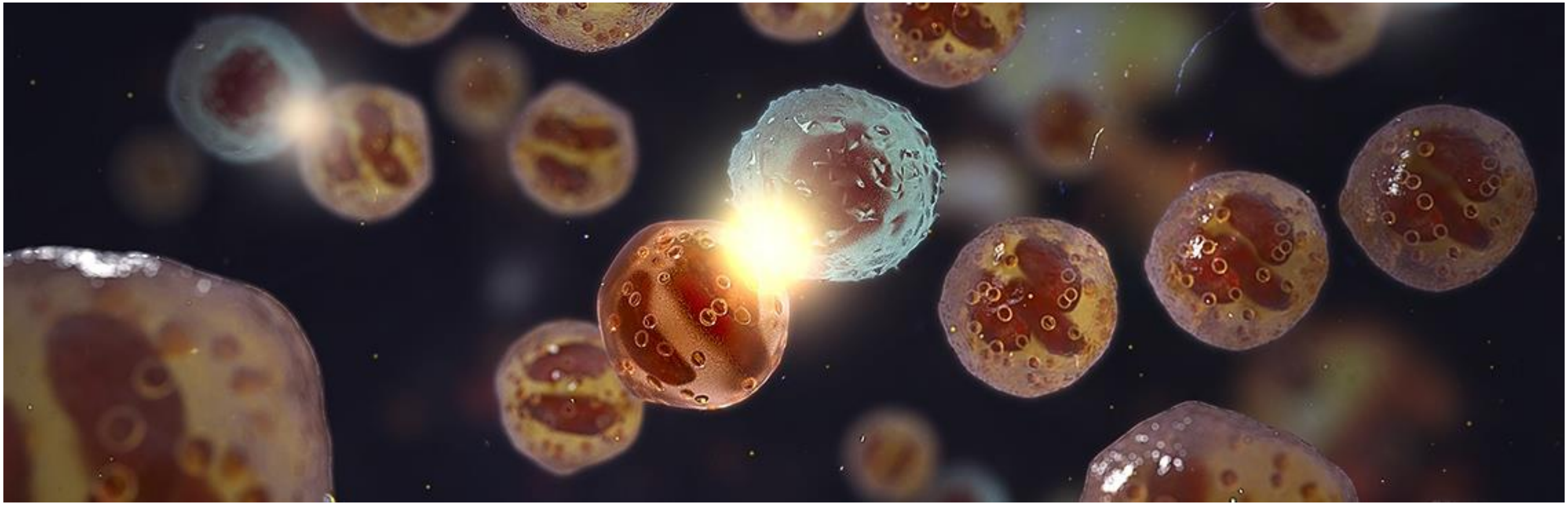
AstraZeneca

Streaming data for health: sensors, wearables, apps and all that

Dr Jim Weatherall

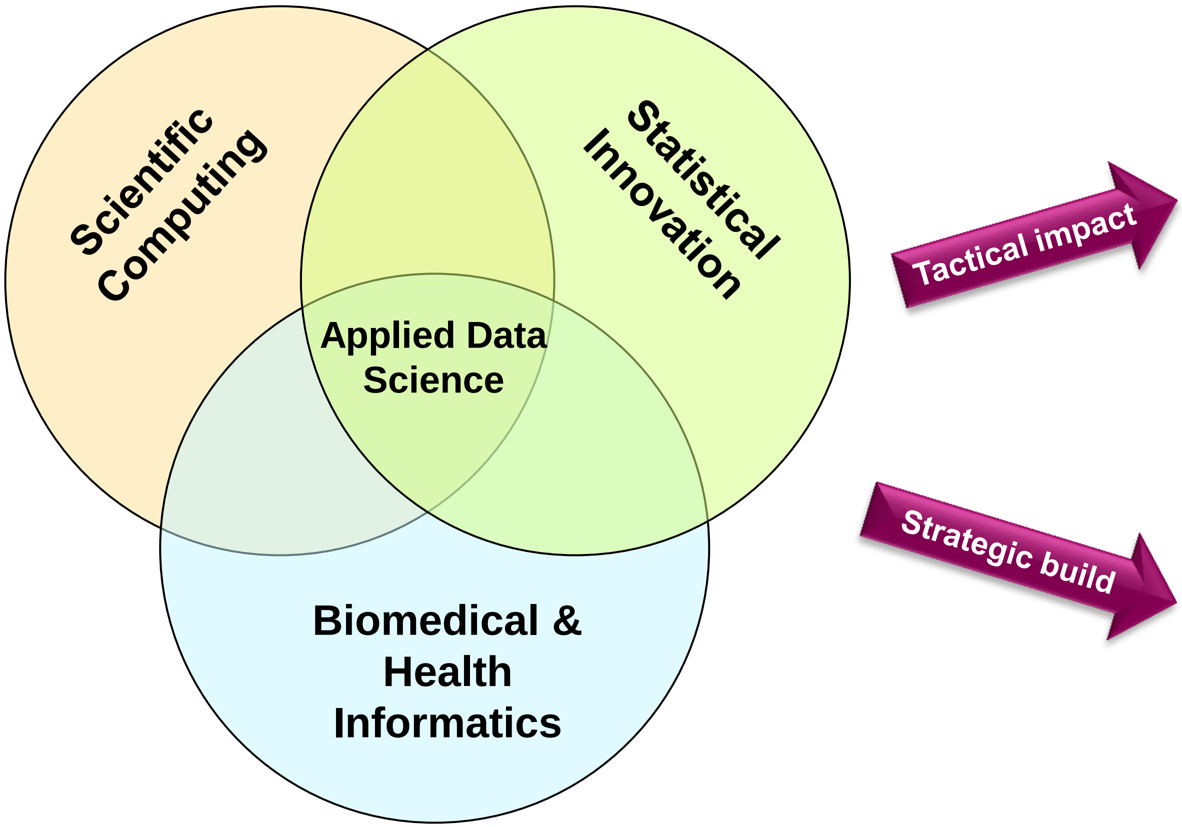
Head, Advanced Analytics Centre, AstraZeneca
Honorary Reader, Health eResearch Centre & School of Computer Science, University of Manchester

StatScale Launch Workshop, London 15th November 2016



The AZ Advanced Analytics Centre (AAC)

Transforming drug development decision making through applied data science



Advanced Analytical support to drug development:

- ✓ Design options & simulations
- ✓ Study data visualisation
- ✓ Effect size benchmarking
- ✓ Data mining
- ✓ Real world data analytics

Focussed strategic projects:

- ✓ Outcomes database
- ✓ Predictive code and visualisations
- ✓ Dynamic drug supply modelling
- ✓ Structured exploratory analysis
- ✓ Treatment pathway analytics

~30 clinical data scientists

Based in UK (Cambridge & Cheshire); Sweden (Gothenburg); US (Washington, DC)

Overview

- The emerging role of technology in clinical research
- Implications for statistics: clinical trials in the balance
- Case studies from personal experience
- Closing thoughts

Digitalization of clinical trials is coming…?

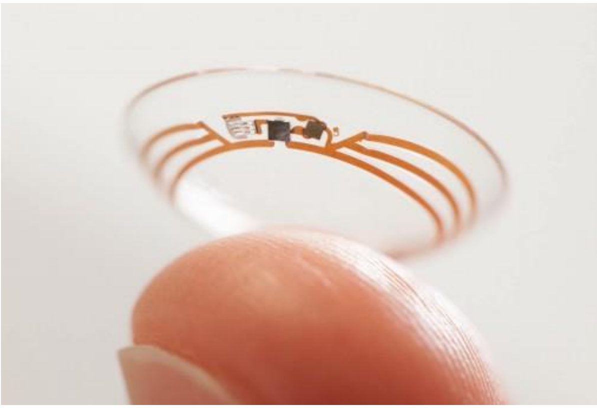
1970

2014

Will sensors be a pharmaceutical ”game changer”?

Wicks, Hotopf, Narayan, Basch, Weatherall & Gray. (2016, under review). “It's a long shot, but it just might work! Perspectives on the future of medicine”. *BMC Medicine*

Smart lens for diabetes (Novartis-Google)



Minimally invasive Technology. Fluid to seep into the sensor and be used to measure blood sugar levels. A wireless antenna, thinner than a human hair, will communicate information to an external device. Google engineers even considered adding LED lights that could warn the wearer for hypoglycemic events, but abandoned the idea.

Technology

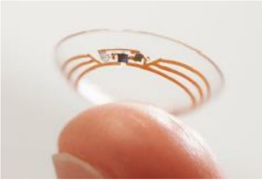
ECG
Respiratory rate
Body Temp
Activity
Coughing

Medical grade ECG

Glucose

GENTAG
Disposable IA system

Metabolic array as glucose, ketones, lactate, uric acid, potassium and sodium

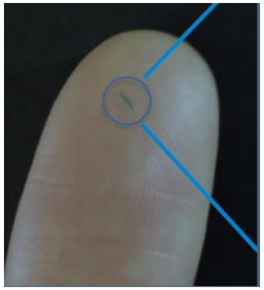


Glucose in tears

3D movement
Fall detection
Respiratory rate
Heart rate

Uric Acid

ECG
Respiratory rate
Activity



Tissue oxygen level
Glucose

Heart Rate
Blood Pressure
Activity
Sleep pattern
Hydration level

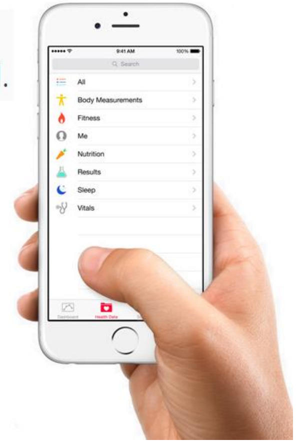
PEF
FEV1

The dawn of eHealth's "Citizen Science"

HealthKit

You're already carrying a powerful medical research tool.

