ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Themed Section: Artificial Intelligence

A Framework for Using Real-World Data and Health Outcomes Modeling to Evaluate Machine Learning–Based Risk Prediction Models

Patricia J. Rodriguez, PhD, David L. Veenstra, PhD, Patrick J. Heagerty, PhD, Christopher H. Goss, MD, Kathleen J. Ramos, MD, Aasthaa Bansal, PhD

ABSTRACT

Objectives: We propose a framework of health outcomes modeling with dynamic decision making and real-world data (RWD) to evaluate the potential utility of novel risk prediction models in clinical practice. Lung transplant (LTx) referral decisions in cystic fibrosis offer a complex case study.

Methods: We used longitudinal RWD for a cohort of adults (n = 4247) from the Cystic Fibrosis Foundation Patient Registry to compare outcomes of an LTx referral policy based on machine learning (ML) mortality risk predictions to referral based on (1) forced expiratory volume in 1 second (FEV_1) alone and (2) heterogenous usual care (UC). We then developed a patient-level simulation model to project number of patients referred for LTx and 5-year survival, accounting for transplant availability, organ allocation policy, and heterogenous treatment effects.

Results: Only 12% of patients (95% confidence interval 11%-13%) were referred for LTx over 5 years under UC, compared with 19% (18%-20%) under FEV_1 and 20% (19%-22%) under ML. Of 309 patients who died before LTx referral under UC, 31% (27%-36%) would have been referred under FEV_1 and 40% (35%-45%) would have been referred under ML. Given a fixed supply of organs, differences in referral time did not lead to significant differences in transplants, pretransplant or post-transplant deaths, or overall survival in 5 years.

Conclusions: Health outcomes modeling with RWD may help to identify novel ML risk prediction models with high potential real-world clinical utility and rule out further investment in models that are unlikely to offer meaningful real-world benefits.

Keywords: machine learning, microsimulation, real-world data.

VALUE HEALTH. 2022; 25(3):350–358

Introduction

Despite the rapid development of new risk prediction models (RPMs) using machine learning (ML) methodologies, few RPMs have been implemented for use in clinical practice.[1–4] A recent systematic review found only 51 applications of artificial intelligence in real-world clinical practice in more than 15 000 ML or artificial intelligence publications identified.[5] One reason for the gap between development and implementation is a lack of evidence on the real-world clinical utility offered by new RPMs: the expected change in downstream patient outcomes when used for decision making in clinical practice.[6–11] Although commonly reported improvements in predictive accuracy are necessary for consideration of novel RPMs, accuracy alone is insufficient for assessing real-world clinical utility because it does not capture the complex clinical context in which the model would be used. Additional consideration is needed for real-world factors that affect RPM utility in clinical practice, including (1) the true, heterogenous current process for making decisions and (2) the downstream patient outcomes associated with clinical decisions.

Novel RPMs are typically compared with a reference model—an existing RPM, biomarker, or clinical guidelines.[12,13] Nevertheless, real-world clinical decision making is heterogenous and often deviates from the reference model, with different clinicians weighing different factors in decisions, including various pieces of evidence, historical experience, and preferences.[4,11,14,15] Clinicians may also have additional pieces of information, such as expensive tests available for a subset of patients and subjective clinical impressions. In such cases, it remains unclear whether an RPM that outperforms a reference model would also outperform usual care (UC).

Furthermore, studies rarely relate changes in the discrimination and calibration properties of an RPM to changes in downstream patient outcomes.[16,17] Some approaches have been proposed, such as considering the balance of false positives and false negatives at a given threshold.[18,19] Nevertheless, in many cases, treatment effects are heterogenous, with not all true positives experiencing the same benefits of treatment and not all false negatives and false positives experiencing the same harms of misclassification. In such cases, discrimination will fail to capture

the expected patient impacts, for example, that a model that better identifies cases with large treatment benefits offers higher clinical utility than a model that better identifies cases with smaller treatment benefits.

Our objective was to compare the expected real-world patient health outcomes (survival) of using a new RPM for decision making in clinical practice to those expected under (1) UC and (2) a reference model. We propose a health outcomes modeling framework that relies on real-world data (RWD) to estimate changes in real-world clinical decisions and linked downstream outcomes when an RPM is used in clinical practice. We leverage RWD to mimic the clinical context in which a novel RPM model would be used, providing a clearer picture of consequences in clinical practice.

We selected lung transplant (LTx) referral decisions in cystic fibrosis (CF) as a case study for 3 primary reasons. First, the standard predictor of short-term mortality in CF, forced expiratory volume in 1 second (FEV_1), has low positive predictive value,[20-25] and we previously developed an ML-based RPM with better discrimination and calibration (Rodriguez et al, unpublished data, 2021). Second, UC for referral decision making is heterogenous, so performance improvements relative to FEV_1 may not be indicative of performance relative to UC.[3,4,26] Third, the relationship between clinical decisions and patient outcomes is complex given the limited transplant availability[27] and heterogenous benefits,[28-30] so additional consideration of downstream outcomes is needed.

Methods

Framework

We propose a general framework to evaluate the expected utility of novel RPMs in real-world clinical practice, with respect to patient outcomes. This framework has 3 tenets:

1. The use of RWD to mimic patterns of real-world clinical practice. Real-world care and decision-making patterns may deviate substantially from expectations (ie, guidelines), which can affect expected outcomes of RPM model use. Leveraging RWD allows the RPM evaluation environment to mirror to the real-world context in which the RPM would be used, including patterns of utilization and decision-making practices under UC.
2. Dynamic decision making to reflect intended use in clinical practice. Rather than assessing risk model use at a single timepoint, such as baseline, the risk model is applied at each encounter over time, using the most recently collected values.
3. Health outcomes modeling to evaluate the downstream patient outcomes resulting from a clinical decision. We expand outcomes considered in RPM evaluation to include the clinical decisions resulting from RPM use and subsequent treatment outcomes, accounting for heterogeneity in treatment effects.

