Pediatric Pulmonology 30:330–344 (2000)

State of the Art

Modern Statistical Techniques for the Analysis of Longitudinal Data in Biomedical Research

Lloyd J. Edwards, PhD*

Summary. Longitudinal study designs in biomedical research are motivated by the need or desire of a researcher to assess the change over time of an outcome and what risk factors may be associated with the outcome. The outcome is measured repeatedly over time for every individual in the study, and risk factors may be measured repeatedly over time or they may be static. For example, many clinical studies involving chronic obstructive pulmonary disease (COPD) use pulmonary function as a primary outcome and measure it repeatedly over time for each individual. There are many issues, both practical and theoretical, which make the analysis of longitudinal data complicated. Fortunately, advances in statistical theory and computer technology over the past two decades have made techniques for the analysis of longitudinal data more readily available for data analysts.

The aim of this paper is to provide a discussion of the important features of longitudinal data and review two popular modern statistical techniques used in biomedical research for the analysis of longitudinal data: the general linear mixed model, and generalized estimating equations. Examples are provided, using the study of pulmonary function in cystic fibrosis research. **Pediatr Pulmonol. 2000; 30:330–344.**

Key words: statistics; biomedical research; cystic fibrosis.

INTRODUCTION

Longitudinal study designs in biomedical research are motivated by the need or desire of a researcher to assess the change over time of an outcome. In a longitudinal study design, the outcome is measured repeatedly over time for every individual in the study. In addition, associated risk factors may be measured repeatedly over time or they may be static. Compared to cross-sectional study designs, longitudinal study designs can be more efficient, less costly, and more robust to model selection, and they can have increased statistical power.[1,2]

In the past decade, an unscientific observation by the author indicates an increase in the medical literature of the reported use of longitudinal studies in clinical research. There are many reasons why there may be an increase in the use of longitudinal study designs in biomedical research. Two important reasons are advances in statistical theory and computer technology, which have made statistical techniques for the analysis of longitudinal data available to statisticians. Longitudinal study designs have a long history in many scientific disciplines. As a result, longitudinal study design terminology is not standardized. Helms[1] provides an excellent summary of longitudinal design terminology. For example, the phrases "longitudinal study design" and "repeated measures design" are often used synonymously. As another example, in survey research, longitudinal cohort studies are usually referred to as panel studies to distinguish them from studies of birth "cohorts."[3] Duncan and Kalton[4] identified four types of longitudinal study designs in survey research: panel survey, repeated survey, rotating panel survey, and split panel survey. However, this paper will not explore longitudinal study designs in survey research.

There are several classes of longitudinal study designs, including prospective (cohort or follow-up) designs, retrospective (case-control) designs, observational designs, and experimental designs. The prospective longitudinal study design is used to collect data on subjects going

Division of Biometry, Duke University Medical Center, Durham, North Carolina.

*Correspondence to: Lloyd J. Edwards, Ph.D., Division of Biometry, Duke University Medical Center, Box #3827, Hanes House, Rm. 249, Corner of Trent Dr. and Erwin Rd., Durham, NC 27710.

Received 22 July 1999; Accepted 30 March 2000.

forward in time. Usually, subjects are selected with and without risk factors, and then they are followed over time to repeatedly measure a defined outcome variable. The retrospective longitudinal study design does just the opposite, i.e., it is used to collect data on subjects going backwards in time where the outcome variable for both cases (those already known to have disease based on their outcome) and controls (those already known to not have the disease) is repeatedly collected backwards in time.

Though the discussion thus far has used time as the longitudinal metameter, it should be noted that other variables may also be used. For example, in a longitudinal study of pulmonary function in children, height may be used as the longitudinal metameter, with interest being in evaluating how pulmonary function changes with changing height. Another example would be a dose response study of a cholesterol-reducing drug, where repeated measures of cholesterol are made on each individual. Of main interest would be how cholesterol changes with changing dose.

There are many issues, both practical and theoretical, which make the analysis of longitudinal data complicated. Such issues include, but are not limited to, correlation between repeated outcome measurements, missing data, irregularly timed data, mixture of static and time-varying covariates, and availability of software for model fitting. In addition, there are specific issues of concern for the biomedical researcher who is not a statistician: statistical methods must address factors such as ease of use; incorporation of biological assumptions and treatment modalities; and strength of interpretation. The aim of this paper is to provide a review of the state-of-the-art of techniques for statistically analyzing longitudinal data in biomedical research. A preponderance of mathematical detail are eschewed so that important concepts can be highlighted. Though the emphasis of this paper is on biomedical research, much of the paper's discussion is applicable to studies other than biomedical research. Examples are provided, using the study of pulmonary function in cystic fibrosis (CF) research.

CONSIDERATIONS IN THE ANALYSIS OF LONGITUDINAL DATA IN CLINICAL STUDIES

As stated above, there are many issues, both practical and theoretical, which make the analysis of longitudinal data complicated. These issues include, but are not limited to, correlation between repeated outcome measurements, irregularly timed data, missing data, mixture of static and time-varying covariates, and availability of software for model fitting.

Abbreviations

ANOVA	Analysis of variance
CF	Cystic fibrosis
COPD	Chronic obstructive pulmonary disease
F_H	Helms-McCarroll approximate F statistic
FEV_1	Forced expired volume in 1 sec
GEE	Generalized estimating equation
MAR	Missing at random
MCAR	Missing completely at random

Correlation Between Outcome Measurements

An extremely important fact regarding measurements repeated on an individual is that the measurements are typically correlated. Though it could happen that repeated measurements on an individual may not be correlated, it is unlikely that repeated measurements on the same individual will actually be independent. If correlation is ignored, it can negatively impact parameter estimation, hypothesis testing, and efficiency of study design.

In standard univariate regression analyses,[5] a fundamental assumption is that the values of the outcome are independent, whether there is one observation per individual or repeated observations per individual. A simple example illustrates what can happen when correlation between observations are ignored in the univariate case: suppose Y_i, $i = 1, \ldots, n$, are correlated normally distributed observations, with mean μ, and variance σ^2. Assume all pairwise correlations are equal to ρ (e.g., repeated measurements of a healthy individual's weight over a short period of time). Suppose we want to construct a 95% confidence interval for the mean μ. Let's isolate our attention to the upper 95% confidence limit:

$$\overline{Y} \pm 1.96\left[\frac{\sigma^2}{n}\{1 + (n-1)\rho\}\right]^{1/2},$$

where

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{k} Y_i.$$

If we were to ignore the correlation and assumed that the Y_is are instead independent, i.e., assume $\rho = 0$, then the upper 95% confidence interval is given by

$$\overline{Y} \pm 1.96\left[\frac{\sigma^2}{n}\right]^{1/2}.$$

If the correlation is ignored, the computed confidence interval could be much smaller than the nominal level, hypothesis tests can have a much higher Type I error, and statistical power can be lower than planned.

Before the availability of more appropriate longitudinal data analysis techniques, a common method of performing regression analyses for repeated measures was to perform regression analyses, say a simple linear

TABLE 1a—Pediatric Patients, National Cystic Fibrosis Patient Registry 1989–1995, Correlation Matrix for Percent Predicted FEV_1

	1989	1990	1991	1992	1993	1994	1995
1989	1.00						
1990	0.71	1.00					
1991	0.67	0.74	1.00				
1992	0.67	0.72	0.76	1.00			
1993	0.65	0.70	0.73	0.81	1.00		
1994	0.66	0.72	0.75	0.80	0.84	1.00	
1995	0.65	0.70	0.72	0.78	0.81	0.91	1.00

model, for each individual using the same number of parameters and then aggregate the parameters, e.g., taking the average of the individual slopes to obtain a measure of group behavior. Accordingly, each separate individual regression analysis would be conducted under the fundamental assumption that the values of the outcomes were independent, i.e., the regression analyses would ignore correlations. As demonstrated in the univariate example above, ignoring correlations between observations leads to bias in even the simplest cases.

The correlation matrix and/or covariance matrix between observations play an important role in the analysis of longitudinal data. Let's consider a straightforward example: the Cystic Fibrosis Foundation National Patient Registry[6] contains yearly data on over 21,000 registered CF patients in the United States. The Registry contains approximately 85% of all diagnosed cases of cystic fibrosis in the United States and more than 90% of all deaths.[7] Specifically, let's consider the 2,982 registered pediatric patients (ages between 6–18) and 2,105 registered adult patients (ages between 18–45) who had FEV_1 measurements for each of the 7 years from January 1, 1989 to December 31, 1995 (complete data). Tables 1a and b present the estimated correlation matrices for percent predicted FEV_1 in pediatric patients and adults (7 rows and 7 columns). The correlations contain the estimated correlation coefficients between pairs of percent predicted FEV_1 measurements on the same individual in two different years.

