Simulated Allocation Model study. Liver Transpl. 2021;27:1058–60.
15. Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. 2007;45:797–805.
16. Biggins SW, Kim WR, Terrault NA, Saab S, Balan V, Schiano T, et al. Evidence-based incorporation of serum sodium concentration into MELD. Gastroenterology. 2006;130:1652–60.
17. Vitale A, Volk ML, Senzolo M, Frigo AC, Cillo U. Estimation of liver transplant related survival benefit: The devil is in the details. Gastroenterology. 2016;150:534–5.
18. Volk M, Marrero JA. Liver transplantation for hepatocellular carcinoma: Who benefits and who is harmed? Gastroenterology. 2008;134:1612–4.
19. Doshi MD, Schaubel DE, Xu Y, Rao PS, Sung RS. Clinical utility in adopting race-free kidney donor risk index. Transplant Direct. 2022;8:e1343.
20. Rao PS, Schaubel DE, Guidinger MK, Andreoni KA, Wolfe RA, Merion RM, et al. A comprehensive risk quantification score for deceased donor kidneys: The kidney donor risk index. Transplantation. 2009;88:231–6.
21. Zhong Y, Schaubel DE, Kalbfleisch JD, Ashby VB, Rao PS, Sung RS. Reevaluation of the Kidney Donor Risk Index. Transplantation. 2019;103:1714–21.
22. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. JAMA. 2017;318:1377–84.
23. Hartman N, Kim S, He K, Kalbfleisch JD. Pitfalls of the concordance index for survival outcomes. Stat Medicine. 2023; 42:2179–90.
24. Pölsterl S. Scikit-survival: A library for time-to-event analysis built on top of scikit-learn. J Mach Learn Res. 2020;21:1–6.
25. Goldberg D, Zarnegarnia Y. Prediction of long-term survival among patients with cirrhosis using time-varying models. Hepatol Commun. 2023;7:e0185.
26. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005;61:92–105.
27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
28. Dekker FW, Ramspek CL, van Diepen M. Con: Most clinical risk scores are useless. Nephrol Dial Transplant. 2017;32:752–5.
29. Kwan JL, Lo L, Ferguson J, Goldberg H, Diaz-Martinez JP, Tomlinson G, et al. Computerised clinical decision support systems and absolute improvements in care: Meta-analysis of controlled clinical trials. BMJ. 2020;370:m3216.

30. Shah RU, Bress AP, Vickers AJ. Do prediction models do more harm than good? Circ Cardiovasc Qual Outcomes. 2022;15: e008667.

31. Harris AHS. Three critical questions that should be asked before using prediction models for clinical decision support. JAMA Network Open. 2019;2:e196661.

32. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. BMC Med. 2015;13:1.

How to cite this article: Goldberg D, Ishwaran H, Potluri V, Harhay M, Vail E, Abt P, et al. Evaluating allograft risk models in organ transplantation: Understanding and balancing model discrimination and calibration. Liver Transpl. 2025;31:956–963. https://doi.org/10.1097/LVT.0000000000000575