Longitudinal Project: EDA & Project Conception

Brenton Graham

2022-11-18

Background

Pulmonary exacerbations (PEx) are a leading cause of morbidity in cystic fibrosis (CF). Treatment response is often suboptimal despite seemingly appropriate antimicrobial therapy. In this study, we seek to determine changes in airway microbiome and clinical response with onset and treatment of a PEx in children and adolescents with CF. Participants hospitalized for PEx were evaluated at admission, hospital discharge and a follow-up clinic visit. Sputum and blood samples were collected along with lung function and PEx score. Quantitative CF cultures were performed. Sputum and plasma were analyzed for inflammatory and protein markers.

Data

Subjects & Repeated Measures

Data from 35 unique subjects and 40 pulmonary exacerbations are present in this data set (i.e., five subjects had exacerbations on two separate occasions). For each exacerbation there are three repeated measures (T1 = Admission, T2 = Hospital Discharge, T3 = Follow-Up Visit). Time points are approximately separated by 10-14 days.

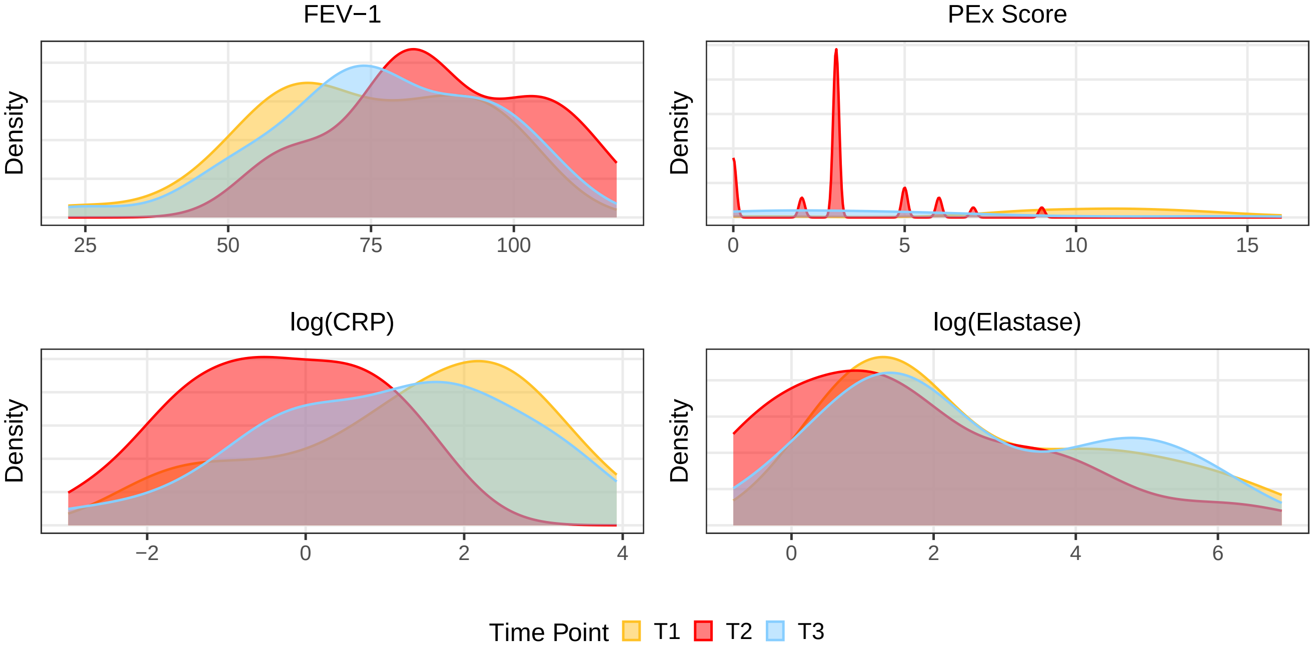
Potential Outcomes

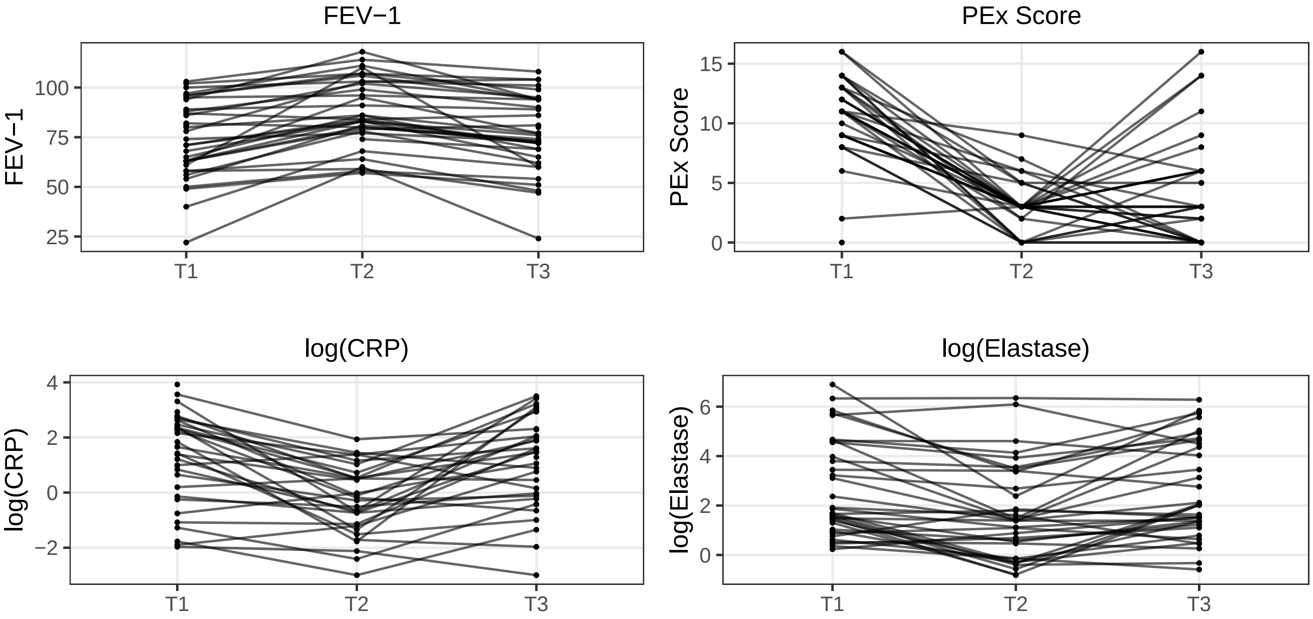
We're interested in modeling several potential response outcomes over time:

- fev1_pred: FEV1, a measure of lung function (the higher the better)
