



Development and external validation of 1- and 2-year mortality prediction models in cystic fibrosis

Sanja Stanojevic^{1,2}, Jenna Sykes³, Anne L. Stephenson^{2,4}, Shawn D. Aaron⁵ and George A. Whitmore⁶

Affiliations: ¹Translational Medicine, The Hospital for Sick Children, Toronto, ON, Canada. ²Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada. ³Keenan Research Centre in the Li Ka Shing Knowledge Institute of St Michael's Hospital, Toronto, ON, Canada. ⁴Adult CF Program, St Michael's Hospital, Toronto, ON, Canada. ⁵The Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON, Canada. ⁶Desautels Faculty of Management, McGill University, Montreal, QC, Canada.

Correspondence: Sanja Stanojevic, Translational Medicine, The Hospital for Sick Children, Toronto, ON, Canada. E-mail: Sanja.Stanojevic@sickkids.ca



@ERSpublications

This clinical tool considers patients' overall health status and the risk of intermittent shock events to predict 1- and 2-year risk of death for patients with cystic fibrosis. The model accurately predicted death for patients in Canada and the UK. <http://bit.ly/2Wbnyaq>

Cite this article as: Stanojevic S, Sykes J, Stephenson AL, *et al.* Development and external validation of 1- and 2-year mortality prediction models in cystic fibrosis. *Eur Respir J* 2019; 54: 1900224 [<https://doi.org/10.1183/13993003.00224-2019>].

ABSTRACT

Introduction: We aimed to develop a clinical tool for predicting 1- and 2-year risk of death for patients with cystic fibrosis (CF). The model considers patients' overall health status as well as risk of intermittent shock events in calculating the risk of death.

Methods: Canadian CF Registry data from 1982 to 2015 were used to develop a predictive risk model using threshold regression. A 2-year risk of death estimated conditional probability of surviving the second year given survival for the first year. UK CF Registry data from 2007 to 2013 were used to externally validate the model.

Results: The combined effect of CF chronic health status and CF intermittent shock risk provided a simple clinical scoring tool for assessing 1-year and 2-year risk of death for an individual CF patient. At a threshold risk of death of $\geq 20\%$, the 1-year model had a sensitivity of 74% and specificity of 96%. The area under the receiver operating curve (AUC) for the 2-year mortality model was significantly greater than the AUC for a model that predicted survival based on forced expiratory volume in 1 s $< 30\%$ predicted (AUC 0.95 *versus* 0.68 respectively, $p < 0.001$). The Canadian-derived model validated well with the UK data and correctly identified 79% of deaths and 95% of survivors in a single year in the UK.

Conclusions: The prediction models provide an accurate risk of death over a 1- and 2-year time horizon. The models performed equally well when validated in an independent UK CF population.

This article has supplementary material available from erj.ersjournals.com

Received: Feb 01 2019 | Accepted after revision: May 06 2019

Copyright ©ERS 2019

Introduction

Survival in patients with cystic fibrosis (CF) has improved steadily over the past three decades, with the median age of survival in many countries now exceeding 45 years [1, 2]. Despite these improvements, CF remains a life-shortening illness with half of all deaths occurring before the age of 35 years [3]. There have been many studies that have developed predictive models for CF mortality [4–10]; however, forced expiratory volume in 1 s (FEV₁) <30% predicted remains the most commonly used indicator of 2-year survival, and CF clinicians generally use the 30% FEV₁ threshold to guide patient referrals for lung transplantation assessment. In addition, there are several methodological limitations to existing prediction models, including age-related bias, the lag time between risk factor measurement and death, and the introduction of confounding by indication when transplant recipients are included in the study population. There is still a great need to develop an accurate prediction model for mortality in order to identify patients who would benefit from expedited referral to a lung transplant programme.

We previously published a clinical scoring tool to predict the 1-year risk of death for CF patients living in Canada [6]. This tool predicted the risk of death in Canadian patients with CF by modelling two components of a patient's disease course, namely, a chronic health index that measured baseline disease severity and a shock index that measured the potential for acute deterioration in health, such as deterioration that occurs at the time of a pulmonary exacerbation. The 1-year model had good internal validity and goodness-of-fit but had not been externally validated. Furthermore, given the wait time required to complete a transplant evaluation and receive a lung transplant, a 2-year predictive risk model would be clinically useful, as this would give adequate lead time to successfully initiate a transplant.

Here we aimed to 1) develop a clinical predictive scoring tool for 1- and 2-year risk of death using updated Canadian CF Registry (CCFR) data, and 2) externally validate the tool using UK CF Registry data.