ResearchKit



Mount Sinai asthma trial: 3500 enrolled in 72 hours

gsk

NOVARTIS

Speed ☺ ←→ Representativeness ☹

A vision: clinical trials now…and in the future

Striking the statistical balance

Now – don't lose		Future – need to add
Defined timepoints	+	Continuous measurement
Multiplicity handling	+	Empirical significance
Hypothesis testing	+	Hypothesis generation
Pre-specification	+	Specified adaptation

Case studies

Asthma

Diabetes

Asthma prevalence

25 million*

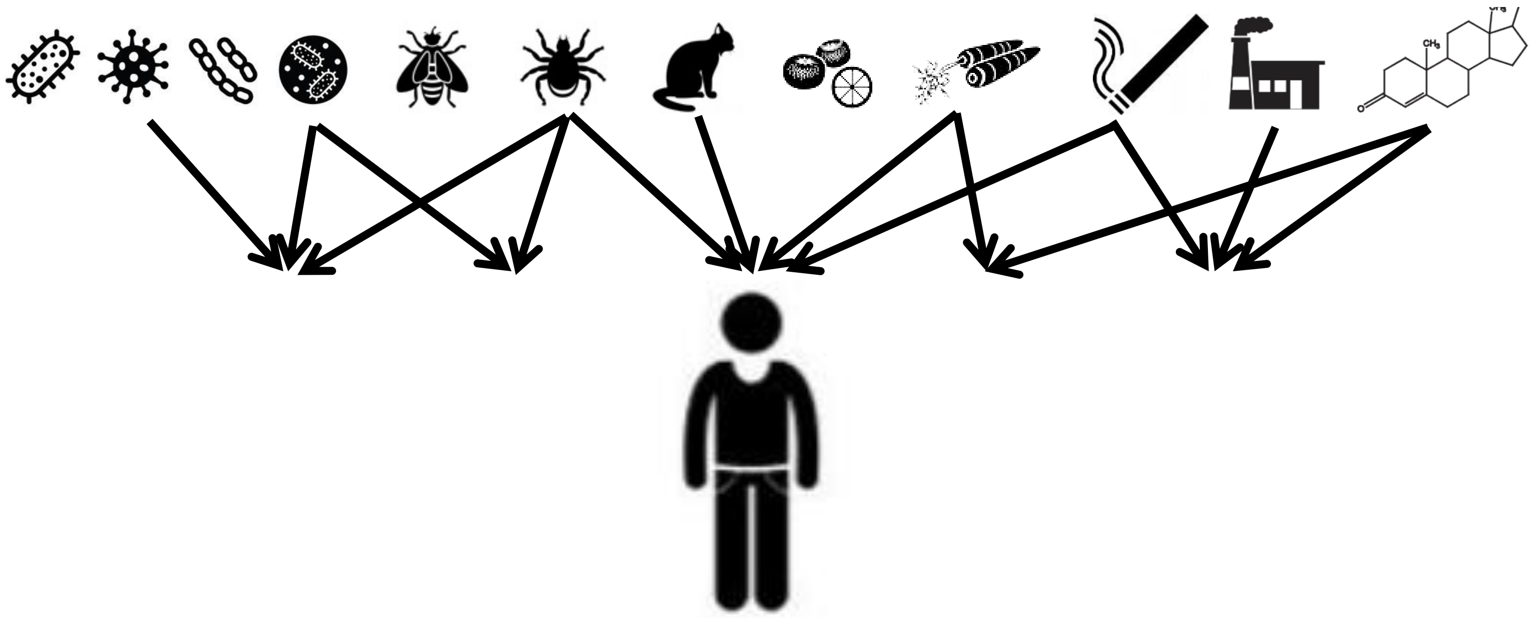
12 million*

3500*

Asthma cost the US about **$56 billion** in medical costs, lost school and work days, and early deaths in 2007*.

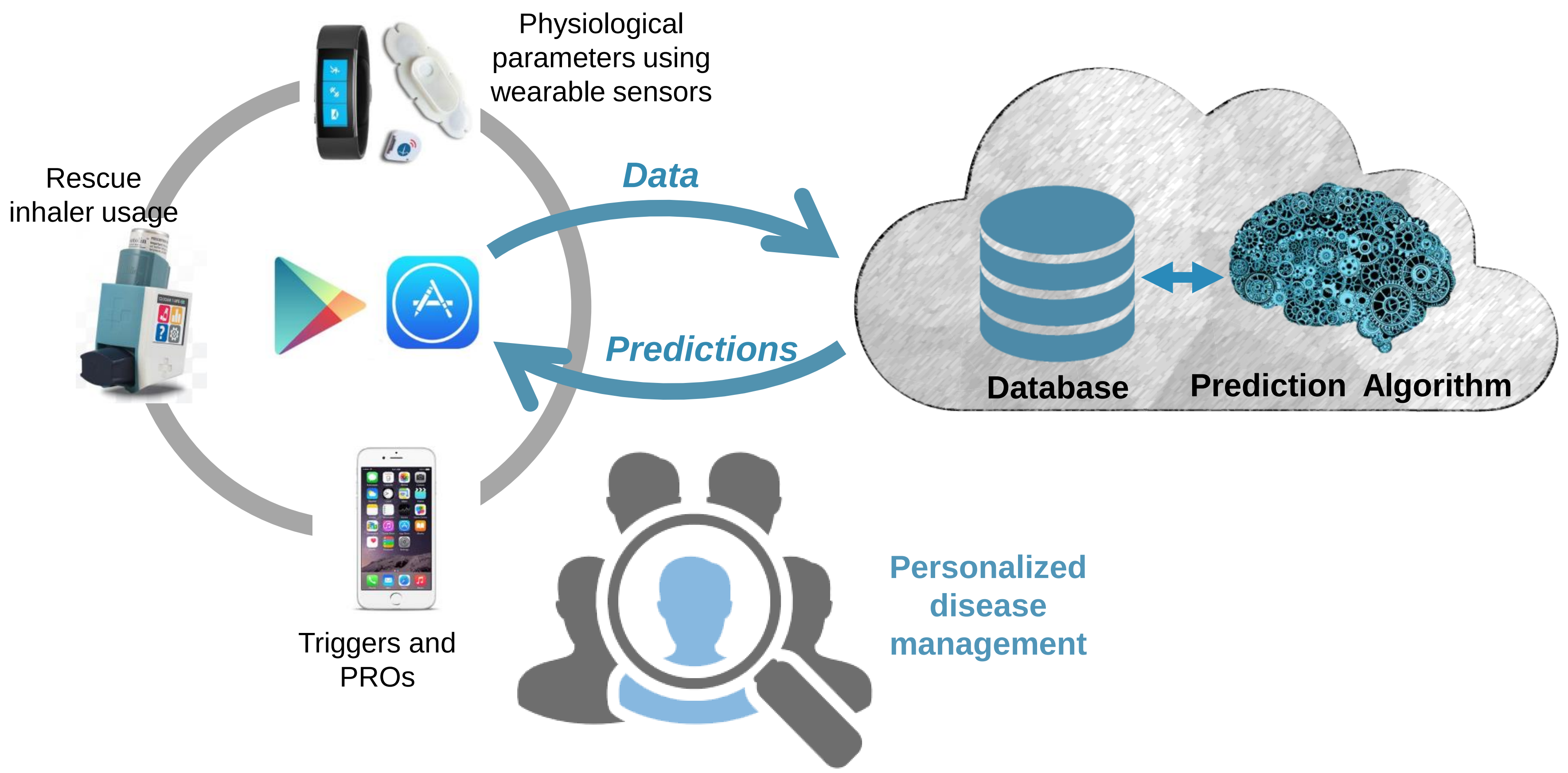
**Centers for Disease Control and Prevention*

Asthma triggers



Asthma originates from the complex interplay between individual’s genetic and environmental factors

