American Journal of Medical Genetics 13:179-195 (1982)

Phenotypic Heterogeneity in Cystic Fibrosis

Charles F. Sing, David R. Risser, William F. Howatt, and Robert P. Erickson

Departments of Human Genetics (C.F.S., D.R.R., R.P.E.) and Pediatrics (R.P.E.) University of Michigan, Ann Arbor, and Department of Pediatrics, Mott Children's Hospital (W.F.H.), University of Michigan, Ann Arbor

We have confirmed heterogenity in CF using a different combination of primary clinical variables than those used in previous studies. Subgroupings of individuals with similar levels of sweat chloride were independent of the clustering based on level of pancreatic enzyme supplementation and degree of pulmonary involvement. Data from families with multiple CF children are consistent with the hypothesis that the genetic etiology of CF involves two or more genes that modify the expression of the primary gene defect.

Key words: cystic fibrosis, phenotype heterogeneity, familial aggregation

INTRODUCTION

Cystic fibrosis of the pancreas (CF) is a common genetic disorder that often leads to serious complications and death at an early age. The presenting symptoms of the disease are usually related to malabsorption and/or pulmonary involvement. The primary clinical sign that confirms a diagnosis of cystic fibrosis is the elevation of sodium and chloride ions in sweat [Wood et al, 1976]. It is well known that there is considerable heterogeneity in the manifestation of CF. Before the sweat test was developed, cystic fibrosis patients were frequently undiagnosed until after the disease process had already done extensive damage that could have been lessened if treatment had followed an earlier diagnosis. For those diagnosed as having CF, there is also considerable variation in the severity of the disease. The mildest extreme is represented by those who are not diagnosed until middle age [Scully et al, 1977].

Received for publication October 4, 1981; revision received January 8, 1982.

Address reprint requests to Dr. Charles F. Sing, Department of Human Genetics, University of Michigan, 1137 E. Catherine Street, Ann Arbor, MI 48109

It is widely accepted that cystic fibrosis is a monogenetically determined disease, but the specific location of the gene or gene products that cause the disease have not been identified. Individuals with cystic fibrosis are usually born to normal parents, although they may have relatives who are affected. Equal incidence in the two sexes and a segregation of ¼ affected children of unaffected parents in a number of early studies [Anderson and Hodges, 1946; Danks et al, 1965] argue for an autosomal recessive gene as the cause of the disease. Despite these convincing data, the high frequency of CF in the Caucasian population and the great variability in disease expression among and within families has suggested to some workers that one or more dominant genes may be involved [Baumann, 1958; Schaap and Cohen, 1976]. It has been suggested that heterogeneity in the severity of the disease may be explained by variability in the expression and penetrance of the gene for cystic fibrosis [Frydman, 1979; Danes and Bearn, 1969]. Schaap and Cohen [1976] give a comprehensive review of the genetics of CF and suggest further that the high incidence of this disease can be explained by an epistatic interaction between dominant alleles at each of two loci. Definitive studies to decide among alternative genetic models to explain the inconsistencies of the observations with a single autosomal recessive causation have yet to be carried out.

The purpose of the studies reported in this paper was to identify those measures of CF patients that discriminate subclasses of the disease. These studies were conducted on data collected from patients seen at the Mott Children's Hospital at the University of Michigan. We identified four distinct subgroups of patients who vary in the severity of

TABLE I. Summary of Variables Studied

Variable[a]	Scale	N	Mean	Standard deviation	Min-max
Sweat chloride	mEq/liter	83	99.0	16.7	65-143
Age of diagnosis	yr	101	2.4	3.6	0-16
Birth weight	gm	74	3,261.4	740.4	1,570-7,500
Exam age	yr	101	9.7	6.5	0-33
Pulmonary grade					
Cough	1-4	104	2.1	1.2	1-4
Rales	1-4	100	1.7	1.3	1-4
Clubbing	1-4	103	2.3	1.1	1-4
Chest shape	1-3	86	1.7	0.8	1-3
Chest film	1-4	102	2.3	1.2	1-4
Gastrointestinal					
Standard enzyme dose	Pills[b]	101	17.9	12.0	0-45
Liver size	cm	81	1.2	2.3	0-10
Laboratory					
SGPT	IU/liter	52	35.2	48.0	8-334
SGOT	IU/liter	56	43.5	39.0	6-291
Alkaline phosphotase	IU/liter	52	153.4	157.9	8-900
Albumin	g%	21	3.9	0.8	1.2-4.0
Total protein	g%	25	7.4	1.4	4.8-10.8

[a] All variables except the sweat value were recorded at the University of Michigan CF clinic at the last exam. All sweat values are those at age of diagnosis either at the University of Michigan (54 cases) or by a reputable CF laboratory (29 cases).

[b] Equivalent number of Viokase pills per day (explanation in test).

their disease. The distribution of these groups among affected sibs suggest needed studies to determine the causes and control of the expression of the traits we have examined.

The Sample Studied

Between 1970 and 1979 we examined a sample of 104 patients at the Mott Children's Hospital of the University of Michigan. Most of the patients are from Southeast Michigan, which includes the Detroit area. However, the sample also includes individuals from within a 500 mile radius, since many patients were referred to the cystic fibrosis center at the Mott Hospital for treatment. Fifty-four patients were diagnosed at the University of Michigan CF Clinic; the remaining 50 were diagnosed before referral to the Mott Children's Hospital.

There are 53 female patients and 51 male patients in the sample distributed among 73 sibships (see Table VII). The diagnosis of CF was confirmed in all cases by sweat sodium level when diagnosed at the University of Michigan or by sweat sodium or chloride when diagnosed elsewhere. The average age of diagnosis was 2.4 yr. The second CF child in the multiplex families studied ($N = 16$) also had an average age of diagnosis of 2.4 yr. The range in age of diagnosis was from birth to 16 yr, with 75% of the patients diagnosed by 7 yr. Only six patients were diagnosed after 10 yr.