Methods

Data source

CCFR data from 1982 to 2015, for patients aged ≥ 6 years were used to develop the 1- and 2-year models (more details can be found in the supplementary material). A detailed description of the clinical variables that were considered in the development of the risk model is included in the supplementary material.

Risk models

The predictive risk models for death as the outcome were developed using threshold regression, which assumes that the decline in lung function over time is a stochastic process explained by two components: 1) a declining chronic health status (health index), and 2) intermittent shocks typically due to pulmonary exacerbations (shock index). The health index and shock index represent separate but complementary influences on mortality risk. The interpretation of the model is based on a joint association between the chronic health index and shock index. A single pool of variables is considered that may be important in none, one or both domains. Patient survival times were time until death censored by date of first transplant or date of next annual visit. The model and methodology have been described in detail in prior publications [6, 11]. Clinical data from the preceding year were used to calculate a patient's chronic health and shock indices for the 1-year risk of death. Backward step-wise elimination was utilised for variable retention until all p-values were <0.05.

A 2-year risk of death prediction was calculated by multiplying 1) the estimated survival probability for the first year and 2) the estimated conditional probability of surviving the second year given survival for the first year. For this second probability, we constructed a separate risk model for the second year using the same predictors from the 1-year model that are lagged by one annual visit (*i.e.* predictors that are 2 years out of date). The construction follows the same statistical procedure used for the 1-year risk, with the final model including only risk factors that are significant at the $p < 0.05$ level.

Oversampling

Since each patient contributes multiple years of observation to the model, the outcome of death in any given year is extremely rare (<1%), thus heavily weighting the model's fitted regression function towards the survivors. To address this issue, we applied a synthetic minority oversampling technique to re-weight the model coefficients. In our application of the technique, we oversampled underrepresented or minority events in the sample, which in our context are death events and undersampled survivor records [12]. On average, this resulted in ~ 1300 deaths (6% of the sample) and 21000 survivor records in each of the sampled populations, compared with 893 deaths (1.6% of the total population) and 55942 survivor records if no oversampling was done. Deaths were oversampled by creating synthetic records using a five-nearest-neighbour approach. For each death, two synthetic records were created by randomly selecting two neighbours from the nearest five neighbours. Survivor records were randomly undersampled such that

10 survivor records were selected for every death. We subsequently used bootstrap sampling to create 1000 samples where deaths were oversampled and re-estimated the coefficients for the 1- and 2-year risk models. The final models represent the average coefficients from the 1000 models.

Sensitivity analyses

In order to evaluate the robustness of our primary results, we performed several sensitivity analyses which involved re-estimating the predictive model 1) excluding pancreatic status as a covariate; 2) excluding patients on cystic fibrosis transmembrane conductance regulator modulator treatments; 3) using pulmonary death as the outcome rather than death from any cause; 4) using death or transplant as the outcome; 5) excluding patients born before 1970; and 6) using clinical data from only 1 year prior rather than up to 3 years prior to predict risk.

External validation of risk models

UK CF Registry data from 2007 to 2013 were used to externally validate the models. Exclusion criteria and variable definitions used for the Canadian dataset were matched with the UK dataset whenever possible. The two main exceptions encountered were as follows. 1) Hospitalisation data in the UK were not available until 2007, thus, patients whose only clinic visit year was in 2007 were excluded because they did not have “look-back” data available; 2) the number of hospital days per year was available in the UK registry rather than the number of hospitalisations. In order to harmonise the data, hospital days from the UK registry were scaled by dividing by 14 days, assuming an average patient spends 2 weeks in hospital for each stay.

The probability of survival at 1 and 2 years was calculated for each patient and each year by directly applying the weighted Canadian model coefficients (*i.e.* the model derived from the oversampled data).

Goodness of fit

Model validation was assessed by looking at both model calibration and discrimination. Discrimination was evaluated using areas under receiver operating characteristic curves (AUC measures). Calibration was assessed by graphical investigation of cumulative Martingale residual plots. Martingale residuals are calculated as differences between actual death outcomes (represented as binary outcomes) and estimated probabilities of death [6].

Results

1- and 2-year predictive models

The Canadian dataset used to build the predictive models contained 56839 annual records for 4993 patients (figure 1a); 893 deaths were recorded. The demographic and clinical characteristics of the study population are listed in table 1.

The 1- and 2-year risk factors for death are summarised in figures 2 and 3 and in supplementary table S1. Positive coefficients in the model were associated with a favourable outcome (*i.e.* lower risk of death), whereas negative coefficients were associated with a greater risk of death. The joint modelling of the chronic health index and shock index allow for some covariates to contribute to both domains. For example, lung function contributes to both overall health status and shock events, such that the combined contribution of low lung function to mortality risk in the shock index is greater in individuals with low overall health status. When deaths were oversampled, the same risk factors were found to be statistically significant, but the magnitudes of the regression coefficients changed (figures 2 and 3).

