Evaluating the Clinical Utility of a New Risk Prediction Model in Cystic Fibrosis

Patricia Rodriguez

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Aasthaa Bansal, Chair

David Veenstra

Patrick Heagerty

Christopher Goss

Program Authorized to Offer Degree: Pharmacy

University of Washington

Abstract

Evaluating the Clinical Utility of a New Risk Prediction

Model in Cystic Fibrosis

Patricia J Rodriguez

Chair of the Supervisory Committee:
Aasthaa Bansal
Associate Professor
Pharmacy

Risk prediction models (RPMs), which estimate the probability of some future event, can inform clinical decisions about appropriate testing and treatment. Novel, machine learning (ML)-based RPMs have demonstrated superior performance for predicting events in numerous clinical applications, but the utility of such models for decision-making in real-world clinical practice remains unclear and uptake has been limited. We explored the development and evaluation of a novel RPM in Cystic Fibrosis (CF), where predictions of short-term mortality can inform decisions about when to refer patients for lung transplantation (LTx). In the first aim, we used real-world data (RWD) and ML approaches to develop a novel RPM for predicting 2-year mortality among adults with CF. We compared the discrimination accuracy and calibration of 8 potential ML models to the biomarker forced expiratory volume in 1 second (FEV_1) alone. Super learner, an ensemble approach, had the highest discrimination accuracy at baseline, with an area under the receiver operating curve (AUC) at baseline of 0.914 (95% CI: 0.898, 0.929), compared to 0.876 (0.858, 0.895) for FEV_1. In the second aim, we considered the potential impact of using the novel ML model for clinical decision-making. Using health outcomes modelling with RWD, we predicted the clinical decisions and downstream health outcomes of three alternative policies for LTx referral: (1) ML-based decisions, (2) FEV_1-based decisions, and (3) usual care (UC) decisions identified in RWD. ML-based referral resulted in more patients referred for LTx (20.4% of patients (19.1%, 21.6)), compared to FEV_1 (19.2% (18.0%, 20.4%)) and UC (12.4% (11.4%, 13.4%)). Of patients who died without referral under usual care, 40% would have been referred under ML and 31% would have been referred under FEV_1. However, given a fixed supply of organs available for transplantation, higher referral rates did not lead to differences in the number of transplants or pre-

transplant deaths. We found no significant difference in 5-year post-transplant or overall 5-year survival among policies. Our work demonstrates the value of using health outcomes modelling with RWD to evaluate the potential real-world clinical utility of novel RPMs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my deep gratitude to all members of my dissertation committee. Each committee member brought a unique perspective that strengthened the quality this work and my capabilities as a researcher. Thank you for your time, ideas, and mentorship. In addition to my dissertation committee, Dr. Kathy Ramos provided substantial guidance on this work. I would also like to thank the CF Foundation and CF patients for sharing data to support this research.

I am privileged to have had many mentors during my time as a PhD student. Dr. Aasthaa Bansal spent many hours discussing potential research ideas, pointing me to topics of interest, and guiding my dissertation work. Dr. Davene Wright supported my ideas and showed me how to be a confident researcher and communicator. All CHOICE faculty members provided invaluable guidance on everything from coursework to career paths throughout the PhD.

I would not have finished this PhD without my brilliant and supportive fellow students at CHOICE.

Finally, my deepest thanks to my family and partner who have been constant supporters through this journey.

Chapter 1. INTRODUCTION

Risk prediction models (RPMs) estimate the probability of some future event, such as disease recurrence or short-term mortality. RPMs can be used to predict individualized patient risk, allowing clinicians to identify high-risk patients for additional testing or treatment. Recently, development of new RPMs has accelerated,[1] particularly models developed using machine-learning (ML) methods.[2]

Unlike traditional methods for RPM development, which rely on scientific knowledge to model clinical pathways, ML methods attempt to harness the power of data to learn observed patterns. ML-based RPMs have demonstrated high performance for predicting events in numerous healthcare applications.[3-6] Flexible ML algorithms can capture complex interactions between variables to better predict outcomes. With increasing access to big, high-dimensional datasets generated through clinical care, such as EHR data, ML has become an attractive and feasible option for the development of new RPMs across clinical domains.

However, despite high expectations about the value of ML and artificial intelligence (AI) for predicting clinical outcomes, applications of ML/AI in clinical practice remain rare.[7-9] A recent systematic review found only 51 applications of AI in real-world clinical practice, among over 15,000 publications on ML or AI identified.[10] This well-known gap between model development and implementation has been termed the *AI chasm*.[11] Numerous reasons for the chasm exist, including small and/or unrepresentative samples used in model development, low model interpretability, lack of validation studies, and logistical implementations challenges.[7, 12, 13] In this dissertation, we focus in particular on the need for evidence about the real-world clinical utility offered by novel RPMs. Current methods for assessing RPM performance focus primarily on improvements in predictive accuracy, rather than improvements in clinical actions and subsequent patient outcomes. While improvements in predictive accuracy are necessary, additional measures are necessary to evaluate real-world clinical utility. As Shah et al. recently described in JAMA,

> *"[R]ealizing the potential benefit of machine learning for patients in the form of better care requires rethinking how model performance during machine learning is assessed. A framework for rigorously evaluating the performance of a model in the context of the subsequent actions it triggers is necessary to identify models that are clinically useful."*[14]

This dissertation explores the development and the rigorous evaluation of a novel ML-based RPM, using Cystic Fibrosis (CF) as a case study. In CF, predictions of short-term morality can inform decisions about when to refer a patient for lung transplant (LTx). Improved risk predictions are critically needed because current approaches to predicting short-term mortality have relatively poor performance.[15-20] In Chapter 2, we use state-of-the-art ML methods to develop candidate risk prediction models for short-term mortality in CF and evaluate models using standard measures of predictive performance. In Chapter 3, we use a framework of health outcomes modelling with real-world data to predict real-world LTx referral decisions and downstream patient health outcomes when using an ML for LTx referral decisions, compared to usual care.

Chapter 2. DEVELOPING A NOVEL MODEL FOR MORTALITY RISK PREDICTION IN CYSTIC FIBROSIS

2.1 INTRODUCTION

Cystic Fibrosis (CF) is a genetic disease that causes death from progressive respiratory failure.[21] While advances in the treatment of CF have led to large improvements in survival, lung transplant (LTx) remains a treatment approach after other therapeutic options have been exhausted.[22-25] Determining the appropriate time for referral for LTx remains challenging because existing models for predicting short-term mortality in CF have poor performance.[26] Many risk prediction models in CF have been attempted using traditional methods of first identifying clinically relevant variables, then using stepwise selection or variable screening based on statistical significance to select a final model.[17-20] However, previously developed risk prediction models typically perform no better than percent predicted forced expiratory volume in 1 second ($ppFEV_1$), a standard test of pulmonary function.[17, 19, 27 15, 28] Although $ppFEV_1$ is a strong biomarker,[15] the accuracy of clinical decision making may be improved if $ppFEV_1$ could be combined with other clinical measures or if the dynamics of longitudinal measures were incorporated into risk prediction models. On its own, $ppFEV_1$ <30%, a common threshold used for referral, has low positive predictive value,[19] and median survival exceeding 6.5 years.[29] A prediction model that could better discriminate between patients who are likely to die within two years and patients likely to survive could be valuable for informing LTx referral decisions.

Machine learning (ML) methods are promising for the development of an accurate multivariate risk prediction model in CF. Machine learning methods encompass a broad range of statistical tools, which vary in flexibility and interpretability.[30, 31] In general, these methods make use of large datasets to empirically develop models based on a large number of candidate predictors, rather than strategies that limit to variables based on pre-specified scientific knowledge. ML methods were recently used to predict short-term mortality in CF using UK registry data, and performed better than $ppFEV_1$ alone.[32] We aim to use ML methods to develop a more accurate model for short-term mortality in adults with CF in the United States.

2.2 METHODS

2.2.1 *Data*

We used 2012-2016 data from the CF Foundation Patient Registry (CFF PR), which collects observational data for all US patients seen at CFF-accredited care centers who consent to participate.[33] The CFF PR is estimated to cover 80% the US CF population, and captures approximately 95% of clinic visits and 90% of hospitalizations for participating patients.[33] Data on patient diagnosis, demographics, encounters, care episodes, and annual summaries are entered electronically by care center staff using information from electronic medical records and patient forms.[33] For this study, the CFF PR data was merged with data from the United Network for Organ Sharing (UNOS). UNOS data was used only to ascertain death and lung transplant outcomes for individuals who entered the lung transplant waiting list. The data merge and initial data cleaning have been described elsewhere.[34] This study was approved by the University of Washington Institutional Review Board (Study #2270) and the Seattle Children's Research Institute (Study #PIROSTUDY15294).

Our study cohort consisted of US CFF PR adults (≥18 years old) who had not undergone LTx by January 1, 2012, and recorded at least one encounter in 2012 (n = 11,524). We excluded individuals without any encounter (n= 901) or annual (n=8) level data in 2011. The final study cohort included 10,615 patients. The first encounter in 2012 was used as the baseline encounter. We split the patient population into training (60%) and validation sets (40%) to validly develop and evaluate prediction models according to guidelines in the TRIPOD statement.[35] Because the registry is estimated to cover >80% of the population of interest, including all CF centers in the US, we did not use an external validation set from a separate data source.

2.2.2 *Outcome*

We considered a binary composite outcome of pre-transplant death or LTx in two years. While death without LTx and transplant are not equivalent, both represent severe outcomes in CF and the use of a composite outcome is common when measuring clinical progression of disease.[20, 32] In

model development we considered outcomes within 2 years of baseline. In model validation, we also considered time-varying outcomes, with 2-year outcome status updated at each encounter from 2012 - 2014.

2.2.3 *Variables*

We aimed to include as many variables as possible in model development, including patient demographics, encounters, episodes, and annual health summaries. We took several pre-processing actions to improve the accuracy and generalizability of our model. For lab and microbiology values, which were frequently missing at any given encounter but generally available in the prior year, we used the most recent available lab and microbiology values in the past 12 months and included an indicator for the whether a lab value was measured at the current visit. We categorized lab values into clinically meaningful categories and used a missingness category when no values were available in the prior 12 months. We excluded lab measures for which >50% of patients had no measures in the last 12 months. Complications and comorbidities were treated as present if documented at any encounter in the prior 12 months and absent otherwise. We excluded very rare complications and bacteria (occurring in <1% of patients). All episode information was summarized for the prior rolling 12-month period (i.e. total hospital nights in prior 12 months).

Our analysis adopts a novel strategy for the use of longitudinal clinical assessments and leverages the availability of individual patient histories for use in prediction. Specifically, we incorporated longitudinal information for height, weight, and pulmonary function tests (PFTs) using partly conditional models, which relate a longitudinal biomarker process to an outcome process. PFTs include $ppFEV_1$, percent predicted forced vital capacity (ppFVC), and FEV1 to FVC ratio (FEV_1/FVC). In order to fully leverage the longitudinal history of each time-varying PFT, we derived individual fitted values from a comprehensive linear mixed model analysis and for each patient at each follow-up time we extracted the best linear unbiased predictor (BLUP) estimator, which both incorporates information from patients' longitudinal histories and can reduce measurement error.[36] Raw and BLUP $ppFEV_1$ values over time are presented graphically for a sample of patients (Figure 1). We view the computation of predicted values as a form of feature-engineering that can be done to incorporate biological or statistical information about a set of predictors prior to their incorporation into machine learning algorithms in order to improve

empirical performance.[37] BLUPs were also used to impute the value at encounters where it was missing, using information collected up to that point. We also included the change in PFT BLUP values over the past 6 months. In a sensitivity analysis, we tested whether the inclusion of residuals (i.e. $ppFEV_1$ -$ppFEV_1$ BLUP) improved performance, which would suggest information loss due to over-smoothing of biomarker values. Additional detail is provided in Appendix A.

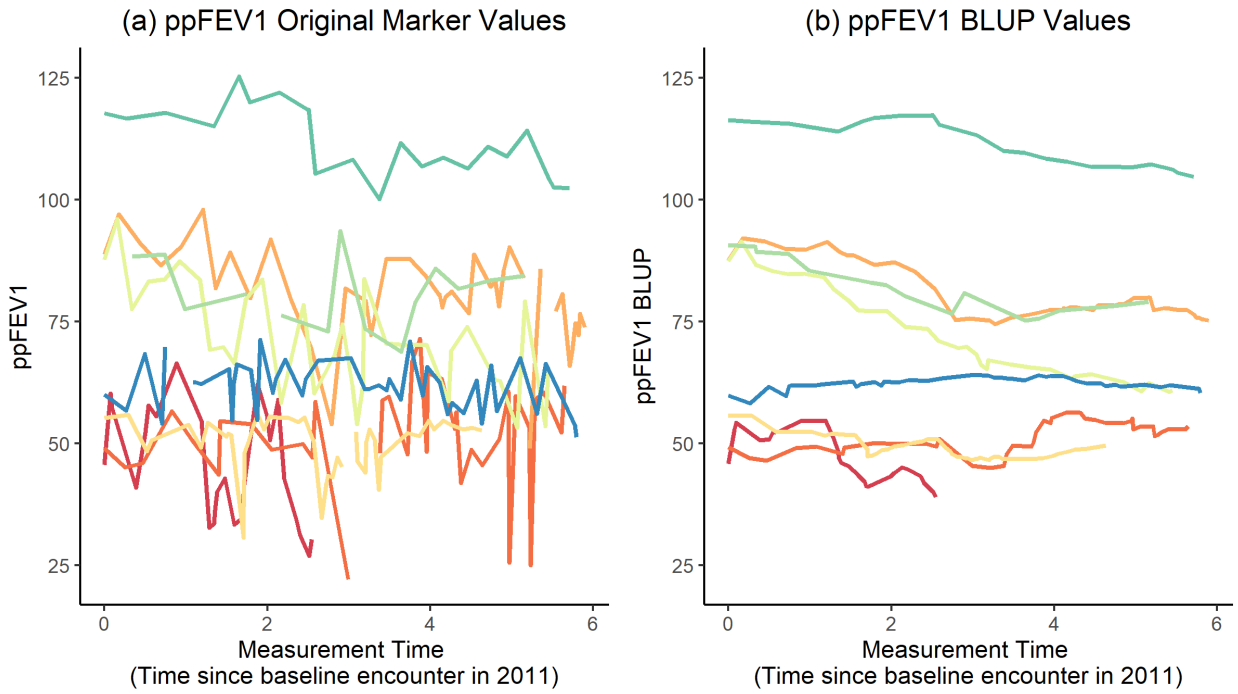


Figure 1: Raw and smoothed ppFEV1 values over time. ppFEV1 over time is presented for 8 randomly sampled patients, each represented by color. Panel a contains raw ppFEV1 measures and panel b contains smoothed best linear unbiased predictor (BLUP) values, which incorporate an individual's ppFEV1 history up to the current measure.

Consistent with previous CF risk prediction models, our primary model considered physiologic factors only, and did not include race or proxies of socio-economic status (SES).[18-20, 32] Despite lower SES being associated with poorer outcomes in many applications, including CF,[38-42] race and SES proxies are rarely included explicitly in clinical risk prediction models because inclusion can be perceived as racial profiling or supporting differential decision-making on the basis of race or SES.[43, 44] To test whether the inclusion of race and SES proxies would improve performance overall or within lower SES subgroups, we tested secondary models that included race and SES proxies (education, insurance type, patient assistance programs). We compared the AUC and

calibration of models that included race and SES proxies to the AUC of our primary models. We also evaluated model AUC within subgroups of educational attainment, a proxy for SES.

Multi-level categorical variables were transformed to binary variables. We included squared terms for age, BMI, height, and PFT values and interactions between $ppFEV_1$ and each of gender, age, delF508 status, and an indicator of pulmonary exacerbation at the current encounter. The primary model considered 202 binary or continuous variables.

2.2.4 *Models*

We considered a diverse set of ML models: lasso,[45] elastic net,[46] ridge regression,[47] random forest,[48] extreme gradient boosting (XGBoost),[49] and support vector machine (SVM).[50] We also used a stacking ensemble, super learner[51], which combines the base ML models in an optimally weighted combination, and has shown high performance in many applications.[51] We chose models that vary in interpretability and flexibility. Lasso, elastic net, and ridge are penalized regression approaches that offer less flexibility but can be interpreted directly. Lasso may be especially appealing in clinical practice because it can perform variable selection, yielding a sparse, easily interpretable model.[45] In contrast, random forest, XGboost, and SVM are highly flexible approaches that may better capture non-linearity in predictors, but have limited interpretation. Finally, super learner can capitalize on the strengths of each base model by creating a weighted ensemble.[51] For all models, tuning parameters were selected through 10-fold cross-validation, then fit to the entire training set. We provide additional detail on tuning parameters and fitting procedures in Appendix B. All analyses were conducted in R.[52]

To aid in interpretability, variable importance measures were calculated for random forests using a permutation approach and XGBoost using frequency. At present, variable importance measures are not available from SuperLearner.[53] As a comparator, we separately considered a model of raw $ppFEV_1$ alone. For observations missing $ppFEV_1$, we imputed $ppFEV_1$ using the $ppFEV_1$ BLUP.

2.2.5 *Evaluation*

We evaluated the performance of all ML models against $ppFEV_1$ alone in the 40% validation set. Discrimination accuracy was evaluated in three ways. First, we evaluated AUC at baseline for all

models. Second, we assessed AUC over time using dynamic predictions, where patients' predicted and true outcome status were updated at each encounter between 2012 and 2014 using the most recently collected information. Dynamic predictions are more reflective of clinical practice, where decisions are made repeatedly over time using new information, rather than at a single timepoint.[27, 54] Third, we assessed sensitivity, positive predictive value (PPV), and negative predictive value (NPV) for fixed specificity levels of 85%, 90%, 95%, and 99%. Confidence intervals were obtained through bootstrapping. Calibration was assessed graphically by comparing observed versus predicted risk deciles at baseline. While our primary analysis defined cases and controls based on 2-year outcomes, we also considered the predicted risk distribution with controls disaggregated into those with events outside the primary 0-2 year window, and consider proximal controls as subjects with subsequent events in 2-3 years (near controls) as compared to those without events through 3 years (far controls).

Because referral for transplant is only considered when patients have elevated risk of short-term mortality, we performed subgroup analyses on the higher risk population, which we define as patients with $ppFEV_1$ <40 at baseline. We separately evaluated AUC and calibration within this subgroup.

2.3 RESULTS

2.3.1 *Population Characteristics*

Characteristics at baseline by outcome status are provided in Table 1. Overall, 8% of the sample experienced death (n= 467) or transplant (n= 362) within 2 years of the baseline 2012 encounter. Characteristics disaggregated by death vs LTx outcomes are given in Appendix C. Those with pre-LTx death or LTx within 2 years had lower BMI, $ppFEV_1$, and ppFVC at baseline, and higher rates of oxygen use. Consistent with CF demographics, the study population is overwhelming white (95.6%). Few patients (1%) were on a cystic fibrosis transmembrane conductance regulator (CFTR) modulator (ivacaftor) at baseline.

Table 1: Baseline Characteristics. Continuous variables reported as mean(sd). Categorical variables reported as n(%). Height, ppFEV1, ppFVC refer to smoothed BLUP values. BMI is calculated from smoothed height and weight BLUP values.

	Alive, pre-LTx in 2 Years (n=9768)	Death or LTx in 2 years (n=847)	Overall (n=10615)
Age	30.0 (10.6)	31.8 (11.0)	30.1 (10.6)
Female	4588 (47.0%)	443 (52.3%)	5031 (47.4%)
Race			
White	9332 (95.5%)	809 (95.5%)	10141 (95.5%)
Black	314 (3.2%)	27 (3.2%)	341 (3.2%)
Hispanic	19 (0.2%)	3 (0.4%)	22 (0.2%)
Other or Unknown	103 (1.1%)	8 (0.9%)	111 (1.0%)
BMI	22.7 (3.93)	20.4 (4.20)	22.5 (4.00)
Height	168 (9.43)	165 (9.86)	168 (9.49)
Mutation Class			
1, 2, or 3	6960 (71.3%)	622 (73.4%)	7582 (71.4%)
4 or 5	1084 (11.1%)	46 (5.4%)	1130 (10.6%)
Other	155 (1.6%)	54 (6.4%)	209 (2.0%)
Unknown	1569 (16.1%)	125 (14.8%)	1694 (16.0%)
F508del Mutation			
Homozygous	4460 (45.7%)	416 (49.1%)	4876 (45.9%)
Heterozygous	3933 (40.3%)	297 (35.1%)	4230 (39.8%)
None	1235 (12.6%)	81 (9.6%)	1316 (12.4%)
Unknown	140 (1.4%)	53 (6.3%)	193 (1.8%)
ppFEV1	65.2 (21.9)	34.9 (15.4)	62.8 (23.0)
ppFVC	79.4 (18.4)	52.7 (15.9)	77.3 (19.6)
Oxygen			
Yes	1248 (12.8%)	564 (66.6%)	1812 (17.1%)
No	8374 (85.7%)	275 (32.5%)	8649 (81.5%)
Unknown	146 (1.5%)	8 (0.9%)	154 (1.5%)
P. aeruginosa	5939 (60.8%)	638 (75.3%)	6577 (62.0%)
S. aureus	5052 (51.7%)	374 (44.2%)	5426 (51.1%)
MRSA	2309 (23.6%)	280 (33.1%)	2589 (24.4%)
MSSA	3807 (39.0%)	212 (25.0%)	4019 (37.9%)
Burkholderia complex	339 (3.5%)	50 (5.9%)	389 (3.7%)
B. cenocepacia	49 (0.5%)	5 (0.6%)	54 (0.5%)
Ivacaftor	114 (1.2%)	6 (0.7%)	120 (1.1%)
Education			
High School	2749 (28.1%)	287 (33.9%)	3036 (28.6%)
College	6040 (61.8%)	451 (53.2%)	6491 (61.1%)
Unknown	979 (10.0%)	109 (12.9%)	1088 (10.2%)
Insurance			
Private	6152 (63.0%)	366 (43.2%)	6518 (61.4%)
Medicaid	2356 (24.1%)	362 (42.7%)	2718 (25.6%)
Other	557 (5.7%)	39 (4.6%)	596 (5.6%)
Medicare	496 (5.1%)	74 (8.7%)	570 (5.4%)
Missing	207 (2.1%)	6 (0.7%)	213 (2.0%)

2.3.2 *Discrimination Accuracy*

All ML models had higher AUC at baseline than $ppFEV_1$ alone, though confidence intervals overlap in some cases (Table 2). Super learner had the highest AUC at baseline, 0.914 (95% CI: 0.899, 0.929) and $ppFEV_1$ had the lowest, 0.877 (0.858, 0.896). When predictions and outcomes were updated dynamically over time, as is done clinically with the lung allocation score, AUC remained fairly consistent (Table 2). Model AUC rankings also remained relatively consistent over time except for SVM and $ppFEV_1$, which alternately performed worst.

