A Novel Tool to Evaluate the Accuracy of Predicting Survival and Guiding Lung Transplantation in Cystic Fibrosis

Aasthaa Bansal1, Nicole Mayer-Hamblett2,3,5, Christopher H Goss4, Lingtak N. Chan1 and Patrick J. Heagerty5

1 The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, University of Washington, Seattle WA, USA
2 Department of Pediatrics, University of Washington, Seattle, WA, USA
3 Seattle Children’s Research Institute, Seattle WA, USA
4 Division of Pulmonary and Critical Care Medicine, Department of Medicine and Pediatrics, University of Washington, Seattle WA, USA
5 Department of Biostatistics, University of Washington, Seattle WA, USA

Corresponding author: Aasthaa Bansal, The Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, Department of Pharmacy, University of Washington, Box 357630, 1959 NE Pacific Ave. H-375B, Seattle, WA, USA, 98195, Tel: +1 (206) 427-5448; E-mail: abansal@uw.edu

Received date: May 25, 2019; Accepted date: June 08, 2019; Published date: June 17, 2019

Copyright: © 2019 Bansal A, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Background: Effective transplantation recommendations in cystic fibrosis (CF) require accurate survival predictions, so that high-risk patients may be prioritized for transplantation. In practice, decisions about transplantation are made dynamically, using routinely updated assessments. We present a novel tool for evaluating risk prediction models that, unlike traditional methods, captures classification accuracy in identifying high-risk patients in a dynamic fashion.

Methods: Predicted risk is used as a score to rank incident deaths versus patients who survive, with the goal of ranking the deaths higher. The mean rank across deaths at a given time measures time-specific predictive accuracy; when assessed over time, it reflects time-varying accuracy.

Results: Applying this approach to CF Registry data on patients followed from 1993-2011 we show that traditional methods do not capture the performance of models used dynamically in the clinical setting. Previously proposed multivariate risk scores perform no better than forced expiratory volume in 1 second as a percentage of predicted normal (FEV1%) alone. Despite its value for survival prediction, FEV1% has a low sensitivity of 45% over time (for fixed specificity of 95%), leaving room for improvement in prediction. Finally, prediction accuracy with annually-updated FEV1% shows minor differences compared to FEV1% updated every 2 years, which may have clinical implications regarding the optimal frequency of updating clinical information.

Conclusions: It is imperative to continue to develop models that accurately predict survival in CF. Our proposed approach can serve as the basis for evaluating the predictive ability of these models by better accounting for their dynamic clinical use.

Keywords: Cystic fibrosis; Lung transplantation; Survival; Risk prediction; Classification accuracy

Introduction

Lung transplantation has been shown to improve survival for some cystic fibrosis (CF) patients whose disease is no longer amenable to more conventional medical therapies [1,2]. However, due to a shortage of donor lungs, a large number of wait-list patients die while awaiting transplantation employing the current allocation system in the US. In 2010-2012, the wait-list mortality rate was 15.4 per 100 wait-list years. Candidates aged 12-17 years had the highest wait-list mortality, at 19.7 deaths per 100 wait-list years, followed by those aged 18-34 years at approximately 18.5 deaths per 100 wait-list years [1].

Despite an increase in the rate of lung transplants over the past decade, wait-list mortality rates continue to rise [3]. Accurate predictions of mortality are necessary so that limited donor lungs may be prioritized to patients who are at the greatest risk of death without transplantation. The goal is to use a patient’s clinical characteristics to calculate the predicted risk of mortality within a specified time period and to rank or classify patients on the wait-list as those who are predicted to die soon versus those who are not.

Although forced expiratory volume in 1 second as a percentage of predicted normal (FEV1%) alone is the currently accepted measure of pulmonary function used for assessing patient prognosis and recommending lung transplantation, its prognostic accuracy has been questioned [4-6]. Several risk prediction (or prognostic) models combining FEV1% with other clinical factors have been proposed to improve survival prediction; however, it is unclear how accurate these models are, since their evaluation has failed to address classification accuracy and instead focused solely on measures of model fit or calibration. For example, Liu et al. developed the most well-known survival prediction model in CF that includes FEV1% and several additional clinical features [5]. They assessed model fit, and concluded that their model was more accurate than FEV1% alone for predicting 5-year mortality. However, only assessing how closely the model fits the observed data is not sufficient when the model is meant for classification [7-9].

