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

Pierre-Régis Burgel [1], Lydie Lemonnier[2], Clémence Dehillotte[2], Jenna Sykes[3,4], Sanja Stanojevic[4,5], Anne L. Stephenson[3,4] and Jean-Louis Paillasseur[6]

Affiliations: [1]University Paris Descartes, Sorbonne Paris Cité, Paris, France. [2]Vaincre la Mucoviscidose, Paris, France. [3]Adult CF Program, Dept of Medicine, University of Toronto, St Michael's Hospital, Li Ka Shing Knowledge Institute, Keenan Research Centre, Toronto, ON, Canada. [4]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada. [5]The Hospital for Sick Children, Division of Respiratory Medicine, Toronto, ON, Canada. [6]Effi-Stat, Paris, France.

Correspondence: Pierre-Régis Burgel, Service de Pneumologie, Hôpital Cochin, 27 Rue du Faubourg St Jacques, 75014 Paris, France. E-mail: pierre-regis.burgel@aphp.fr

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

Factor analysis for mixed data followed by Ward's cluster analysis was 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 CART (classification and regression tree) analysis 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 nonclassic CF, the other subgroup (17% of patients) was composed of patients with classic CF but low rates of Pseudomonas aeruginosa infection and diabetes. An algorithm based on CART analysis of data in 2005 allowed us to identify most French adults with low rates of death. When tested using data from Canadian CF adults in 2005, the algorithm identified 287 out of 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 70 000 patients worldwide [1], and is characterised as a multisystemic disease involving primarily the lung, pancreas and 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 socioeconomic 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 s (FEV_1), low body mass index (BMI), pancreatic insufficiency, colonisation with *Pseudomonas aeruginosa* and/or *Burkholderia cepacia*, massive haemoptysis, and pneumothorax) associated with poor prognosis that were sometimes combined into prognostic scores [13, 15–18]. 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], chronic obstructive pulmonary disease [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 not entitled to loan insurance because of their disease. Our strategy was to use cluster analysis to identify subgroups of French adults with CF experiencing low death rates over 10 years and to use CART (classification and regression tree) 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 CF 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 BMI) 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 aforementioned methodology developed for the French CF Registry.

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 31, 2015 (10-year death). We next used 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 the Canadian CF Registry. Survival and transplantation-free survival were analysed using Kaplan–Meier curves and the log-rank test. Risk of any death or risk of the combined event death without transplantation/occurrence of lung transplantation was analysed by Cox models. Data are presented as median (interquartile range (IQR)) or number (percentage). Analyses were performed using SAS version 9.2 (SAS Institute, Cary, NC, USA).

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. A complete description and definition of the 25 variables selected in this analysis is provided in supplementary table S1. Briefly, these variables included sex, age as of December 31, 2005, BMI (kg·m^{-2}), FEV1 % pred, CF transmembrane conductance regulator (*CFTR*) mutations, pancreatic status, diabetes mellitus, liver cirrhosis, haemoptysis, pneumothorax, airway infection, and variables related to healthcare utilisation and/or therapeutic management (long-term oxygen therapy, noninvasive ventilation (NIV), treatment with oral steroids for >3 months, insulin, number of intravenous antibiotic courses in 2005 and number of hospitalisations ⩾48 h in 2005). Identification of homogeneous subgroups of adults with CF was achieved using FAMD (factor analysis for mixed data) [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 FAMD. Cluster analysis was performed using Ward's method, for which grouping was based on the 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 analysed data using rates of any death at 5 and 10 years. As 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 *versus* 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 (a detailed explanation can be found in the supplementary material). 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 31, 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 a waiting list for transplantation in 2005, death in 2005, previous history of cancer, and missing data for lung function and/or BMI. A flowchart 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 31, 2005 to December 31, 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 supplementary table S1), using FAMD followed by Ward's hierarchical classification. The dendrogram resulting from this analysis is shown in supplementary figure S1. Based on pseudo-F and pseudo-t^2 statistics, and visual assessment of the dendrogram, the data could be optimally classified into seven clusters. Table 2 presents the clinical characteristics of the 1572 adult CF patients according to these seven clusters.