Although our case study considers an ML-based RPM, the framework is equally applicable to RPMs developed using more traditional methods, such as logistic regression or biomarkers used for decision making. Similarly, the framework could be used to compare alternative thresholds for decision making.

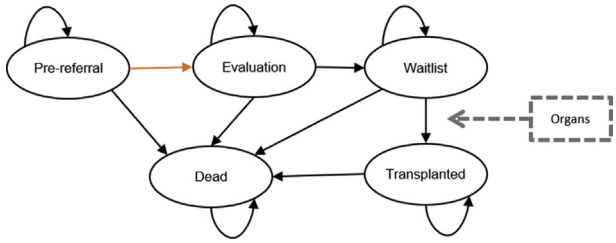
Case Study

Data

We used RWD from the CF Foundation Patient Registry (CFFPR), which collects longitudinal, observational data for all US patients seen at CF Foundation–accredited care centers who consent to participate.[31] Data on patient diagnoses, demographics, encounters, care episodes, and annual visits are entered electronically by CF care center staff using information from electronic medical records and patient forms.[31] The CFFPR covers approximately 80% of the US CF population and includes 95% of clinic visits and 90% of hospitalizations for participating patients.[31] Our cohort included CFFPR adults (≥18 years old) who had not undergone LTx by January 1, 2012, and had at least 1 encounter in both 2011 and 2012 (n = 10 615). Our cohort was followed until December 31, 2016. We previously split cohort data into training (60%) and validation (40%) sets to develop and evaluate the ML model. The 40% validation set (n = 4247) was used in this patient-level simulation.

CFFPR data were linked to United Network for Organ Sharing (UNOS) data, which contains additional waitlist, transplant, and post-transplant information for patients listed for LTx. UNOS data also contain information on donated organs. The data linkage was performed at University of Washington in collaboration with University of Toronto.[32,33] This study was approved by the University of Washington Institutional Review Board (study #2270), St Michael's Hospital, Toronto, Canada (research ethics board #14-148), and the Seattle Children's Research Institute (study #PIROSTUDY15294).

Figure 1. Microsimulation model with 5 mutually exclusive health states: prereferral, evaluation, waitlist, transplanted, and dead. Patients waitlisted before model start begin in the waitlist state; all other patients begin in prereferral. A patient moves from prereferral to evaluation at the time of referral (orange arrow), which varies among policies.



Patient-Level Simulation Model Structure

We developed a patient-level simulation model with 5 mutually exclusive health states: prereferral, evaluation, waitlist, transplanted, and dead (Fig. 1). Patients began in the state corresponding to their status on January 1, 2012: prereferral, evaluation, or waitlist. Patients transitioned from prereferral to evaluation at their time referral, which varied among ML, FEV_1, and UC policies. Evaluation, modeled as a tunnel state, represents the time between referral and waitlisting when evaluation for LTx occurs. Surviving patients transitioned to the waitlist, where they remained until they were matched with an organ for LTx or died before transplant. Transplanted patients remained in the transplanted state until post-transplant death or model end. Transitions between states were determined by individual-specific transition probabilities, described below, that rely on RWD. We used a cycle time of 1 day and a time horizon of 5 years. Modeling was conducted in R.[34]

Interventions: Referral Policies

We considered 3 potential policies for referring patients for LTx: (1) ML model based, (2) reference model (FEV_1) based, and (3) UC. The ML policy used individual risk predictions from a previously developed ML model for 2-year mortality. The ML

model used super learner, an ensemble ML approach that optimally combined multiple underlying models (lasso, elastic net, ridge, XGboost, random forest, and support vector machine).[35,36] ML had higher discrimination at baseline (area under the receiver operating curve 0.914 [95% confidence interval 0.898-0.929]) and over time and better calibration than FEV_1 (baseline area under the receiver operating curve 0.876 [0.858-0.895]). Additional detail is provided in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360.

The ML-based referral policy was intended to reflect the ML model's likely use in clinical practice. We assumed referral would occur at the first clinic visit where a patient's predicted ML risk exceeded the threshold corresponding to 95% model specificity at baseline, which matches the specificity of the common FEV_1 <30% criteria.[21] However, any alternative decision rule could be considered, including different thresholds or more complex rules, such as multiple visits when criteria are met. For the FEV_1-based policy, referral occurred at the first clinic visit with a stable FEV_1 <30% predicted based on the Global Lung Initiative equations for % predicted.[37] For UC, the referral time was determined using RWD. We describe referral in detail in section "Referral (prereferral → evaluation)" mentioned below.

Simulation Population

Our data contain correlated, longitudinal information on patients' visit patterns, lung function, other health factors, predicted ML risk, and pre-LTx survival. Simulating a data set that preserves the complex underlying relationships between these variables would be extremely challenging. Rather than imposing strong and potentially incorrect distributional assumptions to simulate correlated longitudinal data, we use the approach of plasmode simulation, where resampled populations ("plasmodes") are drawn with replacement from observed data.[38,39] In this approach, unmodified RWD for the resampled population are combined with modeling to simulate unknown elements. In our application, we drew 1 resampled population with replacement from observed cohort data on each of 1000 simulation model runs. Resampled patients retained their true, observed covariate history up to the time of transplant, including visit history, ML risk scores, pulmonary function, and pretransplant survival. Outcomes of transplant timing and post-transplant survival were then simulated, using models described below. We also used modeling to synthetically extend patients' pretransplant covariate history in cases where their actual time of transplant occurred earlier than it would have under an alternative policy (ie, when pretransplant covariate history is censored by transplant).[40] We summarize resampled versus modeled elements in Appendix Table S1.2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360.

In general, plasmode simulation is a flexible approach that is useful for preserving the underlying relationships among the potentially hundreds of variables in RWD.[38] Nevertheless, it is also more computationally intensive than a completely simulated population.

Outcomes

We compared referral policies on each of the following outcomes: 5-year overall survival (the sum of time spent in all non-death states), number of 5-year pretransplant deaths, and number of post-transplant deaths. Overall, 5-year survival is intended to capture the population-level impact of using an alternative policy. Because deaths are relatively rare and the impact to overall survival may be small, we separately evaluate 5-year deaths.