- pes_total: PEx score (the lower the better)
- crp: C-reactive protein, a measure of inflammation (the lower the better)
- elastase: An enzyme marker

Table 1: Demographics of study population.

	Overall
	(N=35)
gender	
Male	17 (48.6%)
Female	18 (51.4%)
age	
Mean (SD)	16.4 (2.77)
Median [Min, Max]	16.0 [12.0, 22.0]
bmi	
Mean (SD)	20.0 (2.40)
Median [Min, Max]	19.5 [16.4, 28.7]
genotype	
0 F508del	4 (11.4%)
1 F508del	9 (25.7%)
2 F508del	22 (62.9%)
admit.cf_pathogens	
0	3 (8.6%)
1	19 (54.3%)
2	10 (28.6%)
3	2 (5.7%)
Missing	1 (2.9%)
admit.coinfect	
0	21 (60.0%)
1	9 (25.7%)
Missing	5 (14.3%)


FEV−1
Density
25
50
75
100
PEx Score
Density
0
5
10
15
log(CRP)
Density
−2
0
2
4
log(Elastase)
Density
0
2
4
6
Time Point
T1
T2
T3


FEV−1
FEV−1
100
75
50
25
T1
T2
T3
PEx Score
PEx Score
15
10
5
0
T1
T2
T3
log(CRP)
log(CRP)
4
2
0
−2
T1
T2
T3
log(Elastase)
log(Elastase)
6
4
2
0
T1
T2
T3

Explanatory Variables

In this data set we have a collection of interesting explanatory variables. These include important CF pathogen measures (from both culture and 16S rRNA sequencing) and viral measures at admission. Well known CF pathogens include Pseudomonas aeruginosa, Staphylococcus aureus, Haemophilus influenzae, Stenotrophomonas maltophilia, Achromobacter xylosoxidans and Burkholderia cepacia.

