ERJ Express. Published on December 21, 2018 as doi: 10.1183/13993003.01943-2018

EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Cluster and CART analyses identify large subgroups of adults with cystic fibrosis at low risk of 10-year death

Pierre-Régis Burgel, Lydie Lemonnier, Clémence Dehillotte, Jenna Sykes, Sanja Stanojevic, Anne L. Stephenson, Jean-Louis Paillasseur

Please cite this article as: Burgel P-Régis, Lemonnier L, Dehillotte Clémence, *et al*. Cluster and CART analyses identify large subgroups of adults with cystic fibrosis at low risk of 10-year death. *Eur Respir J* 2018; in press (https://doi.org/10.1183/13993003.01943-2018).

Title: **Cluster and CART analyses identify large subgroups of adults with cystic fibrosis at low risk of 10-year death.**

Authors: Pierre-Régis Burgel[1], Lydie Lemonnier[2], Clémence Dehillotte[2], Jenna Sykes[3], Sanja Stanojevic[4], Anne L. Stephenson[3], Jean-Louis Paillasseur[5]

[1]University Paris Descartes, Sorbonne Paris Cité, Paris, France ; pierre-regis.burgel@aphp.fr

[2]Vaincre la Mucoviscisdose, Paris, France ; llemonnier@vaincrelamuco.org; cdehillotte@vaincrelamuco.org.

[3]University of Toronto, St-Michael's hospital, Li Ka Shing knowledge institute, Keenan research centre, department of medicine, adult CF program, 30, Bond street, 6th floor, Bond Wing, M5B 1W8 Toronto, ON, Canada; University of Toronto, institute of health policy, management and evaluation, Toronto, ON, Canada; SykesJ@smh.ca; stephensona@smh.ca

[4]The hospital for sick children, division of respiratory medicine, Toronto, ON, Canada; University of Toronto, institute of health policy, management and evaluation, Toronto, ON, Canada. sanja.stanojevic@sickkids.ca

[5]Effi-Stat, Paris France ; jean-louis.paillasseur@effi-stat.com

Corresponding author: Pierre-Régis Burgel, MD PhD
Service de Pneumologie, Hôpital Cochin
27 Rue du Faubourg St Jacques
75014 Paris, France
E-mail: pierre-regis.burgel@cch.aphp.fr
Tel: 33 1 58 41 23 67
Fax: 33 1 46 33 82 53

Take-home message: Large subgroups of adults with cystic fibrosis have very low rates of death at 10 years

Abstract

Our goal was to identify subgroups of cystic fibrosis (CF) adults at low risk of death within 10 years.

Factor analysis for mixed data followed by Ward's cluster analysis were conducted using 25 variables from 1572 French CF adults in 2005. Rates of death by subgroups were analysed over 10 years. An algorithm was developed using classification and regression tree (CART) to provide rules for the identification of subgroups of CF adults with low rates of death within 10 years. This algorithm was validated in 1376 Canadian CF adults.

Seven subgroups were identified by cluster analysis in French CF adults, including two subgroups with low (~5%) rates of death at 10 years: one subgroup (22% of patients) was composed of patients with non-classic CF, the other subgroup (17% of patients) was composed of patients with classic CF but low rates of *P. aeruginosa* infection and diabetes. An algorithm based on CART analysis of data in 2005 allowed to identify most French adults with low rates of death. When tested using data from Canadian CF adults in 2005, the algorithm identified 287/1376 (21%) patients at low risk (10-year death, 7.7%).

Large subgroups of CF adults share low risk of 10-year mortality.

Introduction

Cystic fibrosis (CF) is a genetic disease affecting at least 70000 patients in the world [1], and is characterized by a multisystemic disease involving primarily the lung, the pancreas and the liver [2]. CF used to be a devastating disease with premature death in children, but prognosis has improved over the past decades. Recent studies performed in countries with multidisciplinary CF care [3] revealed that almost all CF children now reach adult age [4-6] and that paediatric mortality has become very rare [7, 8]. Despite these improvements, CF adults remain at increased risk of respiratory failure, resulting in death [8] or the need for lung

transplantation [9, 10]. Although CF adults aim to achieve normal lives, they often face socioeconomical barriers in their daily lives. Among these barriers, CF adults are usually excluded from access to loan insurance, because insurance companies and banks assume that all CF adults are at high risk of premature death.

Several studies have examined factors associated with short-term (up to 5 years) prognosis using cohorts of CF patients usually obtained from national CF registries [11-14]. These studies have identified individual factors (e.g., low forced expiratory volume in 1 sec, low body mass index, pancreatic insufficiency, colonization with *P. aeruginosa* and/or *B. cepacia,* massive haemoptysis, pneumothorax) associated with poor prognosis that were sometimes combined into prognostic scores [13][15][16, 17][18]. Importantly, previous studies were generally designed to identify subgroups at high risk of death and/or lung transplantation and there is a lack of data regarding the clinical profile of CF adults at low risk of poor outcome.

Cluster analysis is a generic term for several statistical methods that allow grouping patients who share multiple characteristics [19]. Such exploratory analyses have been successfully implemented in large groups of patients with asthma [20], COPD [21, 22], and non-CF bronchiectasis [23] in which they have identified specific subgroups of patients sharing clinical outcomes (phenotypes) and/or biological pathways (endotypes). In CF adults, we identified only one study that used cluster analysis in 211 patients for identification of patient subgroups [24].

In the present study, we sought to identify subgroups of adult CF patients at low risk of death at 10 years. This objective was based on a request from the French government and private insurance companies to identify criteria that would allow proposing 10-year access to bank loans for adults with CF, who are currently unentitled to loan insurance because of their disease. Our strategy was to use a cluster analysis to identify subgroups of French CF adults with CF experiencing low death rates over 10 years and to use classification and regression

tree (CART) analysis for developing an algorithm that would allow the identification of these subgroups. We then externally tested this algorithm using data from Canadian CF adults.

Methods

Patients

The present study was conducted using the French Cystic Fibrosis Registry, which contains longitudinal data from at least 90% of CF patients in France [25]. The registry collects data once a year in CF patients followed in the network of accredited CF centres. Eligible patients were adult (≥18 years) CF patients living without lung transplantation in the 2005 registry. Patients listed for lung transplantation in 2005 were excluded due to the possibility that they underwent transplantation between their last visit (when the data were sent to the Registry) and the end of the year. Patients with any history of cancer were excluded from the analysis because this comorbidity may have affected prognosis. Patients with missing data for critical prognosis factors (spirometry and/or body mass index were also excluded from the analyses).
Data were further obtained from the Canadian CF Registry, which represents > 95% of Canadians with CF. Data were extracted from 2005 and selection of patients was performed by applying the methodology developed for the French CF Registry (see above).

Statistical analysis plan

First, French CF adults were classified into subgroups based on the results of a cluster analysis of data obtained in the cohort in 2005. The prognostic relevance of the identified subgroups was established by examining their association with death from any cause as per December 31st 2015 (10-year death). We next used classification and regression tree (CART) analysis to develop an algorithm aimed at the identification of clusters containing patients with low rates (~5%) of death at 10 years in the French CF Registry. This algorithm was

externally tested using data obtained in Canadian CF Registry. Survival and transplantation-free survival were analysed using Kaplan-Meier curves and Log-rank test. Risk of any death or risk of the combined event death without transplantation/occurrence of lung transplantation were analysed by Cox models. Data are presented as median (interquartile range, IQR) or n (%). Analyses were performed using SAS 9.2 (SAS Institute Inc., Cary, NC, USA) software.

Cluster analysis

Variables were selected for inclusion in the cluster analysis based on clinical knowledge (i.e., previously described association with prognosis in CF patients) and were usual descriptors of adult CF patients. Complete description and definition of the 25 variables selected in this analysis is provided in the online supplement (**Supplementary Table S1**). Briefly these variables included gender, age as of December 31st 2005, body mass index (BMI, kg/m^2), FEV_1 (% predicted), CFTR mutations, pancreatic status, *diabetes mellitus*, liver cirrhosis, haemoptysis, pneumothorax, airway infection and variables related to healthcare utilization and/or therapeutic management (long-term oxygen therapy, non-invasive ventilation, treatment with oral steroids for more than 3 months, insulin, number of intravenous antibiotic courses in 2005, number of hospitalisation≥48h in 2005). Identification of homogeneous subgroups of adults with CF was achieved using factor analysis for mixed data (FAMD) [26, 27], followed by classification of patients using Ward's agglomerative hierarchical cluster analysis [21, 28]. Briefly, FAMD allowed transformation of linear combinations of the 25 selected variables into 25 new independent variables (eigenvectors) called "components". The eigenvalue of each component is a measure of its variability. A component with an eigenvalue<1 contributes little to explain the relationships between original variables and thus is not subjected to further analysis. We then performed a cluster analysis based on significant components (i.e., with an eigenvalue>1) identified in the FAMD. Cluster analysis was

performed using the Ward's method, for which grouping was based on quantitative measures of similarity procedure (minimum within cluster sum of square), such that subjects in the same cluster were more similar to each other than to subjects in another cluster. We used pseudo F and pseudo t^2 statistics and visual assessment of the dendrogram to determine the optimal number of clusters in the data [21].