While agreement between the predicted and actual...
numbers of deaths represents good model calibration, it does not necessarily translate to good classification accuracy [10].

Two studies have evaluated predicted versus actual mortality, Aaron et al. and Mayer-Hamblett et al. [9,11]. Aaron et al. compared predicted versus actual numbers of deaths and concluded that their model was accurate for prediction of one-year survival [11]. Mayer-Hamblett et al. however, addressed appropriate evaluation of their model using classification error rates, i.e. sensitivity and specificity, for classifying subjects based on 2-year mortality [9]. They showed that their multivariate model did not perform better than FEV1% alone, and that both are inadequate for use in practice. They concluded: “better clinical predictors of short-term mortality among patients with CF are needed” [9].

In addition to classification error rates, appropriate evaluation of a risk model must account for time-varying measurements. In practice, key clinical parameters included in the lung allocation score (LAS) are updated at routine care visits and then used to guide decisions regarding listing status in the US. Interestingly, the utility of updating markers has never been empirically evaluated. We assess time-varying accuracy using appropriate time-dependent classification error rates. Such evaluation could also help assess how often patient information such as FEV1% should be updated in practice before information becomes outdated and impacts accuracy.

It is imperative to continue to develop models that more accurately predict survival for CF patients. Prior studies that have proposed models have either failed to evaluate the accuracy of survival prediction, or have been limited to quantifying model accuracy according to an arbitrary baseline time point that does not reflect clinical practice where physicians update predictions as clinical parameters change. The primary objective of our work is to apply a novel statistical tool for evaluating the accuracy of survival models in CF importantly - and unlike prior evaluations - accounting for the potentially time-varying predictions based on routinely updated clinical assessments. Application of the tool for evaluating existing published survival models in CF is presented, in addition to a demonstration of how the tool can be utilized to identify subgroups of CF patients for whom prediction of survival is more reliable.

Methods

Study cohort

We used the Cystic Fibrosis Foundation (CFF) National Patient Registry (CFFPR) data, which consists of all patients who agreed to participate in the CF Registry and who were seen in a CFF-accredited care center in the US from 1986 through 2012. The CFFPR obtains written and informed consent from each participant. This analysis was approved by the CFFPR oversight committee and was granted Institutional Review Board Exempt Status by the University of Washington Human Subjects Division (due to being anonymized data).

We selected the prevalent subjects in 1993 as our study cohort to match the analysis of Liou et al. [5]. Our study cohort consisted of 17,926 prevalent CF cases on January 1, 1993. We treated the measurements available on this date as baseline FEV1% measurements. 6,028 subjects were excluded due to missing baseline FEV1% measurements (of these, 3,507 were younger than 5.5 years at baseline). After exclusions, 11,254 subjects remained in the study cohort; 3,906 (35%) of these subjects were observed to die during the study period ending on December 31, 2011. 2,414 subjects were recipients of lung transplantation between 1993 and 2011; we censored these subjects at the time of transplantation.

For each subject, we had baseline demographic and diagnosis data and approximately annual measurements of FEV1%. Counting multiple observations per patient, we included 140,651 total annual records. Of these, 16,846 had missing FEV1% measurements. We imputed missing values by carrying forward the last non-missing value for that patient.

Risk models

FEV1% of predicted is a standard measure used in practice to guide recommendation for lung transplantation. We evaluated the following models: (i) a base model containing only FEV1% and (ii) a multivariate model consisting of FEV1%, age, gender, weight, presence of pancreatic sufficiency, Staphylococcus aureus infection and Burkholderia cepacia infection [5]. Predictions from Cox models were summarized into a single baseline risk score and a separate time-varying, updated risk score. For the baseline score, we used 10-fold cross-validation to protect against overfitting. For the time-varying score, we used baseline measurements as training data to develop the Cox model and predicted the score at follow-up times using updated values of FEV1%. We added flexibility to both models by using cubic splines to model continuous variables.

