Main project: Bayesian inference with Expectation Maximisation for the characterisation of antibiotic treatment recovery in Cystic Fibrosis

Side project: Estimation of the variability in Cystic Fibrosis patient’s FEV1 lung function measurements

Master’s thesis of **Tristan Trébaol**

University of Cambridge
Department of Medicine

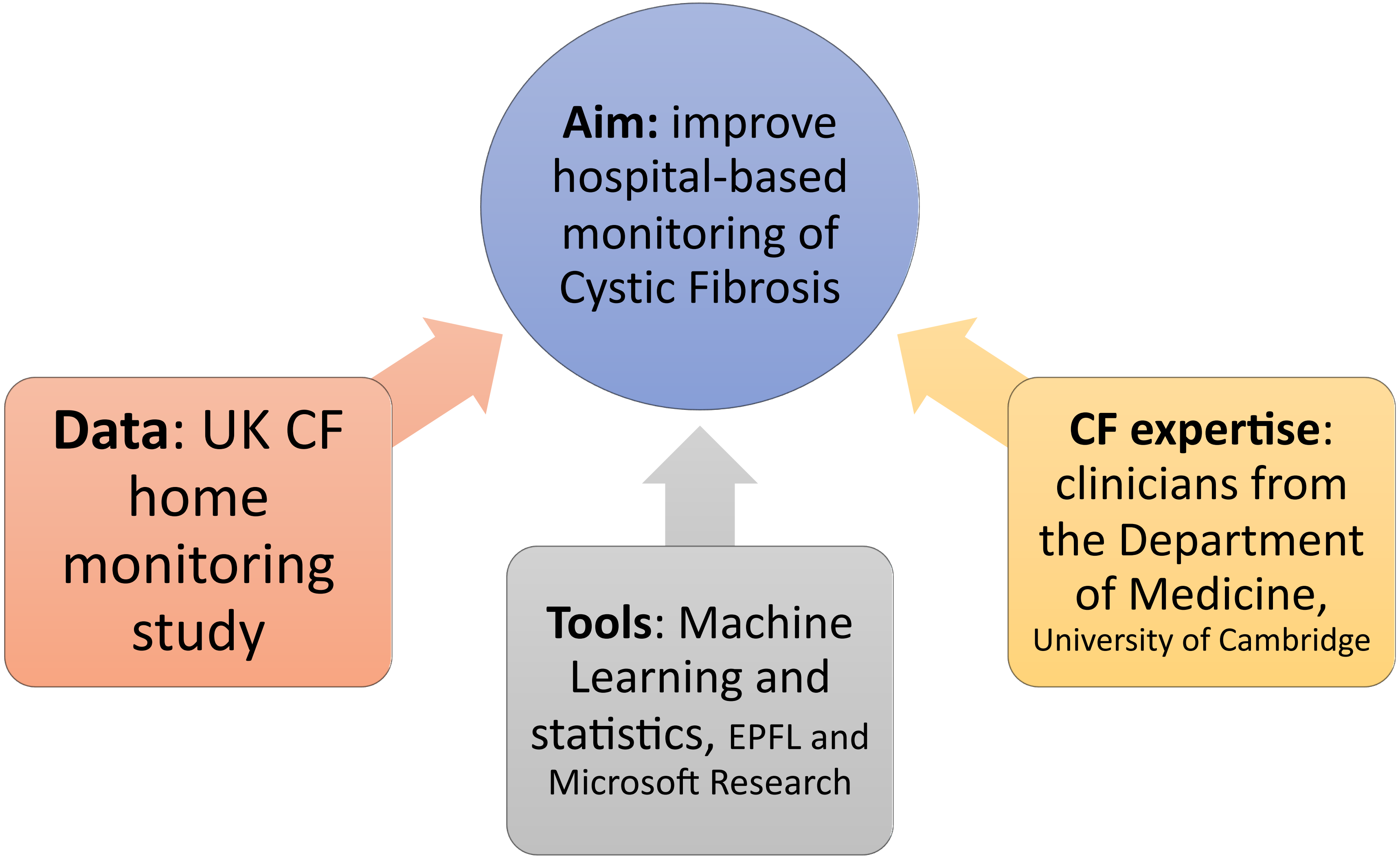
Academic hosts
- Prof. Andres Floto
- Ph.D. Damian Sutcliffe

External Expert
- John Winn, Machine Intelligence Group, Microsoft Research

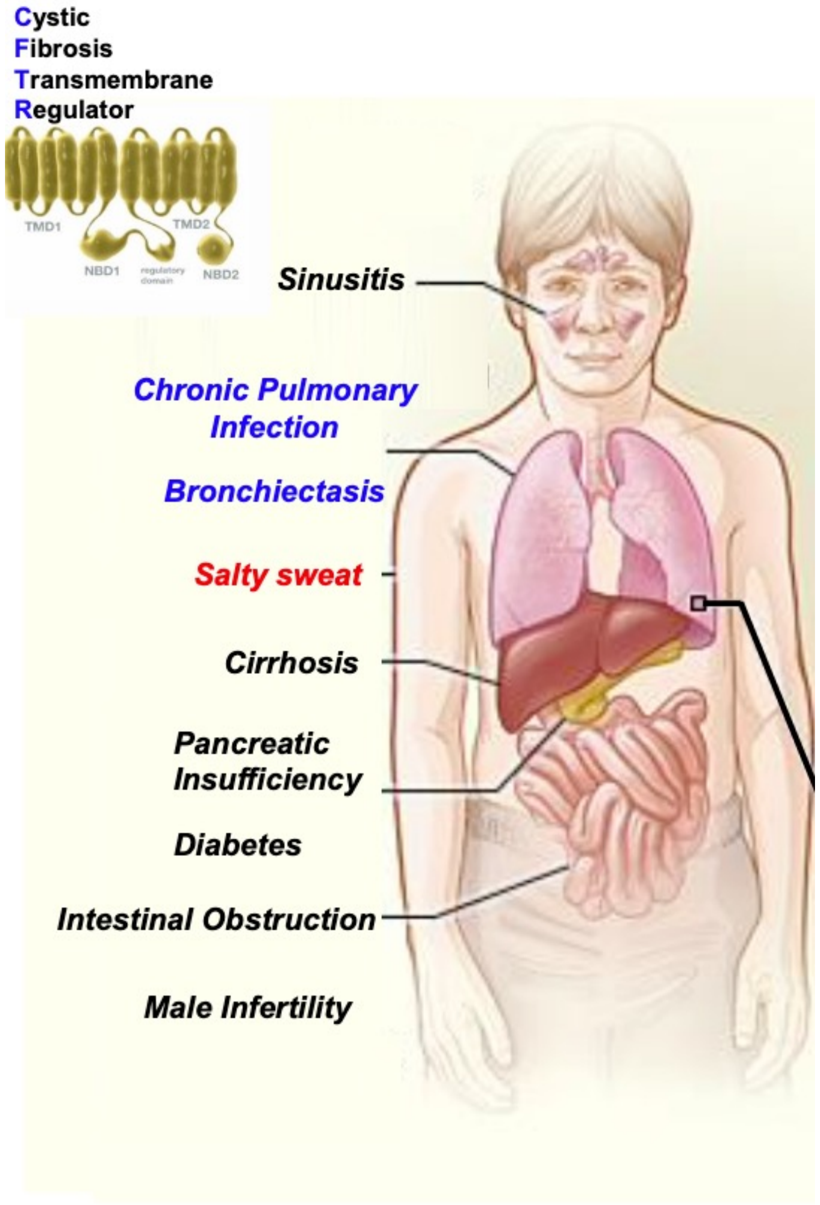
École Polytechnique Fédérale de Lausanne
School of Engineering, School of Computer and Communication Sciences

Supervised by
- Prof. Philippe Müllhaupt, School of Engineering
- Prof. Martin Jaggi, Machine Learning and Optimization Laboratory
- Dr. Mary-Anne Hartley, Machine Learning and Optimization Laboratory

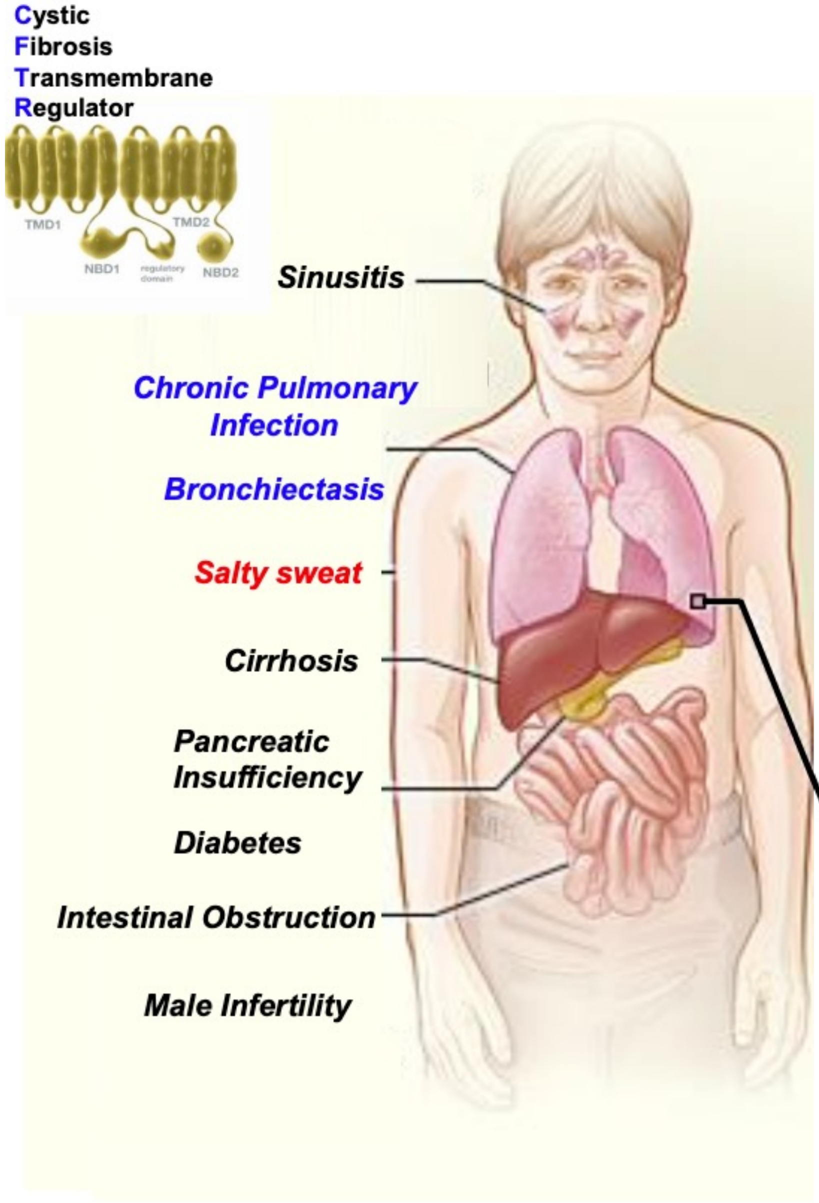
26.08.2021


Aim: improve hospital-based monitoring of Cystic Fibrosis
Data: UK CF home monitoring study
Tools: Machine Learning and statistics, EPFL and Microsoft Research
CF expertise: clinicians from the Department of Medicine, University of Cambridge

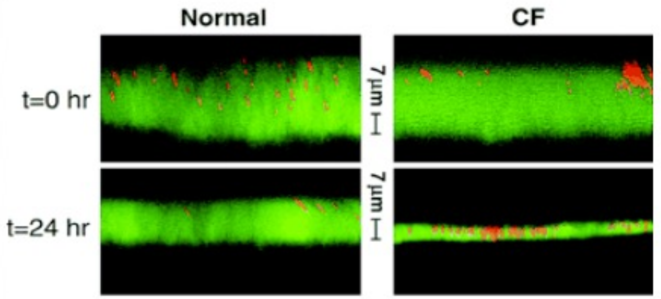
90’000 individuals worldwide



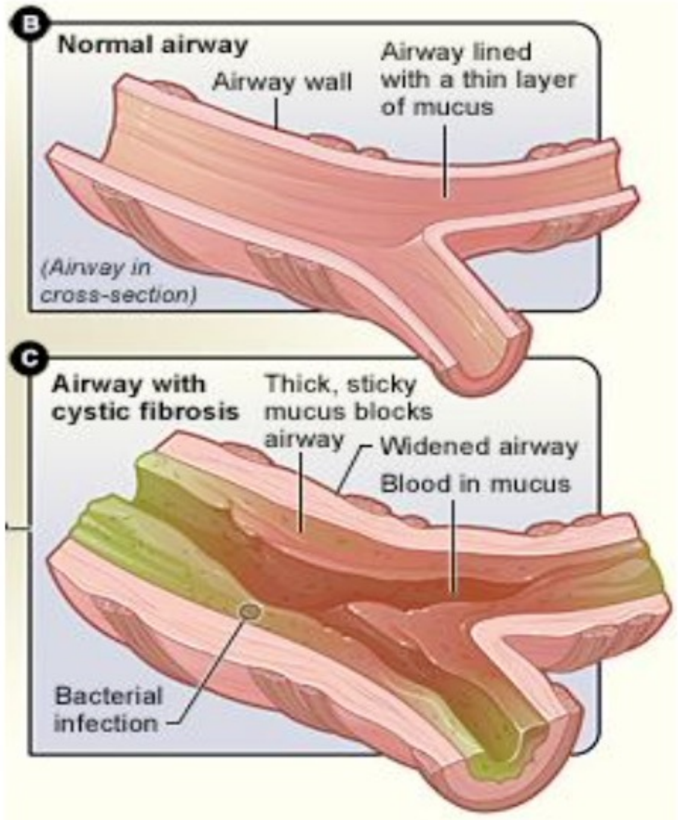
90'000 individuals worldwide



Failure to maintain airway surface liquid



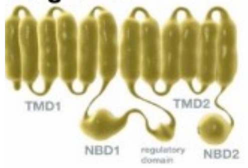
Matsui H *et al. Cell* 1998; 95: 1005-15