Prognostic outcomes

To examine the prognostic relevance of the identified subgroups (clusters) of patients, we examined vital status in the French CF Registry, which contains reliable data on the occurrence of death in CF patients in France [8]. We first analyzed data using rates of any death at 5 and 10 years. Because lung transplantation has become the standard of care in CF patients with severe respiratory insufficiency, we also examined the occurrence of lung transplantation and death rates according to transplantation status (death without lung transplantation vs. death after lung transplantation). Finally, we examined a combined outcome of death without lung transplantation or the occurrence of lung transplantation, as described previously [13].

CART analysis

The development of an algorithm to assign French adult CF patients to the subgroups at low rates of 10-year death (cluster 1 and cluster 2) was achieved using CART analysis, as described previously [22]. Briefly, variables included in this analysis were those selected for the cluster analysis (n=25 variables). Variables retained by the analysis and threshold values of these variables were obtained by CART analysis (see online supplement for detailed explanation). This algorithm was then externally tested using data from the Canadian CF Registry.

RESULTS

Patients

The French CF Registry contained 1942 adult CF patients as of December 31st 2005. A total of 370 subjects were excluded from the analysis; reasons for excluding these patients included a previous history of lung transplantation, being on waiting list for transplantation in 2005, death in 2005, previous history of cancer and missing data for lung function and/or body mass index. A flow chart describing the study population is presented in **Figure 1**. Thus, further analyses were conducted on 1572 adult CF patients; clinical characteristics are presented in **Table 1.** From December 31st 2005 to December 31st 2015, lung transplantation was performed in 402 (25.6%) patients and death occurred in 232 (14.8%) patients, including 134 (8.5%) patients who had not undergone lung transplantation.

Cluster analysis

Cluster analysis was performed on 25 variables (listed in **Table S1**), using FAMD followed by Ward's hierarchical classification. The dendrogram resulting from this analysis is shown in **Figure S1**. Based on pseudo *F* and pseudo t^2 statistics, and visual assessment of the dendrogram, the data could be optimally classified into 7 clusters. **Table 2** shows clinical characteristics in the 1572 adult CF patients according to these 7 clusters. Cluster 1 was composed of young adults (median age, 21.8 years) with classic CF characterized by diagnosis early in life and pancreatic insufficiency, but with mild to moderate impairment in lung function, low rates of chronic *Pseudomonas aeruginosa* airway infection and no diabetes. Cluster 2 was composed of older adults (median age, 30.2 years) with non-classic CF characterized by late diagnosis, high rates of class IV or V CFTR mutations or incomplete genotype, high rates of pancreatic sufficiency, mild to moderate impairment in lung function and moderate rates of chronic *P. aeruginosa* airway infection. Cluster 3 was composed of

patients with classic CF, moderate to severe lung function impairment, high rates of *P. aeruginosa* airway infection; two thirds of these patients had IV antibiotics and one third had diabetes mellitus. Cluster 4 was composed of patients with classic CF, high rates of pneumothorax (and thoracic surgery), and high rates of airway non-tuberculous mycobacteria. Cluster 5 was composed of patients with classic CF and very high rates of respiratory insufficiency treated with long-term oxygen therapy or non-invasive ventilation. Cluster 6 was composed of patients with classic CF and chronic *B. cepacia* airway infection. Cluster 7 was composed of patients with classic CF, very high rates of treated aspergillosis and diabetes mellitus; all these patients were on oral steroids.

Prognostic relevance of the clusters

To assess the prognostic relevance of the 7 clusters identified in the cluster analysis, we next examined rates of death and/or lung transplantation over 10 years in each cluster. Kaplan-Meier analysis of (1) survival (outcome=any death) or (2) transplantation-free survival (outcome=death without lung transplantation/occurrence of lung transplantation) are shown in **Figure 2.** Patients in cluster 1 (classic CF/low *P. aeruginosa*) and cluster 2 (non-classic CF) had the best prognosis with ~5% death at 10 years; 10-15% of these patients underwent lung transplantation. Patients in cluster 3 (classic CF/moderate) and 4 (classic CF/pneumothorax/nontuberculous mycobacteria) had less than 10% deaths at 5 years, which increased to 13.5 and 20.6% at 10 years, respectively. Patients in cluster 5 (classic CF/respiratory insufficiency) had very high rates (67%) of lung transplantation and 45.6% rates of death at 10 years. Patients in cluster 6 (classic CF/*B. cepacia*) had 51.4% death at 10 years with only 32.4% lung transplantation. Patients in cluster 7 (classic CF/ABPA) had 37.3% death at 10 years with 31.4% lung transplantation. **Table 3** summarizes the main descriptors of patients in each cluster and prognostic outcomes in each cluster. Comparison of

rates of any death or rates of death without transplantation/occurrence of lung transplantation are shown in **Figure 3**.

Algorithm for the identification of CF adults at low rates of 10-year death

Because cluster 1 and cluster 2 were shown to have low (~5%) and comparable death rates at 10 years, we next develop an algorithm that would allow to identify these patients (n=613 patients; 262 patients in Cluster 1 + 351 patients in Cluster 2), using data obtain at study entry in 2005. The algorithm is presented in **Figure 4**: patients with a least one factor negatively affecting the prognosis (see list on **Figure 4**) were not considered at low risk; patient who did not have these negative factors were at low risk if they had no *P. aeruginosa* or, when they had P. aeruginosa if they were pancreatic sufficient. The algorithm identified 515/613 (84% of patients in cluster 1 and 2) patients and rate of death in these 515 patients was 3.9% at 10 years. Kaplan-Meier analysis comparing 10-year death at lower risk vs. higher risk according to the algorithm in French patients is presented in **Figure 5A**.

Next, we tested this algorithm using data from adult patients in the Canadian CF registry in 2005 (see **Supplementary Table S3-S4** for characteristics of patients). Three variables necessary for excluding patients at higher risk according to the algorithm were unavailable in the Canadian CF registry in 2005: infection with non-tuberculous mycobacteria, use of systemic steroids for more than 3 months, and use of non-invasive ventilation. The algorithm was thus used on Canadian data without considering these variables: it identified 287/1376 (21%) patients at low risk (10-year death, 7.7%). Kaplan-Meier analysis comparing 10-year death at lower risk vs. higher risk according to the algorithm in Canadian patients is presented in **Figure 5B.**

DISCUSSION

In the present study, we sought to identify subgroups of adults with CF at low risk of mortality at 10-years. We performed a cluster analysis on data from adult CF patients contained in the French CF Registry in 2005 and found that a large proportion (39%) of CF adults had low (~5%) death rates at 10 years. We next used CART analysis to develop an algorithm that allowed the identification of 84% of these patients, using data collected in 2005. The algorithm was then externally tested using data from the Canadian CF Registry in 2005 showing consistent findings. These results indicate that many CF adults have a favourable long-term prognosis, a finding that has important socioeconomical consequences.

An important finding was that 39% of adult CF patients, grouped in cluster 1 (n=262 patients, 17%) and in cluster 2 (n=351 patients, 22%), shared rather good prognoses with 1% of death at 5 years and 5% of death at 10 years. These data indicate that a large proportion of adult CF patients have improved prognosis in the current era, as compared with previous decades, which was in part related to lung transplantation (which occurred in 10-15% patients in these clusters). Most patients with at least one class IV or V mutation were found in cluster 2, confirming previous data showing that patients with residual function mutations often have non-classic CF (characterized by late diagnosis and milder clinical features) and lower risk of death [29, 30]. Cluster 2 was also composed of a large group of patients with incomplete CFTR genotypes, in whom CF diagnosis could be questioned; this finding was in agreement with a previous report showing that patients found in CF registries and less well documented CF diagnosis had usually milder disease [31]. An important and novel finding of our study was that patients in cluster 1, who had a prognosis comparable to patients in cluster 2, had classic CF characterized by high occurrence of two loss of function CFTR alleles, diagnosis early in life and pancreatic insufficiency. Interestingly, these latter patients were younger than patients in other clusters, had higher lung function and nutritional status, very low rates of

chronic *P. aeruginosa* infection and no diabetes. These findings suggest that the prognosis is currently improving in some adults with classic CF. We speculate that these findings are related to improvement in lifelong care (e.g., nutrition, strategies for eradication of *Pseudomonas aeruginosa* infection).

