Machine Learning and Counterfactual Reasoning for "Personalized" Decision-Making in Healthcare

Suchi Saria
Assistant Professor
Computer Science, Applied Math & Stats
and Health Policy
Institute for Computational Medicine

Hossein Soleimani
Postdoctoral Fellow
Computer Science

NSF
GORDON AND BETTY MOORE FOUNDATION
Google
THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH
DARPA

Machine Learning and Personalization

Machine Learning and Personalization

- Relevant to anyone with interest in personalization
  - Domains: education, recommender systems, retail
  - Focus of this talk is on **Medicine**

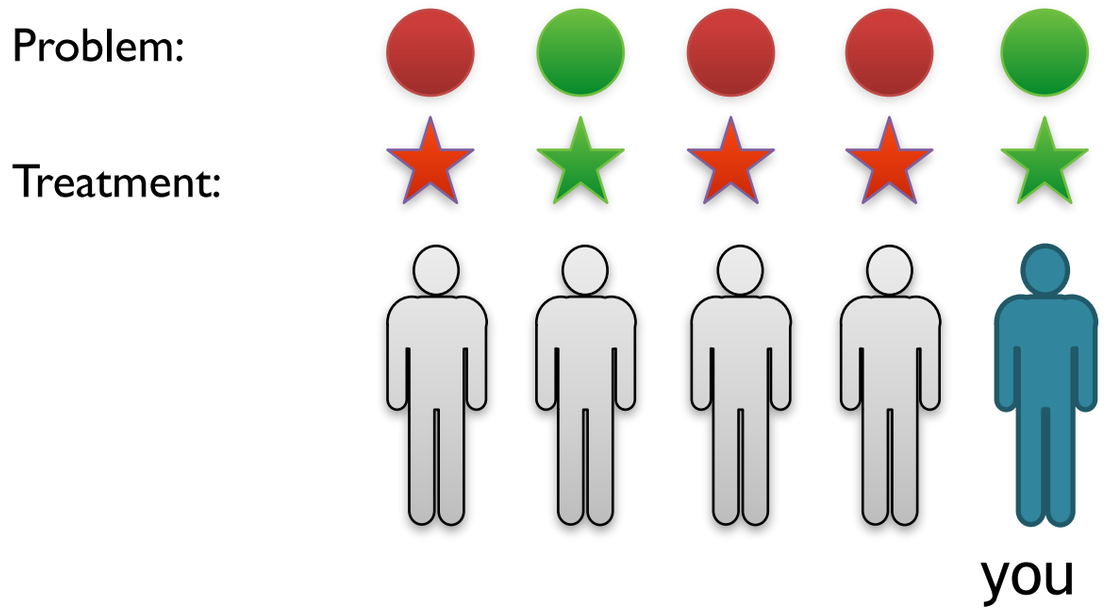
Machine Learning and Personalization

- Relevant to anyone with interest in personalization
  - Domains: education, recommender systems, retail
  - Focus of this talk is on **Medicine**

Style: Rather than a broad survey, we focus on a narrow, core set of ideas that motivates one way of approaching the problem.

Describing key ideas with pointers to papers.

Classical view — Randomized Trials, Clinical Practice Guidelines and *Population models*



- Based on a *coarse* set of characteristics, define a population P.
- Conduct trials to determine Intervention A vs B.
- Define guideline to assign intervention to P.

Often referred to as population models. Does not adequately account for individual-specific variability.

Classical view — Randomized Trials, Clinical Practice Guidelines and *Population models*

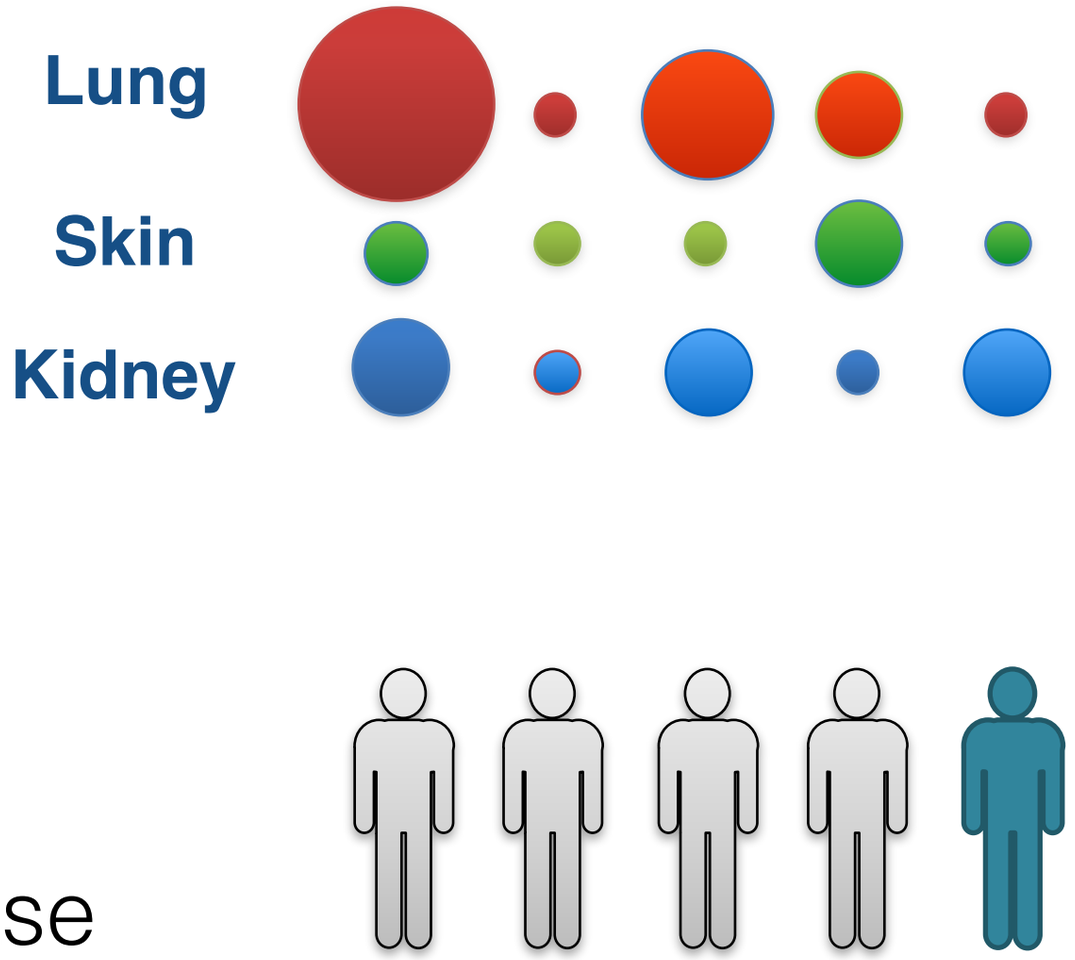
- **Example**: managing high blood pressure in adults

James, Oparil, Carter, et al. 2014

- "Recommendation 8":
  - In population ≥18 with chronic kidney disease (CKD)
  - Initial anti-hypertensive treatment should include:
    - (1) ACEI or (2) ARB
  - Use for **all** CKD patients regardless of race or diabetes status