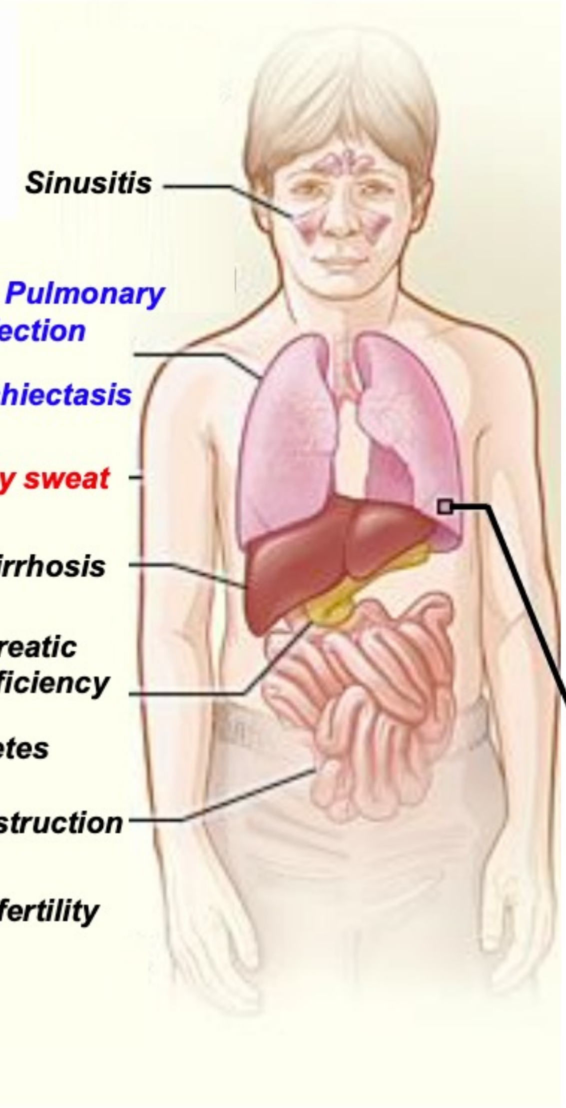


Cystic Fibrosis: pathophysiology

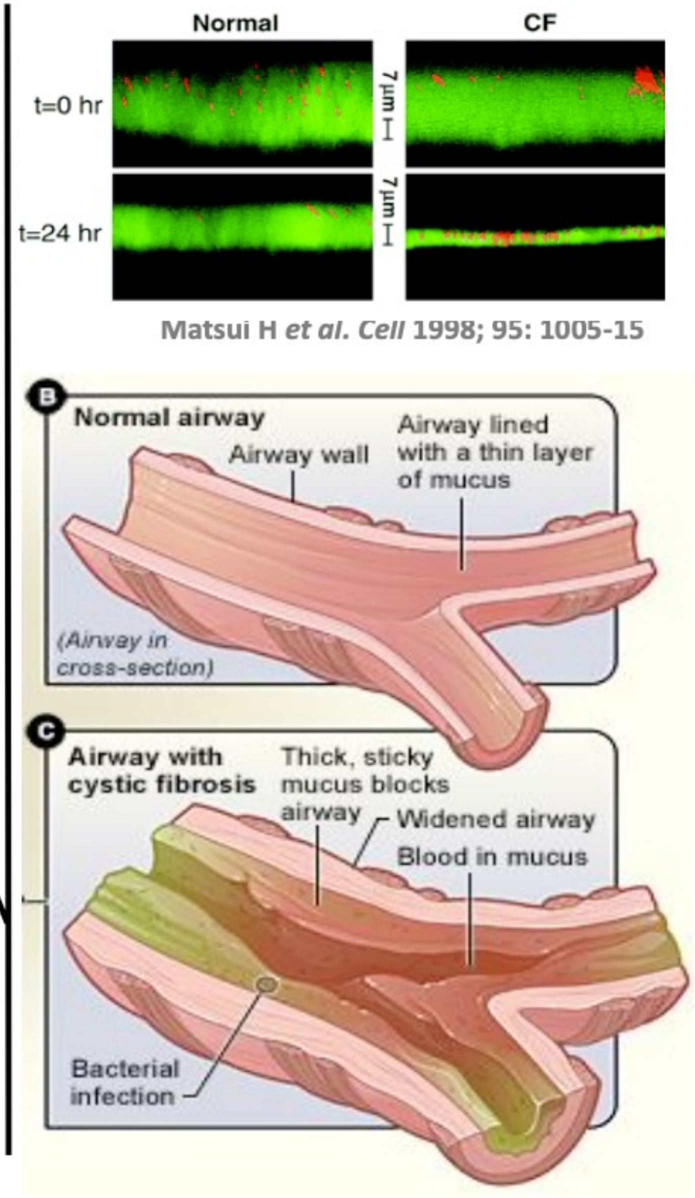
90’000 individuals worldwide

Cystic
Fibrosis
Transmembrane
Regulator

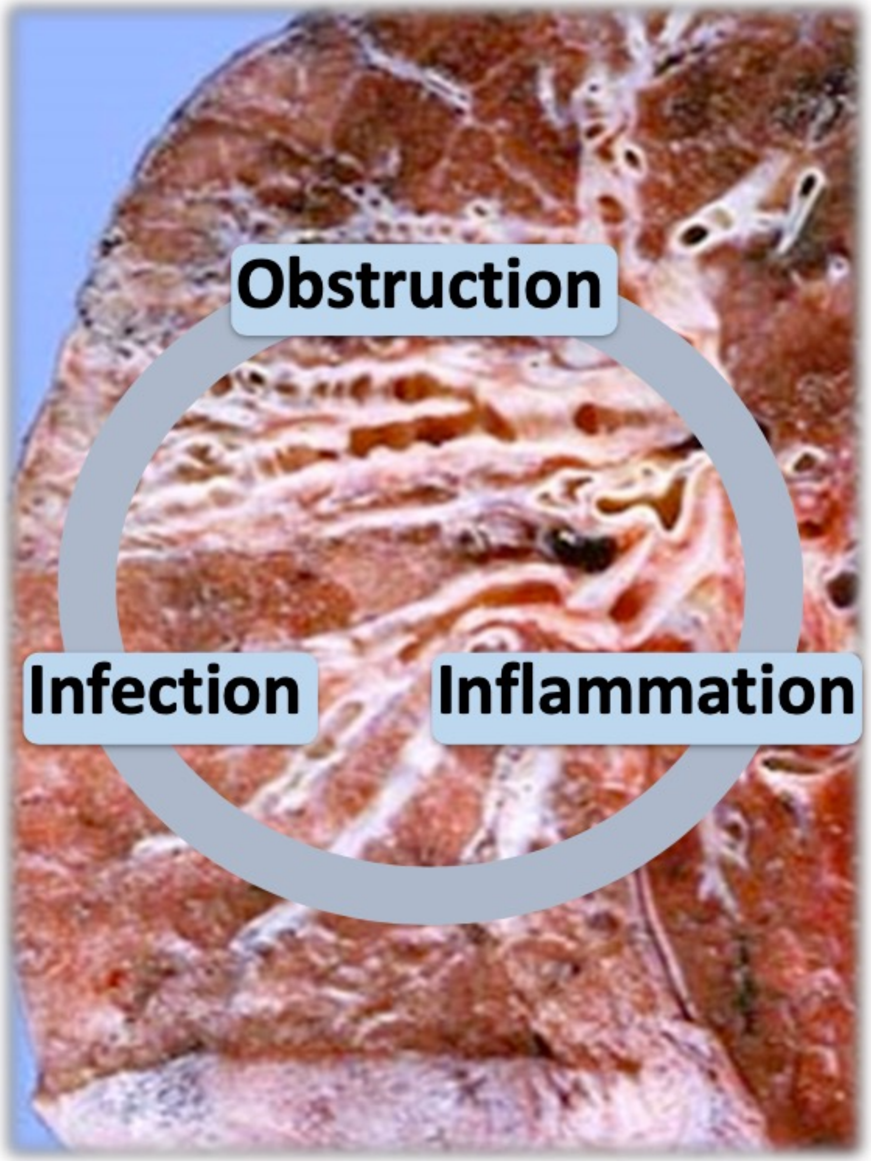




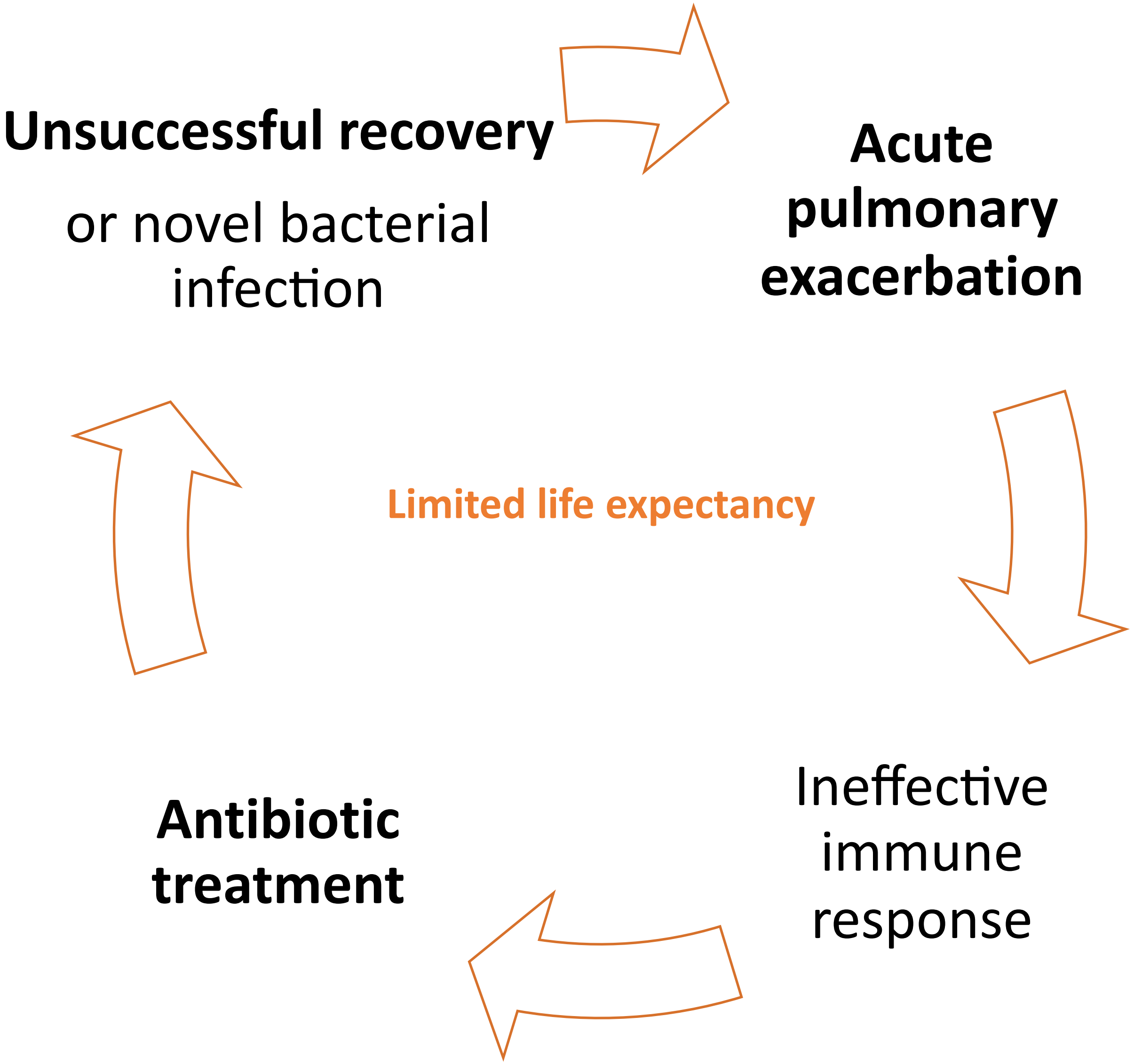
Failure to maintain airway surface liquid



Progressive Inflammatory Lung Damage

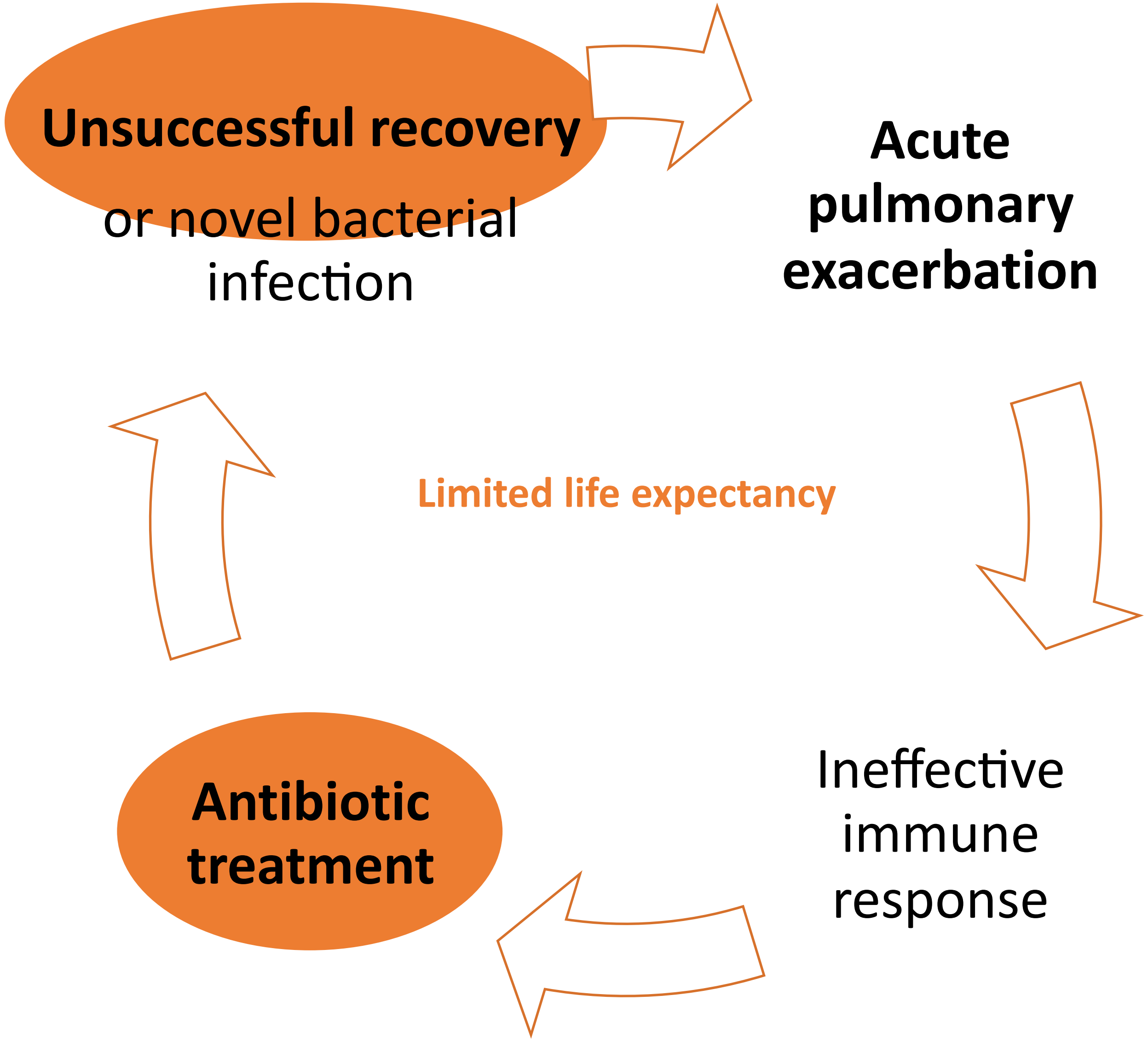


Cystic Fibrosis: pathophysiology



Symptoms:

- **Cough**
- **Fatigue**
- **Lung function diminution**
- **Decrease in oxygen saturation**
- **Breathlessness**
- **Increased volume of purulent sputum**



Symptoms:

- **Cough**
- **Fatigue**
- **Lung function diminution**
- **Decrease in oxygen saturation**
- **Breathlessness**
- **Increased volume of purulent sputum**

Project Breathe biggest CF data set in the world with daily measurements:

1. 2 hospitals in the UK
2. > 300 patients
3. Multidimensional
4. Automatic recording
5. > 500k recordings
6. Ongoing

Smartwatch:

- Calories
- Resting heart rate
- Sleep

Self-reported symptom diary:

- Wellness
- Cough
- Temperature

Spirometer + Pulse Oximeter:

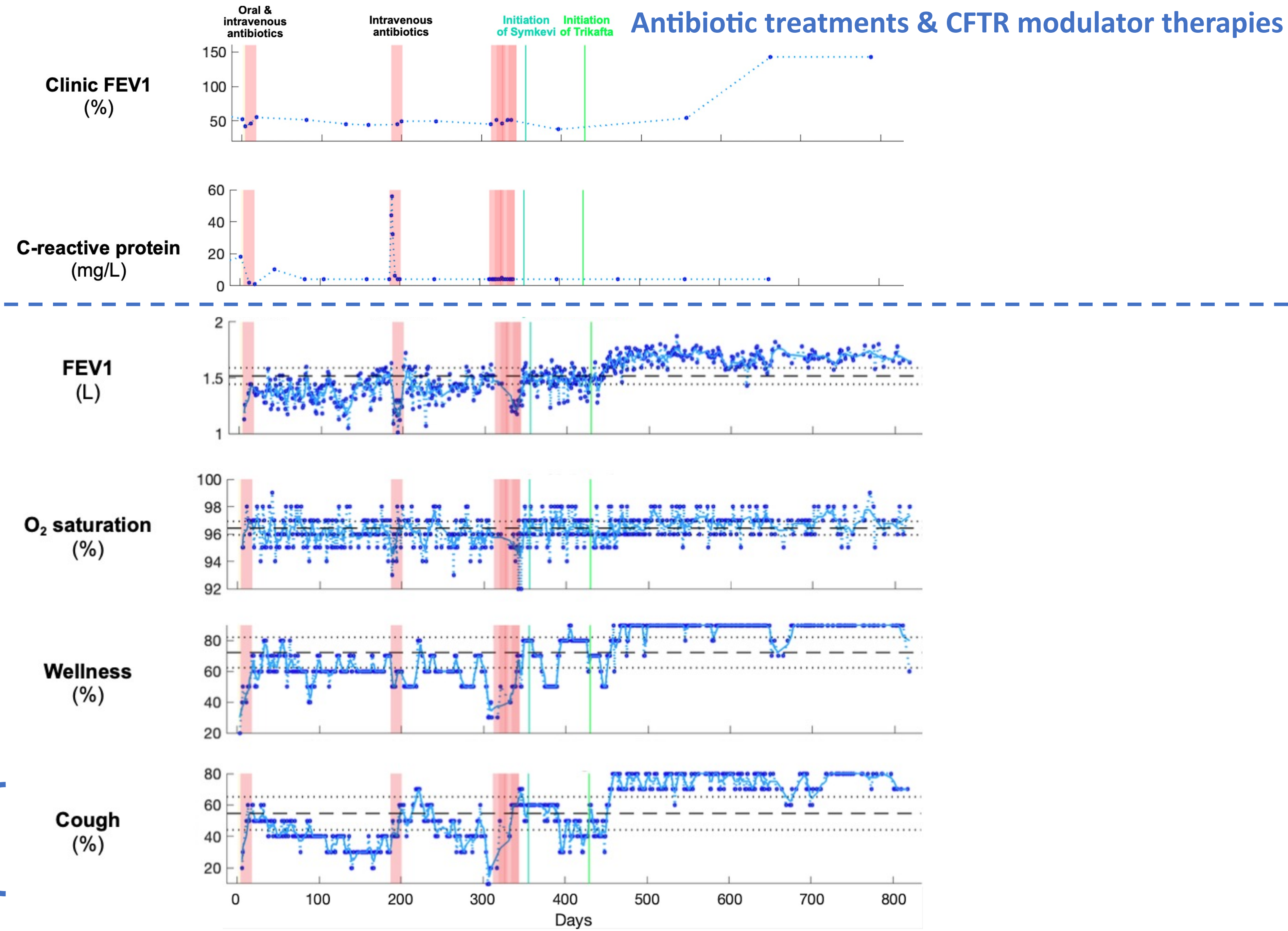
- Lung function (FEV_1, FEF_{2575}, …)
- O_2 saturation
- Pulse rate

Weight

Clinical data:
Bimonthly recordings

Home monitoring data:
Measurements ~3 times per week

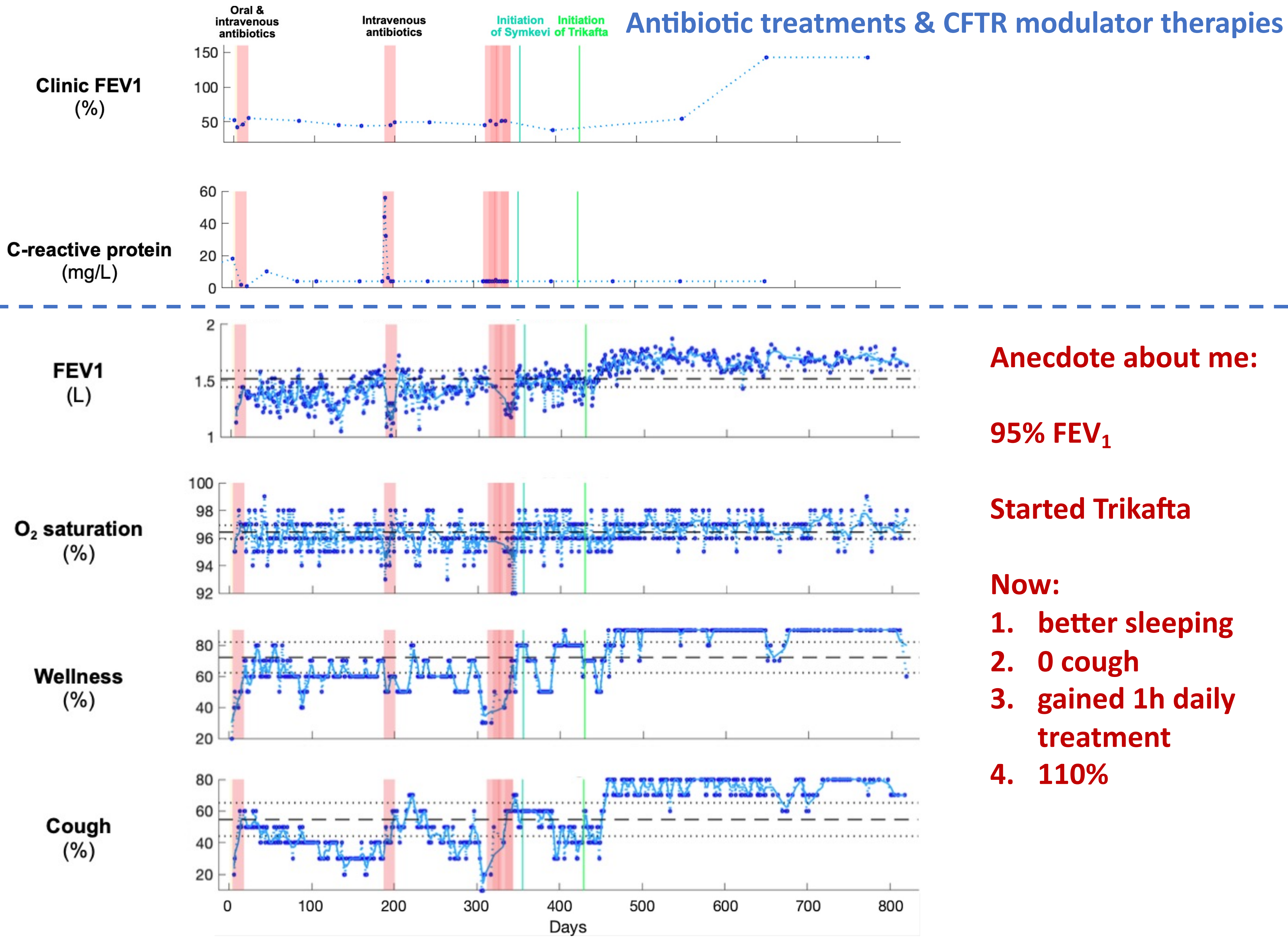
Inverted signals: cough reduction!



Clinical data:
Bimonthly recordings

Home monitoring data:
Measurements ~3 times per week

Inverted signals: cough reduction!



