Department of Statistical Science, Department of Mathematics
UCL Great Ormond Street Institute of Child Health
Bloomsbury Institute of Intensive Care Medicine

# Something's Missing!

## Data Imputation in Critical Care Medicine

Ali Septiandri, Takoua Jendoubi, Alejandro Diaz, Samiran Ray, Edward Palmer

# Chemistry Refresher

# Acid-Base Balance

- Human body is composed principally of water
- Water is highly ionising: $H^+ + OH^-$
- In pure water at 25ºC, the $[H^+]$ and $[OH^-]$ are $1.0 \times 10^{-7} mEq/L$
- Sorenson negative logarithmic pH = 7.0

# Water & Alkalinity

- At 0°C: pH = 7.5 (alkaline)
- At 100°C: pH = 6.1 (acidic)
- Arterial pH = 7.4
  - Acidosis pH < 7.3
  - Alkalosis pH > 7.5

# What determines pH?

- Water dissociation equilibrium
- Weak acid dissociation equilibrium
- Conservation of mass for weak acids
- Bicarbonate ion formation equilibrium
- Carbonate ion formation equilibrium
- Electrical neutrality

# What determines pH?

- $[H^+] \times [OH^-] = K_W$
- $[H^+] \times [A^-] = K_A \times [HA]$
- $[HA] + [A^-] = A_{TOT}$
- $[H^+] \times [HCO_3^-] = K_C \times pCO_2$
- $[H^+] \times [CO_3^{2-}] = K_3 \times [HCO_3^-]$
- $[SID] + [H^+] – [HCO_3^-] – [A^-] – [CO_3^{2-}] – [OH^-] = 0$

# What determines pH?

$$[SID] + [H^+] - K_C\frac{pCO_2}{[H^+]} - \frac{K_A A_{TOT}}{K_A + [H^+]} - K_3\frac{K_C pCO_2}{[H^+]^2} - \frac{K_W}{[H^+]} = 0$$

where SID, $A_{TOT}$, and $pCO_2$ are independent variables and $K_X$ are constants.

