ORIGINAL ARTICLE

Understanding the natural progression in %FEV$_1$ decline in patients with cystic fibrosis: a longitudinal study

David Taylor-Robinson, Margaret Whitehead, Finn Diderichsen, Hanne Vebert Olesen, Tania Pressler, Rosalind L Smyth, Peter Diggle

ABSTRACT

Background Forced expiratory volume in 1 s as a percentage of predicted (%FEV$_1$) is a key outcome in cystic fibrosis (CF) and other lung diseases. As people with CF survive for longer periods, new methods are required to understand the way %FEV$_1$ changes over time. An up to date approach for longitudinal modelling of %FEV$_1$ is presented and applied to a unique CF dataset to demonstrate its utility at the clinical and population level.

Methods and findings The Danish CF register contains 70 448 %FEV$_1$ measures on 479 patients seen monthly between 1969 and 2010. The variability in the data is partitioned into three components (between patient, within patient and measurement error) using the empirical variogram. Then a linear mixed effects model is developed to explore factors influencing %FEV$_1$ in this population. Lung function measures are correlated for over 15 years. A baseline %FEV$_1$ value explains 63% of the variability in %FEV$_1$ at 1 year, 40% at 5 years, and about 30% at 5 years. The model output smooths out the short-term variability in %FEV$_1$ (SD 6.3%), aiding clinical interpretation of changes in %FEV$_1$. At the population level significant effects of birth cohort, pancreatic status and Pseudomonas aeruginosa infection status on %FEV$_1$ are shown over time.

Conclusions This approach provides a more realistic estimate of the %FEV$_1$ trajectory of people with chronic lung disease by acknowledging the precision in individual measurements and the correlation structure of repeated measurements on the same individual over time. This method has applications for clinicians in assessing prognosis and the need for treatment intensification, and for use in clinical trials.

INTRODUCTION

Understanding the long-term natural history of changes in lung function in people with lung diseases is a research priority. In order to do this, objective measures of disease progression are necessary. The per cent predicted forced expiratory volume in 1 s (%FEV$_1$) is commonly used to monitor lung function, and to describe disease severity in cystic fibrosis (CF) and chronic obstructive pulmonary disease (COPD). %FEV$_1$ is used to inform clinical decisions about changing or intensifying treatment, and as an outcome measure in clinical studies. Furthermore %FEV$_1$ has been shown to be related to survival in CF. Kerem et al’s study in 1992 demonstrated that patients with a %FEV$_1$ <50 had a 2-year mortality over 50%, though a more recent study shows that survival rates at low levels of lung function have improved in subsequent cohorts.

Interpreting the significance of changes in %FEV$_1$ in CF to inform patient management and to counsel patients regarding prognosis requires an understanding of the inherent variability of %FEV$_1$ measures within individuals, to determine what constitutes a clinically significant deterioration in %FEV$_1$, rather than a change due to measurement error, or recoverable day-to-day fluctuation in lung function. Furthermore, this variability needs to be understood to make valid inferences about the association between covariates and %FEV$_1$ in observational studies.

As survival in CF improves with successive cohorts, there are many more people surviving into late adulthood. An implication of this, coupled with the availability of long-term follow-up data in CF registers, is that up to date methods should be adopted to interpret the long-term dynamics of lung function in CF. Statistical techniques for...
longitudinal data analysis have been the subject of much methodological development over the past 20 years, and the random intercept and slope model has become a popular analysis framework. While this is often appropriate for relatively short follow-up periods, there are theoretical reasons to suggest that this approach makes assumptions that will lead to incorrect inferences if applied over longer follow-up periods. One central assumption is that the variability in %FEV\(_1\) increases as a quadratic function over time (in proportion to time squared), which leads to estimates that diverge unrealistically over longer time periods. Methods for undertaking these analyses over longer time periods have been described, but have not been commonly applied.

In this study we analyse a unique population-level dataset of people with CF that includes longitudinal %FEV\(_1\) measures taken monthly for up to 30 years. We apply these methods to develop a general model for %FEV\(_1\) decline that goes beyond the popular random-intercept and slope approach, and explicitly describes the variability in %FEV\(_1\) within individuals over time. We show how this could be applied clinically to help interpret the significance of changes in lung function, and at a population level to explore the association of covariates (eg, *Pseudomonas aeruginosa* acquisition) with %FEV\(_1\) decline.

**METHODS**

**Subjects**

All patients aged over 5 years whose %FEV\(_1\) data were entered on the Danish CF database between 1969 and 2010 were eligible. Post-transplant data from patients who had received a lung transplant were excluded. Patients attending the two Danish CF centres (Copenhagen and Aarhus) are seen routinely every month in the outpatient clinic for evaluation of clinical status, pulmonary function and microbiology of lower respiratory tract secretions. It is estimated that coverage of people with CF resident in Denmark is almost complete from 1990 when CF care was centralised. This coverage and the unparallelled frequency of measurement make this a unique dataset for epidemiological research. The study was approved by the Danish Data Inspectorate (Datatilsynet).

**Lung function testing**

The primary outcome for this analysis was %FEV\(_1\). Pulmonary function tests were performed according to international recommendations, measuring FEV\(_1\), expressed as a percentage of predicted values for sex and height using reference equations from Wang or Hankinson.

**Covariates**

Covariates in the analysis were age, sex, genotype coded as the number of Delta F508 alleles (0, 1 or 2), onset of chronic *Pseudomonas* infection (coded 0 or 1 as a time-varying covariate), pancreatic insufficiency determined on the basis of pancreatic enzyme usage (coded 0 or 1 as a baseline covariate), birth cohort (six 10-year cohorts starting at 1948), and CF-related diabetes (CFRD) diagnosed using the WHO criteria (coded 0 or 1 as a time-varying covariate).

**Statistical analysis**

A detailed explanation is given in the online appendix. Repeated %FEV\(_1\) measures on individuals are correlated, and this must be accommodated to obtain valid inferences. We used a linear mixed effects model with longitudinally structured correlation, and contrasted our approach with the widely used random intercept and slope model. We modelled random variation in %FEV\(_1\) over time for an individual subject so that the strength of the correlation of the random variation between two values depends on the corresponding time separation. The model decomposed the overall random variation in the data into three components: between subjects, between times within subjects, and measurement error.

First, we fit a provisional model for the mean response by ordinary least squares and used the empirical variogram of the residuals (see figure E1 in the online appendix) to provide initial estimates for the three components of variation, and for the shape of the correlation function of the between-times-within-subjects component. We then re-estimated all of the model parameters by maximum likelihood estimation, and used generalised likelihood ratio statistics to compare nested models, and Wald statistics to test hypotheses about model parameters. We assessed associations between single or multiple covariates and the population mean %FEV\(_1\) over time, and explored alternatives to a linear function for the population-averaged time trend.

**RESULTS**

**Population characteristics**

The dataset contained 70 448 lung function measures on 479 patients seen between 1969 and 2010 in Denmark (table 1). The median number of %FEV\(_1\) measures per person was 101 (range 2–597). The median follow-up period was 10.5 years (range 0.1–31.5), with a total of 6500 person-years of follow-up. Forty-two patients were followed up for more than 30 years (see also figures E2 and E3 in the online appendix).