Cluster 1 was composed of young adults (median age 21.8 years) with classic CF characterised by diagnosis early in life and pancreatic insufficiency, but with mild to moderate impairment in lung function, low rates of chronic *P. aeruginosa* airway infection and no diabetes. Cluster 2 was composed of older adults (median age 30.2 years) with nonclassic CF characterised by late diagnosis, high rates of class

```
┌─────────────────────────────┐
│      French CF Registry 2005 │
│         n=4758 patients      │
└─────────────────────────────┘
              │
              │        ┌─────────────────────────┐
              ├────────│   Age <18 years n=2816  │
              │        └─────────────────────────┘
┌─────────────────────────────┐
│  Adult patients age ≥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)   │
                       └───────────────────────────────────┘
```
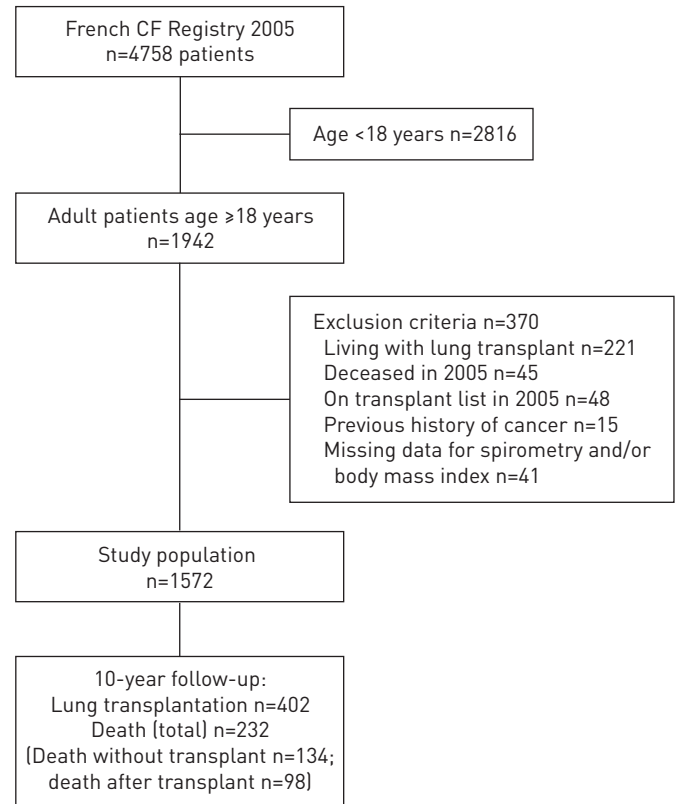
FIGURE 1 CONSORT diagram of the cystic fibrosis (CF) patients included in the study.

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 *i.v.* 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 nontuberculous mycobacteria (NTM). Cluster 5 was composed of patients with classic CF and very high rates of respiratory insufficiency treated with long-term oxygen therapy or NIV. 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 seven clusters identified in the cluster analysis, we next examined rates of death and/or lung transplantation over 10 years in each cluster. Kaplan–Meier curves of survival (outcome: any death) or 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 (nonclassic CF/late diagnosis) 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 cluster 4 (classic CF/pneumothorax/NTM) had <10% death 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% 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/allergic bronchopulmonary aspergillosis) had 37.3% death at 10 years with 31.4% lung transplantation. Table 3 summarises the main descriptors of patients in each cluster and prognostic outcomes in each cluster. Comparisons 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
As 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 us to identify these patients (n=613 patients: 262 patients in cluster 1 and 351 patients in cluster 2) using data obtain at study entry in 2005. The algorithm is presented in

TABLE 1 Main characteristics of adults with cystic fibrosis (CF) in the French CF Registry as per December 31, 2005

Subjects	1572
Male	53.7 (844)
Age years	24.9 (21.0–30.9)
Age at diagnosis years [#]	1.6 (0.2–10.2)
BMI kg·m^{-2}	19.9 (18.3–21.7)
***CFTR* genotype** [¶]	
Class I–III/class I–III	61.1 (961)
At least one class IV or V mutation	11.9 (187)
Class I–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$_1$ % pred	57 (38–79)
Pancreatic insufficiency	82.1 (1291)
Airway infection	
Pseudomonas aeruginosa	61.7 (970)
Burkholderia cepacia	2.8 (44)
MRSA	15.7 (247)
MSSA	45.5 (716)
Nontuberculous mycobacteria	2.3 (36)
Liver cirrhosis	11.0 (173)
Diabetes mellitus	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)
Noninvasive ventilation	4.5 (71)
Courses of *i.v.* antibiotics in 2005 n	1 (0–3)
Hospitalisations in 2005 n	0 (0–1)

