Published in final edited form as:
Biometrics. 2010 June ; 66(2): 594–602. doi:10.1111/j.1541-0420.2009.01288.x.

Uncovering Symptom Progression History from Disease Registry Data with Application to Young Cystic Fibrosis Patients

Jun Yan[1], **Yu Cheng**[2], **Jason P. Fine**[3], and **HuiChuan J. Lai**[4]

[1]Department of Statistics and Institute for Public Health Research, University of Connecticut, Storrs, CT U.S.A.

[2]Department of Statistics and Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA U.S.A.

[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC U.S.A.

[4]Department of Nutritional Sciences, Department of Biostatistics and Medical Informatics, and Department of Pediatrics, University of Wisconsin, Madison, WI U.S.A.

Summary

The growing availability of various disease registry data has brought precious opportunities to epidemiologists to understand the natural history of the registered diseases. It also presents challenges to the traditional data analysis techniques due to complicated censoring/truncation schemes and temporal dynamics of covariate influences. In a case study of the Cystic Fibrosis Foundation Patient Registry data, we propose analyses of progressive symptoms using temporal process regressions, as an alternative to the commonly employed proportional hazards models. Two end points are considered, the prevalence of ever positive and currently positive for Pseudomonas aeruginosa (PA) infection in the lungs, which capture different aspect of the disease process. The analysis of ever PA positive via a time-varying coefficient model demonstrates the lack of fit, as well as the potential loss of information, in the standard proportional hazards analysis. The analysis of currently PA positive yields results which are clinically meaningful and have not previously been reported in the cystic fibrosis literature. Our analyses demonstrate that prenatal/neonatal screening results in lower prevalence of PA infection compared to traditional diagnosis via signs and symptoms, but this benefit attenuates with age. Calendar years of diagnosis also affect the risk of PA infection; patients diagnosed in more recent cohort show higher prevalence of ever PA positive but lower prevalence of currently PA positive.

Keywords

Generalized linear model; Lung infection; Newborn screening and diagnosis; Prevalence; Pseudomonas aeruginosa; Varying-coefficient

1. Introduction

Disease registries for progressive chronic illnesses are unique resources for evaluating the natural history of the diseases at the population level. Such population based registries have existed in Europe for much of the 20th century, with recent emergence in North America oriented towards study of the life history of chronic diseases. Statistical analyses of registry data aid clinical researchers searching for patterns of progressive symptoms, which might lead to better treatments and patient management. Nevertheless, due to lack of appropriate methods and unawareness of key statistical issues, standard techniques, which may be inappropriate or may waste valuable information, are often employed. We investigate some

of the issues in a case study of the Cystic Fibrosis Foundation Patient Registry (CFFPR) data maintained by the Cystic Fibrosis Foundation (CFF).

Cystic Fibrosis (CF) is one of the most common lethal inherited disorders in Caucasians, affecting an estimated 30,000 people in the US (Cystic Fibrosis Foundation, 2008). Chronic lung infections and obstructive lung diseases, eventually leading to respiratory failure, are the primary causes (about 80%) of death in patients with CF (Cystic Fibrosis Foundation, 2008). Pseudomonas aeruginosa (PA), a ubiquitous environmental bacterium, is the most important pathogen that accelerates lung disease and shortens survival of CF patients (Aebi et al., 1995; Liou et al., 2001). Because PA infections can be transient/intermittent, and can be eradicated by treatment with antibiotics, PA culture results often switch between positive and negative states at different visits within a patient. Both the timing of the initial infection, which we denote ever PA positive, and the extent to which infection persists over time, which we denote current PA positive, are important in treatment decisions. Characterizing PA infection patterns is particularly important for pediatric CF patients in their first decade of life because PA infections begin early in life (Cystic Fibrosis Foundation, 2008) and are very difficult to eradicate once they become chronic (Treggiari et al., 2007).

The prevalence of a progressive symptom, whether transient, intermittent, or persistent, provides information for planning health services, allocating health resources, and assessing the relative burden of the disease on mortality and quality of life. The prevalence denotes the proportion of patients in the surviving population either having or having had a symptom. Although ample discussion on statistical analyses of prevalence exists in the literature and some has been based on registry data (e.g., Gail et al., 1999; Verdecchia et al., 2002), there are issues shared by many registry data that are not appropriately addressed by standard techniques commonly employed in the existing analyses. When analyzing the CFFPR data, we address these issues in the context of the prevalence of two symptom measures, ever having PA infection and currently having PA infection, assessing both cumulative and local disease burdens at different ages.

The special features of the CFFPR data, such as nonstandard double censoring schemes and time-varying covariate effects (see Section 2 for more details), require appropriate statistical methods. Prevalence analyses of both ever PA positive and currently PA positive can be situated in a unified framework of the temporal process regression (Fine et al., 2004), which models the prevalence as the mean of a stochastic process of symptom indicator. Compared to the standard survival analysis techniques, this framework is valid under milder assumptions and provides further insights into the covariate effects, owing to the unspecified time-varying covariate coefficients. The proposed model for the prevalence of ever PA infection accommodates double censoring where left-censored cases are typically excluded in standard techniques. It also enables a test of the proportional hazards assumption in analyzing the onset age of first PA infection, which, as will be illustrated, is not appropriate.

Our results generate several findings regarding PA infections in young children with CF. The prevalence of current PA infection is significantly lower, and the onset age of first PA infection is significantly older, in children diagnosed early through prenatal/neonatal screening (SCR) than in those diagnosed later through signs and symptoms of CF. The beneficial effect of SCR on PA infection, however, attenuates with age, and is most evident during the first several years of life. Children diagnosed more recently (1994-2000), when compared to those diagnosed during 1986-1990, acquire first PA infection at earlier ages (i.e., higher prevalence of ever PA positive) but have lower PA positive rates at older/current ages (i.e., lower prevalence of currently PA positive). Further, both differences evolve with age.