Case study: Monitoring the level of disease control in asthma patients



Salient features of the data and modeling

Rescue
Medication

Pollen
Count

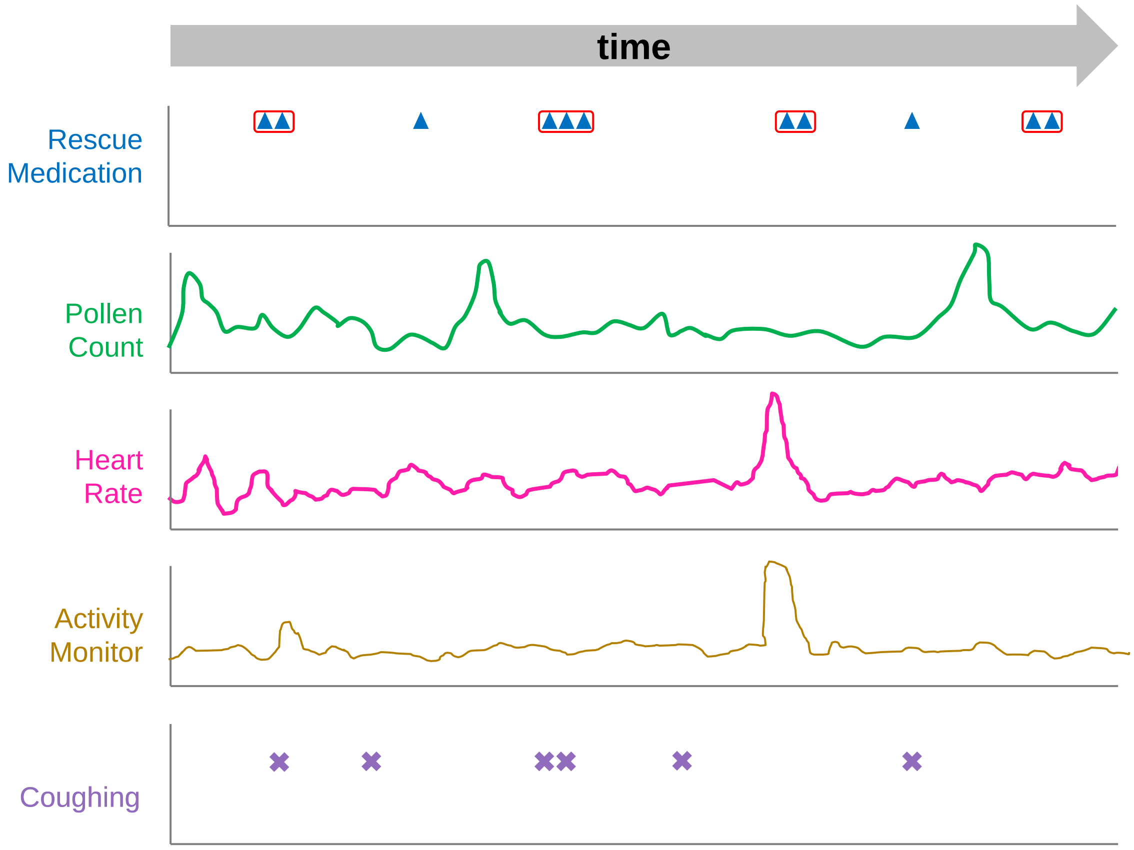
Heart
Rate

Activity
Monitor

Coughing

Acknowledgement: Bhaskar Dutta

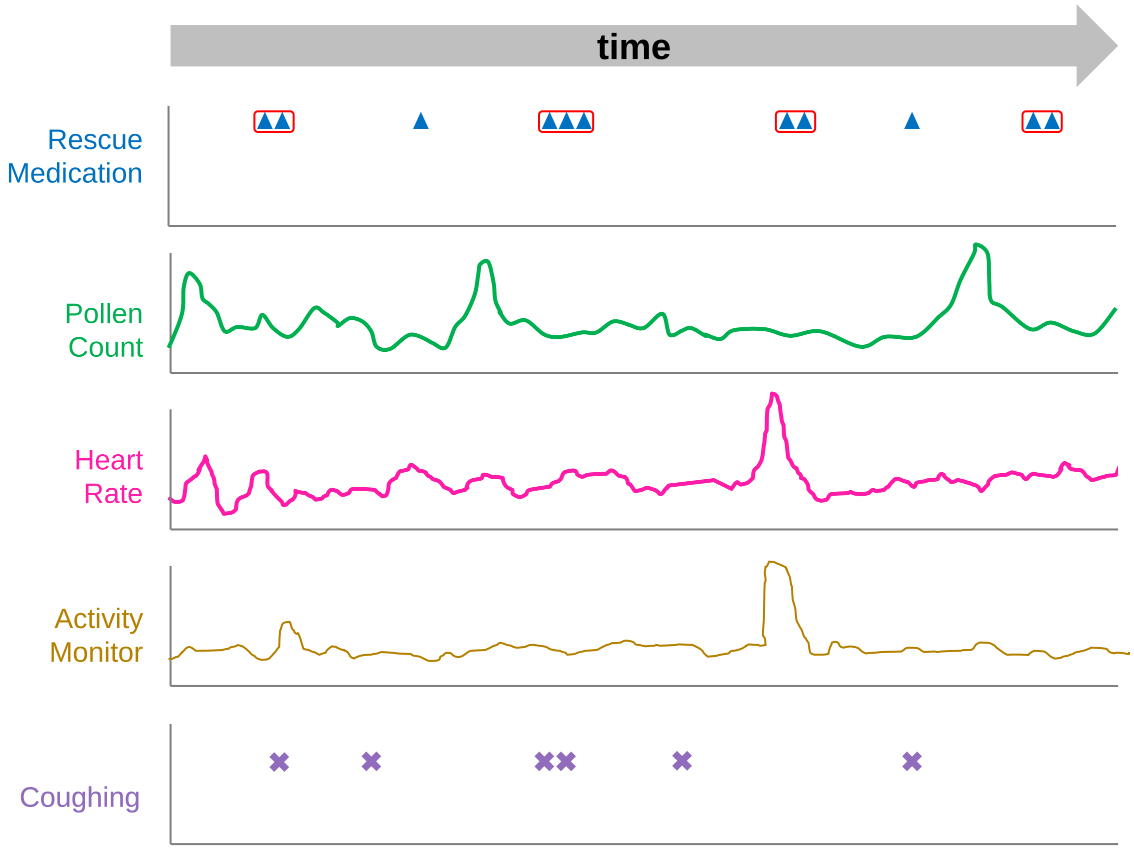
Salient features of the data and modeling



- Uncontrolled asthma

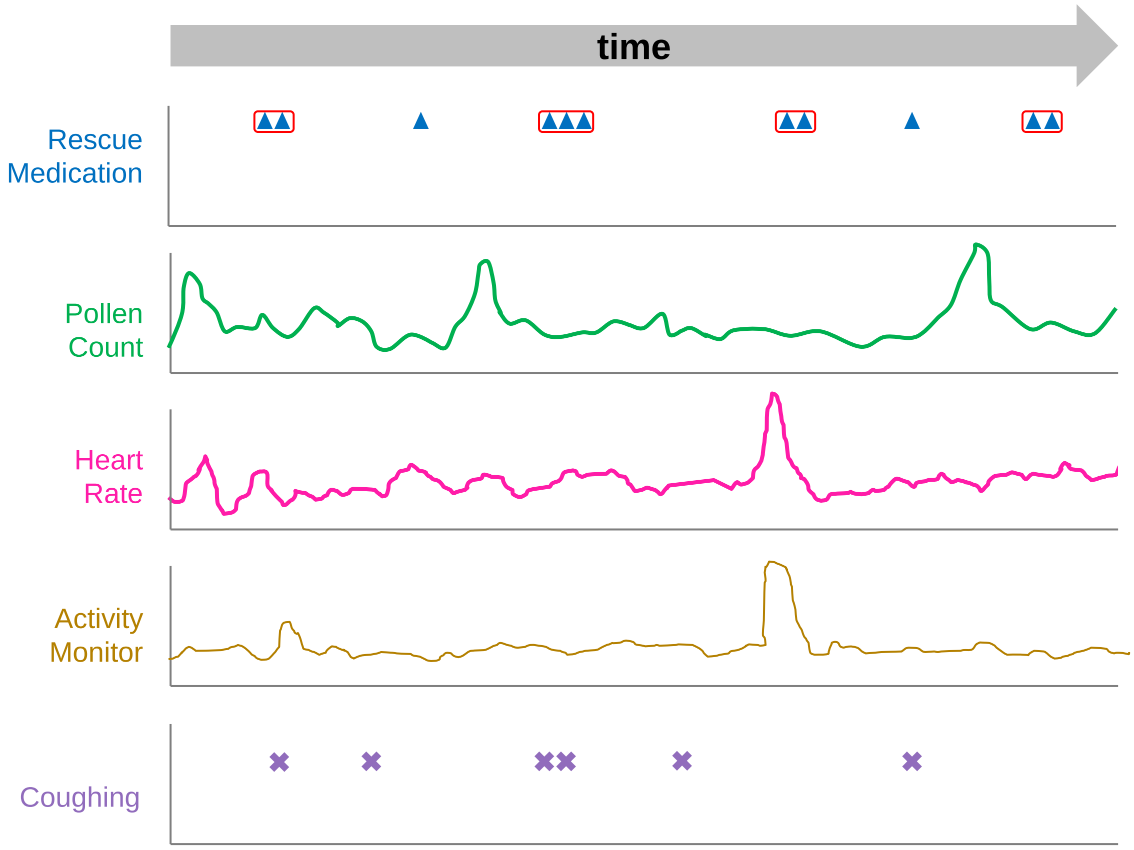
Clinical definition of uncontrolled asthma:
Two or more consecutive inhaler puffs

Salient features of the data and modeling



- Uncontrolled asthma
- Classification problem
  - *Can we predict when/whether a patient is going to have uncontrolled asthma event?*
  - *Which variables are the best personalized predictors (feature importance) of the event?*

Salient features of the data and modeling



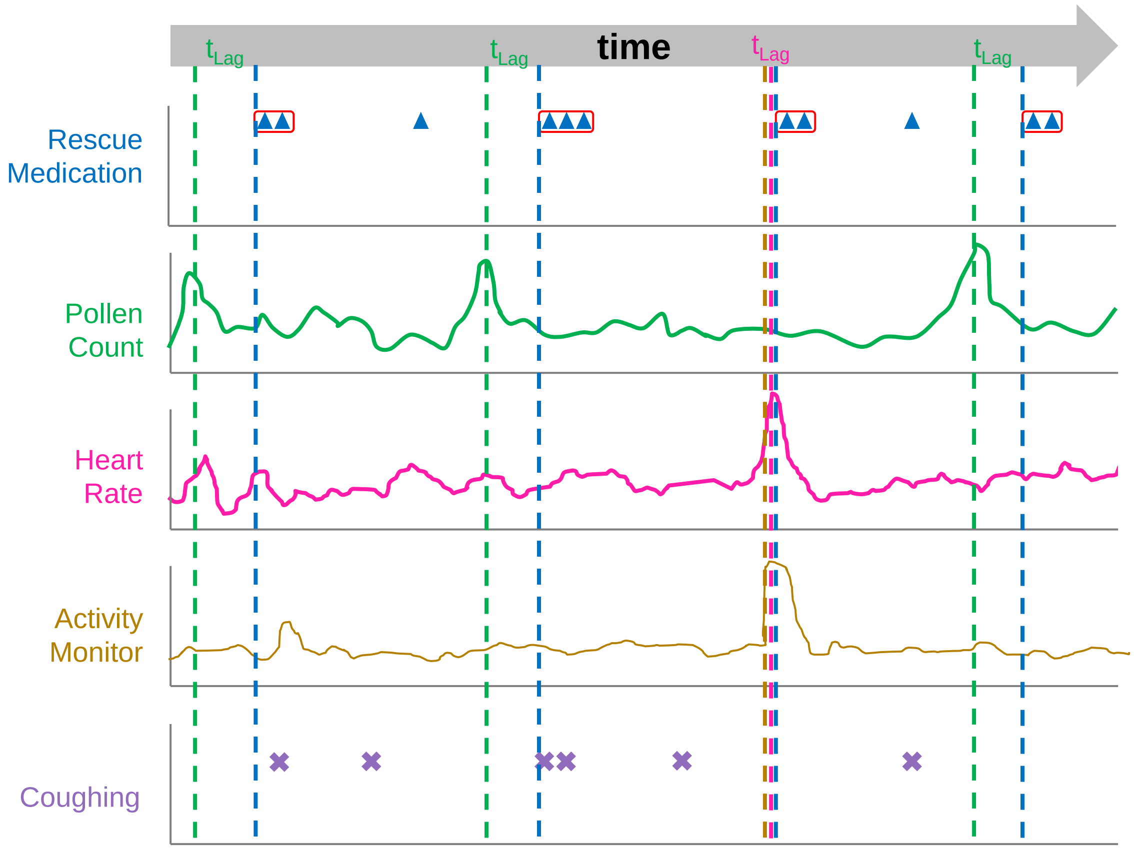
- Uncontrolled asthma
- Classification problem
- Dynamic model

- *Predicted and predictor variables are time series*

$$E = f(x_t^1, x_t^2, \ldots, x_t^n, x_{t-1}^1, x_{t-1}^2, \ldots, x_{t-1}^n, \ldots x_{t-m}^1, x_{t-m}^2, \ldots, x_{t-m}^n)$$

- *For separation of training and test sets, dynamic aspect should be considered*