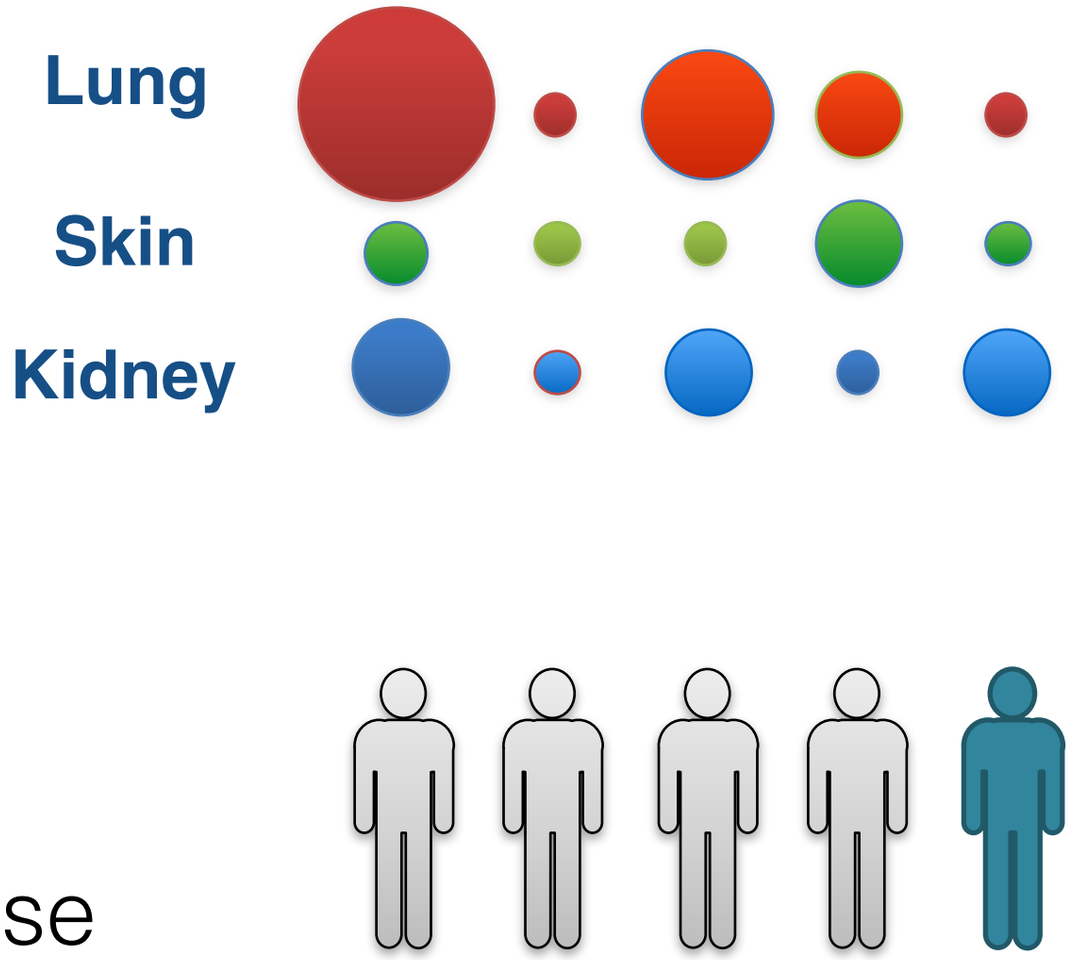
*(1) Indications are **coarse**.*
*(2) **Not relevant to many** in the population — people with multiple diseases or allergies.*