Anecdote about me:

95% FEV_1

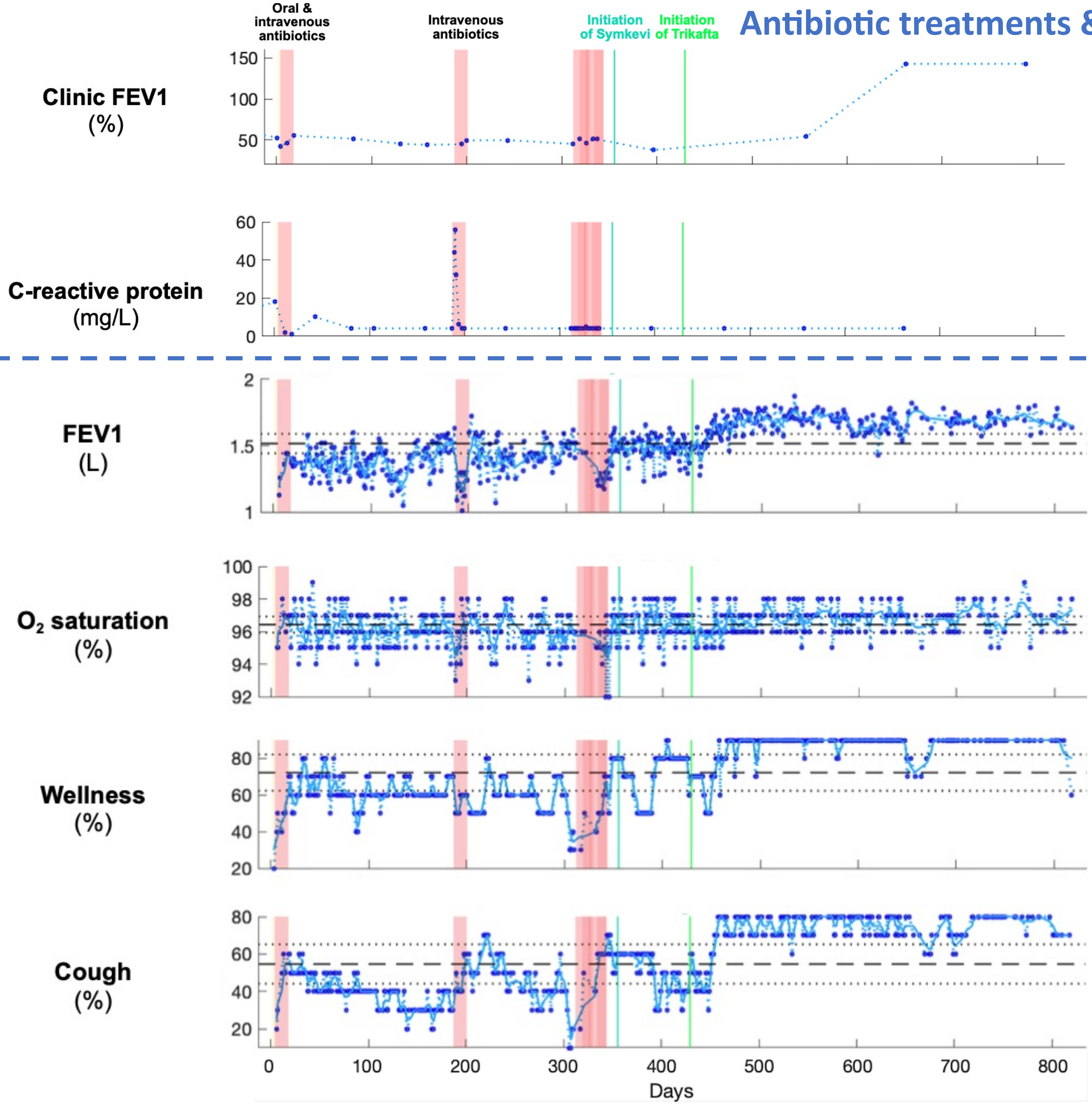
Started Trikafta

Now:

1. better sleeping
2. 0 cough
3. gained 1h daily treatment
4. 110%

Antibiotic treatments & CFTR modulator therapies

Clinical data:
Bimonthly recordings



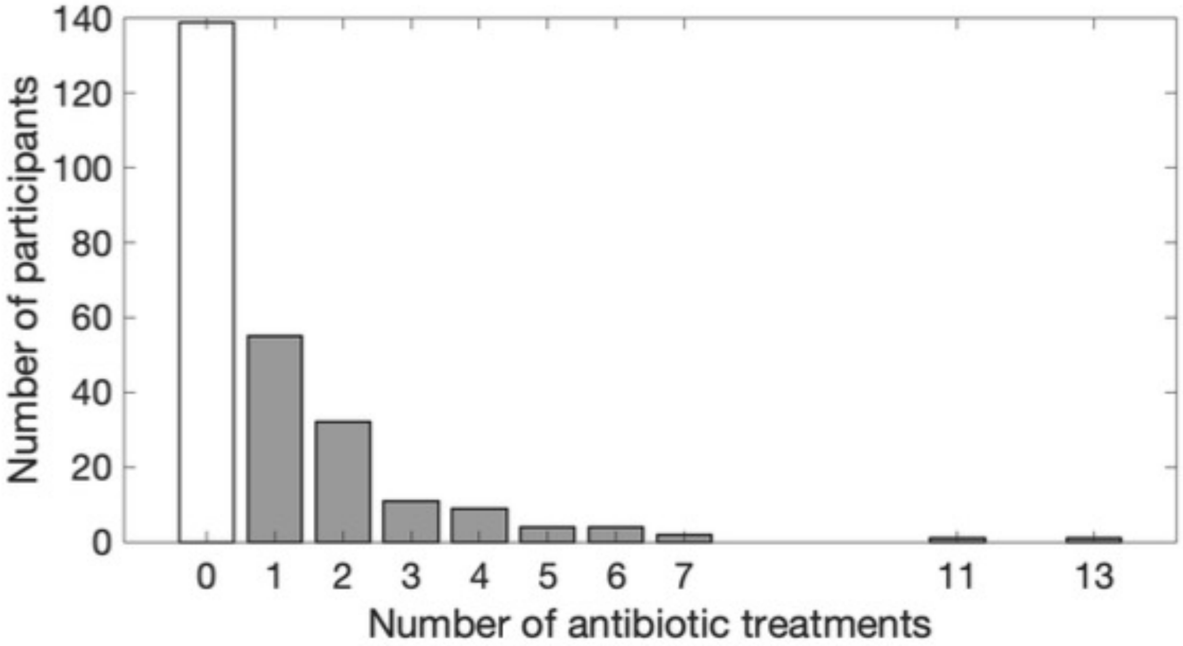
Home monitoring data:
Measurements ~3 times per week

Richness => ML

Inverted signals: cough reduction!

Characteristic	Value (N = 258)
Female	133 (52%)
Age (yr)	31.6 ±9.7
BMI (kg/m²)	23.3 ± 3.1
FEV1 (% of predicted)	69.0 ± 22.5
Sub-grouping	
< 40%	20 (8%)
>= 40% to < 70%	117 (47%)
>= 70% to < 90%	70 (28%)
>= 90%	41 (16%)
Genotype	
F508del homozygous	130 (50%)
F508del heterozygous	105 (41%)
Other	23 (9%)
Prescribed CFTR Modulators	
Triple Therapy	178 (69%)
Symkevi	121 (47%)
Ivacaftor	20 (8%)
Okrambi	6 (2%)

Computations of 1) Female, Age, BMI, Genotype, P. CFTR M. with clinical data, 2) FEV1 with home monitoring or clinical data. P. CFTR M. concern any period within the study.



Ask clinical questions

- How does a recovery look like?
- Are there different types of recoveries?

Answer them with Machine Learning

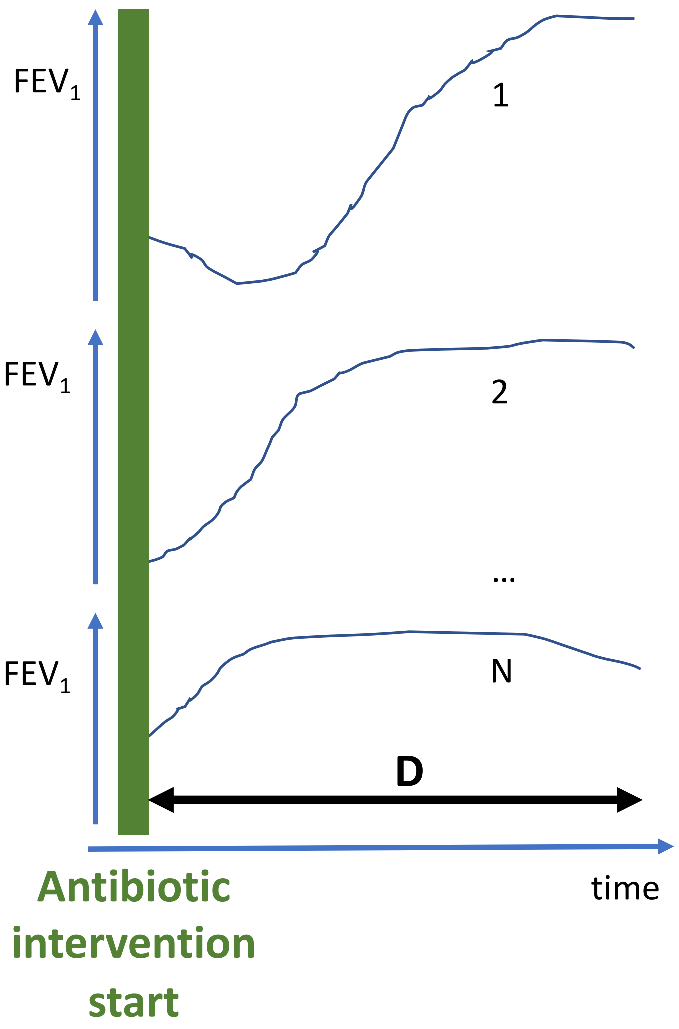
- Bayesian inference algorithm
- Convergence with expectation maximisation

Provide interpretable results to clinicians

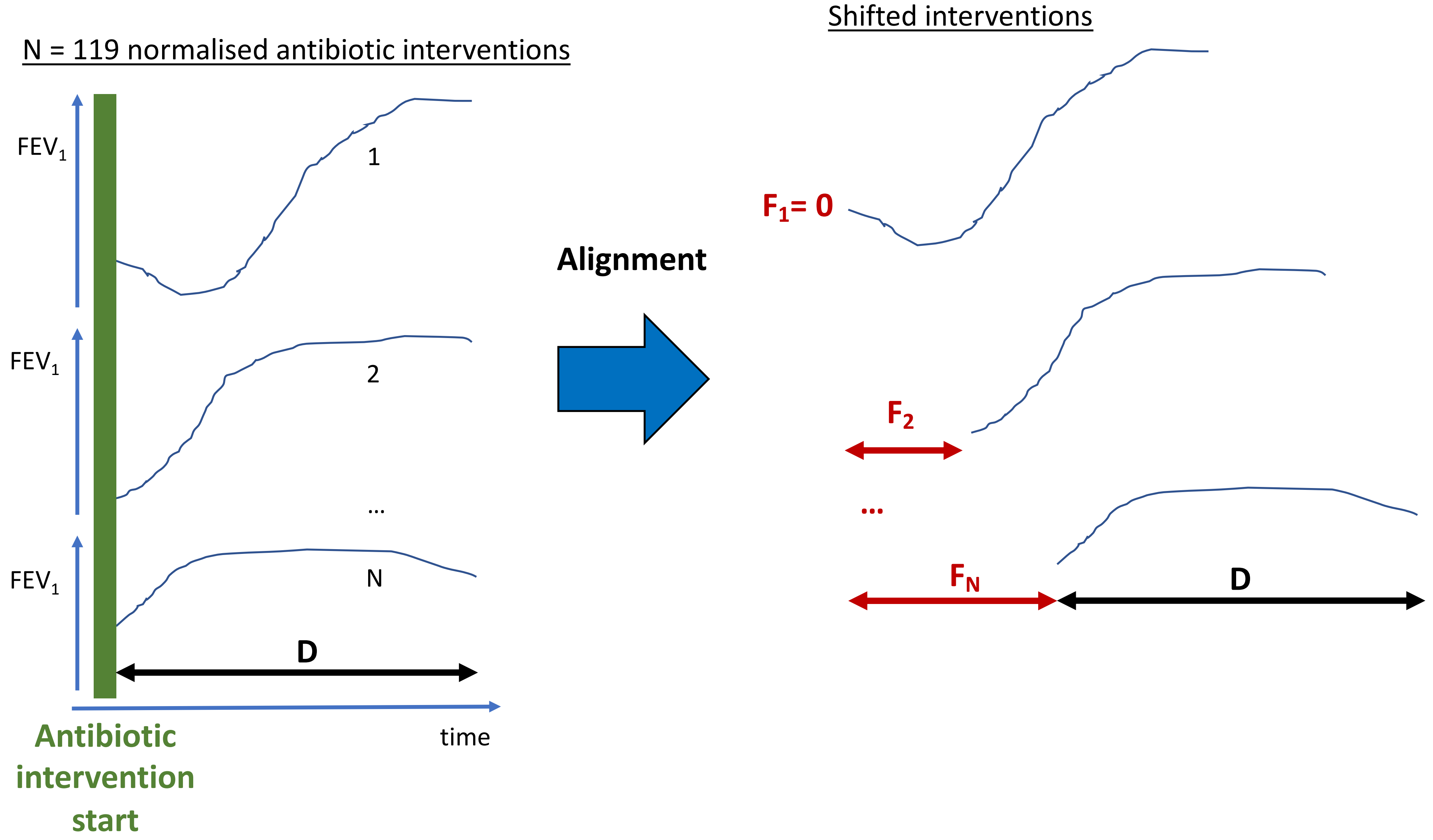
Clinicians feedback:
- Royal Papworth Hospital (UK)
- Hôpital Necker (France)
- Arizona Health Sciences Center (US)

Alignment model: samples

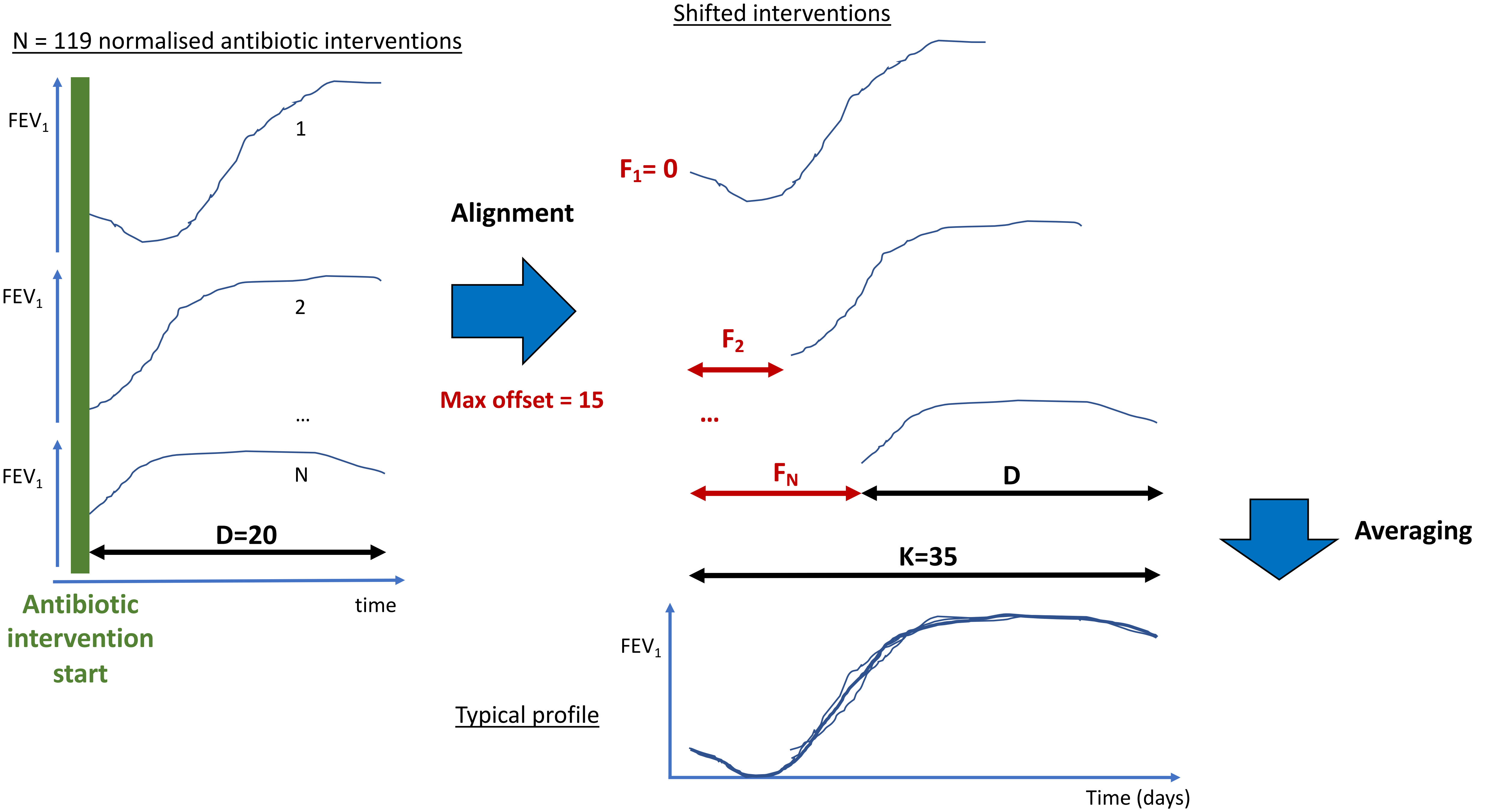
N = 119 normalised antibiotic interventions