**Table 1** Baseline characteristics of the Danish cystic fibrosis (CF) population

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N (%)</td>
<td>7.1 (4)</td>
<td>42 (8.8)</td>
<td>110 (23)</td>
<td>105 (21.9)</td>
<td>141 (29.4)</td>
<td>74 (15.4)</td>
<td>479 (100)</td>
</tr>
<tr>
<td>Women</td>
<td>1 (14.3)</td>
<td>19 (45.2)</td>
<td>48 (43.6)</td>
<td>52 (49.5)</td>
<td>74 (52.5)</td>
<td>42 (15.4)</td>
<td>236 (49.3)</td>
</tr>
<tr>
<td>No. Delta F508 = 0</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (0.9)</td>
<td>4 (3.8)</td>
<td>5 (3.5)</td>
<td>5 (6.8)</td>
<td>15 (3.1)</td>
</tr>
<tr>
<td>No. Delta F508 = 1</td>
<td>2 (26.6)</td>
<td>14 (33.3)</td>
<td>26 (23.8)</td>
<td>24 (22.9)</td>
<td>42 (29.8)</td>
<td>19 (25.7)</td>
<td>127 (26.5)</td>
</tr>
<tr>
<td>No. Delta F508 = 2</td>
<td>5 (71.4)</td>
<td>28 (66.7)</td>
<td>83 (75.5)</td>
<td>77 (73.3)</td>
<td>94 (66.7)</td>
<td>50 (67.6)</td>
<td>337 (70.4)</td>
</tr>
<tr>
<td>Developed chronic <em>Pseudomonas</em></td>
<td>6 (85.7)</td>
<td>31 (73.8)</td>
<td>84 (76.0)</td>
<td>55 (52.4)</td>
<td>20 (14.2)</td>
<td>5 (6.8)</td>
<td>201 (42)</td>
</tr>
<tr>
<td>Missing infection information</td>
<td>0 (0)</td>
<td>5 (11.9)</td>
<td>2 (1.8)</td>
<td>2 (1.9)</td>
<td>1 (0.7)</td>
<td>0 (0)</td>
<td>10 (2.1)</td>
</tr>
<tr>
<td>Pancreatic insufficient</td>
<td>7 (100)</td>
<td>42 (100)</td>
<td>105 (95.5)</td>
<td>99 (94.3)</td>
<td>133 (94.3)</td>
<td>73 (98.6)</td>
<td>459 (95.8)</td>
</tr>
<tr>
<td>Copenhagen</td>
<td>7 (100)</td>
<td>38 (90.5)</td>
<td>83 (75.5)</td>
<td>72 (68.6)</td>
<td>79 (56)</td>
<td>50 (67.6)</td>
<td>329 (68.7)</td>
</tr>
<tr>
<td>Alive</td>
<td>4 (57.1)</td>
<td>27 (64.3)</td>
<td>79 (71.8)</td>
<td>77 (73.3)</td>
<td>132 (93.6)</td>
<td>74 (100)</td>
<td>393 (82)</td>
</tr>
<tr>
<td>Developed CFRD</td>
<td>3 (42.9)</td>
<td>21 (50)</td>
<td>41 (37.3)</td>
<td>31 (29.5)</td>
<td>22 (15.6)</td>
<td>1 (1.4)</td>
<td>119 (24.8)</td>
</tr>
</tbody>
</table>

CFRD, cystic fibrosis related diabetes.
Limitations of random intercept and slope model

The high degree of short-term and long-term variation in predicted %FEV$_1$ is illustrated in figure 1. The standard random intercept and slope model approach is illustrated over long and short follow-up periods in figure 1A,C. This approach assumes that any deviation of an individual’s trajectory from the population mean is linear in time over the whole of the follow-up period apart from independent random errors. One can see that this assumption is reasonable over short time periods, as illustrated by the fit of the shorter dotted-line segments (figure 1A, C), but over longer time periods the individual data traces diverge unrealistically from their fitted linear mean trajectories (long solid line). Our proposed model produces a much closer fit to the data (figure 1B,D), and one that better reflects the relative magnitude of the three estimated components of variation in %FEV$_1$ over time.

Quantifying the variability in %FEV$_1$ over time

The empirical variogram quantifies the variability in the dataset (figure 2A). The intercept at time zero represents measurement error because there can be no true within-person variation at a time lag of zero. Of the total variance in the Danish dataset, about half is due to systematic differences between patients (eg, genotype, sex or pancreatic status), two-fifths is within patients, representing change over time (disease progression), and one-tenth is ‘measurement error’. In practice, this last component represents the combined effects of technical errors, and physiological variability occurring at time intervals less than the monthly interval of measurement, for example, day-to-day variability. This error variance equates to an average SD of 6.3% for repeated measures on the same individual at short time intervals. Figure 2B shows the proportion of the within-person variability in %FEV$_1$ at follow-up time (t), which can be explained by their %FEV$_1$ value at baseline. For example, about 50% of the within-patient variability at t=2.5 years is explained by the baseline measurement, and about 30% at t=5 years. Overall, the dependence on baseline measures gradually decays and is negligible at 15 years.

Clinical utility of our proposed model

The model can be used to guide interpretation of sudden changes in lung function. Consider seeing the person in figure 1B at...
Effect of covariates on lung function in the Danish population

We explored the effect of covariates that have been associated with %FEV₁ in previous studies to demonstrate how this model can be used to answer questions at the population level (see table E1 online appendix for univariate associations). There was no evidence to suggest that covariate effects were nonlinear (see figure E4 in online appendix). The final model included age, Pseudomonas status, pancreatic status, cohort and CFRD (table 2). Note that the estimated covariate effects in table 2 are population-averaged effects, that is, they describe average values of %FEV₁ for sub-populations of individuals sharing the same explanatory characteristics, rather than for any one individual.

The most prominent effects are associated with birth cohort, pancreatic function and the onset of Pseudomonas infection (figure 5). There is clear separation between the three most recent birth cohorts, with a successive increase in the intercept term at age 5 of 66%. Pancreatic insufficiency is associated with a significantly steeper rate of decline of lung function (−0.92% per year, 95% CI −1.7 to −0.3), as is acquisition of Pseudomonas

