UNIVERSITY OF WISCONSIN
DEPARTMENT OF BIOSTATISTICS

Technical Report #85
February 1993

THE INCIDENCE OF CYSTIC FIBROSIS AFTER ACCOUNTING FOR UNDER-DIAGNOSIS

Michael R. Kosorok
Wen-Hsiang Wei
Department of Biostatistics

Philip M. Farrell
Department of Pediatrics

MADISON, WISCONSIN
The incidence of cystic fibrosis after accounting for under-diagnosis

MICHAEL R. KOSOROK, WEN-HSIANG WEI,

Department of Biostatistics, University of Wisconsin,
K6/428 Clinical Science Center, 600 Highland Avenue,
Madison, Wisconsin 53792, U.S.A.

AND PHILIP M. FARRELL

Department of Pediatrics, University of Wisconsin,
H4/458 Clinical Science Center, 600 Highland Avenue,
Madison, Wisconsin 53792, U.S.A.

UNIVERSITY OF WISCONSIN-MADISON

DEPARTMENT OF BIOSTATISTICS

TECHNICAL REPORT 85

SUMMARY

A statistical model for estimating cystic fibrosis (CF) incidence among infants born in the U.S.A. which accounts for under-diagnosis due to death prior to diagnosis is developed and applied to Cystic Fibrosis Foundation Patient Registry data for the years 1989 through 1991. After accounting for 15% under-reporting, the incidence relative to live births of CF among whites is estimated to be 1:3,001 while the incidence among nonwhites is estimated to be 1:10,918. These incidence figures are higher than other published figures. As a by-product of the modeling approach, the underlying average
diagnosis age given survival to diagnosis was estimated to be 3.42 years for whites and 3.63 years for nonwhites, but this difference was not statistically significant and appears to demonstrate that diagnosis effort is approximately the same for whites and nonwhites. Also as a by-product of the modeling approach, CF mortality was estimated to be more severe for females than males and slightly more severe for nonwhites than whites. Excess mortality due to CF appears to increase over the first 30 years of life, but after age 30, mortality for individuals with CF appears to be the same as for normal individuals.

Some key words: Additive hazard; Bootstrap methods; Cystic fibrosis; Computationally intense methods; Genetic defects; Multiple imputation; Nonparametric Maximum Likelihood; Parametric Maximum Likelihood; Incidence studies; Survival analysis.
1. Introduction

Cystic fibrosis (CF) is a common and potentially fatal genetic disorder affecting the lung and gastrointestinal tract. One of the most fundamental gaps in knowledge about the epidemiology of cystic fibrosis concerns the incidence of this disease in North America. Generally, disease incidence figures for genetic disorders affecting infants have been expressed not as absolute rates but in relationship to annual live births, i.e. relative rates (Elandt-Johnson, 1975). By dividing live births in a given year by the determined relative rate for CF, one can calculate a true estimate of the incidence rate or number of persons developing CF expressed per total number of births and per unit of time. Cross-sectional studies in European countries reviewed by Boat et al (1989) have suggested a relative incidence as high as 1:2000 live births, but this figure has been derived from CF prevalence data or the number of persons with this disease expressed per total number of children in the population and extrapolated to annual birth rate. Unfortunately, none of the previous estimates of CF incidence in North America have taken ethnic diversity into account.

Despite its relatively high incidence, CF is difficult to recognize and delays in diagnosis are frequent (Farrell and Mischler, 1992). A glance at the 1992 Cystic Fibrosis Foundation Patient Registry annual data report (Cystic Fibrosis Foundation, 1992) reveals that over half of the CF patients diagnosed in 1991 were diagnosed after the age of 6 months and a few were even diagnosed after the age of 40. Because of the possibility that early medical intervention may benefit CF patients and their families, there is considerable interest in developing screening programs for diagnosing CF as soon after birth as possible (Farrell and Mischler, 1992). Significant progress in developing such a screening program has recently been made at the University of Wisconsin, where a two-tiered screening procedure has been successful at identifying a majority of CF patients before reaching the age of 2 months (Rock et al, 1990, and Farrell, 1993).
Farrell (1993) points out that there are a number of issues to consider before establishing an area-wide CF screening program, including whether or not the regional incidence of CF justifies the expense and whether or not available financial resources are sufficient to provide adequate healthcare management for all diagnosed patients. Clearly, an accurate estimate of the potential number of CF patients born in a given area could help address these issues.

This paper presents a statistical method for estimating the incidence of cystic fibrosis, by race (white and nonwhite) and gender, based on data from the Cystic Fibrosis Foundation Patient Registry for the period between 1 January 1989 and 31 December 1991. This method takes into account both (1) that individuals with CF have a higher mortality than normal individuals and (2) that early death prevents diagnosis of CF for some patients, causing the observed incidence of CF to be lower than the actual incidence at birth. In addition, differences in the diagnosis age distribution between whites and nonwhites will be explored.