# Motivation

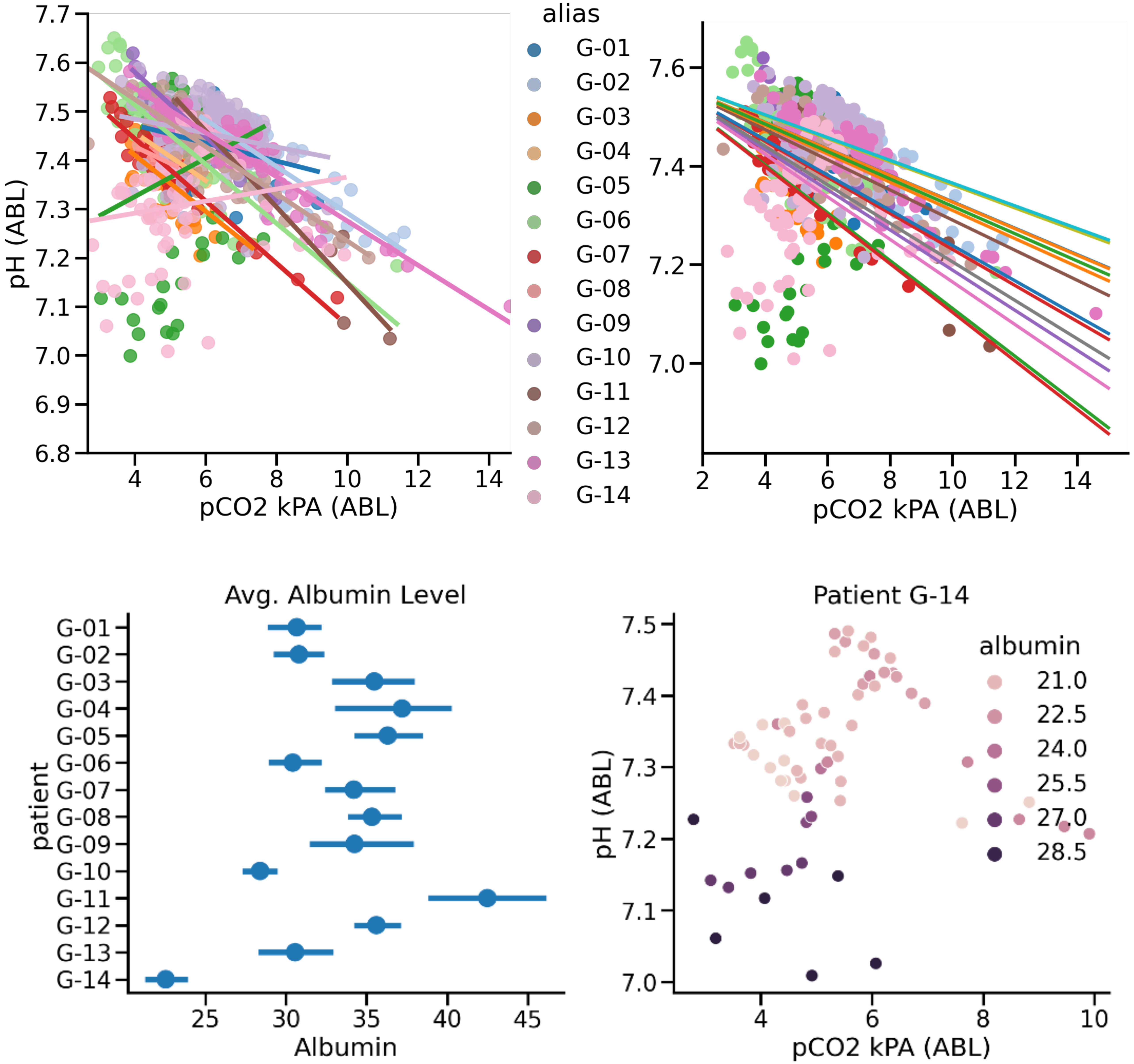
# Issues

- Concentration of $CO_2$ can be easily monitored
- It is not enough to tell the whole story
- Other variables are collected in different frequencies