In Table 1a, the correlations just below the main diagonal consisting of 1.00 represent correlations between successive adjacent years of data. For example, the correlation between 1989 and 1990 percent predicted FEV_1 is 0.71, and the correlation between 1994 and 1995 is 0.91. Similarly, the correlations just below the diagonal representing adjacent years are correlations for percent predicted FEV_1 which are 2 years apart. In general, all correlations are moderate to large, and their magnitudes demonstrate why the correlations should not be ignored. In addition, it appears that the correlations along a diagonal increase with time, i.e., the more recent years of FEV_1 appear to be more correlated than the earlier years for these data. The latter observation suggests that a possible correlation pattern could be modeled which would enhance accuracy and precision in the longitudinal analyses of these data.

TABLE 1b—Adult Patients, National Cystic Fibrosis Patient Registry 1989–1995, Correlation Matrix for Percent Predicted FEV_1

	1989	1990	1991	1992	1993	1994	1995
1989	1.00						
1990	0.84	1.00					
1991	0.80	0.83	1.00				
1992	0.79	0.83	0.83	1.00			
1993	0.74	0.78	0.78	0.83	1.00		
1994	0.72	0.76	0.76	0.83	0.83	1.00	
1995	0.68	0.71	0.71	0.78	0.78	0.89	1.00

On the other hand, correlations for adults are larger than correlations for pediatrics. Also, in contrast to the pediatric correlations, the adult correlations do not exhibit the same pattern of behavior. The adult correlations along the diagonals have more of a tendency to either decrease over time or rise and fall over time. Though a correlation pattern may exist for the adult patients, the pattern is less clear than for the pediatric patients.

Modern longitudinal statistical methods, such as the general linear mixed model[8–10] and generalized estimating equations,[11] use the correlation (or covariance) between observations in modeling longitudinal data. Grady and Helms[12] provide techniques for selecting appropriate covariance matrices in the analysis of longitudinal data. Adjusting for correlation between observations is one reason that modern longitudinal data analysis techniques are more appropriate than some previous methods of analyses.

Data Collection Schedule

It is customary in a longitudinal study design to address the scheduling for collecting repeated measurements. Helms[1] provides us with two very good definitions regarding data collection scheduling in a longitudinal study:

a) A longitudinal study has a *regularly timed schedule* if measurements are scheduled at equal intervals of the longitudinal metameter. A longitudinal study has *regularly timed data* if measurements are actually obtained at regular intervals of the longitudinal metameter.
b) A longitudinal study has a *consistently timed schedule* if every subject has the same schedule, i.e., is scheduled to be evaluated at the same set of longitudinal metameter values, whether or not the schedule is regularly timed. A longitudinal study has *consistently timed data* if every subject is *evaluated* at the same set of longitudinal metameter values.

In the study of many chronic diseases, one may plan on having a regularly timed scheduled, but the actual data collection is *irregularly timed.* For example, in a longitudinal study of pulmonary function in cystic fibrosis with a regularly timed schedule, say at the end of each month for 6 months, cystic fibrosis patients may have unexpected pulmonary exacerbations during the month which may require measurements sometime during the month in addition to at the end of the month. A further example is that patients may miss the window of opportunity of pulmonary measurement and have to be rescheduled for some other time.

Observe that a longitudinal study can have a consistently timed schedule, but the schedule can be irregularly timed. For example, if data were scheduled to be collected at months 1, 3, and 6, then the study would have a consistently timed schedule but the schedule would also be irregularly timed.

Missing Data

Because longitudinal studies are rarely complete due to patient attrition, mistimed visits, premature study termination, death, and other factors, missing data in longitudinal studies can be a difficult problem to overcome. Missing data makes sense only in the context of a regular or consistently timed data collection schedule. Missing data can be classified into two broad categories: randomly missing data, and nonrandomly missing data. Randomly missing data can be further broken down into "missing completely at random" (MCAR) or "missing at random" (MAR). Nonrandomly missing data are often referred to as informatively missing data.

Little and Rubin[13] provide a formal way of classifying missing values: Let Y* be the vector containing the complete set of observations which would have been obtained in the case of no missing values. Let $Y^{(o)}$ denote the vector of actual, observed measurements and $Y^{(m)}$ denote the vector of missing observations which would have been observed but were not, so that $Y^* = (Y^{(o)}, Y^{(m)})$ can be represented as the partitioned vector of $Y^{(o)}$ and $Y^{(m)}$. Let R denote a set of indicator random variables which delineate which observations in Y* are actually observed, i.e., elements of $Y^{(o)}$, and which are missing, i.e., elements of $Y^{(m)}$. Then, using probabilistic arguments, randomly missing (MCAR and MAR) and nonrandomly missing (informatively missing) can be classified as follows:

1) MCAR means that R is independent of both $Y^{(o)}$, $Y^{(m)}$;
2) MAR means that R is independent of $Y^{(m)}$;
3) nonrandomly missing means that R is dependent on $Y^{(m)}$.

Statistical analysis with randomly missing data has been shown to provide a more tractable solution than with nonrandomly missing data. Though there are several methods in the statistics literature for addressing randomly missing data in longitudinal studies, there appear to be essentially two general approaches:[14] using generalized least-squares, and using the test statistics of Wald.[15] However, there is no general consensus on how to analyze longitudinal data with missing values. In addition, since nonrandomly missing data can be even more of a problem than randomly missing data, there simply are no unified approaches to addressing the problem. Woolson et al.[14] support the latter statement: "Little work has been done on the problem of nonrandomly missing longitudinal data by way of formal modeling of the incompleteness, although models do exist for handling special types of completeness such as censoring or truncation of data beyond a certain time period." Remedies for addressing both randomly and nonrandomly missing data will be active areas of research in the statistical community for years to come.

Static and Time-Varying Covariates

In most longitudinal studies, there is an interest in assessing the relationship between the outcome variable and selected covariates (the word "covariates" is used synonymously with independent variables, predictor variables, explanatory variables, and risk factors). In cross-sectional studies, only static variables such as race and gender, and variables measured at a single point in time such as age, height, and weight, can be covariates. However, longitudinal studies allow for the effect of covariates as they change over time, in addition to the use of static covariates.

For example, consider a longitudinal study of pulmonary function in CF. Variables such as age and weight are obvious time-varying covariates. However, other less obvious time-varying covariates may include *Pseudomonas aeruginosa* status, *Hemophilus influenzae* status, or normal flora status. With the ability to use time-varying covariates, CF researchers can more accurately assess the relationship between pulmonary function and time-varying covariates of interest, instead of forcing the time-varying covariate to be static (e.g., using the last observed measure to determine normal flora status).

STATISTICAL METHODS USED IN THE ANALYSIS OF LONGITUDINAL DATA

Though there are several statistical methods which may be used in the analysis of longitudinal data, the primary focus of this paper is to highlight two modern methods which are receiving considerable attention in both the statistical and subject-matter literature: the *general linear mixed model* and *generalized estimating equations.*

Both the *general linear mixed model* (mixed model) and *generalized estimating equations* (GEE) are part of a broader class of techniques called generalized linear models.[11,16,17] There are three basic extensions of generalized linear models, each reflecting the interpretation of the regression parameters for dependent outcomes which are correlated:

Random effects models (mixed model, subject-specific models);
Marginal models (population-average models);
Conditional models.

Both random effect models and marginal models can be referred to as unconditional models. An unconditional model simply means that the expected outcome is modeled as a (linear) function of time and other covariates which represent both within- and between-subject effects. Conditional models, on the other hand, can be described as (linear) models where the outcome appears on the right side of the regression equation (as a predictor) as well as the left side of the regression equation, i.e., the mean or probability of the outcome variable is conditional on the other values of the outcome (in many settings, the conditioning is on the prior value(s) of the outcome). Conditional models are outside the scope of this paper, but the interested reader is directed to Rosner and Munoz[18] and Rosner et al.[19] for further reading.

Random effects models (mixed models) are regression models which are particularly suited for analyzing correlated outcomes which are continuous. The mixed model provides estimation and hypothesis testing for simultaneously modeling both population effects (fixed effects) and random effects (subject-specific effects). Marginal models are particularly relevant when the main focus of a study is investigating the effects of covariates on the population mean. GEE is a method of estimation in marginal models (GEE is *not* a model, but an estimation technique) with correlated outcomes. GEE can be applied to marginal models where the outcome is either continuous or categorical.

For readers without a matrix algebra and/or calculus background, the definitions and assumptions below may be skipped. Instead, the reader may use the Appendix for a more specific and less complex presentation of the mixed model, which may help to facilitate their understanding of this complex modeling technique.