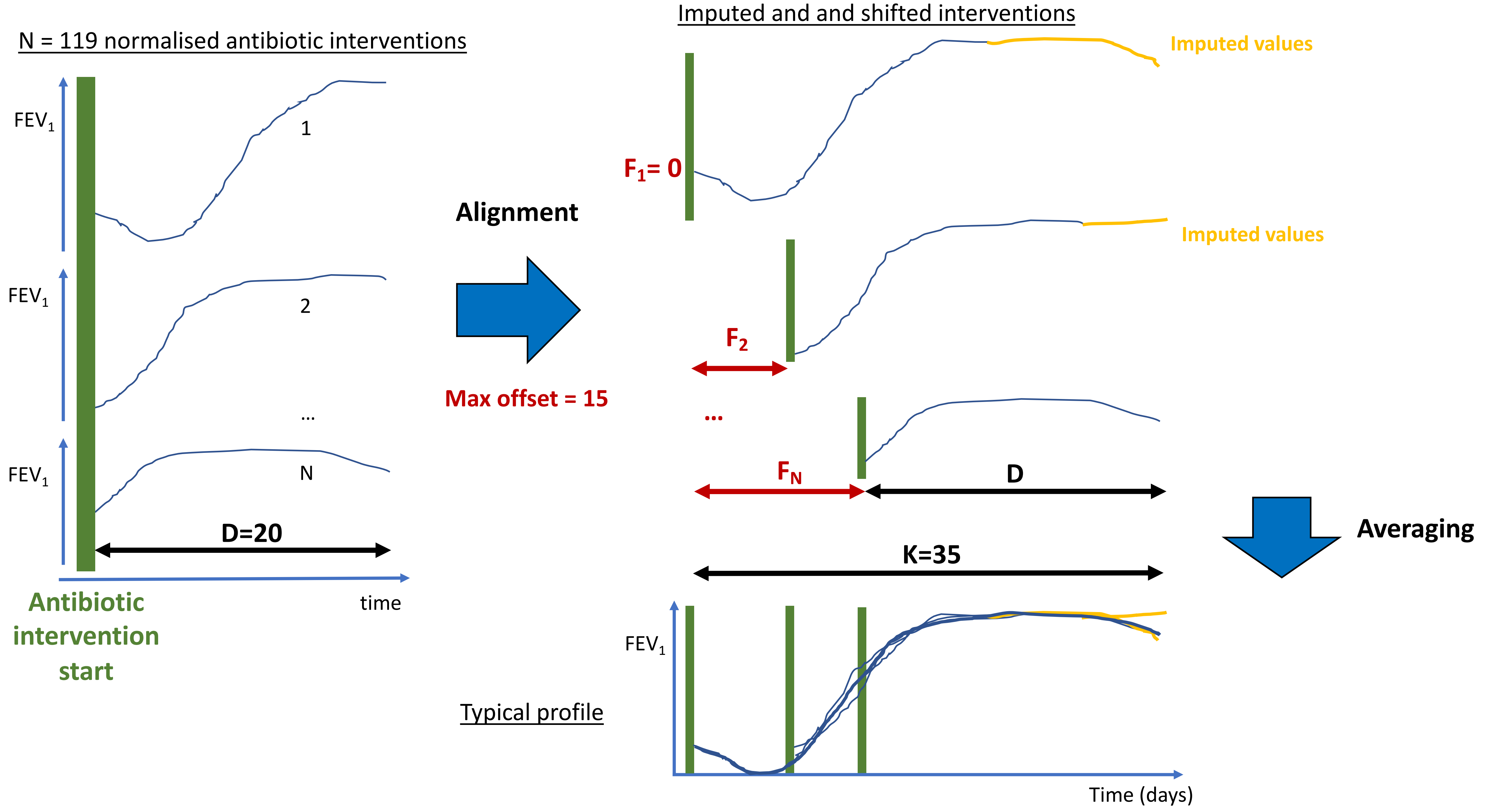


Alignment model: alignment

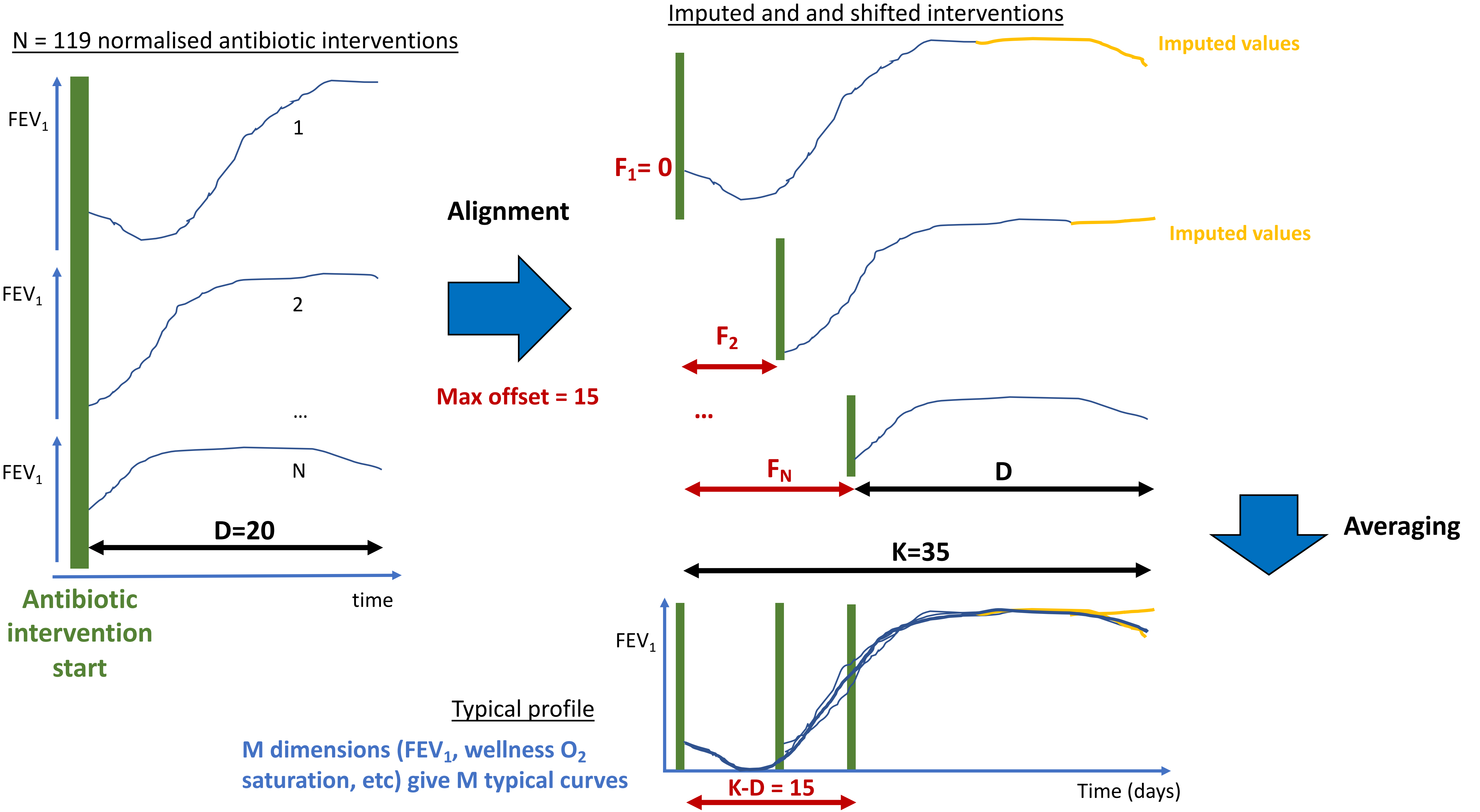


Alignment model: averaging




N = 119 normalised antibiotic interventions
FEV1
1
FEV1
2
...
FEV1
N
D=20
Antibiotic intervention start
time
Alignment
Max offset = 15
Imputed and and shifted interventions
Imputed values
F1= 0
Imputed values
F2
...
FN
D
Averaging
K=35
FEV1
Typical profile
Time (days)

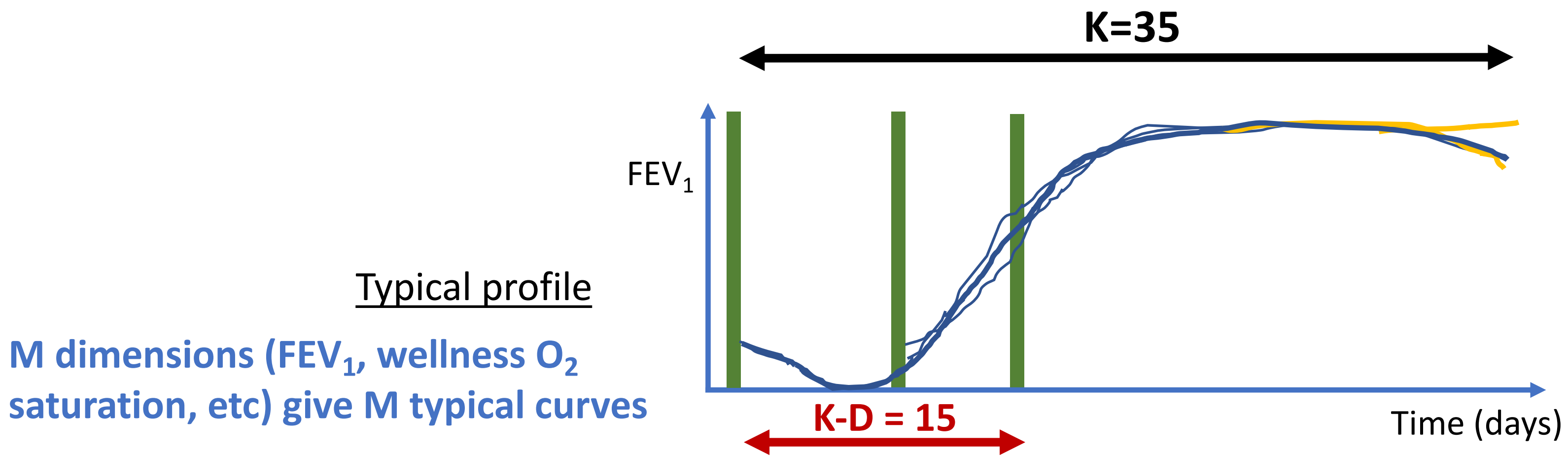
Alignment model: multidimensionality

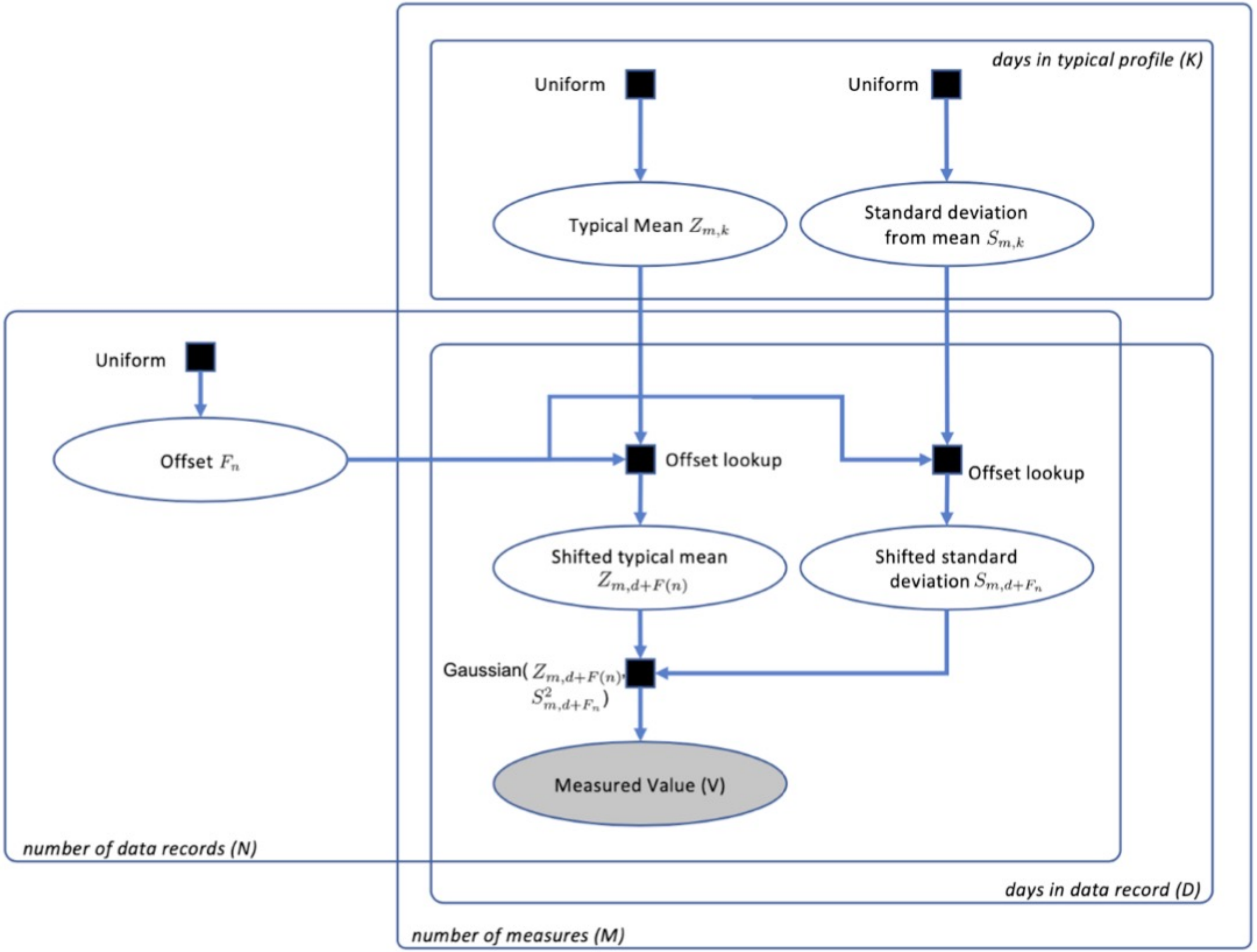


1. For each measure, the recorded values for the period immediately following treatment are a noisy version of a single typical profile.

2. The amount a measured value deviates from this profile is controlled by the position on the profile and is independent from one day to the next.

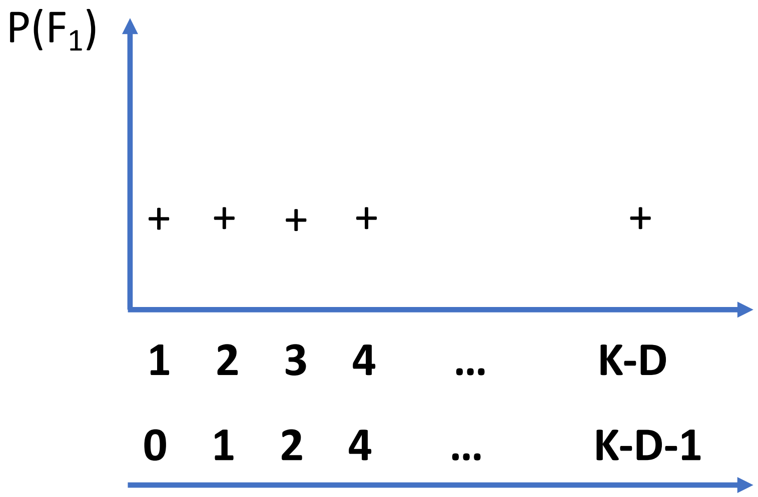
3. The treatment start can happen anytime between day 1 and the maximum allowed offset (K-D).




days in typical profile (K)
Uniform
Uniform
Typical Mean $Z_{m,k}$
Standard deviation from mean $S_{m,k}$
Uniform
Offset F_n
Offset lookup
Offset lookup
Shifted typical mean $Z_{m,d+F(n)}$
Shifted standard deviation $S_{m,d+F_n}$
Gaussian($Z_{m,d+F(n)}$, $S^2_{m,d+F_n}$)
Measured Value (V)
number of data records (N)
days in data record (D)
number of measures (M)

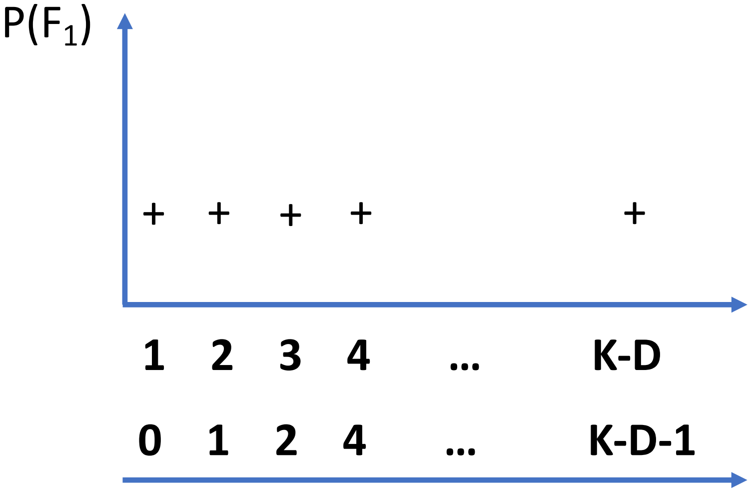
Machine learning model formulation

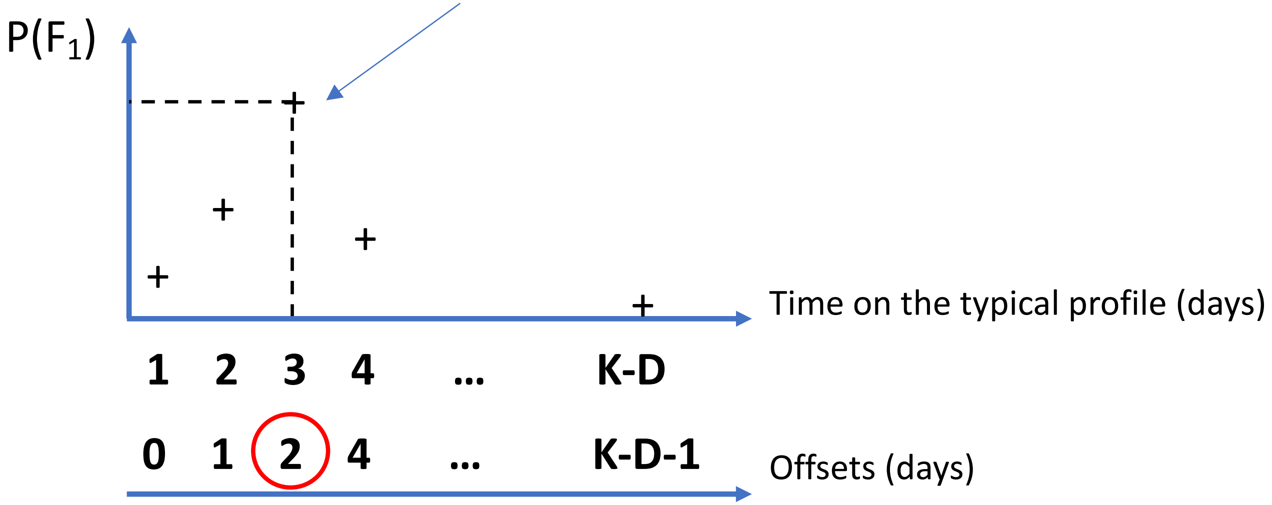
Initialisation: $P(F_N)$ is uniformly initialised over the offsets' span [0; K-D-1]



Initialisation: $P(F_N)$ is uniformly initialised over the offsets' span [0; K-D-1]

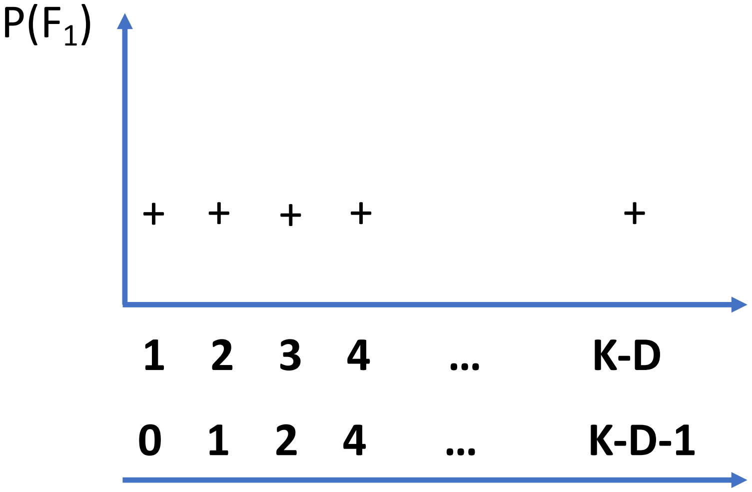
End-state: $$\mathcal{F} : [1;N] \longrightarrow [\overset{\text{no shift}}{0}; \underset{\text{max shift}}{K-D-1}], n \longmapsto F_n, F_n = \underset{[1;K-D]}{\arg\max}(P(F_n))) - 1$$

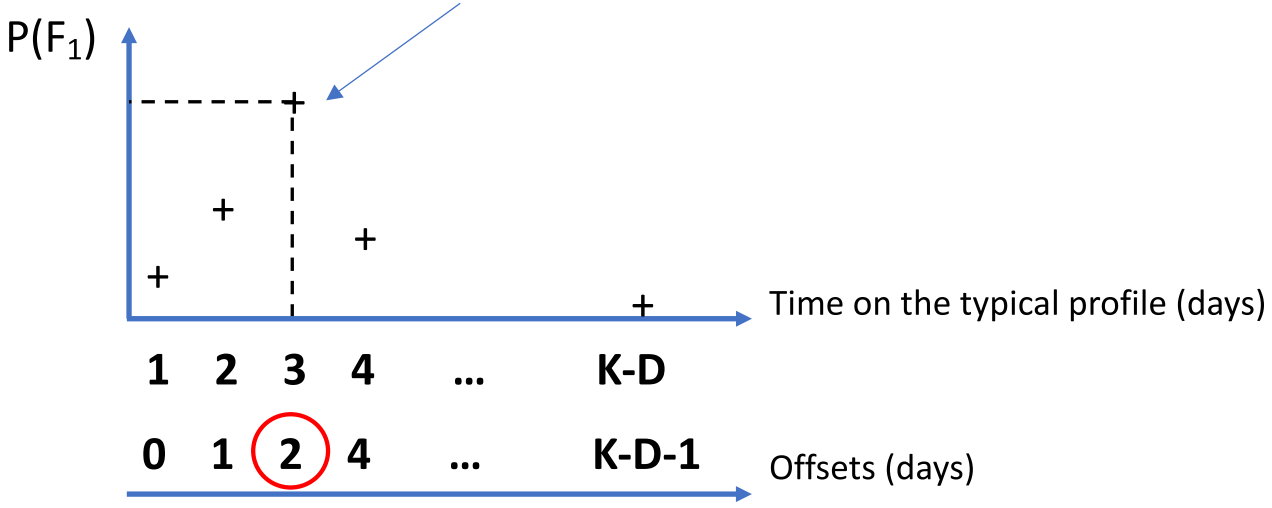




Initialisation: $P(F_N)$ is uniformly initialised over the offsets' span [0; K-D-1]

End-state: $$\mathcal{F} : [1;N] \longrightarrow [\overset{\text{no shift}}{0}; \underbrace{K-D-1}_{\text{max shift}}], n \longmapsto F_n, F_n = \underset{[1;K-D]}{\arg\max}(P(F_n))) - 1$$