ML models also outperformed $ppFEV_1$ at thresholds of decision-making corresponding to fixed levels of baseline specificity, though confidence intervals overlapped (Table 3). Super learner had the highest sensitivity at all levels of fixed specificity. At fixed specificity of 95%, super learner had sensitivity of 60% (51%, 64%), while $ppFEV_1$ had sensitivity of 49% (42%, 55%). Among ML models, SVM had the worst performance at all levels of specificity. For fixed specificity levels ≤95%, all models had PPV <50% (Table 3). PPV point estimates were higher for ML models than $ppFEV_1$, but confidence intervals overlapped. All models predicted higher risk for patients who had an event in 2-3 years (near controls), compared to those without events in 3 years (far controls) (Figure 2).

Table 2: AUC Performance over Time (Dynamic Predictions). AUC is given every 6 months from baseline, using the most recent risk prediction and outcome status at that time, for all patients still at risk of pre-LTx mortality.

Model	Baseline	6 months	12 months	18 months	24 months	30 months	36 months
Super Learner	0.914 (0.898, 0.929)	0.917 (0.903, 0.932)	0.927 (0.913, 0.94)	0.921 (0.907, 0.935)	0.920 (0.906, 0.934)	0.918 (0.904, 0.932)	0.904 (0.875, 0.932)
Ridge	0.91 (0.894, 0.926)	0.916 (0.902, 0.931)	0.927 (0.914, 0.94)	0.919 (0.904, 0.933)	0.916 (0.902, 0.93)	0.912 (0.898, 0.926)	0.901 (0.875, 0.928)
Elastic Net	0.912 (0.897, 0.927)	0.917 (0.904, 0.931)	0.927 (0.914, 0.939)	0.92 (0.906, 0.933)	0.916 (0.902, 0.93)	0.916 (0.901, 0.93)	0.902 (0.875, 0.93)
Lasso	0.913 (0.898, 0.928)	0.918 (0.904, 0.932)	0.927 (0.915, 0.94)	0.92 (0.907, 0.934)	0.917 (0.903, 0.931)	0.916 (0.902, 0.931)	0.904 (0.877, 0.931)
Random Forest	0.91 (0.894, 0.925)	0.914 (0.899, 0.929)	0.922 (0.908, 0.936)	0.916 (0.902, 0.931)	0.918 (0.904, 0.932)	0.918 (0.903, 0.933)	0.899 (0.869, 0.93)
XGBoost	0.911 (0.895, 0.926)	0.917 (0.903, 0.93)	0.919 (0.906, 0.933)	0.916 (0.902, 0.931)	0.917 (0.903, 0.931)	0.914 (0.899, 0.93)	0.901 (0.872, 0.929)
SVM	0.884 (0.865, 0.903)	0.882 (0.863, 0.902)	0.898 (0.881, 0.915)	0.896 (0.880, 0.912)	0.890 (0.872, 0.908)	0.888 (0.869, 0.906)	0.877 (0.845, 0.909)
ppFEV1	0.876 (0.858, 0.895)	0.884 (0.867, 0.901)	0.891 (0.874, 0.908)	0.890 (0.873, 0.907)	0.887 (0.868, 0.905)	0.891 (0.872, 0.909)	0.888 (0.86, 0.915)

Table 3: Sensitivity, Positive Predictive Value, and Negative Predictive Value for Fixed Specificity. Sensitivity, specificity, positive predictive value, and negative predictive value were calculated at baseline for each model. 95% confidence intervals were bootstrapped.

Model	Specificity = 85%	Specificity = 90%	Specificity = 95%	Specificity = 99%
		Sensitivity		
Super Learner	84% (79%, 88%)	77% (71% ,81%)	60% (51% ,64%)	37% (25%, 36%)
Lasso	83% (79%, 86%)	75% (68%, 79%)	59% (51%, 63%)	33% (24%, 35%)
Random Forest	83% (78%, 87%)	75% (69%, 80%)	60% (53%, 64%)	32% (22%, 34%)
Elastic Net	83% (78%, 86%)	76% (68%, 79%)	59% (51%, 63%)	33% (24%, 35%)
Ridge	82% (77%, 86%)	76% (69%, 80%)	60% (48%, 64%)	34% (22%, 34%)
XGBoost	81% (75%, 84%)	74% (68%, 78%)	58% (51%, 62%)	34% (23%, 36%)
SVM	76% (71%, 81%)	72% (64%, 76%)	54% (47%, 58%)	32% (22%, 34%)
$ppFEV_1$	73% (68%, 78%)	66% (60%, 71%)	49% (42%, 55%)	26% (14%, 29%)
		PPV		
Super Learner	32% (29%, 35%)	39% (36%, 42%)	48% (45%, 53%)	68% (66%, 76%)
Lasso	32% (29%, 35%)	38% (35%, 42%)	48% (45%, 53%)	65% (65%, 75%)
Random Forest	31% (29%, 35%)	38% (35%, 42%)	49% (45%, 53%)	66% (64%, 75%)
Elastic Net	31% (29%, 35%)	38% (35%, 42%)	48% (45%, 53%)	65% (65%, 75%)
Ridge	31% (29%, 34%)	39% (35%, 42%)	49% (44%, 53%)	67% (64%, 75%)
XGBoost	31% (28%, 34%)	38% (35%, 41%)	48% (44%, 53%)	66% (65%, 77%)
SVM	31% (27%, 33%)	37% (34%, 41%)	46% (43%, 51%)	64% (60%, 74%)
$ppFEV_1$	29% (26%, 32%)	36% (32%, 39%)	44% (40%, 49%)	60% (55%, 71%)
		NPV		
Super Learner	98% (98%, 99%)	98% (97%, 98%)	97% (96%, 97%)	95% (94%, 95%)
Lasso	98% (98%, 99%)	98% (97%, 98%)	96% (96%, 97%)	95% (94%, 95%)
Random Forest	98% (98%, 99%)	98% (97%, 98%)	97% (96%, 97%)	94% (93%, 95%)
Elastic Net	98% (98%, 99%)	98% (97%, 98%)	96% (96%, 97%)	95% (94%, 95%)
Ridge	98% (98%, 99%)	98% (97%, 98%)	97% (95%, 97%)	95% (93%, 95%)
XGBoost	98% (98%, 99%)	98% (97%, 98%)	96% (96%, 97%)	95% (94%, 95%)
SVM	98% (97%, 98%)	97% (97%, 98%)	96% (95%, 97%)	94% (93%, 95%)
$ppFEV_1$	97% (97%, 98%)	97% (96%, 97%)	96% (95%, 96%)	94% (93%, 94%)

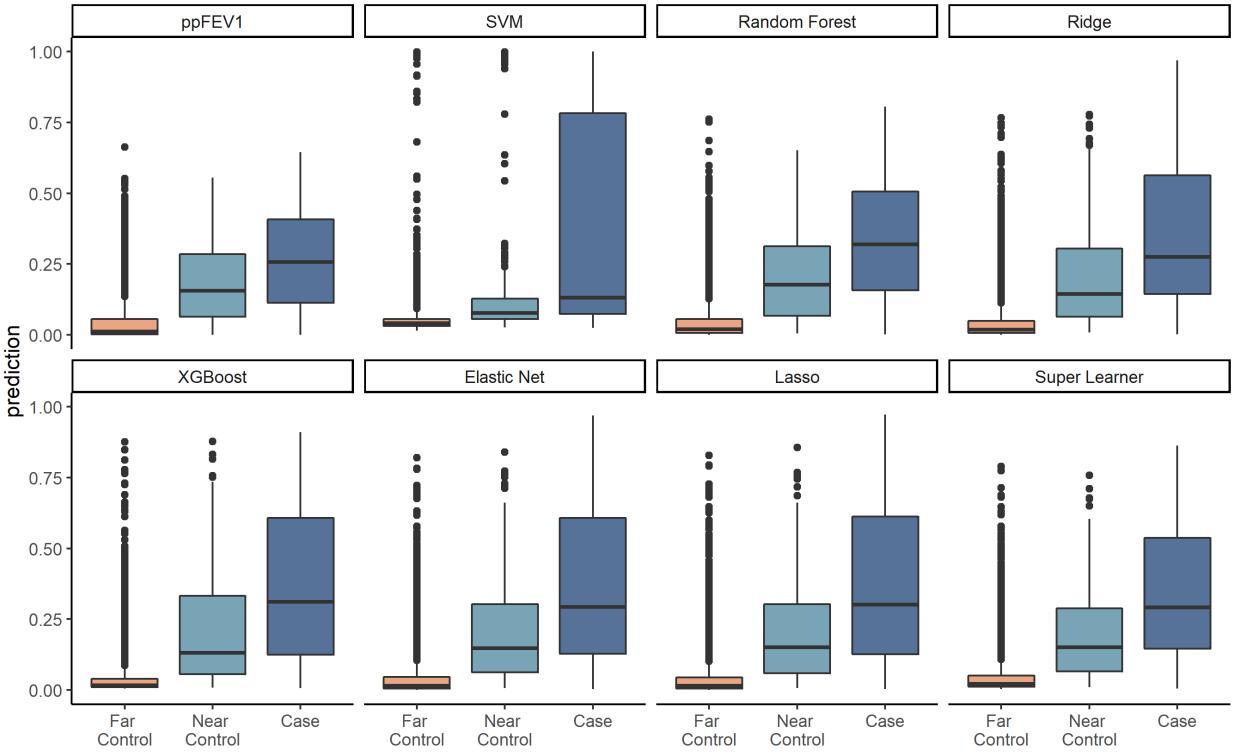


Figure 2: Discrimination by Outcome, Near vs. Far controls. Cases refer to patients who experienced death or LTx within 2 years of baseline. Near controls experienced death or LTx in 2-3 years from baseline. Far controls did not experience death or transplant within 3 years. Y-axis refers to baseline predictions for each model.

2.3.3 *Calibration*

ML models generally had better calibration than raw $ppFEV_1$ (Figure 3). $ppFEV_1$ underpredicted events in the highest risk patients, with the top decile by $ppFEV_1$ having mean predicted probability of death or transplant of 36.6% and observed events in 42.5% of patients. In contrast, ML models better discriminated between cases and controls, such that the observed event rate in the top decile of risk was higher and predicted more accurate risk, with mean predicted probabilities of event by decile closer to observed probabilities. For example, the highest risk decile for the lasso model had mean predicted probability was 47.7% and observed events occurred in 47.4% of patients. The top super learner decile had a mean predicted probability of events of 43.1% and observed events occurred in 47.6% of patients.

Variable importance results, provided in Appendix D, showed that measures of pulmonary function, episodes, and BMI values were most important to predictions in the random forest and XGBoost models. In both models, $ppFEV_1$ was the single most important predictor.

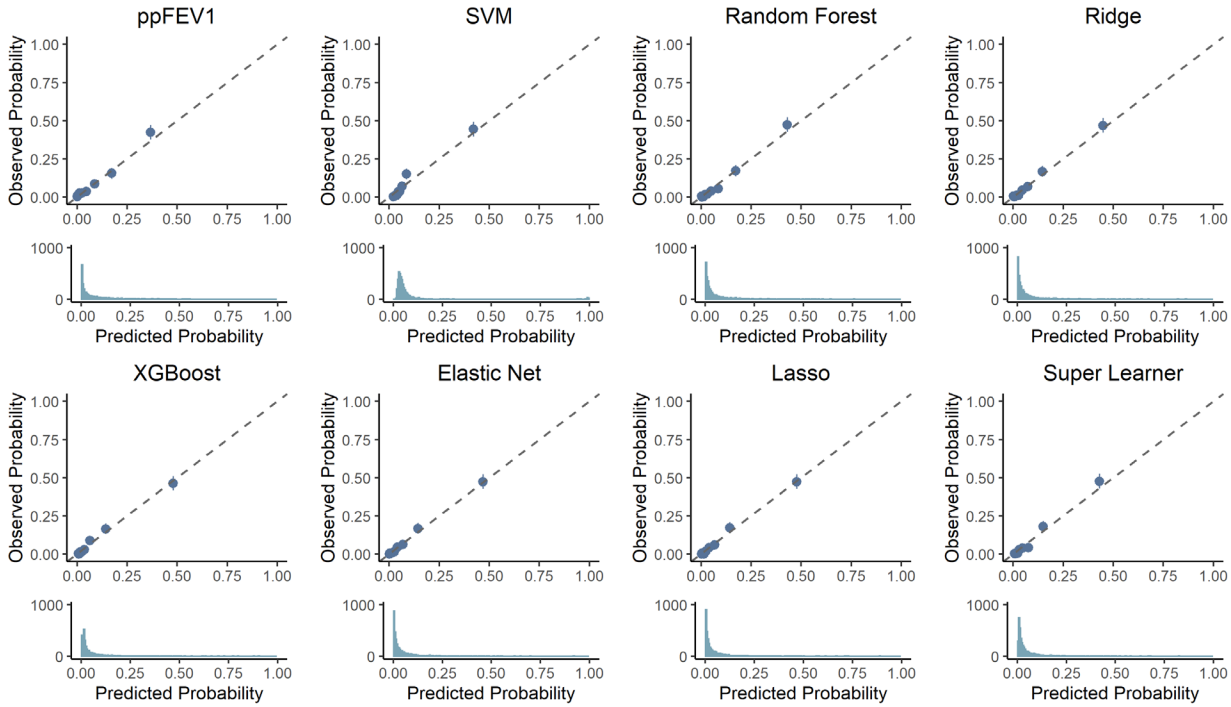


Figure 3: Baseline Calibration. For each model, the top panel compares observed versus predicted probability of 2-year death of LTx at baseline. The diagonal reference line indicates perfect calibration, where the observed probability of events (y-axis) equals the predicted probability of events (x-axis). Each decile of model-specific risk is represented as a point. Points that lie above the dashed line represent underprediction (observed probabilities are higher than predicted) and points below the dashed line represent overprediction (observed probabilities are lower than predicted). The bottom panel contains the distribution of predictions for each model. In all models, most patients had a low (<20%) predicted probability of death or LTx.

2.3.4 *High-Risk Subgroup Analysis*

Compared to the overall cohort, baseline AUC in the higher risk population ($ppFEV_1$ <40% at baseline) was lower for all models, but the difference between the ML models and $ppFEV_1$ was greater (Table 4Error! Reference source not found.). AUC was highest for random forest, 0.80

(0.77, 0.83), and lowest for $ppFEV_1$, 0.73 (0.69, 0.77). Calibration plots for the higher risk subgroup are shown in Figure 4. Among this subgroup, the range of predicted risk for a model with $ppFEV_1$ alone is relatively narrow, with the highest risk decile having predicted risk of 51.3% and observed events in 64.0% of patients. In contrast, the ML-based models spread risk over a wider and more accurate range. For example, the highest lasso decile had a mean predicted probability of 80.9% and observed events in 77.5%.

Table 4: Baseline Performance, High-Risk Subgroup (Baseline $ppFEV_1$ <40)

Model	AUC
Super Learner	0.794 (0.762, 0.827)
Lasso	0.788 (0.755, 0.821)
Elastic Net	0.787 (0.754, 0.821)
XGBoost	0.792 (0.76, 0.825)
Ridge	0.782 (0.748, 0.815)
Random Forest	0.798 (0.766, 0.830)
SVM	0.762 (0.727, 0.798)
ppFEV1	0.731 (0.694, 0.768)

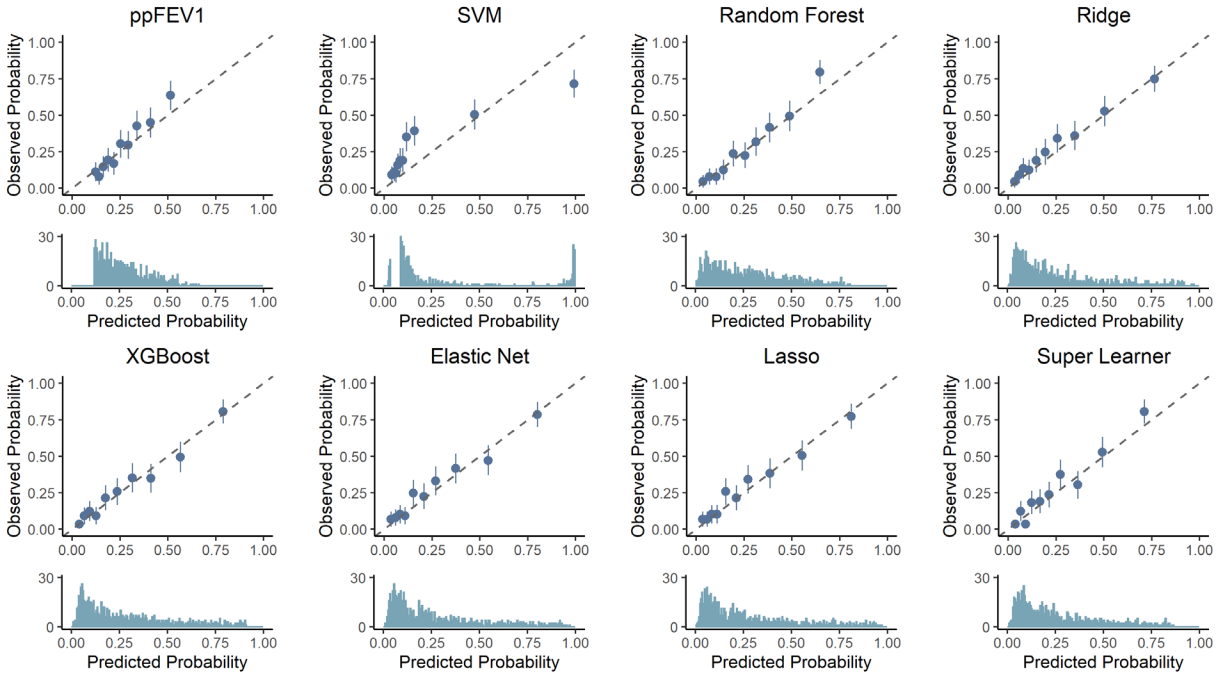


Figure 4: Baseline Calibration, High-Risk Population (Baseline ppFEV$_1$ <40). For each model, the top panel compares observed versus predicted probability of 2-year death of LTx at baseline. The diagonal reference line indicates perfect calibration, where the observed probability of events (y-axis) equals the predicted probability of events (x-axis). Among the high-risk population, each decile of model-specific risk is represented as a point. Points that lie above the dashed line represent underprediction (observed probabilities are higher than predicted) and points below the dashed line represent overprediction (observed probabilities are lower than predicted). The bottom panel contains the distribution of predictions for each model. In all models, most patients had a low (<20%) predicted probability of death or LTx.

2.3.5 *SES Subgroup Analysis*

AUC differed across levels of education attainment, a proxy for SES (Appendix E). 61% of patients had at least some college education. All ML models performed best among those with college education, compared to those with a high school education or unknown educational attainment. Super learner baseline AUC was 0.93 (0.91, 0.95), 0.89 (0.86, 0.92), and 0.89 (0.83, 0.95) among subgroups with college education, high school education, and unknown educational attainment, respectively (Appendix E). The disparity in discrimination accuracy between levels of educational attainment was similar for all models, including ppFEV$_1$ alone. Including SES predictors explicitly in ML models did not impact overall performance, nor did it reduce the disparity in accuracy between levels of educational attainment (Appendix E). For example, super

learner AUC was 0.91 (0.90, 0.93) in versions with and without SES variables. Further, SES variables were not among the 20 most important variables in random forest or XGBoost models.

A secondary analysis of a model that included BLUP residuals as candidate variables had minor differences in AUC (Appendix A). For example, AUC for super learner was 0.913 (0.898, 0.929), compared to 0.914 (0.899, 0.929) in the primary model.

2.4 DISCUSSION

We find that ML-based models had marginally higher discrimination accuracy and sensitivity for fixed specificity than $ppFEV_1$ alone for the prediction of death or LTx in adults with CF in the US. While ML models performed better than $ppFEV_1$ alone, no single model consistently performed best across all evaluation metrics or subgroups. The super learner had the highest AUC and sensitivity for fixed specificity, but poorer calibration overall. Lasso showed better calibration overall, but performed poorly among the higher-risk subgroup.

While our ML models can be used for predictions in all adult patients, the benefits are especially prominent among the more clinically relevant population of higher-risk patients ($ppFEV_1$ <40%). In this population, the difference in AUC between ML models and $ppFEV_1$ was greater. In addition, ML models had better calibration than $ppFEV_1$ and spread risk over a larger and more accurate range.