# Issue

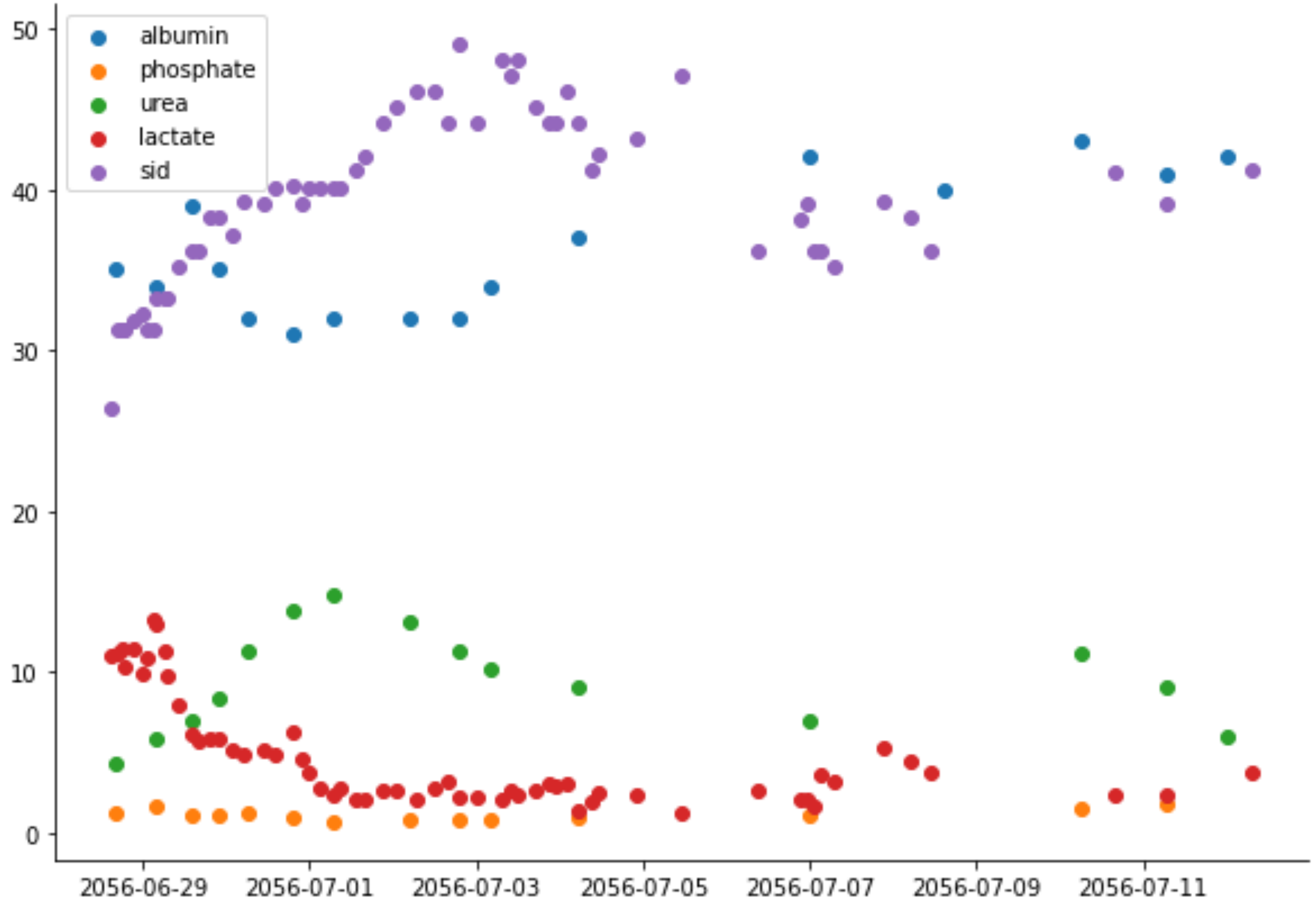
**$CO_2$ is not enough to tell the whole story**

- Model: pH ~ $pCO_2$
  - Left: One OLS model for each
  - Right: Hierarchical with shared intercept
- For patient G-14, albumin level is low most of the time
- It spikes when the anomaly occurs