The patients include 72 Caucasians, three Blacks, one Mexican, and two of mixed parentage. Information is limited on the nationality of the sample, but we know that there are a number of patients with German ancestry. Cystic fibrosis is considered to be a disease that is predominantly confined to the Caucasian race [Cystic Fibrosis Foundation, 1977] and our data support this.

MATERIALS AND METHODS

Data Collection Procedures

All clinical variables were collected at the University of Michigan clinic. The age of examination ranged from 1 to 33 yr ($\bar{x} = 9.6$ yr). Data collection involved a careful review of the physicians' clinic reports, with the information on each patient being transcribed to a standardized form. The review was undertaken either by physicians or assistants under their direct supervision. Clinical data included ratings of the severity of pulmonary involvement, including cough, rales, clubbing, chest shape, and chest film findings. The measures of gastrointestinal involvement included digestive enzyme dose and general physical condition. In addition to sweat test results, SGPT, SGOT, alkaline phosphatase, albumin, and total protein were considered. Although all individuals studied were diagnosed by the sweat test, values were available from the referring physician on only 29 of the 50 patients who were not diagnosed at the University of Michigan clinic. The data available also included information about the disease status of all sibs.

The Variables Studied

The array of variables studied was selected because they are generally considered to reflect the clinical manifestations of CF. All but sweat sodium were evaluated at the time of clinical examination at the University of Michigan CF clinic. The scale of measurement and the mean and standard deviation for each is given in Table I. The indicators of pulmonary damage were discrete categories ranging from least (grade 1) to greatest deviation from normality (3 or 4). Enzyme supplementation was Viokase powder (N = 14), Viokase pills (N = 61), Cotazyme powder (N = 2), Cotazyme capsules (N = 14), Pancrease capsules (N = 2), or not recorded (N = 2). Nine patients were not receiving enzyme supplementation when examined. Standard enzyme dose was expressed in terms of equivalent number of Viokase pills per day for this study. These equivalents were arrived at by multiplying the number of Pancrease pills by 4, Cotazyme powder packets by 2, and teaspoons of Viokase powder by 7, one Cotazyme capsule being the equivalent of one Viokase pill. Liver size was evaluated (by palpation) in cm below the costal margin. Laboratory data were obtained from a serum sample collected at the most recent date of examination.

The deviation of an individual's age at exam from the average of all exam ages for those who share his grade of pulmonary involvement (Table I) was taken as an age-corrected measure of pulmonary severity. This was done separately for each of the five measures of pulmonary function. Individuals with positive deviations, regardless of their severity rating for the pulmonary variable, are assumed to be experiencing a less severe course of the disease than those who deviate negatively. A statistically significant positive correlation was observed between age of diagnosis and each of these variables (eg, 0.51 with exam age deviation for a given grade of x-ray finding). These age corrected measures of severity were then used in all analyses, rather than contrasting the severity rating directly.

Statistical Procedures

The objective of the statistical analyses was to identify those variables that characterize homogeneous subgroups of CF patients. All analyses reported here were carried out using the Michigan Interactive Data Analysis System (MIDAS) on the University's Amdahl 470-V8. Contrasts between patients diagnosed at the University of Michigan and those diagnosed at other institutions were made for each of the analyses reported below, to identify any possible ascertainment biases that might exist in the study sample.

A description of the data was followed by an investigation of the effects of differences in age and sex on each of the continuously varying measures of sweat, birth weight, pulmonary severity, gastrointestinal and laboratory variables. Analyses of the matrix of correlations among these age- and sex-adjusted variables were performed to reduce the dimensionality of the data set. In this way we were able to identify the most parsimonious subset of variables on which to base identification of phenotypic heterogeneity. This was followed by a cluster analysis to identify subclasses of patients. Duran and Odell [1974], and Sneath and Sokal [1973], review the principles of cluster analysis. It is a nonstatistical technique to explore systematically the information contained in a multivariate set of data on the similarities between individuals. As such it is a technique for generating hypotheses rather than for testing hypotheses. A number of combinations of distance measures and clustering algorithms were considered. The greatest discrimination between subgroups of patients and the highest cophenetic correlation was given by the minimum variance algorithm [Ward, 1963] with nearly identical assignments of individuals when either the Euclidean or the correlation measure of distance was employed. The correlation measure used in this study places in the same group individuals who have the highest correlation for the variables studied while the correlations between pairs in different groups is minimized. Principle component analyses were then employed to generate the linear functions that gave the greatest discrimination among groups.

RESULTS

Analyses of the effects of age, sex, and place of diagnosis were carried out on the sweat, birth weight, pulmonary, gastrointestinal and laboratory variables given in Table I. In addition, the exam age at which a patient had achieved a given grade for each of the five measures of pulmonary severity was considered. There were no statistically significant differences in mean levels for any of the 15 continuous variables, or in frequencies of the discrete pulmonary grades, between those diagnosed at the University of Michigan and those diagnosed elsewhere. As might be expected, there was a statistically significant (P = 0.05) relationship between age and a number of the measures considered. These findings are summarized in Table II. Sweat levels increased 1.26 mEq/liter for each year increase in the patient's age at diagnosis. Pulmonary severity grades were predictive of the age when the patient was evaluated. On the average, each grade increase in severity was associated with 1.4 to 6.7 yr increase in the patient's age at examination. Rales and chest film findings progressed most rapidly. The number of years to change to one more severe grade was greatest for chest shape, but it should be pointed out that chest shape is only a three-grade scale, while the other pulmonary variables are four-grade scales. Again, the contrast of regressions between patients diagnosed at the University of Michigan and those diagnosed elsewhere was not stastistically significant at the 0.10 level of probability for any of the variables considered.