This is the first study to use longitudinal feature engineering and a variety of ML approaches for the development of a new risk prediction model for adults with CF in the US. A previous study in the UK also used ML to predict short-term (3-year) mortality in CF.[32] Our methods differ in several respects: the UK study considered a full automated ML pipeline, did not include super learner, and used a population of all ages. Still, they similarly found that among classifiers considered, XGBoost performed relatively well (AUC = 0.87 ±0.02) and SVM performed relatively poorly (AUC = 0.84 ±.03). The most important variables in our random forest and XGBoost models are consistent with previous studies. $ppFEV_1$ was the most important variable in both models, similar to previous studies.[15, 18-20] Other important variables in our models, including BMI, oxygen use, and various indicators of number of clinical encounters have also been identified in previous

models.[19, 20, 32] However, our ML-based models allowed for more flexible inclusion of these variables, which may have improved performance relative to traditional models. The ML models in this study had higher sensitivity for fixed levels of specificity than previous models in the US. A 2002 study by Mayer-Hamblett and colleagues found that for fixed specificity of 95%, their multivariate model had sensitivity of 52% (39%, 64%) and PPV of 33% (24%, 44%), while $ppFEV_1$ had sensitivity of 42% (30%, 55%) and PPV of 28% (19%, 38%). At the same specificity, we found sensitivity of 60% (51%, 64%) and PPV of 48% (45%, 53%) for super learner, and sensitivity of 49% (42%, 55%) and PPV of 44% (40%, 49%) for ppFEV1 alone. However, differences may be due to different populations (all CF patients versus adults), outcome definitions (death versus death without LTx or transplant), and time periods studied.

In addition to the use of ML, our approach expands on previous studies by using dynamically updated information. To our knowledge, previous models in a US population have relied on annualized calendar year data. When used in clinical practice, such annualized data would only be available for last complete calendar year, which may not represent the most recent information on the patient and would become more outdated as the year progresses. In contrast, we used encounter-level data and summarized lab, microbiology, and episode information on a rolling 12-month basis, so that predictions can be updated at every encounter using the most recently collected information. We found AUC was maintained over time when using dynamic predictions, suggesting our model could be valid for use at the point of care.

Our findings should be viewed in the context of several limitations. We relied on registry data, which has certain limitations. Data may be entered incompletely or inconsistently between and within sites. Further, we were limited by the variables available in the CFF PR. The registry currently does not capture measures of 6-minute walk test, pulmonary hypertension, or blood gas (PaO2, PaCO2), all potentially important predictors of mortality.[55-57] Given the time period modeled, the only CFTR modulator included in our models was Ivacaftor, which was only available during the study period for patients with a G551D gating mutation, a small population. Recently approved CFTR modulators will cover the majority of patients with CF and are likely to significantly alter the trajectory of respiratory function in CF.[58, 59] Models will need to be updated as the use of CFTR modulators increases.

ML methods improved performance for all strata of educational attainment, a proxy for SES, but all models, including $ppFEV_1$, had lower discrimination accuracy for those with lower educational attainment. Including education and other proxies of SES in models did not change overall model performance, nor did it reduce the disparities in model performance between levels of educational attainment. SES-related disparities in CF care have been previously documented. Proxies of low SES are significant predictors of non-referral for LTx,[39, 40] and factors associated with lower SES, including distance from transplant center and poor social support, can be barriers to LTx.[42, 60] Further research is needed to explore the potential impact of a risk prediction model on SES-related disparities in care and outcomes in CF.

Additional work is needed to consider the patient health outcomes of decision-making under each model, as misclassification or miscalibration does not have uniform clinical consequences. Miscalibration among the highest risk patients may be inconsequential, as predicted risk for these patients is likely to far exceed the threshold for referral decision-making. By comparison, miscalibration near the decision threshold could have a larger impact on long-term patient outcomes. Finally, while ML models perform better than $ppFEV_1$, actual referral practices are heterogenous, with many clinicians considering factors other than $ppFEV_1$.[28, 39, 40] Therefore, it remains unclear whether ML models would perform better than observed clinical decision-making.

2.5 Conclusion

ML-based models had higher discrimination accuracy than $ppFEV_1$ alone for the prediction of short-term mortality in adults with CF in the US. ML models performed similarly, but AUC was highest for super learner. Additional work is needed to consider the expected patient health outcomes associated with referral decision-making that relies on ML-based risk prediction models.

Chapter 3. A FRAMEWORK FOR USING REAL-WORLD DATA AND HEALTH OUTCOMES MODELLING TO EVALUATE MACHINE-LEARNING BASED RISK PREDICTION MODELS

3.1 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, 7, 10, 28, 39] 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 if an RPM were used for decision-making in clinical practice.[61-66] While improvements in predictive accuracy are necessary for consideration of novel RPMs, prediction accuracy alone is not sufficient for assessing real-world clinical utility because it does not account for the complex clinical context in which the model would be used. Additional consideration is needed for real-world factors that impact RPM utility 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 to a reference model - an existing risk prediction model, biomarker or clinical guidelines.[67, 68] However, real-world clinical practice often deviates from the reference model. Clinical decision-making is frequently heterogenous, with different clinicians weighing different factors in decisions, including various pieces of evidence, historical experience, and preferences.[28, 66, 69, 70] Clinicians also have access to information not included in the risk prediction model, including expensive tests available for only a subset of patients and subjective clinical impressions. In such cases, it remains unclear whether new models that demonstrate superior performance relative to a reference model would also demonstrate superior performance compared to usual care (UC).

Furthermore, studies rarely relate changes in the discrimination and calibration properties of a predictive model to changes in downstream patient outcomes.[2, 71] Some approaches have been proposed, such as considering the balance of false positives and false negatives at a given threshold.[72, 73] However, 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 cases of heterogenous treatment effects, discrimination accuracy will fail to capture the expected patient impacts, for example, that a model which better identifies cases with *large* treatment benefits offers higher clinical utility than a model that better identifies cases with *smaller* treatment benefits.

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 when used in clinical practice.

We selected lung transplant (LTx) referral decisions in cystic fibrosis (CF) as a case study for three primary reasons. First, the standard predictor of short-term mortality in CF, forced expiratory volume in one second (FEV_1), has low positive predictive value, and[15-20] we previously developed an ML-based RPM with better discrimination accuracy and calibration (Chapter 2). Second, UC for referral decision-making is heterogenous, so performance improvements relative to FEV_1 may not be indicative of performance relative to UC.[28, 39, 40] Third, the relationship between clinical decisions and patient outcomes is complex given limited transplant availability[74] and heterogenous transplant benefits,[26, 75, 76] so additional consideration of downstream outcomes is needed.

3.2 METHODS

3.2.1 *Data*

We used RWD from the CF Foundation Patient Registry (CFF PR), which collects longitudinal, observational data for all US patients seen at CFF-accredited care centers who consent to participate.[33] 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.[33] The CFF PR covers approximately 80% the US CF population, and includes 95% of clinic visits and 90% of hospitalizations for participating patients.[33] Our cohort included CFF PR adults (≥18 years old) who had not undergone LTx by January 1, 2012, and had at least one 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 sets (40%) sets to develop and evaluate the ML model. The 40% validation set (n = 4,247) was used in this microsimulation.

CFF PR data was linked to United Network for Organ Sharing (UNOS) data, which contains additional waitlist, transplant, and post-transplant information for patients who were listed for LTx. UNOS data also contains detailed information on donated organs. The data linkage was performed by a team at the University of Washington in collaboration with a team from the University of Toronto.[34, 77] This study was approved by the University of Washington Institutional Review Board (Study #2270), by St Michael's Hospital, Toronto, Ontario, Canada (Research Ethics Board #14-148) and the Seattle Children's Research Institute (Study #PIROSTUDY15294).

3.2.2 *Microsimulation Model structure*

We developed a microsimulation model with 5 mutually exclusive health states: pre-referral, evaluation, waitlist, transplanted, and dead (**Figure 5**). Patients began in the state corresponding with their status on January 1, 2012: pre-referral, evaluation, or waitlist. Patients transitioned from pre-referral to evaluation at their time referral, which varied between ML, FEV_1, and UC policies. Evaluation, modeled as a tunnel state, represents the time between referral and waitlisting when patients are evaluated for LTx. Patients surviving evaluation transitioned to the waitlist, where they remained until they were matched with an organ for transplant or died before transplant. Finally, transplanted patients remained in the transplanted state until their simulated time of post-transplant death. Transitions between states were determined by individual-specific transition probabilities that rely on RWD, described in detail in the following sections. We used a cycle time of 1 day and a time horizon of 5 years. Modeling was conducted in R.[52]

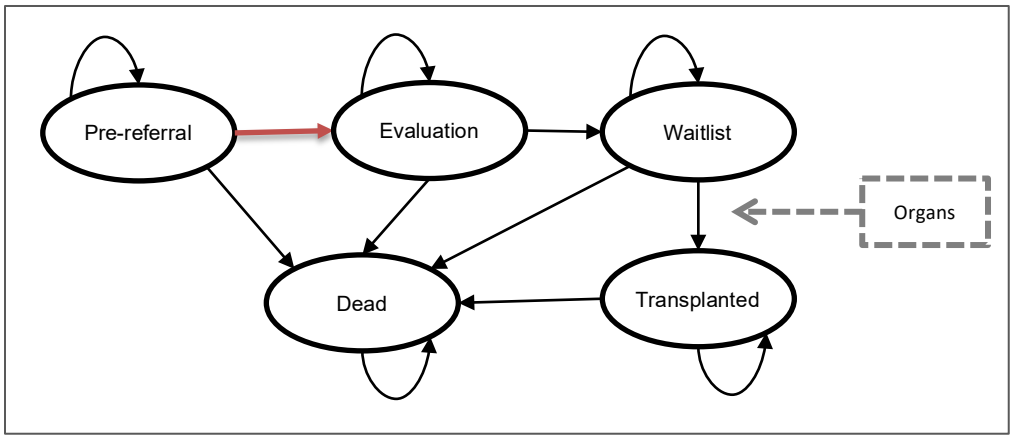


Figure 5: Microsimulation Model. Microsimulation model with 5 mutually exclusive health states: pre-referral, evaluation, waitlist, transplanted, and dead. Patients waitlisted before model start begin in the waitlist state, all other patients begin in pre-referral. A patient moves from pre-referral to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

3.2.3 *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) usual care. 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 combines multiple underlying models.[51, 53] ML had higher discrimination at baseline and over time and better calibration than FEV_1 (Chapter 2).

For ML-based referral decision-making in our microsimulation, we applied a decision rule that reflects the model's likely use in clinical practice. We assumed referral would occur at the first clinic visit where a patient's predicted ML risk exceeds the threshold corresponding to 95% model specificity at baseline. This rule was selected because it matches the specificity of the common FEV_1 <30% criteria.[19] However, any alternative decision rule could be considered, including differing levels of specificity or more complex rules, such as multiple visits or time periods when criteria are met. For FEV_1-based policies, referral occurred at the first clinic visit with a stable FEV_1 < 30% predicted based on the Global Lung Initiative equations for percent predicted.[78] For UC, the referral time was determined used RWD. We describe referral in detail in section 3.2.6.1 below.

3.2.4 *Simulation Population*

RWD often contains numerous longitudinal measures, whose relationships over time are unknown. Our RWD, for example, contains correlated, longitudinal information on patients' visit times, lung function, other health factors, predicted ML risk, and pre-LTx survival. Simulating a dataset that preserves the complex underlying relationships between these variables would be extremely challenging. As an alternative to a simulated population for estimating microsimulation uncertainty, we use the approach of plasmode simulation, where resampled populations ("plasmodes") are drawn from observed data.[79, 80] Unmodified, observed data for the resampled population are combined with modeling to simulate unknown elements. For example, resampled patients could retain their characteristics and longitudinal exposures documented in RWD, while outcomes are simulated using a model. This flexible approach is useful for preserving the complex underlying relationships between the potentially hundreds of variables in RWD.[79] However, it is also more computationally intensive than a completely simulated population.

In each of 1,000 microsimulation model runs, we drew 1 resampled population from observed cohort data. Each resampled patient retained their true, observed covariate history up to the time of transplant, including visit history, ML risk scores, pulmonary function, and pre-transplant survival. Transplant timing and post-transplant outcomes were then simulated, using models described below. We also used modeling to synthetically extend patients' pre-transplant covariate history in cases where their actual time of transplant occurred earlier than it would have under an alternative policy (i.e. when pre-transplant covariate history is censored by transplant).[81] Resampled versus modeled elements are summarized in Appendix F.

3.2.5 *Outcomes*

For each referral policy, we estimated the following outcomes: 5-year overall survival (the sum of time spent in all non-death states), number of 5-year pre-transplant deaths, and number of post-transplant deaths.

3.2.6 ***State Transitions***

3.2.6.1 Referral (Pre-referral → evaluation)

3.2.6.1.1 Model-based policies

For the ML-based policy, LTx referral time was determined using each patient's risk predictions and absolute contraindications to transplant over time. ML risk predictions for each patient were obtained at every pre-LTx clinic visit from model start to model end, using the information collected up to that visit. This approach mirrors true clinical practice, where decisions are made repeatedly at visits over time, using the most up-to-date information.[54] ML-based referral occurred at the first clinic visit where risk exceeded the threshold corresponding to 95% model specificity and no absolute contraindications to LTx (*mycobacterium abscessus* and *burkholderia cenocepacia*) were present. **Figure 6** provides an example of ML and FEV_1 referral under each model for one patient.

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

In instances where patients actually received LTx, observed pre-transplant visit history and ML risk predictions were censored at the patient's observed time of transplant, L_i. Under an alternative policy, a patient's time of referral may not have occurred by time L_i, when pre-transplant history was censored. In this case, we synthetically extended the risk trajectory beyond L_i.[81] We assumed that clinic visits for individual *i*, which represented opportunities for referral, would continue at the same average frequency observed in the 12 months prior to L_i and generated synthetic visits by sampling from a patient-specific normal distribution of visit intervals (days between visits). We separately accounted for pre-transplant deaths (i.e. that an individual may not survive until the synthetic visit times) and truncated synthetic visits at the time of pre-transplant death (described in section 2.5.4 below). Because ML risk predictions relied on a large number of time-varying covariates that were not observable after time L_i, we could not calculate risk directly at these synthetic visits. We instead estimated risk predictions at synthetic visits using a linear mixed effects model. Additional detail and an example are given in Appendix F.

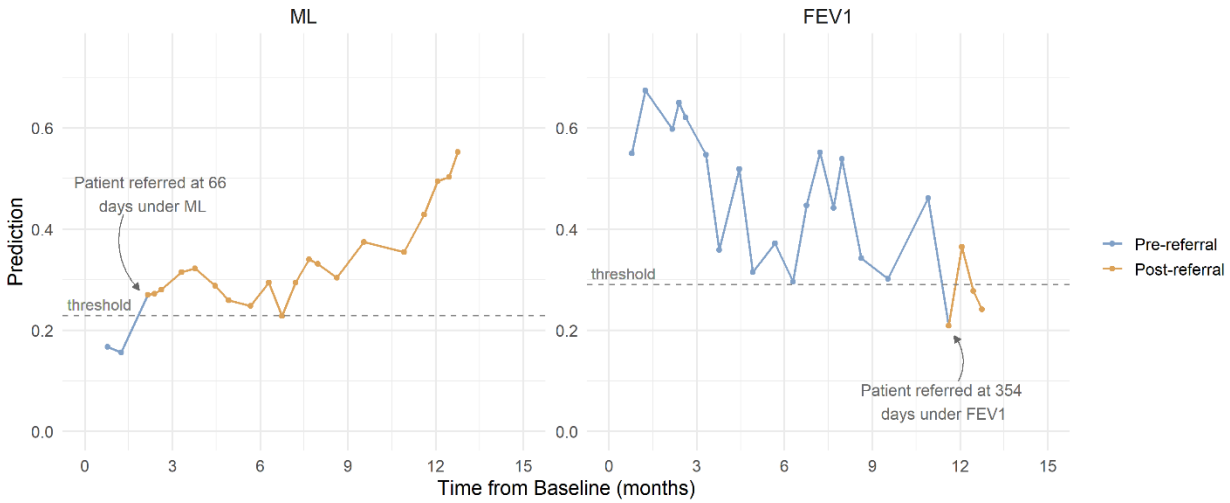


Figure 6: Patient trajectory and referral example. Risk of 2-year mortality from the ML model and FEV1 % predicted for an example patient at each clinic visit. For the ML model, a patient is referred at the first visit where risk exceeds the threshold, denoted by a change in line color. For FEV1, referral occurs at the first visit where FEV1 is lower than 30%, denoted by a change in line color.

3.2.6.1.2 Usual Care

Exact referral dates observed in clinical practice are not recorded. However, categorical transplant status ("not pertinent", "accepted, on the waitlist", "evaluated, final decision pending", "evaluated, rejected", and "had transplantation") is recorded annually in the CFF PR. We used the first year with a status other than "not pertinent" as the UC referral year for each patient. To obtain a more granular date, we defined a subset of visits where referral could have occurred for each patient: clinic visits in the referral calendar year and, if listing occurred in the same calendar year, before the listing date. We randomly selected one of these visits as the patient's UC referral visit within each simulation run. We validated the resulting simulated UC listing time against the observed listing times.

3.2.6.2 Listing (Evaluation → Waitlist)

After a patient is referred for transplant, a rigorous evaluation occurs at the lung transplant center to assess whether the patient is a suitable candidate for transplantation. This includes evaluation of the patient's health, as well as their medical adherence, emotional well-being, social support, and finances.[74] In some cases, a barrier is identified that must be addressed before listing can occur. For example, patients must live within a certain distance of the transplant center and may need to move before they can be listed. Addressing barriers can result in long times from referral to listing

for a subset of patients. Unfortunately, evaluation times are not captured in our data. We attempted to simulate a realistic time from referral to listing that approximates available estimates[40, 82] by sampling evaluation times from a marginal truncated normal distribution with mean of 4.5 months, standard deviation of 4 months, and minimum of 3 weeks. Patients' evaluation time was held constant between policies for each simulation run, but varied between simulation runs.

3.2.6.3 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 three components: the waitlist of patients (defined above), a flow of organs for transplantation, and a policy for matching organs to patients.[81]

3.2.6.3.1 Organ Flow

We relied on historical organ data from UNOS to define a flow of organs available for transplantation. For each simulation run, we sampled from the average annual number of organs available, N_o, and their characteristics (ABO, height) using organs matched to patients in our cohort for the years 2012 - 2016. Organ dates of availability were sampled from a uniform distribution, where all dates were equally likely. Our model created a flow of organs that reflected characteristics of historically observed organ donations to CF patients, but allowed for variability in the exact number of organs and dates of availability.

3.2.6.3.2 Organ Matching

As of 2005, lung allocation policy prioritizes patients based on lung allocation score (LAS), a measure of expected mortality with and without transplant.[83] LAS aims to identify patients with both an urgent need and a large expected survival benefit of transplant. LAS is calculated daily to prioritize patients on the waitlist. Observed LAS measures for waitlisted patients were available in UNOS data for each active day on the waitlist. However, alternative policies sometimes resulted in earlier waitlisting and/or the waitlisting of patients not listed under UC, such that LAS values were not available for all patients or at all necessary time points. We therefore imputed LAS at all timepoints for all patients who would be waitlisted under any policy using a linear mixed effects model. We relied on LAS components that are measured in the CFF PR, and thus available for all patients regardless of listing status (age, FVC, BMI, diabetes). Large changes in LAS are

frequently observed in the days or weeks preceding death or LTx, as patients experience sudden changes in lung function. To capture such changes in LAS in the absence of measures underlying the change, we included a fixed and random effect indicator for whether the patient experienced a death or transplant in the next 30 days. Additional detail is given in Appendix F.

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. While donors and recipients should have similar height (a proxy for lung capacity) no fixed thresholds exist for acceptable donor-recipient height differences.[84] 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 used relative rather than absolute differences in donor-recipient height, as oversize and undersize organs have different acceptability and outcomes.[84]

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

3.2.6.4 Pre-transplant and post-transplant survival

Patients in the pre-referral, evaluation, and waitlist states were at risk of pre-transplant 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 pre-transplant time of death, $(T_i \mid LTx = 0)$, unless LTx occurred first. Similarly, patients who survived for the full five-year period retained their observed pre-transplant survival, $(T_i \mid LTx = 0) > 5$ years, unless LTx occurred first.

Among patients with observed LTx, pre-transplant survival was censored at the observed time of transplant, L_i. In such cases, $(T_i \mid LTx = 0)$ was unknown, but known to be 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 two potential outcomes at any time: (1) survival with no transplant at time *t*, and survival with transplant at time *t*. Only one outcome can be realized for each patient, but information from other 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. That is, two patients with the same LAS had the same propensity for treatment.

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. LAS was centered at 40. We adjusted for gender, age at waitlisting, BMI 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 Appendix F.

Estimating expected pre- and post-transplant residual survival for simulation: For patients with observed transplant whose pre-transplant survival was censored at Li, we obtained their expected time of death in the absence of transplant, conditional on survival and history up to L_i, $(T_i(t) \mid LTx = 0, T_i > t, X_i(t))$. At $t = L_i$, we use the inverse sampling method to obtain $T_i(t)$ if the patient was not transplanted at L_i.

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

where $U \sim Uni(0,1)$, β is a vector of coefficients from the exponential survival model and $X_i(t)$ is a vector of covariate values for individual i at time t, with the transplant indicator set to 0. Additional detail is given in Appendix F.