# Issue

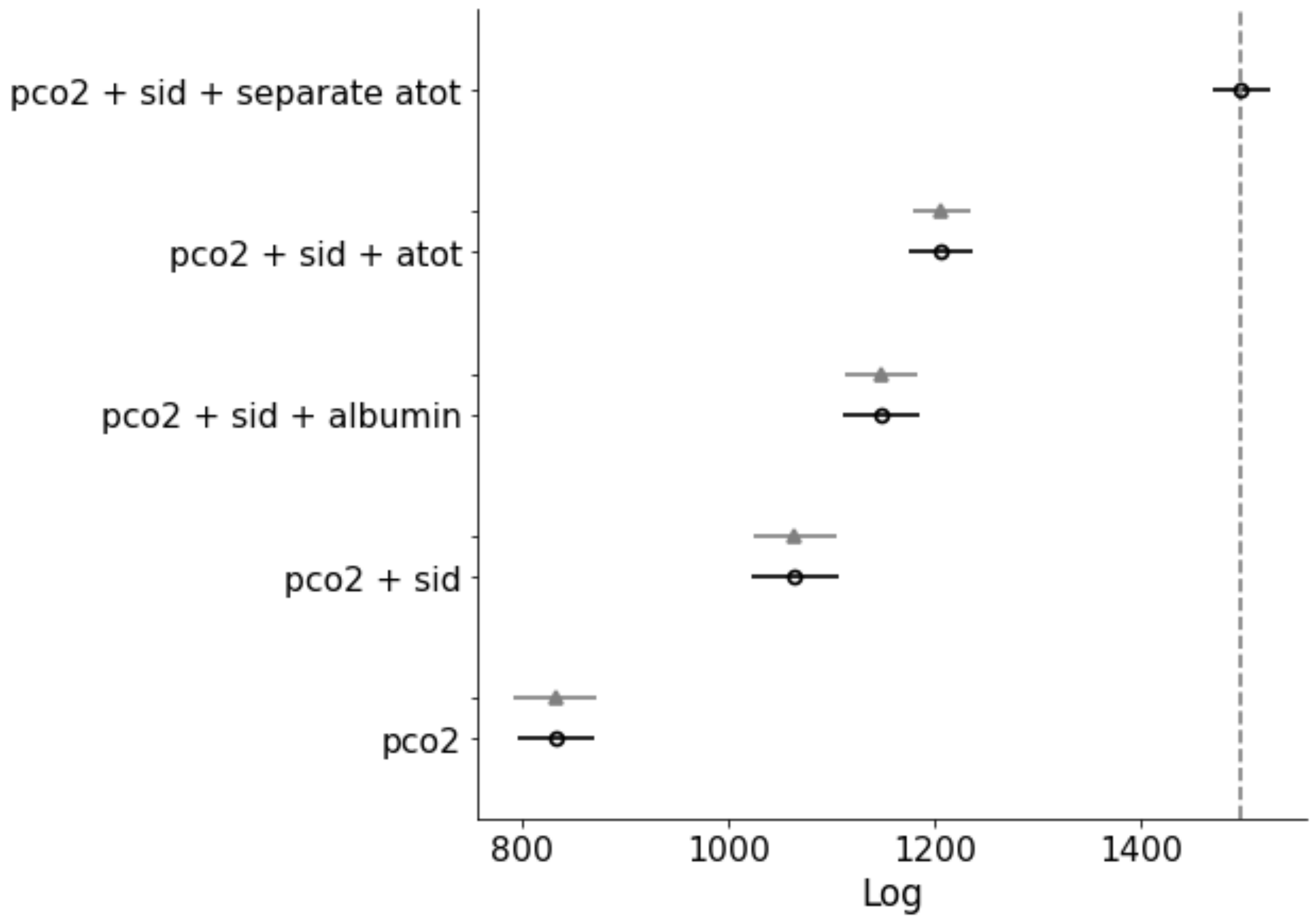
**Data comes in different frequencies**




The healthier the