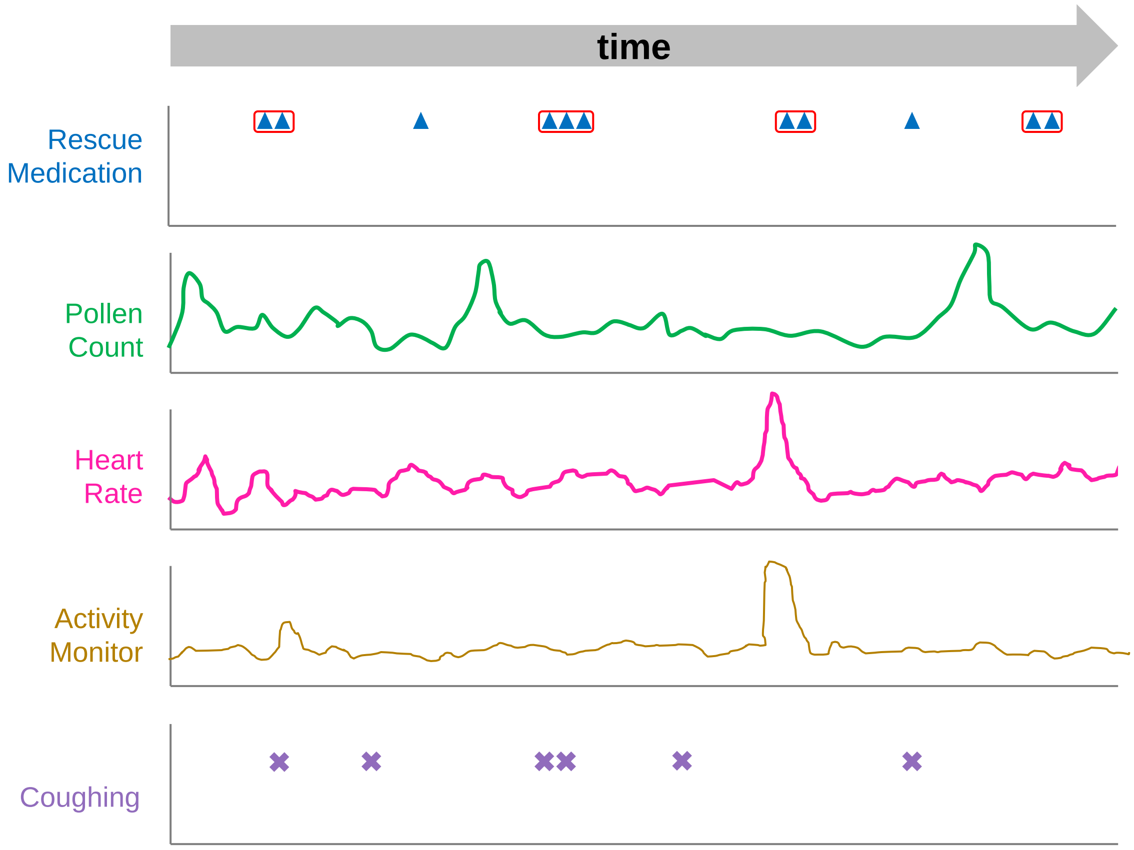
Salient features of the data and modeling



- Uncontrolled asthma
- Classification problem
- Dynamic model
- Variable time lag

- *Features might have different time lags (e.g., pollen count has greater time lag compared to Heart Rate)*
- *Time lags for different variables should be estimated before feeding the model*

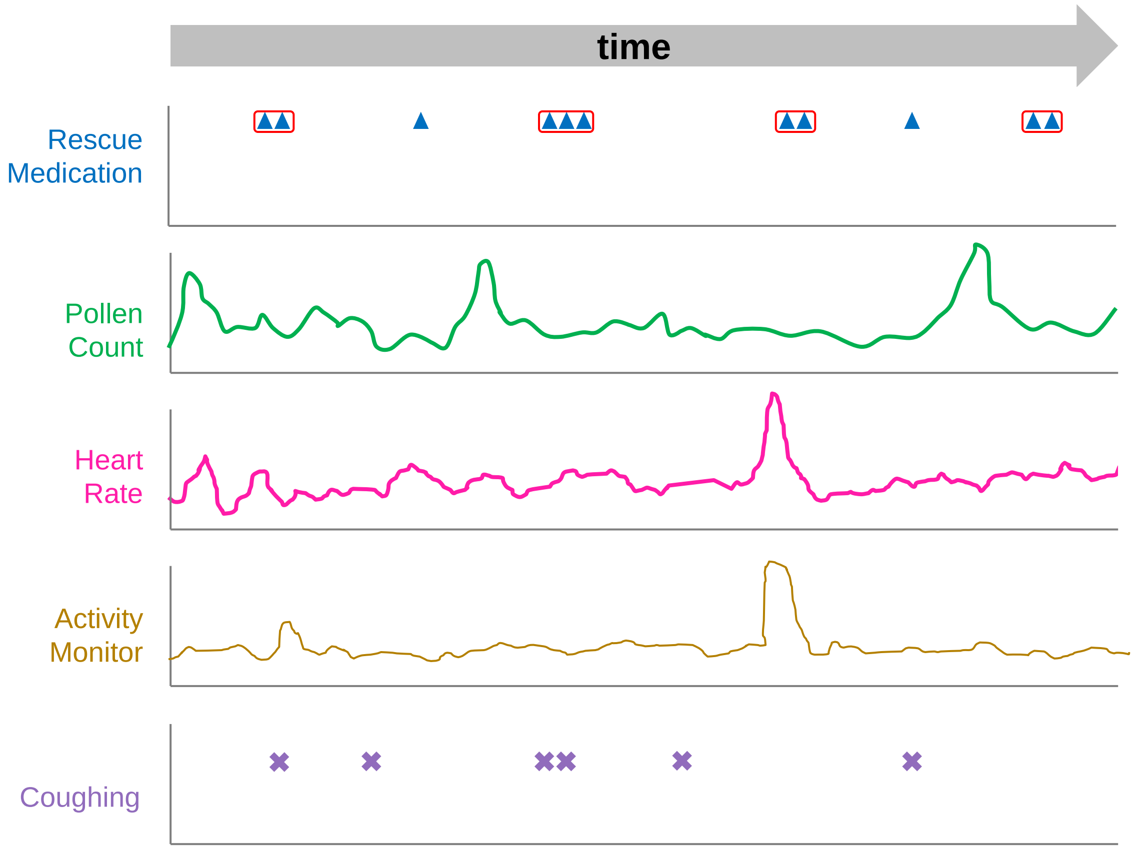
Salient features of the data and modeling



- Uncontrolled asthma
- Classification problem
- Dynamic model
- Variable time lag
- Patient specific model

As asthma triggers differ between patients, personalized models should be developed

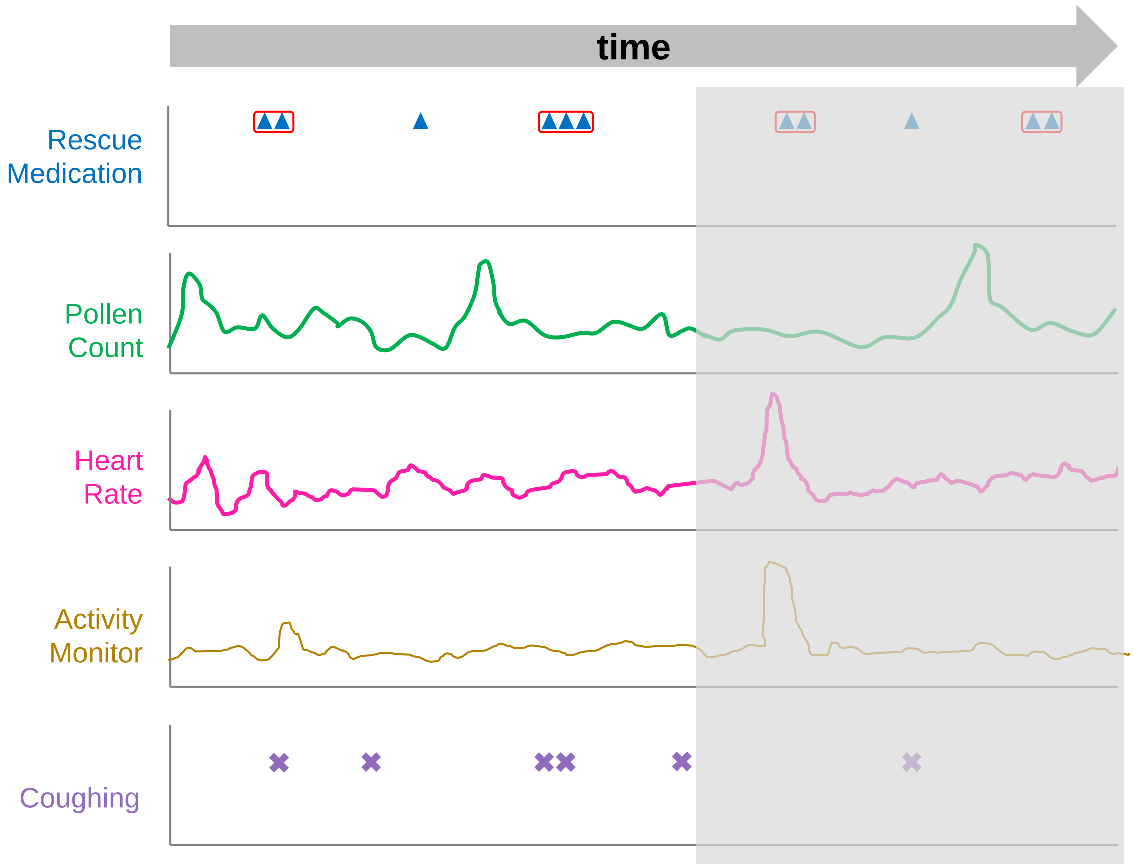
Salient features of the data and modeling



- Uncontrolled asthma
- Classification problem
- Dynamic model
- Variable time lag
- Patient specific model
- Heterogeneous features

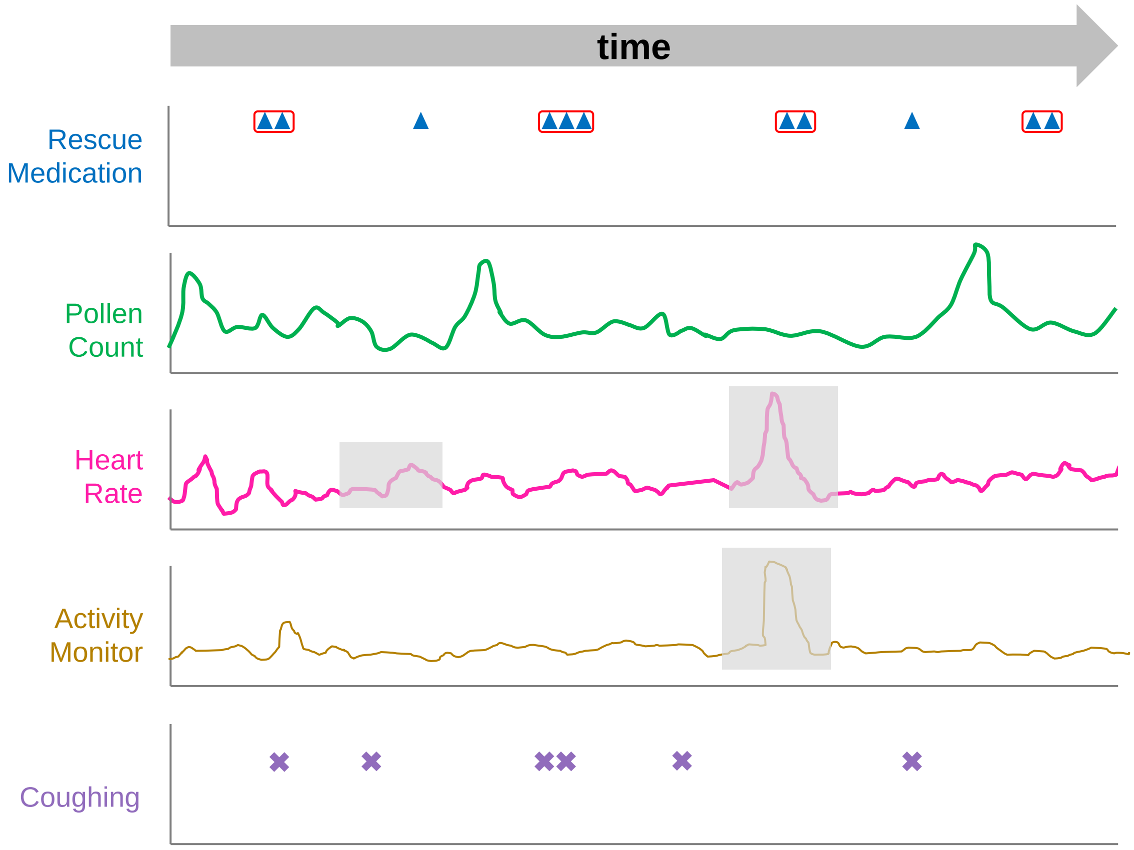
Here Pollen Count, Heart Rate, and Activity Monitor measurements are continuous numbers, but Coughing is categorical.