The paper will begin with a description of the registry data in section 2. Section 3 will then present and discuss the statistical models used for estimation and inference. Section 4 will discuss the nationwide demographic data which is required—in addition to the registry data—for implementation of these models. Section 5 will describe the final statistical and computational methods which implement the models of section 3 to construct estimates and confidence intervals for CF mortality, rate of diagnosis, and incidence at birth among both whites and nonwhites. The results of these analyses will then be presented in section 6, and a final discussion will be given in section 7.

2. THE CYSTIC FIBROSIS FOUNDATION PATIENT REGISTRY DATA

Data from the Cystic Fibrosis Foundation Patient Registry was provided to us for the years 1989, 1990, and 1991, and contains information for all patients reporting to Cystic Fibrosis Centers during that time period. The information provided to us
includes patient gender, race, date of birth, date of CF diagnosis, and date of death for deceased patients, as well as a number of other variables. FitzSimmons (1991) estimates that about 85% of all diagnosed CF patients are reported to the CF Foundation—this is equivalent to 15% under-reporting—and that about 96% of all deaths attributable to CF, as a primary or secondary cause of death, are reported to the Foundation, whether or not the deceased patients were listed in the Registry prior to their deaths. Because of this, and the careful monitoring done by the CF Foundation, it seems reasonable to assume that nearly 100% of the deaths among CF patients who were listed in the Registry during 1989 through the end of 1991 were accurately reported to the Registry.

Some of the patient data were incomplete or obviously flawed. Of the 19,847 records included in these data (one record corresponds to one patient), 18,810 were complete, with a death date (if the person died) after the beginning of 1989, and directly usable in our analyses. Another 971 records had complete information on date of death (or it was known that they had not died yet), date of birth, and had complete information on at least one of race (white or nonwhite), gender, or diagnosis date. In addition, each of these 971 records had a diagnosis date (if it was not missing) which was later than the birth date and had a death date (if the person had died) after the beginning of 1989. Another 66 records did not satisfy the criteria given for these 971 records and were excluded from our analyses. The final usable data set then consisted of 19,781=18,810+971 records.

The data was then recoded for our use. To explain this recoding, we will need to introduce some notation throughout the remainder of this paragraph. We first converted all dates to decimal years. Let $u_0 = 1989.0$ and $u_1 = 1992.0$, then $u_0$ is the beginning of 1989 and $u_1$ is the end of 1991. Let $N = 19,781$ and, for $k = 1 \ldots N$, let $t_k$ be the birth date for individual $k$, $x_k$ the age (in decimal years) at diagnosis of CF,
and \(\tau_k\) the age at death. The variables used in our analyses are the following:

\[
\begin{align*}
v_k &= x_k \vee (u_0 - t_k), \\
\xi_k &= I_{\{v_k = x_k\}}, \\
\alpha_k &= \tau_k \land (u_1 - t_k), \\
\delta_k &= I_{\{\alpha_k = \tau_k\}}, \\
\epsilon_k &= I_{\{\text{person is male}\}},
\end{align*}
\]

and

\[
\zeta_k = I_{\{\text{person is nonwhite}\}},
\]

where

\[
\begin{align*}
y \vee z &= \max\{y, z\}, \\
\min\{y, z\},
\end{align*}
\]

and

\[
I_{\{X\}} = \begin{cases} 1, & \text{if } X \text{ is true}, \\
0, & \text{otherwise}. \end{cases}
\]

Missing data were then suitably recoded so that they could be identified as missing.

As part of section 5, we will discuss the imputation algorithm we used to fill in the missing data for the 971 partially incomplete records. However, the incomplete or faulty records have recently been repaired by the Cystic Fibrosis Foundation (personal communication with Stacey FitzSimmons, 16 February 1994), and future analyses of these data should not require imputation or omission of bad records.

3. THE STATISTICAL MODELS

As we present the model which we developed for generating estimates of CF incidence, we will need to introduce some new notation from time to time. We will begin by assuming that all CF patients diagnosed in the U.S.A. during the time interval
\([u_0, u_1]\) are in the CF Registry data. (Later, we will adjust our incidence estimates by the 15% under-reporting rate mentioned earlier.) We will also assume that we know the gender- and race- (white and nonwhite) specific birth and death rates for all individuals eligible for observation in the USA during the time interval \([u_0, u_1]\). We will also assume that we know the gender- and race- specific death rates for individuals born with cystic fibrosis. Let \(i\) be the subscript for race, with \(i = 0\) for white and \(i = 1\) for nonwhite, and let \(j\) be the subscript for gender, with \(j = 0\) for female and \(j = 1\) for male.