patient, the less data you would see

# Adding Covariates

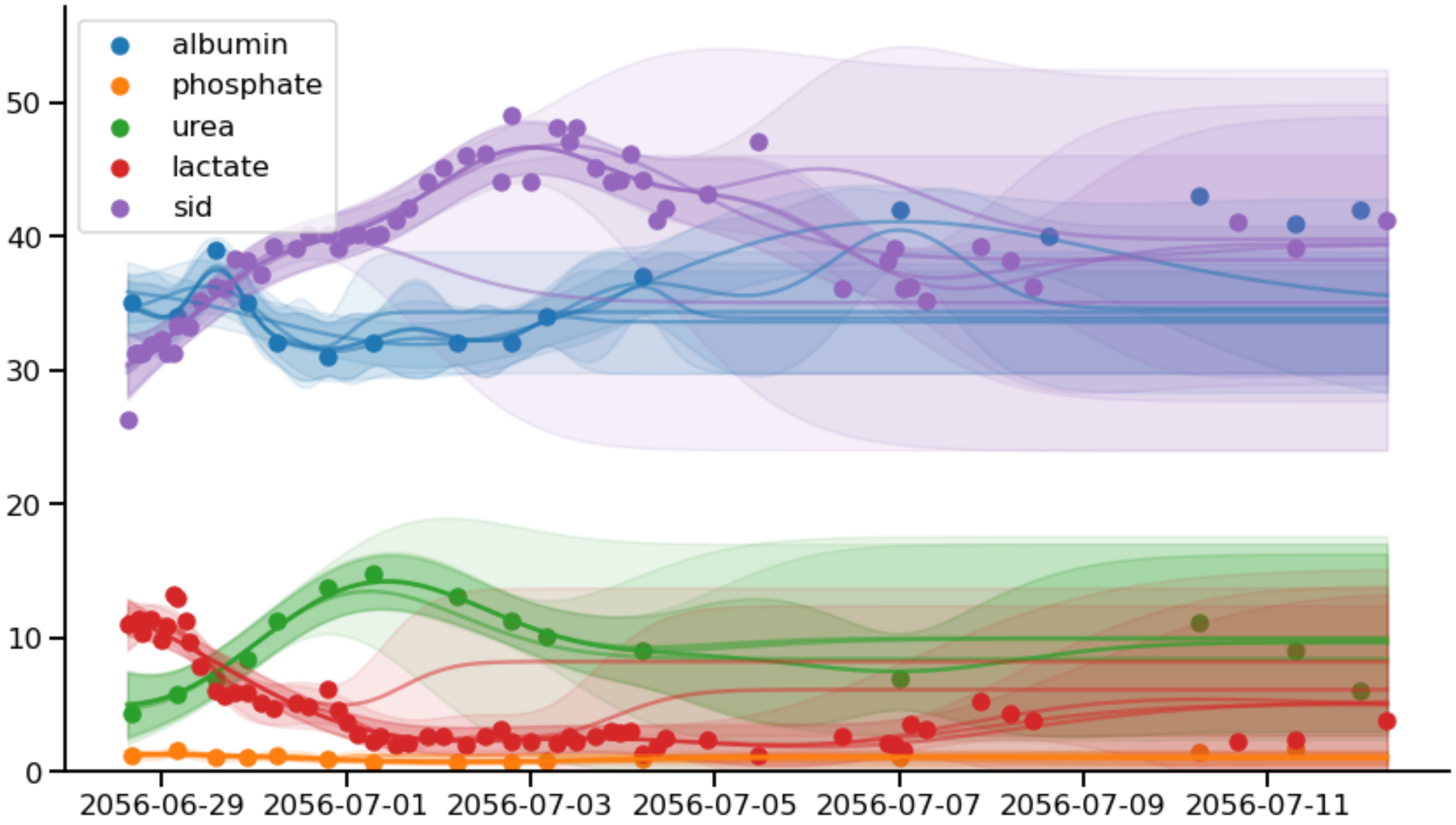
## Last known value