Table 2

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Point estimate</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept at age 5 years</td>
<td>66.02</td>
<td>61.13</td>
<td>70.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>CFRD</td>
<td>−2.47</td>
<td>−3.58</td>
<td>−1.37</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age</td>
<td>−0.26</td>
<td>−0.49</td>
<td>−0.03</td>
<td>0.025</td>
</tr>
<tr>
<td>Cohort=1948 (reference 1968)</td>
<td>1.20</td>
<td>−25.50</td>
<td>27.90</td>
<td>0.930</td>
</tr>
<tr>
<td>Cohort=1958</td>
<td>−0.75</td>
<td>−10.01</td>
<td>8.51</td>
<td>0.874</td>
</tr>
<tr>
<td>Cohort=1978</td>
<td>16.60</td>
<td>10.15</td>
<td>23.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cohort=1988</td>
<td>25.19</td>
<td>19.11</td>
<td>31.27</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cohort=1998</td>
<td>29.81</td>
<td>22.85</td>
<td>36.78</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pancreatic sufficiency</td>
<td>2.78</td>
<td>−10.43</td>
<td>15.99</td>
<td>0.679</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa infection</td>
<td>−0.51</td>
<td>−0.72</td>
<td>−0.29</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age×cohort=1948</td>
<td>−0.63</td>
<td>−0.67</td>
<td>0.61</td>
<td>0.920</td>
</tr>
<tr>
<td>Age×cohort=1958</td>
<td>0.06</td>
<td>−0.23</td>
<td>0.34</td>
<td>0.699</td>
</tr>
<tr>
<td>Age×cohort=1978</td>
<td>−0.72</td>
<td>−1.00</td>
<td>−0.44</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age×cohort=1988</td>
<td>−0.72</td>
<td>−1.09</td>
<td>−0.35</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Age×cohort=1998</td>
<td>0.50</td>
<td>−0.41</td>
<td>1.42</td>
<td>0.280</td>
</tr>
<tr>
<td>Age×pancreatic sufficiency</td>
<td>0.98</td>
<td>0.29</td>
<td>1.67</td>
<td>0.005</td>
</tr>
</tbody>
</table>

CFRD, cystic fibrosis related diabetes.
Cystic fibrosis

Figure 3  Effect of covariates on forced expiratory volume in 1 s as a percentage of predicted (%FEV1). (A) Birth cohort effect in the final model. There is clear separation between the three most recent birth cohorts, with a successive increase in the intercept term at age 5 years. (B) Effect of pancreatic insufficiency and Pseudomonas infection on the predicted population trajectory for a person born in the 1988–1998 cohort. CF, cystic fibrosis; PA, Pseudomonas aeruginosa.

making—the model fit curves in figure 1 provide more realistic estimates of underlying lung function, and more valid criteria for clinical decisions. We propose that this model could be used to develop a real-time smoothing tool embedded in electronic patient records to aid clinical interpretation of spirometry data. We suggest that access to this information would provide some reassurance to patients experiencing lower than expected lung function values, since lung function can recover quite dramatically, and these data suggest that a linear or stepwise decline in lung function over time is not the norm.

We have generated, for the first time to our knowledge, the variance function for %FEV1 in people with CF over long follow-up periods. This precisely quantifies how %FEV1 measures are correlated over time. Furthermore we have done this for the whole CF population of Denmark. This quantifies the degree to which a baseline %FEV1 measure can be used to predict subsequent %FEV1 measures over long follow-up periods, and is likely to be of interest to clinicians and patients. We demonstrate a long-term correlation between levels of %FEV1 within an individual. This suggests that there is long-term predictive value in a high %FEV1 measure—people with CF with a high %FEV1 at baseline are more likely to have a high %FEV1 up to 15 years later than individuals with a lower baseline %FEV1 (figure 2B). However, the predictive value of a %FEV1 measure drops away rapidly over this period. We can say that on average a %FEV1 reading today explains about 63% of the variability in %FEV1 at 1 year, 40% at 3 years, and about 30% at 5 years.

This corroborates Rosenthal’s study,26 which found that baseline %FEV1 explains 66% of the variability in %FEV1 at 1 year, and Mastella et al’s study of European registry data in which differences in lung function at enrolment at age 5, categorised as mild, moderate or severe, tracked through the study to age 40.27 Konstan et al also describe how a lower %FEV1 for a given age can be used to characterise the aggressiveness of lung disease.28 Other studies have shown a high %FEV1 to be an independent risk factor for a greater rate of decline of %FEV1 over the next few years.29 This is not at odds with our findings here; a high %FEV1 can be a risk factor for greater decline in the short term, while still being associated with a relatively higher %FEV1 over the longer term.28

At the population level we show how our approach can be applied to quantify the effect of covariates on changes in lung function. Furthermore, the partitioning of the variability in %FEV1 and the precise description of the correlation structure captured in the model provide important information for sample

DISCUSSION

We describe a novel longitudinal modelling technique specifically aimed at analysing long sequences of repeated measurements, and apply this to %FEV1 from a CF population. We show how this approach could be used to inform patient management, by aiding the interpretation of sudden changes in lung function, and by quantifying the predictive value of a baseline %FEV1 measure up to 15 years later. At the population level, we show how our model can be used to quantify the effect of covariates on populations or sub-populations. Translation of these methods into clinical practice is important because people with CF are living longer, and we have shown how commonly applied approaches are unhelpful over long follow-up periods.

This study quantifies the short-term variability in %FEV1 in this population (SD 6.3%), and demonstrates that %FEV1 measures within individuals are correlated over time lags of 15 years or more. We have also explored the effect of previously studied risk factors for lung function decline in the Danish CF population, and have demonstrated significant effects of birth cohort, pancreatic status and Pseudomonas infection status.

The findings from this study have a number of clinical applications. Quantifying the variability in lung function measures is essential to make correct clinical interpretation.18 Exploiting the unusually high frequency of data collection in Denmark, this study implies that on average a change in %FEV1 of >15% (ie, twice the error SD, to give a 95% confidence range) is likely to represent true within-patient variation over time (disease progression), whereas anything less than this could be due to short-term fluctuation, which may recover. Stanbrook et al19 found a pooled within-subject %FEV1 SD of 4.3% when measured over a 9-day period in 21 stable adults with CF. This population is different to the population in our study, who were measured regardless of clinical status, and one would therefore expect greater variability. Other studies have shown that people with CF, asthma and COPD have more short-term variability in lung function tests22–24 and that more impaired lung function is associated with greater variability.25

Our model can be used to generate an underlying representation of an individual’s ‘true’ lung function trajectory (figure 1B,D) that smooths out the noise inherent in %FEV1 measures. These smoothed traces could be used to inform clinical decision-making—the model fit curves in figure 1 provide more realistic estimates of underlying lung function, and more valid criteria for clinical decisions. We propose that this model could be used to develop a real-time smoothing tool embedded in electronic patient records to aid clinical interpretation of spirometry data. We suggest that access to this information would provide some reassurance to patients experiencing lower than expected lung function values, since lung function can recover quite dramatically, and these data suggest that a linear or stepwise decline in lung function over time is not the norm.

We have generated, for the first time to our knowledge, the variance function for %FEV1 in people with CF over long follow-up periods. This precisely quantifies how %FEV1 measures are correlated over time. Furthermore we have done this for the whole CF population of Denmark. This quantifies the degree to which a baseline %FEV1 measure can be used to predict subsequent %FEV1 measures over long follow-up periods, and is likely to be of interest to clinicians and patients. We demonstrate a long-term correlation between levels of %FEV1 within an individual. This suggests that there is long-term predictive value in a high %FEV1 measure—people with CF with a high %FEV1 at baseline are more likely to have a high %FEV1 up to 15 years later than individuals with a lower baseline %FEV1 (figure 2B). However, the predictive value of a %FEV1 measure drops away rapidly over this period. We can say that on average a %FEV1 reading today explains about 63% of the variability in %FEV1 at 1 year, 40% at 3 years, and about 30% at 5 years.