When considering post-transplant survival, a patient's time of transplant under each policy, $L_{i,p}*$ is determined in each simulation run. This time may vary between policies. For example, a patient could be transplanted at $t = 100$ days under ML and $t = 150$ days under FEV_1. At each of these potential transplant times, the patient could have a different LAS value, which suggests that their expected post-transplant survival differs. Using the same example, if the patient's lung function declined substantially from 100 to 150 days, their expected post-transplant survival may be lower when transplant is given at 150 vs. 100 days. As with pre-transplant survival, we similarly estimated an individual's expected time of death at each $t = L_{i,p}*$ using the inverse sampling method . For post-transplant survival, we set the transplant indicator at t to 1.

3.3 RESULTS

3.3.1 *Validation*

Simulated results for UC were compared to observed data for validation. Among patients listed for LTx, the simulated UC listing date was a median of 8 days earlier than the observed listing date (IQR = 102 days earlier, 157 days later). In observed data, 466 patients died without transplant, compared to 411 (367, 459) in the simulated UC. There were 309 transplants and 65 post-transplant deaths in observed data within the 5-year period, compared to 287 (241, 325) transplants and 41 (24, 59) post-transplant deaths in our simulated UC.

3.3.2 *Clinical Decisions*

Most patients remained too healthy for referral in the 5-year period, regardless of policy. Only 12.4% (11.4%, 13.4%) of patients were referred for LTx under UC (Table 5). 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. While characteristics were not significantly different, average FEV_1 percent predicted at the time of referral was somewhat higher for ML, 31.5% predicted (30.9, 32.2) than UC, 30.9% predicted (29.8, 32) (Table 5). 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. 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%).

Table 5: Patient Characteristics at Time of Referral and Transplant, by Policy. Mean (95% CI) at the time of referral and transplant by policy.

	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)

3.3.3 ***Patient Outcomes***

3.3.3.1 Transplantation

Despite higher referral rates, there was no difference in overall transplantation rates among policies due to real-world constraints in organ supply (Table 5). State membership over time (Figure 7) 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 waitlist, but no change in patients transplanted. At a population level, 0.39 (95% CI: 0.30, 0.44) years (of 5), on average, were spent on the waitlist under ML, compared to 0.41 (95% CI: 0.36, 0.45) under FEV_1 and 0.23 (95% CI: 0.20, 0.27) under UC.

Patient characteristics at the time of LTx were similar among policies (Table 5). While confidence intervals overlapped, patients transplanted under ML were slightly older and had slightly higher LAS at the time of transplant, compared to UC. As a measure, higher LAS is intended to indicate a higher expected short-term benefit of LTx.

While characteristics at the time of transplant were similar overall, the specific patients who received transplant and experienced pre-transplant death differed among policies (Figure 8). Under UC, 309 (277, 341) pre-transplant deaths occurred among patients who were never referred for LTx. Approximately 20.1% of these pre-transplant deaths were averted under ML because patients

were referred and transplanted. However, this was offset by fewer transplants and more pre-transplant deaths among those who received transplant under UC (Figure 8).

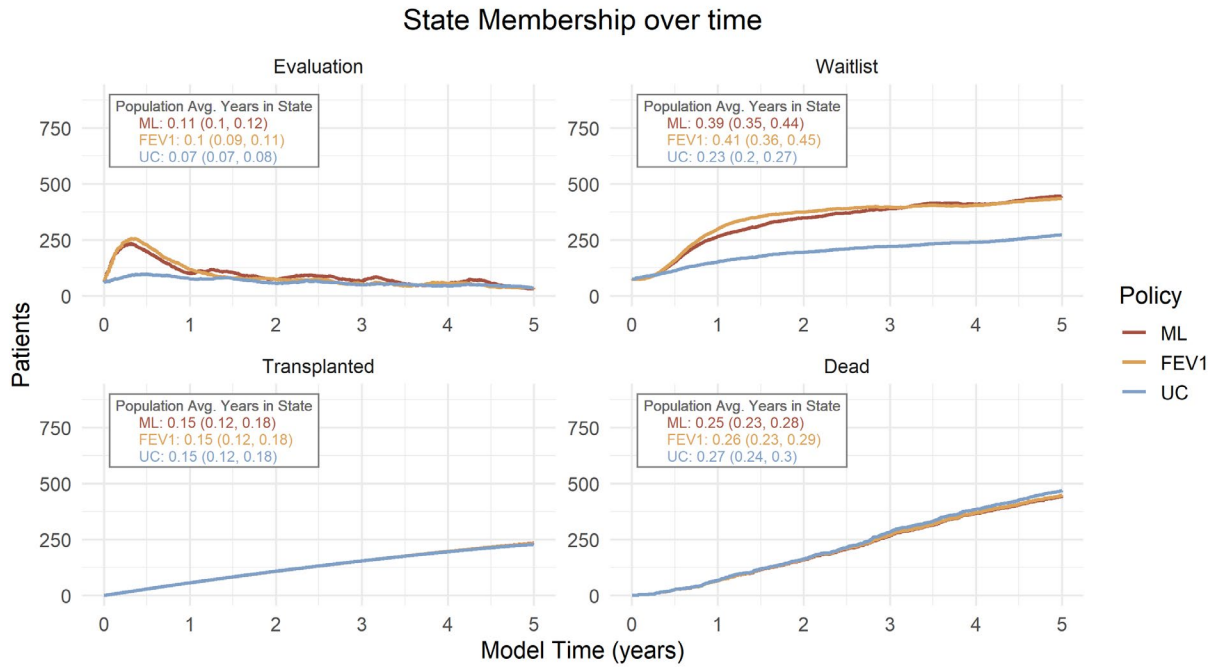


Figure 7: State membership over time, by policy: The average number of patients in each state except pre-referral for the 5-year time horizon. Population average years spent in each state (95% CI) is shown.

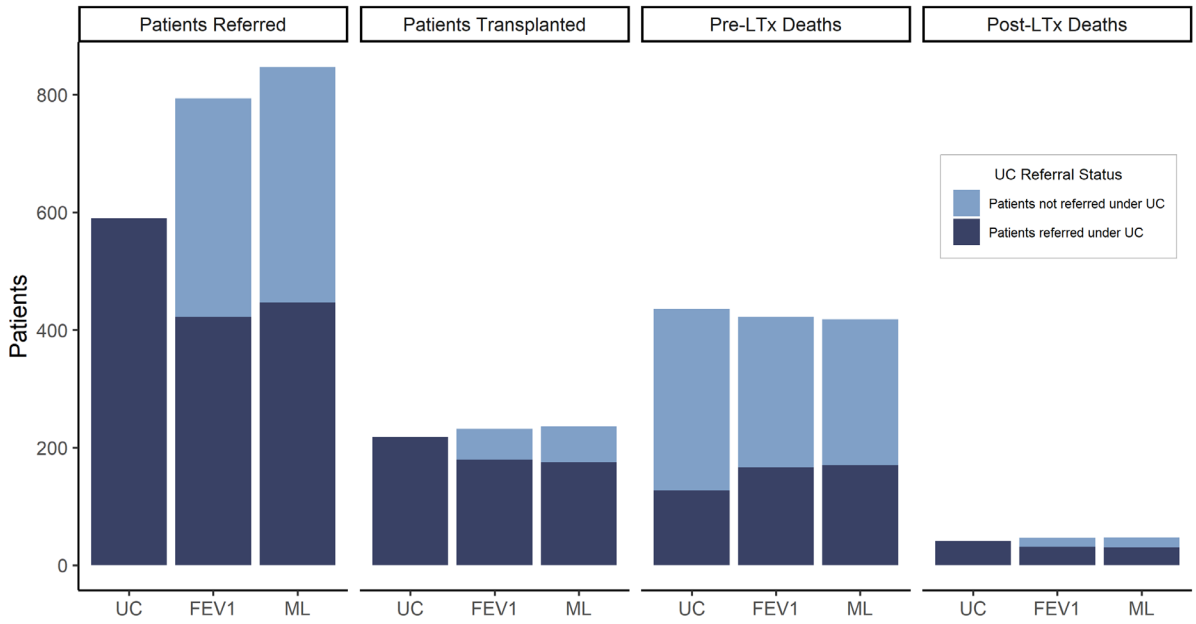


Figure 8: Patient Referral and Outcome, by UC Referral Status. Average number of referrals, transplantations, and pre-LTx deaths given by referral status under UC.

3.3.3.2 Survival

At a population level, these changes resulted in no significant differences in overall 5-year survival, pre-transplant deaths, or post-transplant deaths (Table 6). Overall 5-year survival was approximately 4.7 years under all polices.

Table 6: Expected Outcomes, by Policy

Policy	Model AUC at Baseline*	Pre-transplant 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)

3.4 DISCUSSION

We demonstrated an application of microsimulation 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 for the ML model did not yield differences in expected downstream patient outcomes when used for clinical decision-making. While the ML model did lead to changes in the number of patients referred and the timing of referral, real-world constraints on treatment availability limited the extent to which referral decisions could influence treatment.

We found a significant difference between the clinical decisions expected under FEV_1 alone, the reference model, and those observed in clinical practice. While 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,[28, 39] comparisons to FEV_1 are standard for new models in CF.[18-20]

Our work suggests that additional comparisons to UC are needed to assess model performance. While any new RPM must predict better than an existing RPM to add value, improvements relative to a reference model may be poor proxy for real-world clinical utility when clinical decision-making is heterogenous. In such cases, 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 versus those following usual care.[63, 65, 85, 86] Such studies are expensive and typically undertaken as a final step before implementation.[69] In contrast, our approach uses real-world data 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, including our own. While simulation-based evaluation does not capture the complex ways that clinicians interact with models to make decisions,[69, 70] it can be used as a first step for demonstrating clinical utility before conducting randomized controlled trials.

The use of health outcomes modelling to evaluate new RPMs remains minimal.[2] Statistical approaches for assessing clinical utility based on measures of discrimination accuracy have gained relatively more popularity,[72, 73, 87] 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.[56, 74, 88] Given limited data on the evaluation process, we assumed a marginal distribution of evaluation time with no rejection for listing. This may not accurately reflect the patient-specific factors that lead to longer evaluation times or rejection for listing, including proxies of lower socioeconomic status. However, median simulated listing time was within 10 days of observed listing times, suggesting this assumption was acceptable on average. We conservatively assumed a fixed average organ supply available to CF patients. In reality, the share of all organs allocated to CF patients could increase, as CF patients tend to have better outcomes than patients with other native lung diseases.[76] Finally, we are unable to distinguish between clinician decision-making and patient preferences using RWD. Lower rates of referral under the UC may represent patient preferences for non-referral, rather than clinician decisions not to refer patients. These complexities can be measured through impact evaluation.

3.5 CONCLUSION

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 due to constraints in organs available for transplantation. While constraints in transplant availability are unique to the organ allocation setting, complex real-world factors that impact 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.

Chapter 4. CONCLUSIONS

Machine learning methods with real-world data provide the promise to tackle historically challenging prediction problems and enable accurate, personalized clinical decision-making. By combining flexible ML methods with big, observational data sources, researchers across clinical applications have developed more accurate models for predicting outcomes.[3-6] Yet, despite thousands of publications on ML/AI model development, there are fewer than 100 publications on the use of ML/AI in real-world clinical practice.[10] Better evidence on the clinical utility offered by novel, ML-based RPMs may support greater translation to clinical practice.

This dissertation was an exploration of the development and evaluation of a novel, ML-based RPM in CF. The need for better short-term mortality predictions in CF has been widely noted.[89] A patient registry with a large number of longitudinal measures for the majority of the CF population provided an ideal, real-world data source for ML-based RPM development. We used novel, longitudinal feature engineering methods to remove noise from key predictors and considered a range of ML approaches, which varied in flexibility and interpretability. We leveraged a stacking ensemble approach, Super Learner, to optimally combine all models. While the Super learner ML model lacks interpretability, it predicted 2-year mortality outcomes in adults with CF better than FEV_1 alone. It also spread risk predictions over a wider and more accurate range. Finally, when predictions were updated dynamically at each clinic visit, reflecting likely clinic use, the ML model maintained high predictive performance over time.

However, when applied to real-world data to simulate clinical decision-making and patient outcomes over time, we found that ML-based referral led to different clinical decisions but no change in downstream outcomes, when compared to true usual care. While more patients were referred for LTx using ML, a fixed supply of organs available for transplant meant no change in the number of patients transplanted. Furthermore, the specific individuals who received transplant under the ML-based policy didn't appear to be better transplant candidates. We found no difference in post-transplant deaths between those transplanted under ML-based referral versus UC.

Importantly, the decisions that occurred in usual care differed substantially from FEV_1, a commonly used, guideline-based reference model for new RPMs.[18-20] Our work suggests that performance relative to FEV_1 may be a poor proxy for performance relative to usual care. Considering the implications more broadly, the practice of comparing a new RPM to an existing biomarker, multivariate model, or guidelines is commonplace. Of course, a new RPM must predict better than an existing biomarker, multivariate model, or guidelines to add any value. However, our work suggests that these improvements are not enough. In practice, clinicians often consider a range of factors when making clinical decisions, which can result in decisions that deviate substantially from those suggested by the reference model.[28, 66, 69, 70] Comparisons to usual care are needed to assess performance of new RPMs relative to real-world clinical practice.

It remains challenging to address the interaction between clinicians, patients, and predictions using health outcomes modelling.[69, 70] Ideally, medical decisions are the result of shared decision-making, where clinicians and patients together use the information available to arrive at the optimal decision.[90] Unfortunately, while RWD allowed us to observe the clinical action or inaction (referral, treatment, etc) that occurred, we were unable to determine whether it represented clinician, patient, or shared decision-making. Additional research is needed on the complex interactions between clinicians, patients, and predictions.[91]

Evaluations of clinical utility are needed to identify and support the development of high-value ML models. Clinicians may be more willing or interested in ML adoption when stronger evidence is presented on its benefit in clinical practice.[92] Benefits measured in terms of decisions and patient outcomes are likely more intuitively interpretable than measures like AUC. These evaluations can help to realize the potential value of ML, where clinically useful models are actually applied in clinical practice to enable better decisions and outcomes for patients.

BIBLIOGRAPHY

1. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE clinical predictive model registry: update 1990 through 2015. Diagnostic and prognostic research. 2017;1(1):20.
2. van Giessen A, Peters J, Wilcher B, Hyde C, Moons C, de Wit A, et al. Systematic review of health economic impact evaluations of risk prediction models: stop developing, start evaluating. Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research. 2017;20(4):718-26.
3. Kruppa J, Ziegler A, König IR. Risk estimation and risk prediction using machine-learning methods. Human genetics. 2012;131(10):1639-54.
4. Chen JH, Asch SM. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. New England Journal of Medicine. 2017;376(26):2507-9.
5. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA. 2018;320(1):27-8.
6. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS one. 2017;12(4).
7. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC medicine. 2019;17(1):195.
8. Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. The Lancet Oncology. 2019;20(5):e262-e73.
9. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive care medicine. 2021:1-11.
10. Yin J, Ngiam KY, Teo HH. Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review. Journal of medical Internet research. 2021;23(4):e25759.
11. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. Nature Publishing Group; 2018.
12. Navarro CLA, Damen JA, Takada T, Nijman SW, Dhiman P, Ma J, et al. Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. BMJ open. 2020;10(11):e038832.
13. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine. 2019;25(1):30-6.
14. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. Jama. 2019;322(14):1351-2.
15. Kerem E, Reisman J, Corey M, Canny GJ, Levison H. Prediction of mortality in patients with cystic fibrosis. New England Journal of Medicine. 1992;326(18):1187-91.
16. Aaron SD, Chaparro C. Referral to lung transplantation--too little, too late. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2016;15(2):143-4.
17. Buzzetti R, Alicandro G, Minicucci L, Notarnicola S, Furnari ML, Giordano G, et al. Validation of a predictive survival model in Italian patients with cystic fibrosis. Journal of Cystic Fibrosis. 2012;11(1):24-9.
18. Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. American journal of epidemiology. 2001;153(4):345-52.
19. Mayer-Hamblett N, Rosenfeld M, Emerson J, Goss CH, Aitken ML. Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. American journal of respiratory and critical care medicine. 2002;166(12 Pt 1):1550-5.
20. Nkam L, Lambert J, Latouche A, Bellis G, Burgel P-R, Hocine M. A 3-year prognostic score for adults with cystic fibrosis. Journal of Cystic Fibrosis. 2017;16(6):702-8.

21. Ramos KJ, Smith PJ, McKone EF, Pilewski JM, Lucy A, Hempstead SE, et al. Lung transplant referral for individuals with cystic fibrosis: Cystic Fibrosis Foundation consensus guidelines. Journal of Cystic Fibrosis. 2019.
22. Elborn JS. Cystic fibrosis. The Lancet. 2016;388(10059):2519-31.
23. Stephenson AL, Sykes J, Stanojevic S, Quon BS, Marshall BC, Petren K, et al. Survival comparison of patients with cystic fibrosis in Canada and the United States: a population-based cohort study. Annals of internal medicine. 2017;166(8):537-46.
24. Stephenson AL, Stanojevic S, Sykes J, Burgel P-R. The changing epidemiology and demography of cystic fibrosis. La Presse Médicale. 2017;46(6):e87-e95.
25. Shweish O, Dronavalli G. Indications for lung transplant referral and listing. Journal of Thoracic Disease. 2019:S1708-S20.
26. Thabut G, Christie JD, Mal H, Fournier M, Brugière O, Leseche G, et al. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. American journal of respiratory and critical care medicine. 2013;187(12):1335-40.
27. Bansal A, Mayer-Hamblett N, Goss CH, Chan LN, Heagerty PJ. A Novel Tool to Evaluate the Accuracy of Predicting Survival and Guiding Lung Transplantation in Cystic Fibrosis. Epidemiology (Sunnyvale, Calif). 2019;9(2).
28. 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-.
29. Ramos KJ, Quon BS, Heltshe SL, Mayer-Hamblett N, Lease ED, Aitken ML, et al. Heterogeneity in Survival in Adult Patients With Cystic Fibrosis With FEV1 < 30% of Predicted in the United States. Chest. 2017;151(6):1320-8.
30. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019;380(14):1347-58.
31. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920-30.
32. Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Scientific reports. 2018;8(1):11242.
33. Knapp EA, Fink AK, Goss CH, Sewall A, Ostrenga J, Dowd C, et al. The Cystic Fibrosis Foundation Patient Registry. Design and Methods of a National Observational Disease Registry. Annals of the American Thoracic Society. 2016;13(7):1173-9.
34. Stephenson AL, Ramos KJ, Sykes J, Ma X, Stanojevic S, Quon BS, et al. Bridging the survival gap in cystic fibrosis: An investigation of lung transplant outcomes in Canada and the United States. The Journal of Heart and Lung Transplantation. 2020.
35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Annals of internal medicine. 2015;162(10):735-6.
36. Maziarz M, Heagerty P, Cai T, Zheng Y. On longitudinal prediction with time-to-event outcome: Comparison of modeling options. Biometrics. 2017;73(1):83-93.
37. Heaton J, editor An empirical analysis of feature engineering for predictive modeling. SoutheastCon 2016; 2016: IEEE.
38. Schechter MS, Shelton BJ, Margolis PA, FitzSimmons SC. The association of socioeconomic status with outcomes in cystic fibrosis patients in the United States. American journal of respiratory and critical care medicine. 2001;163(6):1331-7.
39. Ramos KJ, Quon BS, Psoter KJ, Lease ED, Mayer-Hamblett N, Aitken ML, et al. Predictors of non-referral of patients with cystic fibrosis for lung transplant evaluation in the United States. Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society. 2016;15(2):196-203.
40. 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. The

Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation. 2017;36(3):264-71.
41. Quon BS, Psoter K, Mayer-Hamblett N, Aitken ML, Li CI, Goss CH. Disparities in access to lung transplantation for patients with cystic fibrosis by socioeconomic status. American journal of respiratory and critical care medicine. 2012;186(10):1008-13.
42. Lehr CJ, Fink AK, Skeans M, Faro A, Fernandez G, Dasenbrook E, et al. Impact of socioeconomic position on access to the US lung transplant waiting list in a matched cystic fibrosis cohort. Annals of the American Thoracic Society. 2020;17(11):1384-92.
43. Paulus JK, Kent DM. Race and Ethnicity: A Part of the Equation for Personalized Clinical Decision Making? Circulation: Cardiovascular Quality and Outcomes. 2017;10(7):e003823.
44. Paulus JK, Wessler BS, Lundquist CM, Kent DM. Effects of race are rarely included in clinical prediction models for cardiovascular disease. Journal of general internal medicine. 2018;33(9):1429-30.
45. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
46. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-20.
47. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.
48. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
49. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
50. Boser BE, Guyon IM, Vapnik VN, editors. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; 1992.
51. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical applications in genetics and molecular biology. 2007;6(1).
52. Team RC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
53. Polley E, LeDell E, Kennedy C, Laan Mvd. SuperLearner: Super Learner Prediction. 2019.
54. Bansal A, Heagerty PJ. A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making. Medical Decision Making. 2018;38(8):904-16.
55. Venuta F, Tonelli A, Anile M, Diso D, De TG, Ruberto F, et al. Pulmonary hypertension is associated with higher mortality in cystic fibrosis patients awaiting lung transplantation. The Journal of cardiovascular surgery. 2012.
56. Weill D. Lung transplantation: indications and contraindications. Journal of thoracic disease. 2018;10(7):4574.
57. Tuppin MP, Paratz JD, Chang AT, Seale HE, Walsh JR, Kermeeen FD, et al. Predictive utility of the 6-minute walk distance on survival in patients awaiting lung transplantation. The Journal of heart and lung transplantation. 2008;27(7):729-34.
58. Pettit RS, Fellner C. CFTR modulators for the treatment of cystic fibrosis. Pharmacy and Therapeutics. 2014;39(7):500.
59. Burgel P-R, Durieu I, Chiron R, Ramel S, Danner-Boucher I, Prevotat A, et al. Rapid Improvement After Starting Elexacaftor-tezacaftor-ivacaftor in Patients with Cystic Fibrosis and Advanced Pulmonary Disease. American journal of respiratory and critical care medicine. 2021(ja).
60. Barbour KA, Blumenthal JA, Palmer SM. Psychosocial issues in the assessment and management of patients undergoing lung transplantation. Chest. 2006;129(5):1367-74.
61. Dekker FW, Ramspek CL, van Diepen M. Con: Most clinical risk scores are useless. Nephrology Dialysis Transplantation. 2017;32(5):752-5.
62. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:l6927.

63. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. Bmj. 2009;338.
64. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2).
65. Reilly BM, Evans AT. Translating Clinical Research into Clinical Practice: Impact of Using Prediction Rules To Make Decisions. Annals of Internal Medicine. 2006;144(3):201-9.
66. Khalifa M, Magrabi F, Luxan BG. 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. Journal of medical Internet research. 2020;22(7):e15770.
67. Goto T, Camargo 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-e.
68. 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 digital medicine. 2020;3(1):1-9.
69. Kappen TH, Van Loon K, Kappen MA, Van Wolfswinkel L, Vergouwe Y, Van Klei WA, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. Journal of clinical epidemiology. 2016;70:136-45.
70. Bate L, Hutchinson A, Underhill J, Maskrey N. How clinical decisions are made. Br J Clin Pharmacol. 2012;74(4):614-20.
71. Siontis KC, Siontis GC, Contopoulos-Ioannidis DG, Ioannidis JP. Diagnostic tests often fail to lead to changes in patient outcomes. Journal of clinical epidemiology. 2014;67(6):612-21.
72. Vickers AJ, Cronin AM, editors. Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework. Seminars in oncology; 2010: Elsevier.
73. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making. 2006;26(6):565-74.
74. Mitchell AB, Glanville AR. Lung transplantation: a review of the optimal strategies for referral and patient selection. Therapeutic advances in respiratory disease. 2019;13:1753466619880078.
75. Vock DM, Tsiatis AA, Davidian M, Laber EB, Tsuang WM, Finlen Copeland CA, et al. Assessing the causal effect of organ transplantation on the distribution of residual lifetime. Biometrics. 2013;69(4):820-9.
76. Vock DM, Durheim MT, Tsuang WM, Copeland CAF, Tsiatis AA, Davidian M, et al. Survival benefit of lung transplantation in the modern era of lung allocation. Annals of the American Thoracic Society. 2017;14(2):172-81.
77. Ramos KJ, Sykes J, Stanojevic S, Ma X, Ostrenga JS, Fink A, et al. Survival and lung transplant outcomes for individuals with advanced cystic fibrosis lung disease in the United States and Canada: an analysis of national registries. Chest. 2021.
78. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, et al. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. Eur Respiratory Soc; 2012.
79. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Computational statistics & data analysis. 2014;72:219-26.
80. Vaughan LK, Divers J, Padilla MA, Redden DT, Tiwari HK, Pomp D, et al. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. Computational statistics & data analysis. 2009;53(5):1755-66.
81. Thompson D, Waisanen L, Wolfe R, Merion RM, McCullough K, Rodgers A. Simulating the allocation of organs for transplantation. Health Care Management Science. 2004;7(4):331-8.
82. Alkhateeb AA, Lease ED, Mancl LA, Chi DL. Untreated dental disease and lung transplant waitlist evaluation time for individuals with cystic fibrosis. Special Care in Dentistry. 2021.

83. Egan TM, Murray S, Bustami R, Shearon T, McCullough KP, Edwards L, et al. Development of the new lung allocation system in the United States. American Journal of Transplantation. 2006;6(5p2):1212-27.
84. Chambers DC, Yusen RD, Cherikh WS, Goldfarb SB, Kucheryavaya AY, Khusch K, 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. The Journal of Heart and Lung Transplantation. 2017;36(10):1047-59.
85. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
86. Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). BMC medical informatics and decision making. 2011;11(1):1-7.
87. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass). 2010;21(1):128.
88. Lynch III JP, Sayah DM, Belperio JA, Weigt SS, editors. Lung transplantation for cystic fibrosis: results, indications, complications, and controversies. Seminars in respiratory and critical care medicine; 2015: NIH Public Access.
89. Aaron SD, Chaparro C. Referral to lung transplantation- too little, too late. Journal of Cystic Fibrosis. 2016;15(2):143-4.
90. Elwyn G, Frosch D, Thomson R, Joseph-Williams N, Lloyd A, Kinnersley P, et al. Shared decision making: a model for clinical practice. Journal of general internal medicine. 2012;27(10):1361-7.
91. Stryckers M, Nagler EV, Van Biesen W. The need for accurate risk prediction models for road mapping, shared decision making and care planning for the elderly with advanced chronic kidney disease. prilozi. 2016;37(2-3):33-42.
92. Kann BH, Hosny A, Aerts HJ. Artificial intelligence for clinical oncology. Cancer Cell. 2021.
93. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017;77(1):1-17.
94. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 1201. 2020:1-4.
95. Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab-an S4 package for kernel methods in R. Journal of statistical software. 2004;11(9):1-20.
96. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1-22.
97. Patel MS, Kurtzman GW, Kannan S, Small DS, Morris A, Honeywell S, et al. Effect of an automated patient dashboard using active choice and peer comparison performance feedback to physicians on statin prescribing: the PRESCRIBE cluster randomized clinical trial. JAMA network open. 2018;1(3):e180818-e.

APPENDIX A: BLUP SUMMARY

METHODS

We used partly conditional models to (1) smooth biomarker trajectories over time, and (2) impute missing biomarker values. The biomarkers modeled were height, weight, ppFEV1, ppFVC, and FEV1FVC ratio. For each biomarker, we first fit a linear mixed effects model with fixed effects for age (centered at 18), age^2 and a random intercept and age slope using data from 2011-2016 on the training set.

$$\hat{y}_{ij} = \beta_0 + \beta_1 age_j + \beta_2 age_j^2 + \beta_3 male + u_{0i} + u_{1i} age_j + \epsilon_{ij}$$

where $\hat{y}_{ij}$ refers to the biomarker value $\epsilon(height, weight, ppFEV, ppFVC, FEV1FVC)$ for individual i at visit j. We then obtained fitted best linear unbiased predictions (BLUPs) for each individual i at each visit j by using biomarker values collected up to and including visit j. If the biomarker value was missing at the current visit, we used only the values collected prior to the current visit. In addition, we included a measure of 6 month change in our models by first calculating the BLUP for each biomarker at a synthetic visit exactly 6 months prior to the current visit, then calculating the change between the synthetic 6 month prior BLUP and the current visit. We also conducted a sensitivity analysis where we re-fit all ML models on a training set that included the BLUP residuals ($\hat{y}ij - y_{ij}$) for pulmonary function tests. If the biomarker value at the current visit was missing we used the most recent non-missing residual. Individuals with no non-missing residuals since 2011 were excluded from the sensitivity analysis. If the residuals represented important variables in the sensitivity analyses and/or the performance was significantly different, biomarker values may have been oversmoothed, representing a loss of valuable information.

RESULTS

Raw and smoothed (BLUP) biomarker values for 8 randomly selected individuals are presented in Figures A.1-A.5, with panel a containing raw values, including gaps where no biomarker value was recorded, and panel b containing smoothed BLUP values. As shown, BLUPs result in a substantial smoothing of the raw values.

Model baseline AUC for ML models that included BLUP residuals was similar to performance of the primary model (Table A.1). Calibration (Figure A.6) was also relatively unchanged. Residuals were not among the 20 most important variables for random forest (Figure A.7a), but were among the most important variables in XGBoost (Figure A.7b).

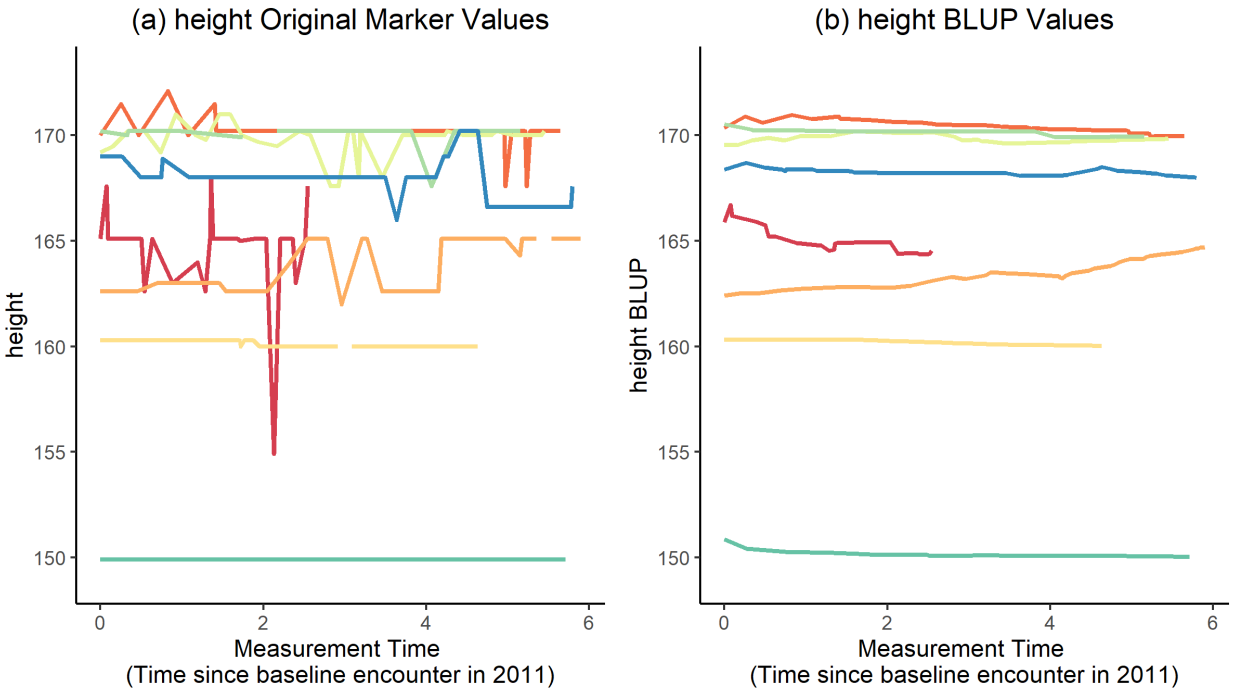


Figure A.1: Height BLUP

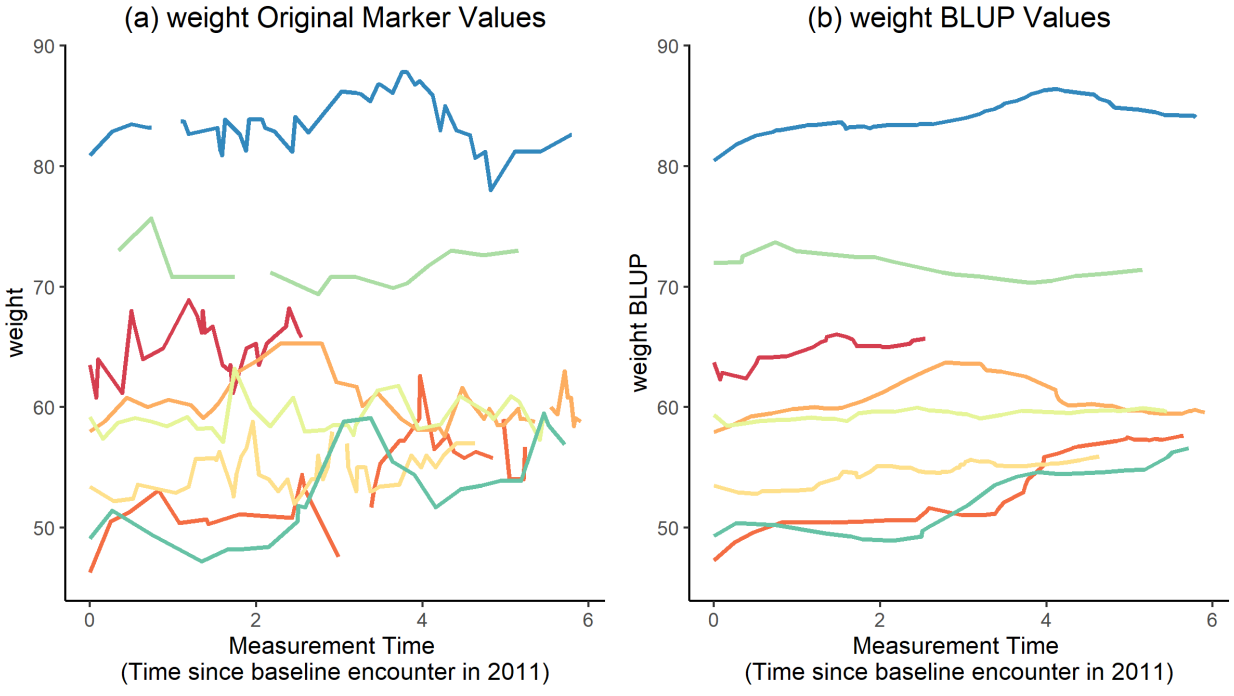


Figure A.2: Weight BLUP

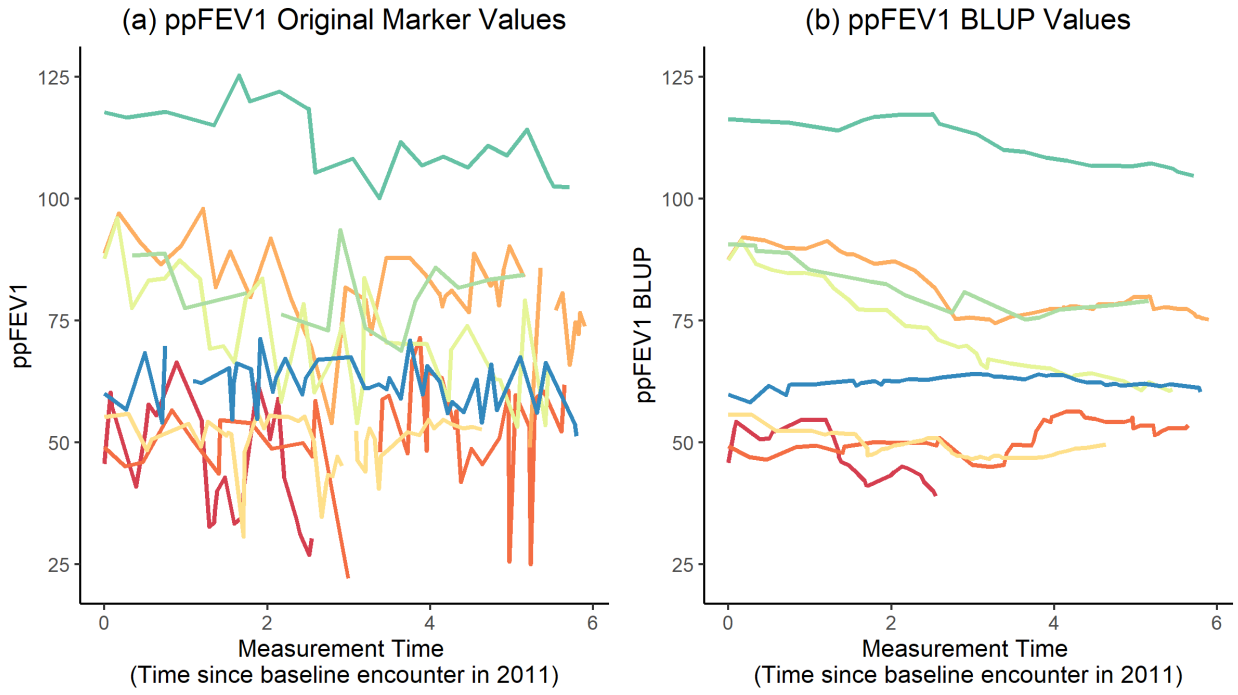


Figure A.3: ppFEV1 BLUP

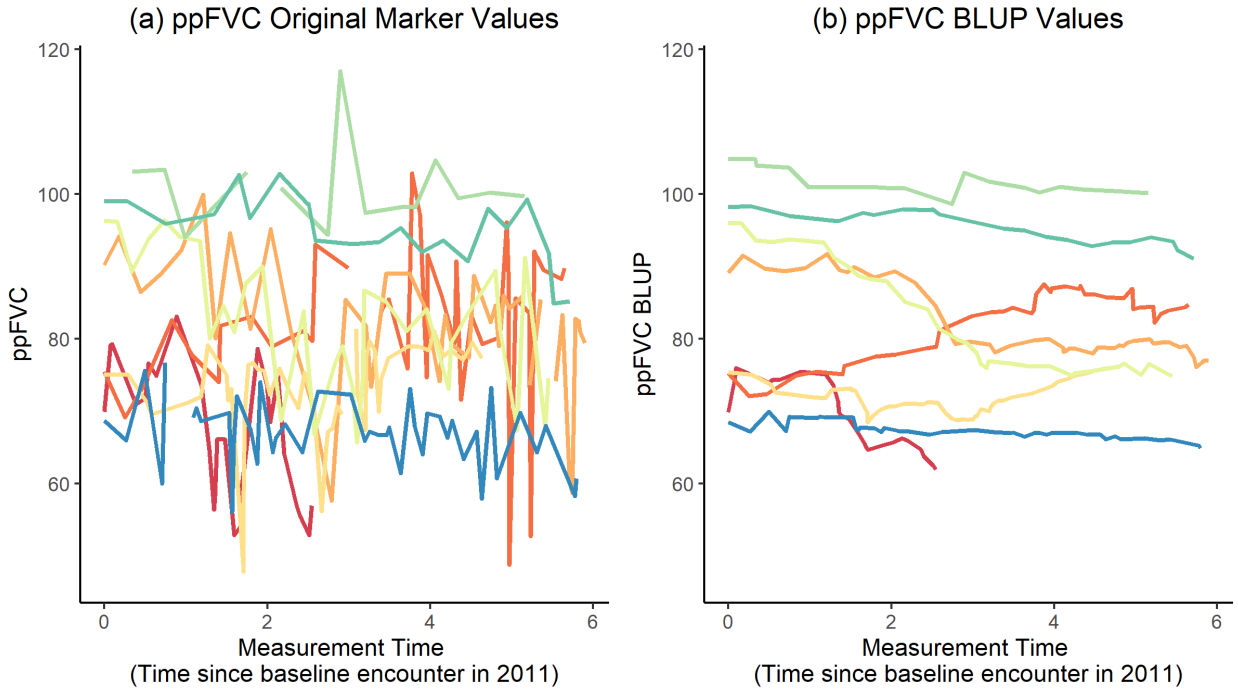


Figure A.4: ppFVC BLUP

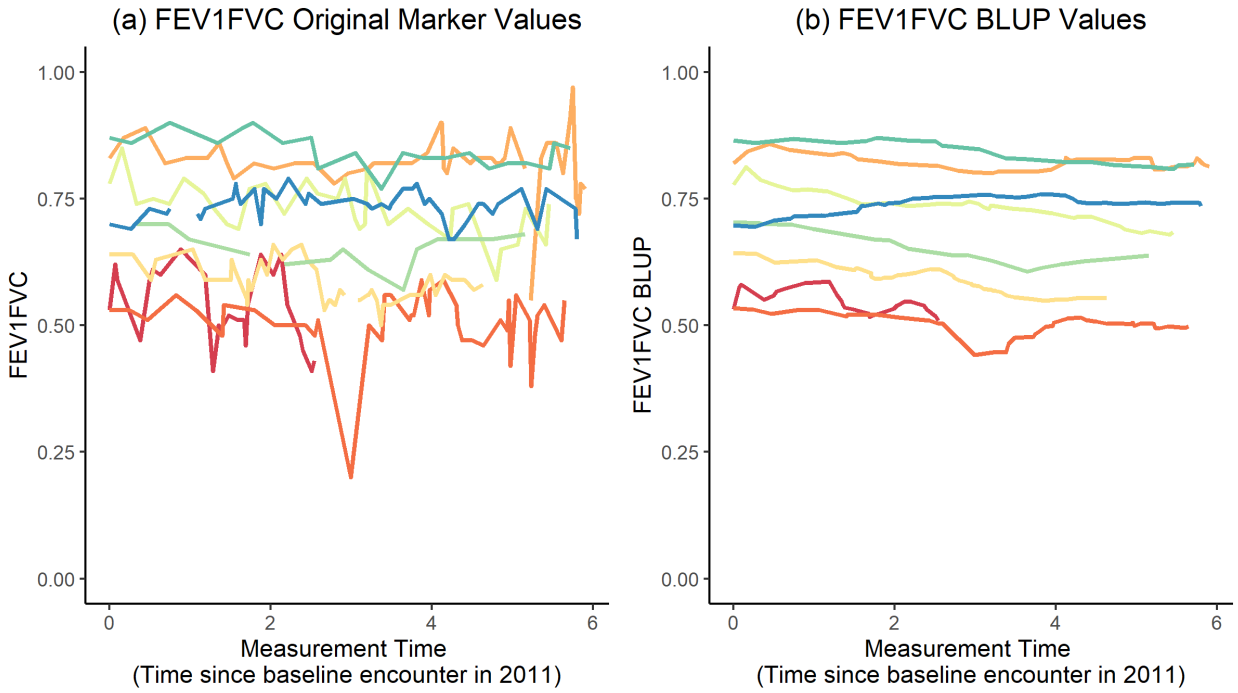


Figure A.5: FEV1FVC

Table A.1: Baseline AUC of Model with BLUP Residuals

Model	AUC (CI)
Super learner	0.91 (0.9,0.93)
Lasso	0.91 (0.9,0.93)
Random Forest	0.91 (0.89,0.93)
Elastic Net	0.91 (0.9,0.93)
Ridge	0.91 (0.89,0.93)
XGBoost	0.91 (0.89,0.93)
SVM	0.89 (0.87,0.91)
ppFEV1	0.88 (0.87,0.9)