imputation



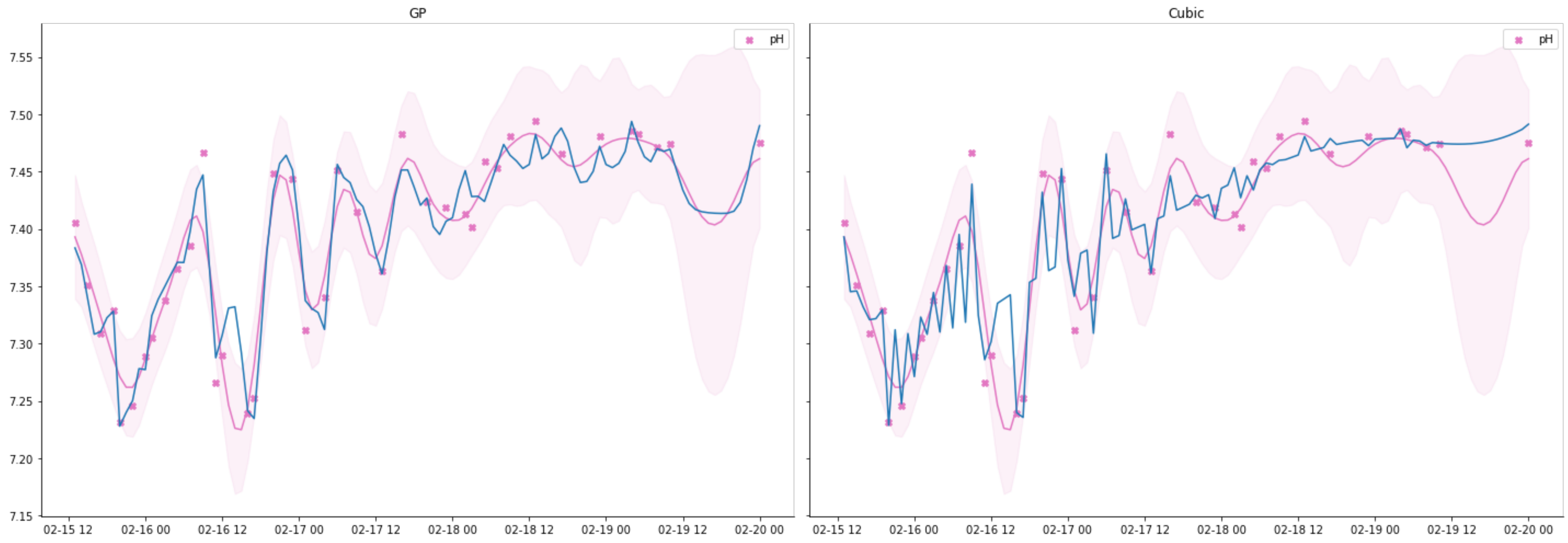

Log pointwise predictive density (Vehtari et al., 2017)