The study also highlighted the fact that despite comprehensive CF care and lung transplantation, approximately 10% of patients belonged to subgroups with poor prognoses at 10 years. Thus, subjects with respiratory insufficiency (cluster 5, 8% of patients) and subjects with *B. cepacia* infection (cluster 6; 2% of patients) had comparable mortality rates at 10 years (45.6 and 51.4%, respectively). Of note, rates of lung transplantation were two times lower in cluster 6 vs. cluster 5, confirming that *B. cepacia* complex infection is still often considered a contraindication to lung transplantation [9].

The present study has several strengths. It was conducted using a nationwide Registry covering over 90% of the French CF population, with limited amount of missing data, which captures reliable data on CF patients (including patients who underwent lung transplantation) and survival [8]. Subgroups of CF patients were constructed using factorial and cluster analyses of variables with clinical relevance in the description of CF adults; importantly the relevance of these subgroups was established using prognostic data (e.g., rates of death and/or lung transplantation) that did not participate in the construction of subgroups. CART analysis allowed identification of a relatively simple algorithm that highlighted patients at low rates of 10-year death in France; importantly this algorithm was externally tested using data from the Canadian CF Registry. We also recognize limitations. The present study was designed to identify subgroups of individuals sharing comparable characteristics and long-term prognosis but was not intended at determining prognosis at the individual level in adults with CF. Thus, belonging to a group considered “not at low risk” does not necessarily mean that the individual prognosis of the patient was poor, but rather that the 10-year risk of death in this

group of patients was considered too high for providing loan insurance by banks and insurance companies. Although our algorithm identified a subgroup of adult CF patients at low risk of death, it did not identify all adults at low risk of death (also see **Supplementary Tables S5-S6**). Some of the data (i.e., infection with non-tuberculous mycobacteria, use of systemic steroids for more than 3 months, and use of non-invasive ventilation) necessary to run the algorithm were unavailable in the Canadian CF Registry in 2005, which may have resulted in the selection of a limited number of patients at higher risk of death by the algorithm. However, the algorithm consistently identified patients at low rates (7.7%) of death at 10 years in Canadian CF adults despite this limitation. The effect of missing variables was likely limited by the low prevalence of patients treated with steroids, having NTM infection or treated with NIV and by the fact that some of these patients were probably excluded by other variables (e.g., patients treated with NIV have usually low FEV_1 and/or oxygen therapy). Finally, the study was limited to CF adults and excluded the paediatric population. This choice was related to the fact that death and lung transplantation have become extremely low in paediatric CF patients in France [8], as in other developed countries [7]. Although our data were obtained using data from France and Canada, data generated from these two countries are likely relevant to other countries with comparable access to specialized CF care and to lung transplantation.

The present study highlighted the recent evolution of prognosis in CF adults. Although studies performed 10 to 15 years ago suggested that intrinsic patient characteristics (e.g., CFTR genotype associated with residual CFTR function) were associated with improved prognosis [29, 30], the present study extends these findings by suggesting that improved prognosis is observed in some adults (cluster 1) with classic CF who were exposed to appropriate CF care (e.g., eradication strategies for *P. aeruginosa* [32]) and who can benefit from lung transplantation. The finding that our algorithm was able to detect a large proportion of

patients with low rates of 10-year death in France and in Canada, led to a recent agreement between the French Government and insurance companies for providing access to loan insurance for CF adults at low risk of 10-year death identified with this algorithm. Although the agreement is limited to a fraction of the overall CF population, it constitutes an important first step on the road to providing normal access to mortgage insurance for all patients with CF [33].

In conclusion, CF has progressively evolved from a devastating paediatric disease, causing early death in children, to a disease where almost all patients reach adult age in developed countries. The present study further indicated that large subgroups of adults with CF have improved longevity with low rates of 10 year-death, confirming that prognosis continues to improve in the CF adult population. These latter data further indicate that lifelong specialized CF care results in better health status and prolonged survival, even in subjects with classic CF. We speculate that the future care of CF patients will presumably be a mix of patients with milder disease (especially younger adults who had lifelong exposure to high quality CF care) and older patients with more severe disease often due to lack of intensive CF care in their early years. Of note, these results were obtained at a time when novel drugs targeting the CFTR defect where not widely available. Future use of these disease modifying agents will likely further reduce the survival gap that still exists between CF patients and the general population.

Funding source: This study was funded by the French CF association Vaincre la Mucoviscidose

References

1. Stephenson AL, Stanojevic S, Sykes J, Burgel PR. The changing epidemiology and demography of cystic fibrosis. *Presse Med* 2017: 46(6 Pt 2): e87-e95.
2. Elborn JS. Cystic fibrosis. *Lancet* 2016: 388(10059): 2519-2531.
3. Elborn JS, Bell SC, Madge SL, Burgel PR, Castellani C, Conway S, De Rijcke K, Dembski B, Drevinek P, Heijerman HG, Innes JA, Lindblad A, Marshall B, Olesen HV, Reimann AL, Sole A, Viviani L, Wagner TO, Welte T, Blasi F. Report of the European Respiratory Society/European Cystic Fibrosis Society task force on the care of adults with cystic fibrosis. *Eur Respir J* 2016: 47(2): 420-428.
4. McCormick J, Mehta G, Olesen HV, Viviani L, Macek M, Jr., Mehta A, European Registry Working Group. Comparative demographics of the European cystic fibrosis population: a cross-sectional database analysis. *Lancet* 2010: 375(9719): 1007-1013.
5. Burgel PR, Bellis G, Olesen HV, Viviani L, Zolin A, Blasi F, Elborn JS. Future trends in cystic fibrosis demography in 34 European countries. *Eur Respir J* 2015: 46(1): 133-141.
6. Burgel PR, Bellis G, Elborn JS. Modelling future trends in cystic fibrosis demography using the French Cystic Fibrosis Registry: update and sensitivity analysis. *Eur Respir J* 2017: 50 (2): 1700763.
7. Urquhart DS, Thia LP, Francis J, Prasad SA, Dawson C, Wallis C, Balfour-Lynn IM. Deaths in childhood from cystic fibrosis: 10-year analysis from two London specialist centres. *Arch Dis Child* 2013: 98(2): 123-127.
8. Martin C, Hamard C, Kanaan R, Boussaud V, Grenet D, Abely M, Hubert D, Munck A, Lemonnier L, Burgel PR. Causes of death in French cystic fibrosis patients: The need for improvement in transplantation referral strategies! *J Cyst Fibros* 2016: 15(2): 204-212.
9. Hirche TO, Knoop C, Hebestreit H, Shimmin D, Sole A, Elborn JS, Ellemunter H, Aurora P, Hogardt M, Wagner TO. Practical guidelines: lung transplantation in patients with cystic fibrosis. *Pulmonary medicine* 2014: 2014: 621342.
10. Thabut G, Christie JD, Mal H, Fournier M, Brugière O, Leseche G, Castier Y, Rizopoulos D. Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. *Am J Respir Crit Care Med* 2013: 187: 1335-1340.
11. Kerem E, Reisman J, Corey M, Canny GJ, Levison H. Prediction of mortality in patients with cystic fibrosis. *N Engl J Med* 1992: 326(18): 1187-1191.
12. Liou TG, Adler FR, Fitzsimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. *Am J Epidemiol* 2001: 153(4): 345-352.
13. Nkam L, Lambert J, Latouche A, Bellis G, Burgel PR, Hocine MN. A 3-year prognostic score for adults with cystic fibrosis. *J Cyst Fibros* 2017: 16(6): 702-708.
14. McCarthy C, Dimitrov BD, Meurling IJ, Gunaratnam C, McElvaney NG. The CF-ABLE score: a novel clinical prediction rule for prognosis in patients with cystic fibrosis. *Chest* 2013: 143: 1358-1364.
15. McKone EF, Goss CH, Aitken ML. CFTR genotype as a predictor of prognosis in cystic fibrosis. *Chest* 2006: 130(5): 1441-1447.
16. Flume PA, Strange C, Ye X, Ebeling M, Hulsey T, Clark LL. Pneumothorax in cystic fibrosis. *Chest* 2005: 128(2): 720-728.
17. Flume PA, Yankaskas JR, Ebeling M, Hulsey T, Clark LL. Massive hemoptysis in cystic fibrosis. *Chest* 2005: 128(2): 729-738.
18. Lewis C, Blackman SM, Nelson A, Oberdorfer E, Wells D, Dunitz J, Thomas W, Moran A. Diabetes-related mortality in adults with cystic fibrosis. Role of genotype and sex. *Am J Respir Crit Care Med* 2015: 191(2): 194-200.
19. Burgel PR, Paillasseur JL, Roche N. Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities. *Biomed Res Int* 2014: 2014: 420134.
20. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, Wardlaw AJ, Green RH. Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008: 178(3): 218-224.
21. Burgel PR, Paillasseur JL, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, Court-Fortune I, Perez T, Carre P, Roche N, Initiatives BSC. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010: 36(3): 531-539.