Scleroderma - an example disease



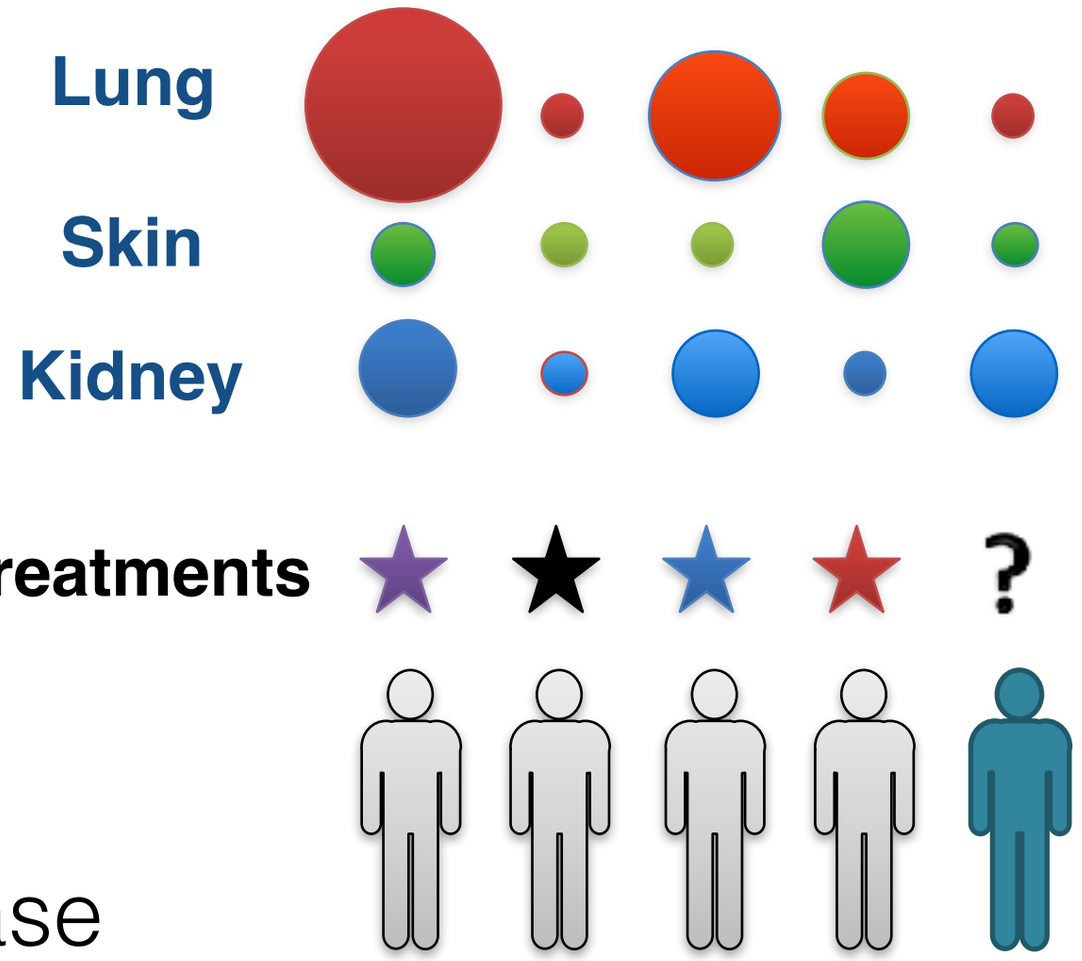
- **Systemic** autoimmune disease

Scleroderma - an example disease



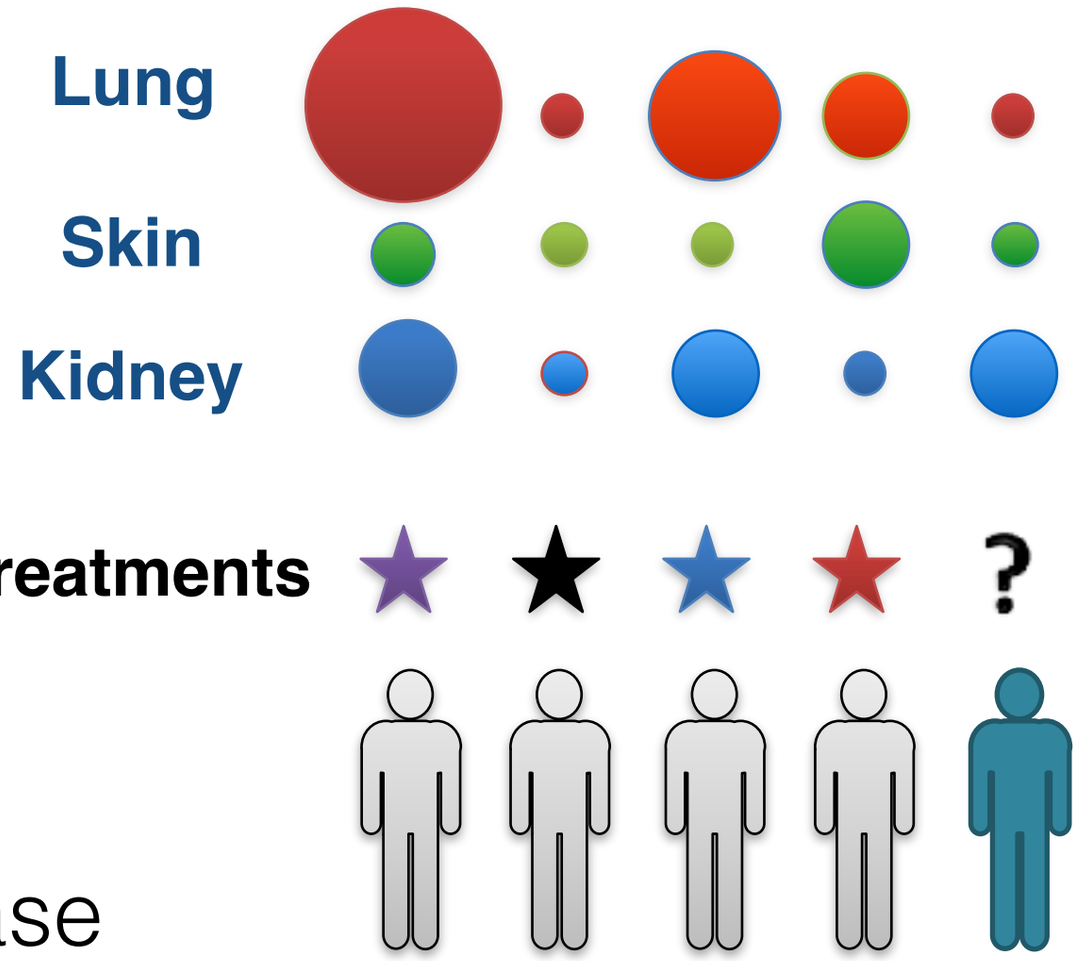
- **Systemic** autoimmune disease
- Affects skin, lung, kidney, intestines, vasculature

Scleroderma - an example disease



- **Systemic** autoimmune disease
- Affects skin, lung, kidney, intestines, vasculature

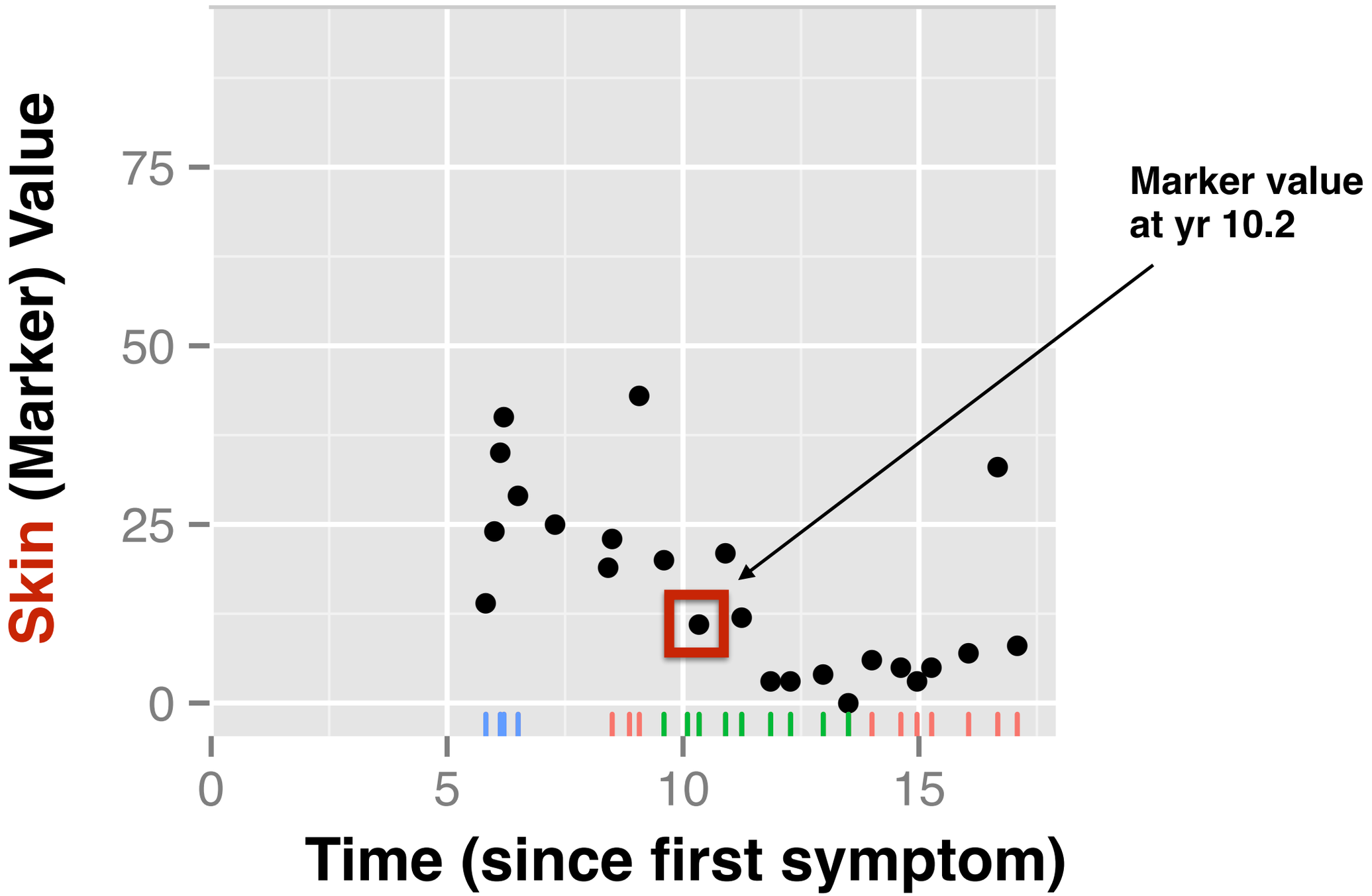
Scleroderma - an example disease



- **Systemic** autoimmune disease
- Affects skin, lung, kidney, intestines, vasculature