State Transitions

Referral (prereferral → evaluation)

Model-based policies. Patients' dynamically updated ML risk predictions and absolute contraindications to transplant (*Mycobacterium abscessus* and *Burkholderia cenocepacia*) were obtained at each of their pretransplant clinic visits from 2012 through 2016.[41] ML-based referral occurred at the first clinic visit where predicted risk exceeded a fixed threshold corresponding to 95% model specificity and no absolute contraindications to LTx were present. An example of ML and FEV_1 referral under each model for 1 patient is provided in Figure 2.

To reflect guidelines, FEV_1-based referral occurred at the first clinic visit where stable FEV_1 was <30% and no absolute contraindications were present. FEV_1 was considered stable when no pulmonary exacerbation was documented at the same visit.

For patients who actually received LTx, observed pretransplant visit history, ML risk predictions, and FEV_1 were censored at their observed time of transplant, L_i. Under an alternative policy, referral may not have occurred by time L_i, when pretransplant history was censored. In such cases, we synthetically extended ML risk and FEV_1 trajectories beyond L_i.[40] We first generated synthetic visit times beyond L_i (ie, referral opportunities), assuming that clinic visits would continue at the same frequency observed in the previous 12 months. We then estimated ML risk and FEV_1 at each synthetic visit using separate linear mixed effects models. Additional detail is given in the Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360. We separately accounted for pretransplant deaths (ie, that an individual may not survive until the synthetic visit) and truncated synthetic visits at the time of pretransplant death (see section "Pretransplant and post-transplant survival").

UC. Exact referral dates observed under UC are not recorded, but categorical transplant status ("not pertinent," "accepted, on the waitlist," "evaluated, final decision pending," "evaluated, rejected," and "had transplantation") is recorded annually in the CFFPR. We used the first year with a status other than "not pertinent" as the UC referral year for each patient and then defined a subset of visits where referral could have occurred: clinic visits in the referral year and before the listing date. On each simulation run, we randomly selected one of these visits as the patient's UC referral date. To test the validity of this assumption, we compared the resulting simulated UC listing time with the observed listing time.

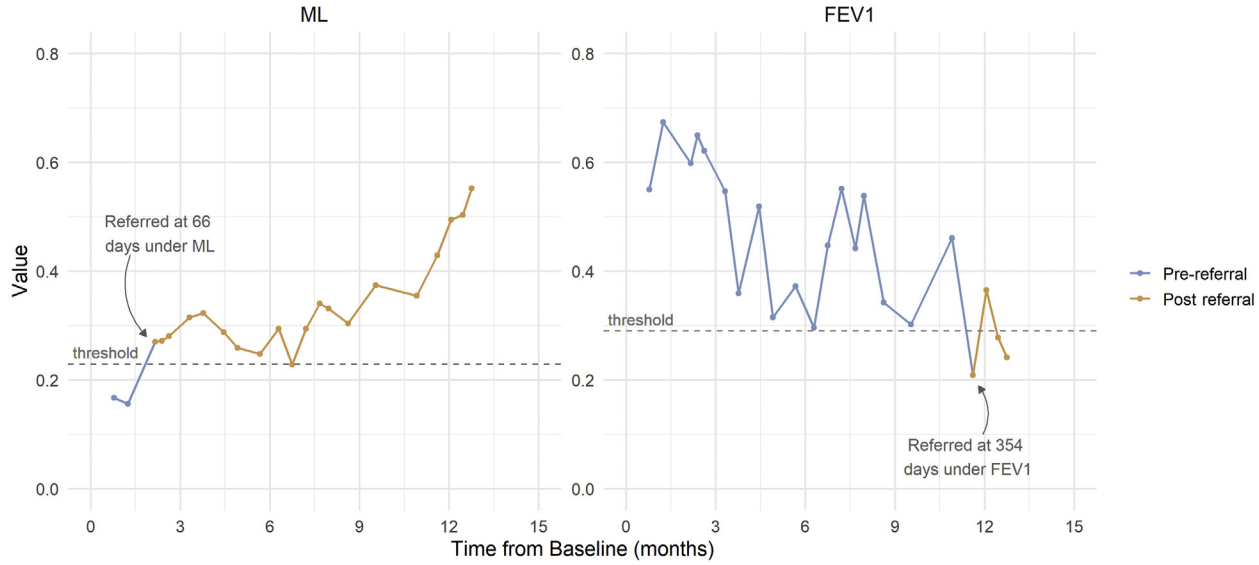
Listing (evaluation → waitlist)

After referral for transplant, patients undergo a rigorous evaluation at an LTx center to assess whether they are suitable candidates for transplant, including evaluation of their health, medical adherence, emotional wellbeing, social support, and finances.[27] Because evaluation times are not captured in our data, we simulated evaluation times to approximate available estimates[26,42] by sampling from a truncated normal distribution (mean 4.5 months, SD 4 months, minimum 3 weeks). Patients' evaluation time was held constant between policies for each simulation run, but varied between simulation runs. Additional detail is given in Appendix 1, section 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360.

Organ allocation (waitlist → transplanted)

We simulated population-level organ allocation to reflect current US policy, whereby new organs are allocated to the highest priority, compatible patient on the waitlist. The allocation process is a deterministic function of 3 components: patients on

Figure 2. Patient trajectory and referral example. Risk of 2-year mortality from the ML model and FEV_1 % predicted for an example patient at each clinic visit. For the ML model, a patient is referred at the first visit where risk exceeds the threshold, denoted by a change in line color. For FEV_1, referral occurs at the first visit where FEV_1 is lower than 30%, denoted by a change in line color.



FEV_1 indicates forced expiratory volume in 1 second; ML, machine learning.

the waiting list (described above), organs available for transplantation, and the policy for matching organs to patients.[40]

Organ flow. We relied on historical organ data from UNOS to define a flow of organs available for transplantation. On each simulation run, we sampled from the average number of organs available annually and their characteristics (ABO, height) using organs matched to patients in our cohort from 2012 to 2016. Organ dates of availability were sampled from a uniform distribution, where all dates were equally likely. We also conducted an expanded organ supply scenario analysis with twice as many organs available annually (Appendix 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360).

