Annals of Epidemiology 23 (2013) 771–777

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

A semiparametric approach to estimate rapid lung function decline in cystic fibrosis

Rhonda D. Szczesniak PhD [a,b,*], Gary L. McPhail MD [b], Leo L. Duan MS [c], Maurizio Macaluso MD, DrPH [a], Raouf S. Amin MD [b], John P. Clancy MD [b]

[a] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH
[b] Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH
[c] Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH

ARTICLE INFO

Article history:
Received 1 April 2013
Accepted 30 August 2013
Available online 5 October 2013

Keywords:
Cystic fibrosis
Nonlinear models
Epidemiologic methods
Spline regression

ABSTRACT

Purpose: Detecting the onset of rapid lung function decline is important to reduce mortality rates in cystic fibrosis (CF) and other lung diseases. The most common approach is conventional linear mixed modeling—estimating a population-level slope of lung function decline and using random effects to address serial correlation—but this ignores nonlinear features of disease progression and distinct sources of variability. The purpose of this article was to estimate patient-specific timing and degree of rapid decline while appropriately characterizing natural progression and variation in CF.
Methods: We propose longitudinal semiparametric mixed modeling and contrast it with the conventional approach, which restricts lung function (measured as forced expiratory volume in 1 second as a percentage of predicted, FEV_1%) to linear decline. Each approach is applied to clinical encounter data from the United States CF Foundation Patient Registry.
Results: Timing and degree of rapid FEV_1% decline vary across patients and as a function of key covariates. Patients experience maximal FEV_1% loss by early adulthood more severe than indicated by conventional slope analysis.
Conclusions: Semiparametric mixed modeling provides a means to estimate patient-specific changes in CF disease progression and may be used to inform prognostic decisions in chronic care settings and clinical studies.

Introduction

Cystic fibrosis (CF) is a common lethal genetic disease that affects more than 70,000 people worldwide [1]. CF is marked by progressive loss of lung function with eventual respiratory failure [2] and has no cure; it often progresses more rapidly during adolescence and early adulthood [3,4] with median survival only into the late 30s [1]. Maintaining lung function over time is essential for survival in CF [5,6]. The percent predicted forced expiratory volume in 1 second (FEV_1%) is the accepted measure of CF lung function [7]. FEV_1% is a key marker in determining disease severity and directing clinical care for each patient [7,8], a primary outcome in therapeutic efficacy studies, and a decisive factor for selecting lung transplant candidates [7].

Across the life span, lung function may decline more rapidly or demonstrate increased variability for many reasons (Fig. 1), ranging from natural disease progression to intermittent infection [3,9]. Well-maintained patient registries have made it possible to characterize lung function decline over the lifetimes of thousands of patients [10–13]. Conventional statistical analyses have focused on the association of disease severity, treatment, and other factors with decline in FEV_1% [9,14–17]. As a result, important epidemiologic findings have emerged and aided our understanding of lung function decline with age; however, few CF epidemiologic studies have used advanced statistical models to characterize lung function changes. A study of the Danish CF Patient Registry found marked improvements in lung function prediction by adding longitudinally structured correlation to account for repeated FEV_1% measurements on patients [18]. The study analysis, however, assumed that population-level lung function declines at the same rate, regardless of age. In contrast, a previous study involving healthy children indicates that lung function changes are not linear over time [19]. Furthermore, longitudinal cohort studies of U.S. CF patients have shown that FEV_1% slopes for adolescents and young adults are

* Corresponding author. Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave. MLC 5041, Cincinnati, OH 45229.Tel.: +1 5138030563; fax: +1 5136367509.
E-mail address: rhonda.szczesniak@cchmc.org (R.D. Szczesniak).

1047-2797/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.annepidem.2013.08.009

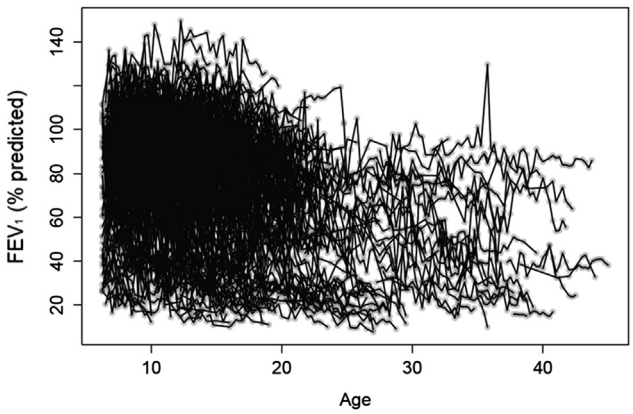

Fig. 1. Individual trajectories of decline in FEV_1% over age (in years) for 350 randomly selected patients from the U.S. CFFPR, 1997–2011.

different, suggesting that lung function change in CF is not linear over time [4,20].