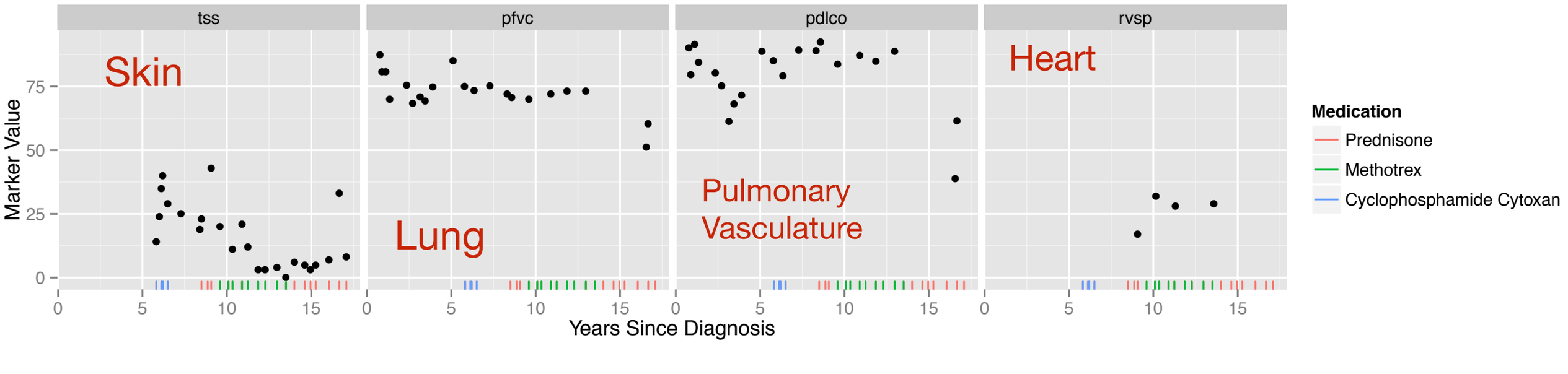
80 other autoimmune diseases —
lupus, multiple sclerosis, diabetes, Crohn's

Tracking an Individual over Time



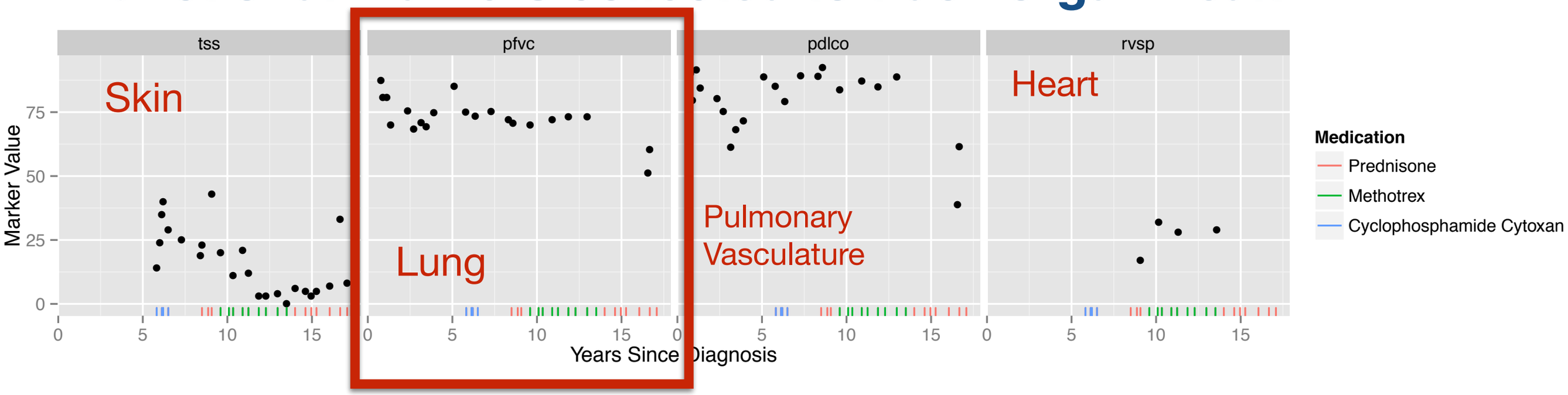
Tracking an Individual over Time

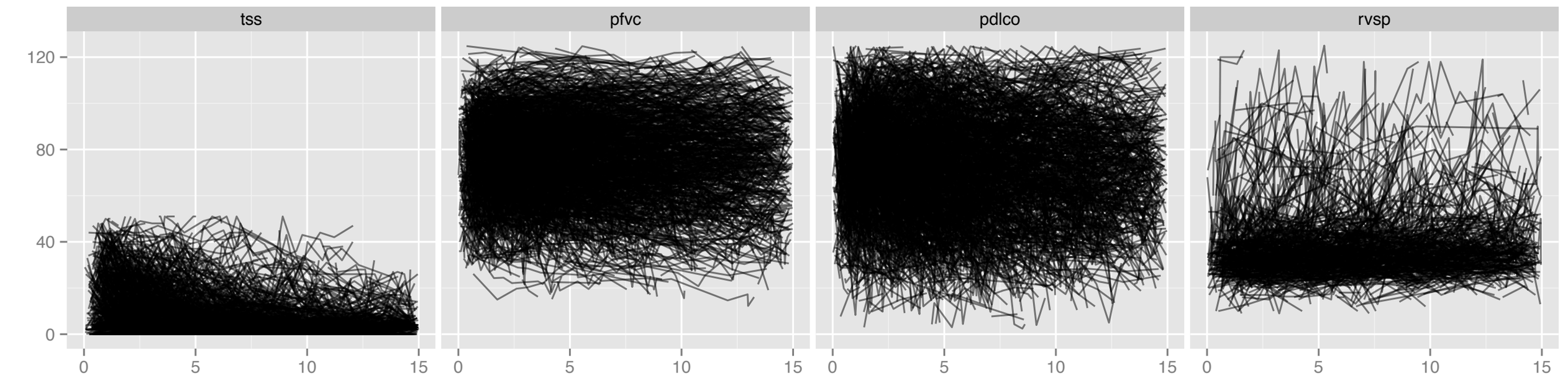
- **Functional markers collected to track organ health**