In general, we'd like to get a sense of whether or not there is anything at admission that can be used to predict the subject's response to treatment. To simplify things, we will test two primary explanatory variables:

1. **Number of CF pathogens detected at admission.** This is the sum of binary detection results from culture. There are some studies suggesting that increased CF pathogens has negative impacts on treatment response.
2. **Bacterial/Viral co-infection at admission.** This is a binary variable that indicates whether or not **both** CF and viral pathogens were detected when the subject was admitted.

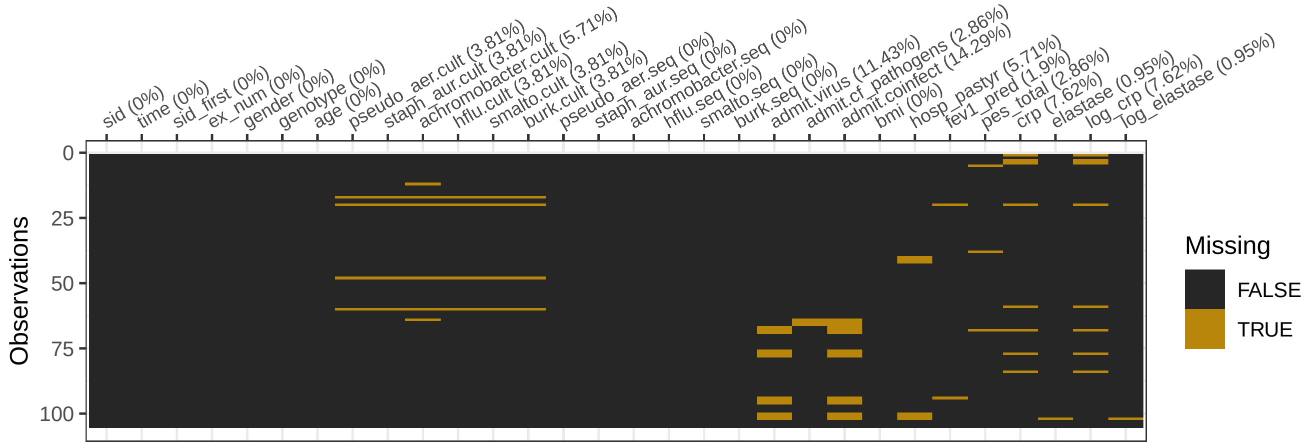
Additional Predictors/Precision Variables

The following variables will be included as covariates in all models.

- Gender
- Age
- BMI
- Number of hospitalizations in the past year
- Genotype: There is a specific genetic mutation of interest in CF patients. This variable indicates the number of F508del mutations

Missing Data

There are some missing data in both our outcome and explanatory variables (shown below).



Analysis

There aren't necessarily any obvious grouping variables in this data set (i.e., there aren't treatment groups for which we'd like to compare). However, we are interested in the trajectory of our outcomes and it would be biologically interesting to assess if certain factors are associated with distinct outcome trajectories. To accomplish this we can use a latent class mixture model to cluster subjects into latent classes based on outcome trajectories. Once classes are assigned, we can use post-hoc analysis methods to characterize risk factor profiles that might be driving the trajectory differences. Baseline clinical markers will be used in this latent class-separation analysis. This will allow us to determine if there are any baseline indicators of treatment response trajectory.

Latent Class Mixed Modeling

In the first step of the analysis we will fit a latent class mixed model. Based on discussions with investigators, the elastase marker might make for the most biologically interesting latent class results. It will be most practical and appropriate to treat time as a class variable for the following reasons. First, there are only three time points, so we do not need to be concerned about the number of parameters we incorporate in the model. Second, modeling time as a class variable allows for the most flexibility. From the spaghetti plots it is obvious that the effect of time on the outcomes is not linear; quadratic relationships likely exists between time and the outcomes. Modeling time as a class variable will allow us to account for this.

Looking at the data structure, we should account for within-subject correlation. We can also consider including a random effect for the time variable since each time point relates to different outcome expectations (i.e., at time point 1 the subject is experiencing an exacerbation; at time point 2 the subject is on antibiotics and symptoms/outcomes are expected to improve). In other words, there should be some between-subject correlation at each time point. This is something that I'm not entirely confident in, however, and need to spend more time thinking about.