Define \(S_{ij}^{c}(y)\) to be the known probability that an individual with CF of race \(i\) and gender \(j\) will live beyond age \(y\). Also define \(B_{ij}(s)\) to be the known right-continuous counting process for the births of individuals of race \(i\) gender \(j\) at time \(s\); ie. in counting process notation (see Fleming and Harrington, 1991), \(\int_{t_1}^{t_2} dB_{ij}(s)\) is the total number of infants of race \(i\) and gender \(j\) born in the interval \((t_1, t_2]\). Now define \(D_{ij}(y)\) to be the unknown probability that someone of race \(i\) and gender \(j\) will be diagnosed on or before age \(y\) given that they survive to age \(y\), and define \(p_{ij}\) to be the unknown probability that an individual of race \(i\) and gender \(j\) is born with cystic fibrosis.

We also need to assume that diagnosis time and survival time are independently distributed for individuals. This assumption is perhaps naive, but CF is difficult to diagnosis (Farrell and Mischler, 1992) and the relationship of diagnosis time to time of death is unclear. This assumption also allows us to use mathematical simplifications which permit us to obtain estimates that would be nearly impossible to get otherwise. With these assumptions and definitions in place, the probability that someone of race \(i\) and gender \(j\) born in year \(s\) (prior to year \(u_1\)) will be diagnosed with CF during the time interval \([u_0, u_1]\), can be shown to be

\[
p_{ij} \int_{(u_0 - s)}^{u_1 - s} S_{ij}^{c}(y) dD_{ij}(y).
\]

Since the number of individuals born is large and the probability of being diagnosed
with CF is small, the number of individuals of race $i$ and gender $j$ diagnosed with CF during $[u_0, u_1]$ is approximately poisson with intensity parameter

$$\int_{-\infty}^{u_1} p_{ij} \int_{(u_0-x)_{\leq 0}}^{u_1-x} S_{ij}^c(y) dD_{ij}(y) dB_{ij}(s) = p_{ij} \int_{0}^{\infty} S_{ij}^c(y) \int_{u_0-y}^{u_1-y} dB_{ij}(s) dD_{ij}(y).$$

In order to estimate the diagnosis age distribution, we will only assign probability mass to observed ages of diagnosis. Let $n_{ij}$ be the total number of individuals of race $i$ and gender $j$ who were diagnosed in the interval $[u_0, u_1]$ and let $x_{ij}^m, m = 1 \ldots n_{ij}$, be the corresponding ages at diagnosis. In this setting, $\{x_{ij}^m, m = 1 \ldots n_{ij}\}'$ is the same as

$$\{x_{k}, \text{ for all } k \text{ such that } e_k = j \text{ and } \xi_k = i\}'$$

but the alternative subscripting will temporarily be more convenient. Now let $\Delta D_{ij}(x)$ denote the probability mass for age $x$ such that $\sum_{m=1}^{n_{ij}} \Delta D_{ij}(x_{ij}^m) = 1$. Now the probability that an individual of race $i$ and gender $j$, who was diagnosed with CF in the interval $[u_0, u_1]$, is diagnosed at age $x$ is proportional to $S_{ij}^c(x) \int_{u_0-x}^{u_1-x} dB_{ij}(s) \Delta D_{ij}(x)$: this is simply the probability of surviving to age $x$ times the total number of individuals born who could have attained age $x$ during the interval $[u_0, u_1]$ times the probability of diagnosis at age $x$ given survival to age $x$.

The joint likelihood of $n_{ij}$ and $x_{ij}^m, m = 1 \ldots n_{ij}$, can now be expressed as

$$\Pr(n_{ij}, x_{ij}^m, m = 1 \ldots n_{ij}) = \Pr(n_{ij}) \Pr(x_{ij}^m, m = 1 \ldots n_{ij}|n_{ij})$$

$$= \exp \left\{ -p_{ij} \int_{0}^{\infty} S_{ij}^c(y) \int_{u_0-y}^{u_1-y} dB_{ij}(s) dD_{ij}(y) \right\}$$

$$\cdot \frac{\{p_{ij} \int_{0}^{\infty} S_{ij}^c(y) \int_{u_0-y}^{u_1-y} dB_{ij}(s) dD_{ij}(y)\}^{n_{ij}}}{n_{ij}!}$$

$$\cdot \prod_{m=1}^{n_{ij}} \left\{ \frac{S_{ij}^c(x_{ij}^m) \int_{u_0-x_{ij}^m}^{u_1-x_{ij}^m} dB_{ij}(s) \Delta D_{ij}(x_{ij}^m)}{\int_{0}^{\infty} S_{ij}^c(y) \int_{u_0-y}^{u_1-y} dB_{ij}(s) dD_{ij}(y)} \right\}.$$