The nature of the model is such that a shock event has a lower risk of causing death for a patient with a good chronic health index than a sicker patient, as demonstrated in figure 4. For example, a patient with a 1-year chronic health index of 4.2, and a shock index of 0.24, has a 1% 1-year risk of death (calculations can be found in the supplementary material). The chronic health and shock indices combine in a similar way to give the probability of 2-year survival conditional on surviving the first year. For example, if the 1-year probability of survival is 99.0% (1.0% risk of death), and the calculated conditional probability of survival for the second year is 97.4% (2.6% risk of death), then the computed conditional 2-year probability of survival is $100 \times (0.990) \times (0.974) = 96.4\%$ (3.6% risk of death).

To examine the trade-off between sensitivity and specificity implied by our prediction model, we selected a single calendar year of Canadian experience (2012) and used a cut-off where the estimated probability of death in the next year was $\geq 20\%$ (table 2). In that year, we identified 110 patients at “high-risk” of death (*i.e.* having a $\geq 20\%$ risk of death), of whom 14 died. Conversely, only five of the 2434 patients identified as “low risk” died. Overall, this cut-point had a sensitivity of 74% and specificity of 96%.

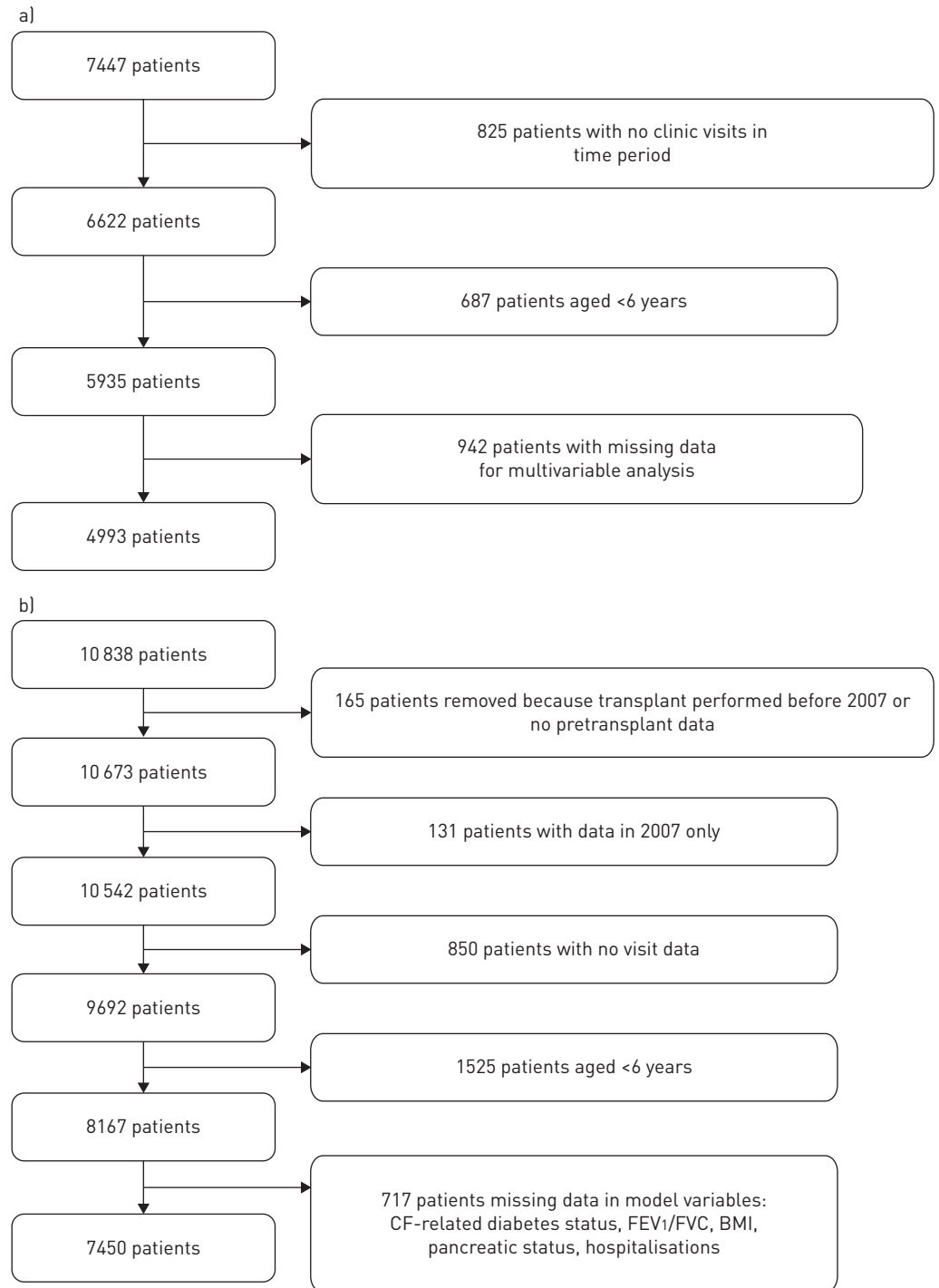


FIGURE 1 a) Consort diagram of Canadian cystic fibrosis (CF) population used to derive prediction models (1982–2015); b) consort diagram of UK CF population used to validate prediction models (2008–2013). FEV₁: forced expiratory volume in 1 s; FVC: forced vital capacity; BMI: body mass index.

In addition, we compared the threshold model results with the traditional model, which uses FEV₁ <30% pred, which is still commonly used as a general indicator for transplant referral, to identify those with >50% risk of death in the next 2 years. The 2-year threshold model was superior (AUC 0.95) to the traditional model (AUC 0.68) (figure 5).

The results of the model were consistent across all sensitivity analyses (supplementary figure S1), with notably wider confidence intervals when the look-back window for clinical data was limited to 1 year.

TABLE 1 Clinical and demographic characteristics of the Canadian and UK cystic fibrosis (CF) populations