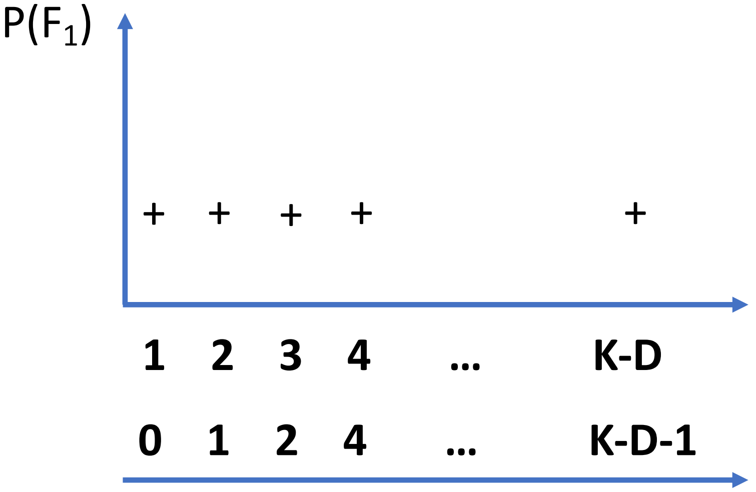


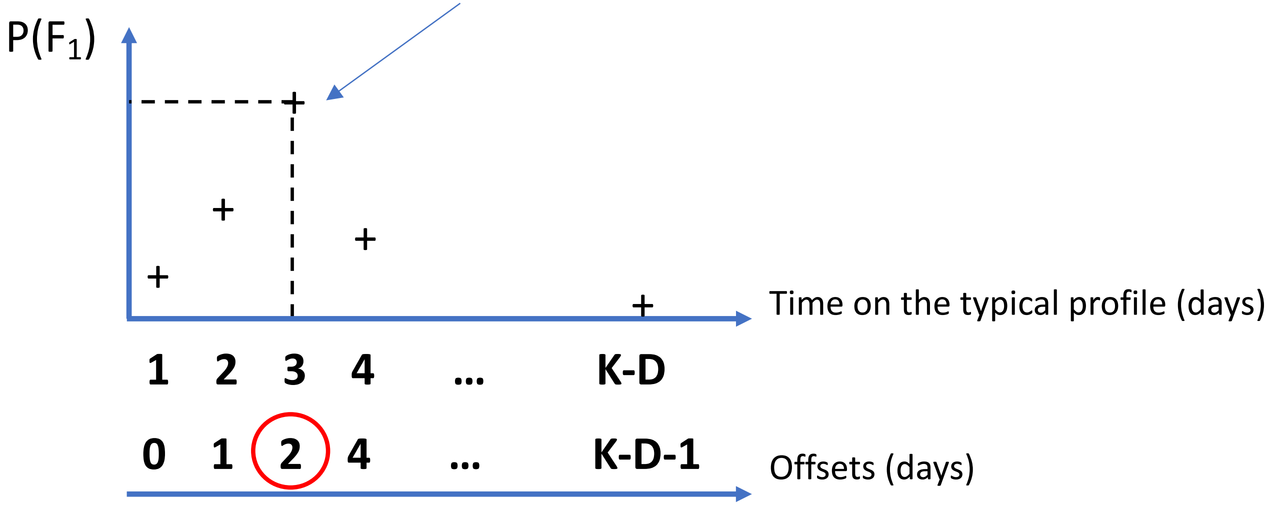
Averaging

$$\forall k \in [1;K], m \in [1;M], \begin{cases} Z'_{m,k} = \dfrac{\sum_{n=1}^{N}\sum_{f=0}^{K-D-1} V_{n,m,k-f}\cdot P(f)\cdot W(k,f)}{\sum_{f=0}^{K-D-1} P(f)\cdot W(k,f)} \\ \\ S'_{m,k} = \dfrac{\left(\sum_{n=1}^{N}\sum_{f=0}^{K-D-1} V_{n,m,k-f}\cdot P(f)\cdot W(k,f)\right)^2}{\sum_{f=0}^{K-D-1} P(f)\cdot W(k,f)} - Z'^2_{m,k} \end{cases}$$

Initialisation: $P(F_N)$ is uniformly initialised over the offsets' span [0; K-D-1]

End-state: $$\mathcal{F} : [1; N] \longrightarrow [\overset{\text{no shift}}{0}; \underset{\text{max shift}}{K - D - 1}], n \longmapsto F_n, F_n = \underset{[1;K-D]}{\arg\max}(P(F_n))) - 1$$





Expectation-step:
Update P(F$_n$)

$$\forall n \in N, Obj_n = \ln(P(F_n|V_n, Z, S)) = -\frac{1}{2} \cdot \sum_{m=1}^{M} \sum_{d=1}^{D} \left[A + \left(\frac{V_{n,m,d} - Z_{m,d+F_n} + B}{S_{m,d+F_n}} \right)^2 \right]$$

Maximisation-step:
Averaging

$$\forall k \in [1; K], m \in [1; M], \begin{cases} Z'_{m,k} = \frac{\sum_{n=1}^{N} \sum_{f=0}^{K-D-1} V_{n,m,k-f} \cdot P(f) \cdot W(k,f)}{\sum_{f=0}^{K-D-1} P(f) \cdot W(k,f)} \\ S'_{m,k} = \frac{\left(\sum_{n=1}^{N} \sum_{f=0}^{K-D-1} V_{n,m,k-f} \cdot P(f) \cdot W(k,f)\right)^2}{\sum_{f=0}^{K-D-1} P(f) \cdot W(k,f)} - Z'^2_{m,k} \end{cases}$$

Given a set of measured values V_n, Z and S fixed from the previous M-step:

$$P(F_n|V_n, Z, S) = \frac{P(V_n|F_n, Z, S) \cdot P(F_n)}{P(V_n, Z, S)}$$

$$P(F_n|V_n, Z, S) = \prod_{m=1}^{M} \prod_{d=1}^{D} P(V_{n,m,d}|F_n, Z_{m,d+F_n}, S_{m,d+F_n})$$

$$\forall n \in N, Obj_n = -\frac{1}{2} \cdot \sum_{m=1}^{M} \sum_{d=1}^{D} \left[A + \left(\frac{V_{n,m,d} - Z_{m,d+F_n} + B}{S_{m,d+F_n}} \right)^2 \right]$$

$$A = \ln(2 \cdot \pi) + 2 \cdot \ln(S_{m,d+F_n})$$

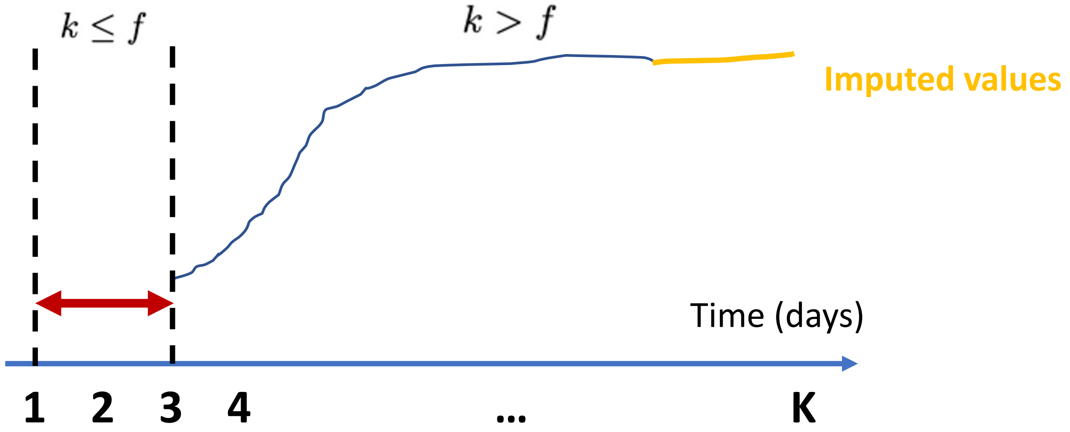
Vertical shift: $$B = \begin{cases} \min(\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_n}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_n} \geq 0 \\ \max(-\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_n}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_n} < 0 \end{cases}$$

Given a set of measured values V_n, and the offset probability distribution F_n fixed from the previous E-step:

$$\begin{cases} Z' = \arg\max_Z(P(F_n|Z,V_n)) \\ S' = \arg\max_S(P(F_n|S,V_n)) \end{cases}$$

$$\forall k \in [1;K], m \in [1;M], \begin{cases} Z'_{m,k} = \dfrac{\sum_{n=1}^{N}\sum_{f=0}^{K-D-1} V_{n,m,k-f}\cdot P(f)\cdot W(k,f)}{\sum_{f=0}^{K-D-1} P(f)\cdot W(k,f)} \\ \\ S'_{m,k} = \dfrac{\left(\sum_{n=1}^{N}\sum_{f=0}^{K-D-1} V_{n,m,k-f}\cdot P(f)\cdot W(k,f)\right)^2}{\sum_{f=0}^{K-D-1} P(f)\cdot W(k,f)} - Z'^2_{m,k} \end{cases}$$

Yes/no function: $$W(k,f) = \begin{cases} 1, & \text{for } k > f \\ 0, & \text{for } k \leq f \end{cases}$$



Updated E-step:

1) Compute the objective function for all classes and interventions.

$$\forall c \in [1;C], \forall n \in [1;N], Obj_{c,n} = -\frac{1}{2} \cdot \sum_{m=1}^{M} \sum_{d=1}^{D} \left[A + \left(\frac{V_{n,m,d} - Z_{c,m,d+F_{c,n}} + B}{S_{c,m,d+F_{c,n}}} \right)^2 \right],$$

where: $A = \ln(2 \cdot \pi) + 2 \cdot \ln(S_{m,d+F_{c,n}})$,

$$B = \begin{cases} \min(\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}} \geq 0 \\ \max(-\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}} < 0 \end{cases}$$

2) Assign the data record n to the best matching curve:

$$\boxed{\forall n \in [1;N], \psi_n = \arg\min_{[1;L]} Obj_{c,n}}$$

Updated E-step:

1) Compute the objective function for all classes and interventions.

$$\forall c \in [1;C], \forall n \in [1;N], Obj_{c,n} = -\frac{1}{2} \cdot \sum_{m=1}^{M} \sum_{d=1}^{D} \left[A + \left(\frac{V_{n,m,d} - Z_{c,m,d+F_{c,n}} + B}{S_{c,m,d+F_{c,n}}} \right)^2 \right],$$

where: $A = \ln(2 \cdot \pi) + 2 \cdot \ln(S_{m,d+F_{c,n}})$,

$$B = \begin{cases} \min(\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}} \geq 0 \\ \max(-\alpha, \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}}), & \text{for } \overline{V}_{n,m} - \overline{Z}_{m,F_{c,n}} < 0 \end{cases}$$

2) Assign the data record n to the best matching curve:

$$\boxed{\forall n \in [1;N], \psi_n = \underset{[1;L]}{\arg\min}\, Obj_{c,n}}$$

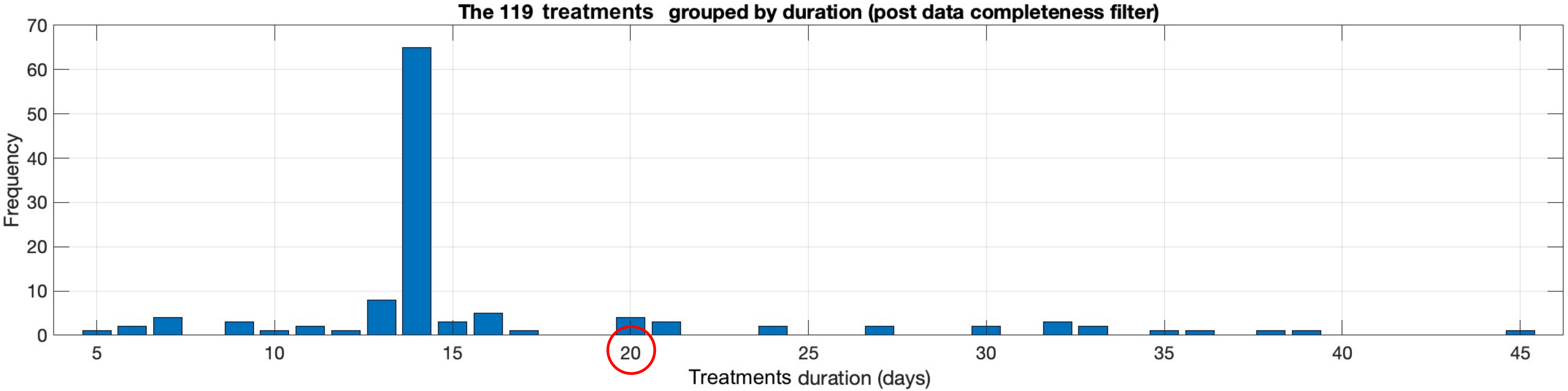
Updated M-step:

$$\forall k \in [1;K], m \in [1;M], \begin{cases} Z'_{c,m,k} = \Psi(c) \cdot \frac{\sum_{n=1}^{N} \sum_{f=0}^{K-D-1} V_{n,m,k-f} \cdot P(f) \cdot W(k,f)}{\sum_{f=0}^{K-D-1} P(f) \cdot W(k,f)} \\ S'_{c,m,k} = \Psi(c) \cdot \frac{\left(\sum_{n=1}^{N} \sum_{f=0}^{K-D-1} V_{n,m,k-f} \cdot P(f) \cdot W(k,f)\right)^2}{\sum_{f=0}^{K-D-1} P(f) \cdot W(k,f)} - Z'^2_{c,m,k} \end{cases}$$

where

$$\boxed{\Psi(c) = \begin{cases} 1, & \text{for } \psi_n = c \\ 0, & \text{for } \psi_n \neq c \end{cases}}$$

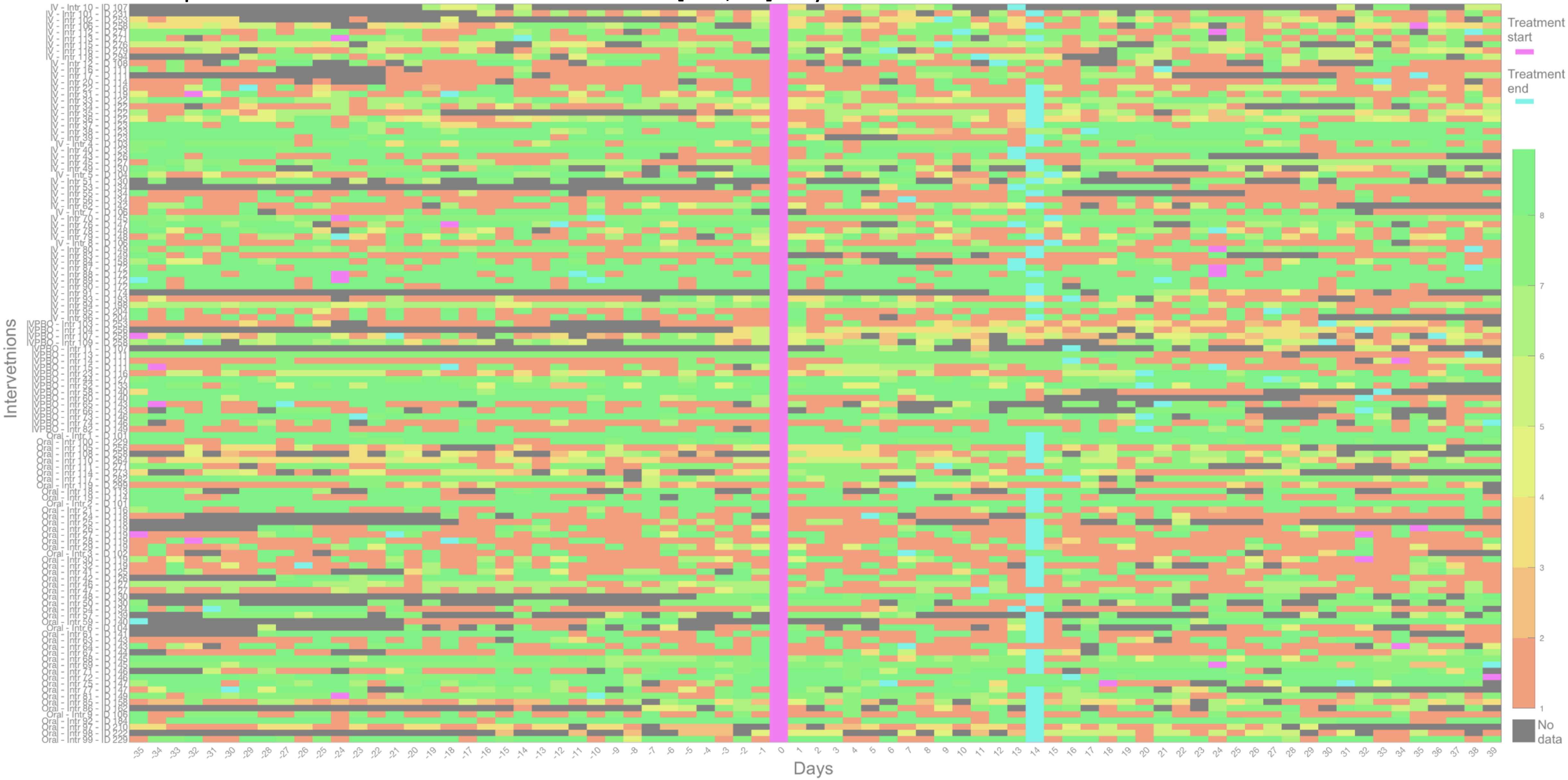
$$W(k,f) = \begin{cases} 1, & \text{for } k > f \\ 0, & \text{for } k \leq f \end{cases}$$


The 119 treatments grouped by duration (post data completeness filter)
Frequency
70
60
50
40
30
20
10
0
5
10
15
20
25
30
35
40
45
Treatments duration (days)

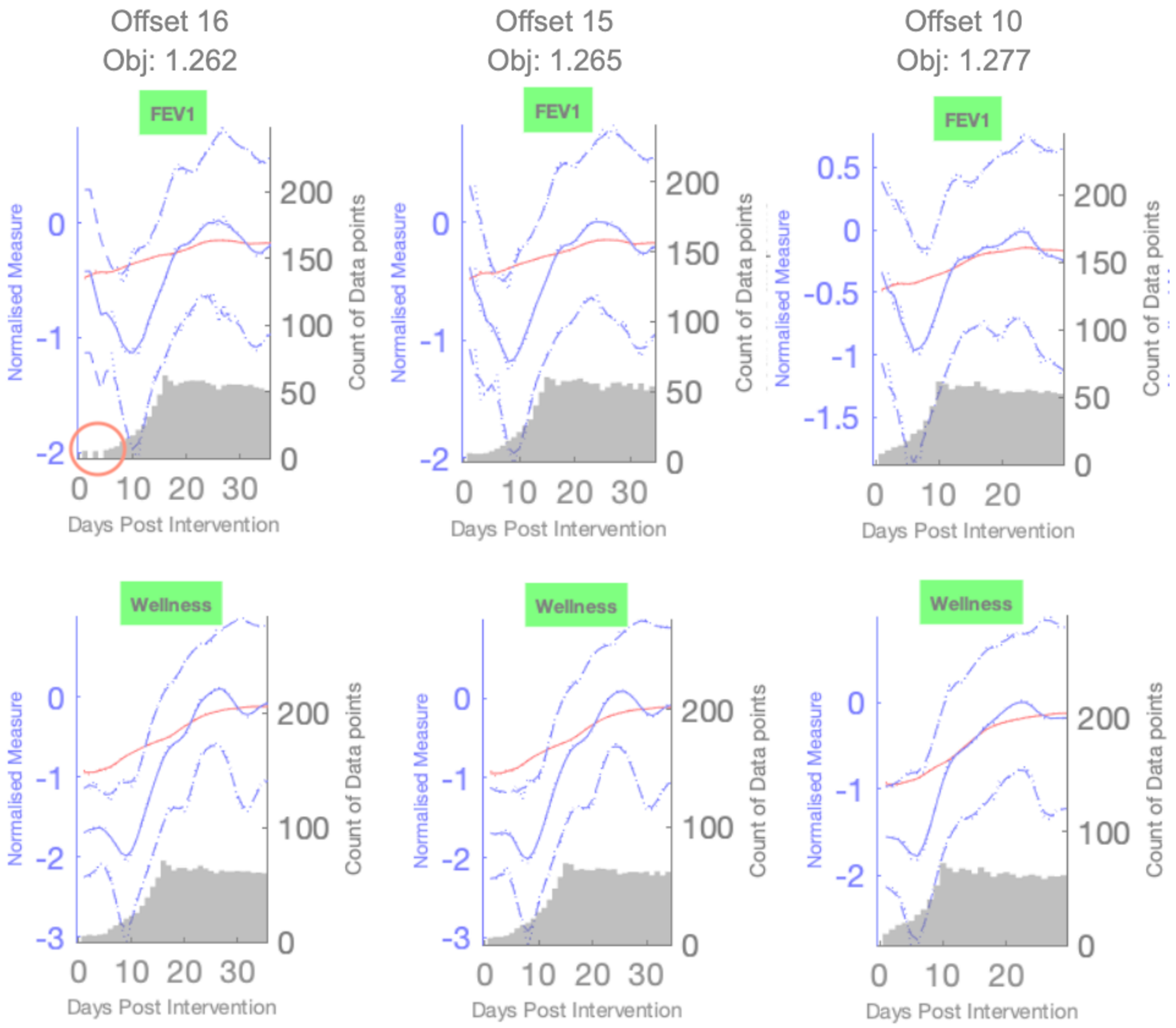
8 measures studied: wellness, cough, FEV1, FEF2575, O2 saturation, pulse rate, temperature, minutes asleep