Missing data analysis: different scenarios



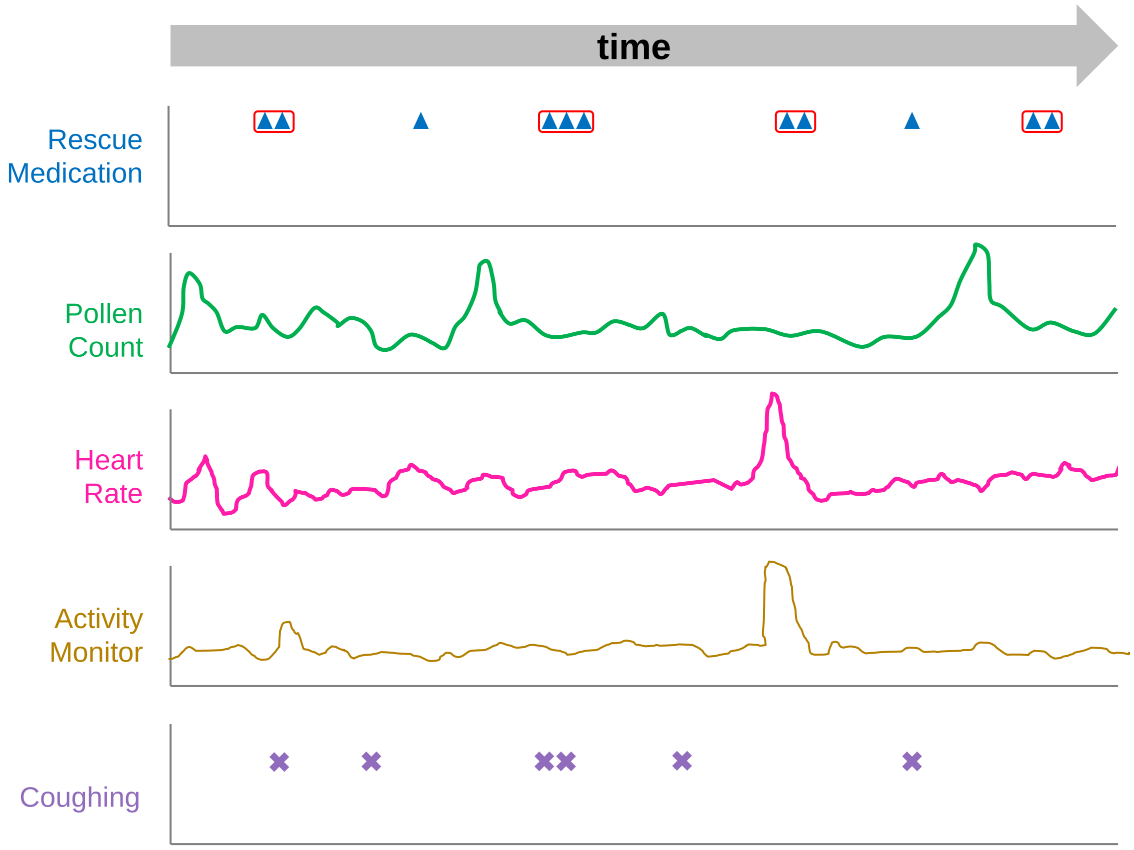
- Loss to follow up:
  - Use available data if feasible

Missing data analysis: different scenarios



- Loss to follow up:
  - Use available data if feasible
- Partially missing measurements:
  - Define max allowable missing data
  - Patient specific median
  - k-nearest neighbor
  - Last value carried forward
  - Potential erroneous imputation

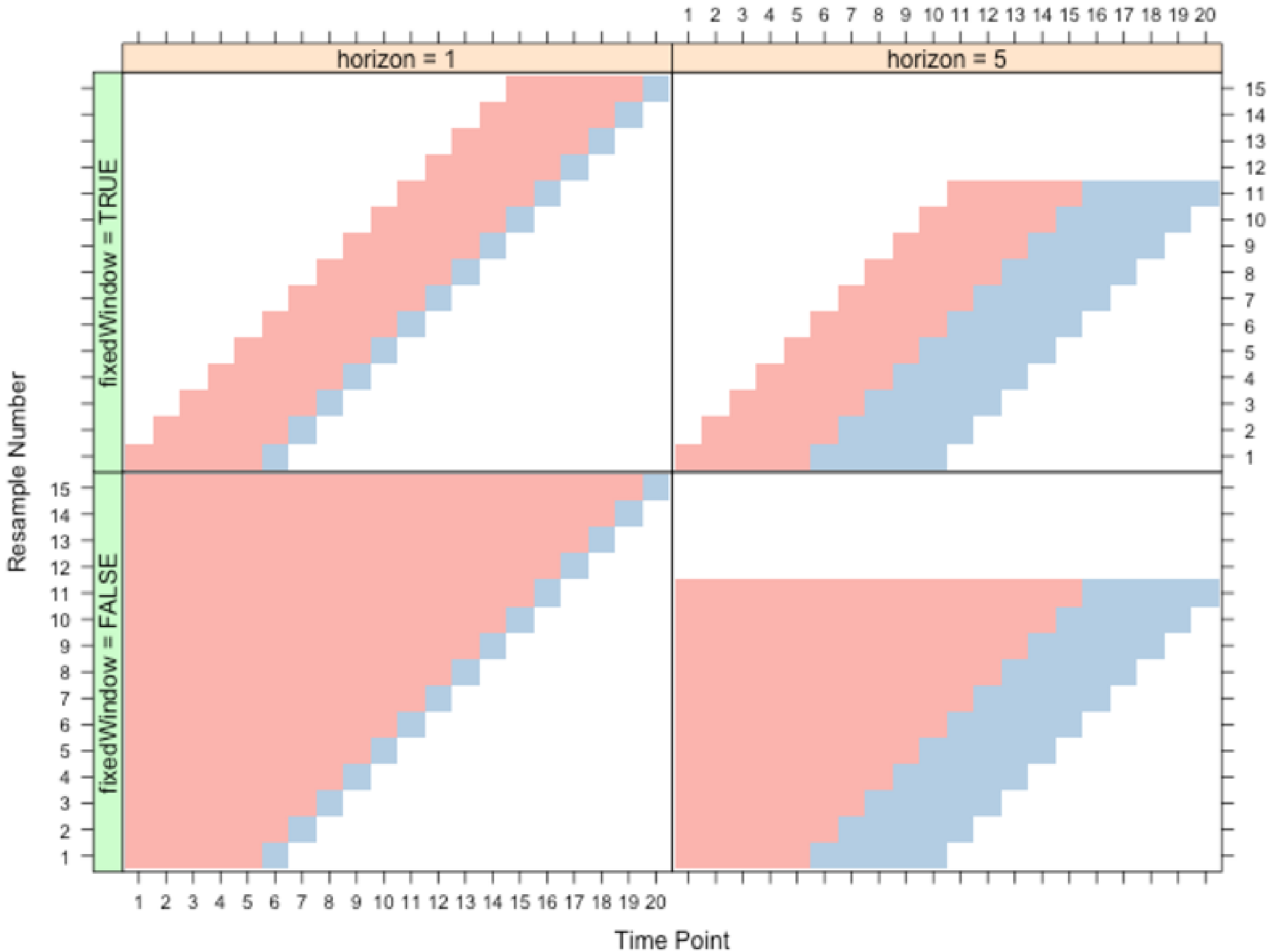
Sanity check and transformation of data



- Data sanity checks:
  - Within feasible range
  - Constant value
  - Null value
- Data transformations
  - Consistent unit of measurement
  - Correct time scale
  - Include new features (temperature difference)
  - Incorporate composite features

Acknowledgement: Bhaskar Dutta

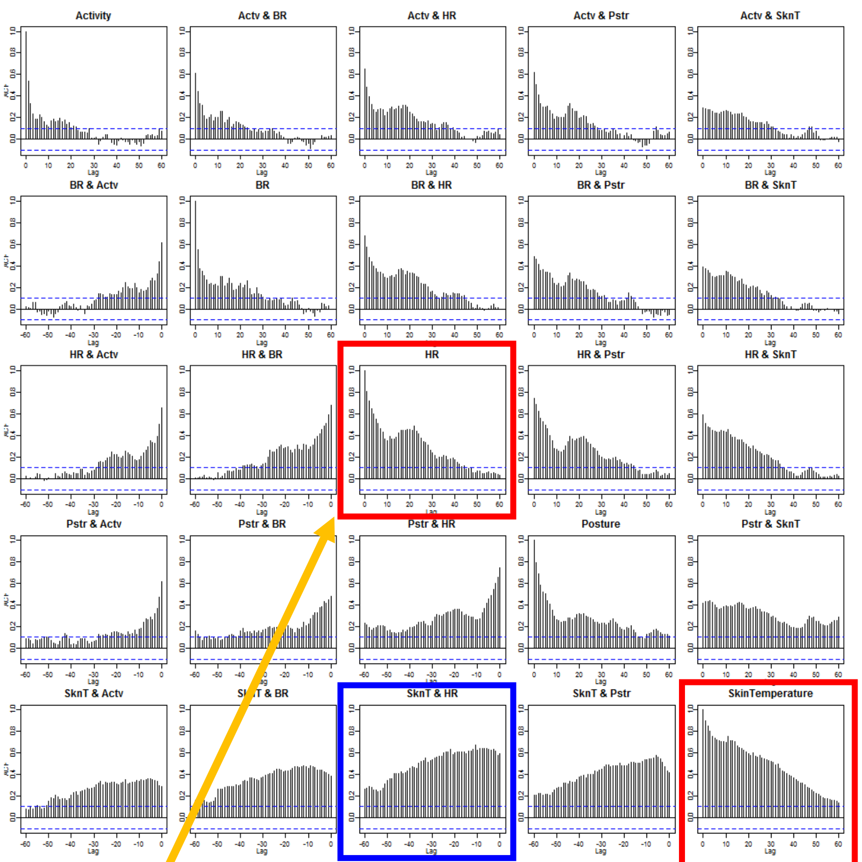
Splitting time series data for training and test set