22. Burgel PR, Paillasseur JL, Janssens W, Piquet J, Ter Riet G, Garcia-Aymerich J, Cosio B, Bakke P, Puhan MA, Langhammer A, Alfageme I, Almagro P, Ancochea J, Celli BR, Casanova C, de-Torres JP, Decramer M, Echazarreta A, Esteban C, Gomez Punter RM, Han MK, Johannessen A, Kaiser B, Lamprecht B, Lange P, Leivseth L, Marin JM, Martin F, Martinez-Camblor P, Miravitlles M, Oga T, Sofia Ramirez A, Sin DD, Sobradillo P, Soler-Cataluna JJ, Turner AM, Verdu Rivera FJ, Soriano JB, Roche N. A simple algorithm for the identification of clinical COPD phenotypes. *Eur Respir J* 2017: 50(5).
23. Aliberti S, Lonni S, Dore S, McDonnell MJ, Goeminne PC, Dimakou K, Fardon TC, Rutherford R, Pesci A, Restrepo MI, Sotgiu G, Chalmers JD. Clinical phenotypes in adult patients with bronchiectasis. *Eur Respir J* 2016: 47(4): 1113-1122.
24. Conrad DJ, Bailey BA. Multidimensional clinical phenotyping of an adult cystic fibrosis patient population. *PloS one* 2015: 10(3): e0122705.
25. Vaincre la Mucoviscidose et Ined. Registre Français de la Mucoviscidose – Bilan des données 2015. [cited January 28th 2017]; Available from: http://www.vaincrelamuco.org/face-la-mucoviscidose/registre-et-muco-en-chiffres/valorisation-des-donnees
26. Wikipedia. Factor analysis of mixed data. 2016 [cited 2018 September 25th]; Available from: https://en.wikipedia.org/wiki/Factor_analysis_of_mixed_data
27. Pagès J. Multiple Factor Analysis for Mixed Data [article in French]. *Rev Statistique Appliquée* 2004: 52: 93-111.
28. Burgel PR, Paillasseur JL, Peene B, Dusser D, Roche N, Coolen J, Troosters T, Decramer M, Janssens W. Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. *PloS one* 2012: 7(12): e51048.
29. McKone EF, Goss CH, Aitken ML. CFTR genotype as a predictor of prognosis in cystic fibrosis. *Chest* 2006: 130: 1441-1447.
30. McKone EF, Emerson SS, Edwards KL, Aitken ML. Effect of genotype on phenotype and mortality in cystic fibrosis: a retrospective cohort study. *Lancet* 2003: 361: 1671-1676.
31. Thomas M, Lemonnier L, Gulmans V, Naehrlich L, Vermeulen F, Cuppens H, Castellani C, Norek A, De Boeck K. Is there evidence for correct diagnosis in cystic fibrosis registries? *J Cyst Fibros* 2014: 13(3): 275-280.
32. Langton Hewer SC, Smyth AR. Antibiotic strategies for eradicating Pseudomonas aeruginosa in people with cystic fibrosis. *The Cochrane database of systematic reviews* 2017: 4: Cd004197.
33. Quon BS, Rowe SM. New and emerging targeted therapies for cystic fibrosis. *Bmj* 2016: 352: i859.

Figure legends

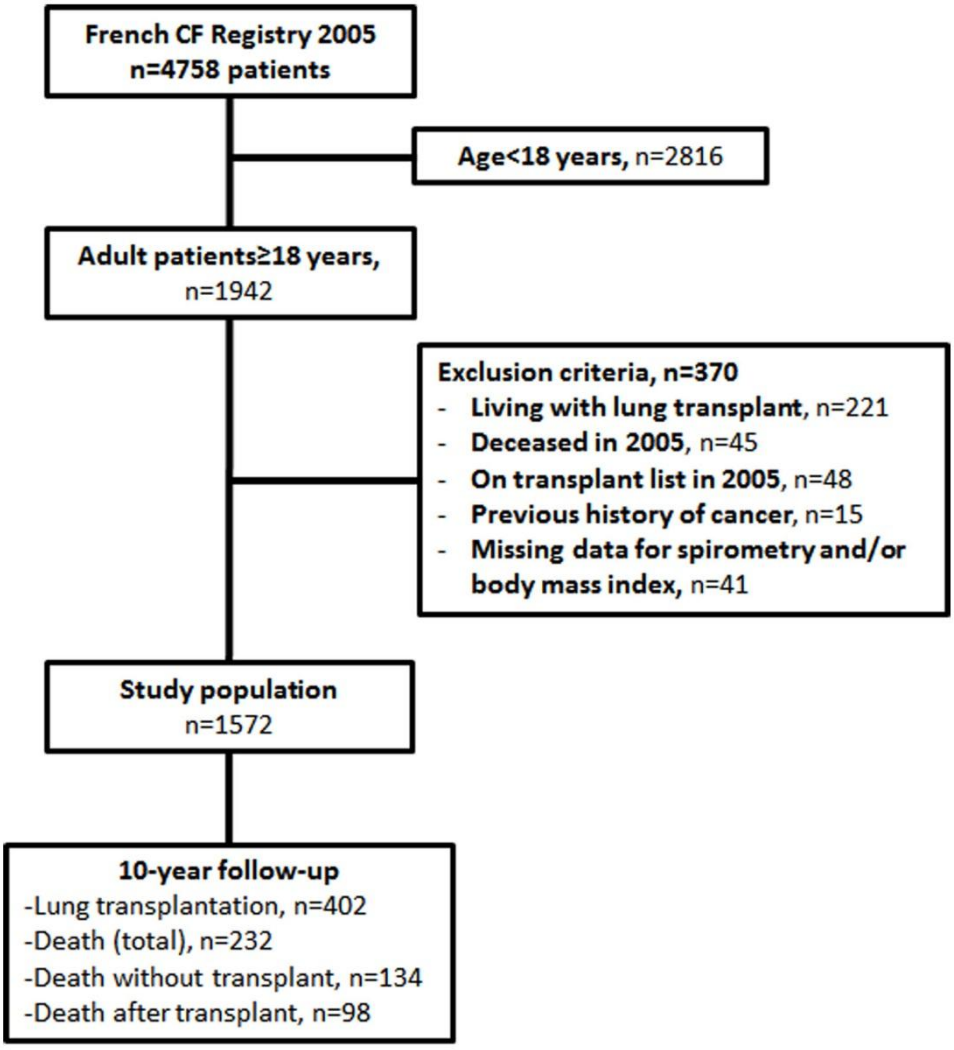
Figure 1. CONSORT diagram of the patients included in the study

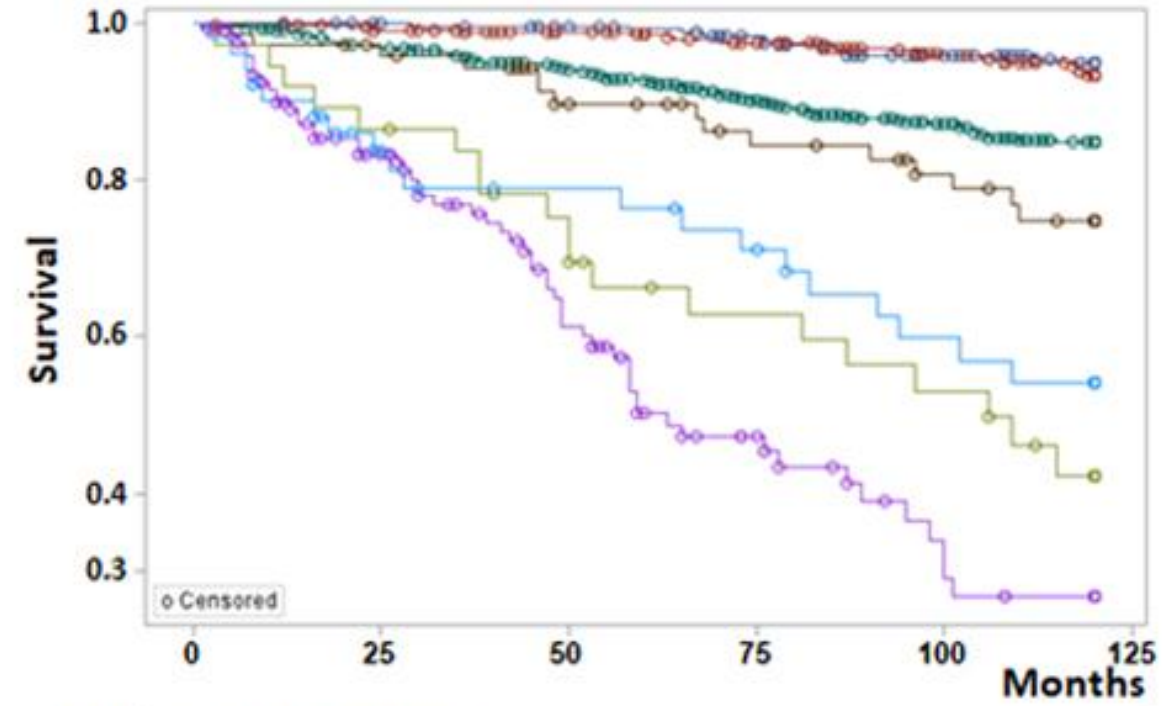
Figure 2. Kaplan–Meier curves for assessing (A) Survival (outcome= any death) or (B) Transplant-free survival (outcome=death without lung transplantation or the occurrence of lung transplantation). Log-Rank test indicated that all subgroups, except cluster 2, had increased risk of outcomes (P<0.001), compared to cluster 1.