The rest of the paper is organized as follows. Section 2 describes the CFFPR data in detail and presents a Cox model with time-varying coefficients for the onset age at first PA positive. Section 3 sketches the temporal process regression method with emphasis in the context of the CFFPR data. Section 4 presents analyses on the two prevalence models, ever PA positive and currently PA positive, the first of which is compared to the time-varying coefficient Cox model. The two PA prevalence models are also contrasted. A discussion concludes in Section 5.

2. CFFPR Data and Preliminary Analysis

The CFFPR is a major data source for epidemiological studies on CF (e.g., Lai et al., 2004; van den Akker-van Marle et al., 2006; Comeau et al., 2007). Currently, data on more than 24,000 people who receive care at CFF-accredited CF Centers are reported to the CFFPR every year (Cystic Fibrosis Foundation, 2008). This paper uses the CFFPR reported during 1986-2005 to investigate several risk factors that influence PA infections in CF patients younger than 10 years of age, a period with greatest potential to benefit from early diagnosis and new therapies and hence, improved health outcomes (Campbell and White, 2005). In addition, to capture a complete follow-up for PA infection after the diagnosis of CF, only patients diagnosed after 1986 were included in the analysis (Lai et al., 2004).

Three risk factors are considered based on their associations with survival and lung disease outcomes demonstrated from previous epidemiological studies (Rosenfeld et al., 1997; Liou et al., 2001; Assael et al., 2002; Lai et al., 2004, 2005): gender, method of CF diagnosis, and calendar year of diagnosis. Methods of CF diagnosis, classified according to common clinical practices that lead to the identification of CF, include four diagnostic categories: patients identified at birth because of an intestinal obstruction known as meconium ileus (MI), patients identified shortly after birth via prenatal/neonatal screening (SCR), patients identified at variable ages because of positive family history (FH) without symptoms, and patients identified at variable ages because of symptoms (SYMP) other than MI. Calendar year of diagnosis is classified into 3 cohorts, 1986-1989, 1990-1993, and 1994-2000, denoted as DX [86-89], DX [90-93], and DX [94-00], respectively; the cutoff years are chosen to reflect medical advances in the diagnosis and treatment of CF, i.e., discovery of the most common gene mutation in 1989 and FDA approval of pulmozyme in 1994.

The onset ages of first PA infection in the CFFPR are characterized with various truncation and censoring schemes. PA infections are evaluated during clinic visits to the CF Centers, which typically occur every 3-6 months. Therefore, the ages at which PA tests are positive or negative are both interval-censored. As is standard in analyses of most symptom data, PA result from the previous visit is carried forward until the next visit. Therefore, the observed indicator process of PA infection jumps only at visit times.

Right and left censoring, as well as left truncation on mortality, are also present in the PA infection data. These patterns are different than those for the usual left truncated right censored data. The left and right censoring times are always observed and the left truncation of infection is indirect, via mortality. The presence of right censoring is obvious, as a patient may not have acquired any PA infection by the age at the end of 2005, or by the age of death. The effect of right censoring at death is small in the analysis for age under 10, because the mortality rate is very low prior to age 10 (Grosse et al., 2006). Left censoring occurs when a patient enters the CFFPR with a positive PA infection, either due to very early acquisition (e.g., positive at time of diagnosis) or delayed entry into the registry. The left truncation occurs due to the delayed diagnosis (thus delayed entry into the CFFPR) and mortality prior to diagnosis. The mortality truncation is universal in registry data because of

the intrinsic nature of data collection. The models proposed in Section 3 deal explicitly with these unique features of the data.

The study population in this paper included 12,822 patients diagnosed between 1986 and 2000, the same population investigated by Lai et al. (2004), but with 5 more years of follow-up data, i.e., 1986-2005. Of those patients, the onset ages of first PA infection of 3,348 were left-censored, 2,426 were right-censored and 413 deaths occurred. Zooming in to the first decade of life, 11,375 patients entered the registry prior to age 10, among which, 2,735 were left-censored, 2,284 were right-censored, and only 21 died prior to age 10. To avoid assuming proportional hazards as in the earlier analysis of Lai et al. (2004), a Cox model with time-varying coefficients (Martinussen and Scheike, 2006) was fitted to examine the association between the three risk factors and the onset age of first PA infection for age period before 10. Note that, as in Lai et al. (2004), the left-censored patients were not included in the fitting.

Figure 1 presents the estimated cumulative coefficients and the 95% pointwise confidence intervals as well as the Hall-Wellner confidence bands for ages before 10. The results were obtained using the companion R package timereg of Martinussen and Scheike (2006), with a start age 0.5 and max age 10. The DX (SCR), male, and DX [86-89] are used as the reference levels for the method of diagnosis, gender, and calendar year of diagnosis, respectively. These results are qualitatively consistent with those of Lai et al. (2004): MI and SYMP patients have greater risks of acquiring PA infection compared with SCR patients; females are at greater risk than males; and patients diagnosed after 1994 have greater risks of acquiring PA infection than those diagnosed before 1989. Figure 1 also suggests that the proportional hazards assumption may not be valid since the cumulative coefficients for DX (SYMP), DX (MI), and DX [94-00] do not seem close to straight lines. Unfortunately, although the timereg package allows for a resampling based significance test and a goodness-of-fit test for each coefficient, these numerical results could not be obtained for the CFFPR data due to the computational burden with large sample sizes and large number of events.