Now if we define $u_{ij}^m \equiv S_{ij}^c(x_{ij}^m) \int_{u_0-x_{ij}^m}^{u_1-x_{ij}^m} dB_{ij}(s)$ and let $\{\psi_{ij}^m\}^2 = \Delta D_{ij}(x_{ij}^m)$, we can estimate $p_{ij}$ and $D_{ij}$ by maximizing $\Pr(n_{ij}, x_{ij}^m, m = 1 \ldots n_{ij})$ with respect to $p_{ij}$ and
\( \psi_{m}^{ij}, m = 1 \ldots n_{ij}, \) subject to \( \sum_{m=1}^{n_{ij}} \{ \psi_{m}^{ij} \}^{2} = 1. \) It can be shown, by using the LaGrange multiplier method, that this results in the estimates

\[
\hat{p}_{ij} = \sum_{m=1}^{n_{ij}} \{ w_{m}^{ij} \}^{-1}
\]

and

\[
\Delta \hat{D}_{ij}(x_{m}^{ij}) = \{ \hat{p}_{ij} w_{m}^{ij} \}^{-1},
\]

for \( m = 1 \ldots n_{ij}. \) It can also be shown, under our restricted assumptions and certain regularity conditions, that \( \hat{p}_{ij} \) is unbiased and consistent for \( p_{ij} \) and that

\[
T_{ij} = \sum_{m=1}^{n_{ij}} \frac{x_{m}^{ij}}{\hat{p}_{ij} w_{m}^{ij}}
\]

is consistent for the expected diagnosis age given survival to diagnosis. \( T_{ij} \) is then a rough measure of the average intensity of diagnosis efforts toward identifying cystic fibrosis in individuals of race \( i \) and gender \( j. \)

This maximum likelihood formulation can readily be adapted to generate other maximum likelihood estimates based on more restricted hypotheses. For example, if we assume that the incidence of CF is the same for males and females of the same race and that the diagnosis age given survival is also the same for males and females of the same race, then a collapsed maximum likelihood can be constructed from which \( p_{i} = p_{i0} = p_{i1} \) and \( D_{i}(y) = D_{i0}(y) = D_{i1}(y) \) can be estimated. If we let \( n_{i} = n_{i0} + n_{i1} \) and \( x_{m}^{i}, i = 1 \ldots n_{i}, \) be the corresponding ages of diagnosis for race \( i \) and both genders combined, then the maximum likelihood estimates can be shown to be

\[
\hat{p}_{i} = \sum_{i=1}^{n_{i}} \{ w_{m}^{i} \}^{-1}
\]

and

\[
\Delta \hat{D}_{i}(x_{m}^{i}) = \{ \hat{p}_{i} w_{m}^{i} \}^{-1},
\]

for \( m = 1 \ldots n_{i}, \) and where

\[
w_{m}^{i} = \sum_{j=0}^{1} S_{ij}^{c}(x_{m}^{i}) \int_{x_{m}^{i}}^{u_{1}-x_{m}^{i}} dB_{ij}(s).
\]
The race-specific mean age of diagnosis given survival then becomes

$$T_i = \sum_{m=1}^{n_i} \frac{x_m^i}{\hat{p}_m w_m^i}.$$  

The unbiasedness and consistency properties of the previous estimators also hold for these estimators.

We cannot use this estimation method directly because we do not know the true values of $w_m^{ij}$ or $w_m^i$. Our approach will be to estimate these values from published national estimates, such as U.S.A. census data, and from CF survival function estimates based on the Registry data. For the remainder of this section, we will assume that we know the race- and gender-specific birth processes $B_{ij}(s)$ and survival functions for normal individuals, $S_{ij}^n(y)$, but that we do not know the survival functions for individuals with CF, $S_{ij}^c(y)$. We will, however, assume that the hazard function for individuals with CF is at least as large as the hazard function for normal individuals—this follows from the earlier observation that mortality is greater among individuals with CF than among normal individuals—and that this hazard, for individual $k$, $k = 1 \ldots N$, has the form

$$\lambda_k^c(s) = \lambda_{ij}^n(s) + \exp\{\beta' z_k\} \sum_{l=1}^{\beta_4} \exp\{\alpha_l\} I_{(s \in J_l)},$$

where $\lambda_{ij}^n(s)$ is the race- and gender-specific hazard for normal individuals,

$$\beta = \{\beta_1, \beta_2, \beta_3\}'$$

$$z_k = \{\epsilon_k, \zeta_k, \epsilon_k \cdot \zeta_k\}'$$

$$J_l = \begin{cases} 
[0, 10], & \text{if } l = 1, \\
(10, 20], & \text{if } l = 2, \\
(20, 30], & \text{if } l = 3, \\
(30, \infty), & \text{if } l = 4,
\end{cases}$$

and both $\epsilon_k$ and $\zeta_k$ are as defined in section 2. This simple, parametric form permits us to estimate the CF hazard function over age intervals with sparse data.
Because we only observe CF mortality among individuals previously diagnosed with CF, and because time periods occurring outside of the interval \([u_0, u_t]\) are not necessarily well represented in the Registry data we are using, we will restrict the age interval for which a person is “at risk for CF mortality” to \(K_k = [v_k, a_k]\). “CF mortality,” for the purposes of this paper, simply refers to mortality from any cause which occurs to an individual diagnosed with CF. The indicator \(\delta_k\), as defined in section 2, is then simply a censoring indicator with \(\delta_k = 1\) when death is observed in the interval \(K_k\) and \(\delta_k = 0\) when death is not observed.