The male patients were 72 gm heavier at birth than females. The male-female comparisons of disease symptoms were not statistically significant for any one of the other 14 continuous variables considered. However, inspection of the differences revealed a trend. For every measure of pulmonary severity, females were younger than males for a given grade of involvement. This concordance might be expected because the five ratings are highly correlated (see below), either because the physician failed to grade indepently each of the symptoms or because they all are measuring a common biological phenomenon. In such a case, one is interested in considering the vector of exam age deviations to determine if sex is affecting the process being measured. The multivariate contrast [Kramer, 1972] between sexes was significant at the 0.03 level of probability. The age at which females achieved a given grade of severity (Table III)

TABLE II. Summary of Statistically Significant Relationships Involving Age of Patient

Dependent variable	Predictor	Linear regression coefficient (± SE)
Sweat	Age of diagnosis	1.26(± 0.45)[a]
Exam age	Cough grade	2.40(± 0.50)[b]
	Rales grade	1.42(± 0.56)[a]
	Clubbing grade	2.31(± 0.54)[b]
	Chest shape grade	6.73(± 0.57)[b]
	Chest film grade	1.43(± 0.56)[a]

[a] Significantly different from zero at the 0.01 level of probability.

[b] Significantly different from zero at the 0.001 level of probability.

TABLE III. The Age Difference Between Males and Females at Which a Given Severity Grade Was Achieved

Variable	Male-female age difference (years)
Cough grade	1.2
Rales grade	1.4
Clubbing grade	0.9
Chest shape grade	1.2
Chest film grade	1.9
Average	1.32

ranged from 1.2 to 2 yr younger than males (average difference equals 1.32 yr). A similar effect was apparent with sweat level adjusted for age of diagnosis, and for standardized enzyme dose, liver size, SGPT, SGOT, alkaline phosphatase, total protein, and albumin, each adjusted for exam age. These data support the observation that females are more severely affected by the CF gene [Kramm et al, 1961].

We next turned to evaluating which variables were of greatest value in discriminating among subgroups of patients. The correlation matrix for the 15 continuously varying measures given in Table IV was investigated for this purpose. The male and female matrices were very similar. Within the pooled matrix certain patterns emerge. The age of the patient when his grade of pulmonary severity was recorded was not independently distributed among the five measures considered. The correlation ranged from 0.92 between age/cough grade and age/clubbing grade to 0.99 for ages associated with grade of chest shape and rales. All other correlations were small and not statistically significant at the 0.01 level of probability except for the relationship between SGOT and SGPT (0.87) and between total protein and enzyme dose (0.55).

A principal component analysis using the correlation matrix [Kshirsagar, 1979] showed that the dependence structure in these data may be reduced to three linear functions, which explained 92% of the total variability represented by the 15 variables considered. Sweat, enzyme dose, and the pulmonary variables contributed the greatest to determining total variability. Because all measures were not taken on every patient, the search for the most parsimonious subset of variables must also include the consideration of sample size. Seventy-seven patients were measured for sweat, enzyme dose, and the five measures of pulmonary involvement. Again, three linear functions of these measures on the larger sample accounted for 89% of the total variability. Ninety-six percent of the variability in the five measures of pulmonary function was explained by one linear function, which weighted each variable nearly identically (coefficients ranged from 0.398 to 0.414). This result reflects the high correlations among the exam age measures of pulmonary severity discussed above and justified the selection of only one of this set of severity measures for further consideration.

The combination of sweat, enzyme dose, and exam age adjusted for grade of x-ray findings was used next in a cluster analysis to identify homogeneous subgroups. Age variability was removed from sweat and enzyme dose and each variable was standardized by subtracting the mean and dividing by the standard deviation prior to conducting the cluster analysis. The resulting clusters are presented in Figure 1. The cluster diagram is a nonstatistical statement about the similarities between individuals. Those cases that are joined at small distances are more similar for the traits considered than those connected at larger distances. Four major clusters of patients are suggested by the distance values among subsets. The average distance between pairs of patients (one from cluster 2 and the second from cluster 3 or 4) is at least 40 times as large as the average distance between those individuals placed within clusters 2, 3, and 4. The average distance between clusters 3 and 4 is approximately ten times greater than the distance between pairs of individuals in either cluster. Individuals within these three clusters are on average at least 4.5 (6.8/1.5) and at most 6.8 times (6.8/1.0) more similar than those 29 individuals assigned to cluster 1.

A statistical summary of the unstandardized variables used to establish these hypothesized groupings is given in Table V. The mean values for sweat, enzyme dose, and exam age adjusted for grade of x-ray finding typify the four clusters. Sweat and enzyme dose contribute the greatest to discrimination among clusters. The clusters are not

TABLE IV. Correlation Matrix

Variable	Sweat	Birth weight	Cough	Rales	Clubbing	Chest shape	Chest x-ray	Dose	Liver size	SGPT	SGOT	Alkaline phospha-tase	Albumin	Total protein
Sweat	1.0000													
Birth weight	.0131 (57)	1.0000												
Exam age/ grade of cough	.1022 (79)	.0715 (73)	1.0000											
Rales	.1329 (79)	.1712 (70)	.9528 (97)	1.0000	↙	ALL[d]								
Clubbing	.0969 (78)	.0812 (72)	.9229 (100)	.9524 (96)	1.0000									
Chest shape	.1483 (65)	.1316 (70)	.9346 (86)	.9876 (82)	.9582 (86)	1.0000								
Chest x-ray	.1350 (78)	.1321 (71)	.9393 (99)	.9845 (95)	.9539 (98)	.9844 (84)	1.0000							
Enzymedose[b]	.1219 (78)	−.1164 (71)	−.0402 (99)	−.0224 (95)	−.0523 (98)	−.0852 (84)	−.0259 (98)	1.0000						
Liver size	.2390 (63)	.0902 (64)	−.1394 (79)	−.0877 (77)	−.1240 (79)	−.0824 (78)	−.0845 (77)	.2545 (77)	1.0000					
SGPT[b]	.0092 (41)	−.0346 (42)	.1402 (51)	.0709 (49)	.1296 (51)	.0615 (51)	.0697 (51)	.0808 (51)	.1348 (45)	1.0000				
SGOT[b]	−.0301 (45)	−.0526 (46)	.0585 (55)	.0532 (52)	.1012 (55)	.0418 (55)	.0268 (55)	−.0466 (55)	.1102 (48)	.8678[d] (51)	1.0000			
Alkaline phosphatase[b]	*.2543* (43)	*−.0608* (43)	*.1455* (52)	*.0698* (49)	*−.0611* (52)	*.0105* (52)	*−.0230* (52)	*.2410* (52)	*.1880* (46)	*.3106* (48)	*.2941* (52)	*1.0000*		
Albumin[b]	−.0846 (18)	−.1896 (20)	.1950 (21)	.0400 (18)	.0623 (21)	.0350 (21)	.0770 (21)	.2312 (21)	−.3581 (18)	−.0178 (18)	−.1359 (21)	.1182 (19)	1.0000	
Total protein[b]	.2256 (21)	.2276 (23)	−.0124 (24)	.0944 (21)	−.1477 (24)	−.0635 (24)	−.0617 (24)	.5483[d] (24)	.4230 (21)	−.3099 (21)	−.4231 (24)	.1233 (21)	.3751 (21)	1.0000

