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

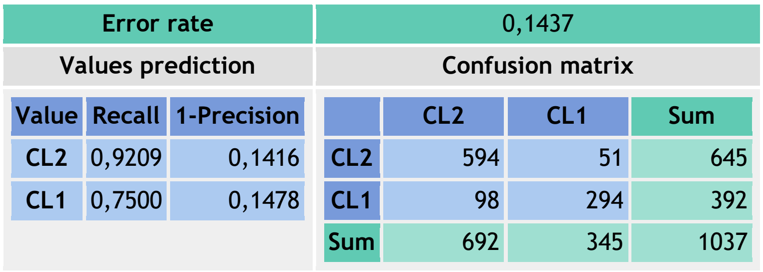
Table S4. Outcome by risk category in 1376 Canadian adults

	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



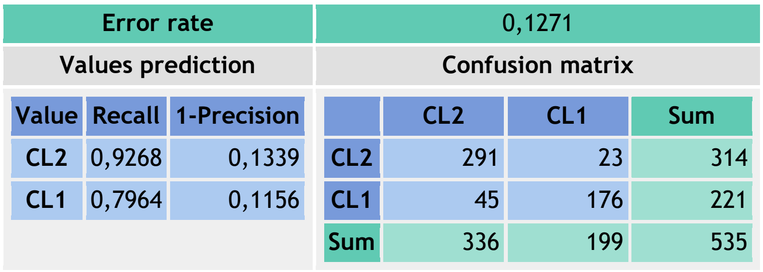
Error rate	0,1437
Values prediction	Confusion matrix

Value	Recall	1-Precision
CL2	0,9209	0,1416
CL1	0,7500	0,1478

	CL2	CL1	Sum
CL2	594	51	645
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	0,9268	0,1339
CL1	0,7964	0,1156

	CL2	CL1	Sum
CL2	291	23	314
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)

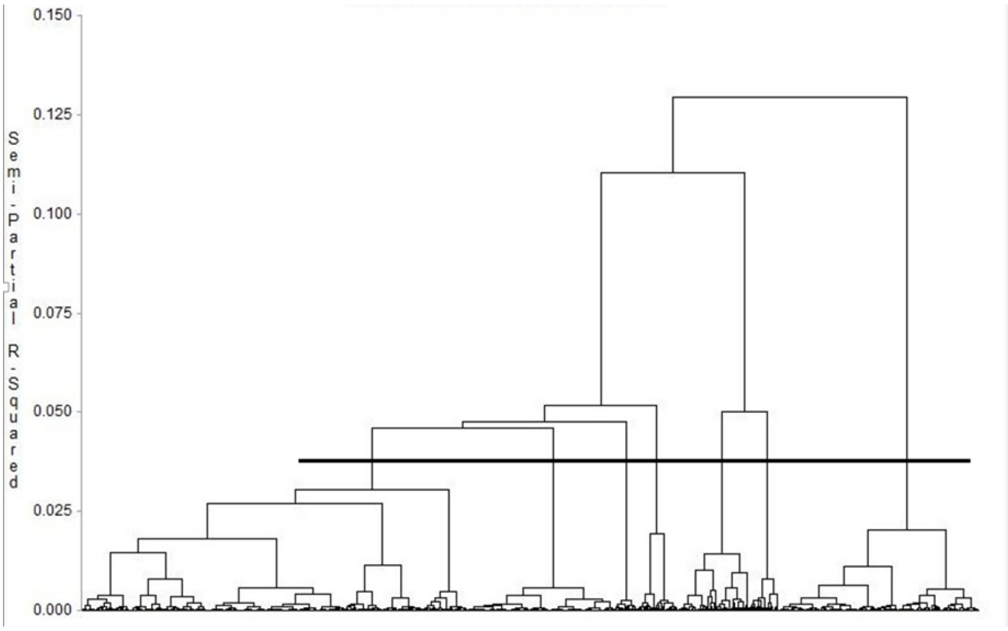
This table can be simplified by examining the concordance between low risk/not low risk according to cluster vs. CART analysis:

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.