Tracking an Individual over Time

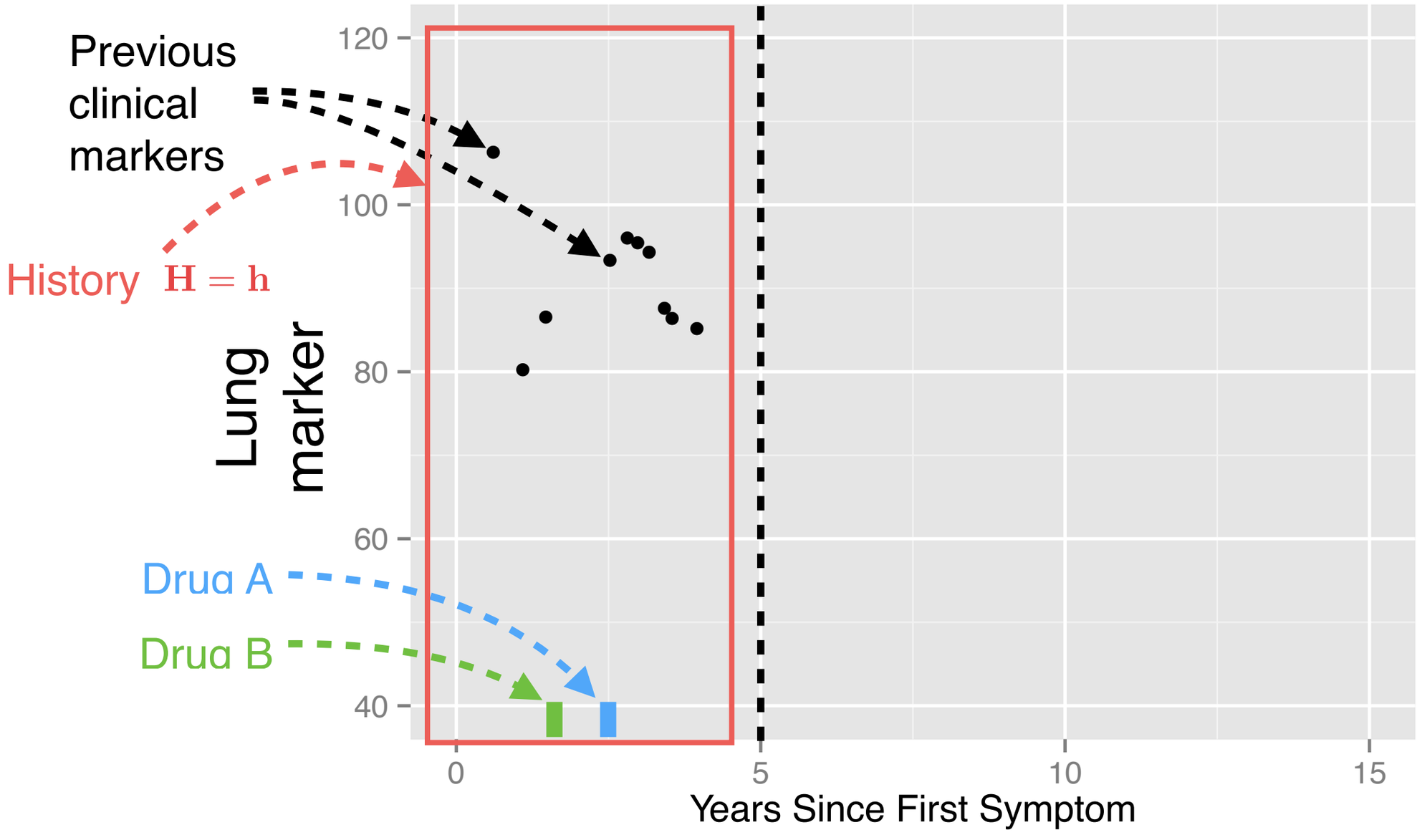
- **Functional markers collected to track organ health**





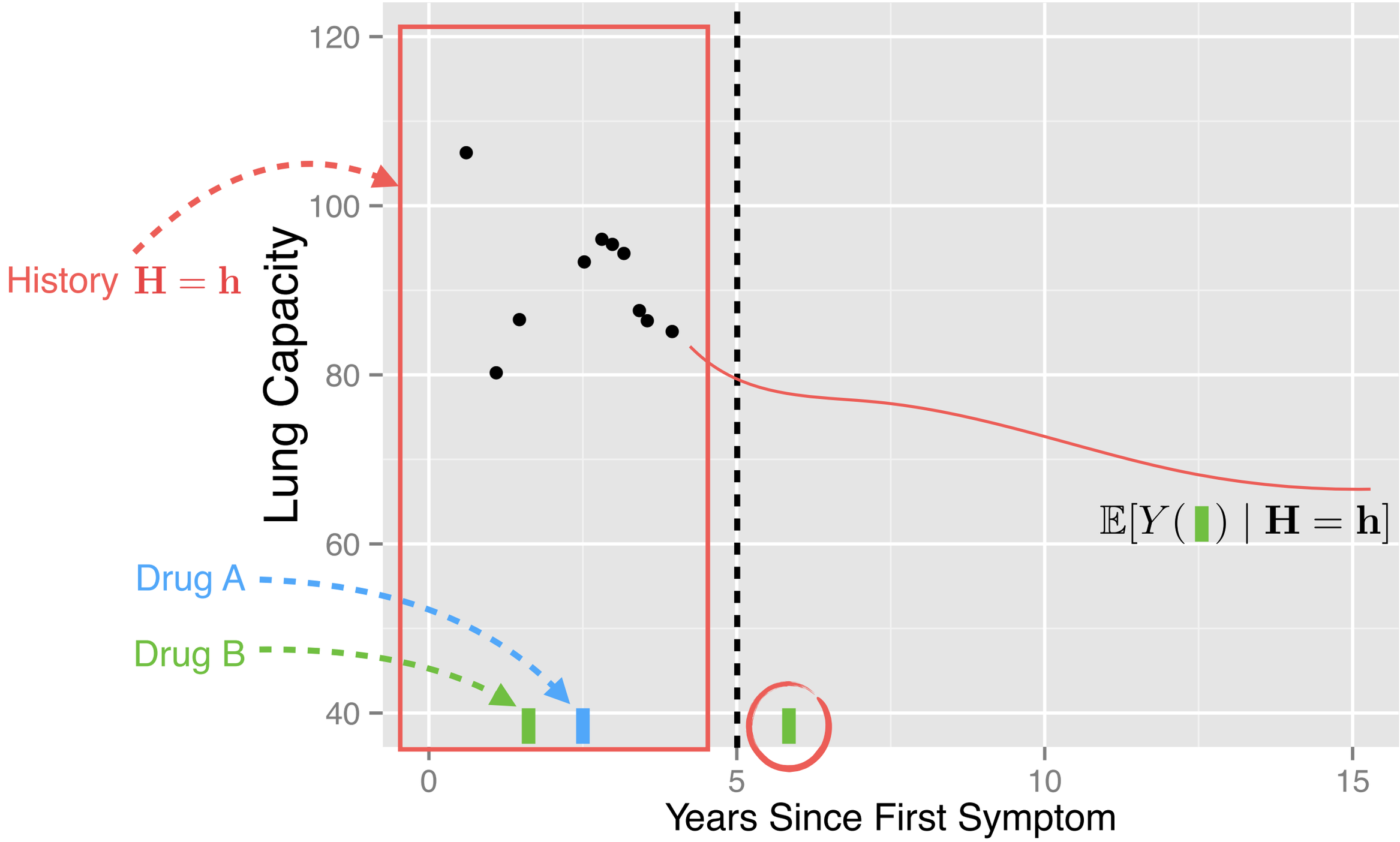
Motivation: Personalized Treatment Planning

- Given what we know about a patient (their history), how should we choose a treatment plan?
  - Should we administer immunosuppressants which can be toxic?