General Linear Mixed Model

The *general linear mixed model,* referred to henceforth as the mixed model, is a multivariate regression method that helps to generalize the analysis of variance (ANOVA) and general linear regression methods. The mixed model[9] is a general statistical technique for analyzing longitudinal data, but can also be used to analyze cross-sectional data. The mixed model is a statistical method for modeling continuous outcome measures as a function of fixed (population) effects, while simultaneously modeling individual subject parameters as random effects. The mixed model can accommodate both time-dependent covariates and static covariates.

Mixed model statistical methods are not especially new. Seminal theoretical papers published by Harville[20] and Laird and Ware[9] helped to popularize the use of mixed models in practice. Although many papers, both theoretical and applied, subsequently appeared in the statistical literature,[21–27] most statistical textbooks do not yet include discussions of mixed models.

We present an abbreviated discussion of general definitions and assumptions used in the formulation of the general linear mixed model.

Definitions and Assumptions

The following discussion provides the general definition and notation of the general linear mixed model for the analysis of incomplete longitudinal data. The mixed model, which contains both fixed and random effects, is given by

$$Y_i = X_i\beta + Z_i d_i + e_i,\ i = 1, \ldots, k,$$

where Y_i is an $n_i \times 1$ vector of n_i observations on the i-th subject; β is a $p \times 1$ vector of unknown, fixed, population parameters; X_i is an $n_i \times p$ known, constant design matrix for the i-th subject; d_i is a $q \times 1$ vector of unknown, random individual parameters. The random parameters are subject-specific, but the vector size is the same from subject to subject; Z_i is an $n_i \times q$ known, constant design matrix for the i-th subject corresponding to the random effects d_i; and e_i is an $n_i \times 1$ vector of random error terms.

For $i = 1, \ldots, k$, it is assumed that the random subject parameters, d_i, have independent, multivariate normal distributions with mean vector zero and covariance D, denoted $d_i \sim \text{NID}(0, D)$, where D is an unknown, positive-definite matrix. Similarly, it is assumed that the vectors of random error terms, e_i, have independent, multivariate normal distributions with mean vector zero and covariance $\sigma^2 I_i$, denoted $e_i \sim \text{NID}(0, \sigma^2 I_i)$, where I_i is an $n_i \times n_i$ identity matrix, and σ^2 is the scalar within-subject variance parameter. It should be noted that the covariance of e_i could be expressed more generally using $\sigma^2 W_i$, where W_i is a known, $n_i \times n_i$ positive-definite matrix. The random subject parameters, d_i, are assumed to be independent of the vector of random error terms, e_i. The correlation between the individual random effects is obtained as a function D.

From the above definitions, it can easily be shown that $E(Y_i) = X_i\beta$ and $\text{Var}(Y_i) = V_i = Z_i D Z_i' + \sigma^2 I_i$, where V_i is the $n_i \times n_i$ positive-definite, symmetric covariance

matrix of Y_i. The covariance matrix V_i can be viewed in a couple of ways when attempting to model it. Since V_i is $n_i \times n_i$, then there are $n_i(n_i - 1)/2$ parameters which require estimating. However, taking advantage of the writing of V_i as a function of D, $V_i = Z_i D Z_i' + \sigma^2 I_i$, allows the flexibility of reducing the number of parameters to $[q(q - 1)/2] + 1$. Also, since each subject is allowed to have unique fixed effect and random effect design matrices X_i and Z_i, the mixed model can accommodate time-dependent covariates and missing and mistimed observations.

Maximum likelihood estimators of the parameters (restricted or unrestricted) in the mixed model generally do not have explicit solutions. Hence, complex iterative computer algorithms are usually required to derive estimates that maximize the likelihood of the observed data. Two frequently used algorithms are the EM (expectation and maximization) algorithm and the Newton-Raphson algorithm. Detailed discussions of these iterative computer algorithms are beyond the scope of this paper, and we refer the interested reader to Lindstrom and Bates.[28] The introduction of SAS Proc MIXED[29] has greatly facilitated the implementation of the mixed model for general practitioners.

Generalized Estimating Equations

Generalized estimating equations (GEEs) are an approach[11,30–32] which specifies only the marginal distribution of the outcome variables. GEE is an estimation technique which estimates a common scale parameter and a working correlation matrix of the outcome variables, treating them as nuisance parameters. GEE can be used for both discrete and continuous outcomes, but is mostly used for discrete outcomes and even then, most real-life applications are correlated binary outcomes.

In specifying only the marginal distribution of the outcome variable, GEE will produce estimates of population parameters only (modeling of population mean only). Hence, GEE cannot be used in settings where subject-specific estimation and hypothesis testing are required. In contrast, the mixed model does provide subject-specific as well as population estimation and hypothesis testing for continuous outcome measures.

We present an abbreviated discussion of general definitions and assumptions used in the formulation of the GEE.

Definitions and Assumptions

The following discussion provides the general definition and notation of the generalized estimating equation (GEE) approach to the analysis of incomplete longitudinal data.

Suppose $Y_1, Y_2, \ldots, Y_k$ are independent vectors with means $\mu_1, \mu_2, \ldots, \mu_k$. For the j-th element of the i-th subject, let

$$E(Y_{ij}) = \mu_{ij},$$
$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta,$$
$$\mathrm{Var}(Y_{ij}) = \phi h(\mu_{ij}),$$

were $g(\cdot)$ is called the link function, $h(\cdot)$ is the variance function, η_{ij} the linear predictor, and ϕ the scale or dispersion parameter. The GEE can be formed by the following:

$$\sum_{i=1}^{k} D_i^T \Sigma_i^{-1} (Y_i - \mu_i) = 0,$$

where

$$D_i = \frac{\partial \mu_i}{\partial \beta}$$

and $\Sigma_i = \mathrm{Var}(Y_i)$. The diagonals of Σ_i are determined by $\mathrm{Var}(Y_{ij}) = \phi h(\mu_{ij})$. To determine the off-diagonal elements first, $\Sigma_i = \phi A_i C_i A_i$, where $A_i = \mathrm{diag}(\sqrt{h(\mu_{ij})})$ and $C_i = \mathrm{Corr}(Y_i)$.

In solving the GEE, the correlation matrix is assumed to be parameterized by an $s \times 1$ vector ρ. An estimate of ρ is plugged into the equation and estimation then proceeds. The assumed correlation structure is called the working correlation matrix, denoted by R_i. In all likelihood, the working correlation matrix R_i may not be identical to the true correlation matrix C_i. Thus, GEE is solved by

$$\sum_{i=1}^{k} D_i^T V_i^{-1} (Y_i - \mu_i) = 0$$

where $V_i = A_i R_i A_i$.

Using GEE, the estimate of β is "nearly efficient relative to the maximum likelihood estimates of β in many practical situations, provided that $\mathrm{Var}(Y_i)$ has been reasonably approximated,"[11] and the estimate of β converges in probability to the true value "even if the covariance structure of Y_i is incorrectly specified."[11] In other words, good estimates of population parameters can be achieved even when the within-subject variances are only roughly approximated. The use of GEE by the practitioner is aided by software such as SAS Proc Genmod.[33]

EXAMPLE OF LONGITUDINAL ANALYSES, USING PULMONARY FUNCTION IN CYSTIC FIBROSIS RESEARCH

We will examine the application of the mixed model and issues arising in the analysis of complex longitudinal

data, using data collected on cystic fibrosis subjects. Since GEE and the mixed model give quite similar results for the population estimates in this example, only the mixed model will be discussed for the sake of clarity and simplicity.

Cystic fibrosis (CF) is the most common lethal autosomal recessive genetic disease among Caucasians. The clinical course of CF varies widely; however, although CF affects multiple organs, the majority of morbidity and mortality in these patients is the result of pulmonary complications. CF is a chronic obstructive pulmonary disease which is both studied and treated within a longitudinal framework. Collecting spirometric data longitudinally on CF subjects has been the norm for many years. However, many of the major results assessing the relationship between CF pulmonary function outcomes such as forced expired volume in 1 sec (FEV_1), forced vital capacity (FVC), and maximum mid-expired flow (MMEF) and possible predictor variables (genotype, pancreatic status, age, gender, and a host of others) have been based on cross-sectional analyses.[34–36]

Longitudinal pulmonary function data from CF patients typically have undesirable characteristics from a statistical viewpoint. Longitudinal CF data are often *irregularly timed*, i.e., obtained at irregular time intervals. The subjects often miss scheduled visits; the available data are *incomplete.* When a visit is missed for reasons related to the underlying disease process, the data are said to be *informatively censored;* a patient who dies before the scheduled end of data collection produces an extreme form of informatively censored data. Most longitudinal CF data have all of these characteristics; any one is sufficient to defeat traditional statistical methods, leading to incorrect statistical analyses in most cases. This may explain the prevalence of cross-sectional analyses in much CF research.