[a] Adjusted for age of diagnosis by linear regression. [b] Adjusted for age at examination by linear regression. [c] Number of pairs of observations. [d] Significantly different from zero at the 0.01 level of probability.

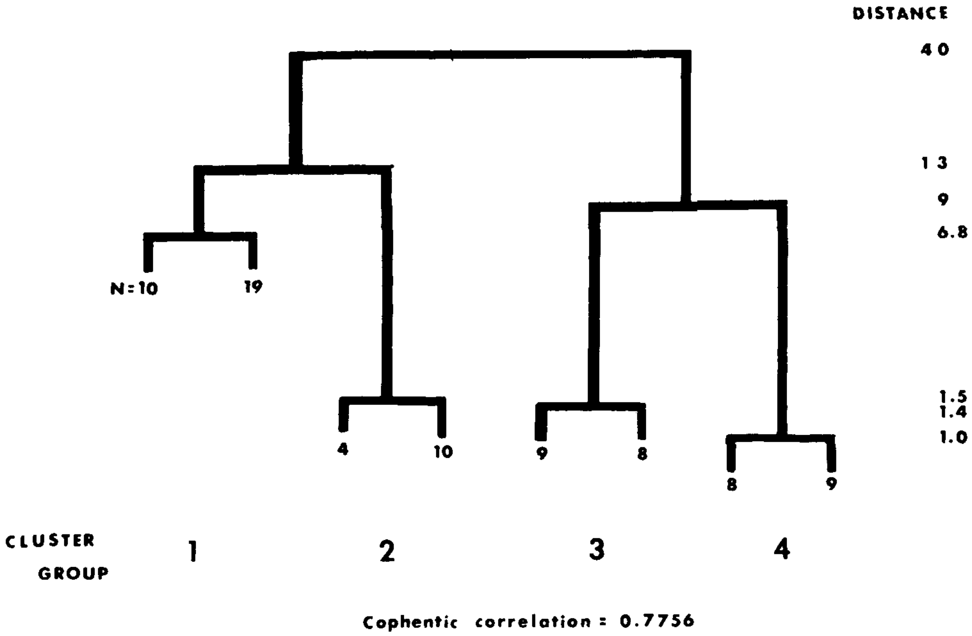

Fig. 1. A cluster analysis using the minimum variance algorithm and the correlation measure of distance.

significantly heterogeneous for variables not included in the cluster analysis. We carried out the cluster analysis on each sex to be certain that the analysis was not sorting out a sex difference. The data presented in Table VI also indicate that the clustering is independent of the sex of the patient.

Figure 2 is a graphic representation of the standardized data being clustered. For example, the centroid for cluster 1 in three-dimensional space deviates high for sweat, low for enzyme dose, and high for the exam age associated with an individual's grade of chest film severity. Sweat level is significantly higher in clusters 1 and 2 than in clusters 3 and 4, while enzyme dose in clusters 1 and 3 is significantly lower than in clusters 2 and 4. Although the test of heterogeneity among all clusters for exam age/chest film grade was not significant (Table V), the pooled contrast of 1 and 3 versus 2 and 4 was significant at the 0.05 level of probability.

The relationship between the variables in the sample of patients that discriminate the four clusters are apparent in the three principal components. They are

$$\lambda_1 = 0.69\,(\text{sweat}) + 0.49\,(\text{dose}) + 0.53\,(\text{exam age}),$$
$$\lambda_2 = 0.00\,(\text{sweat}) - 0.73\,(\text{dose}) + 0.68\,(\text{exam age}),$$
and
$$\lambda_3 = -0.72\,(\text{sweat}) + 0.47\,(\text{dose}) + 0.51\,(\text{exam age}).$$

These orthogonal functions account for 40, 32, and 28%, respectively, of the total variability represented by the three variables. The first is a general measure of the degree of involvement of the three, while the second and third are functions that contrast dose with exam age (λ_2), and sweat with dose and exam age (λ_3). The bivariate scatter of λ_2 and λ_3 for the 77 patients (Fig. 3) suggests that these two functions discriminate among the groups of patients hypothesized from the cluster analysis presented in

TABLE V. Distribution of Variables Among Clusters Identified by Sweat, Exam Age for Chest Film Grade, and Enzyme Dose