Evaluation of model accuracy

Diagnostic accuracy: The traditional diagnostic classification problem is based on a binary outcome, typically the presence or absence of disease. Mayer-Hamblett et al. assessed prognostic accuracy by defining a ‘yes/no’ outcome for death within two years from baseline [9]. Classification errors include false negatives, i.e. patients who were predicted by the risk score to survive for longer than two years were observed to die within two years, and conversely, false positives, i.e. healthier subjects who did survive beyond two years were predicted to die within two years. Minimizing these errors is equivalent to maximizing the sensitivity (or true positive fraction (TPF)) and specificity (or 1 - false positive fraction (FPF)), respectively. The Receiver Operating Characteristic (ROC) curve is a standard tool that plots TPF versus FPF for all possible risk score cut-offs [12-16]. The ROC curve is commonly summarized using the area under the ROC curve (AUC), ranging from 0.5 to 1.0, which indicate no discrimination to perfect discrimination. The AUC is also a concordance statistic and represents the probability that a randomly chosen subject who dies (case) has a higher marker value than a randomly chosen subject who survives (control).

Time-varying prognostic accuracy: Implicit in the use of traditional diagnostic TPF and FPF are current-status definitions of disease. Since we are interested in a setting where outcome status changes with time, precise definitions are necessary to include event timing in definitions of prognostic error rates. Time-dependent ROC curve methods that extend concepts of sensitivity and specificity and characterize prognostic accuracy for survival outcomes have been proposed in the statistical literature and now widely adopted in practice [17,18]. These innovative methods represent a dynamic tool for assessing longitudinal changes that may alter prognosis within a defined time period.

To evaluate both baseline and time-varying measurements, we use the incident-case definitions of Heagerty & Zheng, which are based on a binary classification of the risk set at any time t [18]. That is, among
the patients who are still alive at time $t$, cases are defined as those who die at $t$ and controls as those who survive beyond $t$. The sensitivity and specificity at time $t$ are the error rates in classifying subjects at that time, and can be summarized using a time-dependent ROC curve. The time-dependent AUC, or $AUC(t)$, is then defined as the area under the time-dependent ROC curve and interpreted as the probability that a randomly chosen case who dies at time $t$ has a higher marker value than a randomly chosen control who survives beyond time $t$.

These definitions are appropriate for evaluating the performance of a baseline or time-varying marker in the CF lung transplantation setting, as interest lies in identifying patients who are at the highest risk of death in the near future, so that they may be given priority for limited donor organs. The recipient decision may be made at multiple time points as donor organs become available, but is always only applicable to those subjects who remain alive at those times.

We estimated the time-dependent AUC using a simple nonparametric rank-based approach [19]. The idea behind this approach is to compute for each risk set the binary concordance statistic using only the individual case and associated risk-set controls. For a fixed time $t$, we calculate a percentile for each case in the risk set relative to the controls in the risk set. A perfect marker would have the case value greater than 100% of risk set controls. The mean percentile at time $t$ is calculated as the mean of the percentiles for all cases in a window around $t$. The summary curve, $AUC(t)$, is then estimated as the local average of case percentiles. This nonparametric approach provides both a simple description for marker performance within each risk set and by smoothing individual case percentiles, a final summary curve over time characterizes how accuracy may be changing over time.

Although the AUC is a standard measure of accuracy, it summarizes the sensitivity of a risk score over the entire range of specificities from 0 to 100%. In contrast, clinical decisions are typically made based on a single risk score cut-off that has been shown to perform with high sensitivity and/or high specificity. Therefore, in evaluating a risk score’s performance and its impact for treatment decisions, it is important to also assess the sensitivity at a fixed high specificity. Using the above methods, we obtained a summary curve of sensitivity or TPF for a fixed specificity of 95% or FPF of 5%.

Finally, we assessed the time-varying prognostic accuracy of time-varying, or updated, risk scores using an extension of the approach of Saha-Chaudhuri & Heagerty [19] to accommodate time-varying markers. At any time $t$, the last measured value of the risk score was used as the current risk prediction.

**Stratification by risk group:** We evaluated the performance of annually updated FEV$_1$% measurements in subgroups defined by baseline FEV$_1$% (FEV$_1$% ≤ 30 versus FEV$_1$% > 30), based on the cut-off recommended by the International Society for Heart and Lung Transplantation [20]). Baseline age (≤ 11 years, 12-17 years, and ≥ 18 years), gender, and F508 genotype.

### Results

#### Cohort characteristics

Table 1 summarizes the characteristics of the study cohort. Patients with shorter survival also tended to be older at baseline, have lower FEV$_1$% at baseline, be in lower weight and height percentiles, be slightly less likely to have *Staphylococcus aureus* infection and slightly more likely to have *Burkholderia cepacia* infection.