This corroborates Rosenthal’s study,26 which found that baseline %FEV1 explains 66% of the variability in %FEV1 at 1 year, and Mastella et al’s study of European registry data in which differences in lung function at enrolment at age 5, categorised as mild, moderate or severe, tracked through the study to age 40.27 Konstan et al also describe how a lower %FEV1 for a given age can be used to characterise the aggressiveness of lung disease.28 Other studies have shown a high %FEV1 to be an independent risk factor for a greater rate of decline of %FEV1 over the next few years.29 This is not at odds with our findings here; a high %FEV1 can be a risk factor for greater decline in the short term, while still being associated with a relatively higher %FEV1 over the longer term.28

At the population level we show how our approach can be applied to quantify the effect of covariates on changes in lung function. Furthermore, the partitioning of the variability in %FEV1 and the precise description of the correlation structure captured in the model provide important information for sample
size calculations in longitudinal clinical studies with %FEV₁ as an outcome. Increasingly longitudinal outcomes are being used in randomised control trials, and to undertake an a priori sample size calculation it is essential to have information on the correlation structure. Furthermore, our modelled %FEV₁ trace could be used as an outcome in its own right.

As with other studies of patients with CF, there is a striking cohort effect evident in this population. The treatment of CF lung disease has been transformed over the period captured in this analysis, from 1969 to the present day. Particularly impressive is the improvement in lung function in the post-1998 cohort by comparison with preceding birth cohorts. Although patients in this group are early in their disease progression, the overall picture suggests that new therapeutic strategies are continuing to provide improvements in respiratory function in CF.

Our approach to modelling changes in %FEV₁ can be applied over long follow-up periods. This is in contrast to the widely used random intercept and slope approach that has been applied in studies of CF and COPD over short-term, and longer-term follow-up periods. The development and testing of the new approach is facilitated by the nature of the Danish CF register—to our knowledge there are no other datasets that contain such frequent (monthly) measures of lung function on individuals measured over very long periods (up to 31.5 years). However, the fact that the data are from Denmark does not influence the validity of the methods we have described, since these are essentially context free. Furthermore, this method does not exploit any features of our data that are unique to CF and is equally applicable to other clinical areas that generate long sequences of repeated measurements. As a next step we recommend that this method be applied to longitudinal data collected in other CF registries, such as the UK, to clarify how robust this approach is in terms of predicting changes in %FEV₁ over time, and to better understand how this might inform clinical decision making. Future research could explore the utility of our proposed model in other diseases such as COPD.

A limitation of this study is the likely influence of survivor bias on lung function estimates in the earlier birth cohorts. In the 1948–1978 period, the intercept at age 5 appears significantly lower than in the other cohorts, but there is also a shallower rate of decline of lung function. This is likely to be due to the incomplete capture of patients in earlier cohorts, with censoring due to death leaving only the more stable survivors. This is a common problem in datasets of this type. Fitting the model by maximum likelihood automatically corrects for selection bias that depends on a patient’s observed lung function measurements prior to death, although not for any additional dependence on unmeasured features of their lung function trajectory.

Pancreatic sufficiency had an important effect on the overall rate of decline of lung function (+0.9% per year). In Konstan’s study pancreatic sufficiency was the most important protective factor in the age group 6–8 years (+1.55% per year). The small number of pancreatic-sufficient individuals in the Danish dataset (n=20, 5%) have a notably different lung function phenotype, maintaining near-normal lung function over the period of follow-up (see plot in online appendix). The onset of Pseudomonas infection was associated with a significant increase in the rate of decline of lung function, by around –0.5% per year, similar to that reported in the study by Konstan, in which Pseudomonas colonisation was associated with an increased rate of decline of FEV₁ of –0.31% per year in the 6–8-year-old age group, and –0.22 in the 9–12-year-old age group.

In conclusion, our modelling approach provides a more realistic estimate of the %FEV₁ trajectory in CF, which could be applied in real time to help clinicians interpret the significance of changes in %FEV₁. Furthermore, our approach quantifies the predictive value of a baseline %FEV₁ measure, over three decades. This method is equally applicable to the longitudinal assessment of %FEV₁ in other lung diseases, and can enable more robust comparisons of populations, including groups studied in clinical trials. As people are now living for many decades with these diseases, the development of tools to better understand the natural history of this important outcome will be essential for improved clinical care, as well as being a key research priority.

Acknowledgements We thank Professor Peter Olf Schiøtz for his support in accessing the data for this analysis.

Contributors DTR, MMW, PD, TP, RLS and PD conceived and designed the study. TP and HVD collected the data. DTR undertook the analysis and PD supervised analysis. DTR, MMW, RLS and PD interpreted the results and drafted the paper. All authors contributed to and approved the final draft for publication.

Funding This work was supported by an MRC Population Health Scientist Fellowship to DTR (G0802448). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests None.

Ethics approval The study was approved by the Danish Data inspectorate (Datatilsynet). Danish CF registry data were used, analysed anonymously.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

Cystic fibrosis

Online data supplement

Understanding the natural progression in %FEV1 decline in patients with cystic fibrosis: A longitudinal study

David Taylor-Robinson, Margaret Whitehead, Finn Diderichsen, Hanne Vebert Olesen, Tania Pressler, Rosalind Smyth, Peter Diggle
This data supplement contains additional information on the statistical methods employed in this study. In addition, further plots and results are presented to explore the model fit and the effect of covariates on lung function decline. Finally, R source code is included in an appendix.

**Methods**

**Subjects**

CF was diagnosed as either two known CF-causing mutations in the CF transmembrane conductance regulator (CFTR) gene, and/or two positive sweat-tests together with symptoms compatible with the disease.

The lung function tests are generally measured pre-bronchodilator, although, consistent with normal practice, patients are not instructed to omit their bronchodilator on the day of the measurement.

**Statistical analysis**

**Model**

We use a linear mixed effects model with longitudinally structured correlation. This allows a flexible specification of the mean response and incorporates three qualitatively different components of stochastic variation about the mean response (1, 2).

Let $Y_{ij}$ denote the $j$th repeated measurement (here, %FEV1) on the $i$th patient, and write

$$Y_{ij} = \mu_{ij} + R_{ij},$$

where $\mu_{ij}$ is the mean, population-averaged, response and $R_{ij}$ is the stochastic variation about the mean response.