		Clusters			Level of
N =	1/29	2/14	3/17	4/17	probability
		Class means (S± SD)			F statistic
Cluster Variables					
Sweat[a]	9.63 ± 14.6	11.21 ± 11.1	−13.78 ± 11.2	−10.07 ± 8.3	0.00
Enzyme dose[b]	−8.20 ± 7.3	8.00 ± 9.3	−6.40 ± 7.6	12.06 ± 9.9	0.00
Exam age/chest film	0.79 ± 7.1	−3.41 ± 4.2	1.44 ± 7.1	−0.66 ● 5.3	0.14
Other Variables					
Birth weight	3,353.4 ± 1,082.0	3,295.1 ± 644.19	3,204.4 ± 752.38	3,309.1 ± 403.4	0.97
Age of diagnois	2.55 ± 3.58	1.36 ± 1.82	3.94 ± 4.82	3.00 ● 3.79	0.28
Exam age	10.03 ± 7.23	6.86 ± 4.59	11.88 ± 7.24	8.76 ± 5.85	0.19
Liver size[b]	−0.13 ± 1.84	1.45 ± 4.01	−0.081 ± 1.8	−0.39 ± 1.48	0.20
SGPT[b]	−6.30 ± 23.19	−5.03 ± 19.94	−10.41 ± 12.64	18.49 ± 90.10	0.57
SGOT[b]	2.94 ± 21.99	−14.28 ± 9.28	2.83 ± 26.64	11.00 ± 71.05	0.58
Alkaline phosphatase[b]	−1.50 ± 227.8	−23.58 ± 118.20	−20.66 ± 110.27	−13.38 ± 127.41	0.99
Albumin[b]	−0.229 ± 1.25	0.236 ± 0.48	−0.194 ± 0.00	0.091 ± 0.733	0.86
Total protein[b]	−0.35 ± 1.04	0.78 ± 2.24	−0.72 ± 1.36	0.42 ± 0.957	0.38

[a] As deviations from the linear regression on age of diagnosis.

[b] As deviations from the linear regression on age at examination.

TABLE VI. Distribution of Enzyme Supplementation, Source of Sweat Test, and Pulmonary Severity Grades Among Clusters

Variable		Clusters 1	 2	Class Frequencies 3	 4	Probability[a]
Sweat test						
UM[b]		20	7	10	11	0.66
No UM		9	7	7	6	
Enzyme type						
Viokase pills		18	7	11	13	0.50
Others		11	7	6	4	
Pulmonary grade						
Cough	1-2	23	9	10	12	0.19
	3-4	6	5	7	5	
Rales	1-2	25	9	10	15	0.14
	2-4	3	4	6	2	
Chest shape	1	16	5	4	9	0.14
	2	5	1	4	5	
	3	3	5	4	2	
Chest film	1-2	23	8	9	14	0.19
	3-4	6	6	7	3	
Clubbing	1-2	20	8	9	9	0.51
	3-4	9	5	8	8	
Abdominal	1	17	7	8	11	0.29
muscle	2	4	1	3	0	
	3	4	3	0	4	
Sex	M	17	8	9	7	0.70
	F	12	6	8	11	

[a] Contingency chi-square.
[b] University of Michigan.

Figure 1. The third component separates the patients with high sweat from those with lower sweat, while the second separates those who are taking low doses of enzyme and are older for their chest film grade from those who are taking a higher enzyme dose and are younger for their chest film grade. Data presented in Table VI argue that the clusters are not significantly different for the distribution of patients among sex, source of sweat test, source of enzyme supplementation, or severity grade for any one of the pulmonary measures of severity.

To investigate the distribution of the four groups of patients among full sibs in families, Table VII was constructed. Of the 11 families with at least 2 sibs with CF, 9 segregated in one of only 3 different ways. From Table VII it can be seen that families 28, 85, and 86 all segregate between groups 1 and 2, families 6, 18, and 39 segregate between groups 1 and 3, and families 38, 54, and 84 segregate between groups 3 and 4.

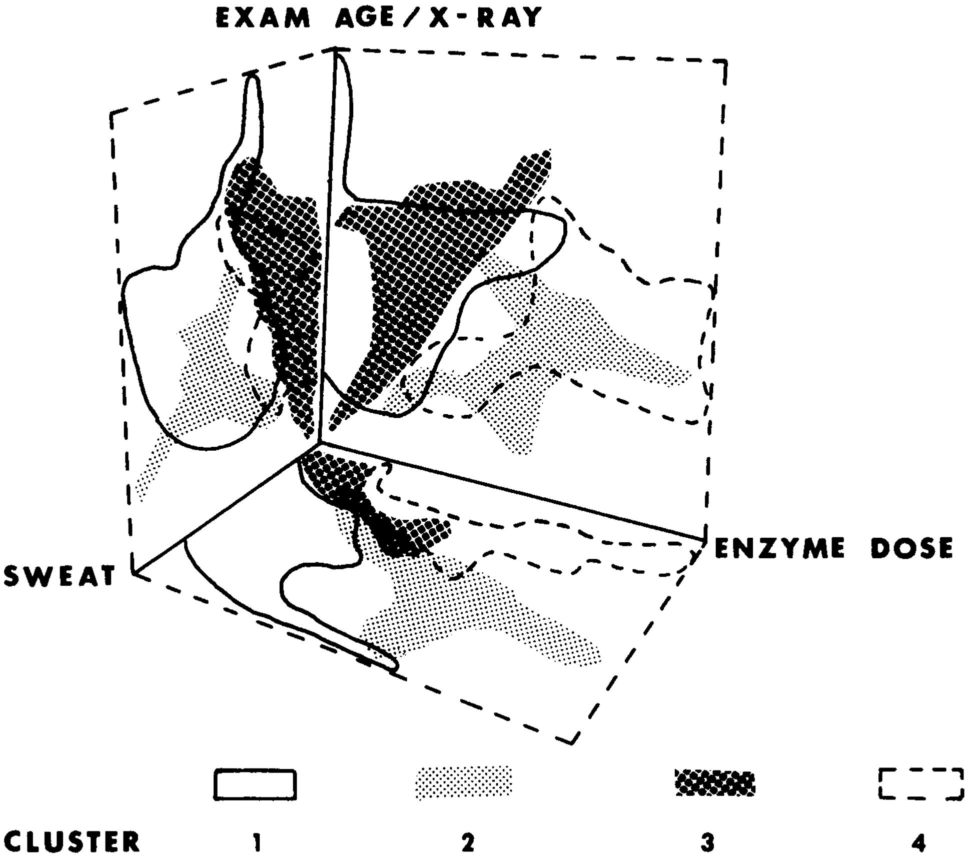

Fig. 2. The distribution of the exam age/chest film and sweat and enzyme dose adjusted for age of diagnosis.