Some studies of non-CF populations comparing cross-sectional analyses with longitudinal analysis of the same data have shown conflicting conclusions.[37–42] For instance, some studies[39,40] have apparently shown the rate of decline in FEV_1 to be significantly greater when using cross-sectional analysis than longitudinal analysis. Other studies,[41,42] however, found the rate of decline in FEV_1 greater using longitudinal analysis than cross-sectional analysis. Pattishall et al.,[41] studying cross-sectional and longitudinal estimates of lung growth in children, found noncomparability of cross-sectional and longitudinal analysis. They concluded that longitudinal studies should be compared with longitudinally collected data, and cross-sectional studies should be compared with data collected cross-sectionally. In addition, van Pelt et al.[40] presented evidence that reference equations based on cross-sectional studies may overestimate longitudinal change; that, in turn, can lead to underestimating the effects of exposure. Longitudinal studies generally have more statistical power than cross-sectional studies and are more robust to model selection.[2]

This CF example also discusses the implications of using cross-sectional methods in the design of controlled clinical trials in cystic fibrosis research when longitudinal data are available. By way of this example, the differences in interpretations one could potentially get from using cross-sectional methods as compared to longitudinal methods of analyses in cystic fibrosis are highlighted.

Issues in CF Research

Several controlled clinical trials per year are conducted on CF subjects by both academic researchers and pharmaceutical companies. Because of the consequences of disease progression in CF, it is important that any controlled clinical trial in CF research have an appropriate study design, statistical analyses, power computation, and sample size determination. As with any controlled clinical trial, a study design which has less power than it should to assess change in CF pulmonary function can seriously undermine study results. In addition, an inappropriate statistical analysis to assess rate of change in pulmonary function in CF subjects can lead to erroneous conclusions regarding the rate of change.

Cross-sectional data analysis techniques have not been issues of concern for CF researchers. Traditional statistical analysis techniques for assessing change in CF pulmonary function have been readily available for many years. These techniques include the use of *t*-tests[37] and general linear regression models.[36,37] Statistical power and sample size computations for the cross-sectional techniques, though at times challenging, have also been available. Statistical software to perform cross-sectional analysis has facilitated the use of cross-sectional methods in CF research.

For CF investigators, general statistical techniques for assessing the relationship between longitudinal pulmonary function outcomes and predictor variables of interest have been lacking. Any general statistical technique for analyzing these relationships has to have features needed by the CF researcher: ease of manipulation, modeling of fixed population effects and random subject effects, the incorporation of biological assumptions and treatment modalities, adjustment for correlated observations (within-subject), and proper handling of irregularly timed and missing observations. In addition, to properly design a controlled clinical trial in CF patients, the investigator must be able to compute power and sample size based on the available longitudinal analysis technique.

Fortunately for CF researchers, the general linear mixed model has the desirable features discussed previously. Applications of the mixed model are increasingly appearing in the literature,[43–45] and the mixed model was

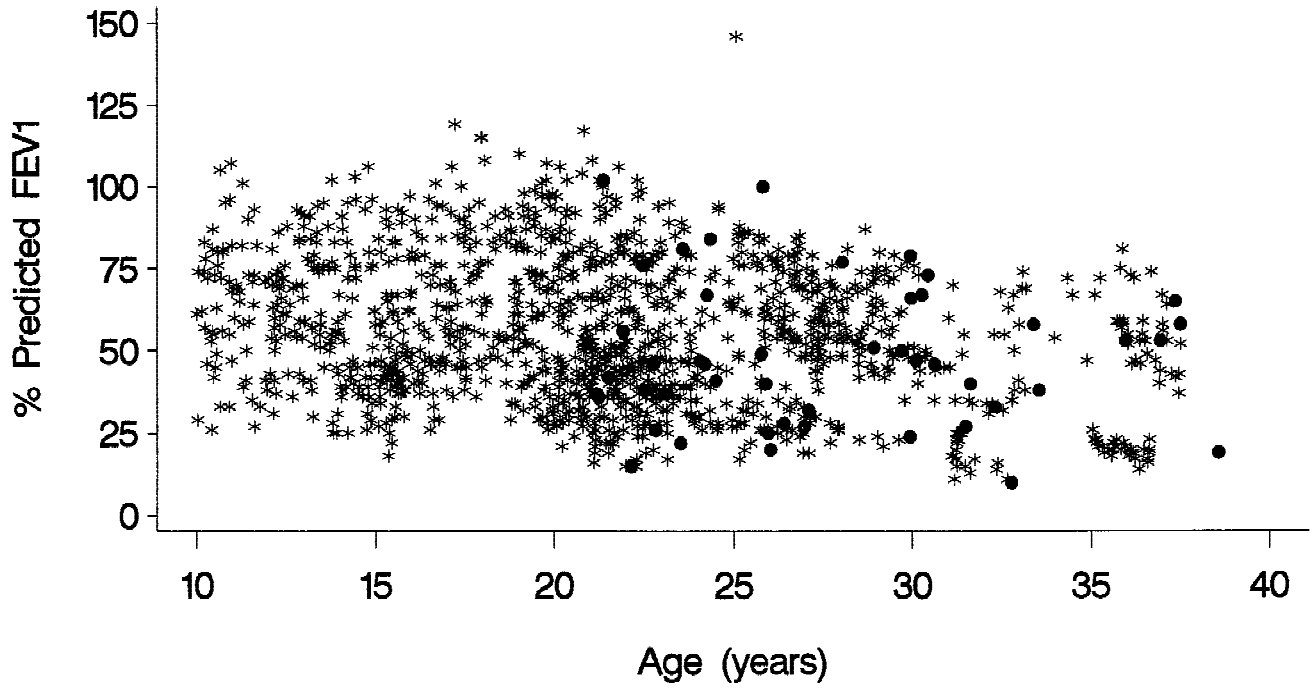

Fig. 1. Longitudinal and cross-sectional adult cystic fibrosis data, 47 subjects and 1,401 observations. *, percent predicted values (1,401 observations); ●, last percent predicted value for each patient.

used as the statistical technique of choice in proposed clinical trials involving the longitudinal analysis of pulmonary function data (this is the author's personal observation as a member of the Clinical Research Committee of the Cystic Fibrosis Foundation[46]).

Data Analysis Using FEV_1 Percent Predicted Values

Longitudinal FEV_1 data were available from clinical follow-up of 47 adult CF patients (23 female, 24 male) seen at the University of North Carolina pulmonary clinic. Complementary data for some of the patients were obtained from other institutions where they had been followed for long-term care. Pediatric measurements (10 < age < 19 years; 10–18) were obtained in a subset of 19 patients, including some data from a subset of patients from other institutions. All subjects were pancreatic insufficient as adults, and all were homozygous for the most common CF mutation, ΔF508. From 4 to 97 (median = 27) measurements were documented per subject.

Figure 1 is a scatter plot of the 1,401 FEV_1 percent predicted values for the 47 homozygous ΔF508 CF patients. The solid circles in Figure 2 represent the last FEV_1 percent predicted value for each patient. Observe that these last values are recorded for each patient after age 18. The values will be used to demonstrate the results of performing a longitudinal analysis vs. a cross-sectional analysis for this group of CF patients. Ideally, it would be helpful to have a unique plot symbol for each of the 47 subjects in Figure 1 so that the reader could get a better view of the individual's longitudinal data. Unfortunately, such a plot would be very crowded and of dubious value. However, Figure 1 does provide the reader with a sense of the complexities facing the CF researcher in analyzing longitudinal pulmonary function outcomes.

Figure 2 is a plot of the linear regression line resulting from a cross-sectional simple linear regression analysis of the 47 last FEV_1 percent predicted values for each patient. Figure 2 also presents the scatter plot of the 47 last FEV_1 percent predicted values for each patient. Table 2 provides the parameter estimates, standard errors (SE), and *P*-values of the intercept and slope resulting from the cross-sectional analysis. The estimate of the model variance is also given. The reader should note that the intercept is centered at age = 25 so that the intercept estimate will be meaningful, i.e., 25 is subtracted from all ages before performing estimation.

Figure 3 is a plot of the population regression line (thick line) and the individual regression lines (thin lines) resulting from a mixed model analysis using the 1,401 FEV_1 percent predicted values for the 47 patients (see Appendix for details). Table 3 provides parameter estimates, SE, and *P*-values of the population intercept and slope obtained from the mixed model analysis. The estimate of the within-subject variance is also given. The intercept is centered at age = 20.