In (1), the mean response is specified as a linear combination of explanatory variables, hence

$$\mu_{ij} = \alpha + \sum_{k=1}^{p} x_{ijk}\beta_k$$

In (2), the $x_{ijk}$ can be any measured values, whether time-constant or time-varying; for example, sex or age. Despite the model’s title, non-linear effects can also be captured. Polynominal time-trends can be defined by including powers of age amongst the $x_{ijk}$. Spline functions can be obtained by including both powers of age and indicator variables at selected time-points, called knots. For example, a model in which $x_{ij1} = \text{age}$ and $x_{ij2} = 0$ for age less than 10, $x_{ij2} = \text{age} - 10$ for age greater than 10, defines a linear spline with a single knot, also called a split-line or broken-stick model, with a change in slope at age 10.
To complete the model specification we decompose the stochastic term $R_{ij}$ in (1) into three components, hence

$$R_{ij} = U_i + W_i(t_{ij}) + Z_{ij},$$  

(3)

where $t_{ij}$ is the $j$th measurement time for the $i$th patient and the three components of $R_{ij}$ are specified as follows. Firstly, $U_i$ describes how the average lung function of the $i$th patient varies about the population-averaged response for all patients with the same values of the explanatory variables $x_{ijk}$, for example all males aged 20 years. The model assumes that the $U_i$ are independent copies of a Normally distributed random variable with mean zero and variance $\nu^2$. Secondly, the stochastic process $W_i(t)$ describes how the actual lung function of the $i$th patient varies over time. The model assumes that the $W_i(t)$ are independent copies of a stationary Gaussian process with mean zero, variance $\sigma^2$, and correlation function $\rho(u) = \text{Corr}\{W_i(t), W_i(t-u)\}$ (2). Typically, $\rho(u)$ decays towards zero as $u$ increases. In the current application, we use an exponential correlation function, $\rho(u) = \exp(-|u|/\phi)$, in which the parameter $\phi$ describes the rate at which the correlation decays towards zero with increasing time-separation, $u$. The exponential correlation function is a special case of the Matérn family, which includes a second parameter that allows the correlation function $\rho(u)$ to assume different shapes if the exponential model does not give a good fit (3). Thirdly, $Z_{ij}$ describes how the imperfectly measured lung function of the $i$th patient at their $j$th measurement time, $t_{ij}$, differs from their underlying actual lung function, i.e. measurement error. In principle, the properties of the measurement error could be estimated directly by repeated measurement of %FEV1 within a single follow-up session. In practice, the $Z_{ij}$ represent the sum of two sources of variation: pure measurement error and within-patient variation in lung-function on shorter time-scales than the shortest time-interval between successive measurement times, $t_{ij}$ and $t_{i,j+1}$. The model assumes that the $Z_{ij}$ are independent copies of a Normally distributed random variable with mean zero and variance $\tau^2$.

Although this is a very flexible model, it does assume that the decomposition of the covariance structure of the $R_{ij}$, as shown in (3), is common to all individuals. Note also that where interest focuses on stochastic variation in the rate of change in %FEV1 rather than on %FEV1 itself, this can also be modeled directly by using a stationary Gaussian process to model a time-varying rate of change, hence in (3)

$$W_i(t) = \int_0^t S(u)du$$

(4)

where $S(u)$ is a stationary Gaussian process.

**Exploratory analysis**

Exploratory analysis consists of identifying a suitable form for the set of mean responses $\mu_{ij}$ and obtaining initial estimates of the parameters in the model for the stochastic terms $R_{ij}$.

For the first of these tasks, we use a combination of ordinary least squares fitting of a
regression model, and kernel smoothing. Ordinary least squares gives unbiased estimates of baseline explanatory variable effects whatever the structure of the \( R_{ij} \), whilst kernel smoothing allows the investigation of possibly non-linear time-trends after adjustment for baseline effects. A kernel smoother is an estimate of the form

\[
s(t) = \sum_{i} \sum_{j} r_{ij} w_{ij}
\]

in which the \( r_{ij} \) are the residuals from the regression on baseline explanatory variables whilst the smoothing weights \( w_{ij} \) are scaled to add to 1 and are proportional to \( f(t - t_{ij}) \), where the kernel function, \( f(u) \), is a probability density function symmetric about \( u = 0 \); a common choice is a Normal probability density function with mean zero and standard deviation \( h \). In exploratory analysis, the value of \( h \) can be chosen subjectively so as to obtain a smoothly varying estimate \( s(t) \).

For the second task, we first re-define the residuals \( r_{ij} \) to adjust for the estimated smooth time-trend \( s(t) \) as well as for baseline explanatory variables. To estimate the covariance structure of these residuals we use the variogram, whose definition is as follows. Let \( v_{ijk} = (r_{ij} - r_{ik})^2/2 \) and \( u_{ijk} = |t_{ij} - t_{ik}| \). Pick a grouping interval \( h \), let \( n_r \) be the number of \( u_{ijk} \) that lie between \((r - 1)h\) and \( rh \) and \( \bar{v}_r \) the sample mean of the corresponding \( v_{ijk} \). A plot of \( \bar{v}_r \) against \((r - 0.5)h \) is called the sample variogram (1). It estimates the function \( V(u) = \tau^2 + \sigma^2(1 - \rho(u)) \), called the theoretical variogram. The sample variance of the residuals estimates the quantity \( \tau^2 + \nu^2 + \sigma^2 \). Hence, as illustrated in Figure 1 below, by sketching a smooth curve to fit the sample variogram we can obtain initial estimates of the variance components \( \tau^2, \nu^2 \) and \( \sigma^2 \), and of the correlation function \( \rho(u) \).

**Confirmatory analysis**

We estimate all of the model parameters by maximum likelihood estimation. This requires numerical maximisation of the log-likelihood function, whose algebraic form is that of the logarithm of a multivariate Normal probability density function with mean vector specified by (2) and a block-diagonal covariance matrix in which the \( i \)th block has diagonal elements \( c_{ij} = \tau^2 + \nu^2 + \sigma^2 \) and off-diagonal elements \( c_{jk} = \nu^2 + \sigma^2 \rho(u_{ijk}) \).

To compare nested models (i.e. one is a special case of the other), we use generalized likelihood ratio tests. If \( L_1 \) and \( L_0 \) denote the maximised values of the log-likelihood for nested models with \( p \) and \( p - q \) parameters, the generalised likelihood ratio test for the goodness-of-fit of the simpler model compares \( D = 2(L_1 - L_0) \) with critical values of the chi-squared distribution on \( q \) degrees of freedom.

To test hypotheses about model parameters, we use Wald tests. These exploit the property that the maximum likelihood estimates are approximately unbiased and Normally distributed, with standard errors that can be computed from the fitted model; for the algebraic details, see Diggle, Heagerty, Liang and Zeger (1).

The fitted theoretical variogram \( V(u) \) gives a graphical representation of the estimated variance components \( \tau^2, \nu^2 \) and \( \sigma^2 \), and of the correlation function \( \rho(u) \). An alternative representation is to transform \( V(u) \) into a function \( R^2(u) \), defined as
\[ R^2(u) = \left(1 - \frac{V(u)}{\sigma^2}\right)^2 \]  

This function is analogous to the conventional \( R^2 \)-value for a fitted regression model in the sense that it measures the proportion of within-patient variation in a person’s lung-function at time \( t+u \) that can be explained by their lung-function at time \( t \).