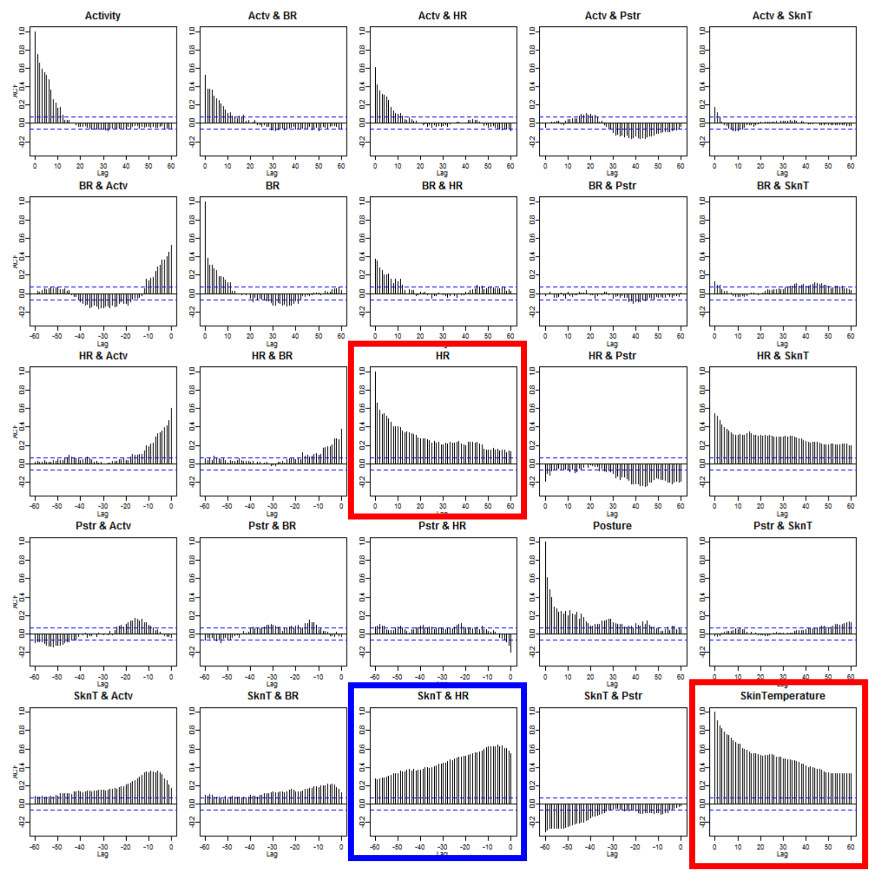
Athanasopoulos and Hyndman, Forecasting: principles and practice

Comparing measurements: correlation and auto-correlation

BioPatch_UserID 13_27 June 2016



BioPatch_UserID 15_29 June 2016



Heart rate

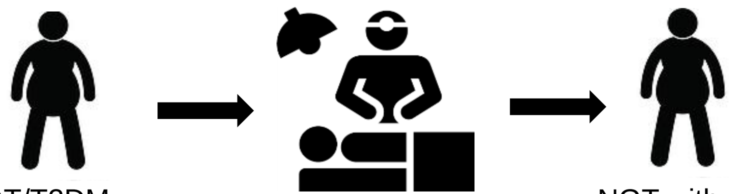
Skin temperature

Smart mobile technology for near real-time data collection

Case study: Continuous glucose monitoring in bariatric surgery patients with dysregulated glucose metabolism – The COLUMBO study

The Trial

Non-interventional study of patients undergoing bariatric surgery



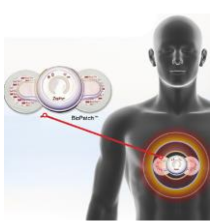
IGT/T2DM
1 month
pre-surgery

Bariatric surgery

NGT with risk of
hypoglycemic events
1 month post-surgery

Continuous data collection

1. Plasma glucose
2. Heart rate
3. Activity
4. Sleeping pattern
5. Food intake



Intended Learnings

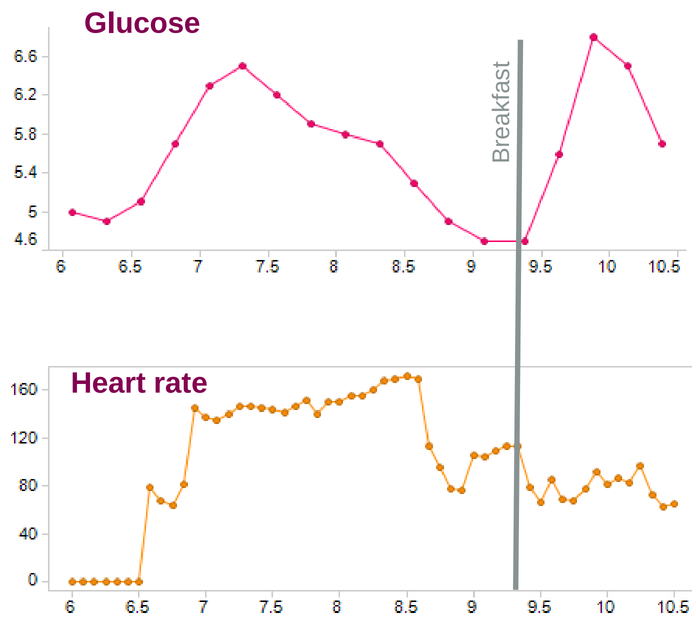
- Increased understanding of glycemic regulation in bariatric surgery patients
- How to conduct the trial operationally
- How to ensure patient privacy
- Ways to analyze continuous biomarkers

Team: Jesper Havsol, Elisabeth Nyman, Lisa Bennett, Martin Karpefors, Cassie Gregson, Björn Carlsson and Cecilia Karlsson

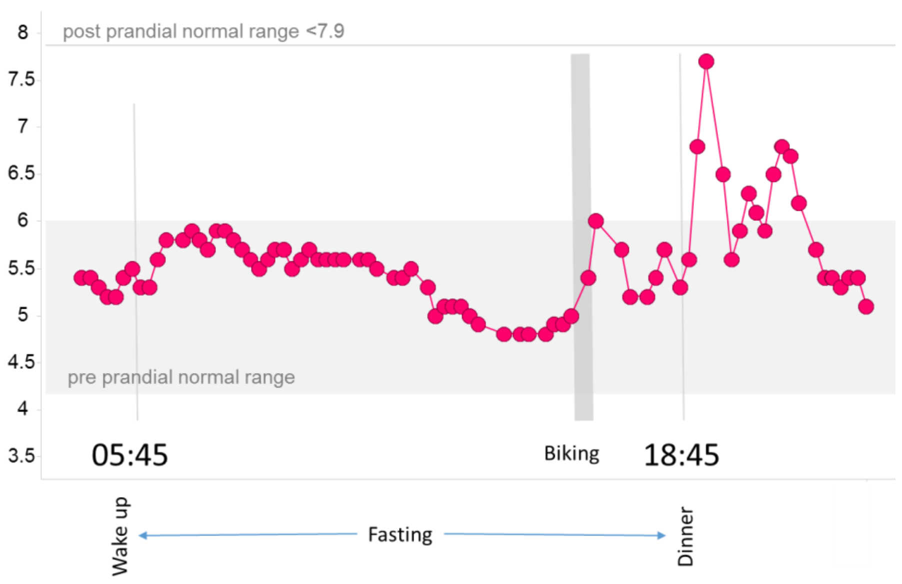
Two live sensor examples generating new insights