The log likelihood for the observed mortality can now be expressed as

\[
\mathcal{L}(\alpha, \beta) = \sum_{k=1}^{N} \delta_k \log \{\lambda_k^*(a_k)\} - \int_0^{z_k} \lambda_k^*(s)ds,
\]

where \(\alpha = \{\alpha_1, \ldots, \alpha_4\}'\). The resulting score vector simplifies to

\[
S(\alpha, \beta) = \sum_{k=1}^{N} \begin{bmatrix}
\exp\{\alpha_1\} \exp\{\beta'z_k\} \left( \frac{I(\alpha_k \in J_1) \delta_k}{\lambda_k^*(a_k)} - \mu(J_1 \cap K_k) \right) \\
\vdots \\
\exp\{\alpha_4\} \exp\{\beta'z_k\} \left( \frac{I(\alpha_k \in J_4) \delta_k}{\lambda_k^*(a_k)} - \mu(J_4 \cap K_k) \right) \\
z_k \exp\{\beta'z_k\} \left( \sum_{l=1}^{4} \frac{I(\alpha_k \in J_l) \exp\{\alpha_l\} \delta_k}{\lambda_k^*(a_k)} - \sum_{l=1}^{4} \exp\{\alpha_l\} \mu(J_l \cap K_k) \right)
\end{bmatrix},
\]

where \(\mu(X)\) is the Lebesgue measure of the set \(X\), eg. \(\mu(J_1 \cap [5.0, 6.5]) = 1.5\) is simply the number of years for which the age interval \([5.0, 6.5]\) overlaps with the interval \(J_1 = [0, 10]\). The second derivatives are straightforward, but lengthy, and will not be presented here.

Maximum likelihood estimates of \(\alpha\) and \(\beta\) can now be obtained and race- and gender- specific survival functions for individuals with CF can be estimated as \(S_{ij}^*(s) = R_{ij}^*(s) S_{ij}^\alpha(s)\), where

\[
S_{ij}^\alpha(s) = \exp \left\{ - \int_0^s \lambda_{ij}^\alpha(t)dt \right\}
\]
and
\[
R_{ij}(s) = \exp \left[ - \exp \left\{ \hat{\beta} \left( \begin{array}{c} i \\ j \end{array} \right) \right\} \sum_{i=1}^{4} \exp \{ \hat{\alpha}_i \} \mu (J_i \cap [0, s]) \right].
\]

These estimates can now be used to generate estimates of the \( w_{ni} \) and \( w_{ni} \) values discussed earlier in this section, which in turn can be used to generate \( \hat{p}_{ij} \), \( T_{ij} \), \( \hat{p}_i \), and \( T_i \), provided we know \( S_{ij}^n(s) \) and \( B_{ij}(s) \).

### 4. National demographic data

This section discusses how we obtain estimates of \( \lambda_{ij}^n(s) \) and \( B_{ij}(s) \) from national demographic data. Values of \( S_{ij}^n(s) \) for integer values of \( s \) were obtained from 1989 U.S.A. life tables (National Center for Health Statistics, 1992a). We interpolated and generated \( \lambda_{ij}^n(s) \) by assuming that the hazard was constant between integer years after age 1. This constant hazard interpolation provides us with continuous approximations of both \( S_{ij}^n(s) \) and \( \lambda_{ij}^n(s) \) and is commonly used in actuarial work (Bowers et al., 1986).

For the period between age 0 and 1, we used the finer resolution of a 1986 life table for all individuals in the U.S.A., presented in Bowers et al (1986), which gave survival function estimates for infants at age 1 day, 7 days, and 28 days, in addition to integer years. We interpolated by again assuming that the hazard rate was constant between time points. We made these infant hazard rates race- and gender-specific by multiplying by a race- and gender-specific factor which made the total hazard during the first year agree with the race- and gender-specific total hazard in the first year as computed from the 1989 life tables. Thus, our final survival functions were race- and gender-specific while also reflecting the rapid change in mortality rate which occurs in the first few weeks of infancy.