	Canada 1982–2015	UK 2008–2013	Canada 2008–2013 [#]
Patients	4993	7450	3342
Annual records	56839	27014	15226
Sex			
Female	2365 (47.4)	3472 (46.6)	1569 (46.9)
Male	2628 (52.6)	3978 (53.4)	1773 (53.1)
Age at diagnosis years	0.7 (0–73.7)	0.422 (0–79.2)	0.7 (0–73.7)
<2 yrs	3186 (63.8)	5222 (70.1)	2131 (63.8)
≥2 yrs	1807 (36.2)	2228 (29.9)	1211 (36.2)
Birth cohort			
<1970	989 (19.8)	580 (7.8)	400 (12.0)
1970–1979	1059 (21.2)	887 (11.9)	484 (14.5)
1980–1989	1260 (25.2)	1966 (26.4)	930 (27.8)
1990–1999	1124 (22.5)	2535 (34.0)	1063 (31.8)
≥2000	561 (11.2)	1482 (19.9)	465 (13.9)
Pancreatic status (ever/never)			
Sufficient	814 (16.3)	941 (12.6)	437 (13.1)
Insufficient	4179 (83.7)	6509 (87.4)	2905 (86.9)
CF-related diabetes (ever/never)			
No	3647 (73.0)	4176 (56.1)	2464 (73.7)
Yes	1346 (27.0)	3274 (43.9)	878 (26.3)
FEV₁ % pred	55.8 (8.7–138.3)	69.5 (9.2–138.9)	70.0 (10.6–136.7)
<40	1786 (35.8)	1361 (18.3)	621 (18.6)
40–69	1307 (26.2)	2412 (32.4)	1048 (31.4)
70–89	991 (19.8)	2111 (28.3)	871 (26.1)
≥90	909 (18.2)	1566 (21.0)	802 (24.0)
FVC % pred	76.4 (13.7–147.4)	83.8 (12.1–146.7)	86.9 (16.7–149.7)
Hospitalisations in prior year[¶]	0 (0–14)	0 (0–19.3)	0 (0–14)
0	2942 (58.9)	4161 (55.9)	2324 (69.5)
1–2	1338 (26.8)	2282 (30.6)	759 (22.7)
≥3	713 (14.3)	1007 (13.5)	259 (7.7)
Nutritional status			
Underweight	1114 (22.3)	1002 (13.4)	438 (13.1)
Normal	3184 (63.8)	5134 (68.9)	2348 (70.3)
Overweight	695 (13.9)	1314 (17.6)	556 (16.6)
Microbiology			
Neither <i>P. aeruginosa</i> nor BCC	1901 (38.1)	3373 (45.3)	1364 (40.8)
<i>P. aeruginosa</i> , no BCC	2168 (43.4)	3684 (49.4)	1563 (46.8)
BCC	924 (18.5)	393 (5.3)	415 (12.4)
Transplant: yes	589 (11.8)	220 (3.0)	197 (5.9)
Vital status: dead	893 (17.9)	355 (4.8)	142 (4.2)

Data are presented as n, n (%) or median (range). FEV₁: forced expiratory volume in 1 s; FVC: forced vital capacity; *P. aeruginosa*: *Pseudomonas aeruginosa*; BCC: *Burkholderia cepacia* complex. Observations on clinical variables for the last clinic visit, unless otherwise stated. [#]: the threshold regression model was developed using the full Canadian dataset from 1982 to 2015. The UK validation dataset had data from 2008 to 2013, so a subset of the Canadian data for the same time period was used for comparison. [¶]: UK hospitalisations are derived by dividing total days in hospital by 14 (2 weeks).