Organ matching. Lung allocation policy in the US prioritizes patients based on lung allocation score (LAS), a measure of expected mortality with and without transplant.[43] The LAS aims to identify patients with both an urgent need and an expected survival benefit of transplant. LAS is calculated daily to prioritize patients on the waiting list. Observed LAS measures for waitlisted patients were available in UNOS data for each active day on the waiting list. However, alternative policies sometimes resulted in earlier waitlisting or the waitlisting of patients not listed under UC, such that LAS values were not available for all patients at all necessary time points. Therefore, we imputed LAS at all timepoints for all patients listed under any policy using a linear mixed effects model. We relied on LAS components that are measured in the CFFPR and thus available for all patients regardless of listing status (age, forced vital capacity, body mass index, diabetes). Large changes in LAS are frequently observed in the days or weeks preceding death or LTx, as patients experience intensive care unit admission or the need for mechanical ventilation. To capture such changes in LAS without access to these variables, we included a fixed and random effect indicator for whether the patient experienced death or transplant in the next 30 days. Additional detail is given in the Appendix 1, section 3.1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360.

Organ-donor compatibility was determined by blood type (ABO) and body size (height). When unavailable, we imputed patient ABO using the empirical distribution of ABO in each simulation run. Although donors and recipients should have similar height (a proxy for lung capacity), no fixed thresholds exist for acceptable donor-recipient height differences.[44] We used the 2.5th and 97.5th percentiles of the historically observed distribution of donor-recipient height difference on each simulation run as bounds for height compatibility.

We assumed no organ decline, no relisting, and all bilateral transplants. We did not account for geographical regions of organ allocation.

Pre-transplant and post-transplant survival

Patients in the prereferral, evaluation, and waitlist states were at risk of pretransplant death. For patients who were never transplanted, complete pre-LTx survival was observed. In our simulation, patients observed to die before LTx retained their observed pretransplant time of death, $(T_i|LTx = 0)$, unless LTx occurred first. Similarly, patients who survived for the full 5-year period retained their observed pretransplant survival, $(T_i|LTx = 0) > 5$ years, unless LTx occurred first.

Among patients with observed LTx, pretransplant survival was censored at the observed time of transplant, L_i. In such cases, $(T_i|LTx = 0)$ was unknown, but greater than L_i. We relied on a potential outcomes model with time-varying transplant exposure to estimate survival in the absence of transplant. Under this model, we assumed that each patient has 2 potential outcomes at any time: (1) survival without transplant at time t and (2) survival with transplant at time t. Only one outcome can be realized for each patient, but information from patients with the same likelihood of treatment at time t can inform the counterfactual outcome. Because transplant is allocated using LAS, we assumed transplant assignment was random among waitlist patients with the same LAS.[28] That is, we assumed 2 patients with the same LAS had the same propensity for treatment.

Table 1. Patient characteristics at time of referral and transplant, by policy.

Characteristic	ML	FEV_1	UC
Characteristics at time of referral			
Patients referred (n)	851 (797-904)	799 (746-851)	518 (477-560)
Age	32.9 (32.1-33.6)	33 (32.3-33.8)	33.4 (32.5-34.4)
FEV_1 % predicted	31.5 (30.9-32.2)	26 (25.8-26.3)	30.9 (29.8-32)
Risk of 2-year mortality	34.8% (34.0%-35.7%)	27.8% (26.5%-29.1%)	30.6% (29.1%-32.2%)
Characteristics at time of LTx			
Patients transplanted (n)	294 (241-345)	292 (241-340)	287 (241-325)
Age	35.9 (34.3-37.5)	35.6 (33.9-37.2)	34.1 (32.8-35.6)
FEV_1 % predicted	29.2 (26-32.6)	28.4 (25.2-31.7)	29.6 (27.4-32.3)
LAS	51.6 (47.3-57.4)	50.5 (46.4-56.3)	47.9 (44.7-51.9)

Note. Mean (95% CI) at the time of referral and transplant by policy.
CI indicates confidence interval; FEV_1, forced expiratory volume in 1 second; LAS, lung allocation score; ML, machine learning; UC, usual care.

Modeling survival conditional on transplant status using observed data. We estimated the impact of time-dependent transplant on survival using an exponential survival model, with time-varying covariates for LAS and LTx status. Higher LAS is intended to indicate a greater benefit of treatment. We adjusted for sex, age at waitlisting, and body mass index at waitlisting. The model was estimated on waitlisted patients in each simulation run, with time measured as time to death since waitlist entry. We provide additional detail in the Appendix 1, section 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360.

Estimating expected pretransplant and post-transplant survival for simulation. For patients with observed transplant whose pretransplant survival was censored at L_i, we obtained expected time of death in the absence of transplant, conditional on survival and history up to L_i, $(T_i|LTx = 0,\ T_i > L_i, X_t)$. At $t = L_i$, we use the inverse sampling method to obtain $T_i|LTx = 0$:

$$T_i(t) = \lambda^{-1}(-log(U_i) * exp(-\beta * X_i(t)))$$

where $U \sim Uni(0,1)$, β is a vector of coefficients, and $X_i(t)$ is a vector of covariate values for individual i at time t, with the transplant indicator set to 0.

When considering post-transplant survival, a patient's simulated time of transplant under each policy, $L^*_{i,p}$, may vary across policies and/or simulation runs. For example, a patient could be transplanted at $t = 100$ days under ML and $t = 150$ days under FEV_1. If their clinical status declined substantially from 100 to 150 days (eg, they were admitted to the intensive care unit with respiratory failure requiring mechanical ventilation), their expected post-transplant survival may be lower when transplant occurs at 150 versus 100 days. To obtain post-transplant survival at each potential transplant time, $L^*_{i,p}$, we again use the inverse sampling method, this time considering transplant at each $t = L^*_{i,p}$ and setting the transplant indicator at t to 1.