In addition to the less intuitive interpretation of cumulative coefficients, the results in Figure 1 are further limited as approximately a quarter of the data are wasted due to the exclusion of left-censored observations. The limitation can be overcome by modeling prevalence instead of hazard, of ever PA positive using the temporal process regression method with an appropriate link function. The temporal process regression method can also be used to model the prevalence of currently PA positive, which is particularly relevant to the burden of present infection and can not be modeled by the standard and time-varying coefficients Cox models. The prevalence model implicitly conditions on a patient being alive and having CF, hence can appropriately handle the PA infection data which feature complicated sampling and censoring schemes, as described next.

3. Temporal Process Regression in the Context of CFFPR

We now apply the temporal process regression method (Fine et al., 2004) to the prevalence of ever PA positive and currently PA positive. Two temporal processes are constructed whose expectations give, respectively, the prevalence of ever PA positive and the prevalence of currently PA positive. Let $Y_{e,i}(t) = 1$ if patient i has ever had PA positive by time t and 0 otherwise. Let $Y_{c,i}(t) = 1$ if patient i has currently PA positive at time t and 0 otherwise. Both temporal processes are continuously observed for patient i during the age interval $[L_i, U_i]$, where L_i is the age at first visit and U_i is the age at last visit in the registry before December 31, 2005. The time window $[L_i, U_i]$ is always observed for each patient. For patients with PA positive at the first visit, $Y_{e,i}(t) = 1$ for all $t \in [L_i, U_i]$ and $Y_{c,i}(t) = 1$ for all

t from L_i until the age at the next visit where it may or may not switch back to 0. These data contain information about the prevalence and will not be thrown away.

The truncation and censoring by death are addressed by the time window $[L_i, U_i]$. If $T_{D,i}$ is time to death for individual i, then individual i is sampled if $T_{D,i} > L_i$, so-called left truncation on mortality. The death time $T_{D,i}$ is also right censored by U_i, which is always observed. Such truncation and censoring of mortality are explicitly dealt with in the models for the prevalence of PA infection in surviving patients discussed below.

Let $Y_i(t)$ (either $Y_{e,i}(t)$ or $Y_{c,i}(t)$) be a 0–1 process whose expectation is the prevalence of interest and $\mathbf{X}_i$ be a $p \times 1$ covariate vector. Consider the time window $[l, u] = [0.5, 10.0]$, where the left end point 0.5 is chosen to avoid numerical problems near age zero caused by the small number of contributing patients and the right end point 10 is chosen to focus on the first decade of life of CF patients. We observe n independent copies of $\{Y_i(t), \mathbf{X}_i : \delta_i(t) = 1\}$, $i = 1,\ldots, n$, where $\delta_i(t) = I(t \in [L_i, U_i])$ is the under-observation indicator. Let $\xi_i(t)$ be the indicator of being alive, i.e., $\xi_i(t) = 1$ if patient i is alive at age t and 0 otherwise. Then, $\mu_i(t) = E\{Y_i(t)|\mathbf{X}_i, \xi_i(t) = 1\}$ is the prevalence conditioning on the covariate vector $\mathbf{X}_i$ and subject i being alive at time t. The temporal process regression model is a varying coefficient generalized linear model (GLM) for each t in $[l, u]$

$$g\{\mu_i(t)\} = \mathbf{X}_i^{\top}\beta(t), \quad (1)$$

where g is a known link function, and $\beta(t)$ is a vector of completely unspecified time-varying coefficients. The model is robust in that only the mean of $Y_i(t)$ is specified instead of the full joint distribution of $Y_i(t)$, for all $t \in [l, u]$.

Note that the proposed prevalence model for PA infection, both current and ever positive, at a given time point t, conditions on death being larger than t, that is, $T_{D,i} > t$. To accommodate the left truncation on mortality, which conditions implicitly on $T_{D,i} > L_i$, one may instead condition on $T_{D,i} > t > L_i$, which gives the original model conditioning on $T_{D,i} > t$, as conditioning on $t > L_i$ is redundant. This can be incorporated in the definition of $\mu_i(t)$ in the model. In the CFFPR data where mortality is rather low, the difference between conditioning on $T_{D,i} > \max(t, L_i)$ and $T_{D,i} > t > L_i$ is very small and can be ignored.

For the response process $Y_{e,i}(t)$ of ever PA positive, model (1) is intrinsically connected with hazard models for the onset time $T_{e,i}$ to first PA positive. The connection is subtle because prevalence at age t is estimated based on all patients who are alive at age t. When there is no death, the prevalence of ever PA positive $\mu_i(t)$ is the complement of the survival function of $T_{e,i}$, $S_i(t)$. Therefore, $\Lambda_i(t) = -\log\{1 - \mu_i(t)\}$, where $\Lambda_i(t) = \int_0^t \lambda_i(s)\,\mathrm{d}s$ is the cumulative hazard function. When g is the complementary log-log (cloglog) link, i.e., $g(u) = \log\{-\log(1-u)\}$, we have $\log \Lambda_i(t) = \mathbf{X}_i^{\top}\beta(t)$, which reduces to the Cox (1972) proportional hazards model for $\beta(t) = \beta$ except for the intercept. Consequently, model (1) with the cloglog link for the prevalence of ever PA positive can be used to check the appropriateness of the proportional hazards models in Lai et al. (2004). In the CF data, where the mortality rate is negligible (0.18% before age 10), such check is approximately valid.

For the response process $Y_{c,i}(t)$ of currently PA positive, $\mu_i(t)$ is not directly connected to a single hazard function as discussed above for ever PA positive. To our knowledge, analysis of currently PA positive prevalence has not previously been reported in the CF literature; the results are economically meaningful, and are distinct from those based on ever PA positive.