We estimated the total number of people alive at the end of 1991 in each one year age group from 0 through 85 by a combination of several methods. First, we used 1990 census data to determine how many were alive over the age of 1.0 years (the census data
for younger ages does not seem to be reliable), by race and gender, at calendar time 1990.5 (U.S. Bureau of the Census, 1990). We assumed that ages were roughly evenly distributed between integer years and then estimated how many would still be alive in each half-year age group at year 1992.0 by integrating the appropriate conditional survival function over the interval \((0.0, 0.5]\). For example, the number of individuals of race \(i\) and gender \(j\) who were in the \((1.0, 1.5]\) age interval in 1990.5 would be multiplied by 

\[
\int_{0.0}^{0.5} \frac{S_{ij}^n(1.0 + s + 1.5)}{S_{ij}^n(1.0 + s)} \, ds
\]

to determine how many individuals were alive in the \((2.5, 3.0]\) age interval at year 1992.0 (the end of year 1991).

In order to determine the number of individuals who were alive in the \((0.0, 2.5]\) age intervals, natality statistics for the years 1989.5 to 1992.0 were obtained from the National Center for Health Statistics (1992b) and the U.S. Bureau of the Census (1992). Race- and gender-specific rates were obtained for half-year intervals. This was accomplished through judicious interpolation and by assuming that natality rates are constant over half-year intervals. As before, the number alive at year 1992.0 in each half-year age group was determined by integrating the appropriate conditional survival function. The first and second half years were then added together to form a race- and gender-specific table of the number of individuals alive at year 1992.0 for each integer age interval from \([0.0, 1.0]\) through \((85.0, 86.0]\) (this required some additional interpolation).

The birth cohorts for each of these age intervals were then calculated by determining how many had to be born in the appropriate birth year in order for there to be the number alive in the corresponding age interval at year 1992.0, assuming that birth rates were uniform during each birth year. The solution to this problem is to divide the number of individuals alive in the age interval \((y, y + 1]\) by \(\int_{0}^{1} S_{ij}^n(y + 1 - s) \, ds\). This
was done for each age interval to determine race- and gender-specific birth cohorts for individuals alive at the end of 1991.

Because the life table used were based on stationary populations (National Center for Health Statistics, 1992a; Bowers et al, 1986), there is a possibility that our birth cohorts might be too large because of the substantial decrease in mortality rates experienced over the last century. However, since we use the same survival functions to go back in time (to determine birth cohort sizes) as well as to go forward in time (to determine the number still alive in the interval $[u_0, u_1]$ who could have been diagnosed with CF) this bias probably cancels itself out in our statistical estimation of CF mortality, rate of diagnosis, and incidence at birth.

5. Statistical and computational methods

This section discusses the statistical and computational methods we used to overcome the missing data problem mentioned in section 2 and how we estimated the uncertainty of our parameter estimates based on the models presented in section 3. In order to minimize bias from under-reporting of CF diagnoses, we felt that it was important to use as much of the data as possible—this is the reason that we did not throw out the 971 partially incomplete, but potentially informative, records. Our approach is a variant of the multiple imputation technique of Rubin (1987) which is computationally feasible for large data sets. The methods discussed were implemented on a Sun Sparcstation 10 in a unix environment with the C programming language and Splus statistical software (Statistical Sciences, Inc., Seattle, WA).

Before proceeding with a description of our imputation method, recall from section 2 that our Registry data consists of two continuous variable $v_k$ and $a_k$ and four dichotomous variables $\xi_k$, $\delta_k$, $\epsilon_k$, and $\zeta_k$. We first needed to form appropriate categories for the continuous variables, so that the complete portion of a partially incomplete record could be matched as much as possible to the corresponding portion of a
complete record. Diagnosis age was divided into four categories corresponding to the
four age intervals [0.00, 0.25], (0.25, 1.00], (1.0, 6.0], and (6.0, \infty): these intervals were
formed on the basis of the distribution of diagnosis age in the Registry data. Death
age was divided into four categories corresponding to the same death age categories as
were formed for the survival analysis model: [0.0, 10.0], (10.0, 20.0], (20.0, 30.0], and
(30.0, \infty). Matching was then done on the basis of these categories for the continuous
variables as well as the dichotomous race and gender variables.

A single imputation of the entire usable Registry data was accomplished by first
randomly permuting the order of the complete records and then filling in missing values
for each incomplete record as follows: a random record from the permuted complete
set was selected as a starting point; if this record matched the complete portions of the
partially complete record, the incomplete portions were filled in with the corresponding
values from the complete record; if this complete record did not match the partially
complete record, the permuted complete file was searched sequentially until a matching
record was identified—if the end of the file was reached before a match was obtained,
the search continued from the beginning of the file. This imputation procedure can be
accomplished quite rapidly with the use of pointer variables and random access files.
Clearly this method is only appropriate for data sets where all levels of all categories
are reasonably well represented; fortunately, this is the case for the Registry data.

After generating a single imputation, the models of section 3 were applied to obtain
the desired parameter estimates. The entire imputation procedure was then repeated
several times until 20 imputations were performed. The mean of each parameter es-
timate for these twenty imputations was calculated to determine the final parameter
estimates. It can be shown that, conditional on the data, each of these imputations
is independent and identically distributed and that, within each imputation, the im-
puted values for all partially complete records of the same categorical combination are
independent and identically distributed conditional on the specific permutation of the complete records. This procedure therefore is slightly less informative than the usual multiple imputation procedure but is computationally much more efficient.