8 measures studied: wellness, cough, FEV1, FEF2575, O2 saturation, pulse rate, temperature, minutes asleep

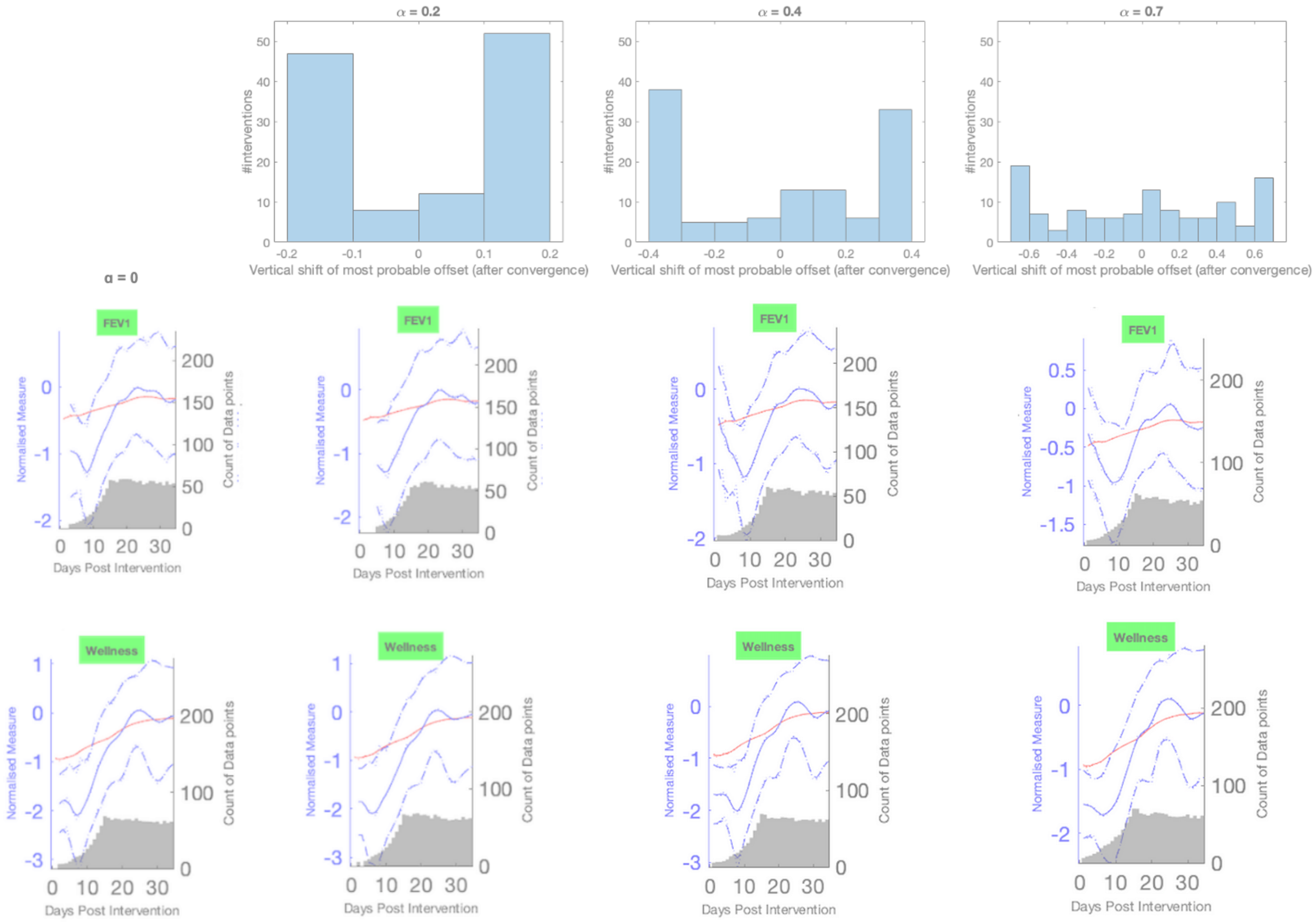
Heatmap of the number of recorded measures [-35;39] days around **treatment start**.



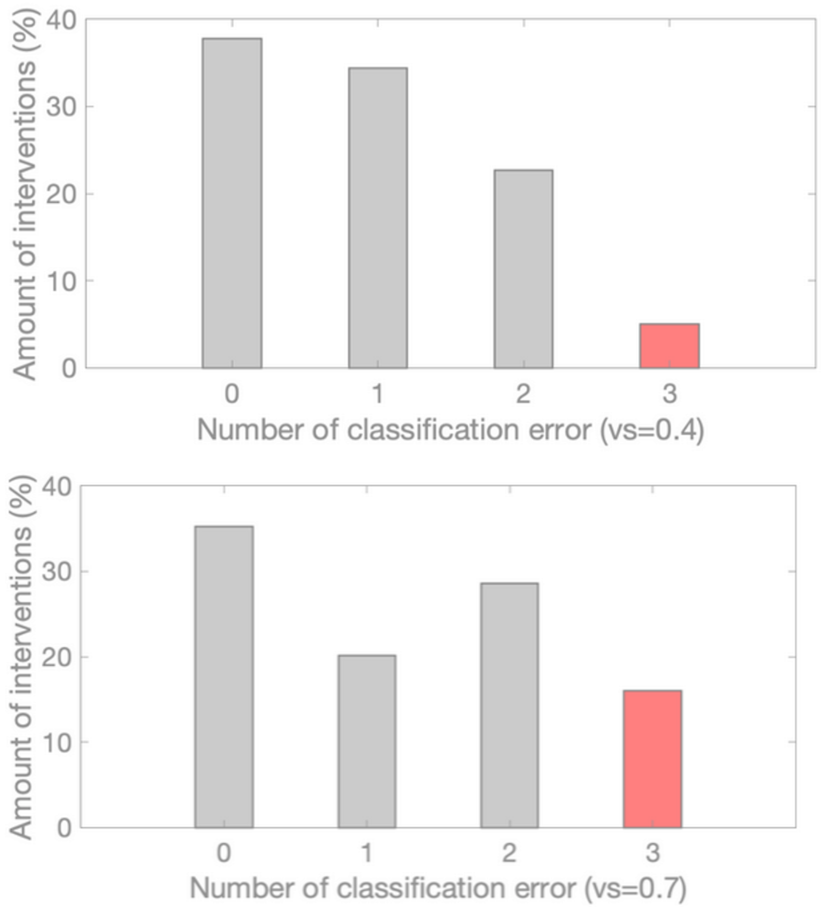
Maximum offset: 15



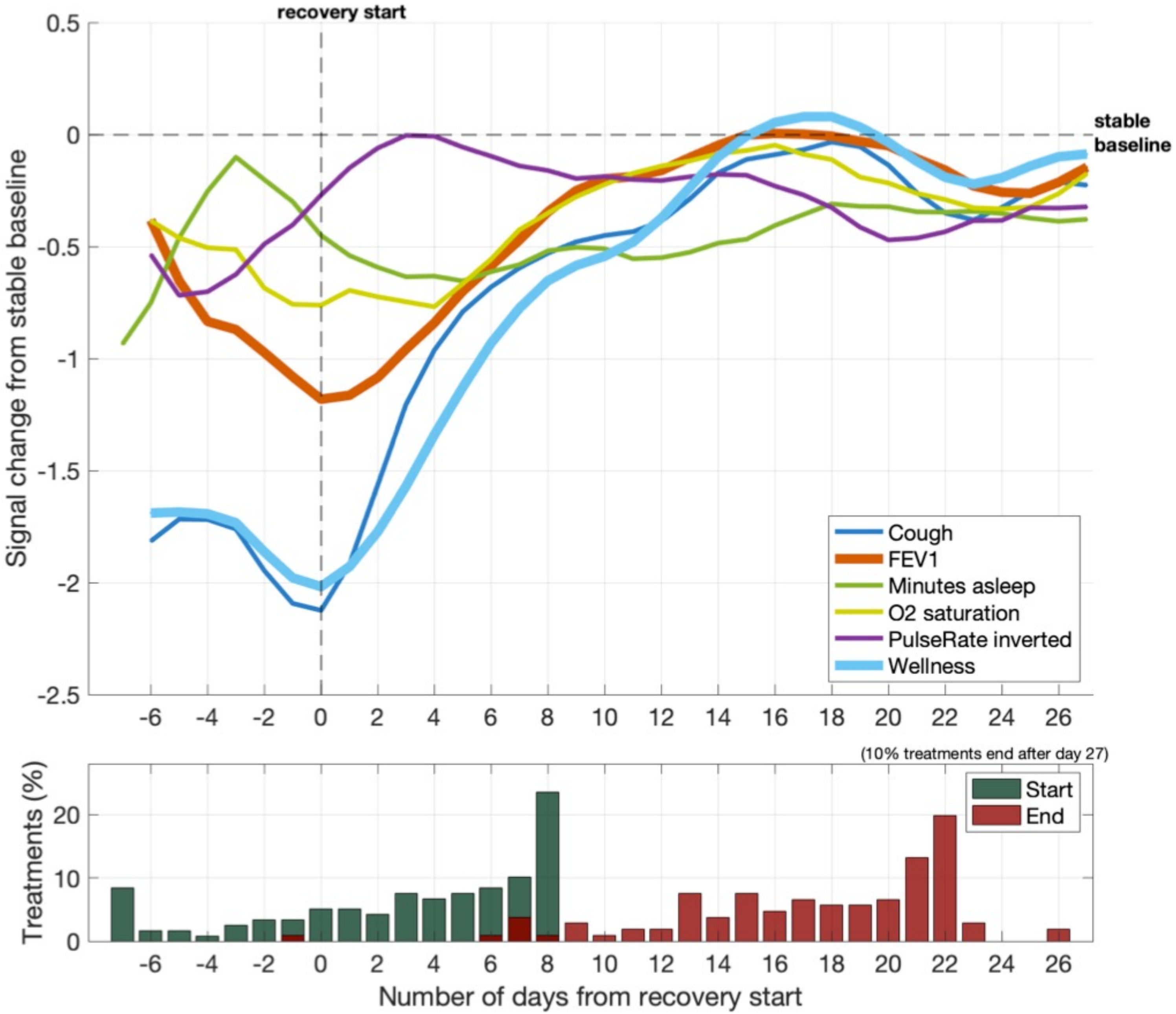
1 latent curve

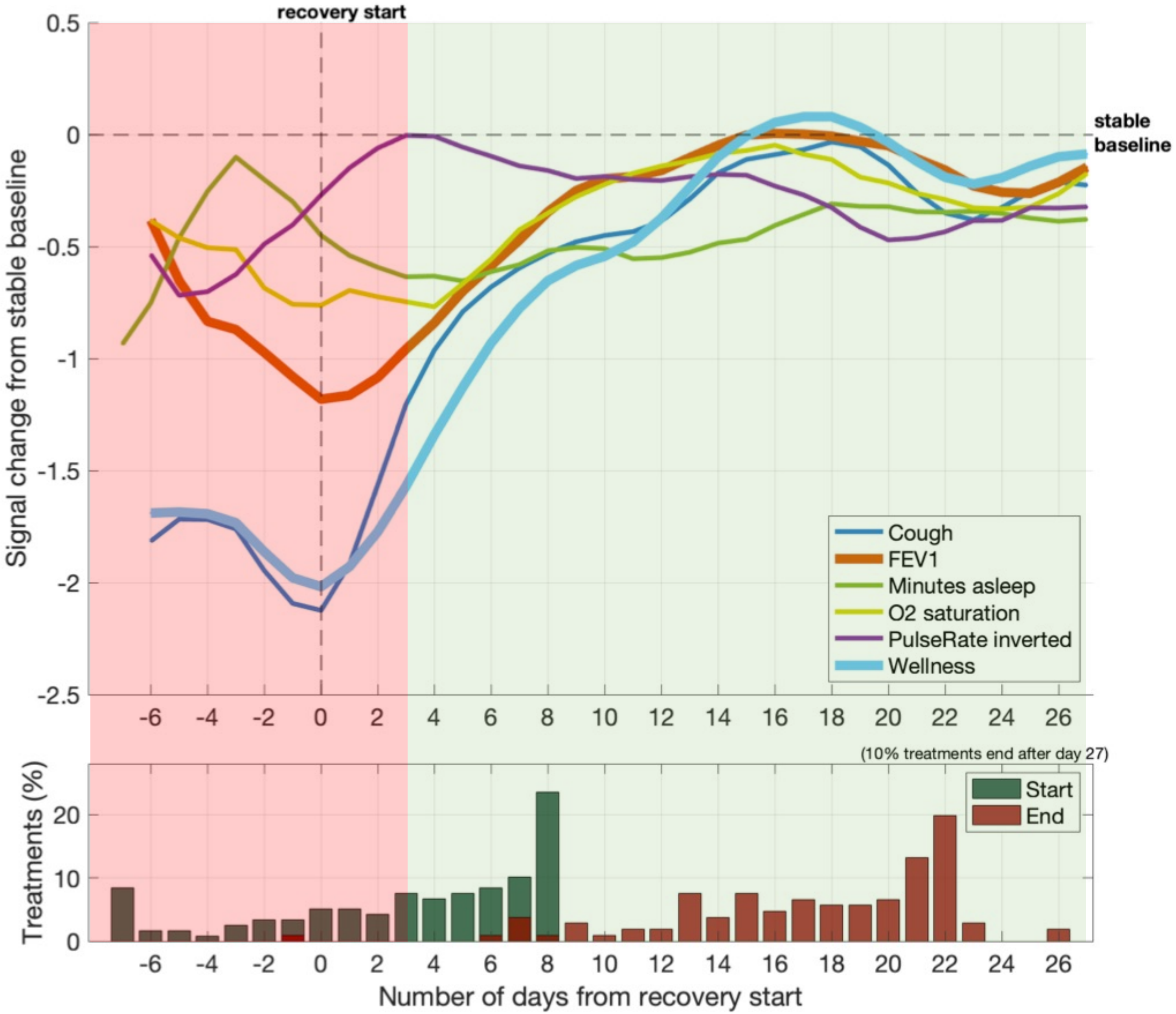


2 latent curves



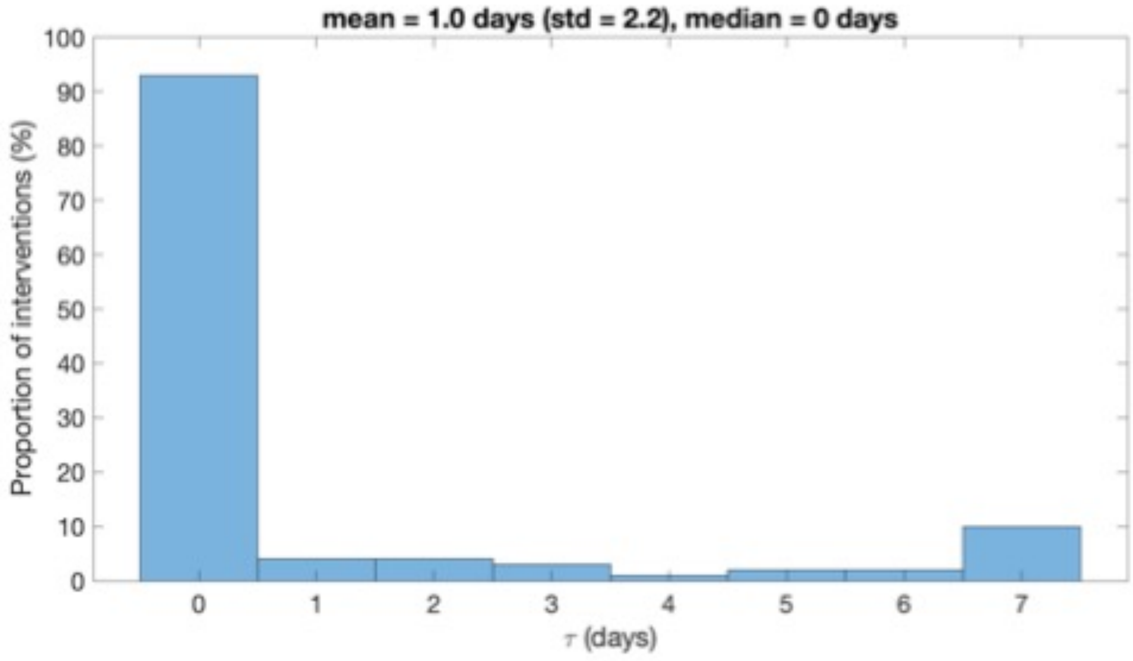
Typical profile



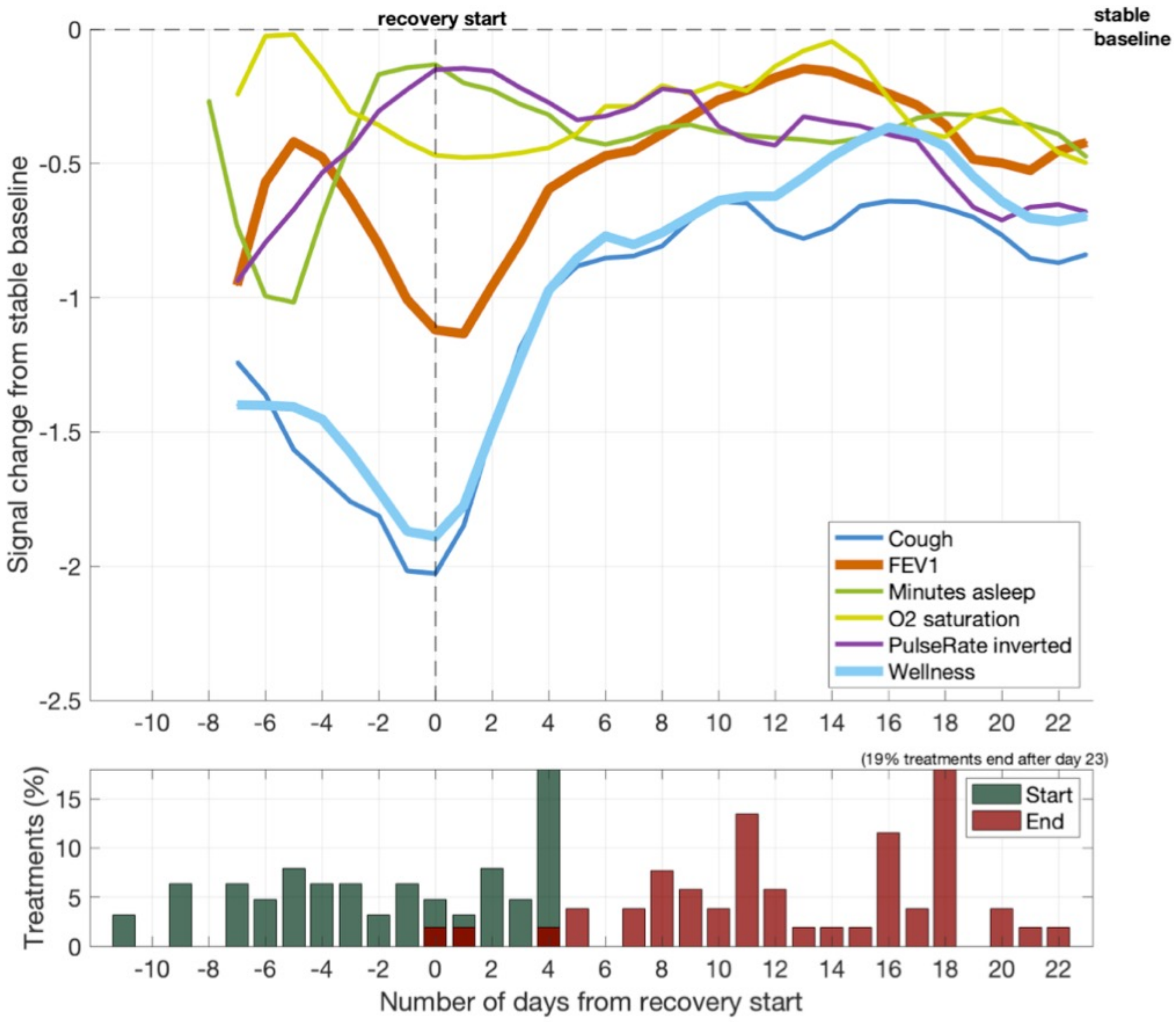

recovery start
0.5
0
-0.5
-1
-1.5
-2
-2.5
Signal change from stable baseline
stable baseline
Cough
FEV1
Minutes asleep
O2 saturation
PulseRate inverted
Wellness
-6 -4 -2 0 2 4 6 8 10 12 14 16 18 20 22 24 26
(10% treatments end after day 27)
Treatments (%)
20
10
0
Start
End
-6 -4 -2 0 2 4 6 8 10 12 14 16 18 20 22 24 26
Number of days from recovery start