Results

Unless otherwise noted, results are presented as estimate (95% confidence interval).

Validation

Among patients listed for LTx, the simulated UC listing date was a median of 9 days earlier than the observed UC listing date (interquartile range 102 days earlier to 157 days later). In observed data, 466 patients died without transplant, compared with 458 (400-520) in the simulated UC. A total of 309 transplants and 65 post-transplant deaths were observed within the 5-year period, compared with 287 (244-327) transplants and 41 (24-59) post-transplant deaths in our simulated UC.

Clinical Decisions

Most patients remained too healthy for referral in the 5-year period, regardless of policy. Only 12.4% of patients (11.4%-13.4%) were referred for LTx under UC (Table 1). By comparison, a uniform application of FEV_1 resulted in significantly more patients referred, 19.2% (18.0%-20.4%). Referral rates were somewhat higher for ML, 20.4% (19.1%-21.6%). On average, ML resulted in earlier referral than UC, when patients were relatively healthier. Characteristics at the time of referral were not significantly different (statistically or clinically), including average FEV_1 for ML, 31.5% predicted (30.9-32.2), and UC, 30.9% predicted (29.8-32) (Table 1). Among patients referred under both ML and UC, ML referral occurred 129 (82-176) days earlier, on average.

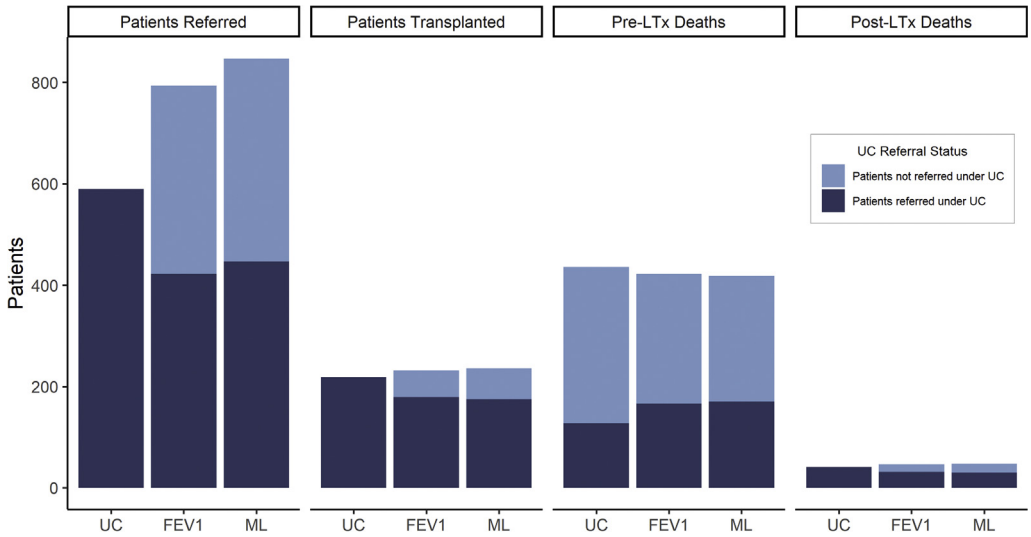
Many patients missed for referral under UC would have been referred by a policy with systematic decision making using either FEV_1 or ML (Fig. 3). Of patients who died without being referred for LTx under UC, ML would have referred 40.0% (35.3%-44.5%) and FEV_1 would have referred 31.2% (26.9%-35.6%).

Patient Outcomes

Transplantation

Despite higher referral rates, there was no difference in overall transplantation rates among policies because of real-world constraints in organ supply (Table 1). State membership over time (Fig. 4) shows that, given a fixed supply of organs available for transplant, relatively higher referral rates under both ML and FEV_1 led to increased patients on the waiting list, but no change in patients transplanted. At a population level, 0.39 (0.30-0.44) years (of 5), on average, were spent on the waiting list under ML, compared with 0.41 (0.36-0.45) under FEV_1 and 0.23 (0.20-0.27) under UC.

Figure 3. Patient referral and outcome, by UC referral status. Average number of referrals, transplantations, and pre-LTx deaths and post-LTx deaths in 5 years, by referral status under UC.

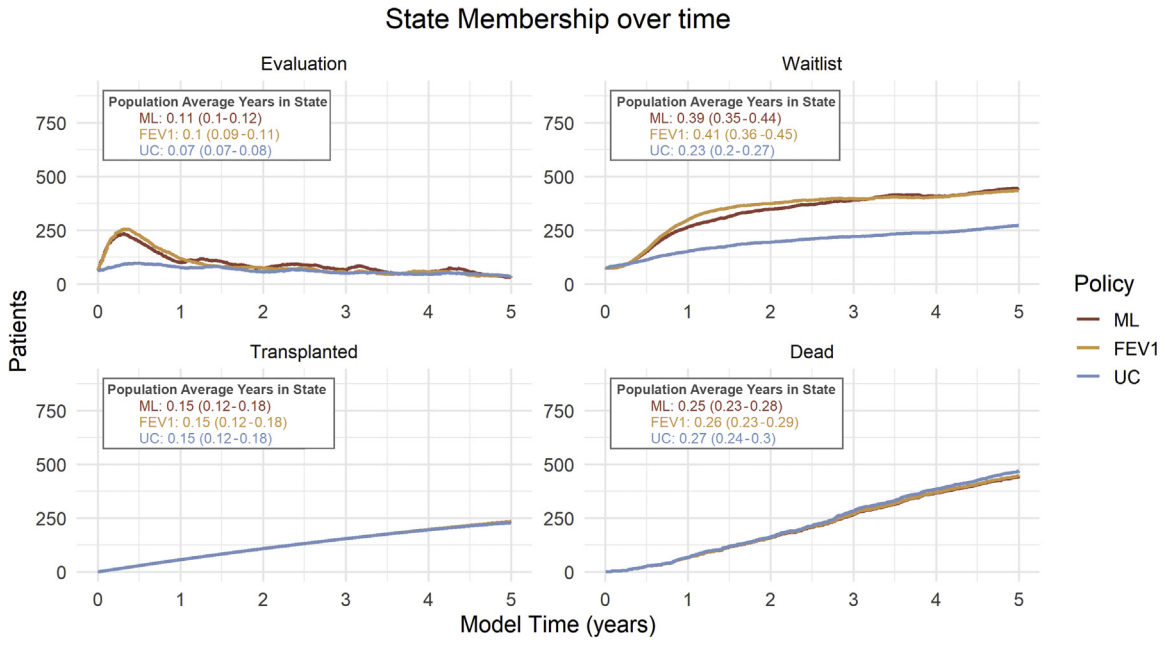


FEV_1 indicates forced expiratory volume in 1 second; LTx, lung transplant; ML, machine learning; UC, usual care.