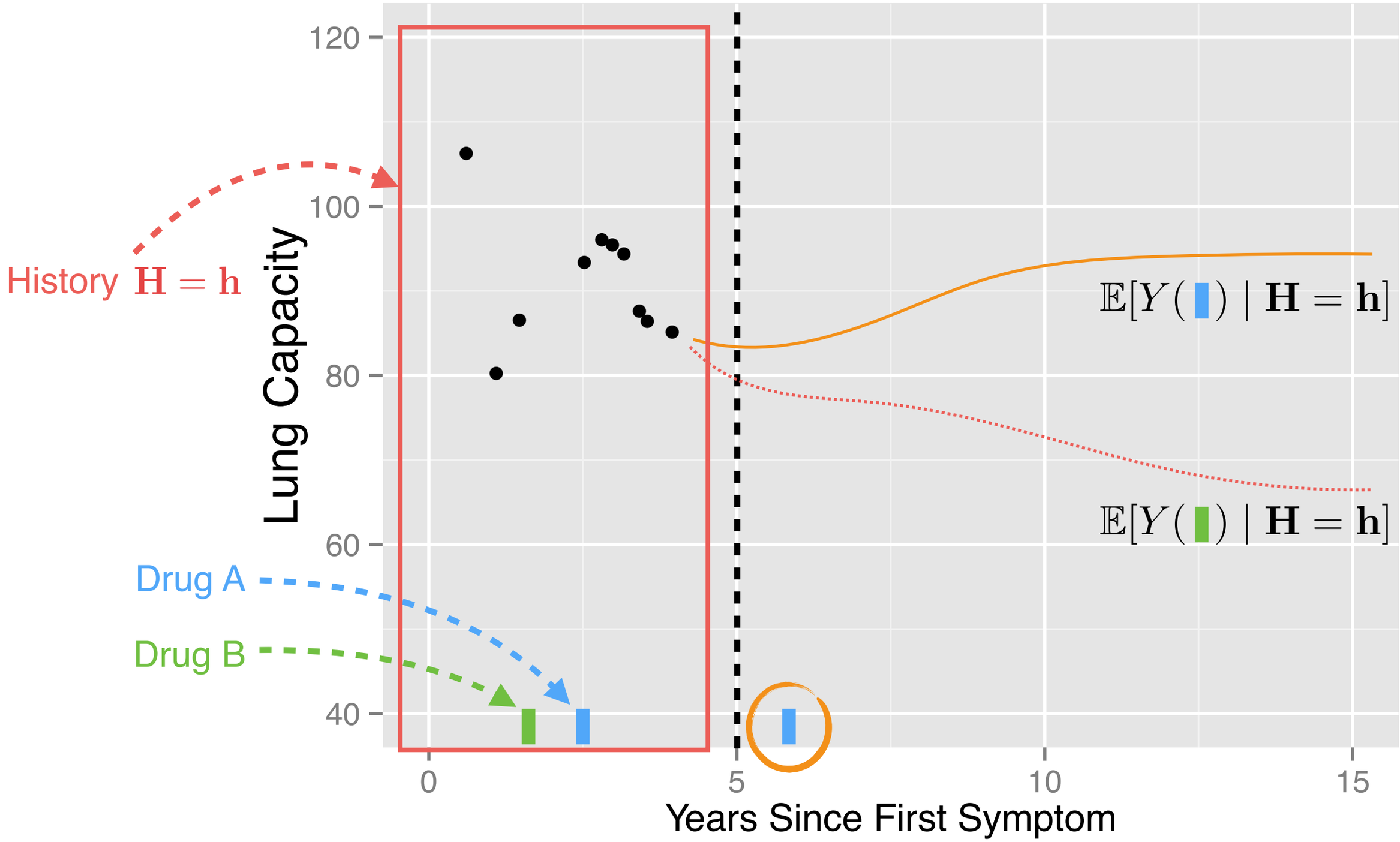
Personalized Treatment Planning

- *Can we simulate a trial*?