Recovery start - offset

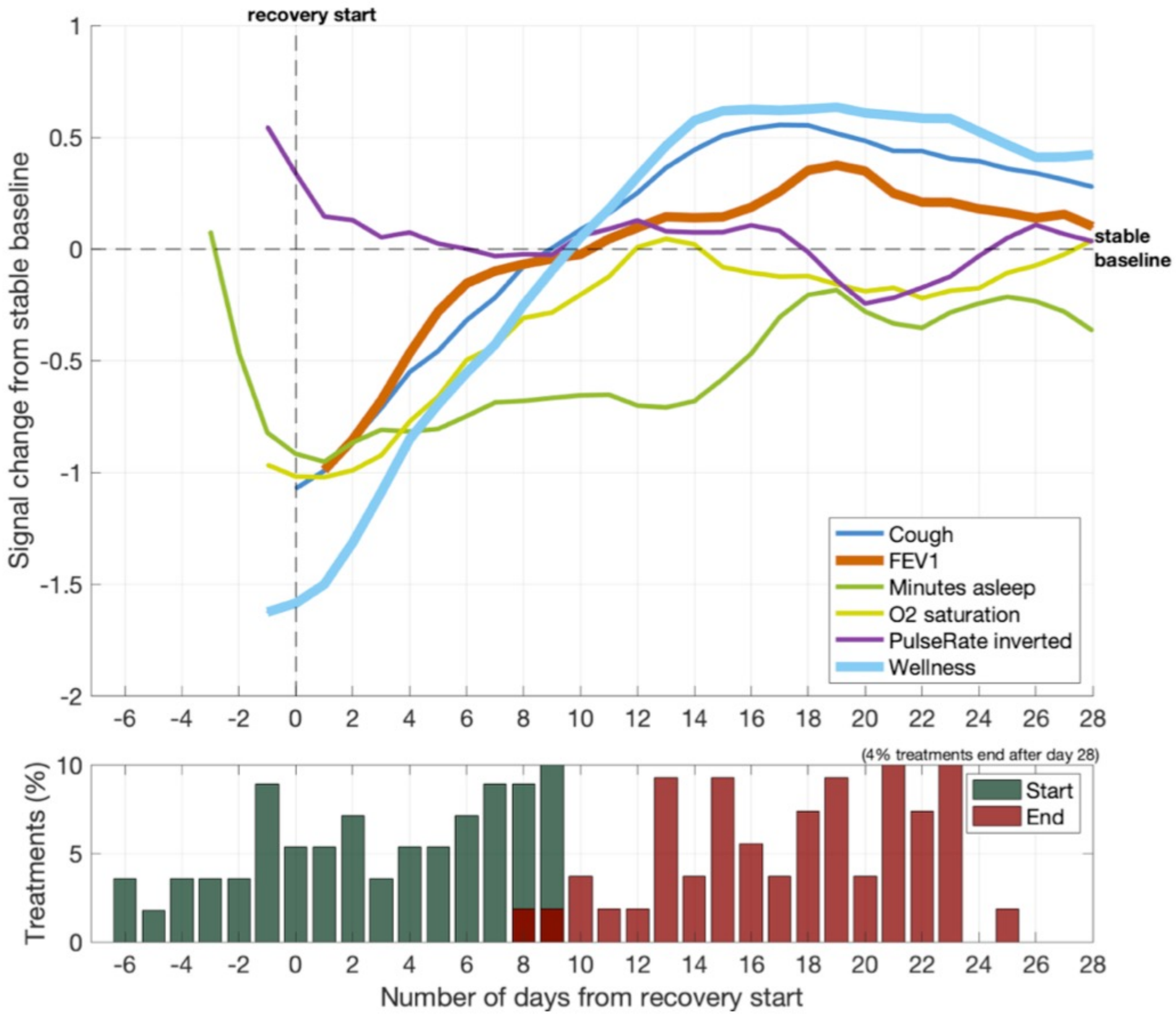
$$\forall n \in N, \tau = \begin{cases} k_R - F_n & \text{for } F_n < k_R \\ 0 & \text{for } F_n \geq k_R \end{cases}$$



Class 1: 53% of antibiotic interventions



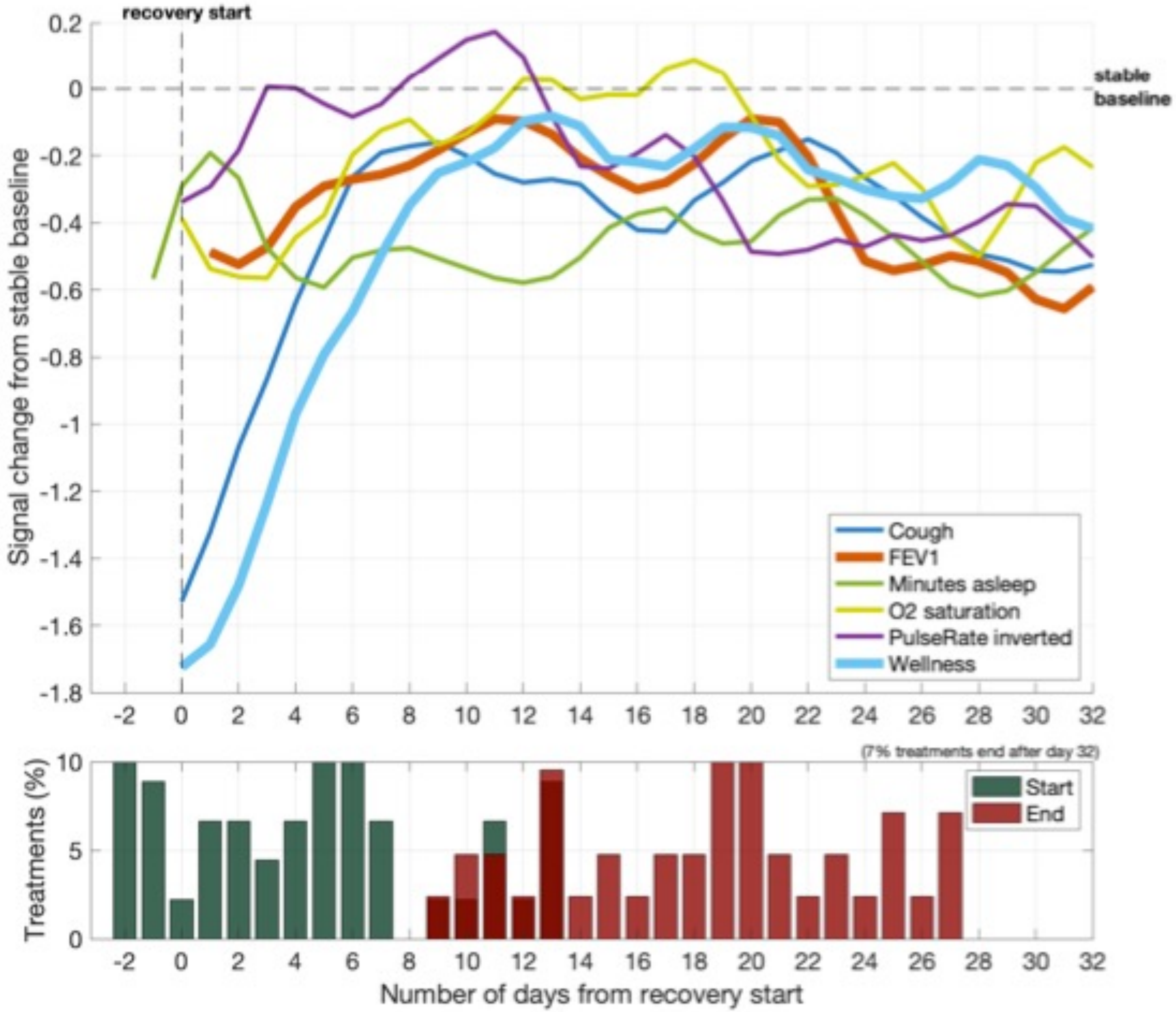
Class 2: 47% of antibiotic interventions

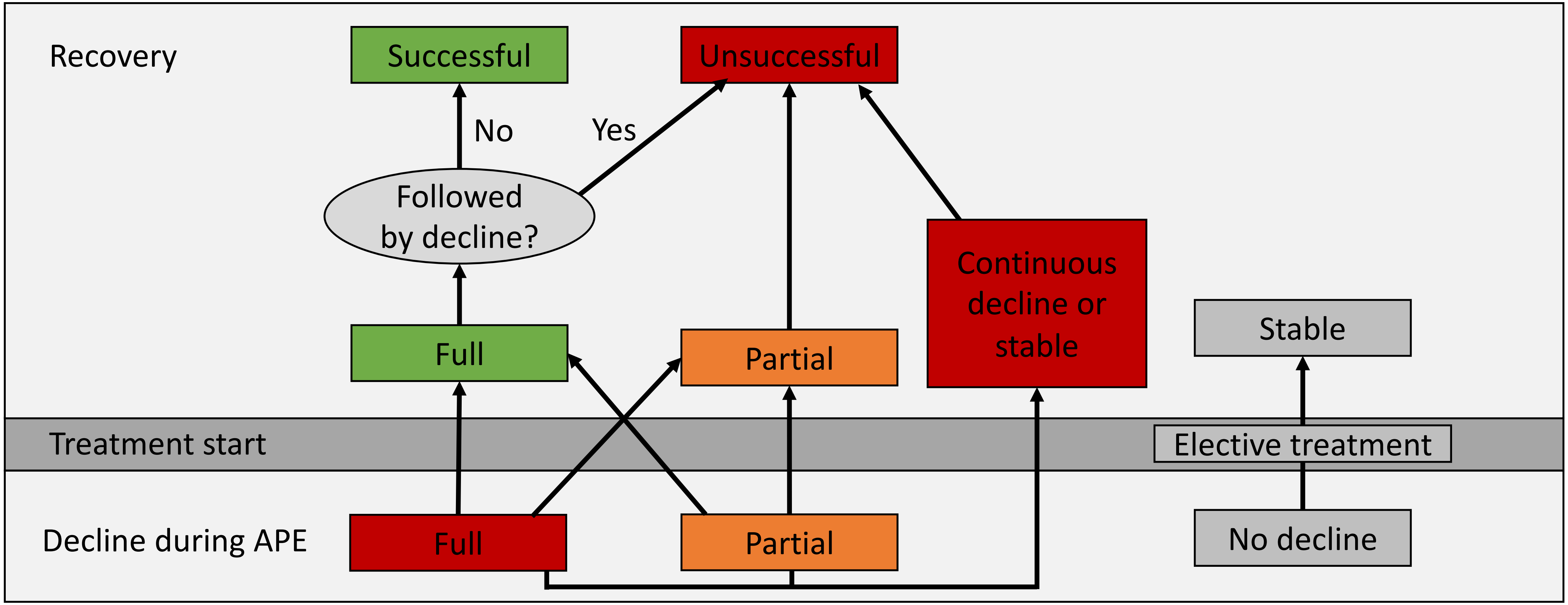


Relation with patients characteristics

	Exacerbation Class	
	Class 1 (n=57)	Class 2 (n=56)
Stable FEV1 (median ± IQR)	2.2 ± 1	1.9 ± 1.2
p value	0.11	0.11
BMI (median ± IQR)	22 ± 4.5	21 ± 3.8
p value	0.74	0.74
Age (median ± IQR)	32 ± 10	35 ± 11
p value	0.24	0.24
CRP on admission (median ± IQR)	4 ± 14	4 ± 20
p value	0.28	0.28
CRP Stable (median ± IQR)	0 ± 0	0 ± 4
p value	0.15	0.15
Time to teatment response (median + IQR)	1 ± 5	0 ± 0.5
p value	0.001	0.001
Female (%)	56	70
p value	0.03	0.04
Chronic P. aeruginosa infection (%)	65	84
p value	<0.001	0.003
Chronic S. aureus infection (%)	16	20
p value	0.46	0.43
Number of IV treatments (%)	2.5	3.5
p value	0.09	0.004
Number of antibiotic treatments (%)	3.6	4.9
p value	0.02	0.001

30% of antibiotic interventions




Recovery
Successful
Unsuccessful
No
Yes
Followed by decline?
Continuous decline or stable
Full
Partial
Stable
Treatment start
Elective treatment
Decline during APE
Full
Partial
No decline

Conclusion

Machine Learning can characterise:

- Typical recovery profile
- Different types of recoveries
    - Full recoveries
    - Partial recoveries
    - Recoveries with decline

Actionable information for the clinician

1. Patients with higher amount of treatments are more likely to experience successful recoveries. Already known, hence validates the ML approach
2. Prognosis of recovery quality: A high increase in subjective parameters (cough and wellness) not followed by physiological signals (FEV1, O2 saturation) can indicate the beginning of an unsuccessful recovery.

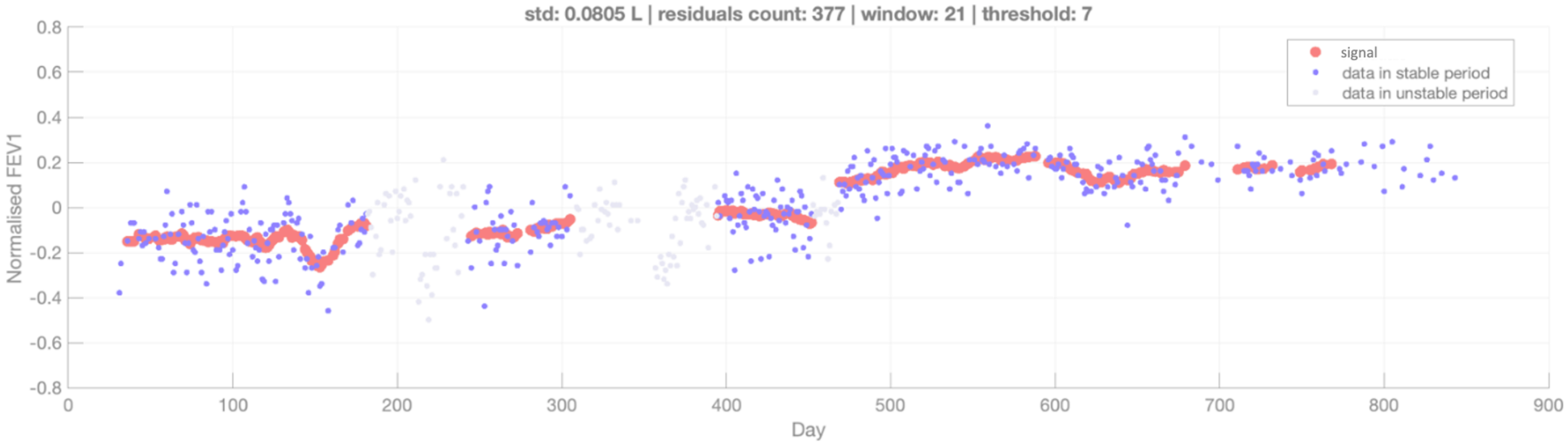
ML promise:

- **More data = more complex relations = more actionable items for clinicians**
- **Model flexibility**

Future work:

1. Analyse the impact of multiple competing features, including antibiotic choice, patient microbiology on the quality of recovery.
2. Infer long-term outcomes of combined treatments and therapies (*in particular CFTR modulators*).