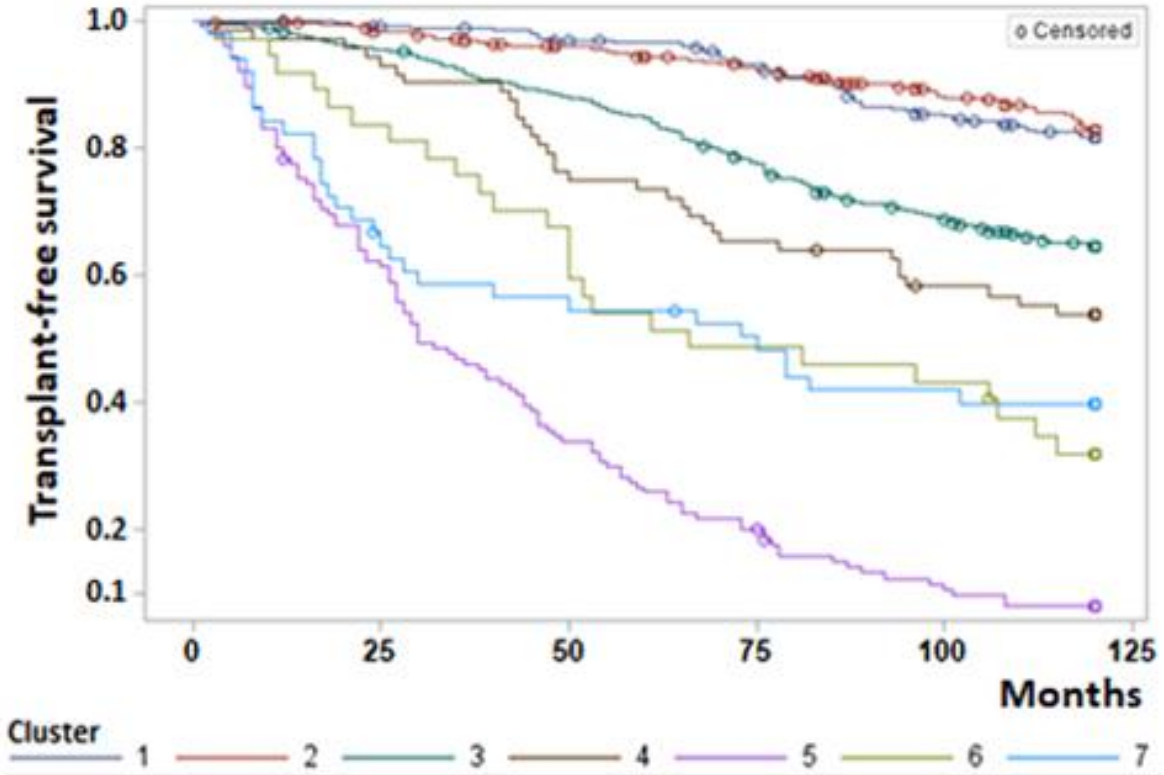
Figure 3. Relative risks of any death (A and C) or death without transplantation/occurrence of lung transplantation (B and D) at 5 years (A and B) and 10 years (C and D) among clusters of adult CF patients. Horizontal bars show hazard ratios and 95% confidence intervals of risks between classes. For example, subjects in cluster 3 have a 9.8-fold (95% CI 2.4–40.4) increased risk of any death at 5 years, when compared with subjects in cluster 1.

Figure 4. Algorithm for identification of CF adults at a low risk vs. not low risk of death at 10 years in French CF adults. The algorithm was obtained based on CART analysis of variables initially used for the cluster analysis (see Methods), with the objective of classifying the larger number of patients in the low risk subgroups (cluster 1 and 2). †, these variables were unavailable for external validation of the algorithm in the Canadian CF registry

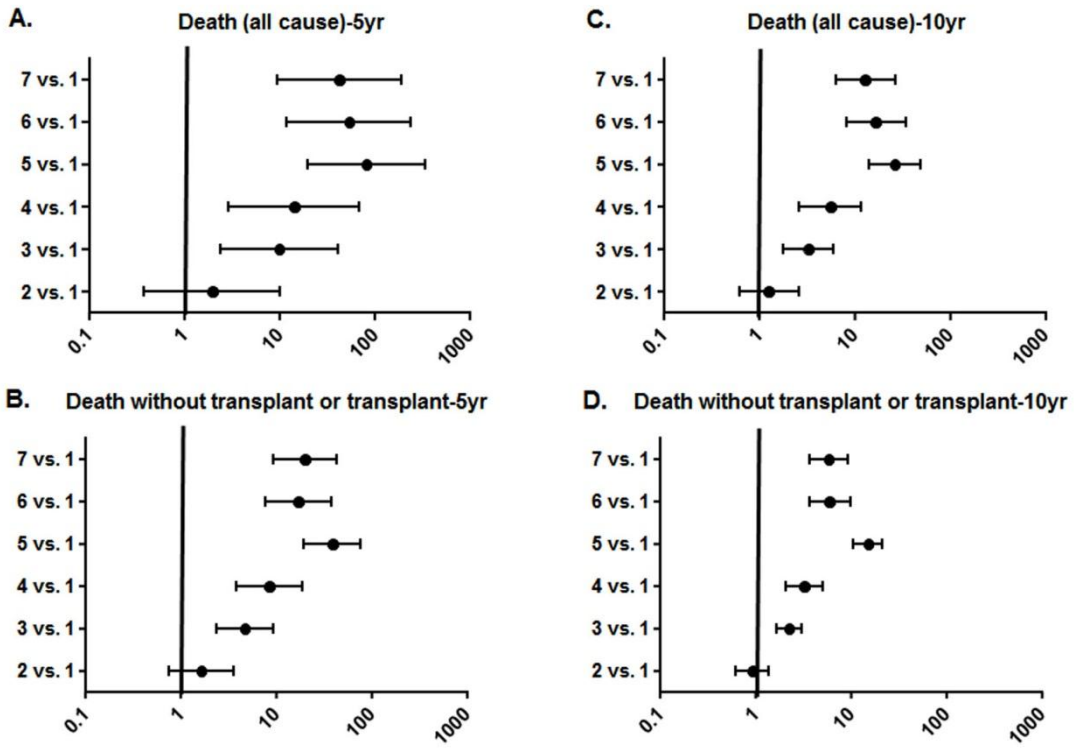
Figure 5. Kaplan-Meier analysis for assessing death at 10 years in subjects at low risk vs. not low risk identified by the algorithm. A. French CF registry. B. Canadian CF registry.

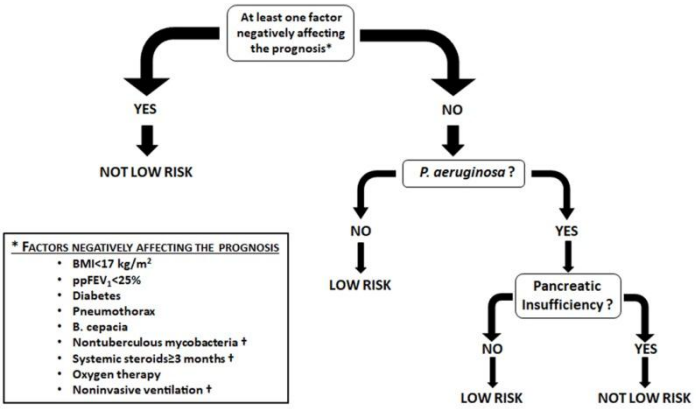

French CF Registry 2005
n=4758 patients
Age<18 years, n=2816
Adult patients≥18 years,
n=1942
Exclusion criteria, n=370
- Living with lung transplant, n=221
- Deceased in 2005, n=45
- On transplant list in 2005, n=48
- Previous history of cancer, n=15
- Missing data for spirometry and/or body mass index, n=41
Study population
n=1572
10-year follow-up
-Lung transplantation, n=402
-Death (total), n=232
-Death without transplant, n=134
-Death after transplant, n=98