Under the assumptions of conditional independence $\{Y_i(t) \perp \delta_i(t)\}|\{\mathbf{X}_i, \xi_i(t) = 1\}$ and positive probability of complete data $\Pr\{\delta_i(t) = 1|\mathbf{X}_i(t), \xi_i(t) = 1\} > 0$, $t \in [l, u]$, the

parameters $\beta(t)$ at each time $t \in [l, u]$ can be estimated using the n_t observations available at this time, where $n_t = \Sigma_{i=1}^{n} \delta_i(t)$. In particular, $\beta(t)$ can be estimated using quasi-likelihood approach (McCullagh and Nelder, 1989) by solving the quasi-score equation:

$$\sum_{i=1}^{n} \delta_i(t)\, \mathbf{D}_i\{\beta(t)\}\, \mathbf{V}_i^{-1}\{\beta(t)\}\{Y_i(t) - \mu_i(t)\} = 0, \quad (2)$$

where $D_i\{\beta(t)\} = d\mu_i(t)/d\beta(t)$, and $\mathbf{V}_i^{-1}\{\beta(t)\}$ is a working weight function, possibly random. Combining $\widehat{\beta}(t)$ from estimating equation (2) for all $t \in [l, u]$, we obtain an estimator of the time-varying coefficients $\beta(t)$ in model (1), without assuming any functional form; see Fine et al. (2004) for details. In practice, particularly for point processes, there are only finitely many jump points. Therefore, the estimator $\widehat{\beta}(t)$ jumps at those M times where $\{Y_i(t), \mathbf{X}_i, \delta_i(t); i = 1, \ldots, n\}$ jump. Since $Y_i(t)$'s are piecewise constant, so too is $\widehat{\beta}(t)$. Finding $\widehat{\beta}(t)$ involves solving (2) at M points. When M is big, as in the CFFPR data, pointwise estimation can be performed on a fine grid of the interval $[l, u]$.

Two important hypotheses tests are considered in analyzing the CFFPR data: a regression coefficient is not significant $H_0 : \beta_j(t) = 0$, $t \in [l, u]$, and a regression coefficient is constant $H_c : \beta_j(t) = \beta_j$, $t \in [l, u]$. Two significance tests are constructed for H_0. The first one is an integral test based on the standardized weighted average of $\widehat{\beta}_j(t) : T^* = T/\widehat{\Sigma}_T^{1/2}$, where $T = \int_l^u \widehat{\beta}_j(t)\, W(t)\, dt$ with some weight function $W(t)$ and $\widehat{\Sigma}_T$ is the variance estimate of T. The second one is the supremum of the weighted process $S = \sup_{t \in [l,u]} |\widehat{\beta}_j(t)\, W(t)|$, whose distribution can be approximated by bootstrapping the influence functions of $\widehat{\beta}_j(t)\, W(t)$. For the goodness-of-fit test H_c, a time-independent estimate $\widehat{\beta}_j = \left\{\int_l^u H(t)\, dt\right\}^{-1} \int_l^u \widehat{\beta}_j(t)\, H(t)\, dt$ is first constructed, where $H(t)$ is a weight function, chosen as $\widehat{\mathrm{var}}\{\widehat{\beta}_j(t)\}^{-1/2}$ in this analysis to downweight time points where there are more variability. An integral statistic and a supremum statistic can then be constructed as in the case of significance test, with $\widehat{\beta}_j(t) - \widehat{\beta}_j$ in place of $\widehat{\beta}_j(t)$. The weight function $W(t)$ influences the power of the test statistics; see Fine et al. (2004) for details. In the supreme test, $W(t)$ is chosen as $\widehat{\mathrm{var}}\{\widehat{\beta}_j(t)\}^{-1/2}$. In the integral test, $W(t)$ is chosen as $I(t \in \mathcal{I})\left[\widehat{\mathrm{var}}\{\widehat{\beta}_j(t)\}\right]^{-1/2}$, where $\mathcal{I}$ is a subinterval of $[l, u]$. The analyses in Section 4 use $[l, u]$ or one of its three equally spaced subintervals as $\mathcal{I}$. The latter weight scheme has substantial power to detect linear and quadratic alternatives to H_c.

4. Prevalence Analysis Results

4.1 Ever PA Positive

For prevalence of ever PA positive, as discussed in Section 3, we use the cloglog link function to formulate $\log[-\log\{1 - \mu_i(t)\}] = \mathbf{X}_i^{\mathrm{T}} \beta(t)$, where $\mu_i(t)$ is the conditional prevalence of ever PA positive at age t given covariate vector $\mathbf{X}_i$. Since the mortality rate before age 10 is negligible (21 deaths, 0.18%), the time-varying covariate coefficients $\beta(t)$ can be used to test the goodness-of-fit of proportional hazards. In contrast to the Cox model results in Figure 1, patients with PA positive at their first visits are included in this ever PA positive model. Exploratory analyses with univariate models and two-variate models stratified by three diagnostic cohorts are performed before including all three risk factors in a single joint model. They suggest an additive model without interactions.

The nonparametric coefficient estimates in the three-factor model are graphically presented in Figure 2(a) with the same scales. The intercept, after being transformed by the inverse link function of cloglog, corresponds to the estimated prevalence of ever PA positive for male patients diagnosed via SCR during 1986-1989. The intercept is generally increasing, although this constraint is not required by the prevalence models and not enforced by the estimation procedure. Recall that the ever PA positive model reduces to a proportional hazards model when the hazard ratios are time-independent, given that the mortality before age 10 is negligible. The plots in Figure 2(a) suggest that proportional hazards may not hold since coefficients such as DX (SYMP), DX [90-93], and DX [94-00] are highly nonconstant over age. These results are similar qualitatively to those in Figure 1.