Several of the coefficients, including FEV₁ and the change in FEV₁, increased in magnitude when death or transplant was used as the outcome, rather than censoring survival at time of transplant.

External validation using UK data

The threshold model was externally validated using data from 2007 to 2013, which included 7450 patients with 27014 annual records from the UK patient registry (figure 1b). The UK and Canadian populations were similar with respect to most clinical and demographic characteristics (table 1). Canada had more patients infected with *Burkholderia cepacia* complex, while the UK had more cases of CF-related diabetes. In 2012, the model identified 409 UK patients as high-risk of death (*i.e.* those with a ≥20% risk of death), of whom 73 died (table 3). Only 336 out of the 6029 UK patients identified as low-risk died. The Canadian-derived survival model exhibited similar sensitivity and specificity when applied to UK patients

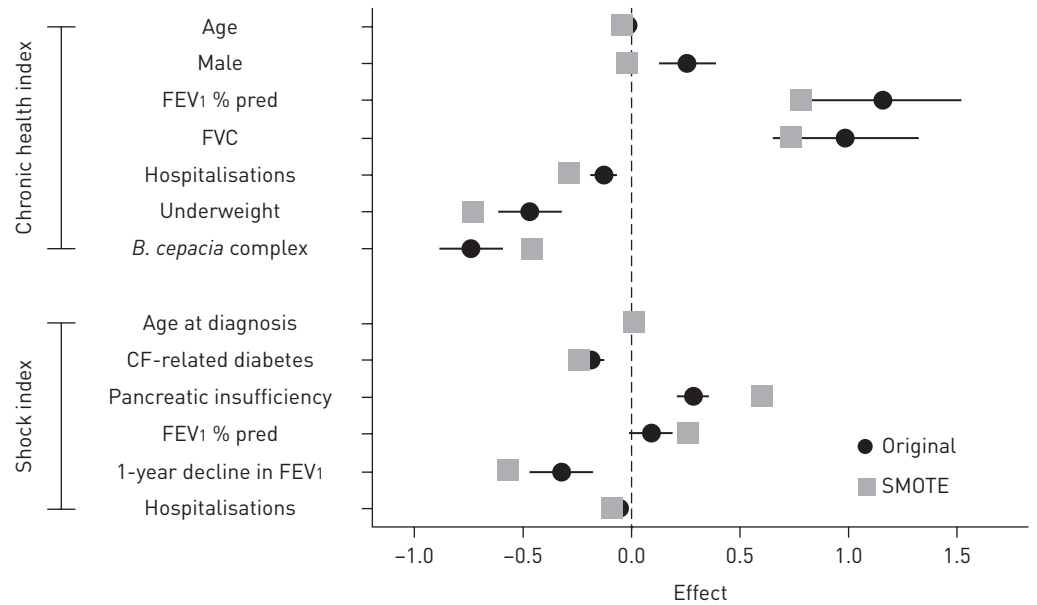


FIGURE 2 Forest plot of original and synthetic minority oversampling technique (SMOTE) estimates for the chronic health index and shock index in the Canadian 1-year predictive model. FEV1: forced expiratory volume in 1 s; FVC: forced vital capacity; *B. cepacia*: *Burkholderia cepacia*; CF: cystic fibrosis.

compared to when it was applied to Canadian patients (Canadian sensitivity 74%, UK sensitivity 79%; Canadian specificity 96%, UK specificity 95%).

Goodness of fit

The overall 1-year model discrimination was good; the AUC was 0.92 using the unweighted threshold regression coefficients and increased to 0.95 using the oversampled coefficients for the 1-year model. These values imply that if we select two patients at random, one who survives 1 year and one who dies in the first year, there is a 95% probability that the model will assign a higher risk of death to the patient who dies compared to the one who survives. Similarly, the AUC for the 2-year model improved from 0.86