# Proposal

# Time Series Cross-Validation with GP



# GP vs Cubic Spline Interpolation




Interpolate the covariates and predict the pH

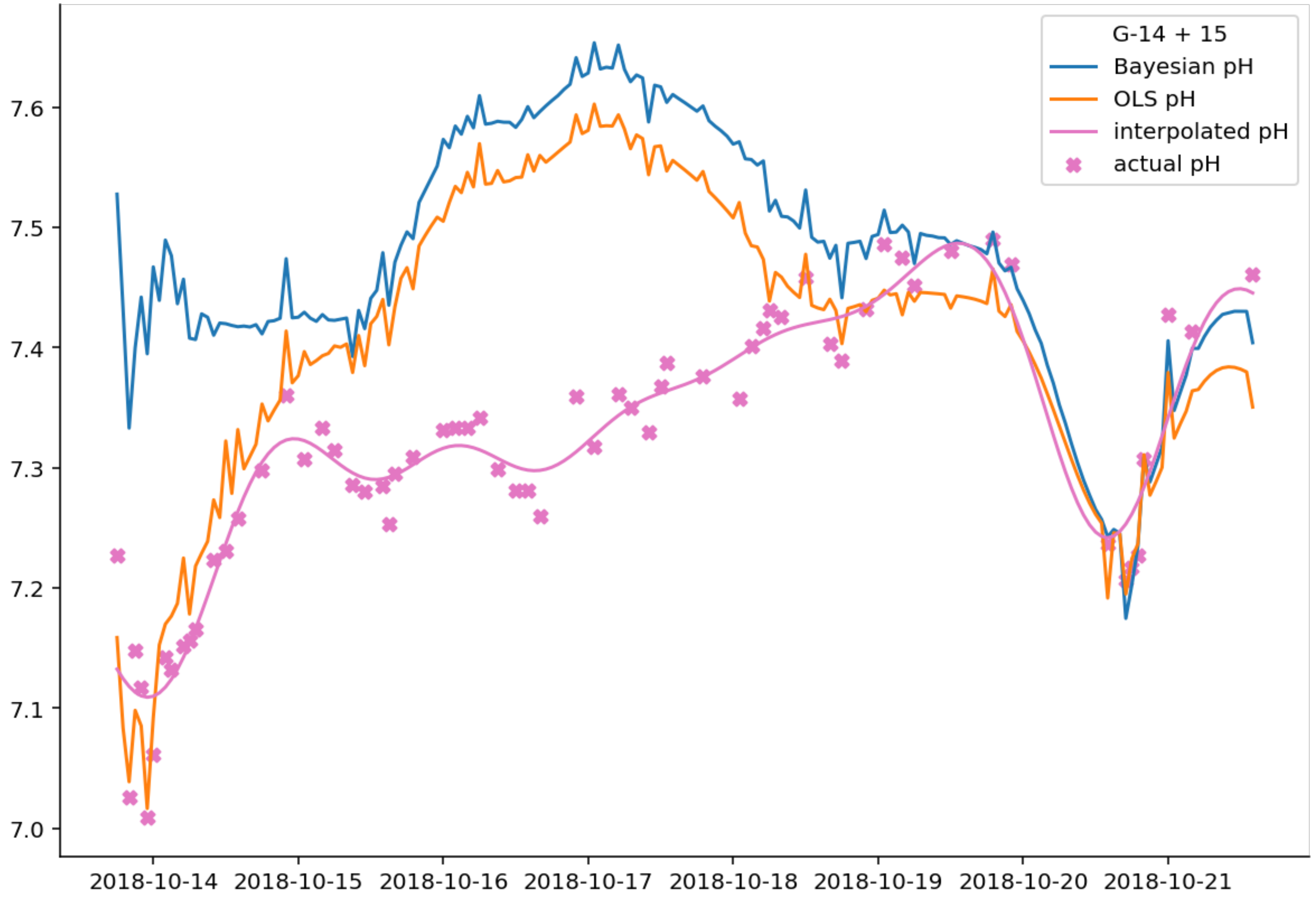
# Cross-Validation

## Leave-One-Patient-Out

- Pink line is GP-interpolated pH
- Blue line is OLS on interpolated covariates
- The difference is not significant