The significance test results for each coefficient using integral test statistic T^* and supremum statistic S are summarized in the upper panel of Table 1. The integral test statistics are computed on the whole interval $[l, u]$ and on every one of the three equally divided subintervals (i.e., approximately 0.5-4, 4-7 and 7-10, respectively). Their signs suggest directions of the covariate effects. The p-values of the supremum test statistic are computed from 1000 bootstrap samples (Fine et al., 2004). These results suggest that all coefficients, except that for the DX (FH) group, are significantly nonzero, and the levels of significance differ among the three subintervals. For example, the risk differences between DX [90-93] and DX [86-69] are highly significant and positive in the 3rd subinterval, moderately significant and positive in the 2nd subinterval, and negative albeit insignificant in the 1st subinterval.

Table 2, upper panel, summarizes the fitted constant parametric submodels for each of the covariates and the associated goodness-of-fit test results. The integral and supremum goodness-of-fit tests suggest that the gender effect is age-independent but the coefficients of DX (SYMP), DX [90-93], and DX [94-00] are time-varying instead of constant, confirming the visual impression from Figure 2(a). The protective effect of SCR on ever PA positive over SYMP is highly significant at earlier ages but attenuates at older ages. Attenuation is also observed for the DX (MI) and DX (FH) groups in Figure 2(a), but such attenuation is not statistically significant.

4.2 Currently PA Positive

For prevalence of currently PA positive, we fit the model with both logit link and cloglog link. Since the results are similar in nature, only results with logit link are reported here. In particular, the model is expressed as $\log \frac{\mu_i(t)}{1-\mu_i(t)} = \mathbf{X}_i^\top \beta(t)$, with $\mu_i(t)$ interpreted as the conditional prevalence of currently PA positive at age t given covariate $\mathbf{X}_i$. Each component of $\beta(t)$ can be interpreted as a log odds ratio at age t per unit increase in the corresponding covariate, as in a logistic regression model. As with ever PA positive, exploratory analyses for currently PA positive also suggest an additive effects model without interactions.

The nonparametric coefficient estimates in the three-factor model are plotted in Figure 2(b). The intercept, after being transformed using the inverse link function of logit, is interpreted as the prevalence of currently PA positive for male patients diagnosed via SCR during 1986-1989. Under the cloglog link, the estimated intercept is lower than that in Figure 2(a) at all age points, which is expected since $Y_c(t) \leq Y_e(t)$. The differences between the diagnosis method groups and the reference SCR group and the gender difference exhibit similar patterns to those in the ever PA positive model in Figure 2(a). A striking difference, however, in the coefficient estimates of the diagnostic cohort groups was observed between Figure 2(a) and Figure 2(b), which is described in greater detail below.

The lower panels in Table 1 and Table 2 summarize the significance tests, constant fit, and goodness-of-fit tests for modeling the prevalence of currently PA positive. The difference between the diagnostic groups to the reference SCR group are all significant throughout the age period. The magnitudes of the differences attenuate with age and are more pronounced than those observed in the ever PA positive model. This is especially obvious for the DX (SYMP) and the DX (MI) groups. Specifically, the gap between SCR and SYMP narrows continuously between age 6 and 10, and so does that between SCR and MI. These time-varying phenomena are statistically significant, as indicated by strong rejections by the goodness-of-fit of the constant fit, except for the DX (FH) effect. The elevated risk of females is confirmed by the significance tests, and a constant fit seems to adequately describe the effect. The advantages of the diagnostic cohort groups DX [90-93] and DX [94-00] compared to the reference group DX [86-89] are both significant. Their constant fits are strongly rejected.

Unlike the ever PA positive model, the DX [90-93] and DX [94-00] groups showed a significantly lower, instead of a higher, prevalence of currently PA positive compared to the DX [86-89] group. The opposite trend in PA prevalence comparing DX [94-00] to DX [86-89] between the ever PA positive model and currently PA positive model may be explained by the change in clinical practice toward more aggressive identification and treatment for PA (Treggiari et al., 2007). Compared to patients diagnosed prior to 1990 (i.e., DX [86-89]), patients diagnosed more recently (i.e., DX [94-00]) are likely to be cultured more frequently (e.g., every 3 rather than 6 months), resulting in earlier detection of the first PA infection and hence, a higher prevalence of ever PA positive at earlier ages. Nevertheless, compared to the DX [86-89] group, the DX [94-00] group is also more likely to receive earlier and more aggressive antibiotic therapy for PA, which increases the chance to eradicate PA (Valerius et al., 1991; Frederiksen et al., 1997), leading to a lower prevalence of currently PA positive at older ages. Similar but less marked patterns are observed for patients diagnosed in DX [90-93] versus the reference group.

5. Discussion

This paper has explored two notions of prevalence in the early respiratory development of CF patients using a national registry database. Two complementary prevalence definitions were employed, which describe the cumulative and current risks of infection in the surviving population. A unifying temporal regression framework was employed which illuminated the underlying dynamics of infection in CF patients and the evolving differences amongst diagnostic method and cohort groups. The framework addresses limitations in the default survival analyses and provides some new insights not available with the existing approaches. It subtly differs from the recent approaches for marginal mean functions accommodating terminal events such as death (Ghosh and Lin, 2000, 2002) because the prevalence of interest implicitly conditions on being alive.