Data are presented as n, % (n) or median (interquartile range). BMI: body mass index; *CFTR*: CF transmembrane conductance regulator; FEV$_1$: forced expiratory volume in 1 s; MSSA: methicillin-susceptible *Staphylococcus aureus*; MRSA: methicillin-resistant *S. aureus*. [#]: missing n=70; [¶]: see supplementary table S2 for details on classification of *CFTR* mutations.

figure 4: patients with a least one factor negatively affecting the prognosis (see inset list in 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 out of 613 (84% of patients in clusters 1 and 2) patients and the rate of death in these 515 patients was 3.9% at 10 years. Kaplan–Meier analysis comparing 10-year death at lower risk *versus* 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 (characteristics of patients are given in supplementary tables S3 and S4). Three variables necessary for excluding patients at higher risk according to the algorithm were unavailable in the Canadian CF Registry in 2005: infection with NTM, use of systemic steroids for >3 months and use of NIV. The algorithm was thus used on Canadian data without considering these variables: it identified 287 out of 1376 (21%) patients at low risk (10-year death: 7.7%). Kaplan–Meier analysis comparing 10-year death at lower risk *versus* 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

TABLE 2 Characteristics of 1572 adults with cystic fibrosis (CF) according to the seven subgroups identified by cluster analysis

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Subjects	262	351	674	72	125	37	51
Male	65 (170)	50 (176)	49 (328)	57 (41)	62 (78)	62 (23)	55 (28)
Age years	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 years[#]	1.1 (0.1–7.5)	17.4 (4.3–31.4)	0.7 (0.1–3.8)	1.9 (0.1–11.7)	0.7 (0.2–2.7)	0.9 (0.2–5.6)	0.9 (0.2–5.6)
***CFTR* genotype**							
Class I–III/class I–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 or 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)
FEV1 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)
FEV1 % pred	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	93.1 (244)	37.6 (132)	97.0 (654)	86.1 (62)	96.0 (120)	94.6 (35)	86.3 (44)
BMI 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)
Nontuberculous mycobacteria	0	0	0	47.2 (34)	0.8 (1)	0	2.0 (1)
Comorbidities							
Liver cirrhosis	11.8 (31)	2.6 (9)	13.6 (92)	11.1 (8)	17.6 (22)	16.2 (6)	9.8 (5)
Diabetes mellitus	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)
Noninvasive ventilation	0	0.3 (1)	0	0	52.8 (66)	0	7.8 (4)
Patients with *i.v.* antibiotics in 2005	26.7 (70)	31.1 (109)	69.6 (469)	63.9 (46)	93.6 (117)	73.0 (27)	72.5 (37)
Courses of *i.v.* antibiotics per patient in 2005	0 (0–1)	0 (0–1)	2 (0–3)	1 (0–3)	4 (3–5)	1 (0.5–3)	2 (0–4)
Patients hospitalised in 2005	14.9 (39)	18.8 (66)	32.8 (221)	43.1 (31)	76.8 (99)	37.8 (14)	51.0 (26)
Hospitalisations per patient in 2005	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	1.1 (3)	0.6 (2)	1.3 (9)	19.4 (14)	1.6 (2)	2.7 (1)	0