Patient characteristics at the time of LTx were similar among policies (Table 1). Although confidence intervals overlapped, patients transplanted under ML were slightly older and had slightly higher LAS at the time of transplant than UC. As a measure, higher LAS is intended to indicate a higher expected short-term benefit of LTx.

Although characteristics at the time of transplant were similar overall, the specific patients who received transplant and experienced pretransplant death differed among policies (Fig. 3). Under UC, 309 pretransplant deaths (277-341) occurred among patients who were never referred for LTx. Approximately 20.1% of these pretransplant deaths were averted under ML because patients were referred and transplanted. However, this was offset by fewer transplants and more pretransplant deaths among those who received transplant under UC (Fig. 3).

Figure 4. State membership over time, by policy. The average number of patients in each state except prereferral for the 5-year time horizon. Population average years spent in each state (95% CI) is shown.



CI indicates confidence interval; FEV_1, forced expiratory volume in 1 second; ML, machine learning; UC, usual care.

Table 2. Expected outcomes by policy.

Policy	Model AUC at baseline*	Pretransplant deaths	Post-transplant deaths	Overall 5-year survival
ML	0.914 (0.898-0.929)	383 (332-436)	47 (29-69)	4.75 (4.72-4.77)
FEV_1	0.876 (0.858-0.895)	389 (339-442)	46 (29-69)	4.74 (4.71-4.77)
UC	—	411 (367-459)	41 (24-59)	4.73 (4.70-4.76)

AUC, area under the receiver operating curve; FEV_1, forced expiratory volume in 1 second; LAS, lung allocation score; ML, machine learning; UC, usual care.
*Model AUC at baseline was previously measured in model assessment (Rodriguez et al, unpublished data, 2021).

Survival

At a population level, these changes resulted in no significant differences in overall 5-year survival, pretransplant deaths, or post-transplant deaths (Table 2). Overall, 5-year survival was approximately 4.7 years under all polices.

Expanded organ availability scenario

In a scenario with twice as many organs available, 441 transplants (404-479) occurred under ML, compared with 412 (376-442) under FEV_1 and 367 (338-394) under UC (Appendix Table S2.1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360). Accordingly, fewer pretransplant deaths occurred under ML (281 [266-315]) than UC (359 [345-381]) (see Appendix Table S2.2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2021.11.1360). Overall, 5-year survival was slightly higher for ML (4.77 [4.75-4.79]) than UC (4.74 [4.73-4.76]), but confidence intervals overlapped.

Discussion

We demonstrated an application of patient-level simulation modeling to estimate the real-world impact of using a novel, ML-based RPM for decision making in clinical practice. We found that improvements in discrimination and calibration for ML did not yield differences in expected downstream patient outcomes when used for clinical decision making. Although ML did lead to changes in the number of patients referred and referral timing, real-world constraints on organ availability limited the extent to which referral decisions could influence transplant. Nevertheless, in a scenario of expanded organ availability, higher referral rates under ML led to more transplants and fewer pretransplant deaths.

We found a significant difference between the clinical decisions expected under FEV_1 alone, the reference model, and those observed in clinical practice. Although 799 patients (19.2%) would have been referred within the 5-year period under FEV_1, only 519 (12.4%) were actually referred in UC. Despite documented differences between clinical decision making and FEV_1,[3,4] comparisons to FEV_1 are standard for new models in CF.[21,24,25]

Our work suggests that additional comparisons to UC are needed to assess model performance. Although any new RPM must predict better than an existing RPM to add value, improvements relative to a reference model may be a poor proxy for real-world clinical utility when clinical decision making is heterogenous. RWD can be used to develop a real-world UC comparator.

Currently, the primary approach for assessing a model's real-world clinical utility is an impact evaluation study—a cluster-randomized trial, where patient outcomes are compared for groups of clinicians with access to a novel model with those using UC.[8,10,45,46] Such studies are typically undertaken as a final step before implementation.[14] In contrast, our approach uses RWD to assess the potential clinical impact in the relatively early model evaluation stage. This approach can rule out further investment in models that have limited usefulness in real-world settings. Although simulation-based evaluation does not capture the complex ways that clinicians interact with models to make decisions,[14,15] it can be used as a first step for demonstrating clinical utility before conducting RCTs. Furthermore, the approach could be extended to include costs and utility measures for cost-effectiveness analysis.

The use of health outcomes modeling to evaluate a new diagnostic test is not new.[47] Nevertheless, health outcomes modeling to evaluate new RPMs specifically remains minimal.[16] In contrast, statistical approaches for assessing clinical have gained relatively more popularity,[18,19,48] but do not generally capture the clinical context in which models would be used.

Our simulation involves several important assumptions. We considered only absolute contraindications to LTx, which may have resulted in over-referral of patients under FEV_1 and ML policies. Many contraindications are relative and vary by center, with larger and more experienced centers willing to accept more complex cases.[27,49,50] We assumed a marginal distribution of evaluation time with no rejection for listing, which may not accurately reflect patient-specific factors that influence evaluation times or rejection. Nevertheless, median simulated listing time was within 10 days of observed times, suggesting that this assumption was acceptable on average. Finally, we are unable to distinguish between clinician decision making and patient preferences using RWD. Lower rates of referral under UC may represent patient preferences for nonreferral, rather than clinician decisions not to refer patients. These complexities can be measured through impact evaluation.