**Goodness-of-fit**

To test the overall goodness-of-fit of the final model, we analyse the residuals as follows. Firstly, plots of residuals against fitted values should show random scatter. Secondly, the residuals should have approximately the same covariance structure as the fitted model, which we check by comparing their sample variogram with the theoretical variogram of the model.

**Computation**

All analyses were carried out using the R open-source software environment (www.r-project.org). Maximum likelihood estimation, generalized likelihood ratio tests and Wald tests used the lme() function within the nlme package, together with the exponential class of correlation functions. Variogram calculations used a specially written R function. The R code follows.
Figure E1 A typical example of a theoretical variogram

Dashed horizontal lines represent a partitioning of the variance into three components, reading from bottom to top, $\tau^2 = 0.15$, $\sigma^2 = 0.6$ and $\nu^2 = 0.25$. Solid line is the curve $\tau^2 + \sigma^2 \{1 - \rho(u)\}$.  

Variation between subjects ($Ui$)

Variation between times within subjects ($Wi(t)$)

Measurement error ($Zij$)
Results

Visualising the dataset

Figure E2 is a histogram of year of birth for the people in the dataset. There is evidence of selective recruitment to the dataset in the earlier cohorts.

**Figure E2: Histogram of year of birth**

![Histogram of Year of Birth](image)

Figure E3 shows the frequency and length of follow up for people in the Danish dataset.

**Figure E3: Histogram of frequency and duration of follow up**

![Histogram of number of FEV1 measures](image)  ![Histogram of years of follow up](image)
Exploring the form for the population average

Figure E4 shows all the %FEV1 measures over time in a scatterplot, with an added mean smoother in red. The mean smoother does not take into account the correlation of repeated measures within individuals. There appears to be a fairly linear decline in mean %FEV1 to about age 25, where the mean stabilizes, before becoming more erratic at older ages where the sample size is smaller. A piecewise ordinary least squares (OLS) regression (blue line) with a change in slope at age 25 provides an improved fit over a straight line, and by eye one can see that this fits the smoother well. However, when one fits the same piecewise mean in the longitudinal model using MLE, then the change in slope at age 25 is not significant (green line). This indicates that the levelling off of the mean smoother to some extent reflects selective drop out (death). The final longitudinal model implicitly takes this drop out into account and generates the parameter estimates that one would expect to see if dropout had not occurred. We therefore modelled the population average as a straight line.

Fig E4: All data with smoothed mean trend

In a recent study McKay et al propose a novel approach to modelling lung function decline in an adult population(4), involving a linear mixed effect model with a cubic spline to account for non-linear population-averaged decline in lung function with age and a standard random intercept and slope model to account for within-patient variability. The spline gives a more flexible set of models for the population-averaged
trajectory, but our analysis of the Danish dataset shows that the random intercept and slope model is too rigid to capture the pattern of within-patient variability in %FEV1 over longer periods (Figure 1, main manuscript). Our approach can easily be combined with the spline model to describe non-linear population-averaged decline, since the form of the population average and the within-patient correlation structure are separate issues, but as illustrated above this was not needed for the Danish dataset. Note also that an adequate model for the correlation structure in the data is necessary for prediction at the individual patient level.

Assessing Model fit

Figure E5 below compares the empirical variogram fit to the theoretical variogram plotted using the MLE estimates from the lme() function in R, with an exponential correlation. This shows that the modelled correlation function approximates reasonably to the empirical correlation in the dataset.

Figure E5: Comparison between empirical variogram and MLE variogram estimate
The estimated variance ($\%$FEV$_1^2$) components derived from the modelled variogram are as follows: total variance=625, within-person variance=229, between-person variance=356, error variance=40.
Residual diagnostics

Figure E6 below plots the standardised residuals against the fitted values. There are no trends in the residuals, and there is no evidence of non-constant variance.

Figure E6: Scatterplot of standardised residuals versus fitted values
Figure E7 shows 4 simulated realisations of trajectories from the fitted model, and compares them to 4 real data traces. This demonstrates that the model is generating individual level predictions that have a similar form to the real data.

**Figure E7: Simulated realisations from the final model**
### Univariate associations

Table E1: Univariate associations between covariates and %FEV1

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Value (%FEV1)</th>
<th>Std.Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genotype</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>90.530</td>
<td>6.656</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (years)</td>
<td>-0.785</td>
<td>0.454</td>
<td>0.0837</td>
</tr>
<tr>
<td>Number delta F508=1</td>
<td>-2.801</td>
<td>7.075</td>
<td>0.6923</td>
</tr>
<tr>
<td>Number delta F508=2</td>
<td>-5.758</td>
<td>6.803</td>
<td>0.3978</td>
</tr>
<tr>
<td>age:Number delta F508=1</td>
<td>-0.070</td>
<td>0.466</td>
<td>0.8805</td>
</tr>
<tr>
<td>age:Number delta F508=2</td>
<td>-0.219</td>
<td>0.458</td>
<td>0.6327</td>
</tr>
<tr>
<td><strong>Pancreatic insufficiency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>92.101</td>
<td>7.144</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (years)</td>
<td>-0.060</td>
<td>0.362</td>
<td>0.868</td>
</tr>
<tr>
<td>Pancreatic insufficiency</td>
<td>-6.822</td>
<td>7.244</td>
<td>0.3468</td>
</tr>
<tr>
<td>age:Pancreatic insufficiency</td>
<td>-0.926</td>
<td>0.366</td>
<td>0.0114</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>84.907</td>
<td>1.680</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (years)</td>
<td>-1.015</td>
<td>0.077</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Male</td>
<td>1.747</td>
<td>2.393</td>
<td>0.4657</td>
</tr>
<tr>
<td>age:Male</td>
<td>0.091</td>
<td>0.105</td>
<td>0.3868</td>
</tr>
<tr>
<td><strong>Cohort</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>70.278</td>
<td>2.375</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age2</td>
<td>-0.640</td>
<td>0.084</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cohort&gt;=1948 (reference 1968)</td>
<td>9.082</td>
<td>13.988</td>
<td>0.5165</td>
</tr>
<tr>
<td>cohort&gt;=1958</td>
<td>1.252</td>
<td>4.860</td>
<td>0.7968</td>
</tr>
<tr>
<td>cohort&gt;=1978</td>
<td>15.062</td>
<td>3.387</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cohort&gt;=1988</td>
<td>22.492</td>
<td>3.159</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cohort&gt;=1998</td>
<td>26.813</td>
<td>3.640</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age:cohort&gt;=1948</td>
<td>-0.149</td>
<td>0.337</td>
<td>0.6571</td>
</tr>
<tr>
<td>age:cohort&gt;=1958</td>
<td>-0.024</td>
<td>0.149</td>
<td>0.8717</td>
</tr>
<tr>
<td>age:cohort&gt;=1978</td>
<td>-0.538</td>
<td>0.142</td>
<td>0.0001</td>
</tr>
<tr>
<td>age:cohort&gt;=1988</td>
<td>-0.356</td>
<td>0.178</td>
<td>0.0459</td>
</tr>
<tr>
<td>age:cohort&gt;=1998</td>
<td>0.884</td>
<td>0.464</td>
<td>0.057</td>
</tr>
<tr>
<td><strong>Pseudomonas</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>84.714</td>
<td>1.224</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (years)</td>
<td>-0.738</td>
<td>0.081</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>onset of pseudomonas</td>
<td>-0.392</td>
<td>0.104</td>
<td>0.0002</td>
</tr>
<tr>
<td><strong>CFRD</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>85.670</td>
<td>1.198</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (years)</td>
<td>-0.928</td>
<td>0.054</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>CFRD</td>
<td>-2.861</td>
<td>1.219</td>
<td>0.019</td>
</tr>
<tr>
<td>age:CFRD</td>
<td>0.024</td>
<td>0.064</td>
<td>0.701</td>
</tr>
</tbody>
</table>
Figure E8 plots the raw data disaggregated by pancreatic status. A mean smoother is plotted to the disaggregated data, as are the population averaged estimates from the univariate longitudinal model (parameters in table above).