Data are presented as n, % (n) or median (interquartile range). *CFTR*: CF transmembrane conductance regulator; FEV1: forced expiratory volume in 1 s; BMI: body mass index; MSSA: methicillin-susceptible *Staphylococcus aureus*; MRSA: methicillin-resistant *S. aureus*. [#]: missing n=70 (cluster 1 missing n=12; cluster 2 missing n=18; cluster 3 missing n=26; cluster 4 missing n=5; cluster 5 missing n=7; cluster 6 missing n=1; cluster 7 missing n=1); [¶]: see supplementary table S2 for details on classification of *CFTR* mutations.
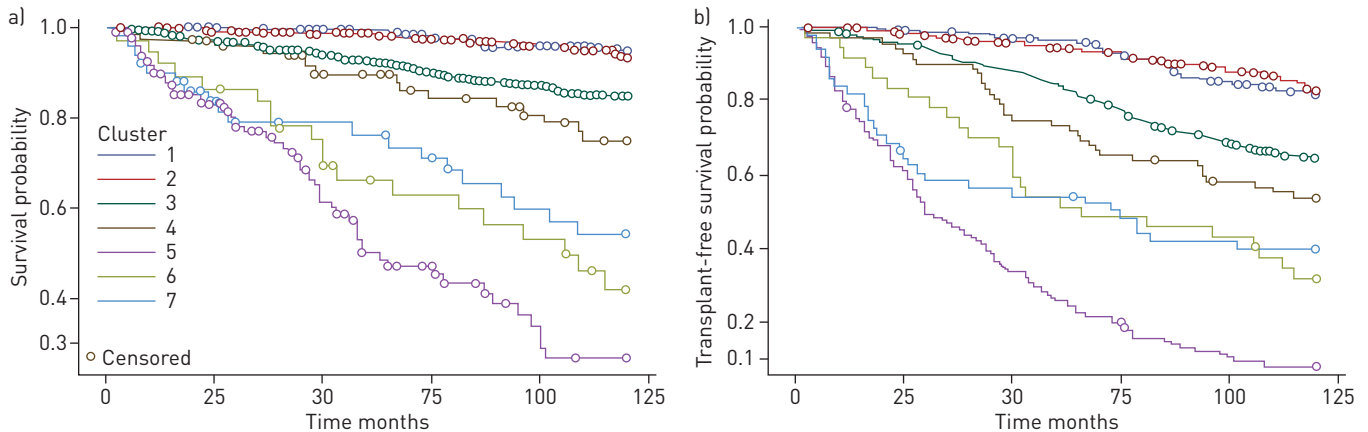
FIGURE 2 Kaplan–Meier curves for assessing a) survival (outcome: any death) or b) transplant-free survival (outcome: death without lung transplantation/occurrence of lung transplantation). The log-rank test indicated that all subgroups, except cluster 2, had increased risk of outcomes (p<0.001) compared with cluster 1.

Registry in 2005 and showed consistent findings. These results indicate that many CF adults have a favourable long-term prognosis, a finding that has important socioeconomic 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% death at 5 years and 5% death at 10 years. These data indicate that a large proportion of adult CF patients have improved prognosis in the current era, 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 nonclassic CF (characterised by late diagnosis and milder clinical features) and lower risk of death [15, 29]. 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 with less well-documented CF diagnosis usually had milder disease [30]. 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 characterised 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 and strategies for eradication of *P. aeruginosa* infection).

The study also highlighted the fact that despite comprehensive CF care and lung transplantation, ∼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 2 times lower in cluster 6 *versus* 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 >90% of the French CF population, with a 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 recognise limitations. The present study was designed to identify subgroups of individuals sharing comparable characteristics and long-term prognosis, but was not intended to determine 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

TABLE 3 Main descriptors of the seven subgroups identified by cluster analysis according to increasing death rate at 10 years