To date, the most complex model describing FEV_1% decline has been a quadratic curve [17]. Although such parametric models detect global changes in lung function, they fail to capture the potentially complex form of patient-specific FEV_1% decline [21,22]. Like many time series, it is sensible to assume that correlations between repeated FEV_1% measurements are smaller for observations that are further apart in time. Several structures are available to model decaying correlation, as long as the measurements are equally spaced [23]; however, most chronic disease registries pose the additional challenge that encounter-level data captured for each patient is irregularly spaced, complicating the task of modeling the correlation of repeated measures.

Semiparametric regression [24], an alternative to conventional regression, provides a more flexible mean structure to analyze longitudinal data through penalized regression splines and mixed modeling. This representation provides smooth estimates of nonlinear mean functions and derivatives while preventing overfitting. Our objective was to develop a flexible and accurate model for predicting rapid CF lung function decline. We exploited new statistical methods and their adaptation for longitudinal data analysis, overcoming the limitations of conventional FEV_1% progression models.

Methods

Study data

Data were acquired from the U.S. Cystic Fibrosis Foundation Patient Registry (CFFPR), which has been tracking outcomes for more than 40 years, and contain demographic and clinical information from all accredited CF centers in the United States. We excluded patients younger than 6 years because such patients tend to have unreliable pulmonary function testing and patients aged older than 45 years because they tend to be subjects with milder phenotypes that are unrepresentative of typical CF disease progression [20]. Data acquired after organ transplantation were excluded. Because many relevant predictors of FEV_1% decline were consistently documented beginning in 1997, we considered data from 1997 to 2011. The FEV_1% measure obtained for analysis was adjusted for gender and height using standard reference equations [25,26]. Predictors of FEV_1% decline, available in the database and identified from previous literature [9,17,20], included age, gender, birth cohort, and first recorded (baseline) FEV_1%; time-varying predictors included CF-related diabetes (with or without fasting hyperglycemia), positive cultures for methicillin-resistant *Staphylococcus aureus* (*MRSA*), *Pseudomonas aeruginosa* (*Pa*), *Burkholderia cepacia* (*B. cepacia*), and socioeconomic status (SES) (having only state/federal or no insurance was used as a proxy). Covariate interactions with age were examined in all models. To include time-varying covariates that correspond to FEV_1% measurements, maximum quarterly values of each covariate were used for analysis. The relationship between maximum quarterly FEV_1% and each potential covariate was first assessed using scatter plot smoothing [27]. Exploratory graphical analyses, model validation, and sensitivity analyses are available in Supplementary data.

Conventional model

The most frequently applied parametric linear mixed effects model when assessing FEV_1% decline in CF cohorts uses subject-specific random effects to characterize the between- and within-subject variations previously described. Let Y_{ij} be the FEV_1% value for the i^{th} patient observed at time t_{ij}, $i = 1,\ldots I$, and $j = 1,\ldots,n_i$. The conventional linear mixed effects model can be expressed as

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \sum_{k=1}^{p} \theta_k x_{ijk} + \gamma_{0i} + \varepsilon_{ij} \quad (1)$$

The population-level mean response is modeled as a linear function of time, t_{ij}, using intercept and slope coefficients β_0 and β_1, respectively. The terms x_{ijk}, $k = 1,\ldots, p$, are measurements of p covariates assumed to enter the model linearly, and each has corresponding parameter θ_k (where covariates may be time-varying). Model (1) accounts for longitudinal correlation using a subject-specific random intercept term, γ_{0i}, and measurement error, ε_{ij}.

Semiparametric mixed effects model

We used penalized regression splines with a cubic truncated power basis [28] to provide smooth estimates of the longitudinal course of FEV_1%. The combination of the fixed effects terms and penalized splines provided a continuous function for estimating the FEV_1% trend over time (expressed as age, in years):

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{l=1}^{L} b_l (t - \kappa_l)_+^3,$$

where β_0, β_1, β_2, and β_3 are the traditional coefficients for intercept, linear, quadratic, and cubic terms, respectively. The expression $\sum_{l=1}^{L} b_l (t - \kappa_l)_+^3$ represents basis functions for the penalized cubic splines evaluated at time t. Number and location of knots, k_l ($l = 1,\ldots,L$), for the corresponding spline basis functions may be chosen using fit statistics. We then incorporate $f(t)$ into a linear mixed effects model. The resulting semiparametric mixed effects model is

$$Y_{ij} = f(t_{ij}) + \sum_{k=1}^{p} \theta_k x_{ijk} + \gamma_{0i} + \varepsilon_i(t_{ij}), \quad (2)$$

where $\varepsilon_i(t_{ij})$ reflects stochastic variation about the mean FEV_1% response. We characterize within-patient variation over time into: an exponential correlation function and include the aforementioned measurement error from model (1). Further details on model, covariance structures, and knot selection are included in Supplementary data. This approach combines the semiparametric model with a more appropriate longitudinal structure for time series error.

Model evaluation and fit

We estimated overall trends in FEV_1% and the first derivative with respect to time for each model's mean response function, holding all