<table>
<thead>
<tr>
<th></th>
<th>Overall (n=11,254)</th>
<th>Died within 1 year (n=287)</th>
<th>Died 1-5 years (n=1,077)</th>
<th>Survived at least 5 years (n=9,146)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age on Dec 31, 1992, Mean (SD)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>18.0 (8.9)</td>
<td>24.2 (9.6)</td>
<td>22.0 (8.2)</td>
<td>16.9 (8.5)</td>
</tr>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>5,187 (46%)</td>
<td>144 (50%)</td>
<td>529 (49%)</td>
<td>4,185 (46%)</td>
</tr>
<tr>
<td>Male</td>
<td>6,067 (54%)</td>
<td>143 (50%)</td>
<td>548 (51%)</td>
<td>4,961 (54%)</td>
</tr>
<tr>
<td><strong>Race, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>10,858 (96%)</td>
<td>281 (98%)</td>
<td>1,031 (96%)</td>
<td>8,819 (96%)</td>
</tr>
<tr>
<td>African American</td>
<td>320 (3%)</td>
<td>6 (2%)</td>
<td>40 (4%)</td>
<td>259 (3%)</td>
</tr>
<tr>
<td>Other</td>
<td>76 (1%)</td>
<td>0 (0%)</td>
<td>6 (1%)</td>
<td>68 (1%)</td>
</tr>
<tr>
<td><strong>Genotype, n (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔF508 homozygous</td>
<td>4,346 (39%)</td>
<td>55 (19%)</td>
<td>281 (26%)</td>
<td>3,856 (42%)</td>
</tr>
<tr>
<td>ΔF508 heterozygous</td>
<td>3,139 (28%)</td>
<td>31 (11%)</td>
<td>165 (15%)</td>
<td>2,819 (31%)</td>
</tr>
<tr>
<td>Other</td>
<td>883 (8%)</td>
<td>11 (4%)</td>
<td>63 (6%)</td>
<td>742 (8%)</td>
</tr>
<tr>
<td>Missing</td>
<td>2,886 (26%)</td>
<td>190 (66%)</td>
<td>568 (53%)</td>
<td>1,729 (19%)</td>
</tr>
<tr>
<td><strong>FEV$_1$%, Mean (SD)</strong></td>
<td>68.5 (28.0)</td>
<td>32.4 (19.2)</td>
<td>40.8 (20.1)</td>
<td>74.5 (25.4)</td>
</tr>
<tr>
<td><strong>Weight percentile, Mean (SD)</strong></td>
<td>28.5 (26.6)</td>
<td>8.7 (16.1)</td>
<td>12.4 (16.5)</td>
<td>30.5 (26.6)</td>
</tr>
</tbody>
</table>
Table 1: Summary of baseline (1993) subject characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Height percentile, Mean (SD)</th>
<th>Staphylococcus aureus status, n (%)</th>
<th>Burkholderia cepacia status, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29.3 (26.3)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>17.5 (23.0)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>19.1 (22.4)</td>
<td>Not cultured</td>
<td>Not cultured</td>
</tr>
<tr>
<td></td>
<td>30.6 (26.5)</td>
<td>3,001 (27%)</td>
<td>346 (3%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42 (15%)</td>
<td>33 (11%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>242 (22%)</td>
<td>88 (8%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,559 (28%)</td>
<td>195 (2%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7,014 (62%)</td>
<td>9,669 (86%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>234 (82%)</td>
<td>243 (85%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>793 (74%)</td>
<td>947 (88%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5,468 (60%)</td>
<td>7,832 (86%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,239 (11%)</td>
<td>1,239 (11%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 (4%)</td>
<td>11 (4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42 (4%)</td>
<td>42 (4%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,119 (12%)</td>
<td>1,119 (12%)</td>
</tr>
</tbody>
</table>

Assessment of time-varying prognostic accuracy of base model using baseline and updated measurements

Figure 1a shows the estimated AUC(t) or mean percentile for FEV₁% baseline and annually updated measurements. Not surprisingly, the performance of a baseline FEV₁% measurement declines over time, from 0.87 (95% CI (0.84, 0.88)) at baseline (1993) to 0.62 (95% CI (0.59, 0.65)) 20 years later (2012). In contrast, an annually updated FEV₁% consistently maintains an AUC of approximately 0.90 over time. Although AUC=0.90 is typically considered to be excellent performance, it does not translate to FEV₁% having adequate performance in this setting. 90th percentile means that out of 100 patients, the case marker value is higher than 90 control marker values; however, it also means that 9 other controls will be prioritized ahead of the case. Figure 1b shows the TPF for a fixed FPF of 5%. Again, an annually updated FEV₁% has better performance than baseline FEV₁%; however, a TPF of 45% is likely inadequate for clinical practice.