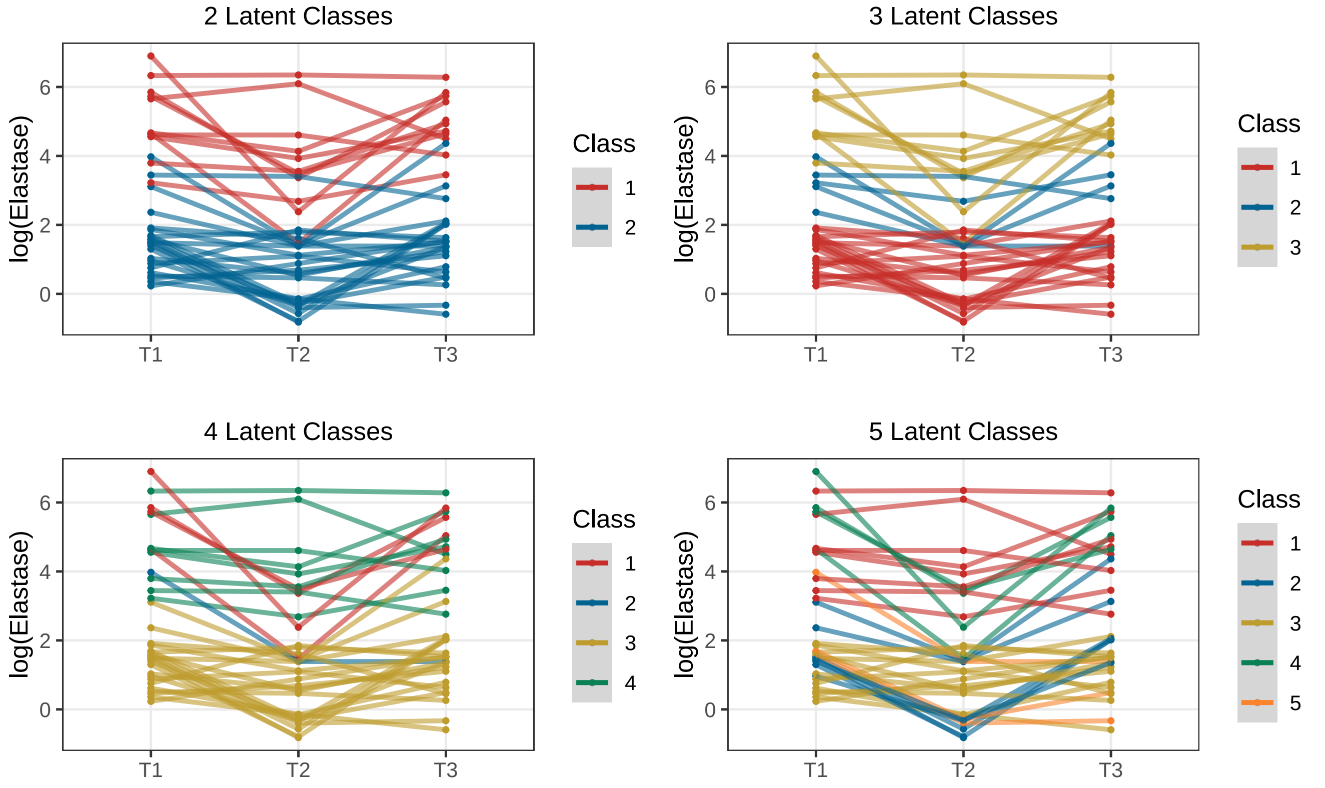
Latent Class Model Selection

This method requires users to specify the number of latent classes that should be created. As such, there is a model selection step. Below are results from models using 1-5 latent classes. Standard information criteria can be used to select the optimal number of latent classes.

```
##    G    loglik      AIC   %class1   %class2  %class3  %class4  %class5
## m1 1 -172.0657 358.1315 100.00000
## m2 2 -163.2205 348.4409  31.42857 68.571429
## m3 3 -158.4389 346.8777  57.14286 14.285714 28.57143
## m4 4 -157.1762 352.3524  11.42857  2.857143 62.85714 22.85714
## m5 5 -153.2515 352.5029  22.85714 20.000000 37.14286 11.42857 8.571429
```

Latent Class Visualization

Below we can see how these clusters look based on number of clusters and outcome trajectory.



Mixed Model Results

Below is an example of the results that can be viewed for each model. I still need to look into which number of classes makes the most sense. We can use the table above, which displays AIC, BIC, and the number of subjects assigned to each latent class, to decide on this. I also need to figure out how to display random effects estimates.