Inferences about time-varying effects are the cornerstone of the analyses presented in Section 4. Testing may be carried out globally after explicitly averaging across time, or locally, via pointwise tests which control for multiple testing, as with simultaneous confidence bands. Temporal process regression permits a formal evaluation of time-dependence, with confidence bands pinpointing those times at which a significant effect is present and profiling the magnitudes of the effects over time. In a survival analysis framework for the onset age of first PA positive, which defines the ever PA positive outcome, existing Cox models with time-varying coefficients can be applied, but in a somewhat artificial way, with left censored observations excluded due to the lack of readily available software. Recent regression methods for doubly censored survival data (Zhang and Li, 1996; Ren and Gu, 1997; Cai and Cheng, 2004), with time-independent coefficients,

have not been implemented in standard statistical packages. Temporal process regression accommodates left censored observations naturally, taking advantage of the additional information on the observation window which is generally available in registry data. One might expect substantial increases in power to detect covariate effects in scenarios similar to the CF registry, where roughly 25% of observations are left censored. More rigorous numerical and theoretical studies are needed to demonstrate these efficiency gains.

For currently PA positive, disease prevalence may be considered as a transit state in a multistate model. Standard approaches to multistate models, like those in Andersen et al. (1995), model the transition intensities instead of the prevalences associated with the states. It may be difficult to interpret covariate effects on the transition intensities in terms of the prevalences, owing to the complicated relationships between these quantities. Other recent work aims at directly modelling such prevalence type quantities to ease the model interpretation. Ghosh and Lin (2002) model the marginal means for recurrent events using survival type models, while Pepe and Couper (1997) consider longitudinal data models for prevalences. The former methods are not applicable to PA prevalence in CFFPR. Temporal process regression (Fine et al., 2004) permits survival type analyses of prevalence with more flexible assessments of covariate effects than the longitudinal data analyses in Pepe and Couper (1997), generalizing the latter methods to continuous time set-ups with time-dependent coefficients.

Care is needed in interpreting the pointwise results in the CF registry analysis. For example, at very young ages, the prenatal/neonatal screening program evidences highly significant differences relative to traditional diagnosis via naturally occurring symptoms, with respect to both ever PA positive and current PA positive. By age 10, however, these differences have substantially attenuated and are not significant. From a clinical point of view, the reduction in disease burden at early ages is meaningful, potentially leading to better long term mortality and morbidity outcomes. Nevertheless, further analysis is needed to elucidate this issue, that is, to quantify the extent of the long term benefits of early detection, which has important policy implications. The interpretation of the temporal effect of diagnostic cohort is also rather subtle, given that the effects on ever PA positive and current PA positive are in opposite directions; see Section 4.2.

In the CF community, there is interest in distinguishing short transient infections from chronic persistent infections. Persistent infections reflect a more advanced disease state than do transient infections (Lee et al., 2004; Treggiari et al., 2007), and are most costly to manage. While the focus of this paper has been on all infections combined, understanding the effects of risk factors on infection patterns is also of interest. For example, in a univariate analysis for the SCR group, the prevalence of ever PA positive is roughly 0.20 at age 2 and 0.40 at age 4, while the prevalence of current PA positive is roughly 0.12 at age 2 and 0.16 at age 4. This gives that roughly 60% of ever positive patients at age 2 are currently infected while 40% of ever positive patients at age 4 are currently infected. These percentages vary across diagnostic method. Further work is needed to formally analyze such differences using the temporal regression methodology. This is a topic for future research.

It is worth pointing out that the different frequency in collecting the data can influence the determination of the disease onset age. Had the tests been taken more frequently in certain patient groups, for example, as a result of more severe symptoms, then biases in the analyses might arise. In the CFF registry, this is not an issue, as per protocol, the monitoring scheme is the same for all patients. Regarding the interpretation of the analyses, more frequent monitoring would yield more accurate determination of PA onset and remission times and hence the prevalence endpoints would be more accurately profiled. Given that visits occur

fairly regularly (2 to 4 times per year), however, the monitoring scheme would seem to have a relatively minor impact on the analyses.

Last observation carried forward method is used to obtain the information on PA infection at each visit date. This is an approximation of the true indicator process, but is clinically relevant because treatment is based on the observed indicator process and one is interested in modeling the clinical process. Nevertheless, changes in observation schemes may result in different observed indicator processes, therefore, the modeling of the observed process does not necessarily reflect the true underlying event process.

Acknowledgments

The authors thank Dr. Preston W. Campbell from the Cystic Fibrosis Foundation and the Cystic Fibrosis Foundation Committee for providing the Registry data. The authors thank Dr. Philip Farrell for reviewing the manuscript and providing comments related to lung infection in cystic fibrosis patients. J. Yan's research is partially supported by U.S. National Science Foundation grant DMS 0805965.

References

Aebi C, Bracher R, Liechti-Gallati S, Tschäppeler H, Rüdeberg A, Kraemer R. The age at onset of chronic Pseudomonas aeruginosa colonization in cystic fibrosis — prognostic significance. European Journal of Pediatrics. 1995; 154:S69–73. [PubMed: 8529715]

Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical Models Based on Counting Processes. Springer-Verlag Inc; 1995.

Assael BM, Castellani C, Ocampo MB, Iansa P, Callegaro A, Valsecchi MG. Epidemiology and survival analysis of cystic fibrosis in an area of intense neonatal screening over 30 years. American Journal of Epidemiology. 2002; 156:397–401. [PubMed: 12196308]

Cai T, Cheng S. Semiparametric regression analysis for doubly censored data. Biometrika. 2004; 91:277–290.

Campbell PW, White TB. Newborn screening for cystic fibrosis: An opportunity to improve care and outcomes. Journal of Pediatrics. 2005; 145:S2–S5. [PubMed: 16202776]

Comeau AM, Accurso FJ, White TB, Campbell PW, Hoffman G, Parad RB, Wilfond BS, Rosenfeld M, Sontag MK, Massie J, Farrell PM, O'Sullivan BP. Guidelines for implementation of cystic fibrosis newborn screening programs: Cystic fibrosis foundation workshop report. Pediatrics. 2007; 119:e495–e518. [PubMed: 17272609]

Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B: Methodological. 1972; 34:187–220.