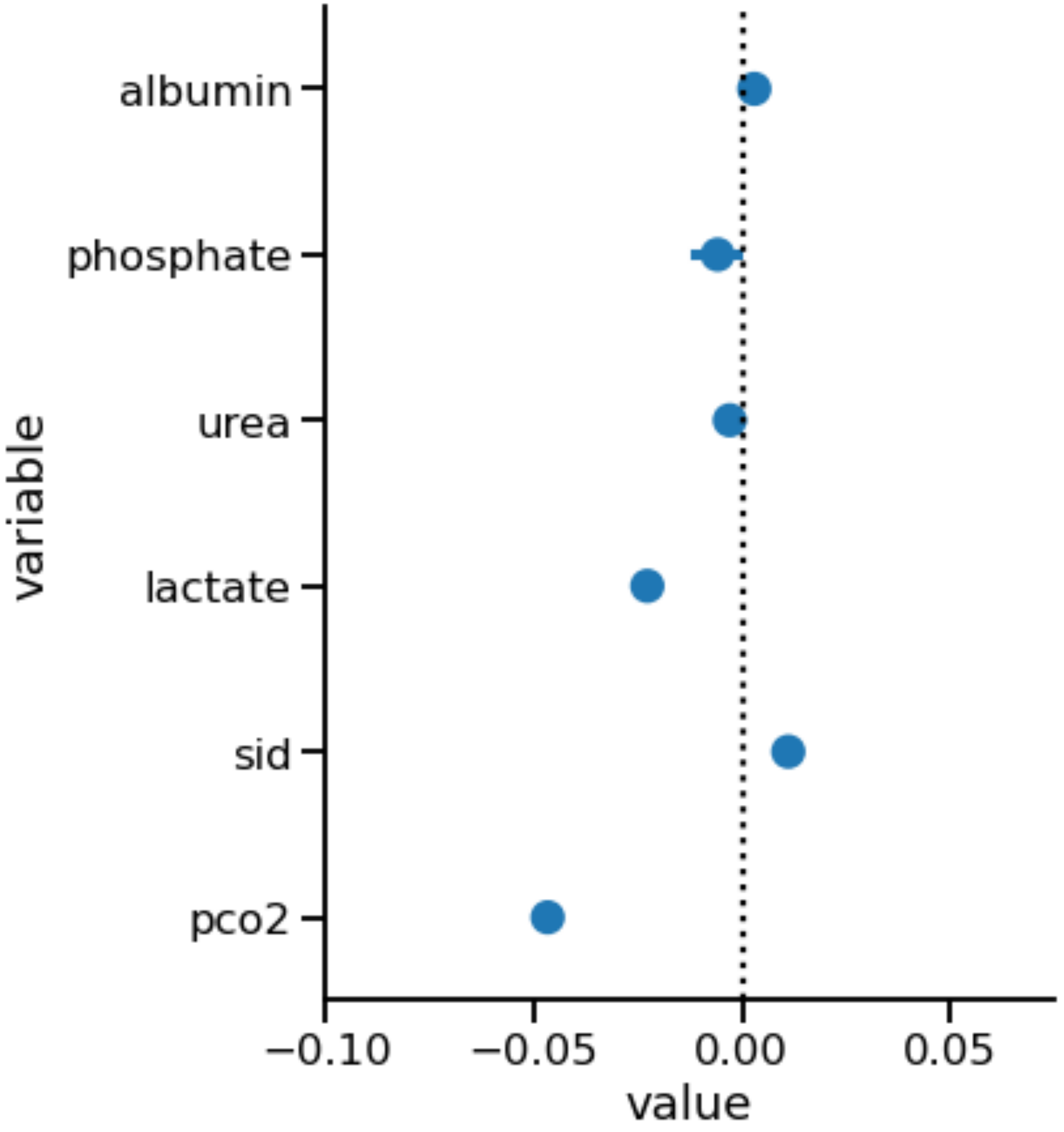
# Systematic Issue

**We might still be missing something**



# Systematic Issue

**Albumin should not have a positive slope**



# Future Work

# Future Work

- Find a way to incorporate uncertainties in the covariates into the final model
- Explore other methods: Sparse Gaussian Process, Hierarchical model with AR(1)
- Apply the analysis to a larger dataset

# MIMIC-III

- ICU dataset
- Has downstream tasks
- De facto standard for studies in data imputation

# What if we don’t have to impute at all?

- Discussed by Lipton et al. (2016), Yoon et al. (2017), Che et al. (2018)
- Using an indicator of missingness that will be used as an input to the model for downstream tasks, e.g. mortality prediction, probability of getting discharged within N-days
- Have been compared with forward-filling and zero imputation

# Masking and Time Interval

- Motivation: Informative sampling → missingness
- We also need the time interval since the last data acquisition

$\boldsymbol{X}$: Input time series (2 variables);
$\boldsymbol{s}$: Timestamps for $\boldsymbol{X}$;
$\boldsymbol{M}$: Masking for $\boldsymbol{X}$;
$\boldsymbol{\Delta}$: Time interval for $\boldsymbol{X}$.

$$\boldsymbol{X} = \begin{bmatrix} 47 & 49 & NA & 40 & NA & 43 & 55 \\ NA & 15 & 14 & NA & NA & NA & 15 \end{bmatrix} \qquad \boldsymbol{M} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{s} = \begin{bmatrix} 0 & 0.1 & 0.6 & 1.6 & 2.2 & 2.5 & 3.1 \end{bmatrix} \qquad \boldsymbol{\Delta} = \begin{bmatrix} 0.0 & 0.1 & 0.5 & 1.5 & 0.6 & 0.9 & 0.6 \\ 0.0 & 0.1 & 0.5 & 1.0 & 1.6 & 1.9 & 2.5 \end{bmatrix}$$

**Figure 2.** An example of measurement vectors $x_t$, time stamps $s_t$, masking $m_t$, and time interval $\delta_t$.

# Challenges

- Most of the recent work is using deep learning
- Deep learning only provides point estimates
- Still need to account for uncertainty
- Can we combine deep learning and classical (Bayesian) statistical approach?

# Thank you