We also assessed the performance of a multivariate model that combined FEV₁% with a number of clinical variables. We note that this model is very similar to Liou et al.’s model [5], which they proposed as a better predictor of mortality, with the exception that our model excluded diabetes mellitus and number of acute exacerbations, as we did not have data on these variables for this analysis. Figure 2 confirms the findings of Mayer-Hamblett et al. [9] and shows that adding clinical variables to the FEV₁% base model does not improve performance beyond using FEV₁% alone, using baseline or updated measurements.

Finally, an assessment of the performance of annually updated FEV₁% measurements compared to those updated every two years shows slightly worse performance of the latter in even years when the measurement is a year old. Performance in odd years is exactly the same, since FEV₁% is always updated in those years (Table available in online supplemental material).
Stratification by risk group

Figure 3 presents Kaplan-Meier survival curves for subgroups defined by baseline FEV₁%, baseline age, gender, and F508 genotype. We see a large gap in survival probabilities by baseline FEV₁%. Patients with FEV₁% ≤ 30 have poor prognosis, with an estimated 5-year survival probability of 35%, compared to 90% in the FEV₁%>30 subgroup. The second panel shows worse survival with increasing age, with estimated 5-year survival probabilities of 95%, 87% and 76% in the ≤ 11 years, 12-17 years, and ≥18 years subgroups, respectively. Gender and F508 genotype have little impact on survival probabilities. This result regarding F508 genotype is likely due to the low rate of genotyping in this cohort (66% of the patients who died within one year were not genotyped).

Discussion

Commonly used measures of model fit and calibration are insufficient for evaluating risk prediction models in CF when the goal is to prioritize patients for lung transplantation and when a patient’s clinical status is changing over time. Therefore, we presented an approach that uses time-dependent classification error rates that can be used to characterize the potential performance of a survival model, while accounting for the time-varying nature of the prediction itself.

In the CF setting, we found that updated measurements of FEV₁% have consistent performance over time, whereas the performance of a baseline measurement declines over time. Thus, previously reported estimates of the accuracy of FEV₁% alone do not capture its true performance in a clinical setting. It is clear that patient information should be updated over time to maintain classification accuracy; however, it is also evident that neither FEV₁% alone nor existing multivariate models are adequate for use in practice.

Being able to evaluate a model’s time-varying accuracy may also help guide clinical practice and policy with regards to the frequency of updating patient information. A comparison of 1-year versus 2-year measurements of FEV₁%, for example, showed minor differences in performance.

A limitation of the analysis is that it was done in an earlier cohort, in order to compare performance with Liou et al.’s model [5]. A more up-to-date cohort could change the results, but it would not take away from the key point – that time-varying approaches are better than the traditional approaches currently used in CF.
Conclusion

In this study we demonstrated that using a statistical evaluation approach that is closely tied to the clinical goal of using predicted risk as a score to rank patients as a function of time can significantly change the conclusions drawn about a risk prediction model’s performance. As new models are developed, perhaps incorporating novel biomarkers, the proposed approach could be used to accurately assess their predictive ability. As shown, standard methods may underestimate their performance by not capturing how these models will be used dynamically within the clinical setting. We note that our focus here is on risk prediction models, assuming that patients are added to a lung transplantation wait-list based on their expected benefit from transplantation. In practice, any risk prediction should be coupled with assessments of treatment benefit. It is imperative to continue to develop models that accurately predict survival in CF. Our proposed approach can serve as the basis for evaluating the predictive ability of these models by better accounting for their dynamic clinical use.

Acknowledgments

The authors thank the Cystic Fibrosis Foundation for allowing the use of the registry data.

Funding

This project was supported by the NIH (R37-CA218413, R01-HL072966, UL1TR000423, HL103965, R01-A1101307, P30-DK089507), the PhRMA Foundation, the Cystic Fibrosis Foundation, and the FDA (R01-FD003704).

References