The hypothesis of random distribution of the four types among families is rejected at the 0.05 level of probability. High sweat values aggregate in families 28, 85, and 86, while low values aggregate in families 38, 54, and 84. Within these two groups of families the same relationship between enzyme dose and exam age separates cluster 1 from cluster 2 as separates 3 from 4. A third set of families (6, 18, and 39) have an aggregation of CF children receiving lower than average enzyme supplementation and greater than average exam ages when x-rayed. Children within these families vary markedly for sweat level (clusters 1 and 3). These data suggest that in families with sweat level heterogeneity the enzyme dose and exam age are homogeneous among children. In contrast, those siblings with relatively homogeneous sweat levels consist of children that vary greatly for the level of enzyme supplementation and age when the x-ray was evaluated.

A segregation analysis was carried out to test the fit of a single-locus, recessive genetic model to the distribution of CF within sibships. This analysis followed the a priori method of Bernstein [1925] to correct for ascertainment of sibships through an affected child. The segregation analysis is shown in Table VIII. The recessive hypothe-

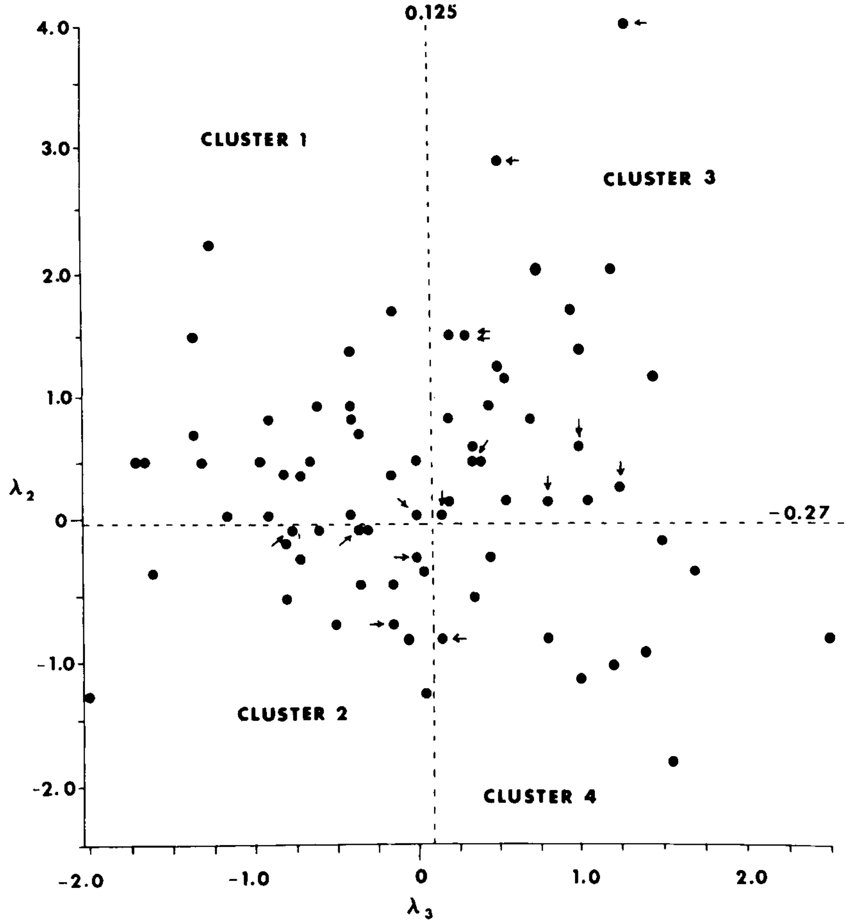

Fig. 3. The distribution of the second (λ_2) and third (λ_3) principal components. Arrows point to individuals who are misclassified by the principal components. The direction of the arrow indicates the group assigned by the cluster analysis summarized in Figure 1.

sis was not rejected by these data ($\chi^2 = 2.45$, 1 df). The chi-square for heterogeneity of the segregation parameter among families of different sizes was 4.5570 (6 df), not significant at the 0.10 level of probability. This result, which documents that these data are not inconsistent with a single-locus recessive inheritance model, may be compared with the segregation analysis performed by Danks et al [1965]. Danks et al rejected the recessive inheritance model when "probable" cases were excluded from the analysis. However, when "probable" cases were included, the results were very compatible with recessive inheritance.

DISCUSSION

The characteristics of the 104 patients examined in this study are similar to those of other groups of cystic fibrosis patients [Nelson, 1969; Shwachman, 1972]. All patients who have had sweat tests have sweat levels above the 60 mEq/liter level, which in-

TABLE VII. Distribution of Hypothesized Phenotypes Among Siblings With CF

	Phenotypic class			
Family number	1	2 most severe	3 least severe	4
28	1	1		
85	1	1		
86	2	1		
6	1		1	
18	1		1	
39	1		1	
54			1	1
84			1	1
39			1	1
59		1	1	
11	2			

TABLE VIII. A Segregation Analysis Based on a Single Locus Model of Recessive Inheritance

Family size (s)	Number of families (N_s)	Observed number affecteds (R_s)	Expected number affecteds ($N_s \bullet [E(R_s)]$)	Variance [$V(R_s)$]
2	28	33	32.00	0.1224
3	18	26	23.35	0.2629
4	11	19	16.09	0.4200
5	3	6	4.92	0.5917
6	5	11	9.12	0.7759
7	2	3	4.04	0.9702
8	1	1	2.22	0.1723

dicates a diagnosis of CF. Most were diagnosed during childhood. There was a variety of levels of severity for the pulmonary variables and also variation in the amount of pancreatic supplementation required. The patients studied are considered to be representative of CF.