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 7	Cluster 5	Cluster 6
Subjects	17 (262)	22 (351)	43 (674)	5 (72)	3 (51)	8 (125)	2 (37)
10-year rate of any death	4.60	5.40	13.50	20.80	37.30	45.60	51.40
Proposed name	Classic CF/ low *P. aeruginosa*	Nonclassic 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
Courses of *i.v.* antibiotics per 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*
Outcome 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)
Outcome 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 presented as % (n) or %, unless otherwise stated. CF: cystic fibrosis; *P. aeruginosa*: *Pseudomonas aeruginosa*; NTM: nontuberculous mycobacteria; ABPA: allergic bronchopulmonary aspergillosis; *B. cepacia*: *Burkholderia cepacia*; MRSA: methicillin-resistant *Staphylococcus aureus*.
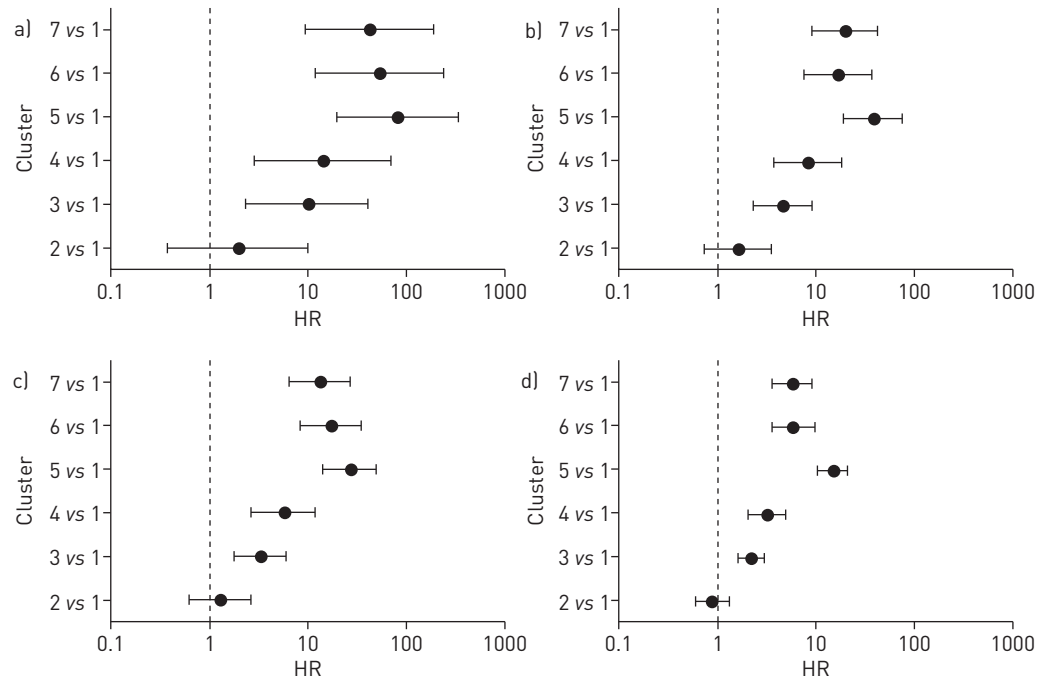
FIGURE 3 Relative risks of a, c) any death and b, d) death without transplantation/occurrence of lung transplantation at a, b) 5 and c, d) 10 years among clusters of adult cystic fibrosis patients. Horizontal bars show hazard ratios and 95% confidence intervals of risks between classes (*e.g.* subjects in cluster 3 have a 9.8-fold (95% CI 2.4–40.4-fold) increased risk of any death at 5 years when compared with subjects in cluster 1).