More generally, RWD, including that used in our study, presents issues with missingness and infrequent data collection for some patients. We used imputation approaches to address missingness at multiple levels, including in longitudinal biomarkers and LAS values. Nevertheless, to the extent individuals with missingness are unlike those with complete data, the results of our analysis may be biased. Imputation strategies for longitudinal measures and time to event outcomes in RWD are an important topic of future research. Additionally, although heterogeneity in our study was established through use of RWD, the impact of heterogeneity on downstream outcomes is an important area for future research.

Conclusions

We used a health outcomes modeling framework with RWD to assess the potential real-world clinical utility of a novel, ML-based RPM for LTx referral decisions in CF. We found differences in clinical decisions under the RPM versus UC, but no change in downstream patient outcomes because of constraints in organs available for transplantation. The ML and FEV_1 policies effectively increased early referral compared with UC, supporting systematic approaches to referral decisions to increase access to the expertise

and treatment available at LTx centers. Efforts to expand organ availability may be necessary to reap clinical benefits from earlier referral of patients with CF. Although constraints in transplant availability are unique to the organ allocation setting, complex real-world factors that affect current clinical decisions and outcomes are common across clinical applications. Health outcomes modeling with RWD can be used to account for these complex real-world factors. When conducted as part of RPM model evaluation, this approach can identify novel, ML-based RPMs that are likely to benefit patients in real-world clinical practice and rule out further investment in RPMs with limited benefits.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2021.11.1360.

Article and Author Information

Accepted for Publication: November 16, 2021

Published Online: December 22, 2021

doi: https://doi.org/10.1016/j.jval.2021.11.1360

Author Affiliations: The Comparative Health Outcomes, Policy & Economics (CHOICE) Institute, University of Washington, Seattle, WA, USA (Rodriguez, Veenstra, Bansal); Department of Biostatistics, University of Washington, Seattle, WA, USA (Heagerty); Division of Pulmonary, Critical Care and Sleep Medicine, Department of Medicine, University of Washington, Seattle, WA, USA (Goss, Ramos); Division of Pulmonology, Department of Pediatrics, University of Washington, Seattle, WA, USA (Goss).

Correspondence: Aasthaa Bansal, PhD, The Comparative Health Outcomes, Policy & Economics (CHOICE) Institute, University of Washington, 1959 NE Pacific Str, HSB H-375, Box 357630, Seattle, WA 98195-7630, USA. Email: abansal@uw.edu
Patricia J. Rodriguez, PhD, MPH, Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, 1959 Pacific Str, HSB H-375, Box 357630, Seattle, WA 98195-7630, USA. Email: prodrig@uw.edu

Author Contributions: *Concept and design:* Rodriguez, Veenstra, Heagerty, Goss, Bansal
Acquisition of data: Goss, Ramos
Analysis and interpretation of data: Rodriguez, Heagerty, Goss, Ramos, Bansal
Drafting of the manuscript: Rodriguez, Ramos
Critical revision of the paper for important intellectual content: Rodriguez, Veenstra, Heagerty, Goss, Ramos, Bansal
Statistical analysis: Rodriguez, Heagerty, Bansal
Obtaining Funding: Heagerty
Supervision: Goss, Veenstra, Bansal

Conflict of Interest Disclosures: Drs Veenstra and Heagerty reported receiving grants from the National Institutes of Health during the conduct of the study. Dr Veenstra is an editor for *Value in Health* and had no role in the peer review process of this article. Dr Goss reported receiving grants from the Cystic Fibrosis Foundation, the European Commission, the National Institutes of Health (National Heart, Lung, and Blood Institute), and the National Institutes of Health (National Institute of Diabetes and Digestive and Kidney Diseases and National Center for Research Resources) during the conduct of the study; personal fees from Gilead Sciences and Novartis outside the submitted work; grants from National Institutes of Health and the US Food and Drug Administration outside the submitted work; speaking honoraria from Vertex Pharmaceuticals for talk at UK LEAD conference outside the submitted work; and serving as US lead in a phase 2 trial of novel therapy for cystic fibrosis for a clinical trial with Boehringer Ingelheim outside the submitted work. Dr Ramos reported receiving grants from the National Institutes of Health and the Cystic Fibrosis Foundation during the conduct of the study and grants from the CHEST Foundation in partnership with Vertex Pharmaceuticals outside the submitted work. Dr Bansal reported grants from the National Cancer Institute of the National Institutes of Health during the conduct of the study. No other disclosures were reported.

Funding/Support: This research was partially supported by the National Cancer Institute of the National Institutes of Health (NIH) (under R37-CA218413). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Role of Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: We thank the Cystic Fibrosis Foundation for providing data from the Cystic Fibrosis Foundation Patient Registry for this project. Additionally, we would like to thank the patients, care providers, and clinic coordinators at CF Centers throughout the US for their contributions to the CF Foundation Patient Registry.