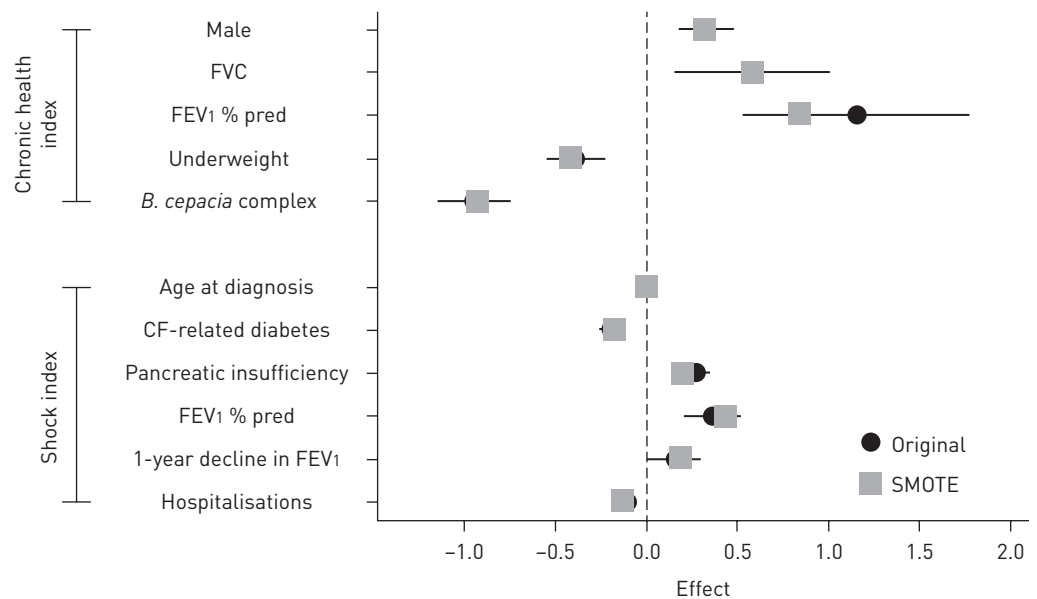


FIGURE 3 Forest plot of original and synthetic minority oversampling technique (SMOTE) estimates for the chronic health index and shock index in the Canadian 2-year predictive model (conditional on 1-year survival). FVC: forced vital capacity; FEV1: forced expiratory volume in 1 s; *B. cepacia*: *Burkholderia cepacia*; CF: cystic fibrosis.

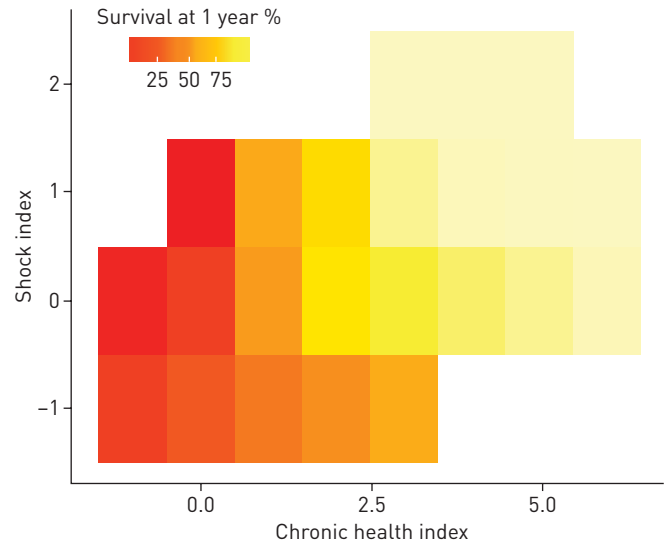


FIGURE 4 Heat map of 1-year survival probability for the Canadian cystic fibrosis population, showing how the probability varies with the chronic health index and exacerbation shock index. Larger values of both indices generally represent improved survival prospects.

to 0.90 after oversampling. Martingale residuals were used to plot cumulative actual deaths and cumulative estimated probabilities of death for the unweighted model. The plots indicate good calibration and predictive power of the model (supplementary figure S2). The same 1- and 2-year oversampled adjusted models derived from the Canadian data performed well when validated using the UK dataset with AUCs of 0.95 and 0.91, respectively (supplementary figure S3).

Discussion

We have developed an updated clinical tool that predicts 1- and 2-year mortality in Canadian CF patients, and demonstrated that the tool is applicable to the UK CF population. The tool is novel in that it uses two distinct components (chronic health index and shock index) to determine the risk of death using accessible clinical parameters. As a result, clinicians will be better able to identify those individuals at high risk of death, thus allowing time for consideration of transplant referral or listing, or more aggressive therapy and follow-up.

This analysis is based on comprehensive longitudinal registry data and the final model validated well with external data on multiple levels. Our validation using UK data confirmed that the Canadian model could be applied to the UK data to identify patients at greater risk of death without the need for adjustments to the model or its coefficients. Another advantage of our model is that we used predictor variables readily available in most CF registries, so that the model may be applied to other countries. Nonetheless, we have not validated our model outside of Canada and the UK and use of this model in other countries may require adjustments in predictor variables or regression coefficients to account for structural differences in healthcare systems and CF patient populations.