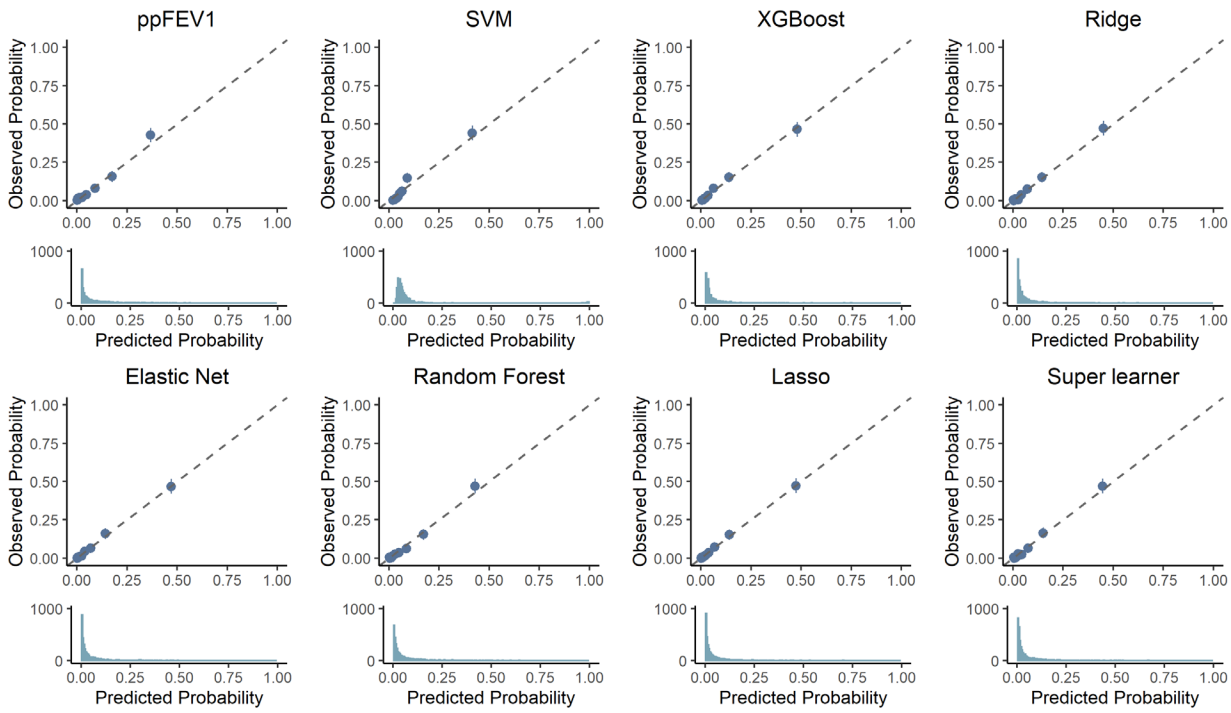


Figure A.6: Calibration Plots for Model with BLUP residuals

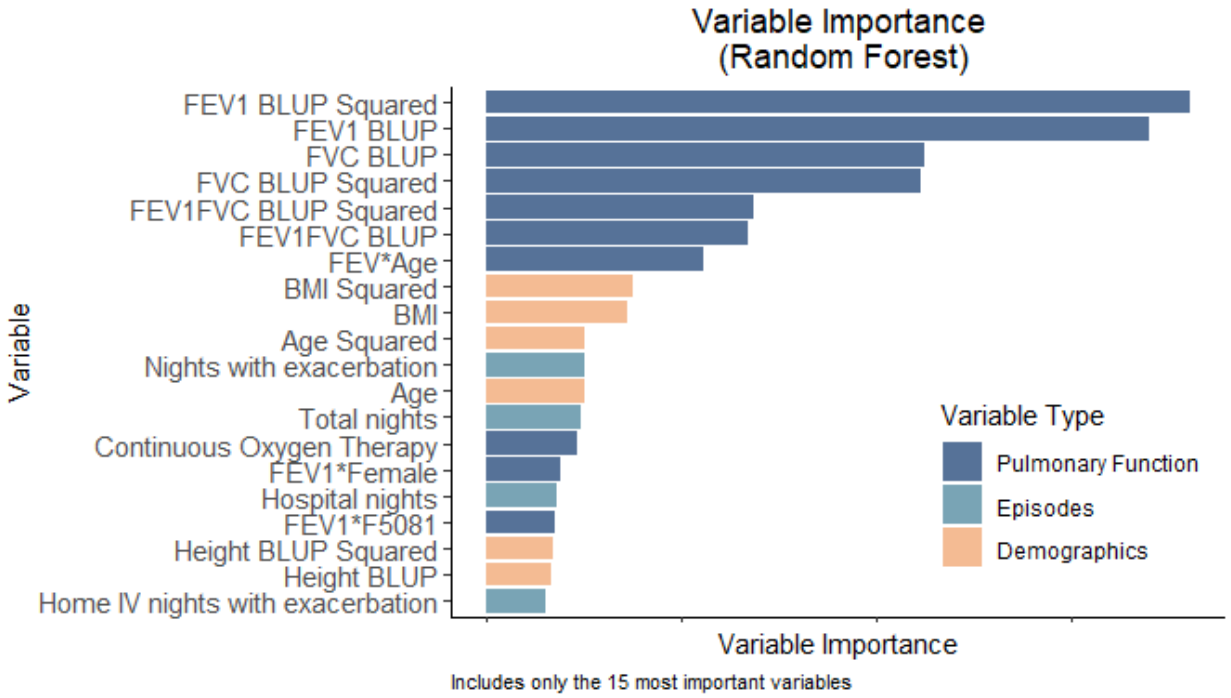


Figure A.7a: Random Forest Variable Importance

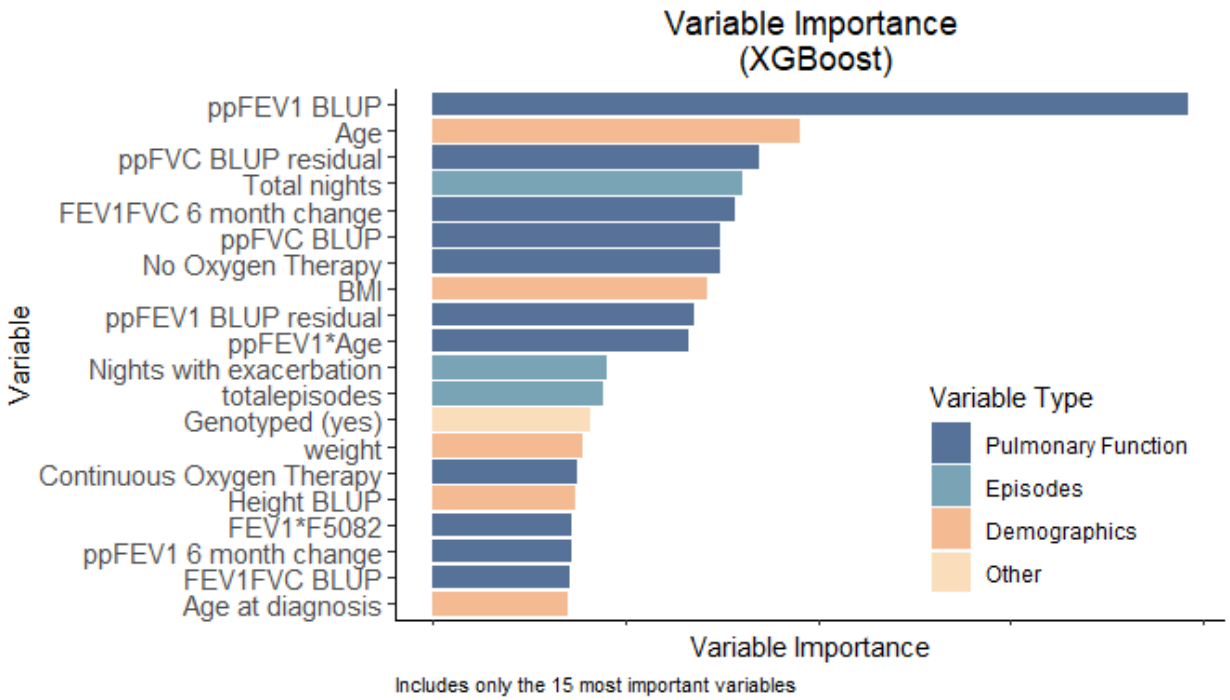


Figure A.7b: XGBoost Variable Importance

APPENDIX B: DETAILS ON MODEL FITTING AND TUNING PARAMETER CHOICES

We considered the following models: lasso, elastic net, ridge, random forest, extreme gradient boosting machine (XGBoost), support vector machine (SVM), and super learner. Super learner is a stacking ensemble, which uses the predictions from all base learners (i.e. all other models) to create an optimally weighted combination of predictions.[51, 53] For random forest, GBM, and SVM models, we included more than one variant, with different tuning parameter values. Random forest was fit using the ranger package, with multiple values of m, the number of variables considered at each split ($0.5\sqrt{p}, \sqrt{p}, 2\sqrt{p}$).[93] XGBoost used 3 values for depth (depth $\in$ (1,2,3)), 2 values for the number of trees (5000, 8000) and a learning rate, η, of 0.001.[94] SVM was fit using Kernlab, with a radial basis kernel and cost values of 0.1, 1, and 10.[95] For random forest, GBM, and SVM, we report results for a single variant, based on performance. Ridge, elastic net and lasso were fit using glmnet[96] with minimum λ values from 10-fold cross-validation. For elastic net, the mixing parameter, α, was set to 0.5, indicating an equal mix of lasso and ridge penalties. The super learner was selected through 10-fold cross-validation, optimizing AUC, using the SuperLearner package.[53]

APPENDIX C: BASELINE CHARACTERISTICS DISAGGREGATED BY OUTCOME

Table C.1: Baseline Characteristics disaggregated by outcome. Continuous variables reported as mean(sd). Categorical variables reported as n(%). Height, ppFEV1, ppFVC refer to smoothed BLUP values. BMI is calculated from smoothed height and weight BLUP values**.**

	Alive, pre-LTx in 2 Years (n=9,768)	Death in 2 Years (n=481)	LTx in 2 Years (n=366)	Overall (n=10,615)
Age	30.0 (10.6)	32.0 (11.9)	31.6 (9.70)	30.1 (10.6)
Female	4588 (47.0%)	262 (54.5%)	181 (49.5%)	5031 (47.4%)
Race				
White	9332 (95.5%)	447 (92.9%)	362 (98.9%)	10141 (95.5%)
Black	314 (3.2%)	25 (5.2%)	2 (0.5%)	341 (3.2%)
Hispanic	19 (0.2%)	3 (0.6%)	0 (0%)	22 (0.2%)
Other or Unknown	103 (1.1%)	6 (1.2%)	2 (0.5%)	111 (1.0%)
BMI	22.7 (3.93)	20.6 (4.02)	20.2 (4.43)	22.5 (4.00)
Height	168 (9.43)	164 (9.51)	167 (10.2)	168 (9.49)
Mutation Class				
1, 2, or 3	6960 (71.3%)	339 (70.5%)	283 (77.3%)	7582 (71.4%)
4 or 5	1084 (11.1%)	28 (5.8%)	18 (4.9%)	1130 (10.6%)
Other	155 (1.6%)	34 (7.1%)	20 (5.5%)	209 (2.0%)
Unknown	1569 (16.1%)	80 (16.6%)	45 (12.3%)	1694 (16.0%)
F508del Mutation				
Homozygous	4460 (45.7%)	226 (47.0%)	190 (51.9%)	4876 (45.9%)
Heterozygous	3933 (40.3%)	165 (34.3%)	132 (36.1%)	4230 (39.8%)
None	1235 (12.6%)	57 (11.9%)	24 (6.6%)	1316 (12.4%)
Unknown	140 (1.4%)	33 (6.9%)	20 (5.5%)	193 (1.8%)
ppFEV1	65.2 (21.9)	38.8 (16.9)	29.7 (11.2)	62.8 (23.0)
ppFVC	79.4 (18.4)	55.3 (17.5)	49.3 (12.8)	77.3 (19.6)
Oxygen				
Yes	1248 (12.8%)	305 (63.4%)	259 (70.8%)	1812 (17.1%)
No	8374 (85.7%)	174 (36.2%)	101 (27.6%)	8649 (81.5%)
Unknown	146 (1.5%)	2 (0.4%)	6 (1.6%)	154 (1.5%)
P. aeruginosa	5939 (60.8%)	356 (74.0%)	282 (77.0%)	6577 (62.0%)
S. aureus	5052 (51.7%)	223 (46.4%)	151 (41.3%)	5426 (51.1%)
MRSA	2309 (23.6%)	151 (31.4%)	129 (35.2%)	2589 (24.4%)
MSSA	3807 (39.0%)	130 (27.0%)	82 (22.4%)	4019 (37.9%)
Burkholderia complex	339 (3.5%)	37 (7.7%)	13 (3.6%)	389 (3.7%)
B. cenocepacia	49 (0.5%)	4 (0.8%)	1 (0.3%)	54 (0.5%)
Ivacaftor	114 (1.2%)	4 (0.8%)	2 (0.5%)	120 (1.1%)
Education				
High School	2749 (28.1%)	198 (41.2%)	89 (24.3%)	3036 (28.6%)
College	6040 (61.8%)	219 (45.5%)	232 (63.4%)	6491 (61.1%)
Unknown	979 (10.0%)	64 (13.3%)	45 (12.3%)	1088 (10.2%)
Insurance				
Private	6152 (63.0%)	173 (36.0%)	193 (52.7%)	6518 (61.4%)
Medicaid	2356 (24.1%)	238 (49.5%)	124 (33.9%)	2718 (25.6%)
Other	557 (5.7%)	22 (4.6%)	17 (4.6%)	596 (5.6%)
Medicare	496 (5.1%)	42 (8.7%)	32 (8.7%)	570 (5.4%)
Missing	207 (2.1%)	6 (1.2%)	0 (0%)	213 (2.0%)

APPENDIX D: VARIABLE IMPORTANCE

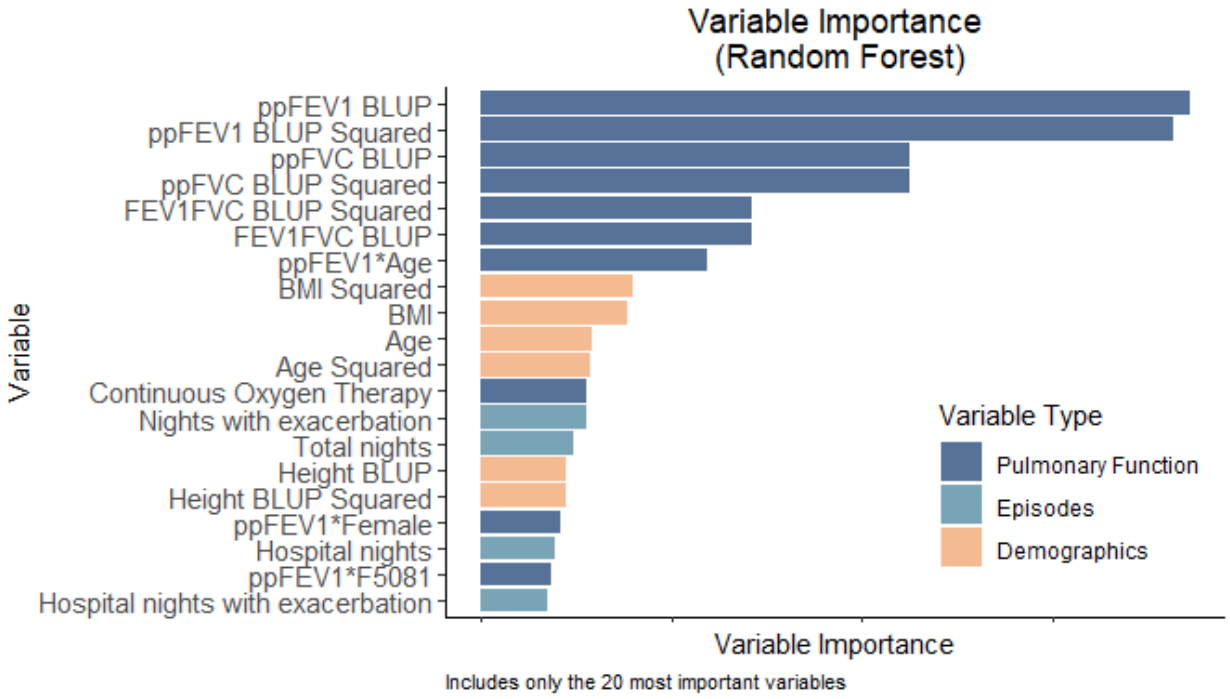


Figure D.1a: Random Forest Variable Importance

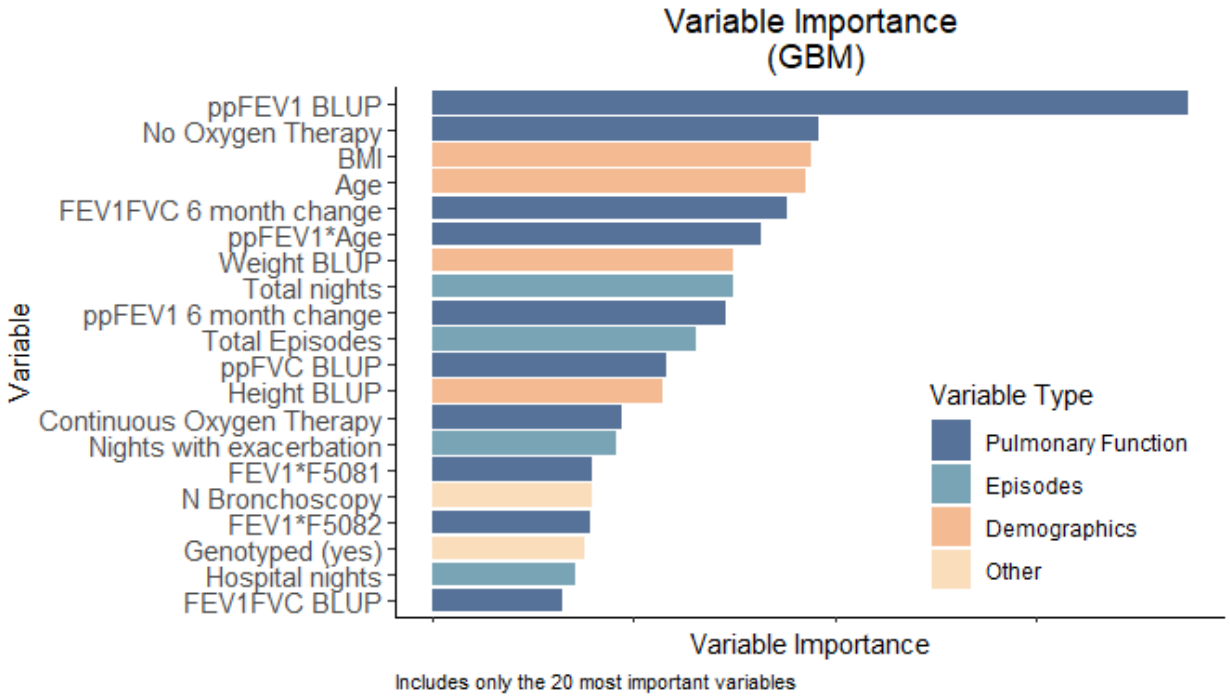


Figure D.1b: GBM Variable Importance

APPENDIX E: SES

METHODS

Several variables related to SES are available in the CFF PR, including race, insurance type, education, and patient assistance program. Consistent with previous models, these variables were not included in our primary analysis. In a secondary analysis, we re-fit all ML models on a dataset that included SES-related variables. We evaluated the discrimination accuracy and calibration of these secondary models, and compared to the results from our primary analysis. In addition, we compared AUC of the primary models and secondary models by levels of educational attainment. Due to the small number of events and patients of races other than white, we did not evaluate performance by race.

RESULTS

The AUC of ML models that included SES-related variables was not different from the AUC of the primary model at baseline or over time (Table D.2). Calibration was also similar to the primary model (Figure D.1). Variable importance measures for random forest and GBM show that SES-related variables did not rank highly. No SES-related variables appeared in the top 20 most important variables for either model (Figures D.2a and D.2b).

AUC results by educational attainment for the primary model and sensitivity analysis are shown in Table D.3. All ML models had the highest performance among those with at least some college education, 61% of patients, and lower performance among those with high school only or unknown educational attainment. ppFEV1 also had the lowest performance for those with high school education, but had similar performance among those with unknown and college education. The differential in AUC between subgroups with college and high school levels of education was similar for all models, including ppFEV1. However, ML models improved performance among all subgroups of educational attainment. The super learner AUC for the worst performing subgroup (high school education) was slightly higher than the ppFEV1 AUC for the best performing subgroup of ppFEV1, suggesting that performance gains are experienced across SES subgroups, but the disparity in performance between SES groups is not reduced with the use of ML models.

The secondary model that included SES-related variables in ML models showed minimal difference in performance by education level compared to the primary model.