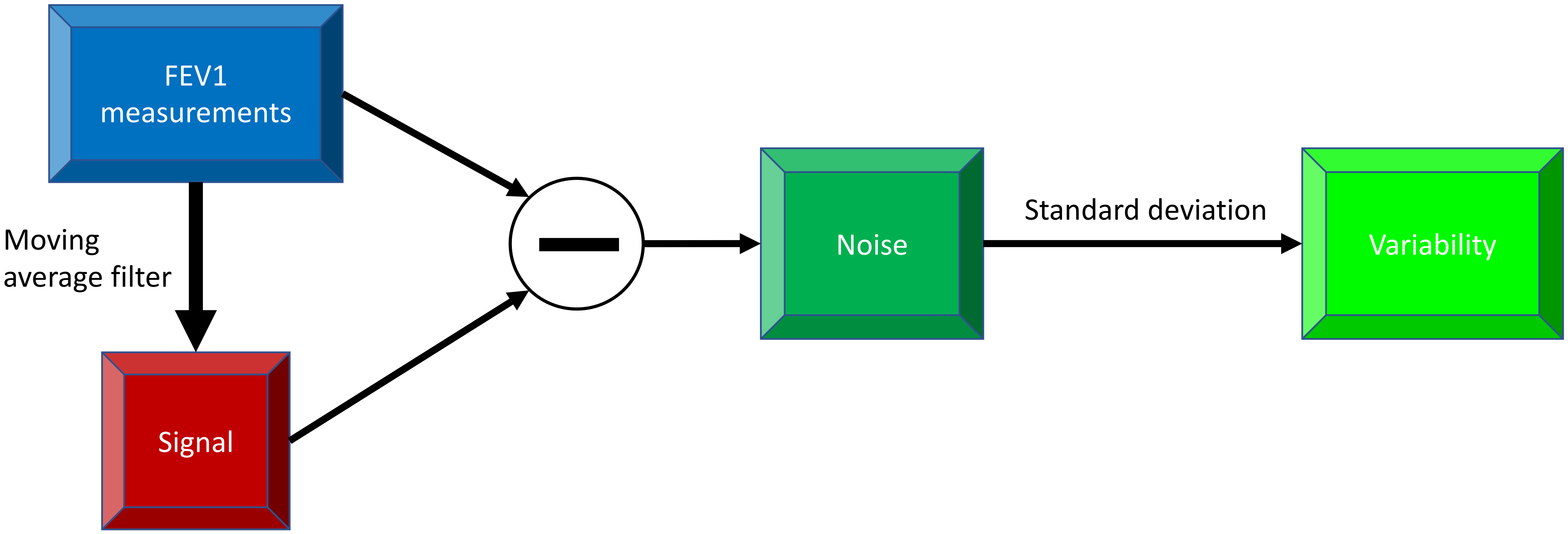
Side project: Estimation of the variability in Cystic Fibrosis patient's FEV1 lung function measurements



Method

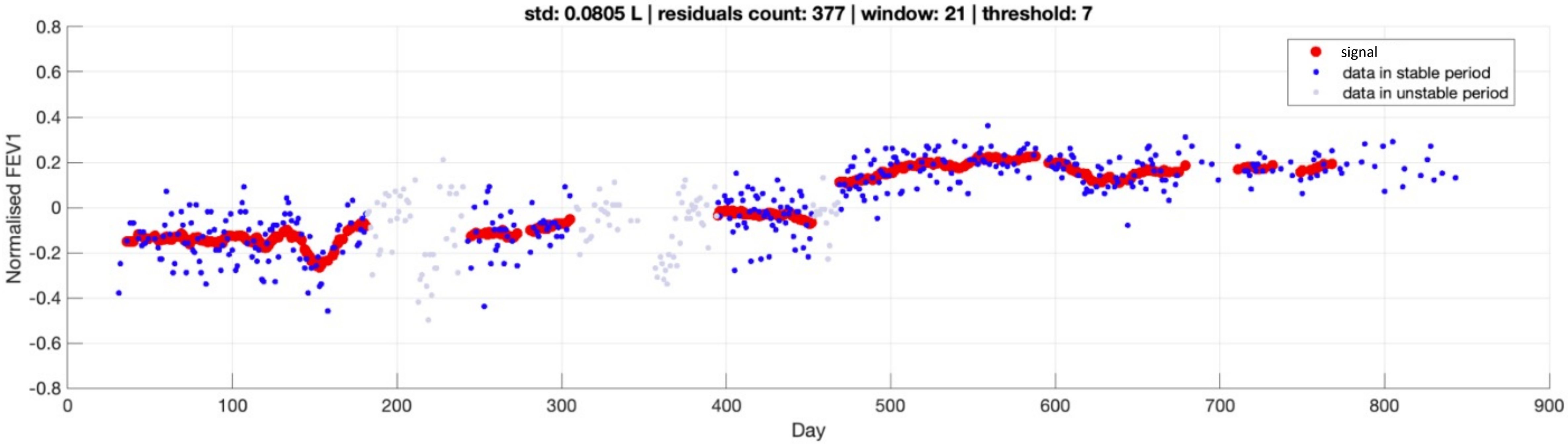
$$\text{measurement} = \text{signal} + \text{noise (L)}$$

$$\text{noise} = \text{measurement - signal (L)}$$



Model result

	Initial values	Values after stable period filter
Number of patients	226	220
Number of measurements	21036	16517



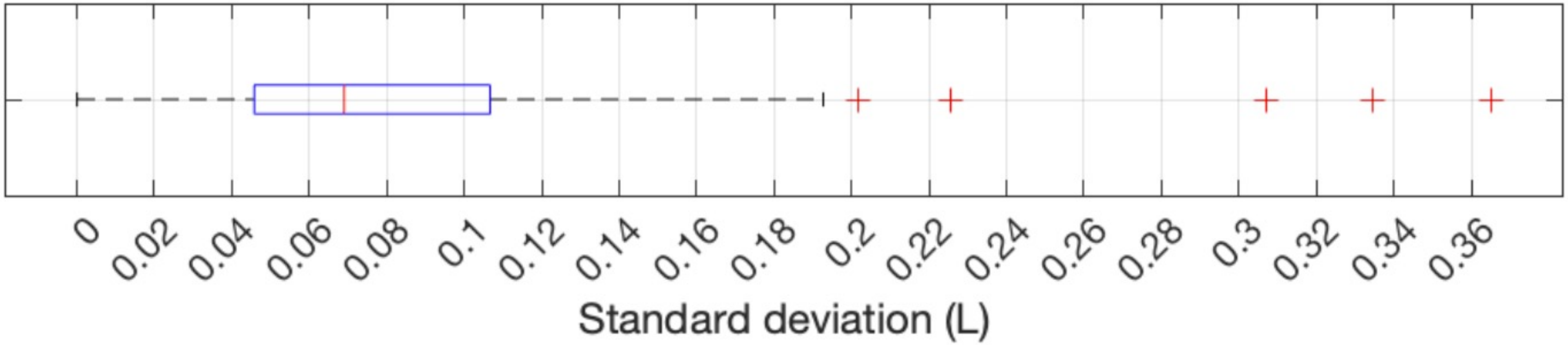
Model result

Variability: 310mL

window size (days)	7	15	21	31	
threshold (days)	3	5	7	10	
threshold/window	43%	33%	33%	32%	
#nonzero r. patients/#stable patients	81%	72%	70%	66%	
#residuals/#stable measurements	77%	79%	76%	74%	max diff (L)
standard deviation (L)	0.067	0.075	0.078	0.080	0.0130
99.5th percentile (L)	0.220	0.234	0.247	0.258	0.0382
0.5th percentile (L)	-0.247	-0.286	-0.287	-0.291	0.0439
97.5th percentile (L)	0.130	0.146	0.149	0.156	0.0261
2.5th percentile (L)	-0.135	-0.153	-0.157	-0.164	0.0288
95th percentile (L)	0.100	0.113	0.118	0.123	0.0229
5th percentile (L)	-0.100	-0.112	-0.117	-0.120	0.0200
std dev. 95% CI lower bound	0.066	0.074	0.077	0.079	
std dev. 95% CI upper bound	0.068	0.076	0.079	0.081	

Variability is *extremely diverse* among patients!

Distribution of patients' std



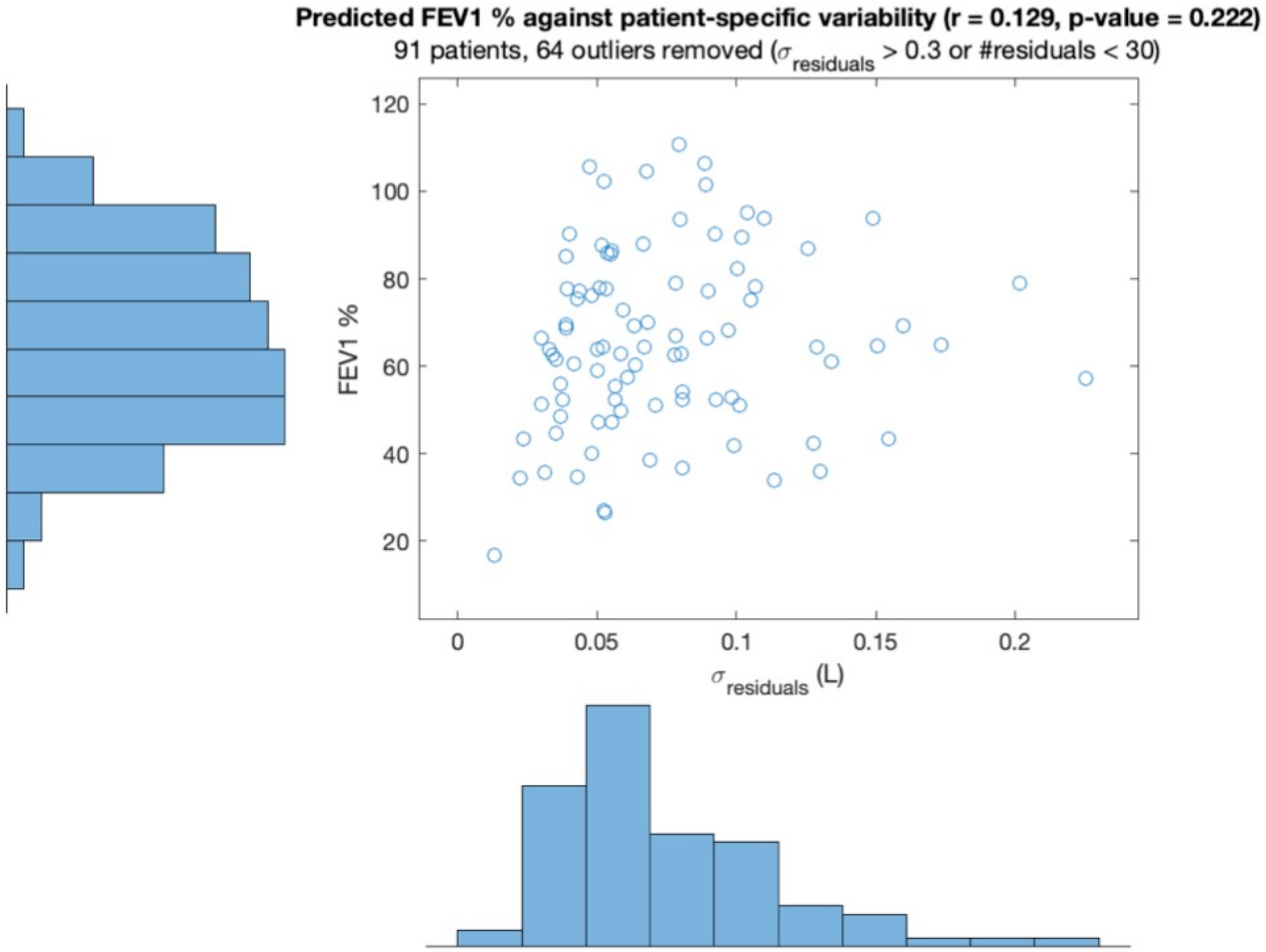
Median variability : 276 mL, interquartile range [184; 428] mL

~~Global variability: 310mL~~

Patient specific variability:
[184; 428] mL

Correlation between variability and predicted FEV1%

No correlation



Equality of variance tests:

F-test: $\sigma_1^2 > \sigma_2^2$

Levene's test: $\sigma_1^2 \neq \sigma_2^2$

Effect on CFTR modulators

Group	Measurements	Number of patients	Number of stable residuals	Std dev. (L)
A	Prior Symkevi (no therapy)		1151	0.084
B	During Symkevi	76	3006	0.075
C	During Triple Therapy		1967	0.069

Test	Groups involved	Relative change in std dev.	upper-tail F-test p-value	Levene's test p-value
1	A-C	-17%	<0.001	<0.001
2	B-C	-7%	<0.001	0.0276
3	A-B	-11%	<0.001	0.0695

Test	Measurements	Number of patients	Number of stable residuals	Std dev. (L)	Relative change in std dev.	upper-tail F-test p-value	Levene's test p-value
4	Prior Triple Therapy* During Triple Therapy	152	5632 3759	0.079 0.069	-13%	<0.001	<0.001
5	Prior Symkevi** During Symkevi	95	1388 3440	0.082 0.074	-10%	<0.001	0.1489

*includes patient histories with no therapy, Symkevi, Orkambi, **includes patient histories with no therapy, Ivacaftor

Significant

Trikafta < Symkevi

Trikafta < No therapy

Trikafta < Symkevi + Ivacaftor + Orkambi + No therapy

Not significant

Symkevi < No therapy

Symkevi < Ivacaftor + No therapy

Main project: Bayesian inference with Expectation Maximisation for the characterisation of antibiotic treatment recovery in Cystic Fibrosis

Side project: Estimation of the variability in Cystic Fibrosis patient’s FEV1 lung function measurements

Thank you for your contribution!

University of Cambridge
Department of Medicine

Academic hosts
- Prof. Andres Floto
- Ph.D. Damian Sutcliffe

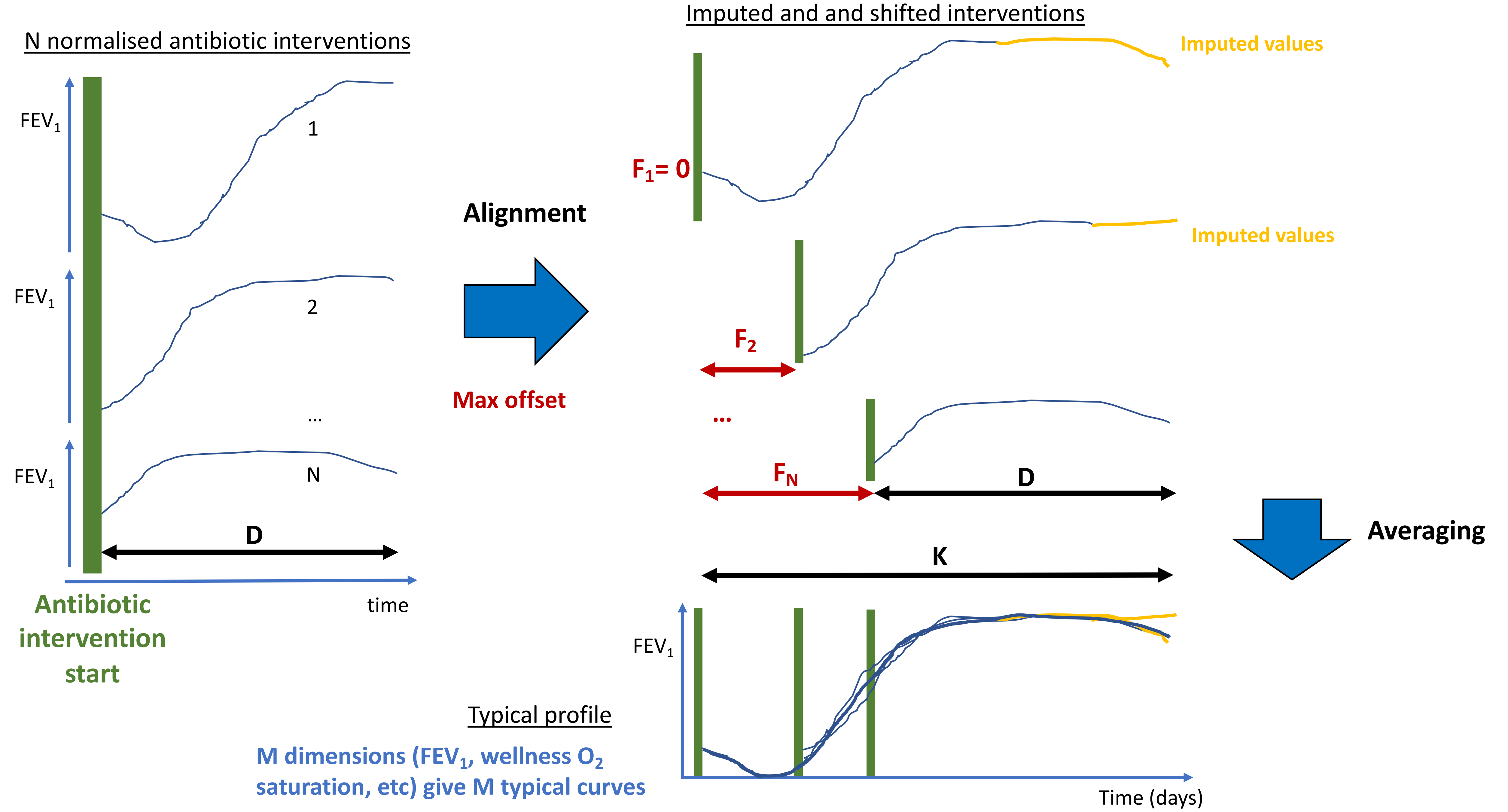
External Expert
- John Winn, Machine Intelligence Group, Microsoft Research

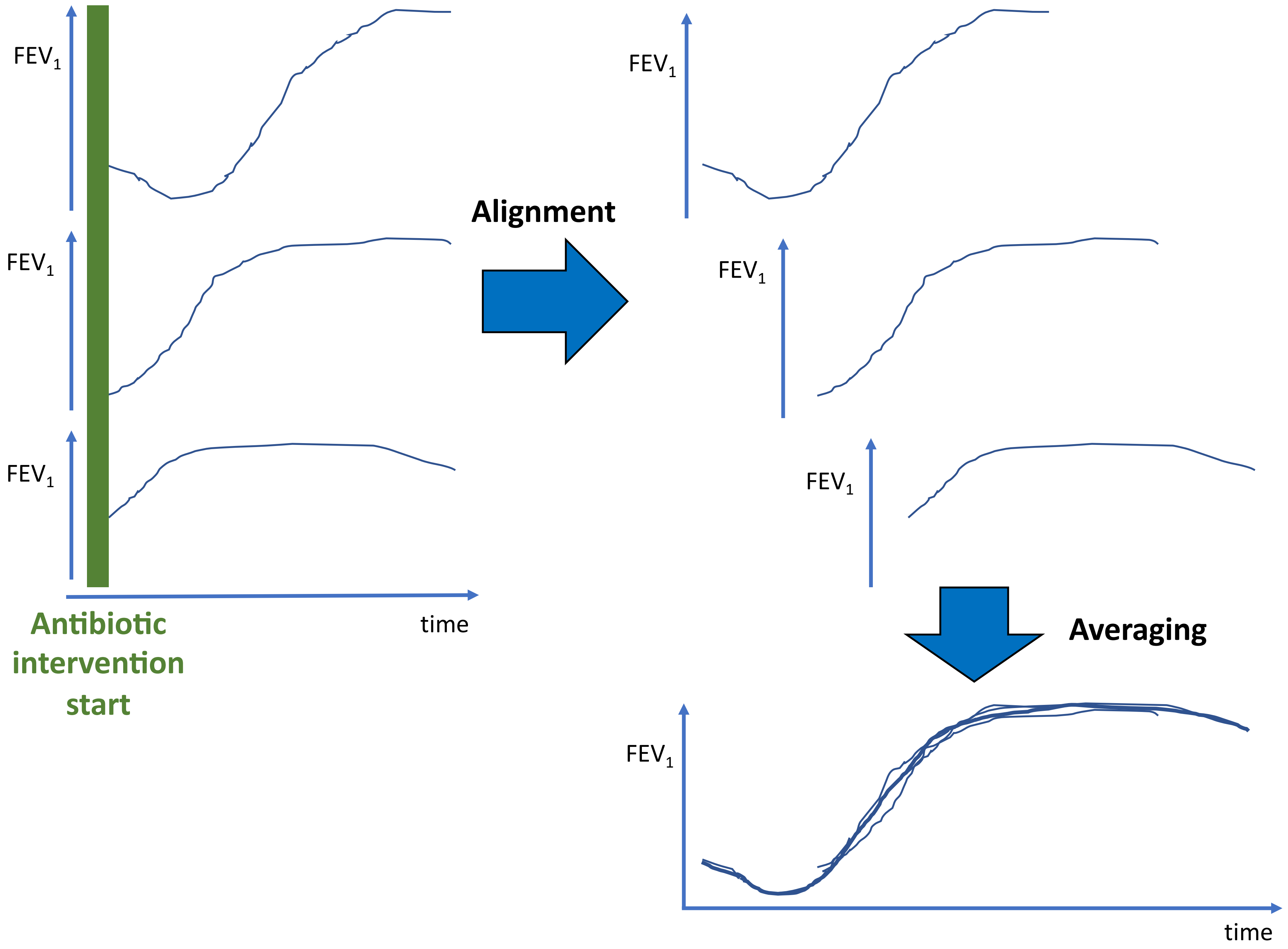
École Polytechnique Fédérale de Lausanne
School of Engineering, School of Computer and Communication Sciences

Supervised by
- Prof. Philippe Müllhaupt, School of Engineering
- Prof. Martin Jaggi, Machine Learning and Optimization Laboratory
- Dr. Mary-Anne Hartley, Machine Learning and Optimization Laboratory

26.08.2021

Without numbers





FEV$_1$
FEV$_1$
FEV$_1$
time
Antibiotic intervention start
Alignment
FEV$_1$
FEV$_1$
FEV$_1$
Averaging
FEV$_1$
time