Like all clinical prediction tools there is a trade-off between correctly predicting death and minimising the number of patients who are identified as high risk. We used a conservative cut-off of 20% risk of mortality in tables 2 and 3 to identify those patients who were at a relatively high risk of death in order to ensure that the opportunity for transplant discussions or more aggressive therapy is not missed. Finally, the model can easily be incorporated into a web-based or electronic health record platform to provide clinicians with a calculated survival probability score to facilitate discussions with patients.

TABLE 2 Summary of the sensitivity and specificity for the 1-year risk of death for a randomly selected year (2012) for the Canadian cystic fibrosis population

	Died in year 1	Alive at 1 year	
Risk of death 20–99%	14	96	PPV 13%
Risk of death 0–19.9%	5	2429	NPV 100%
	Sensitivity 74%	Specificity 96%	

PPV: positive predictive value; NPV: negative predictive value.

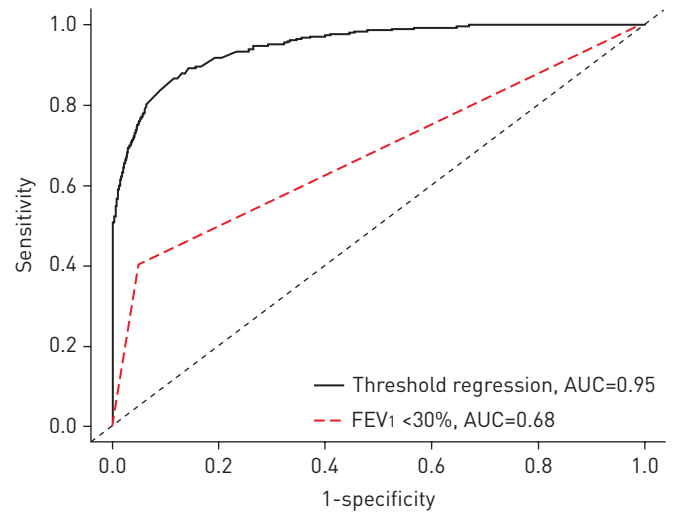


FIGURE 5 Comparison of the threshold regression 2-year risk model and the traditional forced expiratory volume in 1 s (FEV₁) <30% risk model. AUC: area under the curve.

As the rate of death in the CF population is relatively low, we purposely oversampled deaths, and re-estimated the model coefficients to ensure that the model correctly estimated the probability of death. In both the original and oversampled analysis, the model AUC was higher than reported for similar prediction models, especially the commonly used 30% FEV₁ pred cut-off [8–10]. After applying the oversampling technique, the coefficients were all found to be in the same direction and of similar magnitude as the original model results, indicating that the original model is robust. Overall, patients at the extremes of the risk scale remain at the extremes regardless of whether the oversampling technique is used, whereas there is a downweighting of the probability of death for patients in the middle of the scale. The distinct advantage of the oversampling technique is that the coefficients are weighted toward deaths, and therefore more accurately calculate a probability of death.

Our updated risk models include both a 1-year and a conditional 2-year risk of death. The regression coefficients for variables change in magnitude and even direction between the 1-year and conditional 2-year models, highlighting that more distant clinical measures have less of an impact on survival prospects. For example, a drop in FEV₁ during the year immediately prior portends a greater 1-year risk of death; however, if a patient survives the immediate first year after an FEV₁ decline, their conditional survival during the second year is no longer adversely affected by their remote FEV₁ decline.

In our models, pancreatic status produced the counterintuitive result that patients who were pancreatic-sufficient had a higher risk of death. This effect might be caused by our treatment of pancreatic status as a time-dependent variable in the model, with each annual review being an independent observation. Furthermore, pancreatic insufficiency was derived from reported pancreatic enzyme use which is only a proxy variable and may have been misclassified for some patients in both directions. In the sensitivity analysis which excluded coefficients for other covariates, results were very similar whether pancreatic status was excluded or not.