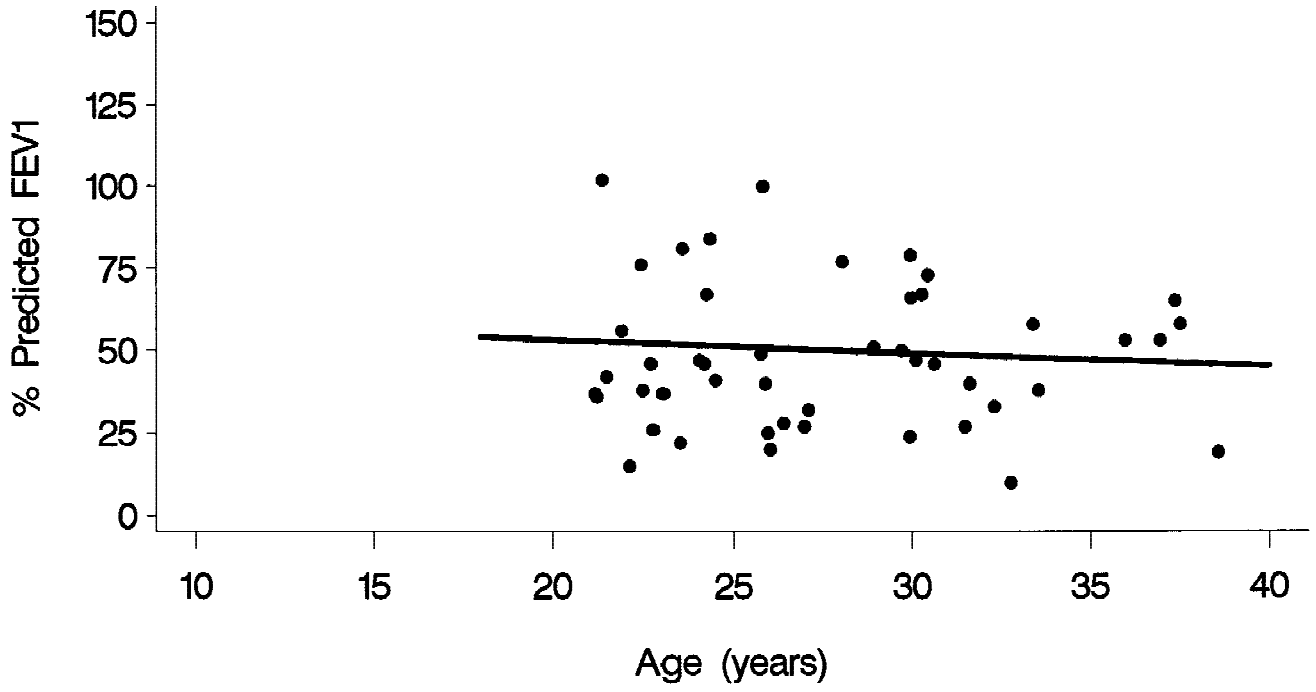

Fig. 2. Cross-sectional analysis regression line and scatter plot, N = 47 subjects with cystic fibrosis. ●, last FEV, percent predicted value for each patient.

TABLE 2—Cross-Sectional Regression Model for FEV_1, Percent Predicted, Intercept Estimated at Age = 25 Years

Parameter	Estimate	SE	*P*-value
Intercept	49	3.7	0.0001
Age	−0.2	0.7	0.75
σ^2	485		

The estimate of the slope of the population regression line from the mixed model is over 10 times that of the estimate of the slope from the cross-sectional analysis. In addition, the conclusion from the cross-sectional analysis is that there is no statistically significant rate of decline in FEV_1 percent predicted over time for this group of adult CF subjects. In contrast, using a mixed model analysis, there is a statistically significant rate of decline in FEV_1 percent predicted over time for this group. Just as important, interpretation of the magnitude of rates of decline from both the cross-sectional and longitudinal analyses would affect the planning of a controlled clinical trial.

Figure 4 is a plot of the linear regression line resulting from a cross-sectional linear regression analysis for the 23 female and 24 male patients. Table 4 provides parameter estimates, SE, and *P*-values of the population intercepts and slopes obtained from the cross-sectional analysis. For this analysis, there are no significant differences between males and females in the intercept or slope. Note that the estimates of the slopes for females (1.1) and males (−0.9) are in opposite directions. Even though the slopes are not statistically different than zero, interpretation of the estimates of rates of decline for males and females would affect the planning of a controlled clinical trial.

Figure 5 is a plot of the population regression lines (thick dashed and solid lines) and the individual regression lines (thin dashed and solid lines) resulting from the mixed model analysis using the 1,401 FEV_1 percent predicted values for the 23 female and 24 male patients. Table 5 provides parameter estimates, SE, and *P*-values of the population intercept and slope obtained from the mixed model analysis. The estimate of the within-subject variance is also given. For this analysis, there are no significant differences between males and females in the intercepts or slopes.

A note of caution should be sounded here. The previous analyses are not presented as an exhaustive or complete modeling (curvilinear or nonlinear) of FEV_1 percent predicted. Although other covariates such as genotype, pancreatic status, gender, and/or polynomial age may be considered when assessing the relation between age and FEV_1, we have limited our analyses to two examples: one using a simple linear mixed model with age (in years) as the time-varying covariate, and the second using both gender and age. In addition, only patients who survived to adulthood were included in this analysis, so that parameter estimates are affected by survival bias and do not represent the whole CF population.

POWER AND SAMPLE SIZE CONSIDERATIONS

This section illustrates how very different statistical power and sample size computations may be obtained

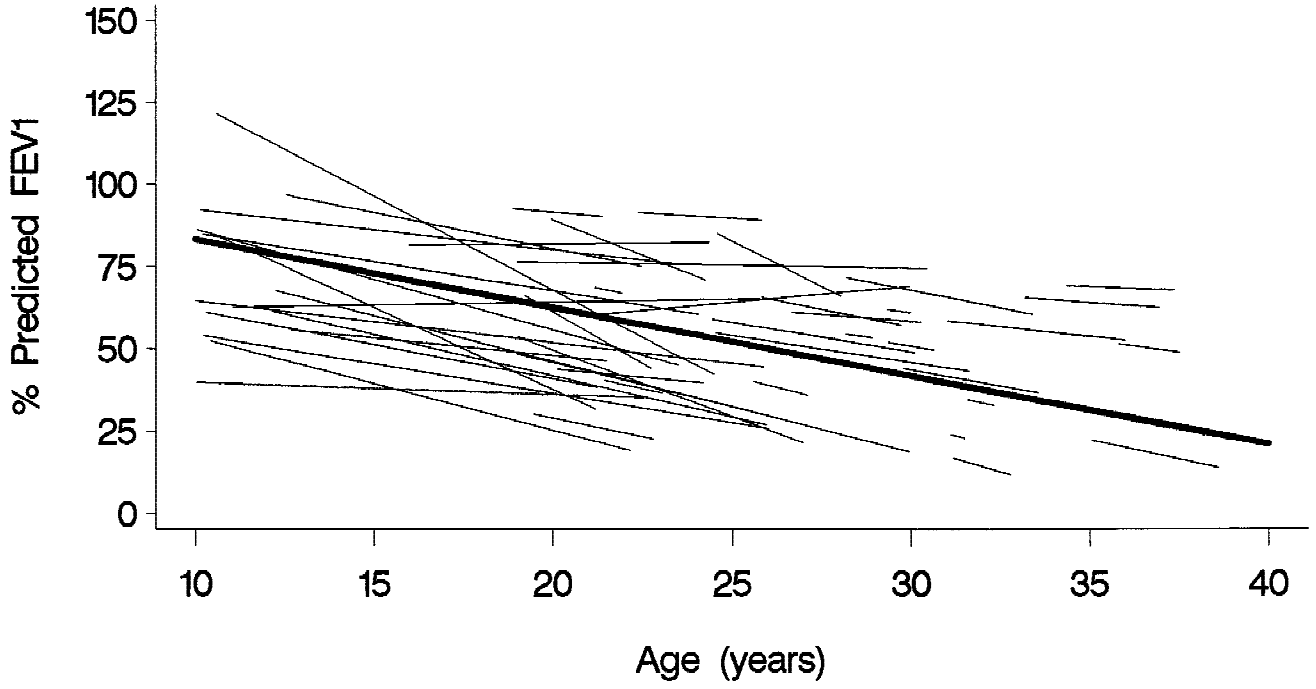

Fig. 3. Mixed model analysis. Population and individual regression lines, 47 subjects with cystic fibrosis and 1,401 observations.