Physical exercise generate glucose peaks of the same magnitude as food intake

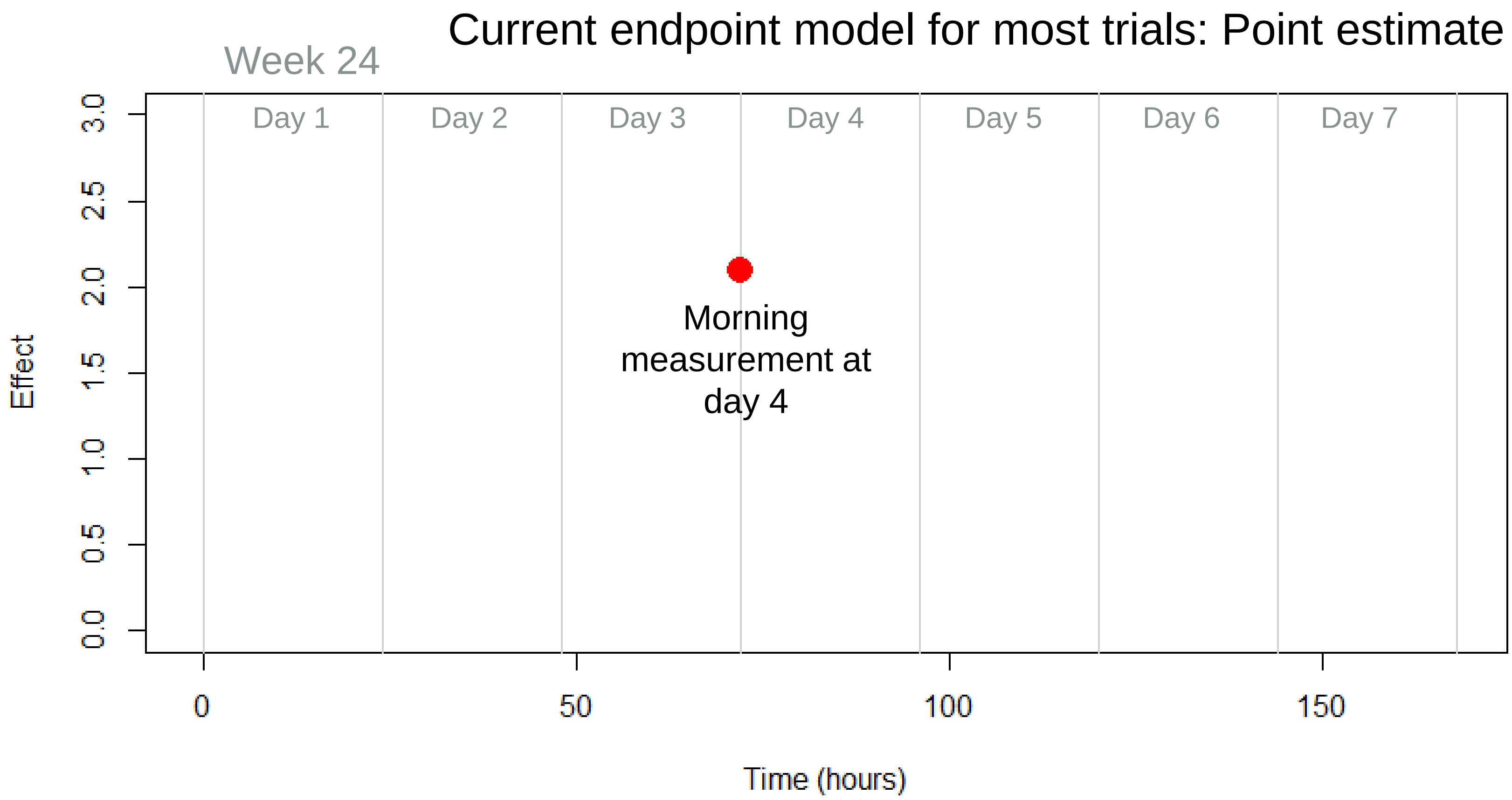
Two-hour morning run before breakfast



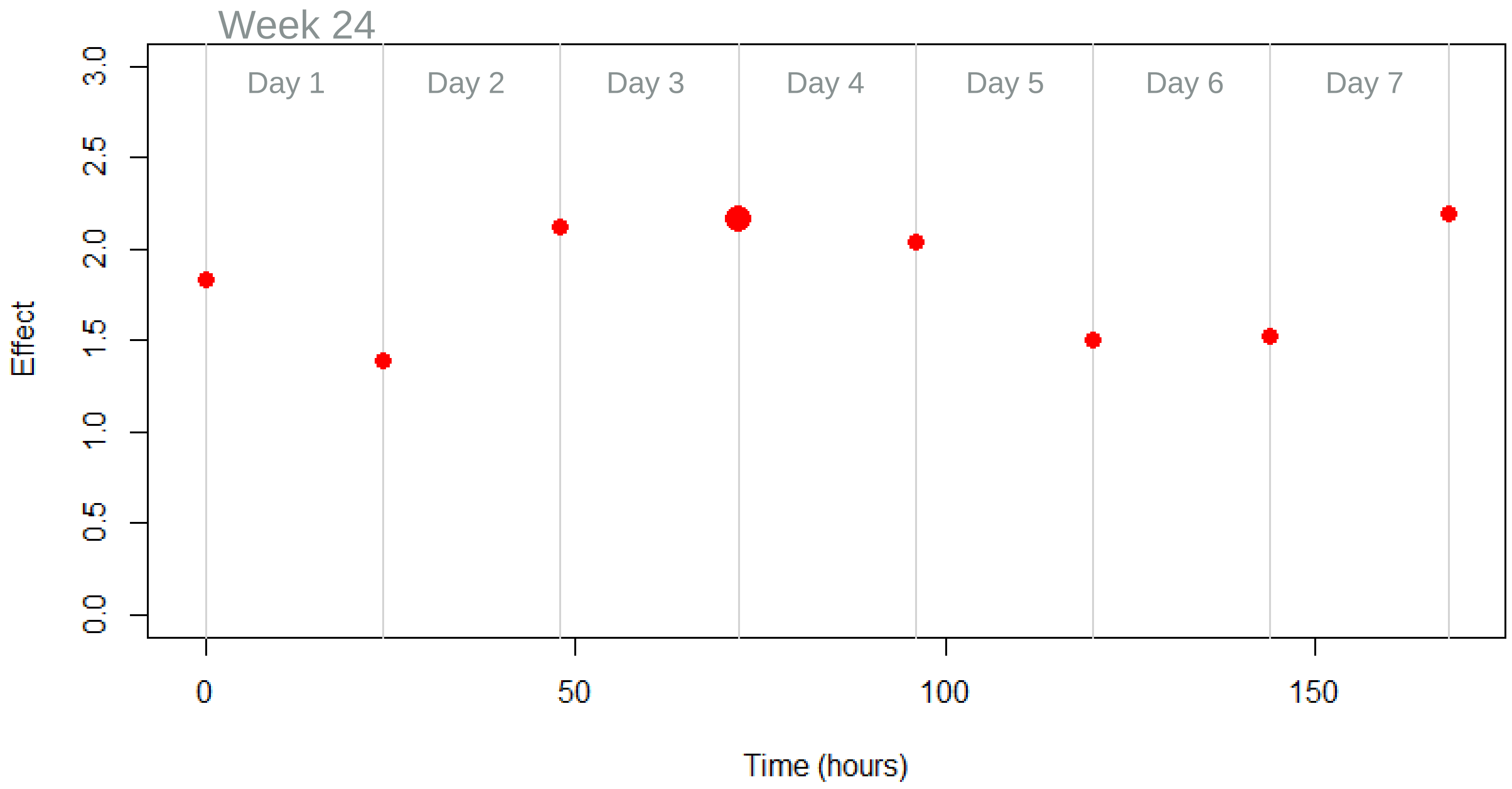
Almost all day fasting



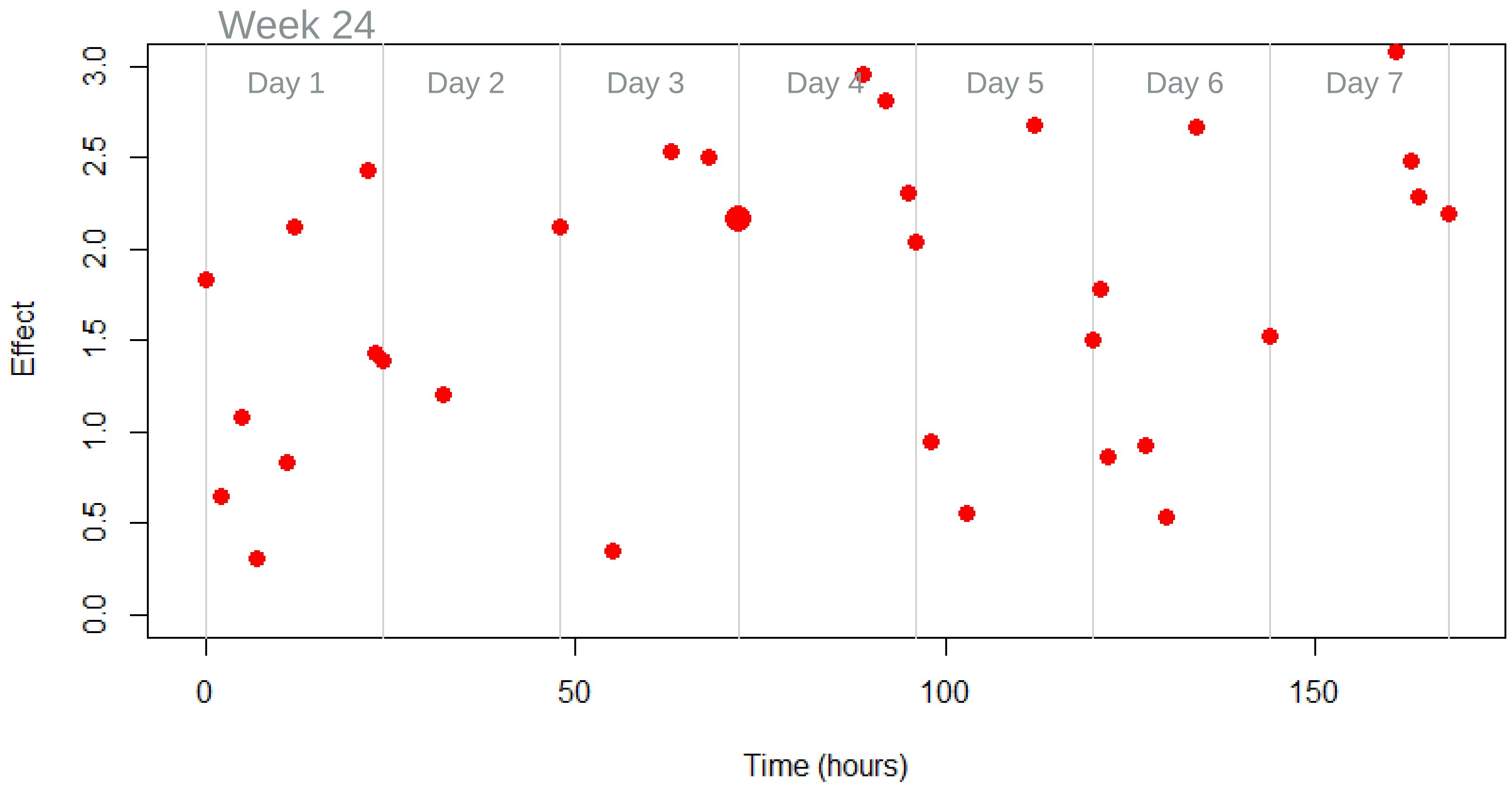
Example of sensor data benefit (simulated):