The risk factors identified in our analysis are similar to the N_{KAM} *et al.* [8] model, which used a simpler logistic regression model for predicting 3-year risk of death using French registry data. In addition to patient characteristics, the French model included treatments. Treatment data were not available in the Canadian registry until 2015, and therefore not used in our analysis. The other main difference between our analysis and the French prognostic score was their use of death and/or transplant as the outcome,

TABLE 3 Summary of the sensitivity and specificity for the 1-year risk of death for a randomly selected year (2012) for the UK cystic fibrosis population

	Died in 1 year	Alive at 1 year	
Risk of death 20–99%	73	336	PPV 18%
Risk of death 0–19.9%	19	5693	NPV 100%
	Sensitivity 79%	Specificity 95%	

PPV: positive predictive value; NPV: negative predictive value.

rather than death alone. The predictors associated with either death or transplant, especially lung function, may differ depending on the outcome. We chose to only include deaths prior to transplant as the outcome, as the clinical features and risk factors for death post-transplant are quite distinct. As indicated in the sensitivity analyses, this approach may introduce bias due to informative censoring; patients who are referred and ultimately go on to receive a transplant may be different from those who do not.

Our prediction model does have some limitations. Although the majority of predictor variables were obtained in the previous 12 months (for the 1-year model), or in the 12-month interval 1 year back (for the conditional 2-year model), the annual measurements recorded in the Canadian registry do not reflect acute deterioration in health status. In future, it would be useful to utilise encounter-based records and re-calibrate our model. The study population includes observations spanning >30 years, during which the prognosis and treatment of CF have changed considerably. In particular this has implications for the inclusion of *B. cepacia* complex in the model; there are multiple known strains of *B. cepacia* complex and not all are associated with increased mortality. The majority of patients in both Canada and the UK would have been infected with *B. cenocepacia*, which may not represent the strains of *B. cepacia* complex common in patients living with CF today. In cases where the specific strain of *B. cepacia* complex is not known, the model may overestimate mortality risk. Despite these limitations the model is clinically plausible, robust and closely mimics historical experience with CF mortality.

In conclusion, the mortality prediction models provide accurate risks of death over 1- and 2-year time horizons. The calculated probability of death may provide physicians and patients with important information with which to access key clinical decisions, such as transplant discussion and referral when the risk of death in the next 2 years is considered high. The model can be easily integrated into electronic health records or patient registries to automatically calculate the risk of death, which would facilitate use.

Acknowledgements: We are grateful to CF Canada and the UK Cystic Fibrosis Trust for use of the data in this study, and for the patients and families for allowing their data to be collected and used for research studies

Conflict of interest: None declared.

References

- 1 Stephenson AL, Sykes J, Stanojevic S, *et al.* Survival comparison of patients with cystic fibrosis in Canada and the United States: a population-based cohort study. *Ann Intern Med* 2017; 166: 537–546.
- 2 Stephenson AL, Tom M, Berthiaume Y, *et al.* A contemporary survival analysis of individuals with cystic fibrosis: a cohort study. *Eur Respir J* 2015; 45: 670–679.
- 3 Cystic Fibrosis Canada. 2012 Annual Report: The Canadian Cystic Fibrosis Registry. Toronto, Cystic Fibrosis Canada, 2014. Available from: www.cysticfibrosis.ca/de/action/download?downloads=16&file=dl_Registry+2012+-+English+FINAL+FOR+WEB.pdf
- 4 Liou TG, Adler FR, Fitzsimmons SC, *et al.* Predictive 5-year survivorship model of cystic fibrosis. *Am J Epidemiol* 2001; 153: 345–352.
- 5 Kerem E, Reisman J, Corey M, *et al.* Prediction of mortality in patients with cystic fibrosis. *N Engl J Med* 1992; 326: 1187–1191.
- 6 Aaron SD, Stephenson AL, Cameron DW, *et al.* A statistical model to predict one-year risk of death in patients with cystic fibrosis. *J Clin Epidemiol* 2015; 68: 1336–1345.
- 7 Mayer-Hamblett N, Rosenfeld M, Emerson J, *et al.* Developing cystic fibrosis lung transplant referral criteria using predictors of 2-year mortality. *Am J Respir Crit Care Med* 2002; 166: 1550–1555.
- 8 Nkam L, Lambert J, Latouche A, *et al.* A 3-year prognostic score for adults with cystic fibrosis. *J Cyst Fibros* 2017; 16: 702–708.
- 9 Keogh RH, Seaman SR, Barrett JK, *et al.* Dynamic prediction of survival in cystic fibrosis: a landmarking analysis using UK patient registry data. *Epidemiology* 2019; 30: 29–37.
- 10 Alaa AM, van der Schaar M. Prognostication and risk factors for cystic fibrosis *via* automated machine learning. *Sci Rep* 2018; 8: 11242.
- 11 Aaron SD, Ramsay T, Vandemheen K, *et al.* A threshold regression model for recurrent exacerbations in chronic obstructive pulmonary disease. *J Clin Epidemiol* 2010; 63: 1324–1331.
- 12 Chawla NVB KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357.