The differences between males and females in the age at which each achieved a given grade of severity (Table III) indicates that females tend to be more severely affected by the CF gene than males. On the average females achieved a given grade of severity 1.32 yr ahead of males. Furthermore, the females' age of diagnosis (2.2 yr) was younger than that of males (2.6 yr). The implication is that females may be characteristically more severely affected by the CF gene than males, a finding alluded to in other studies [Kramm et al, 1961; Blacharsh, 1977]. Studies that suggest an excess of surviving males [Shwachman and Kulczycki, 1958; Danks et al, 1965] provide further evidence of a sex difference in severity. The reasons for this finding are at present unknown.

A correlation matrix is presented in Table IV, in which all of the primary variables examined in this study are compared. It is clear that the severity of pancreatic involvement is not correlated with the severity of pulmonary involvement, a finding that has been indicated by other studies [Shwachman et al, 1956, 1965]. It may be further noted from the correlation matrix that all of the measures of pulmonary severity are highly correlated with one another, which is not unexpected since each is probably a measure of the same phenomenon.

The four groups of patients identified by the cluster analysis clearly indicate that this is a heterogeneous disease. The heterogeneity identified here is based on three variables, sweat level, enzyme dosage, and examination age for chest x-ray class. The shape of the cluster diagram (Fig. 1) and the mean values of the cluster variables for each cluster (Table V) argue that there are four quite distinct groups of patients in this data set. The nature of the clustering and the distance between clusters defined by the mean values are consistent with the coefficients of the second (λ_2) and third (λ_3) principal components (Fig. 3). It is highly relevant for the study of CF (and other diseases where phenotypic heterogeneity is an issue) that investigators realize that although the variables of interest may not be correlated over an entire data set (Table IV) documents that the three variables used here are uncorrelated), important information about the variability of patients may reside in the multivariate dimensions of the data set. The CF analysis presented here serves as an example of such a possibility.

Numerous other studies have appeared that review heterogeneity in CF, including heterogeneity in sweat levels, pancreatic involvement, and pulmonary involvement. Gaskin et al, [1980] discuss heterogeneity in a group of CF patients, all diagnosed by sweat electrolyte levels. These patients were divided into two groups on the basis of pancreatic involvement, those with and those without steatorrhea. It was found that those without any apparent pancreatic damage have better pulmonary function, while those with some pancreatic damage have poorer pulmonary function. This finding was confirmed in the present study. In each of the classes where the enzyme dose is low, indicting a milder pancreatic disease, the exam age adjusted for age differences in grade of chest film abnormality is high, indicating a milder degree of involvement of the pulmonary disease. It should also be pointed out that, even within the group of individuals whose pulmonary function is only mildly affected for age (group 3), the pulmonary involvement may still be severe enough to cause death. Because the pulmonary measure of severity reflects the exam age at which a level of damage was achieved, such a finding simply indicates that death due to respiratory failure will occur at an older age in this group of patients. The study by Oppenheimer [1972] reviews autopsied cases of patients dying of complications of cystic fibrosis. He notes that in 8 of 99 cases there was no pancreatic lesion, even in the older cases. These data suggest that the information contributed by sweat level to the discrimination among children (cluster 1 and 2 different from 3 and 4) is independent of the relationship between pancreatic damage and pulmonary function that separates clusters 1 and 3 from 2 and 4.

Other studies have detected heterogeneity based on characteristics that are completely independent from those considered in the current study. These include the studies by Danes et al, who have investigated heterogeneity in cell culture status in groups of CF patients. Danes and Bearn [1969] and Danes and Flensborg [1971] recognized differences in cell culture types between different groups of patients. When the different cell culture classes were then compared to the level of clinical severity, it was shown that one of the cell culture classes, the ametachromatic class, or class III, is a group with an earlier age of diagnosis [Danes and Flensborg, 1971], a poorer prognosis, and more severe involvement [Bearn and Danes, 1978; Danes, 1972].

Principal components analysis describes the relationships between variables that discriminate the groups identified by the cluster analysis. When the second and third principal components are scattered against one another (Fig. 3) it can be seen that each of the cluster groups is located primarily in one quadrant of the scatter. Only 15 of the 77 cases (19.5%) were misclassified when λ_2 and λ_3 were used as discriminators. Thus, these functions and the associated cutpoints of $\lambda_2 = -0.27$ and $\lambda_3 = 0.12$ may be expected to correctly classify unstudied patients into the groups identified by the cluster analysis over 80% of the time. Prospective studies to evaluate these functions are in progress by our group.

The four "phenotypes" identified by the cluster analysis do not appear to be segregating randomly within families. Although the numbers are small, two kinds of familial aggregation of clusters are suggested by these data. In the first, all CF siblings had either higher or lower than average sweat levels and in the second the siblings share levels of enzyme dose and pulmonary involvement. The independent aggregation of these predictors of severity is consistent with a familial basis, either genetic or environmental in origin, for the heterogeneity among the clusterings of CF patients. Furthermore, the independence of the familial aggregation of these two phenotypes is consistent with the hypothesis that the cause of cystic fibrosis may involve two or more genes that modify the control of the expression and penetrance of a single cystic fibrosis gene. A study in progress of a larger number of families with at least two CF children will provide data for the comparison of alternative models for the distribution of the phenotypes suggested by the cluster analysis presented here.

To investigate the course of disease in the subgroups suggested by this study a preliminary followup was carried out. These results are presented in Table IX. It is apparent that group 2, which has the more severe combination of values for the discriminating variables, also had an earlier age of diagnosis, a higher death rate since the last examination, an earlier age of death, and that they died more recently after the last examination than did those in group 3, a less severely affected group of patients.

The analyses carried out in this study may be applied to achieve a better understanding of CF. Having identified these different groups of patients will enable one to prognosticate on the basis of only a small number of variables. These findings may also be used to relate pathogenetically the various aspects of the disease; they provide clues for the development of a genetic model that can explain all of the manifestations of the disease.