REFERENCES

1. Wessler BS, Paulus J, Lundquist CM, et al. Tufts PACE clinical predictive model registry: update 1990 through 2015. *Diagn Progn Res*. 2017;1(1):20.
2. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
3. Ramos KJ, Quon BS, Psoter KJ, et al. Predictors of non-referral of patients with cystic fibrosis for lung transplant evaluation in the United States. *J Cyst Fibros*. 2016;15(2):196–203.
4. Ramos KJ, Somayaji R, Lease ED, Goss CH, Aitken ML. Cystic fibrosis physicians' perspectives on the timing of referral for lung transplant evaluation: a survey of physicians in the United States. *BMC Pulm Med*. 2017;17(1):21.
5. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res*. 2021;23(4):e25759.
6. Dekker FW, Ramspek CL, van Diepen M. Con: most clinical risk scores are useless. *Nephrol Dial Transplant*. 2017;32(5):752–755.
7. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness [published correction appears in *BMJ*. 2020;369:m1312]. *BMJ*. 2020;368:l6927.
8. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606.
9. Steyerberg EW, Moons KG, van der Windt DA, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381.
10. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144(3):201–209.
11. Khalifa M, Magrabi F, Gallego Luxan B. Evaluating the impact of the grading and assessment of predictive tools framework on clinicians and health care professionals' decisions in selecting clinical predictive tools: randomized controlled trial. *J Med Internet Res*. 2020;22(7):e15770.
12. Goto T, Camargo Jr CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning–based prediction of clinical outcomes for children during emergency department triage. *JAMA Network Open*. 2019;2(1):e186937.
13. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *NPJ Digit Med*. 2020;3(1):148.
14. Kappen TH, van Loon K, Kappen MA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol*. 2016;70:136–145.
15. Bate L, Hutchinson A, Underhill J, Maskrey N. How clinical decisions are made. *Br J Clin Pharmacol*. 2012;74(4):614–620.
16. van Giessen A, Peters J, Wilcher B, et al. Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. *Value Health*. 2017;20(4):718–726.
17. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol*. 2014;67(6):612–621.
18. Vickers AJ, Cronin AM. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. *Semin Oncol*. 2010;37(1):31–38.
19. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574.
20. Kerem E, Reisman J, Corey M, Canny GJ, Levison H. Prediction of mortality in patients with cystic fibrosis. *N Engl J Med*. 1992;326(18):1187–1191.
21. Mayer-Hamblett N, Rosenfeld M, Emerson J, Goss CH, Aitken ML. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. *Am J Respir Crit Care Med*. 2002;166(12 Pt 1):1550–1555.

22. Aaron SD, Chaparro C. Referral to lung transplantation–too little, too late. *J Cyst Fibros*. 2016;15(2):143–144.
23. Buzzetti R, Alicandro G, Minicucci L, et al. Validation of a predictive survival model in Italian patients with cystic fibrosis. *J Cyst Fibros*. 2012;11(1):24–29.
24. Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. *Am J Epidemiol*. 2001;153(4):345–352.
25. Nkam L, Lambert J, Latouche A, Bellis G, Burgel PR, Hocine MN. A 3-year prognostic score for adults with cystic fibrosis. *J Cyst Fibros*. 2017;16(6):702–708.
26. Liu Y, Vela M, Rudakevych T, Wigfield C, Garrity E, Saunders MR. Patient factors associated with lung transplant referral and waitlist for patients with cystic fibrosis and pulmonary fibrosis. *J Heart Lung Transplant*. 2017;36(3):264–271.
27. Mitchell AB, Glanville AR. Lung transplantation: a review of the optimal strategies for referral and patient selection. *Ther Adv Respir Dis*. 2019;13:1753466619880078.
28. Thabut G, Christie JD, Mal H, et al. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. *Am J Respir Crit Care Med*. 2013;187(12):1335–1340.
29. Vock DM, Tsiatis AA, Davidian M, et al. Assessing the causal effect of organ transplantation on the distribution of residual lifetime. *Biometrics*. 2013;69(4):820–829.
30. Vock DM, Durheim MT, Tsuang WM, et al. Survival benefit of lung transplantation in the modern era of lung allocation. *Ann Am Thorac Soc*. 2017;14(2):172–181.
31. Knapp EA, Fink AK, Goss CH, et al. The Cystic Fibrosis Foundation Patient Registry. Design and methods of a national observational disease registry. *Ann Am Thorac Soc*. 2016;13(7):1173–1179.
32. Ramos KJ, Sykes J, Stanojevic S, et al. Survival and lung transplant outcomes for individuals with advanced cystic fibrosis lung disease living in the United States and Canada: an analysis of national registries. *Chest*. 2021;160(3):843–853.
33. Stephenson AL, Ramos KJ, Sykes J, et al. Bridging the survival gap in cystic fibrosis: an investigation of lung transplant outcomes in Canada and the United States. *J Heart Lung Transplant*. 2021;40(3):201–209.
34. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021. https://www.R-project.org/.
35. Polley E, LeDell E, Kennedy C, van der Laan M. SuperLearner: Super Learner Prediction. https://CRAN.R-project.org/package=SuperLearner; 2021.
36. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1). article25.
37. Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J*. 2012;40(6):1324–1343.
38. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
39. Vaughan LK, Divers J, Padilla M, et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal*. 2009;53(5):1755–1766.
40. Thompson D, Waisanen L, Wolfe R, Merion RM, McCullough K, Rodgers A. Simulating the allocation of organs for transplantation. *Health Care Manag Sci*. 2004;7(4):331–338.
41. Bansal A, Heagerty PJ. A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. *Med Decis Making*. 2018;38(8):904–916.
42. Alkhateeb AA, Lease ED, Mancl LA, Chi DL. Untreated dental disease and lung transplant waitlist evaluation time for individuals with cystic fibrosis. *Spec Care Dentist*. 2021;41(4):489–497.
43. Egan TM, Murray S, Bustami R, et al. Development of the new lung allocation system in the United States. *Am J Transplant*. 2006;6(5 Pt 2):1212–1227.
44. Chambers DC, Yusen RD, Cherikh WS, et al. The registry of the International Society for Heart and Lung Transplantation: thirty-fourth adult lung and heart-lung transplantation report—2017; focus theme: allograft ischemic time. *J Heart Lung Transplant*. 2017;36(10):1047–1059.
45. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*. 2012;98(9):691–698.
46. Wallace E, Smith SM, Perera-Salazar R, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). *BMC Med Inform Decis Mak*. 2011;11(1):62.
47. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. *J Clin Epidemiol*. 2009;62(12):1248–1252.
48. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
49. Weill D. Lung transplantation: indications and contraindications. *J Thorac Dis*. 2018;10(7):4574–4587.
50. Lynch 3rd JP, Sayah DM, Belperio JA, Weigt SS. Lung transplantation for cystic fibrosis: results, indications, complications, and controversies. *Semin Respir Crit Care Med*. 2015;36(2):299–320.