TABLE 3—Mixed Model Regression Model for FEV_1, Percent Predicted, Intercept Estimated at Age = 20 Years

Parameter	Estimate	SE	*P*-value
Intercept	62	3.4	⩽0.0001
Age	−2	0.34	⩽0.0001
σ^2	114		

using cross-sectional vs. longitudinal methods for a chronic lung disease like cystic fibrosis.

General Framework

Since cystic fibrosis is a deadly, chronic pulmonary disease, it is very important to design the most effective and most powerful controlled clinical trials possible. To do this, the use of longitudinal statistical methods will have to become a mainstay in the design of controlled clinical trials in CF research.

In practice, when designing a controlled clinical trial, the CF researcher will know the approximate number of subjects he/she can reasonably expect to obtain (constraints may be due to budget concerns or subject availability) and know a minimum number of measurement occasions he/she can expect. What remains is for the investigator to compute the statistical power associated with the respective sample size to detect a difference of sufficient importance. To compute statistical power, several assumptions must be made. The assumptions should be well-stated for the acceptance of the statistical power computations.

A general framework for computing statistical power can be given as follows:

1) The outcome variable or efficacy variable of interest should be clearly stated.
2) It is assumed that a rough estimate of a (minimal) clinically significant difference to detect can be provided. Where this estimate is obtained, it should be stated clearly.
3) It is assumed that an estimate of the relevant variance can be provided. Where this estimate is obtained, it should also be made clear.
4) It is assumed that the statistical procedure for hypothesis testing has been determined and is clearly communicated.
5) It is assumed that a specified sample size is provided and a Type I error level, α, is specified.

From the above assumptions, a statement of statistical power can be made. Similarly, sample size computations can be placed in the same general framework by changing the phrase "sample size" in item 5 above to "power."

In computing statistical power or sample size for a planned controlled clinical trial, estimates are needed which have been derived from a similar study design or at least from procedures which have similar assumptions. For a planned longitudinal controlled clinical trial in CF research, estimates should be derived from a longitudinal analysis. In addition to estimates of the relevant variance and difference to detect, the computation of statistical power for longitudinal designs typically requires an estimate of the correlation between successive measure-

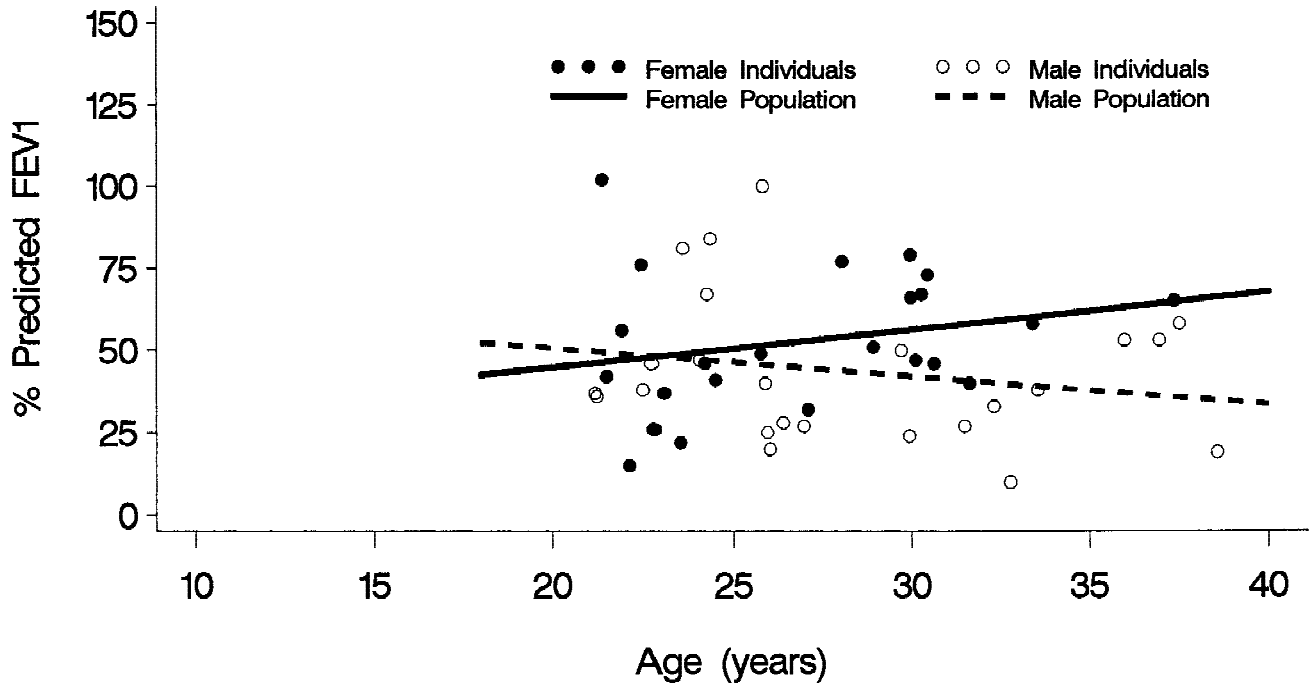

Fig. 4. Cross-sectional analysis. Regression lines and scatter plot by gender in 47 subjects.

TABLE 4—Cross-Sectional Regression Model for FEV_1, Percent Predicted, Gender and Age Included as Covariates, Intercept Estimated at Age = 25 Years

Parameter	Estimate	SE	*P*-value
Intercept	50	4.9	0.0001
Gender	−4	7.2	0.5590
Age	1.1	1.03	0.2763
Gender* age	−2.0	1.32	0.1392
σ_2	460		

ments for an individual. The National Cystic Fibrosis Patient Registry[6] provides an excellent source of data for obtaining initial estimates needed in computing statistical power for longitudinal study designs in CF research.

Since general statistical techniques for the analysis of complex longitudinal data have been lacking for the CF researcher, computation of statistical power and sample size have also been lacking. Appropriate methods for computing statistical power for the classic general linear multivariate model (repeated measures design with no missing observations) are discussed by Muller et al.[47] Muller et al.[47] also provide free computer software (using the Interactive Matrix Language software in SAS[48]), with very good documentation for computing statistical power for the classic general linear multivariate model.

For appropriate statistical power computations in the mixed model, Helms[1] discusses using the Helms-McCarroll approximate F statistic, denoted F_H. The approximate power is expressed as a function of the fixed effect regression parameters, β, the random effect variance matrix, D, the within-subject error variance, σ^2, and both fixed effect and random effect design matrices, X and Z.

Example of Power Analysis

Suppose CF researchers wanted to propose a controlled clinical trial aimed at alleviating the rate of decline in percent predicted FEV_1, using the group of adult CF subjects discussed in the previous section. Consider for the moment the estimation results in Table 2 (cross-sectional analysis). Since the effect size (slope with respect to age) is small and the variance is large, a larger number of CF subjects would be required to have adequate power to detect a clinically meaningful effect in a proposed cross-sectional design. The power we have to detect the observed slope of −0.2 (%/year) using a two-sided *t*-test with level of significance $\alpha = 0.05$ is approximately 0.5. The sample size needed to detect the observed slope with a power of 0.8, $\alpha = 0.05$, is approximately 100 subjects. The approximate sample size needed for a power of 0.9 is approximately 130. Hence, in the planning of a controlled clinical trial using cross-sectional analyses, we would have to double or nearly triple our sample size simply to have reasonable power to detect the observed slope of −0.2 (%/year).

Now consider the estimation results in Table 3 (longitudinal analysis with the mixed model). In a proposed longitudinal study design where the mixed model is used, since the effect size (slope with respect to age) is large and the variance is small, improved statistical power is