Survival
1.0
0.8
0.6
0.4
0.3
Censored
0
25
50
75
100
125
Months


Transplant-free survival
1.0
0.8
0.6
0.4
0.2
0.1
Censored
0
25
50
75
100
125
Months

Cluster
1
2
3
4
5
6
7


A.
Death (all cause)-5yr
7 vs. 1
6 vs. 1
5 vs. 1
4 vs. 1
3 vs. 1
2 vs. 1
0.1
1
10
100
1000
C.
Death (all cause)-10yr
7 vs. 1
6 vs. 1
5 vs. 1
4 vs. 1
3 vs. 1
2 vs. 1
0.1
1
10
100
1000
B.
Death without transplant or transplant-5yr
7 vs. 1
6 vs. 1
5 vs. 1
4 vs. 1
3 vs. 1
2 vs. 1
0.1
1
10
100
1000
D.
Death without transplant or transplant-10yr
7 vs. 1
6 vs. 1
5 vs. 1
4 vs. 1
3 vs. 1
2 vs. 1
0.1
1
10
100
1000


At least one factor negatively affecting the prognosis*
YES
NOT LOW RISK
NO
P. aeruginosa ?
NO
LOW RISK
YES
Pancreatic Insufficiency ?
NO
LOW RISK
YES
NOT LOW RISK
* FACTORS NEGATIVELY AFFECTING THE PROGNOSIS
• BMI<17 kg/m²
• ppFEV1<25%
• Diabetes
• Pneumothorax
• B. cepacia
• Nontuberculous mycobacteria †
• Systemic steroids≥3 months †
• Oxygen therapy
• Noninvasive ventilation †

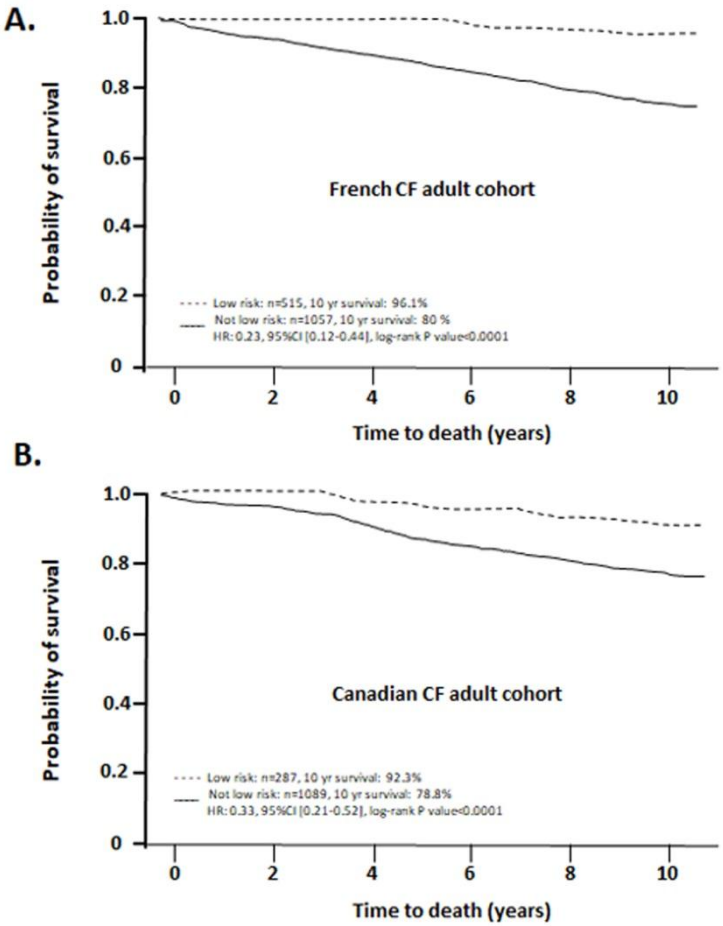

A.
Probability of survival
1.0
0.8
0.6
0.4
0.2
0
French CF adult cohort
Low risk: n=515, 10 yr survival: 96.1%
Not low risk: n=1057, 10 yr survival: 80 %
HR: 0.23, 95%CI [0.12-0.44], log-rank P value<0.0001
0
2
4
6
8
10
Time to death (years)
B.
Probability of survival
1.0
0.8
0.6
0.4
0.2
0
Canadian CF adult cohort
Low risk: n=287, 10 yr survival: 92.3%
Not low risk: n=1089, 10 yr survival: 78.8%
HR: 0.33, 95%CI [0.21-0.52], log-rank P value<0.0001
0
2
4
6
8
10
Time to death (years)

Table 1. Main characteristics of 1572 adults with cystic fibrosis in the French CF Registry as per December 31st 2005

Variables	French CF adults n=1572
Male, % (n)	53.7 (844)
Age, year	24.9 [21.0; 30.9]
Age at diagnosis, year*	1.6 [0.2; 10.2]
Body Mass Index, kg/m^2	19.9 [18.3; 21.7]
CFTR genotype*	
- Class I, II, III/Class I, II, III	61.1 (961)
- At least one class IV, V mutation	11.9 (187)
- Class I, II, III/unclassified	11.2 (176)
- Unclassified/unclassified	1.6 (25)
- Incomplete genotype	
-all	14.1 (223)
-one CFTR mutation	8.7 (136)
-no CFTR mutation	5.4 (87)
FEV_1, L	1.95 [1.32; 2.82]
FEV_1, % predicted	57 [38; 79]
Pancreatic insufficiency, % (n)	82.1 (1291)
Airway infection	
-Pseudomonas aeruginosa	61.7 (970)
-Burkholderia cepacia	2.8 (44)
-MRSA	15.7 (247)
-MSSA	45.5 (716)
-Non-Tuberculous Mycobacteria	2.3 (36)
Liver cirrhosis, % (n)	11.0 (173)
Diabetes mellitus, % (n)	24.6 (387)
-insulin treated	14.4 (227)
Haemoptysis in 2005	10.5 (165)
Pneumothorax in 2005	3.5 (55)
Azithromycin	50.9 (800)
Oral steroids	3.5 (55)
Long-Term Oxygen Therapy	8.8 (139)
Non-Invasive Ventilation, % (n)	4.5 (71)
IV antibiotic courses in 2005, n	1 [0; 3]
Hospitalizations in 2005, n	0 [0;1]

***** missing n=70; MRSA: methicillin-resistant *S. aureus*; MSSA : methicillin-susceptible *S. aureus*

** See **Supplementary Table S2** for details on classification of CFTR mutation

All data are % (n) or median [Quartile 1-Quartile 3] unless otherwise stated.

Table 2. Characteristics of 1572 adults with cystic fibrosis according to the 7 subgroups identified by cluster analysis