Cystic Fibrosis Foundation. Patient registry 2008 annual report. Annual report, Cystic Fibrosis Foundation; Bethesda, Maryland: 2008.

Fine JP, Yan J, Kosorok MR. Temporal process regression. Biometrika. 2004; 91:683–703.

Frederiksen B, Koch C, Hoiby N. Antibiotic treatment of initial colonization with Pseudomonas aeruginosa postpones chronic infection and prevents deterioration of pulmonary function in cystic fibrosis. Pediatric Pulmonology. 1997; 23:330–335. [PubMed: 9168506]

Gail MH, Kessler L, Midthune D, Scoppa S. Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. Biometrics. 1999; 55:1137–1144. [PubMed: 11315059]

Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. Biometrics. 2000; 56:554–562. [PubMed: 10877316]

Ghosh D, Lin DY. Marginal regression models for recurrent and terminal events. Statistica Sinica. 2002; 12:663–688.

Grosse SD, Rosenfeld M, Devine OJ, Lai H, Farrell PM. Potential impact of newborn screening for cystic fibrosis on child survival: A systematic review and analysis. The Journal of Pediatrics. 2006; 149:362–366. [PubMed: 16939748]

Kosorok MR, Zeng L, West SE, Rock MJ, Splaingard ML, Laxova A, Green CG, Collins J, Farrell PM. Acceleration of lung disease in children with cystic fibrosis after Pseudomonas aeruginosa acquisition. Pediatric Pulmonology. 2001; 32:277–287. [PubMed: 11568988]

Lai H, Cheng Y, Cho H, Kosorok MR, Farrell PM. Association between initial disease presentation, lung disease outcomes, and survival in patients with cystic fibrosis. American Journal of Epidemiology. 2004; 159:537–546. [PubMed: 15003957]

Lai HJ, Cheng Y, Farrell PM. The survival advantage of patients with cystic fibrosis diagnosed through neonatal screening: Evidence from the United States Cystic Fibrosis Foundation registry data. Journal of Pediatrics. 2005; 147:S57–63. [PubMed: 16202784]

Lee TW, Brownlee KG, Denton M, Littlewood JM, Conway SP. Reduction in prevalence of chronic Pseudomonas aeruginosa infection at a regional pediatric cystic fibrosis center. Pediatric Pulmonology. 2004; 37:104–110. [PubMed: 14730654]

Liou TG, Adler FR, FitzSimmons SC, Cahill BC, Hibbs JR, Marshall BC. Predictive 5-year survivorship model of cystic fibrosis. American Journal of Epidemiology. 2001; 153:345–352. [PubMed: 11207152]

Martinussen, T.; Scheike, T. Dynamic regression models for survival data. Springer; New York: 2006.

McCullagh, P.; Nelder, JA. Generalized Linear Models. Second edition. Chapman & Hall Ltd; 1989.

Pepe MS, Couper D. Modeling partly conditional means with longitudinal data. Journal of the American Statistical Association. 1997; 92:991–998.

Ren J-J, Gu M. Regression M-estimators with doubly censored data. The Annals of Statistics. 1997; 25:2638–2664.

Rosenfeld M, Davis R, FitzSimmons S, Pepe M, Ramsey B. Gender gap in cystic fibrosis mortality. American Journal of Epidemiology. 1997; 145:794–803. [PubMed: 9143209]

Treggiari MM, Rosenfeld M, Retsch-Bogart G, Gibson R, Ramsey B. Approach to eradication of initial Pseudomonas aeruginosa infection in children with cystic fibrosis. Pediatric Pulmonology. 2007; 42:751–756. [PubMed: 17647287]

Valerius NH, Koch C, Hoiby N. Prevention of chronic Pseudomonas aeruginosa colonisation in cystic fibrosis by early treatment. Lancet. 1991; 21:725–726. [PubMed: 1679870]

van den Akker-van Marle ME, Dankert HM, Verkerk PH, Dankert-Roelse JE. Cost-effectiveness of 4 neonatal screening strategies for cystic fibrosis. Pediatrics. 2006; 118:896–905. [PubMed: 16950979]

Verdecchia A, Angelis GD, Capocaccia R. Estimation and projections of cancer prevalence from cancer registry data. Statistics in Medicine. 2002; 21:3511–3526. [PubMed: 12407687]

Zhang C-H, Li X. Linear regression with doubly censored data. The Annals of Statistics. 1996; 24:2720–2743.