**Figure E8: Univariate effect of pancreatic status**

![Effect of pancreatic status on FEV1 in Danish CF population](image)

**Illustration of cohort effect in multivariate analysis**

The two plots below illustrate the data disaggregated by cohort for the two contemporary cohorts (1988-98 and post-1998). Five individual traces are picked out in each plot, and the population average trajectory is plotted from the final multivariate model.
Figure E9: Scatterplot of data for post-1998 cohort with 5 individual traces randomly picked out, and linear population estimate from final model superimposed

Figure E10: Scatterplot of data for 1988-1998 cohort with 5 individual traces randomly picked out, and linear population estimate from final model superimposed

Appendix references


**R code**

```r
# generating spaghetti plots for 10 individuals
ns <- 10
samp <- sample(d$id, ns)
sampd <- subset(d, d$id %in% samp)
ids <- unique(sampd$id)

plot(sampd$age[sampd$id == ids[1]], sampd$fev1[sampd$id == ids[1]], type="l", ylim=c(0,150), xlab="Age", main="REAL DATA", ylab="% Predicted FEV1")
for (i in 2:ns)
  lines(sampd$age[sampd$id == ids[i]], sampd$fev1[sampd$id == ids[i]], col=i)

# generating variogram
variogram <- function(id, time, residual, u.max=NULL, u.increment=1) {
  # variogram function adapted from the geoR library, to deal with
  # longitudinal data-sets with long individual time series, geoR
  # library must be installed before use.
  #
  # Arguments:
  #   id: identifier for individual subjects
  #   time: time at which measurement is made
  #   residual: corresponding residual from model for mean response profiles
  #   u.max: maximum time-separation at which variogram is estimated (optional, but recommended)
  #   u.increment: increment between successive time-separations (not used when u.max=NULL)
  #
  # Result: a list with components:
  #   u: time-separations at which variogram is estimated
  #   v: corresponding variogram estimates
  #   n: number of pairs contributing to each variogram estimate
  #   sigmasquared: variogram-based estimate of process variance (sum of between-subject
  #                  and within-subject components)
  #
  # NOTES: 1. when data include replicated measurements at a common time within one or more subjects,
  #         the geoR library generates warning messages...these can safely be ignored in the
  #         present context
  #         2. when data include subjects with only one non-missing response, variogram calculation will fail
  #
  nid<-length(id)
  nt<-length(time)
  nr<-length(residual)
  check1<-c(nid-nt, nid-nr)
  if (max(abs(check1))>0) print("Bad data: unequal lengths amongst id, time and response")
  check2<-table(id)
  if (min(check2)<2) print("Bad data: at least one subject with only 1 response")
  library(geoR)
  uid<-unique(id)
  nid<-length(uid)
  if (length(u.max)==0) {
    u1<-(min(time)
    u2<-(max(time)
    h<-(u2-u1)/40
    u<-(length(u.max))
  } else {
    u<-(u.increment*((1:round(u.max/u.increment))-0.5))
  }
```
nu <- length(u)
u.all <- NULL
v.all <- NULL
n.within <- rep(NA, nid)
mean.within <- rep(NA, nid)
var.within <- rep(NA, nid)
for (i in 1:length(uid)) {
take <- (id == uid[i])
if (sum(take) > 1) {
x <- time[take]
y <- rep(0, length(x))
z <- residual[take]
xyz <- as.geodata(cbind(x, y, z))
vario <- variog(xyz, option = "cloud", messages = FALSE)
u.all <- c(v.all, vario[u])
v.all <- c(v.all, vario[v])
n.within[i] <- length(is.na(z))
mean.within[i] <- mean(z[!is.na(z)])
var.within[i] <- var(z[!is.na(z)])
}
sigmasquared <- sum(var.within[n.within >= 2] * (n.within[n.within >= 2] - 1)) / sum(n.within[n.within >= 2] - 1) + var(mean.within[!is.na(mean.within)])
nugget <- min(u.all) == 0
u.breaks <- c(0, u + 0.5 * u.increment)
if (nugget == TRUE) {
u.breaks <- c(0, u, breaks)
u <- c(0, u)
}
nu <- length(u)
v <- rep(NA, nu)
n <- rep(0, nu)
if (nugget == TRUE) {
take <- u.all == 0
n[1] <- sum(take)
v[1] <- mean(v.all[take], na.rm = TRUE)
} else {
take <- (u.all == u.breaks[1]) & (u.all <= u.breaks[2])
n[1] <- sum(take)
v[1] <- mean(v.all[take], na.rm = TRUE)
}
for (i in 2:nu) {
take <- (u.all == u.breaks[i]) & (u.all < u.breaks[i + 1])
n[i] <- sum(take)
v[i] <- mean(v.all[take], na.rm = TRUE)
}
list(u = u, v = v, n = n, sigmasquared = sigmasquared)
}

# Functions required for exploratory analysis
average.by.age <- function(x, y, x.lowest, x.increment, x.highest) {
nbreaks <- ceiling((range(x)[2] - range(x)[1]) / x.increment)
breaks <- x.lowest + x.increment * (0:nbreaks)
yvec <- rep(0, nbreaks)
for (i in 1:nbreaks) {
take <- (x <= breaks[i]) & (x <= breaks[i + 1])
yvec[i] <- mean(y[take], na.rm = TRUE)
}
xvec <- (breaks[2:(nbreaks + 1)] + breaks[1:nbreaks]) / 2
list(x = xvec, y = yvec)
}
data <- d
smooth.trend <- average.by.age(data$age, data$fev1, 5, 1, 60)
plot(data$age, data$fev1, pch = ".", xlab = "age (years)", ylab = "%FEV1")
lines(smooth.trend$x, smooth.trend$y, type = "l", lwd = 2, col = "red")

# add columns to data for smoothed response and residuals
yfit <- smooth.trend$y
xfit <- smooth.trend$x
N <- dim(data)[1]
ysmooth<-rep(NA,N)
for (i in 1:N) {
  take<-floor(data$age[i])==floor(xfit)
  if (sum(take)>0) ysmooth[i]<-yfit[take]
}
data$smooth<-ysmooth
data$res<-data$fev1-data$smooth
d$res <- data$res
x<-d$age
z<-d$res
id<-d$id
data.v<- variogram(id,x,z,u.max=30)