	Cluster 1, n=262	Cluster 2, n=351	Cluster 3, n=674	Cluster 4, n=72	Cluster 5, n=125	Cluster 6, n=37	Cluster 7, n=51
Male, % (n)	65 (170)	50 (176)	49 (328)	57 (41)	62 (78)	62 (23)	55 (28)
Age, year	21.8 [19.9; 26.4]	30.2 [23.7; 40.4]	24.4 [20.7; 29.9]	24.5 [20.3; 33.1]	25.2 [22.3; 29.9]	23.5 [21.7; 27.8]	25.5 [21.0; 33.3]
Age at diagnosis, year*	1.1 [0.1; 7.5] *missing, n=12*	17.4 [4.3; 31.4] *missing, n=18*	0.7 [0.1; 3.8] *missing, n=26*	1.9 [0.1; 11.7] *missing, n=5*	0.7 [0.2; 2.7] *missing, n=7*	0.9 [0.2; 5.6] *missing, n=1*	0.9 [0.2; 5.6] *missing, n=1*
CFTR genotype							
- Class I, II, III/Class I, II, III	78.2 (205)	4.6 (16)	79.4 (535)	69.4 (50)	75.2 (94)	75.7 (28)	64.7 (33)
- At least one class IV, V mutation	0.4 (1)	44.7 (157)	1.5 (10)	12.5 (9)	3.2 (4)	2.7 (1)	9.8 (5)
- Other mutation combinations*	12.2 (32)	17.3 (61)	10.5 (71)	9.7 (7)	12.0 (15)	18.7 (7)	15.7 (8)
- Incomplete genotype	9.2 (24)	33.3 (117)	8.6 (58)	8.3 (6)	9.6 (12)	2.7 (1)	9.8 (5)
FEV_1, L	2.63 [1.85; 3.30]	2.35 [1.57; 3.24]	1.87 [1.32; 2.56]	1.55 [1.14; 2.44]	0.94 [0.73; 1.21]	1.61 [1.24; 2.44]	1.65 [1.19; 2.38]
FEV_1, % predicted	71.8 [50.5; 88.1]	70.8 [48.2; 92.7]	54.4 [38.9; 73.8]	41.9 [33.6; 66.2]	25.8 [20.8; 32.2]	46.1 [33.3; 67.4]	48.9 [33.3; 67.8]
Pancreatic insufficiency, % (n)	93.1 (244)	37.6 (132)	97.0 (654)	86.1 (62)	96.0 (120)	94.6 (35)	86.3 (44)
Body Mass Index, kg/m^2	20.3 [19.0; 21.8]	21.5 [19.7; 23.7]	19.5 [18.0; 21.0]	18.8 [17.2; 20.9]	18.0 [16.7; 19.6]	19.0 [17.7; 21.2]	20.0 [17.9; 22.8]
Airway infection							
Pseudomonas aeruginosa	28.6 (75)	39.9 (140)	82.0 (553)	65.3 (47)	86.4 (108)	32.4 (12)	68.6 (35)
Burkholderia cepacia	0	0	0	0	5.6 (7)	100 (37)	0
MSSA	89.7 (235)	38.5 (135)	34.7 (234)	51.4 (37)	28.0 (35)	45.9 (17)	45.1 (23)
MRSA	2.3 (6)	8.0 (28)	23.4 (158)	15.3 (11)	19.2 (24)	18.9 (7)	25.5 (13)
Non-Tuberculous Mycobacteria	0	0	0	47.2 (34)	0.8 (1)	0	2.0 (1)
Comorbidities							
Liver cirrhosis, % (n)	11.8 (31)	2.6 (9)	13.6 (92)	11.1 (8)	17.6 (22)	16.2 (6)	9.8 (5)
Diabetes mellitus, % (n)	0.4 (1)	0.3 (1)	36.1 (243)	12.4 (9)	40.8 (51)	35.1 (13)	41.2 (21)
Haemoptysis in 2005	0.4 (1)	3.1 (11)	15.3 (103)	13.9 (10)	22.4 (28)	10.8 (4)	15.7 (8)
Pneumothorax in 2005	0	0.3 (1)	0	54.2 (39)	10.4 (13)	2.7 (1)	2.0 (1)
Treated aspergillosis	17.9 (47)	15.7 (55)	27.3 (184)	22.2 (47)	35.2 (44)	27.0 (10)	66.7 (34)
Treatment							
Pancreatic enzyme	92.0 (241)	39.3 (138)	95.8 (646)	86.1 (62)	97.6 (122)	94.6 (35)	86.3 (44)
Azithromycin	20.6 (54)	33.0 (116)	67.5 (455)	38.9 (28)	78.4 (98)	51.4 (19)	58.8 (30)
Oral steroids	0	0	0	0	2.4 (3)	2.4 (1)	100 (51)
Long-Term Oxygen Therapy	0	1.4 (5)	1.6 (11)	8.3 (6)	81.6 (102)	8.1 (3)	23.5 (12)
Non-Invasive Ventilation, % (n)	0	0.3 (1)	0	0	52.8 (66)	0	7.8 (4)
Patients with IV antibiotics in 2005	26.7 (70)	31.1 (109)	69.6 (469)	63.9 (46)	93.6 (117)	73.0 (27)	72.5 (37)
IV antibiotic course/patient in 2005, n	0 [0; 1]	0 [0; 1]	2 [0; 3]	1 [0; 3]	4 [3; 5]	1 [0.5; 3]	2 [0; 4]
Patients Hospitalized in 2005, % (n)	14.9 (39)	18.8 (66)	32.8 (221)	43.1 (31)	76.8 (99)	37.8 (14)	51.0 (26)
Hospitalization/patient in 2005, n	0 [0; 0]	0 [0; 0]	0 [0; 1]	0 [0; 1]	2 [1; 4]	0 [0; 1]	1 [0; 2]
Thoracic surgery in 2005, % (n)	1.1 (3)	0.6 (2)	1.3 (9)	19.4 (14)	1.6 (2)	2.7 (1)	0

All data are % (n) or median [Quartile 1-Quartile 3] unless otherwise stated.

Table 3. Main descriptors of the 7 subgroups identified by cluster analysis according to increasing death rates at 10 years

	Cluster 1 n=262 (17%)	**Cluster 2** n=351 (22%)	**Cluster 3** n=674 (43%)	**Cluster 4** n=72 (5%)	**Cluster 7** n=51 (3%)	**Cluster 5** n=125 (8%)	**Cluster 6** n=37 (2%)
10-year rate of any death	**4.6 %**	**5.4%**	**13.5%**	**20.8%**	**37.3%**	**45.6%**	**51.4%**
Proposed name	Classic CF/ Low P. aeruginosa	Non-classic CF/ Late diagnosis	Classic CF/ Moderate	Classic CF/ Pneumothorax/NTM	Classic CF/ ABPA	Classic CF/ Respiratory insufficiency	Classic CF/ *B. cepacia*
Respiratory disease							
Lung Function	Mild	Mild	Moderate/severe	Severe	Moderate/severe	Very severe	Moderate/severe
IV antibiotic course/year (median)	0	0	2	1	2	4	1
Respiratory Complications	Very low	Very low	Haemoptysis 15%	Haemoptysis 13.9% Pneumothorax 54.2%	Haemoptysis 15% ABPA/steroids	Haemoptysis 22.4% Pneumothorax 10.4 %	Haemoptysis 15.7%
Pancreatic insufficiency rates	Very high	Low	Very high	Very high	Very high	Very high	Very high
Diabetes rates	Very low	Very low	High	Low	High	High	High
Airway infection	Low P. aeruginosa	Low P. aeruginosa	High P. aeruginosa	P. aeruginosa High NTM	P. aeruginosa MRSA	P. aeruginosa Low B. cepacia	High B. cepacia
Outcomes at 5 years							
Any death	0.8 (2)	1.4 (5)	7.1 (48)	9.7 (7)	21.6 (11)	36.8 (46)	32.4 (12)
Death without transplantation	0.4 (1)	1.1 (4)	4.9 (33)	6.9 (5)	19.6 (10)	18.4 (23)	21.6 (8)
Lung transplantation	3.1 (8)	4.3 (15)	10.4 (70)	19.4 (14)	25.5 (13)	55.2 (69)	24.3 (9)
Death after lung transplantation	0.4 (1)	0.3 (1)	2.2 (15)	2.8 (2)	2.0 (1)	18.4 (23)	10.8 (4)
Outcomes at 10 years							
Any death	4.6 (12)	5.4 (19)	13.5 (91)	20.8 (15)	37.3 (19)	45.6 (57)	51.4 (19)
Death without transplantation	3.1 (8)	3.7 (13)	7.4 (50)	9.7 (7)	27.5 (14)	23.2 (29)	35.1 (13)
Lung transplantation	14.5 (38)	11.4 (40)	27.6 (186)	36.1 (26)	31.4 (16)	67.2 (84)	32.4 (12)
Death after lung transplantation	1.5 (4)	1.7 (6)	6.1 (41)	11.1 (8)	9.8 (5)	22.4 (28)	16.2 (6)

All data are % (n) unless otherwise stated.

Online supplementary data

Table S1. List and definition of variables included in the cluster analysis

Variable name	Variable definition in the French CF Registry
Gender	Male/female
Body mass index	Kg/m^2, at the time of last visit of the year
Age	As per December 31th 2005
CFTR mutation class I, II, III	0, 1 or 2 alleles
CFTR mutation class IV, V	0, 1, or 2 alleles
CFTR mutations unclassified	0, 1 or 2 alleles
Liver Cirrhosis	Yes/No
Pancreatic status	Pancreatic insufficiency/Pancreatic sufficiency
Haemoptysis	Any kind, yes/no
Pneumothorax	Any, yes/no
Diabetes mellitus treated	Insulin and/or oral treatment
Diabetes mellitus (untreated)	Diabetes, no treatment
FEV_1, % predicted*	Last spirometry of the year
Surgical procedure	Any surgical procedure in 2005 (excluding chest tube insertion for pneumothorax)
Intravenous antibiotics	Number of courses in 2005
Hospitalisation	Number of hospitalization in 2005
P. aeruginosa	Present/Absent**
B. cepacia	Present/Absent
Non tuberculous mycobacteria	Present/Absent
MSSA	Present/Absent
MRSA	Present/Absent
Long-term oxygen therapy	Yes/no
Non-invasive ventilation	Yes/no
Oral steroids	Prescribed for more than 3 months in 2005
Azithromycin	Prescribed for more than 3 months in 2005

* % predicted are based on equations by Knudson et al. [1]

**At least one positive culture in the past 12 months

Classification of CFTR mutations

Classification of CFTR mutations in the French CF registry was based on the functional classification by Welsh and Smith [2]and subsequent literature [3-5]. It included class I, II, III mutations and class IV or V mutations. When the functional consequences of a specific CFTR mutation was unknown, the mutation was considered unclassified. Uncomplete genotypes were genotypes with one or two unidentified CFTR mutations.