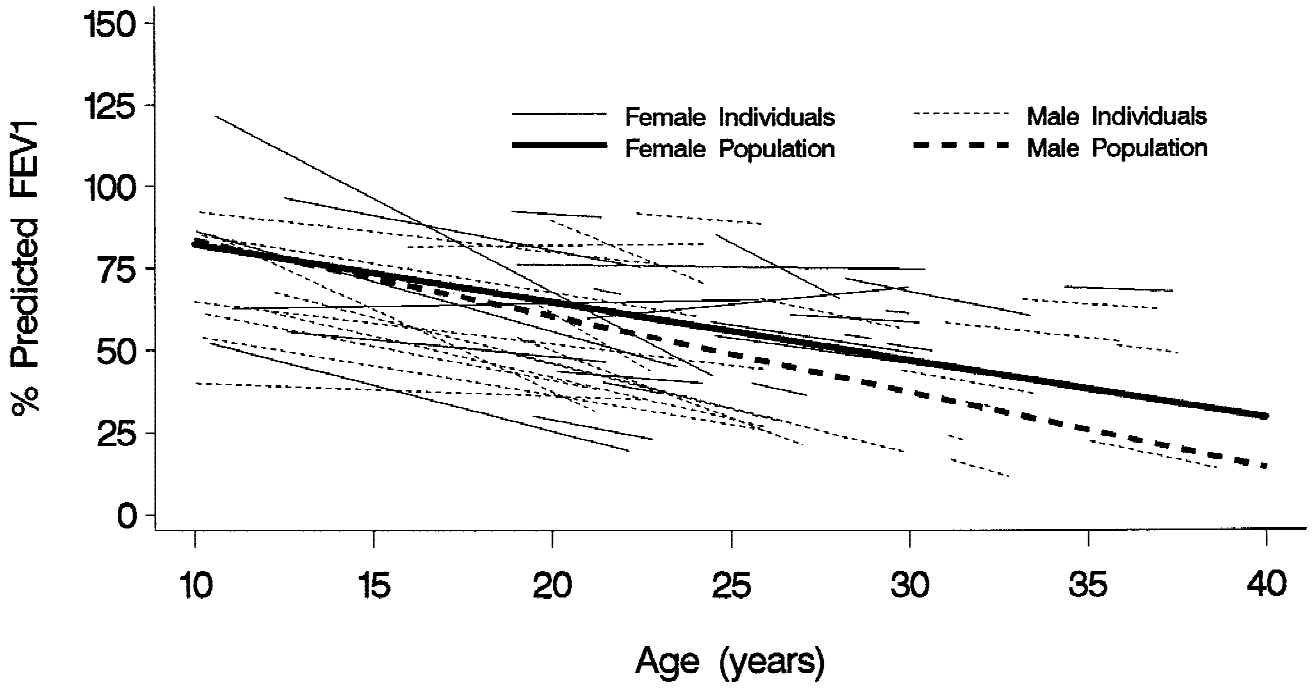

Fig. 5. Mixed model analysis. Population and individual regression lines by gender, 47 subjects with cystic fibrosis and 1,401 observations.

TABLE 5—Mixed Model Regression Model for FEV_1, Percent Predicted, Gender and Age Included as Covariates, Intercept Estimated at Age = 20 Years

Parameter	Estimate	SE	*P*-value
Intercept	65	4.9	0.0001
Gender	−4	7.0	0.5290
Age	−1.8	0.53	0.0017
Gender* age	−0.5	0.71	0.4361
σ^2	108		

achieved. The power we have to detect the observed slope of −2 (%/year) using a one-sided F-test with level of significance $\alpha = 0.05$ is approximately 0.99 (see Helms[1] for more details). Presently, computing sample size using the mixed model is not fully developed and will be omitted here.

It is clear that for this cystic fibrosis example, we have more statistical power, and estimation results are more useful using longitudinal rather than cross-sectional methods. In a study with human subjects it is often less expensive to measure each subject several times than to get an equal number of single measurements from a greater number of subjects.[1,49] Hence, we might expect greater cost savings in using longitudinal methods in the design of controlled clinical trials in CF research based on these results.

Finally, it is noted that both the cross-sectional and longitudinal methods demonstrate large power (>0.95) for detecting the observed differences in rates of decline between males and females.

CONCLUSIONS

This paper has presented an overview of some of the fundamental concepts involved in the analysis of complex longitudinal data. Two of the most commonly used statistical methods for the analysis of longitudinal data were highlighted: the *general linear mixed model* (mixed model) and the method of *generalized estimating equations* (GEEs). An example using pulmonary function data in cystic fibrosis research was used to illustrate the application of the mixed model.

The mixed model and GEE are both very advanced and complex statistical techniques. A comparison of the mixed model and GEE can be found in Park.[50] Though a detailed discussion of this comparison is beyond the scope of this paper, Park[50] demonstrates that results from GEE can differ from that of the mixed model when there are missing observations and/or the covariance matrix is structured. Also, both the mixed model and GEE are lacking in accessible techniques for performing assessment of goodness-of-fit, model assumptions such as normality (in the case of the mixed model), and other regression diagnostics. However, active research in the statistical literature provides encouragement that these limitations will be remedied in the near future.

The example given in this paper involved using simple linear models, and therefore the full complexity of the procedures was understated for the sake of clarity and simplicity. It is important that the reader who is inexperienced and/or not trained in the use of advanced longi-

tudinal statistical methods enlists the services of a trained statistician in attempting to apply the mixed model and/or GEE to complex longitudinal data.

In the near future, longitudinal data analysis techniques will be used to reevaluate some of the most fundamental assumptions about the relationship between correlated outcomes and predictor variables of interest, including treatment modalities for clinical trial participants. In addition, longitudinal data analysis techniques will be used to reevaluate the estimation of parameters used in sample size and statistical power determinations, with obvious implications for the future design of many observational studies and clinical trials.

At present, it is clear that in order to develop more effective and more powerful observational studies and controlled clinical trials, longitudinal statistical methods should be used more often. The examples described in this paper illustrate how a cross-sectional analysis of pulmonary function outcomes obtained from CF subjects should be considered inadequate, when the study design and/or data collection are longitudinal.

Often, there is a lag between the development of advanced statistical techniques and their widespread use. Such has been the case in the development and application of advanced longitudinal statistical techniques such as the mixed model and GEE. Now that the statistical methodology exists and the computer software is readily available to accommodate longitudinal designs of observational studies and controlled clinical trials, it is important that these methods are employed when appropriate. Researchers and practitioners are encouraged to exploit the advances in general statistical methods for the analysis of complex longitudinal outcomes in designing more efficient and more powerful observational studies and controlled clinical trials.

REFERENCES

1. Helms RW. Intentionally incomplete longitudinal designs: I. Methodology and comparison of some full span designs. Stat Med 1992;11:1889–1993.
2. Zeger SL, Liang K-L. An overview of methods for the analysis of longitudinal data. Stat Med 1992;11:1825–1839.
3. Dwyer JH, Feinleib M, Lippert P, Hoffmeister H. Statistical models for longitudinal studies of health. New York: Oxford University Press; 1991. 383 p.
4. Duncan GJ, Kalton G. Issues of design and analysis of surveys across time. Int Stat Rev 1987;55:97–117.
5. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied regression analysis and multivariable methods, 3rd ed. Pacific Grove: Doxbury Press. 1998. 787 p.
6. Cystic Fibrosis Foundation. United States national cystic fibrosis patient registry. Bethesda, MD: Cystic Fibrosis Foundation National Office.
7. FitzSimmons SC. The changing epidemiology of cystic fibrosis. J Pediatr 1993;122:1–9.
8. Harville DA. Extension of the Gauss-Markov theorem to include the estimation of random effects. Ann Stat 1976;4:384–395.
9. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–974.
10. Laird NM, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the EM algorithm. J Am Stat Assoc 1987;82:97–105.
11. Diggle PJ, Liang K-L, Zeger SL. Analysis of longitudinal data. Oxford: Oxford University Press; 1994. 247 p.
12. Grady JJ, Helms RW. Model selection techniques for the covariance matrix for incomplete longitudinal data. Stat Med 1995;14: 1397–1416.
13. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley; 1987. 278 p.
14. Woolson R, Clarke W, Leeper JD. Missing data in longitudinal studies. In: Dwyer JH, Feinleib M, Lippert P, Hoffmeister H, editors. Statistical models for longitudinal studies of health. New York: Oxford University Press; 1991. p 207–300.
15. Wald A. Tests of statistical hypothesis concerning several parameters when the number of observations is large. Trans Am Math Soc 1943;54:426–482.
16. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc [A] 1972;135:370–384.
17. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman & Hall; 1989. 511 p.
18. Rosner B, Munoz A. Conditional linear models for longitudinal data. In: Dwyer JH, Feinleib M, Lippert P, Hoffmeister H, editors. Statistical models for longitudinal studies of health. New York: Oxford University Press; 1991. p 115–131.
19. Rosner B, Munoz A, Tager I, Speizer FE, Weiss ST. Use of a generalized autoregressive model for the analysis of longitudinal data. Stat Med 1985;4:457–467.
20. Harville DA. Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 1977; 72:320–338.
21. Woolson RF, Leeper JD, Clarke WR. Analysis of incomplete data from longitudinal and mixed longitudinal studies. J R Stat Soc [A] 1978;141:242–252.
22. Andrade DA, Helms RW. ML estimation and LR tests for the multivariate normal distribution with general linear model mean and linear-structure covariance matrix: k-population, complete data case. Commun Stat Theor Methods 1986;13:89–108.
23. Andrade DA, Helms RW. ML estimation for the multivariate normal distribution with general linear model mean and linear-structure covariance matrix: one population, complete data case. Commun Stat Theor Methods 1986;15:1927–1955.
24. Fairclough DI, Helms RW. A mixed linear model with linear covariance structure: a sensitivity analysis of the maximum likelihood estimators. J Stat Comput Simul 1986;25:205–236.
25. Jeske DR, Harville DA. Prediction-interval procedures and (fixed-effects) confidence-interval procedures for mixed linear models. Commun Stat Theor Methods 1988;17:1053–1087.
26. Cressie N, Lahiri SN. The asymptotic distribution of REML estimators. Unpublished manuscript (preprint no. 91-20). Ames, IA: Department of Statistics, Iowa State University; 1991.
27. Littell RC, Milliken GA, Stroup WW, Wolfinger RD. SAS® system for mixed models. Cary, NC: SAS Institute, Inc.; 1996. 633 p.
28. Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. J Am Stat Assoc 1987;83:1014–1022.
29. SAS Institute, Inc. The MIXED procedure, in SAS/STAT software, changes and enhancements through release 6.12. Cary, NC: SAS Institute, Inc.; 1997. p 571–703.