#plotting theoretical variogram trace (red line in figure 5)

x <- seq(0,30,0.1)
torsq <- 0.1500497* 16.41201^2
sigsq <- (1-0.1500497)* 16.41201^2
musq <- 18.87013^2
phi <- 6.7211035
y5 <- torsq + sigsq*(1-exp(-x/phi))
lines(x,y5,col="red",lwd=2)
lines(x,rep(torsq+sigsq+musq ,length(x)),col="red",lwd=2,lty=2)

#final multivariate model specification – this runs overnight on this dataset on a 2.8GHz Intel Core 2 Mac Book Pro (8GB memory), running 64-bit version of R

library(nlme)
exp1 <- corExp(value=c(7,(40/300)),form=~age2|id,nugget=T)
exp1 <- Initialize(exp1,d)
m15co <- lme(fev1 ~ sex+DM2+ age2*(cohort + PIb) +age3, data=d, random= ~ 1|id, method="ML",correlation=exp1)
summary(m15co)
intervals(m15co)

# simulation of data from a class of longitudinal models, irregularly
# observed in time
#
mu<-function(t,theta) {
  #
  # Arguments
  #     t: time (vector of non-negative real numbers)
  # theta: vector of parameters that define the mean response as a function of time
  #
  # Result
  #   Vector giving the values of the mean response at times corresponding to
  #   each element of the vector t
  #
  # Comment
  #   The following code is indicative only: it defines the mean response as
  #   linear in time, with intercept theta[1] and slope theta[2] but can be
  #   replaced by any other code that operates on vectors t and theta supplied
  #   as arguments
  #
  theta[1]+theta[2]*t
}

vmat<-function(t,nusqA,nusqB,rhoAB,tausq,sigmasq,phi,kappa=0.5) {
  #
  # Arguments
  #   t: time (vector of non-negative real numbers)
  #   nusqA: variance of random intercept
  #   nusqB: variance of random slope
  #   rhoAB: correlation between random intercept and slope
  #   tausq: measurement error variance
  #   sigmasq: variance of serially correlated component
  #   phi: scale parameter of Matern correlation function
  # kappa: shape parameter of Matern correlation function (defaults to exponential)
  #
  # Result
  #   Variance matrix of sequence of measurements at times corresponding to
  #   each element of the vector t
  nt<length(t)
vAB<-matrix(c(nusqA,rep(rhoAB*sqrt(nusqA*nusqB),2),nusqB),2,2)
xmat<-cbind(rep(1,nt),t)
\[ V_1 = x \text{mat} \times (\nu \text{AB}) \times (x \text{mat}) \]
\[ V_2 = \text{sigmasq} \times \text{matern}(\text{abs}(\text{outer}(t, t, \nspace)), \phi, \kappa) \]
\[ V_3 = \text{tausq} \times \text{diag}(\text{rep}(1, n)) \]
\[ V = V_1 + V_2 + V_3 \]

```r
simulate <- function(t, theta, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa = 0.5) {
  # Arguments
  # t: time (vector of non-negative real numbers)
  # theta: vector of parameters that define the mean response as a function of time
  # nusqA: variance of random intercept
  # nusqB: variance of random slope
  # rhoAB: correlation between random intercept and slope
  # tausq: measurement error variance
  # sigmasq: variance of serially correlated component
  # phi: scale parameter of Matern correlation function
  # kappa: shape parameter of Matern correlation function (defaults to exponential)
  # Result
  # Vector containing simulated realisation for a single subject
  # mean.vector <- mu(t, theta)
  # var.matrix <- vmat(t, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa)
  # rmvnorm(1, mean.vector, var.matrix)
}
```

```r
simulate.balanced <- function(ns, t, theta, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa = 0.5) {
  # Arguments
  # ns: number of subjects
  # t: time (vector of non-negative real numbers), common to all subjects
  # theta: vector of parameters that define the mean response as a function of time
  # nusqA: variance of random intercept
  # nusqB: variance of random slope
  # rhoAB: correlation between random intercept and slope
  # tausq: measurement error variance
  # sigmasq: variance of serially correlated component
  # phi: scale parameter of Matern correlation function
  # kappa: shape parameter of Matern correlation function (defaults to exponential)
  # Result
  # Matrix containing simulated realisations for ns subjects (rows) at
  # nt times (columns)
  # mean.vector <- mu(t, theta)
  # var.matrix <- vmat(t, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa)
  # rmvnorm(ns, mean.vector, var.matrix)
}
```

```r
simulate.unbalanced <- function(ns, tlist, theta, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa = 0.5) {
  # Arguments
  # ns: number of subjects
  # tlist: times (list of vectors of non-negative real numbers) unique to each subject
  # theta: vector of parameters that define the mean response as a function of time
  # nusqA: variance of random intercept
  # nusqB: variance of random slope
  # rhoAB: correlation between random intercept and slope
  # tausq: measurement error variance
  # sigmasq: variance of serially correlated component
  # phi: scale parameter of Matern correlation function
  # kappa: shape parameter of Matern correlation function (defaults to exponential)
  # Result
  # List containing simulated realisation for each subject
  # result <- as.list(1:ns)
  for (i in 1:ns) {
    mean.vector <- mu(tlist[[i]], theta)
    var.matrix <- vmat(tlist[[i]], nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa)
    result[[i]] <- rmvnorm(1, mean.vector, var.matrix)
  }
}
```
```r
result
#

## code to generate appendix figure 7

theta <- c((66.02327 + 25.18973), (-0.26051+-0.72041))
nusqA <- 16.33969^2
nusqB <- 0
rhoAB <- 0.1664673*15.57728^2
tausq <- (1-0.1664673)*15.57728^2
phi <- 5.9045003
kappa <- 0.5

ns <- 4

samp <- sample(d$id, ns)
sampd <- subset(d, d$id %in% samp)
ids <- unique(sampd$id)
tlist <- as.list(1:ns)
for (i in 1:ns)
  { tlist[[i]] <- sampd$age[sampd$id == ids[i]] }
x <- simulate.unbalanced(ns, tlist, theta, nusqA, nusqB, rhoAB, tausq, sigmasq, phi, kappa = 0.5)

par(mfrow=c(1,2))

plot(tlist[[1]], x[[1]], type = "l", ylim = c(0,120), xlab = "Age", main = "MODEL", ylab = "% Predicted FEV1")
for (i in 2:ns)
  lines(tlist[[i]], x[[i]], col = i)

plot(sampd$age[sampd$id == ids[1]], sampd$fev1[sampd$id == ids[1]], type = "l", ylim = c(0,120), xlab = "Age", main = "REAL DATA", ylab = "% Predicted FEV1")
for (i in 2:ns)
  lines(sampd$age[sampd$id == ids[i]], sampd$fev1[sampd$id == ids[i]], col = i)
```