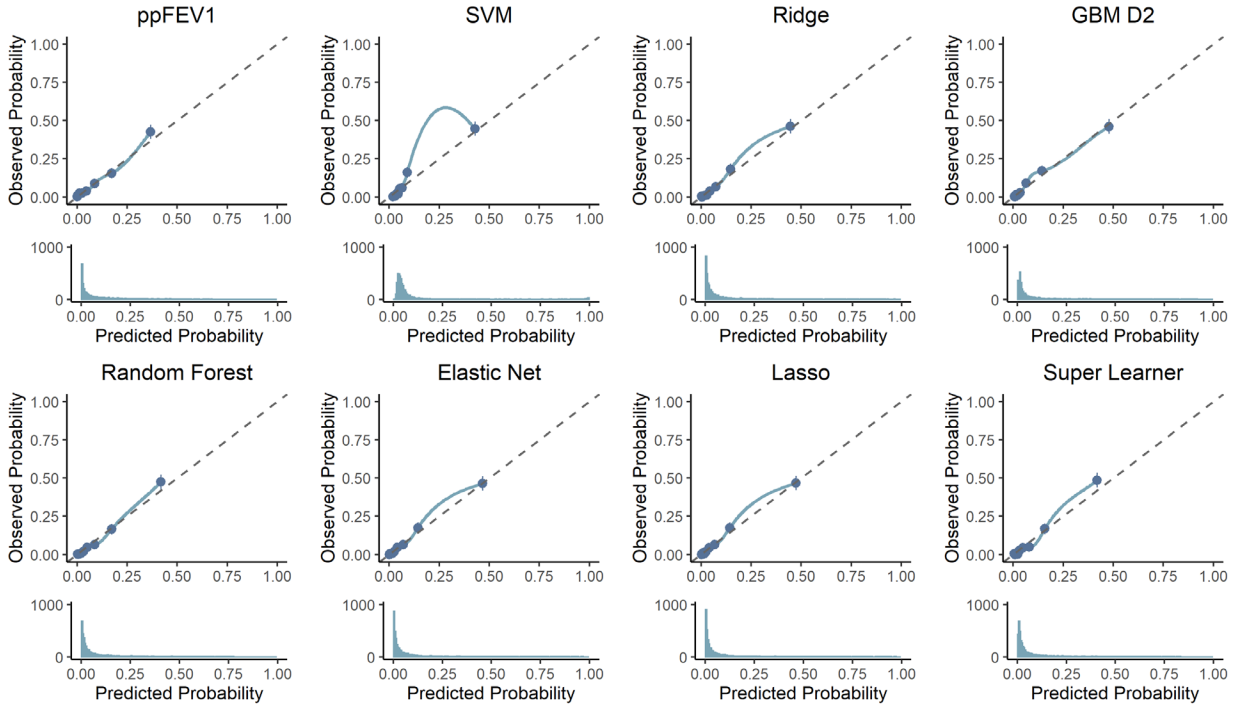
Table D.1: Baseline Characteristics by Educational Attainment. Continuous variables reported as mean(sd). Categorical variables reported as n(%). Height, ppFEV1, ppFVC refer to smoothed BLUP values. BMI is calculated from smoothed height and weight BLUP values.

	High School (n=3036)	College (n=6491)	Unknown (n=1088)	Overall (n=10615)
Age	27.4 (9.91)	30.7 (10.4)	34.5 (11.7)	30.1 (10.6)
Female	1308 (43.1%)	3218 (49.6%)	505 (46.4%)	5031 (47.4%)
Race				
White	2822 (93.0%)	6286 (96.8%)	1033 (94.9%)	10141 (95.5%)
Black	169 (5.6%)	136 (2.1%)	36 (3.3%)	341 (3.2%)
Other or Unknown	32 (1.1%)	63 (1.0%)	16 (1.5%)	111 (1.0%)
Hispanic	13 (0.4%)	6 (0.1%)	3 (0.3%)	22 (0.2%)
BMI	22.0 (4.16)	22.7 (3.77)	23.0 (4.69)	22.5 (4.00)
Height	167 (9.50)	168 (9.40)	167 (9.77)	168 (9.49)
Mutation class				
1,2, or 3	2202 (72.5%)	4668 (71.9%)	712 (65.4%)	7582 (71.4%)
4 or 5	247 (8.1%)	747 (11.5%)	136 (12.5%)	1130 (10.6%)
Other	76 (2.5%)	87 (1.3%)	46 (4.2%)	209 (2.0%)
Unknown	511 (16.8%)	989 (15.2%)	194 (17.8%)	1694 (16.0%)
F508				
Homozygous	1441 (47.5%)	2991 (46.1%)	444 (40.8%)	4876 (45.9%)
Heterozygous	1141 (37.6%)	2649 (40.8%)	440 (40.4%)	4230 (39.8%)
None	385 (12.7%)	769 (11.8%)	162 (14.9%)	1316 (12.4%)
Unknown	69 (2.3%)	82 (1.3%)	42 (3.9%)	193 (1.8%)
ppFEV1				
Mean (SD)	60.0 (23.1)	64.8 (22.7)	58.3 (22.7)	62.8 (23.0)
Imputed (n(%))	401 (13.2%)	651 (10.0%)	140 (12.9%)	1192 (11.2%)
ppFVC				
Mean (SD)	75.1 (20.2)	79.0 (19.2)	72.9 (19.6)	77.3 (19.6)
Imputed (n(%))	404 (13.3%)	656 (10.1%)	141 (13.0%)	1201 (11.3%)
Oxygen				
Yes	666 (21.9%)	971 (15.0%)	175 (16.1%)	1812 (17.1%)
No	2340 (77.1%)	5483 (84.5%)	826 (75.9%)	8649 (81.5%)
Unknown	30 (1.0%)	37 (0.6%)	87 (8.0%)	154 (1.5%)
P. aeruginosa	1788 (58.9%)	4120 (63.5%)	669 (61.5%)	6577 (62.0%)
S. aureus	1712 (56.4%)	3190 (49.1%)	524 (48.2%)	5426 (51.1%)
MRSA	880 (29.0%)	1475 (22.7%)	234 (21.5%)	2589 (24.4%)
MSSA	1194 (39.3%)	2436 (37.5%)	389 (35.8%)	4019 (37.9%)
Burkholderia complex	120 (4.0%)	246 (3.8%)	23 (2.1%)	389 (3.7%)
B.cenocepacia	17 (0.6%)	34 (0.5%)	3 (0.3%)	54 (0.5%)
Ivacaftor	19 (0.6%)	88 (1.4%)	13 (1.2%)	120 (1.1%)
Insurance				
Private	1093 (36.0%)	4815 (74.2%)	610 (56.1%)	6518 (61.4%)
Medicaid	1438 (47.4%)	1049 (16.2%)	231 (21.2%)	2718 (25.6%)
Other	230 (7.6%)	304 (4.7%)	62 (5.7%)	596 (5.6%)
Medicare	199 (6.6%)	260 (4.0%)	111 (10.2%)	570 (5.4%)
Missing	76 (2.5%)	63 (1.0%)	74 (6.8%)	213 (2.0%)
Outcome				
Alive in 2 years	2749 (90.5%)	6040 (93.1%)	979 (90.0%)	9768 (92.0%)
Death in 2 years	198 (6.5%)	219 (3.4%)	64 (5.9%)	481 (4.5%)
LTx in 2 years	89 (2.9%)	232 (3.6%)	45 (4.1%)	366 (3.4%)

Table D.2: AUC of Model with SES-related variables

learner	Baseline	6 Months	12 Months	18 Months	24 Months	30 Months	36 Months
Super Learner	0.91 (0.9, 0.93)	0.92 (0.9, 0.93)	0.93 (0.91, 0.94)	0.92 (0.91, 0.93)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.9 (0.88, 0.93)
Lasso	0.91 (0.9, 0.93)	0.92 (0.9, 0.93)	0.93 (0.91, 0.94)	0.92 (0.91, 0.93)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.9 (0.88, 0.93)
Elastic Net	0.91 (0.9, 0.93)	0.92 (0.9, 0.93)	0.93 (0.91, 0.94)	0.92 (0.91, 0.93)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.9 (0.87, 0.93)
Random Forest	0.91 (0.9, 0.93)	0.91 (0.9, 0.93)	0.92 (0.91, 0.94)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.9 (0.87, 0.93)
GBM	0.91 (0.89, 0.93)	0.92 (0.9, 0.93)	0.92 (0.91, 0.93)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.91 (0.9, 0.93)	0.9 (0.87, 0.93)
Ridge	0.91 (0.89, 0.92)	0.92 (0.9, 0.93)	0.93 (0.91, 0.94)	0.92 (0.9, 0.93)	0.92 (0.9, 0.93)	0.91 (0.9, 0.93)	0.9 (0.87, 0.93)
SVM	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.9 (0.88, 0.92)	0.9 (0.88, 0.92)	0.89 (0.88, 0.91)	0.89 (0.87, 0.91)	0.88 (0.85, 0.91)
ppFEV1	0.88 (0.86, 0.9)	0.88 (0.87, 0.9)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.89 (0.87, 0.9)	0.89 (0.87, 0.91)	0.89 (0.86, 0.92)

Figure D.1: Calibration Plots for Model with SES-related variables. For each model, the top panel compares observed versus predicted probability of 2-year death of LTx at baseline. The diagonal reference line indicates perfect calibration, where the observed probability of events (y-axis) equals the predicted probability of events (x-axis). Each decile of model-specific risk is represented as a point. Points that lie above the dashed line represent underprediction (observed probabilities are higher than predicted) and points below the dashed line represent overprediction (observed probabilities are lower than predicted). The bottom panel contains the distribution of predictions for each model. In all models, most patients had a low (<20%) predicted probability of death or LTx



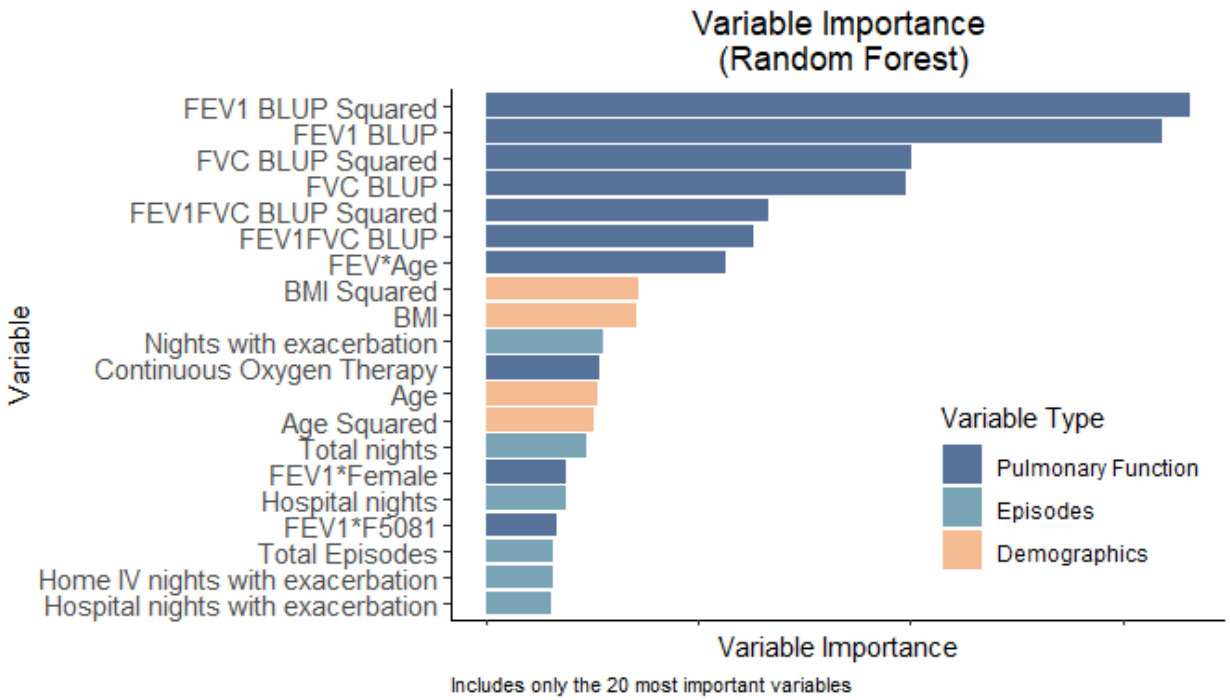


Figure D.2a: Random Forest Variable Importance with SES

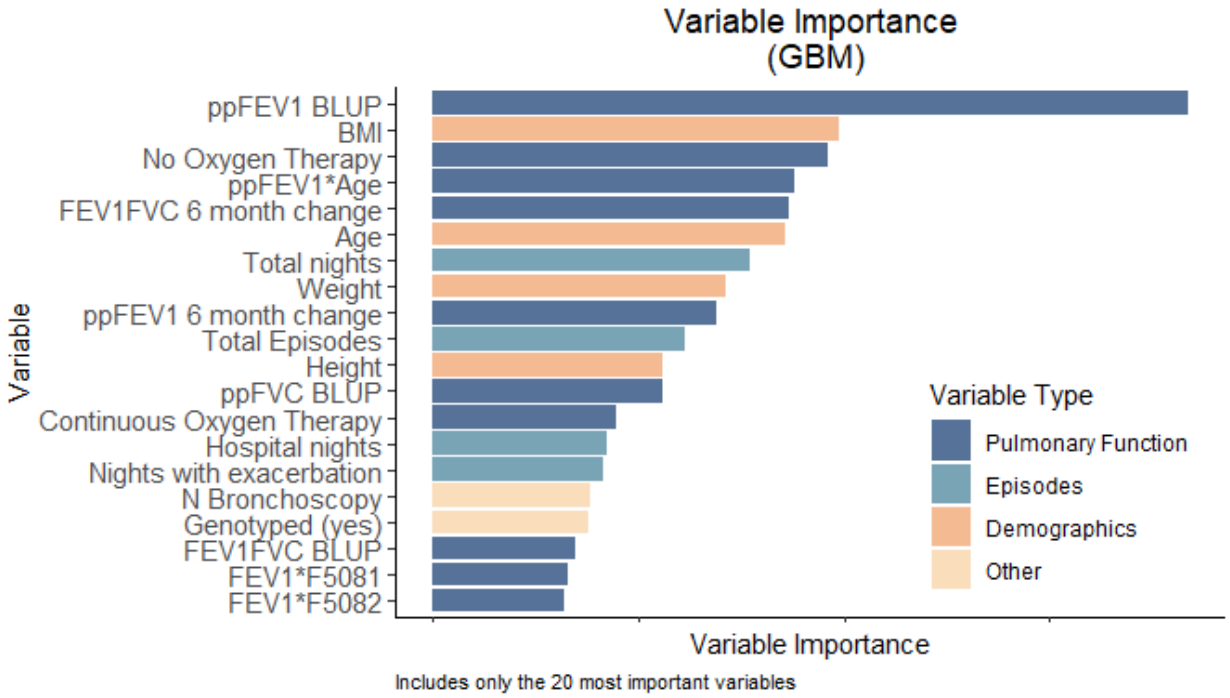


Figure D.2b: GBM Variable Importance with SES

Table D.3: Baseline AUC by Educational Attainment

Primary Model without SES Variables			
learner	**College**	**High School**	**Unknown**
Super Learner	0.93 (0.91, 0.95)	0.89 (0.86, 0.92)	0.9 (0.84, 0.96)
Lasso	0.93 (0.91, 0.95)	0.89 (0.86, 0.92)	0.9 (0.84, 0.95)
Elastic Net	0.93 (0.91, 0.95)	0.89 (0.86, 0.91)	0.9 (0.84, 0.95)
Random Forest	0.92 (0.9, 0.94)	0.88 (0.86, 0.91)	0.89 (0.83, 0.95)
GBM	0.93 (0.91, 0.94)	0.88 (0.86, 0.91)	0.9 (0.84, 0.96)
Ridge	0.92 (0.9, 0.94)	0.89 (0.86, 0.91)	0.9 (0.84, 0.96)
SVM	0.9 (0.87, 0.92)	0.86 (0.83, 0.9)	0.87 (0.81, 0.94)
ppFEV1	0.89 (0.86, 0.91)	0.84 (0.81, 0.88)	0.89 (0.84, 0.95)
Secondary Model with SES Variables			
learner	**College**	**High School**	**Unknown**
Super Learner	0.93 (0.91, 0.95)	0.89 (0.86, 0.92)	0.89 (0.83, 0.95)
Lasso	0.93 (0.91, 0.95)	0.89 (0.86, 0.91)	0.9 (0.84, 0.95)
Elastic Net	0.93 (0.91, 0.95)	0.89 (0.86, 0.91)	0.9 (0.84, 0.95)
Random Forest	0.93 (0.91, 0.95)	0.88 (0.86, 0.91)	0.89 (0.82, 0.95)
GBM	0.93 (0.91, 0.94)	0.88 (0.85, 0.91)	0.9 (0.84, 0.96)
Ridge	0.92 (0.9, 0.94)	0.88 (0.86, 0.91)	0.9 (0.84, 0.96)
SVM	0.9 (0.88, 0.93)	0.87 (0.84, 0.9)	0.87 (0.8, 0.94)
ppFEV1	0.89 (0.86, 0.91)	0.84 (0.81, 0.88)	0.89 (0.84, 0.95)

APPENDIX F: MICROSIMULATION MODEL DETAIL

1.0 REFERRAL

1.1 Model-based

For each individual i, at each clinic visit j, predicted risk of 2-year mortality for the ML model, $\hat{y}_{ij}$ is compared to a decision threshold corresponding to 95% model specificity at baseline, k, (0.229) which was identified in previous work. Referral for lung transplant evaluation for individual i under the ML model occurs at $R^*_{i,ML}$, the first visit where $\hat{y}_{ij} > k$.

We use a similar procedure for the reference model, FEV_1. FEV_{1ij} is obtained for each individual i at each clinic visit j. If FEV_1 was not measured at visit j, we used the FEV_1 BLUP, which is described in detail in our previous work. In short, we impute FEV_{1ij} using a linear mixed effects model that incorporates all FEV_1 values up to visit j. For lung transplant referral decisions, we apply a decision threshold of stable FEV_1 <30% predicted. We consider any FEV_1 value that did not occur during an exacerbation to be stable. For FEV_1, higher values indicate higher lung function and lower risk of death so referral, $R^*_{i,FEV}$, occurs when FEV_1 is *below* the decision threshold.

We refer to the simulated time of referral as R^*_{ip} where $p \in (ML, FEV)$. If individual i received a transplant, $\hat{y}_{ijm}$ and FEV_{1ij} are censored at the time of observed transplant, L_i. In some cases, R^*_{ip} may not occur before L_i and is therefore censored. In such cases, we need to synthetically extend $\hat{y}_{ijm}$ and FEV_{1ij} beyond L_i, assuming the patient was not transplanted, to determine R^*_{ip}. This requires both (1) an observation process - future visits for patient i, which represent opportunities for referral, and (2) marker trajectory ($\hat{y}_{ijm}$ and FEV_{1ij} for $t > L_i$).

According to guidelines, CF patients should have visits at a CF center at least quarterly. Unsurprisingly, observed data suggests that visits occur much more frequently as individuals become sicker, including in the time prior to transplant. In our microsimulation framework, this means that opportunities for referral should occur more frequently than once per quarter for

individuals who were sick enough to receive transplant. We therefore use individual-specific visit times for the synthetic observation process beyond L_i. We assume that individual i would continue to have visits at the same frequency observed in the 12 months prior to L_i. We calculate the visit intervals, $d_{i,j=2} \ldots d_{i,j=J}$ as number of days between each successive visit. We then create synthetic future visits by iteratively sampling from a normal distribution of d_i, $\mathcal{N}(\mu_i, \sigma_i^2)$, and adding the sampled, synthetic visit interval until model end or pre-transplant death. Note that time of pre-LTx death is incorporated through a separate process, such that the visit history for individual i will be truncated at their expected time of pre-transplant death.

We impute $\hat{y}_{ijm}$ and FEV_{1ij} at each synthetic future visit using linear mixed effects models,

$$\hat{y}_{ijm} = \beta_0 + \beta_1 age + \beta_2 time + b_0 + b_1 time, \text{ and}$$

$$FEV_{1ij} = \beta_0 + \beta_1 age + \beta_2 time + b_0 + b_1 time$$

where age refers to age at baseline, $time$ refers to time since baseline and b_0 and b_1 refer to individual random intercept and slope, respectively.

We provide an example of a risk trajectory with imputed values in Figure F.1.