CONCLUSIONS

We have confirmed the heterogeneity in CF based on a group of the primary clinical variables. However, the present study has used a different combination of variables to define groupings than those used in previous studies. Further study will be needed to establish whether these variables are the most efficient predicators of the course and severity of the disease and to develop a genetic model to explain the heterogeneity identified.

TABLE IX. Preliminary Followup, 1980 Statistics ± SD

					Living			Died		
							Years			Years
			Age of				since		Age at	after
Group	Hypothesis	N	diagnosis	Exam age	N	Age	exam	N	death	exam
2	Most	14	1.36	6.86	10	10.7	4.1	4	7.75	0.25
	severe		(± 1.82)	(± 4.49)		(± 4.2)	(± 3.15)		(± 3.86)	(± 0.5)
3	Least	17	3.94	11.88	14	16.1	4.1	3	12.70)	(± 1.33)
	severe		(± 4.82)	(± 7.24)		(± 7.4)	(± 2.7)		(± 11.7)	(± 1.53)

ACKNOWLEDGMENTS

This study was supported by DOE grant E (2828), NIH grant 1-T32-GM-07123 Postdoctoral Genetics Training grant, and Cystic Fibrosis Foundation Center grant.

REFERENCES

Andersen DH, Hodges RG (1946): Celiac syndrome. V. Genetics of cystic fibrosis of the pancreas with a consideration of etiology. Am J Dis Child 72:323–339.

Bauman T (1958): Mucoviscidosis as a recessive and irregular dominant hereditary disease: A clinical and genetic study. Helv Paediatr Acta 13:1–79.

Bearn AG, Danes BS (1978): Cystic fibrosis. In Litwin S (ed): "Genetic Determinants of Pulmonary Disease." New York, Marcel Dekker.

Bernstein F. (1925): Zusammenfassende Betrachtungen uber die erblichen Blutstrukturen des Menschen. Z Induk Abstamm Verenbungsl 37:237.

Blacharsh C (1977): Dental aspects of patients with cystic fibrosis: A prelimimary clinical study. J Am Dent Assoc 95:106–110.

Cystic Fibrosis Foundation (1977): "Report of the Patient Registry." Table US-3.

Danes BS (1972): Cell culture and Cystic Fibrosis. In Bergsma D (ed): Part XIII. "GI Tract Including Liver and Pancreas." Huntington, New York: Robert E. Krieger for The National Foundation—March of Dimes, BD: OAS VIII (2): 114–118.

Danes BS, Bearn AG (1969): Cystic Fibrosis of the pancreas: A study in cell culture. J Exp Med 129:775–794.

Danes BS, Flensborg EW (1971): Cystic Fibrosis: Cell culture studies on a Danish population. Am J Hum Genet 23:297–302.

Danks DM, Allan J, Anderson GM (1965): A genetic study of fibrocystic disease of the pancreas. Ann Hum Genet 28:323–340.

Duran BS, Odell PL (1974): "Cluster Analysis: A Survey." New York: Springer.

Frydman MI (1979): Epidemiology of Cystic Fibrosis: A review. Chronic Dis 32:211–219.

Gaskin KD, Gurwitz MC, Levison H, Forstner G (1980): Improved pulmonary function in cystic fibrosis patients without pancreatic insufficiency. In Sturgess JM (ed): "Perspectives in Cystic Fibrosis," Proceedings of the 8th International Cystic Fibrosis Congress, Toronto, Ontario, Canada.

Kramer CY (1972): A first course in methods of multivariate analysis. Virginia: Virginia Polytechnic Institute and State University.

Kramm ER, Crane MM, Brown ML, Sirken MG (1961): Characteristics of patients with cystic fibrosis discharged from hospital in 1957: Estimates for the United States. Pediatrics 28:128–138.

Kshirsagar AM (1979): In Owen DB (ed): "Multivariate Analysis." Statistics: Textbooks and Monographs.

Nelson WE (1969): Vaughan VC, McKay RJ (eds): "Textbook of Pediatrics." Philadelphia: W.B. Saunders.

Oppenheimer EH (1972): Absence of pancreatic lesions in cystic fibrosis. In Bergsma D (ed): part XIII. "GI Tract Including Liver and Pancreas." Huntington, New York: Robert E. Krieger for The National Foundation—March of Dimes, BD: OAS VIII (2): 108–112.

Schaap T, Cohen MM (1976): A proposed model for the inheritance of cystic fibrosis. In Mangos JA, Talamo RL (eds): "Cystic Fibrosis: Projections into the Future (International Conference on Cystic Fibrosis, Jerusalem, Israel)." New York: Stratton Intercontinental Book Corp., pp 291–304.

Scully RE, Galdabini JJ, McNeely BU (1977): Case records of the Massachusetts General Hospital: Weekly clinicopathological exercises. N Engl J Med 296:1519–1526.

Shwachman H (1972): The heterogeneity of cystic fibrosis. In Bergsma D (ed): Part XIII. "GI Tract Including Liver and Pancreas." Huntington, New York: Robert E. Krieger for The National Foundation—March of Dimes, BD: OAS VIII (2): 107.

Shwachman HL, Dooley RR, Guilmette F, Patterson PR, Weil C, Leubner H (1956): Cystic fibrosis of the pancreas with varying degrees of pancreatic insufficiency. J. Dis Child 92:347–368.

Shwachman HL, Kulczycki LL (1958): Long-term study of one-hundred five patients with cystic fibrosis. J Dis Child 96:6–15.

Shwachman HL, Kulcyzycki LL, Khaw KT (1965): Studies in cystic fibrosis. Pediatrics 36:689–699.

Sneath PHA, Sokal RR (1973): "Numerical Taxonomy." San Francisco: WH Freeman.

Ward JH Jr (1963): Hierarchical grouping to optimize objective function. J Am Stat Assoc 58:236–244.

Wood RE, Boat TF, Doershuk CF (1976): State of the Art: Cystic Fibrosis. Am Rev Respir Dis 113:833–878.