The variability of these final parameter estimates were then obtained by a composite bootstrap procedure. A single bootstrap of the entire procedure was obtained by selecting a random sample with replacement from the complete usable data set of the same size as the usable data ($N=19,781=18,810+971$). The resulting bootstrapped data set may have a different proportion of partially complete to complete records, but this approach approximately simulates the original sampling scheme if we make the "missing at random" assumption (Rubin, 1987). The bootstrapped data set was then divided into complete and partially complete portions and the previously discussed imputation procedure was performed twenty times and average parameter estimates were obtained. This entire procedure was repeated a total of twenty times.

The final parameter estimates were the average parameter estimates from twenty random imputations on the original, non-bootstrapped data; but variability for these estimates were obtained from the twenty composite bootstrap estimates. Since the parameter estimates from both the non-bootstrapped data and the 20 bootstrapped samples should be asymptotically normal, 95% Confidence intervals were obtained based on the $t$-distribution with 19 degrees of freedom. The validity of the normality approximation was evaluated with quantile-quantile (qq) plots. Differences between male and female incidences and average diagnosis ages were tested with paired $t$-tests: the difference to be tested came from the non-bootstrapped data while the variability of this difference was estimated with the sample variance of the corresponding paired differences in the bootstrapped samples.

6. Results

We will begin by presenting the results of the survival analysis estimates. Parameter
Table 1: Parameter estimates and 95% confidence intervals for gender- and race-specific additional hazard resulting from cystic fibrosis.

<table>
<thead>
<tr>
<th>term</th>
<th>parameter</th>
<th>estimate</th>
<th>95% Confidence interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender (male)</td>
<td>exp{β₁}</td>
<td>0.71</td>
<td>[0.64, 0.79]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>race (nonwhite)</td>
<td>exp{β₂}</td>
<td>1.53</td>
<td>[0.99, 2.35]</td>
<td>0.053</td>
</tr>
<tr>
<td>gender × race</td>
<td>exp{β₃}</td>
<td>0.91</td>
<td>[0.41, 2.05]</td>
<td>0.820</td>
</tr>
<tr>
<td>[0,10] years</td>
<td>exp{α₁}</td>
<td>0.0059</td>
<td>[0.0046, 0.0075]</td>
<td>-</td>
</tr>
<tr>
<td>(10,20] years</td>
<td>exp{α₂}</td>
<td>0.0161</td>
<td>[0.0134, 0.0192]</td>
<td>-</td>
</tr>
<tr>
<td>(20,30] years</td>
<td>exp{α₃}</td>
<td>0.0310</td>
<td>[0.0284, 0.0338]</td>
<td>-</td>
</tr>
<tr>
<td>(30,∞) years</td>
<td>exp{α₄}</td>
<td>1.16×10⁻¹⁴</td>
<td>[0.15, 9.09]×10⁻¹⁴</td>
<td>-</td>
</tr>
</tbody>
</table>

estimates and 95% confidence intervals are given in table 1. exp{αᵢ}, i = 1...4, are the additional mortality hazard experienced by a white female with CF, over and above the hazard experienced by a normal (non-CF) white female on the given age interval. exp{β₁} is the effect (as a proportion) of being male; exp{β₂} is the effect (as a proportion) of being nonwhite; and exp{β₃} is the interaction effect, multiplied by the marginal affects, of being both male and nonwhite. Table 1 also presents p-values of the t-tests assessing the significance of the β estimates (the null hypothesis is β = 0). The evidence for females experiencing a higher mortality due to CF than males is substantial, while the evidence for nonwhites experiencing a higher mortality due to CF than whites is only marginal. The excess mortality due to CF for all individuals seems to increase until about age 30, after which mortality rates appear to be unaffected by cystic fibrosis. The qq-plots for these parameter estimates indicate some skewness, particular in α₁, but these problems seem relatively minor.

Parameter estimates and approximate 95% confidence intervals for the incidence and average diagnosis age are given in table 2. The incidence is expressed as an inverse, i.e., the number given is the number of births necessary for one occurrence of CF. The p-value for the paired t-test of the difference in incidence for white males and females is 0.058 while the p-value for the gender difference among nonwhites is 0.893. Although
Table 2: Parameter estimates and 95% confidence intervals (CI) for inverse incidence and average age of diagnosis.