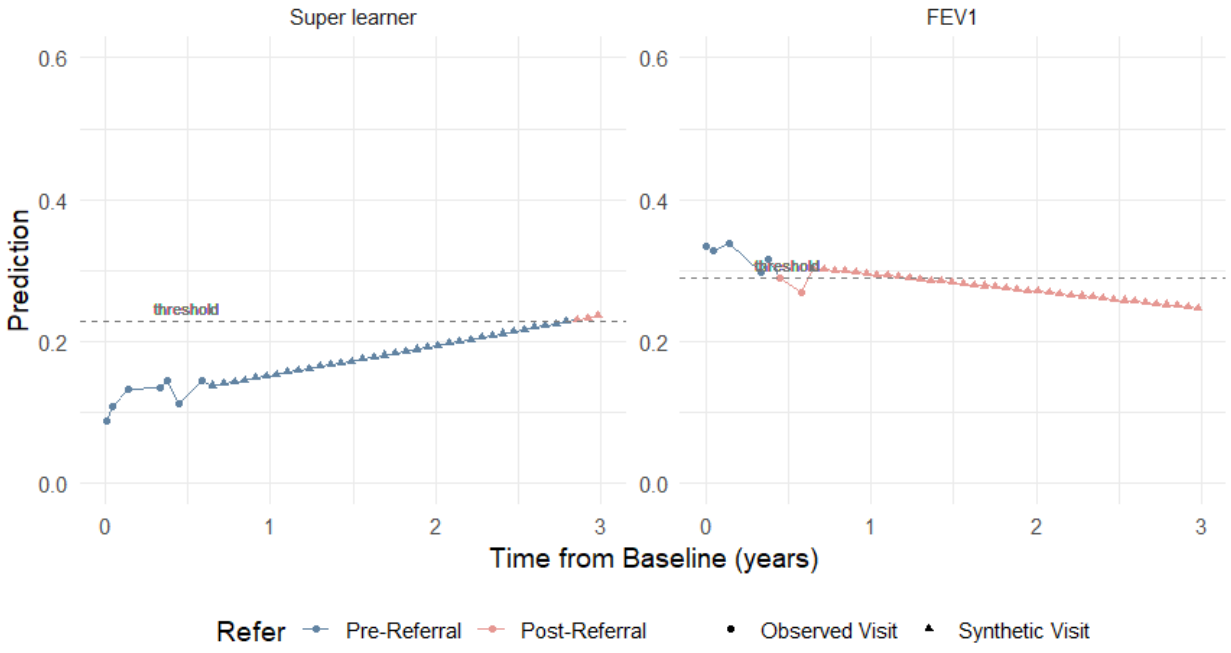


Figure F.1: Risk and FEV1 Trajectories for an Example Patient. Risk of 2-year mortality for the ML model, and FEV1 % predicted for an example patient at each clinic visit. Circles denote observed clinic

visits; triangles denote synthetic clinic visits, which are simulated for times beyond the patient's observed time of transplant, Li, using the patient's observed visit frequency in the 12 months preceding Li. Referral decisions occur at the decision threshold corresponding to 95% model specificity for ML and 30% predicted for FEV1, indicated by the dashed line. For ML models a patient is referred at the first visit where risk exceeds the threshold, denoted by a change in line color. For FEV1, referral occurs at the first visit where FEV1 is lower than the 30% predicted, denoted by a change in line color.

2.0 Evaluation

Limited data are available on time between referral and listing. A recent study on a small sample (N = 51 CF patients) at the University of Washington found an average evaluation time of 170.0 day (IQR: 86.0–252.0).[97] We simulated evaluation time to reflect this estimate using a truncated normal distribution with mean 4.5 months, standard deviation of 4 months, and minimum of 3 weeks (Figure F.2). Patients' evaluation time is held constant between policies for each simulation run, but varies between simulation runs. That is, a patient with a simulated evaluation time of 6 weeks on a given simulation run will have an evaluation time of 6 weeks for SoC and all model-based policies. However, the patient will have a different evaluation time under each simulation run.

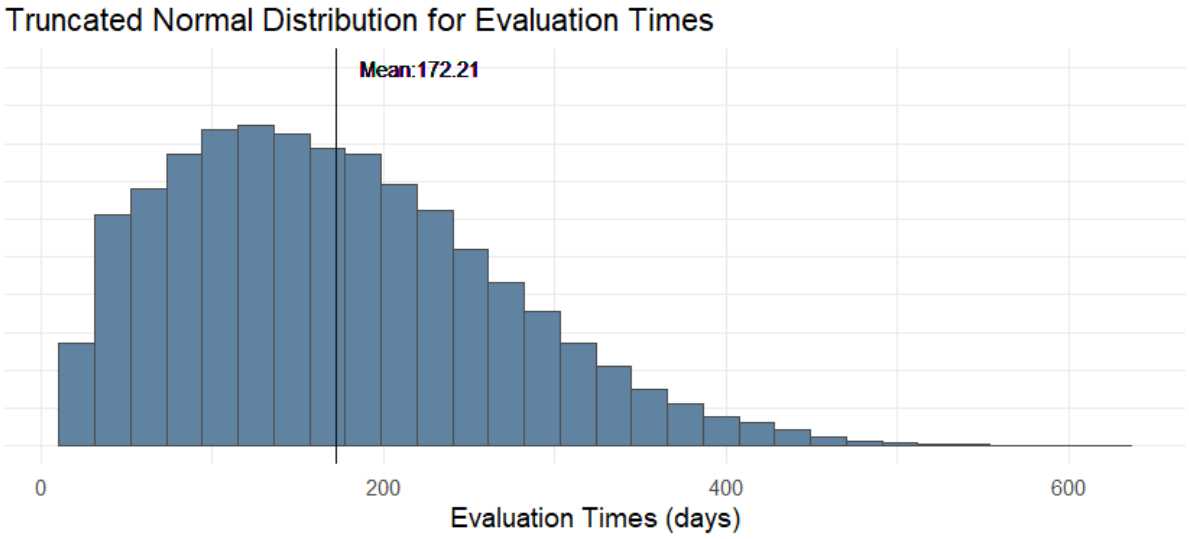


Figure F.2: Simulated distribution of evaluation times.

We validated the resulting listing times in our UC model against observed listing times. On average, patients are listed somewhat earlier in our model (median difference: 9 days). Differences may be attributable to the fact that we do not account for rejection. In observed data, some patients

may be rejected on first referral, but re-referred and listed at a later date. This creates an extremely long time from first referral to listing for a small subset of patients.

3.0 LUNG TRANSPLANT

3.1 Organs

We considered an organ 'population' of organs matched to patients in our cohort between 2012 and 2016. We estimated the annual number of organs available each year during this period. For each simulation run, we sample the number of organs available in each calendar year, N_o from a normal distribution. We sample N_o organ characteristics (height, ABO) from the full organ population and assign each sampled organ a date of availability from a uniform (1,365) distribution. There was no evidence of seasonality in organ donation in historical data (Figure F.3). In a secondary analysis, we considered a hypothetical scenario where twice as many organs are available annually.

Table F.1: Organ Characteristics Continuous variables presented as mean (95% CI). Count variables presented as n(%).

Characteristic	Value
BMI	25.9 (17.5,39.3)
Height	171.5 (154.8,188)
Male	209 (58.1%)
ABO A	141 (39.2%)
ABO AB	7 (1.9%)
ABO B	35 (9.7%)
ABO O	177 (49.2%)

For historical organs, we also calculate donor-recipient height difference as observed donor height - observed recipient height. We use the 2.5th and 95.5th percentile of the empirical donor-recipient height distribution as bounds for compatible heights in each simulation run (Figure F.4). This allows for some variability in acceptable height differences. The empirical distribution reflects the fact that oversize organs (donor taller than recipient) are more acceptable than undersize organs (donor smaller than recipient)

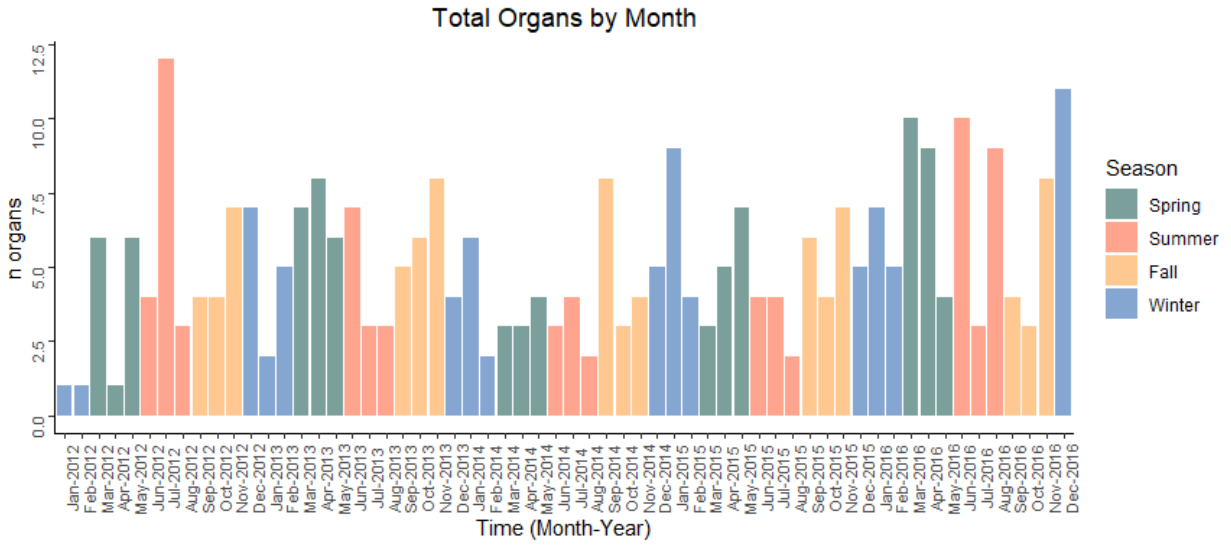


Figure F.3: Organ Seasonality

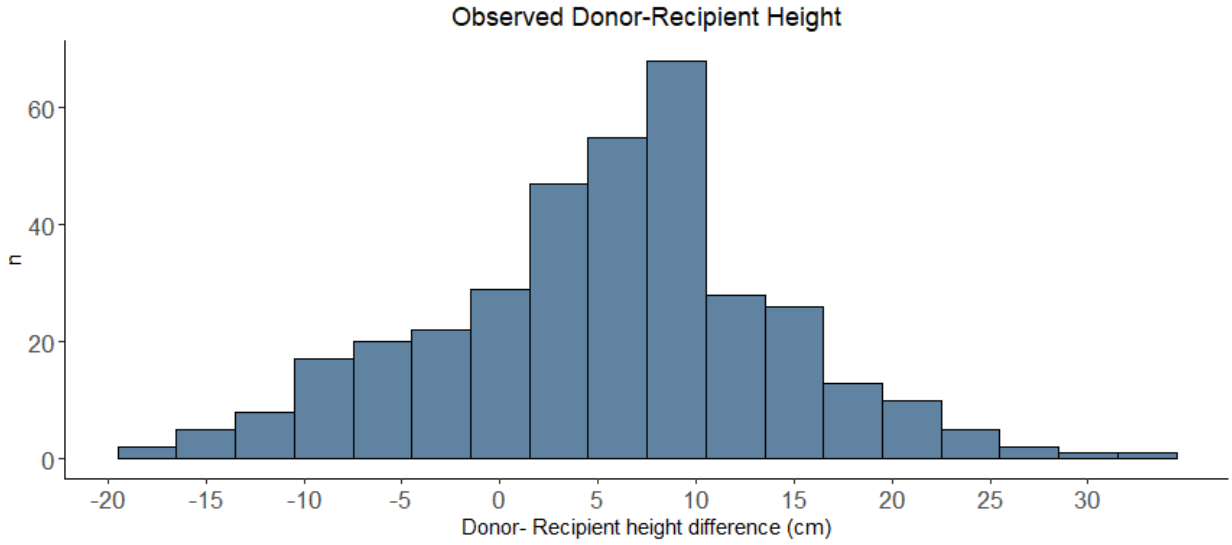


Figure F.4: Empirical Donor- Recipient Height. Height in cm. Positive difference indicates donor taller than recipient. Negative difference indicates donor shorter than recipient.

3.2 LAS Trajectories

LAS trajectories are measured daily for individuals on the waitlist and recorded in UNOS data. We observe LAS_{it} from $t =$ observed listing time to $t =$ observed death T_i, observed LTx L_i, or censoring time C, whichever occurs first. Alternative referral policies may result in listing of individuals who were never listed and/or earlier or later times of listing, so LAS measures are not

available for all required individuals and timepoints. We therefore use a linear mixed effects model to impute LAS at all time points. We include LAS components that are available for all patients in the CFF PR and do not include variables that are only available for waitlist patients (for example, functional status and 6 minute walk test). Patients tend to experience sudden changes in LAS in the days or weeks before LTx or death. Such changes are reflected in measures available in the UNOS data (functional status, 6 minute walk test, etc.), but are not available for all patients or at all timepoints. To capture the sudden changes without these variables, we instead use a fixed and random effect for whether the patient experienced either death or transplant within 30 days.

We use a linear mixed effects model to impute LAS:

$$LAS_{it} = \beta_0 + \beta_1 age_t + \beta_2 BMI_t + \beta_3 FVC_t + \beta_4 I(CFRD_t) + \beta_5 I(outcomein30days_t) + b_0 + b_1 I(outcomein30days_t)$$

Where age is centered at 18 years, time-varying BMI is calculated using the height and weight BLUP values at time t, FVC_t is the time-varying FVC BLUP, $I(CFRD)_t$ is an indicator for CF-related diabetes at time t, and $I(outcomein30days)$ is an indicator for whether the patient had an outcome of death or LTx in $t + 30$ days. b_1 and b_2 are random intercept and outcome effects. Empirical and fitted LAS values for a random sample of patients are given in Figure F.5.

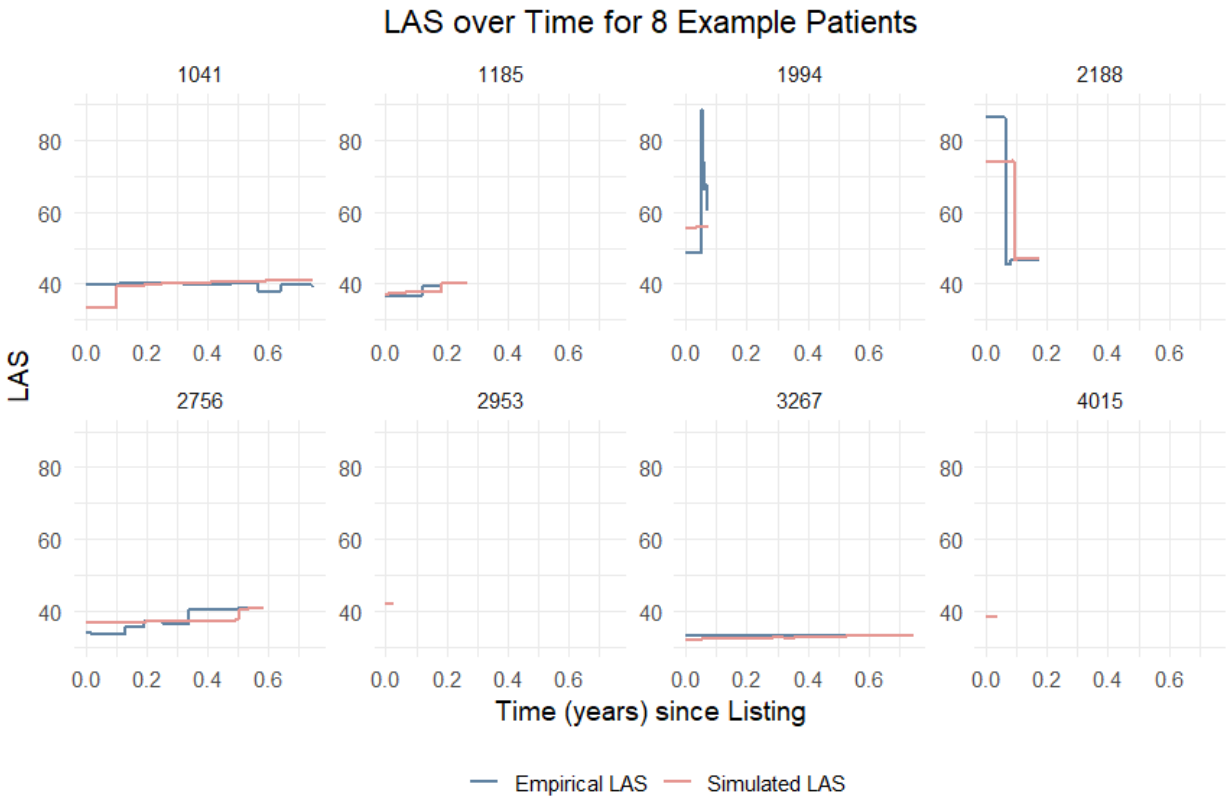


Figure F.5: Empirical and Simulated LAS over time for example patients.

4.0 PRE- AND POST- LTX SURVIVAL

We used a time-dependent survival modeling approach to estimate survival without transplant, and with transplant at time t, which varies between simulation runs as a result of priority position on the waitlist and organ supply. If pre-ltx death was observed, we use the observed time, $(T_i|LTx = 0)$. However, in cases where time of transplant, L_i occurred before $(T_i|LTx = 0)$, $(T_i|LTx = 0)$ is censored at L_i and must be imputed to assess how long the patient would survive in the absence of transplant, given survival up to L_i.

As a motivating example, consider a patient, i, who was observed to receive a lung transplant at $L_i = 100$ (days). Now consider that, under an alternative referral policy, additional high-risk patients are added to the waitlist at $t < 100$. Some of these patients may be higher priority (by LAS) than patient i at $t = 100$. Therefore, under this alternative policy, an organ available at $t = 100$ would go to one of the new, higher priority patients instead of patient i. Patient i would remain

on the waitlist past $t = 100$, however, it's unclear how long patient i would survive beyond $t = 100$ in the absence of transplant. Specifically, whether patient i would survive long enough to receive an organ in the future. To address this issue, we use a potential outcomes framework to estimate pre-transplant survival at time $t = L_i$, $(T_i(t)|LTx(t) = 0)$. The same framework can be employed to estimate individual i's survival if they were alternatively transplanted at another time, for example, at $t = 150$, $(T_i(t)|LTx(t) = 1)$.

Modeling survival conditional on transplant status using observed data

We use a potential outcomes model, where each individual on the waitlist has one potential time of death, T_i if they do not receive transplant at t and one potential time of death if they receive transplant at time t, given survival up t and no transplant before t.

$$T_{i,LTx}(t) = \begin{cases} T_{i,0}(t): \text{Potential time of death without transplant at } t \\ T_{i,1}(t): \text{Potential time of death with transplant at } t \end{cases}$$

For each individual, we can observe only one of these outcomes - they either were or were not transplanted at t. However, we can use information from other individuals to predict their survival under a counterfactual scenario where their transplant status and/or timing differed. Because transplant is allocated based on LAS, we assumed transplant was random within strata of LAS. Two patients with the same LAS should have the same likelihood of receiving any organ.

We utilized the LAS predictions over time $\widehat{LAS_{it}}$ described above and a time-dependent indicator of transplant, Ltx_{it}. Using a counting process, we modeled time-dependent survival using an exponential model. We also tested a Weibull model and generalized gamma model. The exponential model was determined to be an appropriate simplification when comparing visual fit, AIC, and BIC values. Time was measured as time on the waitlist. Because time on the waitlist could vary by policy, we used the time since first waitlist entry among UC, FEV1, and ML policies.

$$h(t) = \lambda * exp(\beta_1 LTx_t + \beta_2 LAS_t + \beta_3 age + \beta_4 gender + \beta_5 BMI)$$

where LTx and LAS vary over time and gender and BMI are measured at baseline (waitlist entry). LAS was centered at 40 and age was centered at 18.

Estimating expected pre- and post-transplant residual survival for simulation

We used the inverse sampling method to obtain potential times of death at time t, given X and LAS_t and changing the transplant indicator at t. For an individual with an observed transplant, we used $t = L_i$. We wish to obtain a time of death if the individual was not transplanted at t, given survival without transplant up to t.

$$T_i(t) = \lambda^{-1}(-log(U_i) * exp(-(\beta_1 * 0 + \beta_2 LAS_t + \beta_3 age + \beta_4 gender + \beta_5 BMI)))$$

where $t = L_i$, the transplant indicator is set to 0, and $U \sim Uni(0,1)$.

Similarly, we obtained estimated survival at the simulated time of transplant under each policy, L_{ip}^* in each simulation run, where $p \in (ML, FEV_1, UC)$. Even under UC, the simulated time of transplant (L_{ip}^* for $p = UC$) may differ somewhat from the observed time of transplant (L_i) for any individual, given differences between an individual's true and simulated UC evaluation time, random variation the organ flow, and simplifications of the simulated organ allocation process. We simulate post-transplant survival at each $t = L_{ip}^*$ using LAS_t and setting the transplant indicator at t to 1:

$$T_i(t) = \lambda^{-1}(-log(U_i) * exp(-(\beta_1 * 1 + \beta_2 LAS_t + \beta_3 age + \beta_4 gender + \beta_5 BMI)))$$

where $t = L_{ip}^*$ and the transplant indicator is set to 1. In any given simulation run, we fix U_i for each individual, such that variation in $T_i(t)$ is not attributable to differing values of U.

5.0 SUMMARY OF PLASMODE AND SIMULATED DATA

The data elements that were resampled from observed data ("plasmode") versus modeled are summarized in Table F.2.

Table F.2: Summary of plasmode versus modeled elements in microsimulation. *C* refers to the time of censoring for all patients (12/31/2016). L_i refers the observed time of transplant for individual *i*. T_i refers to the time of death for individual *i*.

Component	Plasmode (resampled) Elements	Modeled Elements
Baseline characteristics (X) ∈(age, gender, ABO)	Plasmode	-
Time-varying marker (X_t) ∈ (ML risk score, FEV_1)	Plasmode until L_i	Simulated after L_i using linear mixed effects model
Pre-LTx observation process (clinic visit times)	Plasmode until *Li*	Simulated after L_i using visit frequency in 12 months prior to L_i
Time-varying LAS (LAS_t)	-	Simulated using linear mixed effects model
Pre-Ltx survival ($T_i(t)$| t, LTx_t=0, LAS_t, X)	If $T_i < min(L_i, C)$, plasmode	If $L_i < C$, simulated using inverse sampling from time-dependent survival model
Post-Ltx survival ($T_i(t)$| t, LTx_t=1, LAS_t, X)	-	Simulated using inverse sampling from time-dependent survival model

VITA

Tricia Rodriguez received her PhD in Health Economics and Outcomes Research in June 2021 from the Comparative Health Outcomes, Policy, and Economics Institute (CHOICE) at the University of Washington. Tricia previously earned a Master's in Public Health, with a concentration in Health Metrics and Evaluation at the University of Washington. Her work focuses on the intersection of health economics and data science.