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 (supplementary tables S5 and S6). Some of the data (*i.e.* infection with NTM, use of systemic steroids for >3 months and use of NIV) necessary to run the algorithm were unavailable in the Canadian CF Registry in 2005, which may have resulted in the selection
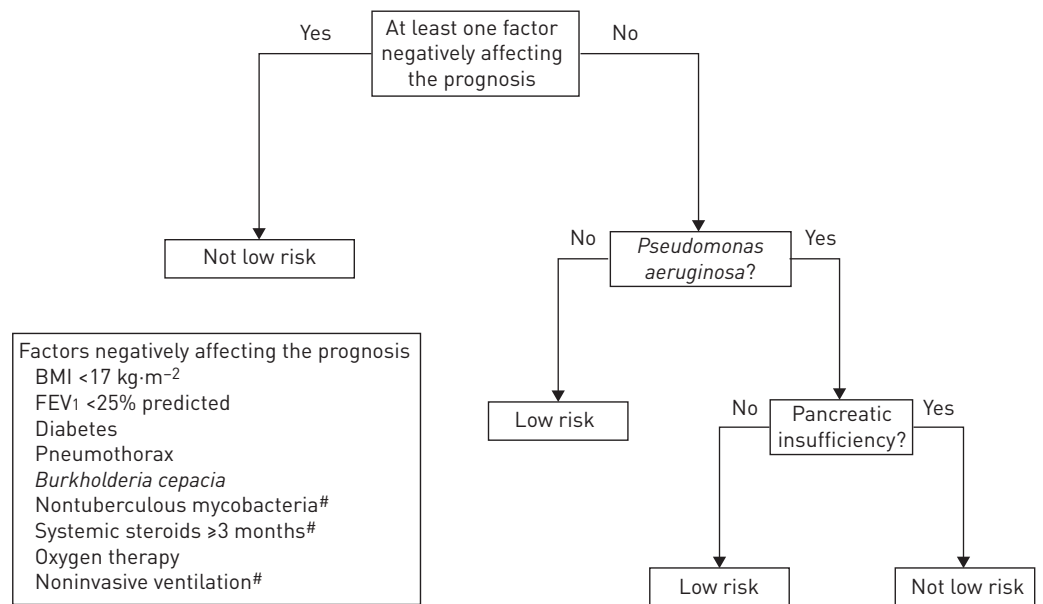
FIGURE 4 Algorithm for identification of cystic fibrosis (CF) adults at low risk *versus* not low risk of death at 10 years in French CF adults. BMI: body mass index; FEV1: forced expiratory volume in 1 s. The algorithm was obtained based on CART (classification and regression tree) analysis of variables initially used for the cluster analysis (see Methods), with the objective of classifying the larger number of patients in the low-risk subgroups (clusters 1 and 2). #: these variables were unavailable for external validation of the algorithm in the Canadian CF Registry.
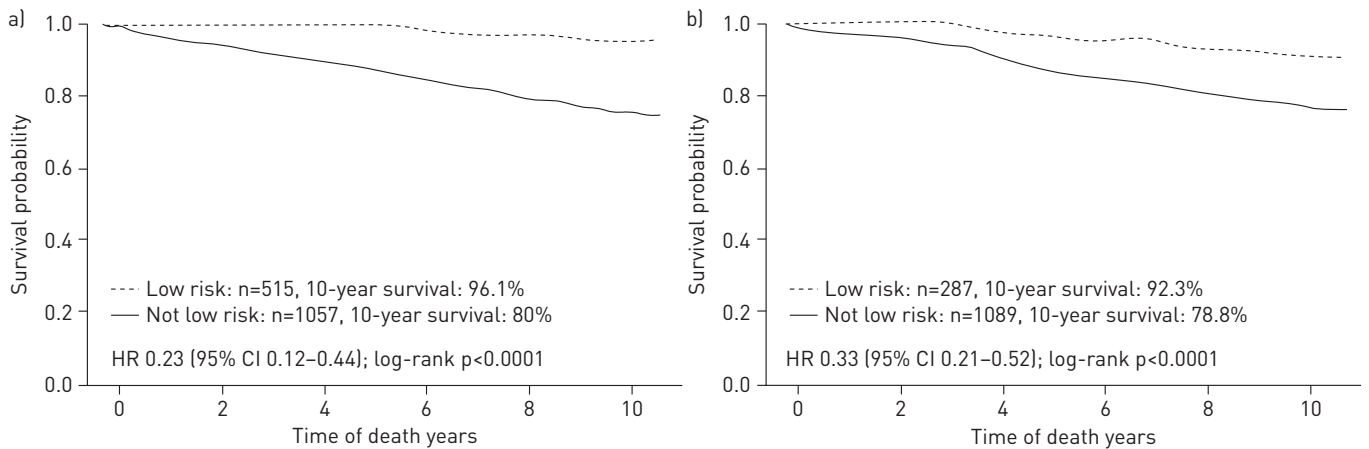
FIGURE 5 Kaplan–Meier analysis for assessing death at 10 years in cystic fibrosis (CF) subjects at low risk *versus* not low risk identified by the algorithm: a) French CF Registry and b) Canadian CF Registry.

by the algorithm of a limited number of patients at higher risk of death. However, the algorithm consistently identified patients with 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 usually have 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 specialised CF care and to lung transplantation.

The present study highlighted the recent evolution of prognosis in CF adults. Although studies performed 10–15 years ago suggested that intrinsic patient characteristics (*e.g. CFTR* genotype associated with residual CFTR function) were associated with improved prognosis [15, 29], 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* [31]) 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 to provide 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 [32].

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 specialised 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 defects in *CFTR* were 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.

References

1   Stephenson AL, Stanojevic S, Sykes J, *et al.* The changing epidemiology and demography of cystic fibrosis. *Presse Med* 2017; 46: e87–e95.

2   Elborn JS. Cystic fibrosis. *Lancet* 2016; 388: 2519–2531.

3   Elborn JS, Bell SC, Madge SL, *et al.* 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: 420–428.

4   McCormick J, Mehta G, Olesen HV, *et al.* Comparative demographics of the European cystic fibrosis population: a cross-sectional database analysis. *Lancet* 2010; 375: 1007–1013.

5   Burgel PR, Bellis G, Olesen HV, *et al.* Future trends in cystic fibrosis demography in 34 European countries. *Eur Respir J* 2015; 46: 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: 1700763.