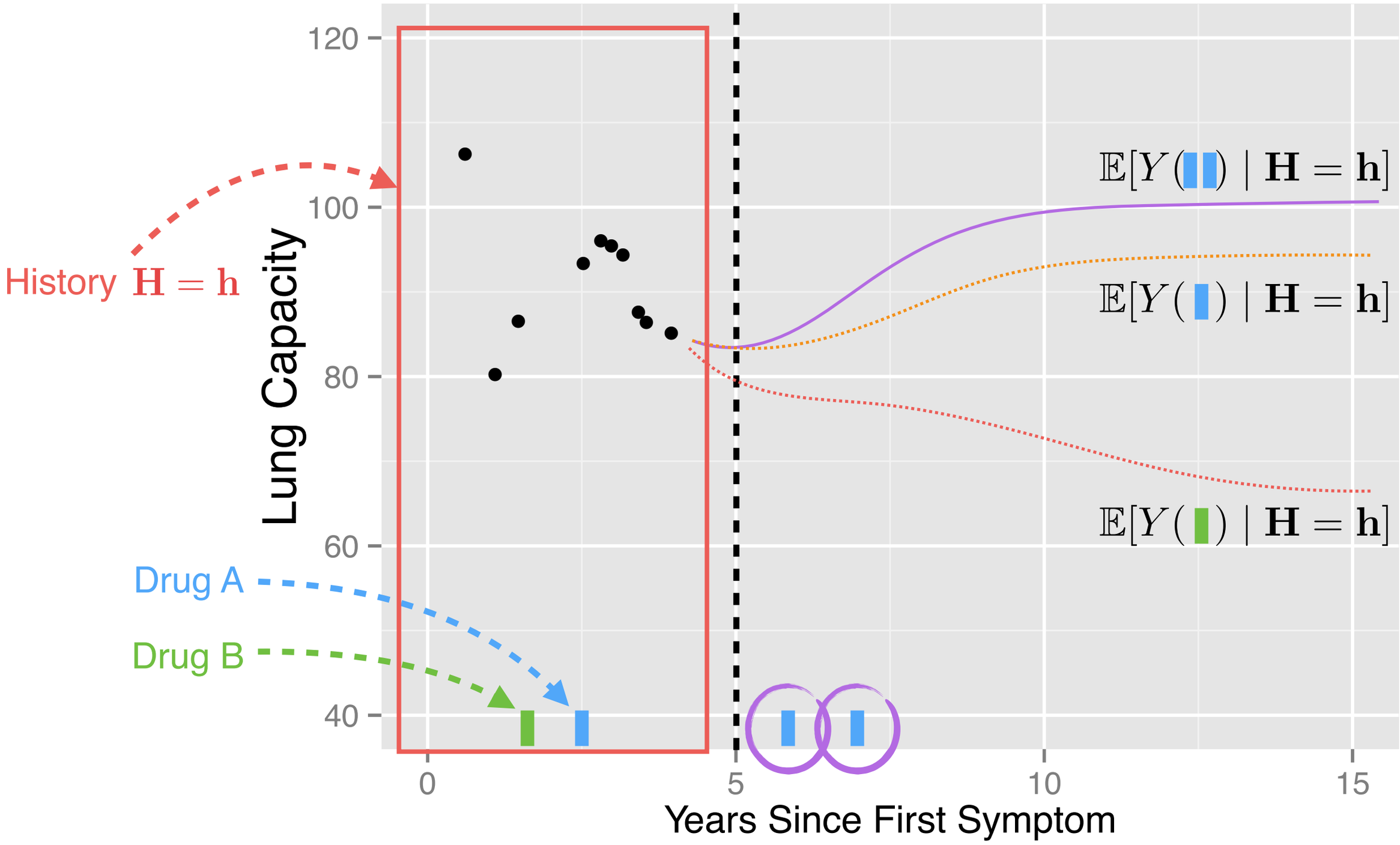
Personalized Treatment Planning

- *Can we simulate a trial*?



Personalized Treatment Planning

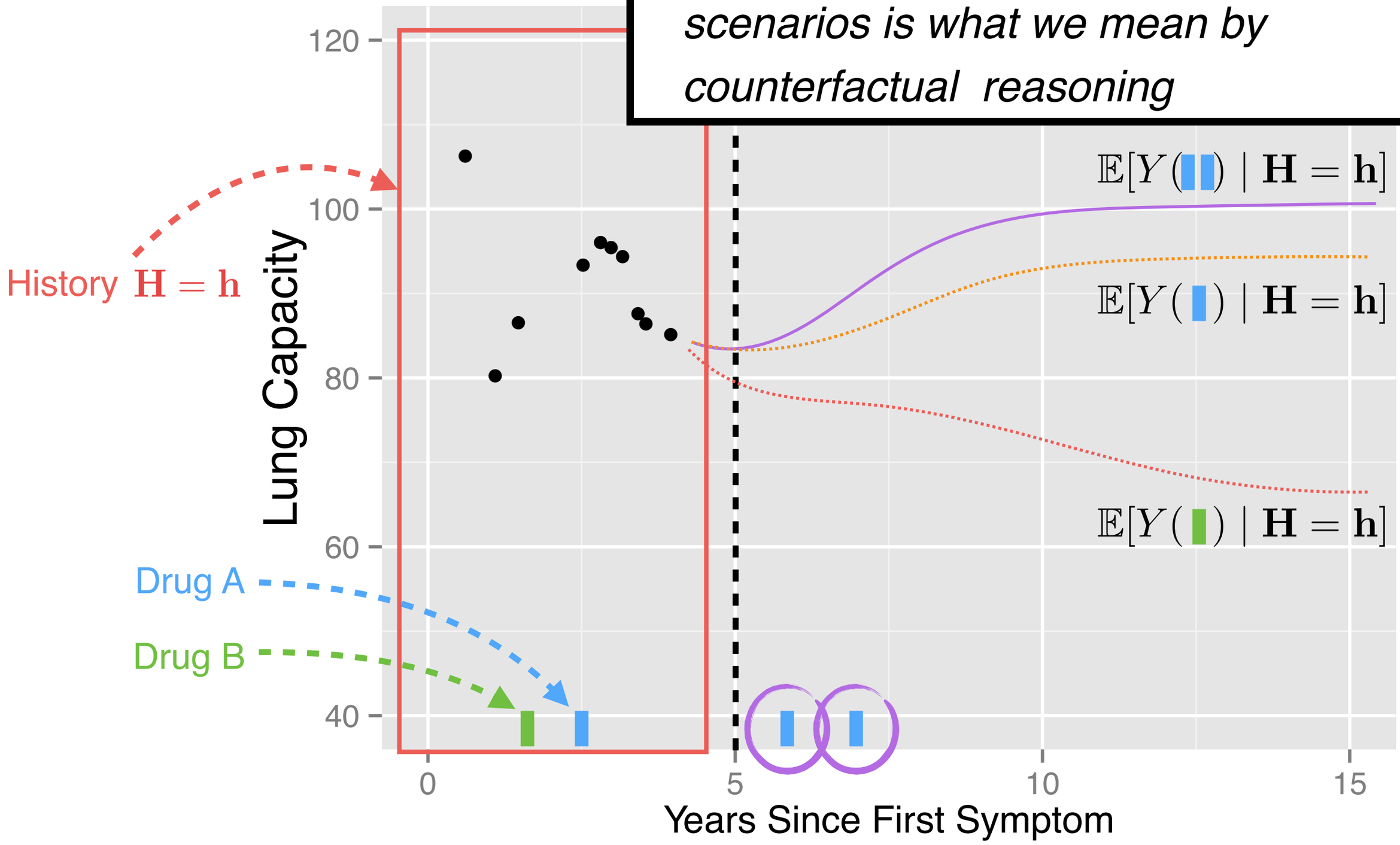
- *Can we simulate a trial*?



Personalized Treatment Planning

- *Can we simulate a trial*?

- *This task of estimating the course (the outcome) under different scenarios is what we mean by counterfactual reasoning*



Outline

#1 Challenges with naive application of off-the-shelf predictive methods.

#2 The use of counterfactual reasoning for personalization
- BG: Potential Outcomes Framework
- BG: SWIGs

#3 Learning from noisy, observational traces
- Classical approaches that treat as discrete time data work poorly
- Treat as functional data
- BG: Gaussian Processes

#4 CGPs — Counterfactual Reasoning from Traces
- Define framework
- Example applications