<table>
<thead>
<tr>
<th>parameter</th>
<th>race and gender</th>
<th>estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>inverse incidence</td>
<td>white males</td>
<td>3434</td>
<td>[3280,3587]</td>
</tr>
<tr>
<td></td>
<td>white females</td>
<td>3638</td>
<td>[3453,3823]</td>
</tr>
<tr>
<td></td>
<td>nonwhite males</td>
<td>12658</td>
<td>[8667,16648]</td>
</tr>
<tr>
<td></td>
<td>nonwhite females</td>
<td>13066</td>
<td>[10449,15683]</td>
</tr>
<tr>
<td>average diagnosis age</td>
<td>white males</td>
<td>3.45</td>
<td>[2.73,4.16]</td>
</tr>
<tr>
<td></td>
<td>white females</td>
<td>3.41</td>
<td>[3.03,3.78]</td>
</tr>
<tr>
<td></td>
<td>nonwhite males</td>
<td>3.41</td>
<td>[1.78,5.04]</td>
</tr>
<tr>
<td></td>
<td>nonwhite females</td>
<td>3.82</td>
<td>[2.56,5.08]</td>
</tr>
</tbody>
</table>

This difference for whites is almost statistically significant, the actual magnitude of the difference is small and probably negligible. The p-value for the paired t-test of the difference in average diagnosis age between white males and females is 0.901 while the p-value for the gender difference among nonwhites is 0.710. For these reasons, we used incidence estimates from the likelihood discussed in section 3—based on the assumption of there being no gender difference in incidence or average diagnosis age—to form the final estimates and 95% confidence intervals given in table 3. Also in table 3 are the incidence estimates adjusted for 15% under-reporting. The confidence intervals for the average diagnosis ages of whites and nonwhites overlaps a lot, indicating that the average diagnosis age does not depend on race or gender. The qq-plots for these estimates again indicate some skewness, but this appears to be only a minor problem.

7. DISCUSSION

Our incidence estimate of 1:3001 for whites is higher than that computed by the Cystic Fibrosis Foundation (FitzSimmons, 1991). This may be due to the fact that we adjusted for under-diagnosis due to pre-diagnosis mortality. Gregg et al (1993), on the other hand, estimate the incidence to be 1:3421 for the state of Wisconsin, including both whites and nonwhites, based on a genetic model which should reflect
Table 3: Final parameter estimates and 95% confidence intervals (CI) for inverse incidence and average age of diagnosis after combining males and females. An estimate for the inverse incidence adjusted for 15% under-reporting is also given.

<table>
<thead>
<tr>
<th>parameter</th>
<th>race</th>
<th>estimate</th>
<th>95% CI</th>
<th>adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>inverse incidence</td>
<td>whites</td>
<td>3531</td>
<td>[3400,3662]</td>
<td>3001</td>
</tr>
<tr>
<td></td>
<td>nonwhites</td>
<td>12845</td>
<td>[11638,14051]</td>
<td>10918</td>
</tr>
<tr>
<td>average diagnosis age</td>
<td>whites</td>
<td>3.42</td>
<td>[2.97,3.87]</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>nonwhites</td>
<td>3.63</td>
<td>[2.84,4.42]</td>
<td>-</td>
</tr>
</tbody>
</table>

actual birth incidence. The difference between our estimate and this estimate may be due to the fact that the estimate of Greg et al is based on both whites and nonwhites combined, and the incidence in nonwhites is extremely low. This difference may also indicate that the actual under-reporting rate is less than 15%. The fact that these two estimate are as close as they are lends credence to both estimation procedures. Our incidence estimate of 1:10,918 among nonwhites is substantially higher than other published figures (FitzSimmons, 1991; Kulczycki and Schaaf, 1974), which range from 1:14,000 to 1:17,033 for blacks and even lower for other races. It should be pointed out that the majority of nonwhites in the U.S.A. are blacks, so our estimates for nonwhite populations largely reflect black population characteristics.

The fact that there are no significant differences in the average diagnosis ages among whites, nonwhites, males, and females, may indicate that efforts to identify CF are independent of race and gender. However, the prognosis for CF, in terms of excess mortality, appears to be much more serious for females than males and somewhat more serious for nonwhites than whites—this is in agreement with results from the Cystic Fibrosis Foundation (1992). Another interesting observation is that, although the mortality rate for individuals with CF appears to be substantially greater than for normal individuals over the first 30 years of life, this difference seems to disappear after age 30.
Because our confidence intervals do not take into consideration the uncertainty of either the demographic estimates or the estimated under-reporting rate, the actual uncertainty of our estimates is unclear. This issue, and other statistical as well as non-statistical issues, need to be more carefully evaluated in future studies. However, the results of this study are interesting, reasonably accurate, and potentially useful for making decisions in cystic fibrosis screening policy.

Acknowledgements

This research was supported by grants DK34108 and RR03186-07 from the National Institutes of Health. We would like to thank the Cystic Fibrosis Foundation for providing the Patient Registry data, without which this study would not have been possible. We also thank Stacey FitzSimmons, Andy Schettino, and Derrick Pressly—all of the Cystic Fibrosis Foundation—for their technical assistance, and we again thank Stacey FitzSimmons for several insightful discussions.

References