30. Zeger SL, Liang K-Y, Self SG. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44: 1049–1060.
31. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033–1048.
32. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
33. SAS Institute, Inc. The GENMOD procedure, in SAS/STAT software, changes and enhancements through release 6.12. Cary, NC: SAS Institute, Inc.; 1997. p 249–347.
34. Kerem E, Corey M, Kerem B, Rommens J, Markiewicz D, Levison H, Tsui L, Durie P. The relation between genotype and phenotype in cystic fibrosis—analysis of the most common mutation (ΔF_{508}). N Engl J Med 1990;323:1517–1522.
35. Knowles MR, Church NL, Waltner WE, Yankaskas JR, Gilligan P, King M, Edwards LJ, Helms RW, Boucher RC. A pilot study of aerosolized amiloride for the treatment of lung disease in cystic fibrosis. N Engl J Med 1990;322:1189–1194.
36. Johansen H, Nir M, Hoiby N, Koch C, Schwartz M. Severity of cystic fibrosis in patients homozygous and heterozygous for the ΔF508 mutation. Lancet 1991;337:631–634.
37. Glindmeyer HW, Diem JE, Jones RN, Weill H. Noncomparability of longitudinally and cross-sectionally determined annual change in spirometry. Am Rev Respir Dis 1982;125:544–548.
38. Burrows B, Lebowitz MD, Camilli AE, Knudson RJ. Longitudinal changes in forced expiratory volume in one second in adults. Am Rev Respir Dis 1986;133:974–980.
39. Dontas AS, Jacobs DR, Corcondilas A, Keys A, Hannan P. Longitudinal versus cross-sectional vital capacity changes and affecting factors. J Gerontol 1984;39:430–438.
40. Vollmer WM, Johnson LR, McCamant LE, Buist AS. Methodologic issues in the analysis of lung function data. J Chronic Dis 1987;40:1013–1023.
41. Pattishall EN, Helms RW, Strope GL. Noncomparability of cross-sectional and longitudinal estimates of lung growth in children. Pediatr Pulmonol 1989;7:22–28.
42. van Pelt W, Borsboom GJJM, Rijcken B, Schouten JP, van Zomeren BC, Quanjer PH. Discrepancies between longitudinal and cross-sectional change in ventilatory function in 12 years of follow-up. Am J Respir Crit Care Med 1994;149:1218–1226.
43. Knowles MR, Hohneker KW, Zhou Z, Olsen JC, Noah TL, Hu P-C, Leigh MW, Engelhardt JF, Edwards LJ, Jones KR, Grossman M, Wilson JM, Johnson LG, Boucher RC. A controlled study of adenoviral-vector-mediated gene transfer in the nasal epithelium of patients with cystic fibrosis. N Engl J Med 1995;333:823–831.
44. Konstan MW, Byard PT, Hoppel CL, Davis PB. Effect of high-dose ibuprofen in patients with CF. N Engl J Med 1995;332:848–854.
45. Corey M, Edwards LJ, Levison H, Knowles M. Longitudinal analysis of pulmonary function decline in patients with cystic fibrosis. J Pediatr 1997;131:809–814.
46. Bethesda, MD: Cystic Fibrosis Foundation National Office.
47. Muller KE, LaVange LM, Ramey SL, Ramey CT. Power calculations for general linear multivariate models including repeated measures applications. J Am Stat Assoc 1992;87:1209–1226.
48. SAS Institute, Inc. SAS/IML software: usage and reference, version 6, first edition. Cary, NC: SAS Institute, Inc.; 1989. 501 p.
49. Helms RW. Longitudinal designs and their statistical analysis in pediatric pulmonary research. Pediatr Pulmonol 1990;9:69–71.
50. Park T. A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. Stat Med 1993;12:1723–1732.

APPENDIX: EXAMPLE OF MIXED MODELS FOR THE ANALYSIS OF PERCENT PREDICTED FEV_1 IN CF

Example 1

We first discussed a simple linear mixed model which is used in the example of data analysis in this paper. We then discussed a mixed model which can be used to test the difference between two treatments.

Suppose the CF researcher is interested in using the mixed model to determine the linear rate of decline (with respect to age) in percent predicted FEV_1 for a group of CF subjects. We will use a simple linear mixed model where the fixed population parameter vector, β, has two elements: an intercept and slope. In addition, we will assume each subject has their own intercept and slope, d_i, i.e., there are two random effects for each subject: a random intercept and a random slope.

The mixed model equation for the i-th subject may be written as:

$$\%\text{PredFEV}_{1i} = \beta_0 1_i + \beta_1 \text{AGE}_i + d_{0i} 1_i + d_{1i} \text{AGE}_i + e_i,\ i = 1, \ldots, k,$$

where $\%\text{predFEV}_{1i}$ is an $n_i \times 1$ vector of n_i measures of percent predicted FEV_1 over time on the i-th CF subject. For most scenarios in CF research, n_i will vary from subject to subject. The mixed model formulation accommodates the differing numbers of observations per subject. β_0 and β_1 are the fixed population parameters (scalars) representing the population intercept and slope (β is a 2×1 vector); 1_i is an $n_i \times 1$ vector of 1s. AGE_i is an $n_i \times 1$ vector of ages for the i-th CF subject. Hence, X_i is an $n_i \times 2$ fixed effect design matrix with two columns: a column of 1s and a column of ages. d_{0i} and d_{1i} are the random intercept and slope parameters specific to the i-th CF subject (d_i is a 2×1 vector). Z_i is an $n_i \times 2$ random effect design matrix for the i-th CF subject and is identical to X_i. e_i is an $n_i \times 1$ vector of random error terms for the i-th CF subject.

Rejection of the fixed-effect null hypothesis H_0: $\beta_1 = 0$ indicates a statistically significant rate of decline (slope) in FEV_1 percent predicted. Similarly, rejection of the fixed-effect null hypothesis H_0: $\beta_0 = 0$ indicates a statistically significant intercept.

In the mixed model, the structure of the random effects covariance matrix, D, and the covariance matrix $\text{Var}(Y_i)$ can be defined or modeled in several ways. The concept of actually modeling the covariance matrices adds to the complexity of using the mixed model. A complete discussion on modeling the covariance matrices is beyond the scope of this paper, but the interested reader is re-

ferred to Grady and Helms.[12] When in doubt, a general rule of thumb (when practical) is to use unstructured covariance matrices for D. An inspection of the estimated unstructured covariance matrices may be used to determine whether patterns exists.

Example 2

Often the CF researcher is interested in the comparison of two groups, such as the comparison of males to females or the comparison of active treatment to placebo. Let us assume that the CF researcher wishes to compare males and females. Define the dummy variable GENDER = 1 if the subject is male; GENDER = 0 if female. A mixed model equation which may be used to test for gender differences can be written as:

$$\%\text{PredFEV}_{1i} = \beta_0 1_i + \beta_1 \text{GENDER}_i + \beta_2 \text{AGE}_i + \beta_3 \text{GENDER}_i * \text{AGE}_i + d_{0i} 1_i + d_{1i} \text{AGE}_i + e_i,\ i = 1, \ldots, k,$$

where GENDER_i is the $n_i \times 1$ vector of values indicating gender (this is either all 0s or all 1s for the i-th subject); and $\text{GENDER}_i * \text{AGE}_i$ represents the interaction of gender and age (multiplication of the dummy variable GENDER by AGE).

In this case, rejection of the fixed-effect null hypothesis H_0: $\beta_3 = 0$ indicates a statistically significant difference in the rates of decline (slopes) in FEV_1 percent predicted between males and females. Similarly, rejection of the fixed-effect null hypothesis H_0: $\beta_1 = 0$ indicates a statistically significant difference in the levels (intercepts) of FEV_1 percent predicted between males and females.