Table S2. Classification of the main CFTR mutations (i.e., with frequencies≥0.3% in the 2015 French CF Registry)

Class I	Class II	Class III	Class IV	Class V
W1282X	F508del	G551D	D1152H	3849+10kbC>T
W846X	I507del	G1244E	R117H	A445E
R553X	N1303K	S1255P	R117C	2789+5G>A
R1162X	L206W	G1349D	R334W	3120+1G>A
R1066C	G85E	S945L	R347H	
G542X	S549N	G551S	R347P	
E60X		R560T	R352Q	
E585X			S1251N	
711+1G>T				
621+1G>T				
394delTT				
3659delC				
2183AA>G				
1811+1.6kbAG				
1078delT				
1717-1G>A				

Table S3. Characteristics of 1376 Canadian adults with CF in 2005.

Variable	Categories	Frequency / Median	% / IQR
N	Overall	1,376	100.0%
Sex	Female	634	46.1%
	Male	742	53.9%
Age in 2005 (yrs)	Median (IQR)	26.8	21.7-34.4
Genotype	Homozygous dF508	669	48.6%
	Heterozygous dF508	554	40.3%
	Other	146	10.6%
	Missing	7	0.5%
BMI	Median (IQR)	21.6	19.8-24.0
FEV1 percent predicted	Median (IQR)	62.3	45.4-80.5
Negative Factors	BMI<17 kg/m^2	41	3.0%
	FEV1<25% predicted	57	4.1%
	CF related diabetes	300	21.8%
	Pneumothorax	19	1.4%
	B. cepacia complex	200	14.5%
	Long-term O_2 therapy	81	5.9%
Pancreatic Status	Pancreatic sufficient	171	12.4%
	Pancreatic insufficient	1205	87.6%

Table S4. Outcome by risk category in 1376 Canadian adults

	5-years		10-years	
Outcome	**Not low risk (N=1089)**	**Low Risk (N=287)**	**Not low risk (N=1089)**	**Low Risk (N=287)**
Any death	128 (11.8%)	9 (3.1%)	231 (21.2%)	22 (7.7%)
Death w/o transplant	92 (8.4%)	6 (2.1%)	160 (14.7%)	15 (5.2%)
Death post-transplant	36 (3.3%)	3 (1.0%)	71 (6.5%)	7 (2.4%)
Transplanted*	162 (14.9%)	13 (4.5%)	244 (22.4%)	25 (8.7%)
Lost to follow-up	9 (0.8%)	7 (2.4%)	76 (7.0%)	23 (8.0%)

Classification and Regression Tree (CART) analysis

CART analysis was conducted in the French CF Registry cohort (n=1572 patients) using the Tanagra 1.4 (Lyon, France) software. As recommended in the software instruction, the analysis was first conducted in a learning set representing two third of the cohorts (n=1037). This set was split into a growing set (n=694) and a pruning set (n=343). The confusion matrix is presented below showing an error rate of 0.14, indicating that 86% (n=888) of the patients were allocated to the appropriate group (low risk vs. not low risk) using the CART-determined algorithm.

Confusion matrix of the CART learning set in the French CF Registry cohort

Error rate			0,1437			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		CL2	CL1	Sum
CL2	0,9209	0,1416	CL2	594	51	645
CL1	0,7500	0,1478	CL1	98	294	392
			Sum	692	345	1037

Next the algorithm was tested in the remaining 535 patients (which data did not contribute to the construction of the algorithm). CART-determined algorithm allowed for classification of 87% of patients in the appropriate group (see below).

Confusion matrix of the CART validation set in the French CF registry cohort

Error rate			0,1271			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		CL2	CL1	Sum
CL2	0,9268	0,1339	CL2	291	23	314
CL1	0,7964	0,1156	CL1	45	176	221
			Sum	336	199	535

Table S5. Concordance of CART defined low-risk/not low risk classification with clusters

Cluster analysis	CART analysis	
Clusters	Low risk n=515	Not low risk n=1057
Cluster 1 (low risk)	35.5% (183) 70%	7.5% (79) 30%
Cluster 2 (low risk)	52.2% (269) 77%	7.8% (82) 23%
Cluster 3 (not low risk)	12.2% (63) 9%	57.8% (611) 91%
Cluster 4 (not low risk)	0.0% (0)	6.8% (72)
Cluster 5 (not low risk)	0.0% (0)	11.8% (125)
Cluster 6 (not low risk)	0.0% (0)	3.5% (37)
Cluster 7 (not low risk)	0.0% (0)	4.8% (51)

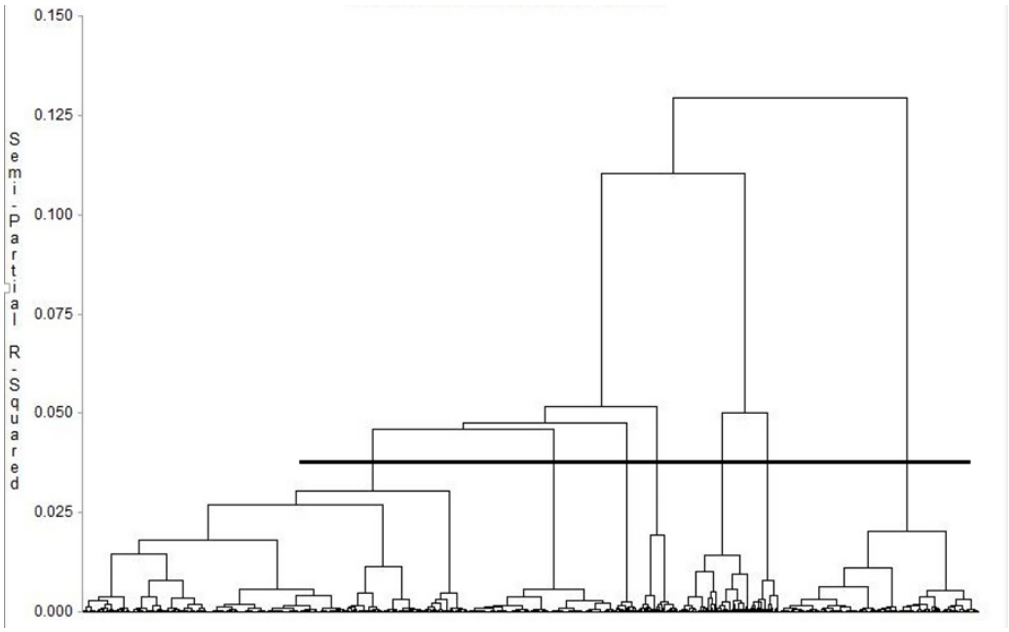
This table can be simplified by examining the concordance between low risk/not low risk according to cluster vs. CART analysis:

Table S6. Concordance of CART defined low-risk/not low risk vs. cluster-analysis defined low-risk/not low risk

		CART analysis		
		Low risk	**Not low risk**	**Total**
Cluster analysis	**Low risk (cluster 1-2)**	452 (28.8%)	161 (10.2%)	613 (39.0%)
	Not low risk (cluster 3-7)	63 (4.0%)	896 (57.0%)	959 (61.0%)
	Total	515 (32.8%)	1057 (67.2%)	1572 (100%)

Based on this table, the following metrics can be calculated for CART analysis performance for classification of low risk/not low risk as defined by cluster analysis:
Sensitivity=87.8%, Specificity=84.8%
Positive predictive value (PPV)=73.7%; Negative predictive value 93.4%

Figure S1. Dendrogram illustrating the results of the cluster analysis in 1572 adults with CF. Subjects were classified using agglomerative hierarchical cluster analysis based on the main components identified by factor analysis for mixed data (FAMD, see Methods section). Each vertical line represents an individual subject and the length of vertical lines represents the degree of similarity between subjects. The horizontal line identify the cut-off for choosing the optimal number of clusters (n=7) in the data.



References

1. Knudson RJ, Lebowitz MD, Holberg CJ, Burrows B. Changes in the normal maximal expiratory flow-volume curve with growth and aging. *Am Rev Respir Dis*. 1983;127:725-34. 10.1164/arrd.1983.127.6.725

2. Welsh MJ, Smith AE. Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell*. 1993;73:1251-4. 10.1016/0092-8674(93)90353-r

3. Castellani C, Cuppens H, Macek MJ, et al. Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. *J Cyst Fibros*. 2008;[Epub ahead of print].

4. Green DM, McDougal KE, Blackman SM, et al. Mutations that permit residual CFTR function delay acquisition of multiple respiratory pathogens in CF patients. *Respir Res*. 2010;11:140. 10.1186/1465-9921-11-140

5. McKone EF, Emerson SS, Edwards KL, Aitken ML. Effect of genotype on phenotype and mortality in cystic fibrosis: a retrospective cohort study. *Lancet*. 2003;361:1671-6.