7   Urquhart DS, Thia LP, Francis J, *et al.* Deaths in childhood from cystic fibrosis: 10-year analysis from two London specialist centres. *Arch Dis Child* 2013; 98: 123–127.

8   Martin C, Hamard C, Kanaan R, *et al.* Causes of death in French cystic fibrosis patients: the need for improvement in transplantation referral strategies! *J Cyst Fibros* 2016; 15: 204–212.

9   Hirche TO, Knoop C, Hebestreit H, *et al.* Practical guidelines: lung transplantation in patients with cystic fibrosis. *Pulm Med* 2014; 2014: 621342.

10 Thabut G, Christie JD, Mal H, *et al.* Survival benefit of lung transplant for cystic fibrosis since lung allocation score implementation. *Am J Respir Crit Care Med* 2013; 187: 1335–1340.

11 Kerem E, Reisman J, Corey M, *et al.* Prediction of mortality in patients with cystic fibrosis. *N Engl J Med* 1992; 326: 1187–1191.

12 Liou TG, Adler FR, Fitzsimmons SC, *et al.* Predictive 5-year survivorship model of cystic fibrosis. *Am J Epidemiol* 2001; 153: 345–352.

13 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.

14 McCarthy C, Dimitrov BD, Meurling IJ, *et al.* 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: 1441–1447.

16 Flume PA, Strange C, Ye X, *et al.* Pneumothorax in cystic fibrosis. *Chest* 2005; 128: 720–728.

17 Flume PA, Yankaskas JR, Ebeling M, *et al.* Massive hemoptysis in cystic fibrosis. *Chest* 2005; 128: 729–738.

18 Lewis C, Blackman SM, Nelson A, *et al.* Diabetes-related mortality in adults with cystic fibrosis. Role of genotype and sex. *Am J Respir Crit Care Med* 2015; 191: 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, *et al.* Cluster analysis and clinical asthma phenotypes. *Am J Respir Crit Care Med* 2008; 178: 218–224.

21 Burgel PR, Paillasseur JL, Caillaud D, *et al.* Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur Respir J* 2010; 36: 531–539.

22 Burgel PR, Paillasseur JL, Janssens W, *et al.* A simple algorithm for the identification of clinical COPD phenotypes. *Eur Respir J* 2017; 50: 1701034.

23 Aliberti S, Lonni S, Dore S, *et al.* Clinical phenotypes in adult patients with bronchiectasis. *Eur Respir J* 2016; 47: 1113–1122.

24 Conrad DJ, Bailey BA. Multidimensional clinical phenotyping of an adult cystic fibrosis patient population. *PLoS One* 2015; 10: e0122705.

25 Vaincre la Mucoviscidose et Ined. Registre Français de la Mucoviscidose – Valorisation des données. [French Cystic Fibrosis Registry – Data Review.] 2015. www.vaincrelamuco.org/face-la-mucoviscidose/registre-et-muco-en-chiffres/valorisation-des-donnees Date last accessed: January 28, 2017.

26 Wikipedia. Factor analysis of mixed data. 2016. https://en.wikipedia.org/wiki/Factor_analysis_of_mixed_data Date last accessed: September 25, 2018.

27 Pagès J. Analyse factorielle de donnees mixtes. [Multiple factor analysis for mixed data.] *Rev Stat Appl* 2004; 52: 93–111.

28 Burgel PR, Paillasseur JL, Peene B, *et al.* Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality. *PLoS One* 2012; 7: e51048.

29 McKone EF, Emerson SS, Edwards KL, *et al.* Effect of genotype on phenotype and mortality in cystic fibrosis: a retrospective cohort study. *Lancet* 2003; 361: 1671–1676.

30 Thomas M, Lemonnier L, Gulmans V, *et al.* Is there evidence for correct diagnosis in cystic fibrosis registries? *J Cyst Fibros* 2014; 13: 275–280.

31 Langton Hewer SC, Smyth AR. Antibiotic strategies for eradicating *Pseudomonas aeruginosa* in people with cystic fibrosis. *Cochrane Database Syst Rev* 2017; 4: CD004197.

32 Quon BS, Rowe SM. New and emerging targeted therapies for cystic fibrosis. *BMJ* 2016; 352: i859.