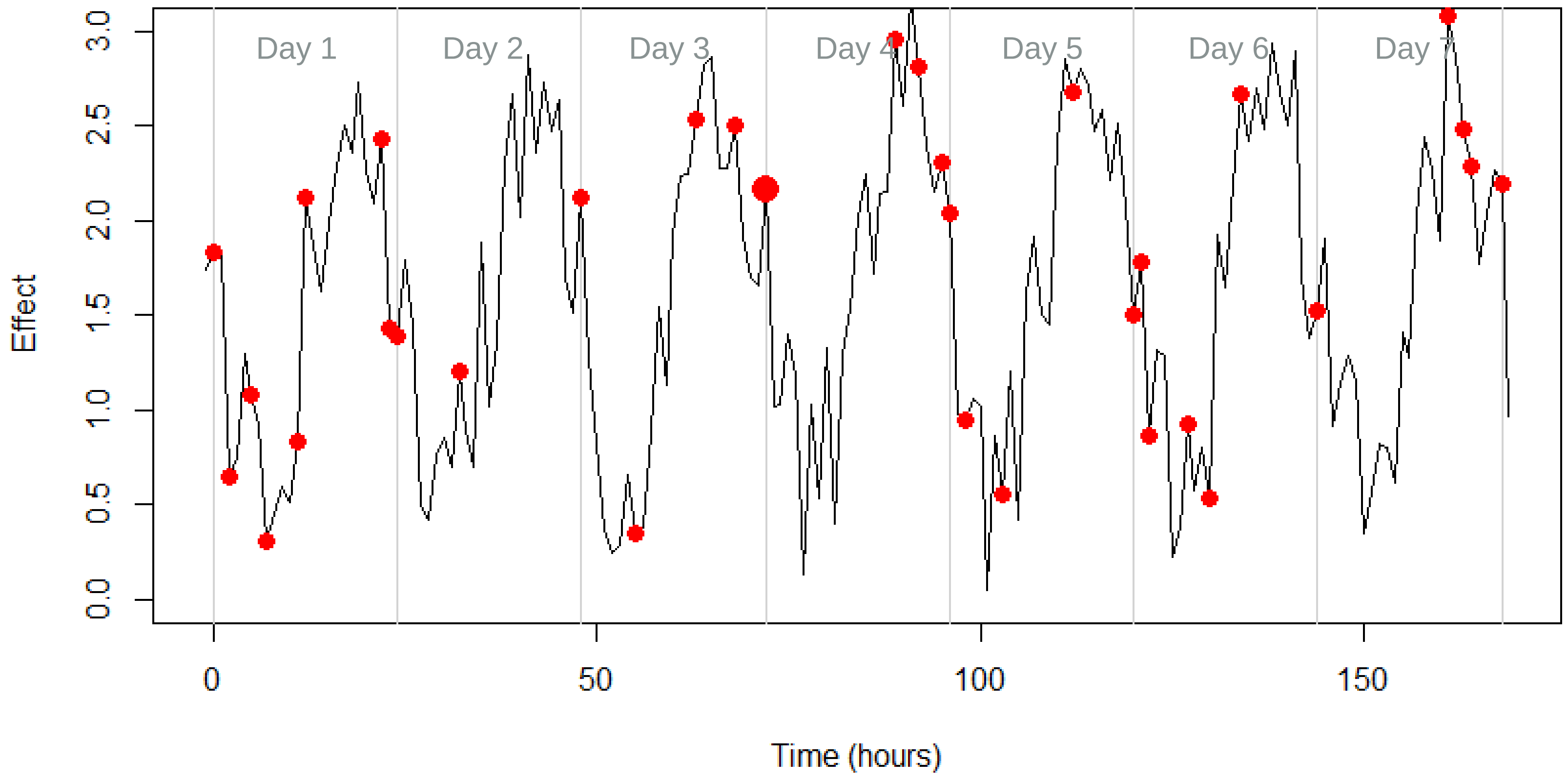
But, there is an inter-day variability



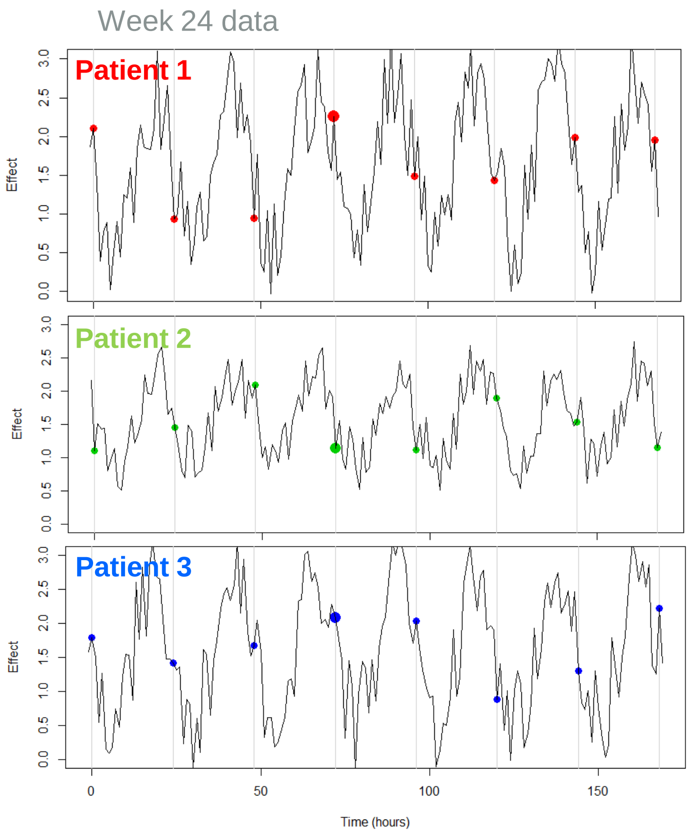
And, there is an inter-measurement variability

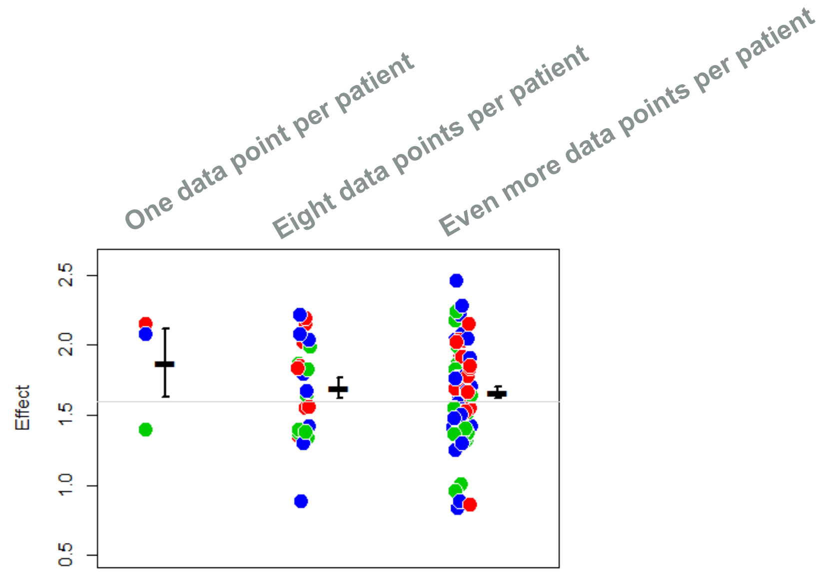


More data points enables modeling approaches and disease understanding



In essence: Better precision with more data





$$\mu + \delta_i + \vartheta_{ij} + \omega_{ijk}$$

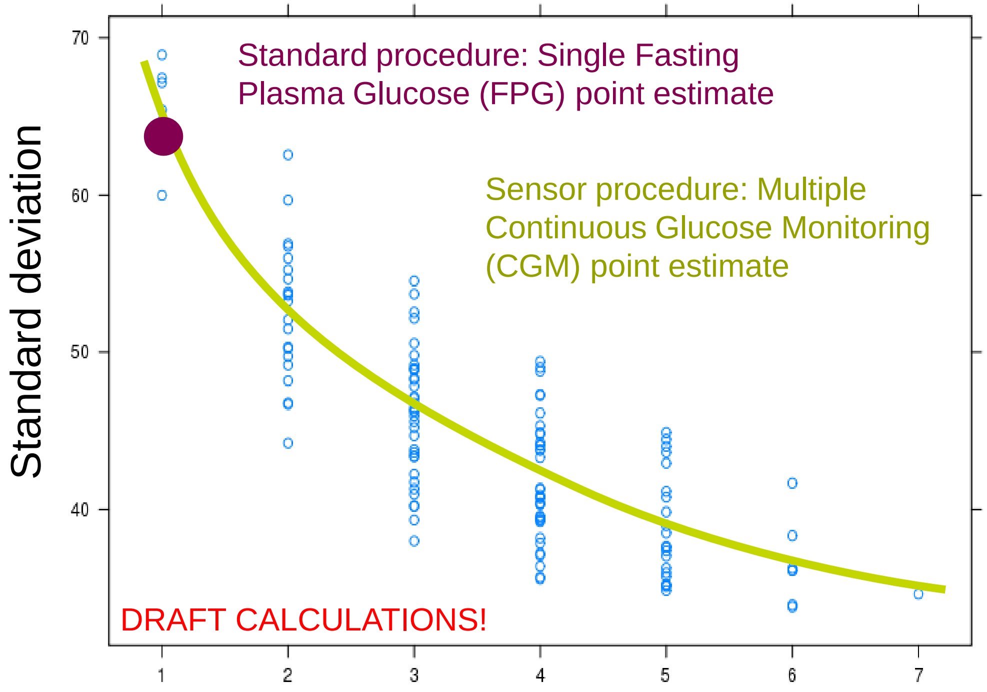
Global mean

Inter individual variability

Inter day variability

Inter measurement variability

Multiple sensor values decrease variance, resulting in smaller clinical trial sample size



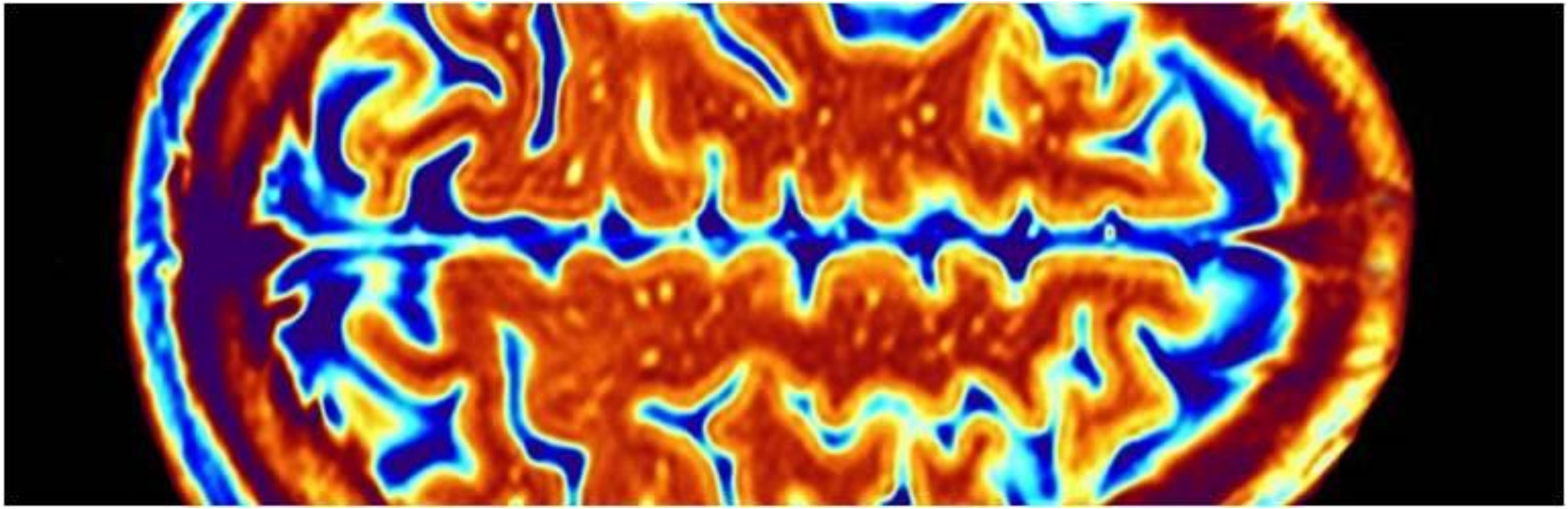
Number of data points used for calculation

Data used in this example is taken from an internal CGM study using Dexcom devices.

To resemble the fasting state in the FPG measurements, the CGM values at 04:00 were used

Thanks to Jesper Havsol and Fredrik Öhrn

Closing thoughts



Streaming data in health
A grab bag of challenges

The wisdom of the crowd

Heterogeneity

Missingness – how to handle?

Trade-off between piloting and trialling

Latent class inference

Changepoint detection

Decision analytics – how much is enough?

What is the purpose of statistics here?

Seasonality

What are we really measuring?

Closed feedback loops

Is dirtiness a virtue?

Correlations & 'proxyness'

Predictive analytics and test sets

Time lag

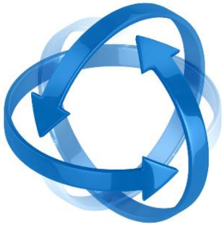
Making sense of big data…key challenges

1 Contextualisation

2 Intelligent aggregation

3 Applicability

4 Agility




Questions
Answers