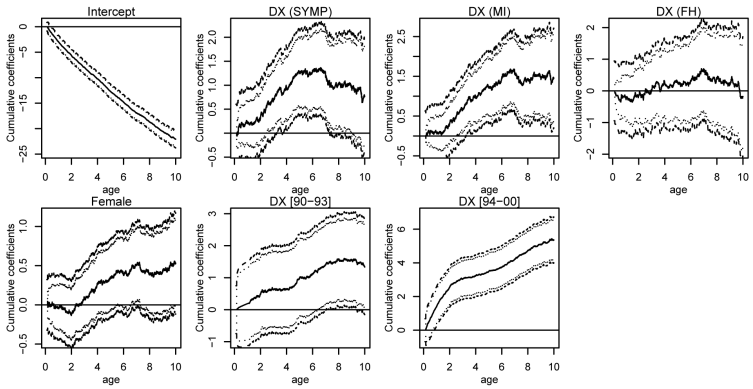
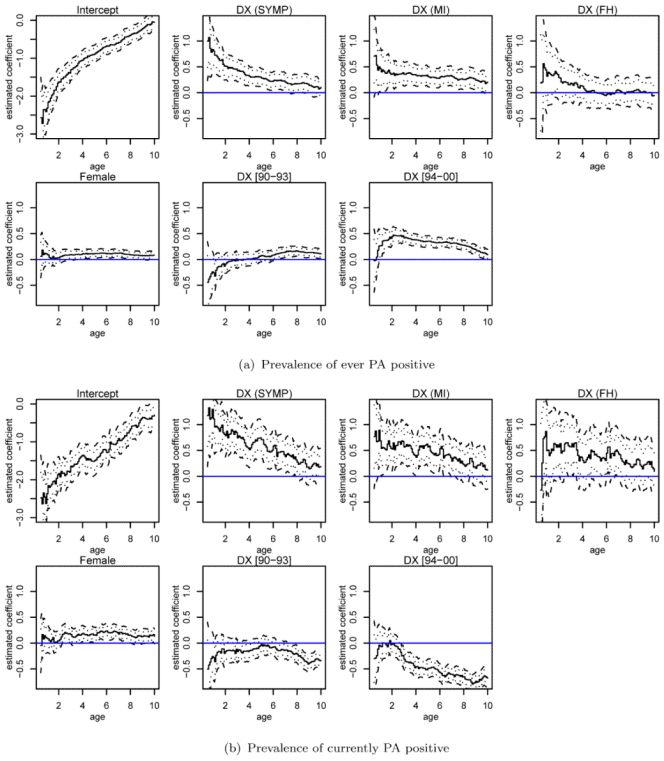


Figure 1.
Estimated cumulative coefficients, with 95% pointwise confidence interval and 95% Hall-Wellner confidence band in a time-varying coefficient Cox model.



(a) Prevalence of ever PA positive

(b) Prevalence of currently PA positive

Figure 2.

Estimated regression coefficients in the prevalence model of ever PA positive and currently PA positive. Solid lines are the nonparametric estimates. Dotted lines are pointwise 95% confidence intervals. Dashed lines are 95% confidence bands.

Table 1

Significant test for each regression coefficient in temporal process regression models. Stat: test statistic; Pval: p-value.

Risk Factor	Integral		Integral-1		Integral-2		Integral-3		Supremum	
	Stat	Pval	Stat	Pval	Stat	Pval	Stat	Pval	Stat	Pval
Prevalence of ever PA positive										
DX (SYMP)	5.746	0.000	7.020	0.000	4.784	0.000	2.761	0.006	6.371	0.000
DX (MI)	5.753	0.000	4.868	0.000	5.284	0.000	4.242	0.000	5.277	0.000
DX (FH)	0.886	0.376	2.011	0.044	0.147	0.883	0.158	0.874	2.495	0.161
Female	4.058	0.000	2.023	0.043	4.319	0.000	3.617	0.000	4.739	0.000
DX [90-93]	2.171	0.030	−1.332	0.183	1.795	0.073	4.579	0.000	4.971	0.001
DX [94-00]	10.604	0.000	8.357	0.000	10.485	0.000	7.541	0.000	10.177	0.000
Prevalence of currently PA positive										
DX (SYMP)	8.388	0.000	10.338	0.000	7.234	0.000	3.499	0.000	7.275	0.000
DX (MI)	6.275	0.000	6.904	0.000	5.705	0.000	2.940	0.003	5.334	0.000
DX (FH)	3.915	0.000	4.079	0.000	3.196	0.001	2.130	0.033	3.682	0.012
Female	5.445	0.000	2.827	0.005	5.394	0.000	4.293	0.000	5.218	0.000
DX [90-93]	−4.932	0.000	−2.980	0.003	−2.166	0.030	−6.330	0.000	7.785	0.000
DX [94-00]	−13.534	0.000	−3.040	0.002	−11.898	0.000	−15.714	0.000	14.145	0.000

Table 2

Constant fit and goodness-of-fit test for constant fit for each regression coefficient in temporal process regression models. Est: estimate; SE: standard error; Stat: test statistic; Pval: p-value.

Risk Factor	Constant Fit		Integral-1		Integral-2		Integral-3		Supremum	
	Est	SE	Stat	Pval	Stat	Pval	Stat	Pval	Stat	Pval
Prevalence of ever PA positive										
DX (SYMP)	0.295	0.051	4.834	0.000	−1.259	0.208	−3.863	0.000	3.625	0.002
DX (MI)	0.313	0.054	1.477	0.140	0.089	0.929	−1.474	0.140	1.983	0.420
DX (FH)	0.070	0.080	2.136	0.033	−1.669	0.095	−1.018	0.309	2.067	0.354
Female	0.095	0.023	−0.888	0.375	1.874	0.061	−0.325	0.745	1.511	0.808
DX [90-93]	0.069	0.032	−3.765	0.000	−0.321	0.749	4.075	0.000	3.507	0.010
DX [94-00]	0.312	0.029	2.631	0.009	2.580	0.010	−4.373	0.000	6.166	0.000
Prevalence of currently PA positive										
DX (SYMP)	0.552	0.066	5.280	0.000	0.543	0.587	−5.590	0.000	3.909	0.005
DX (MI)	0.436	0.069	2.758	0.006	1.080	0.280	−3.558	0.000	2.727	0.209
DX (FH)	0.391	0.100	1.663	0.096	0.090	0.929	−1.724	0.085	1.839	0.813
Female	0.156	0.029	−1.346	0.178	1.970	0.049	−0.231	0.818	2.207	0.525
DX [90-93]	−0.186	0.038	0.182	0.856	3.731	0.000	−3.283	0.001	4.210	0.000
DX [94-00]	−0.485	0.036	8.103	0.000	−1.553	0.120	−7.252	0.000	6.723	0.000