Outline

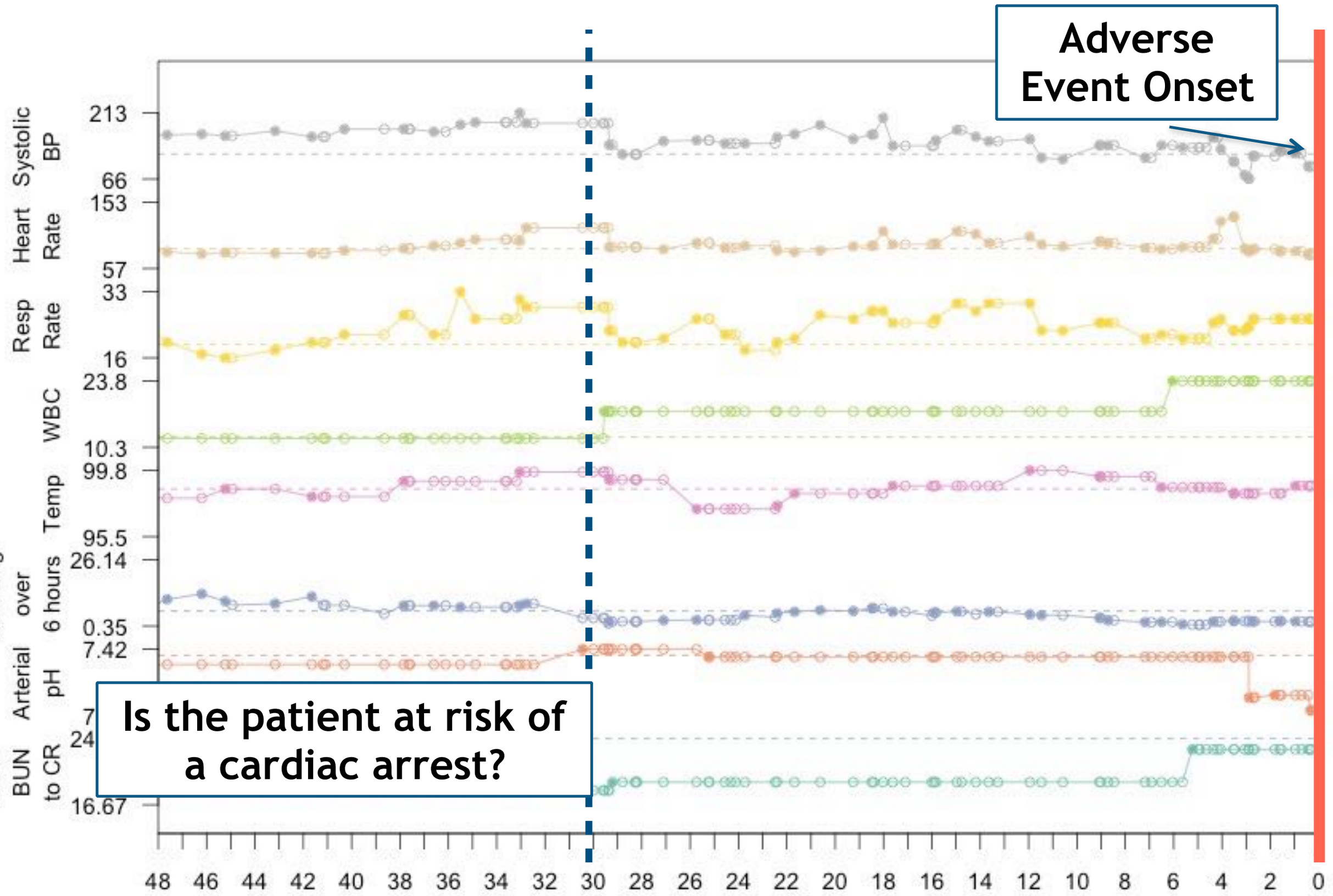
#1 Challenges with naive application of off-the-shelf predictive methods.

#2 The use of counterfactual reasoning for personalization
- BG: Potential Outcomes Framework
- BG: SWIGs

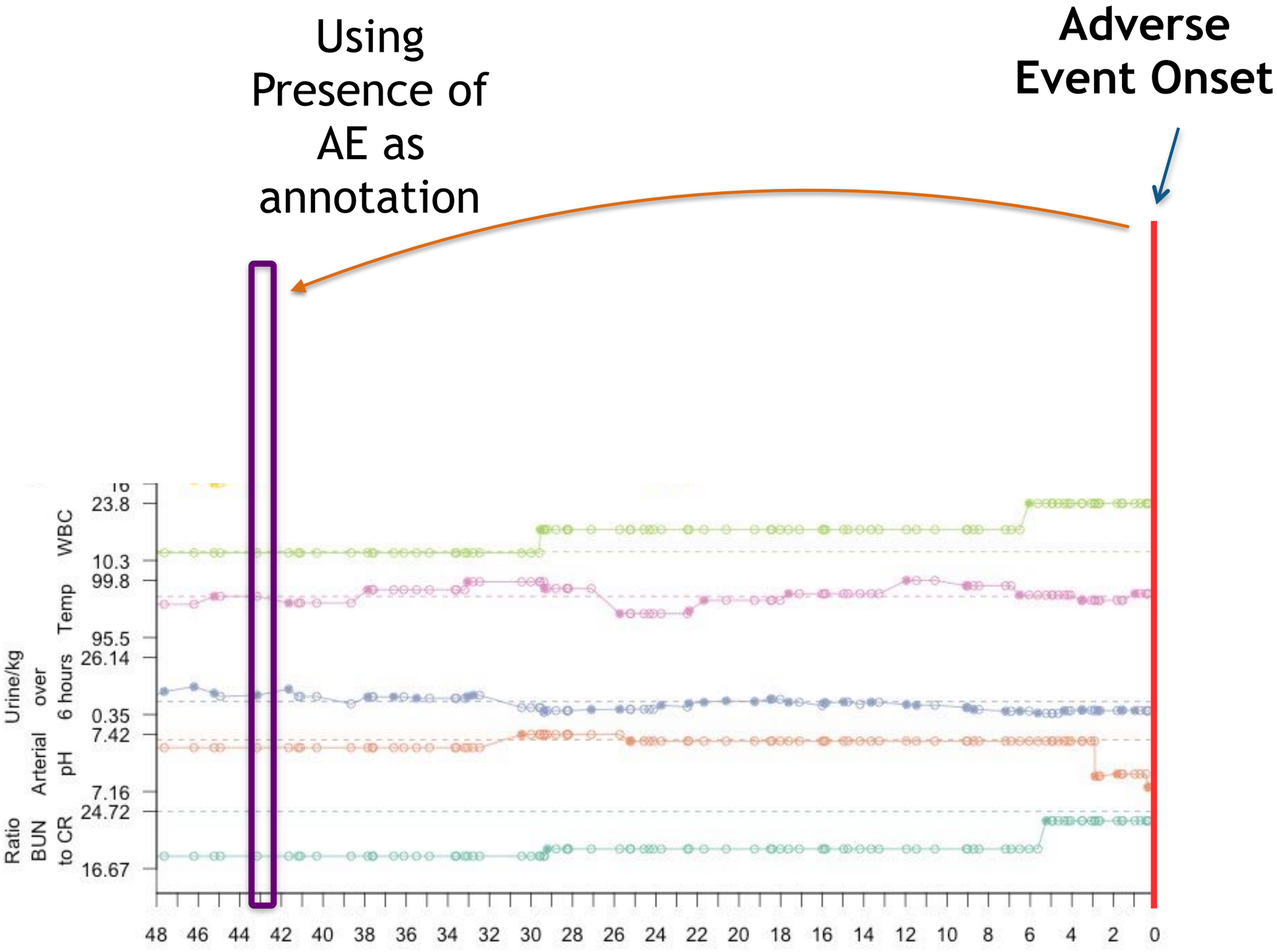
#3 Learning from noisy, observational traces
- Classical approaches that treat as discrete time data work poorly
- Treat as functional data
- BG: Gaussian Processes

#4 CGPs — Counterfactual Reasoning from Traces
- Define framework
- Example applications

Example: Continuous Monitoring and Detection



Use supervised learning for distinguishing patients with AE from those without



Pneumonia Severity Index: Risk of Mortality

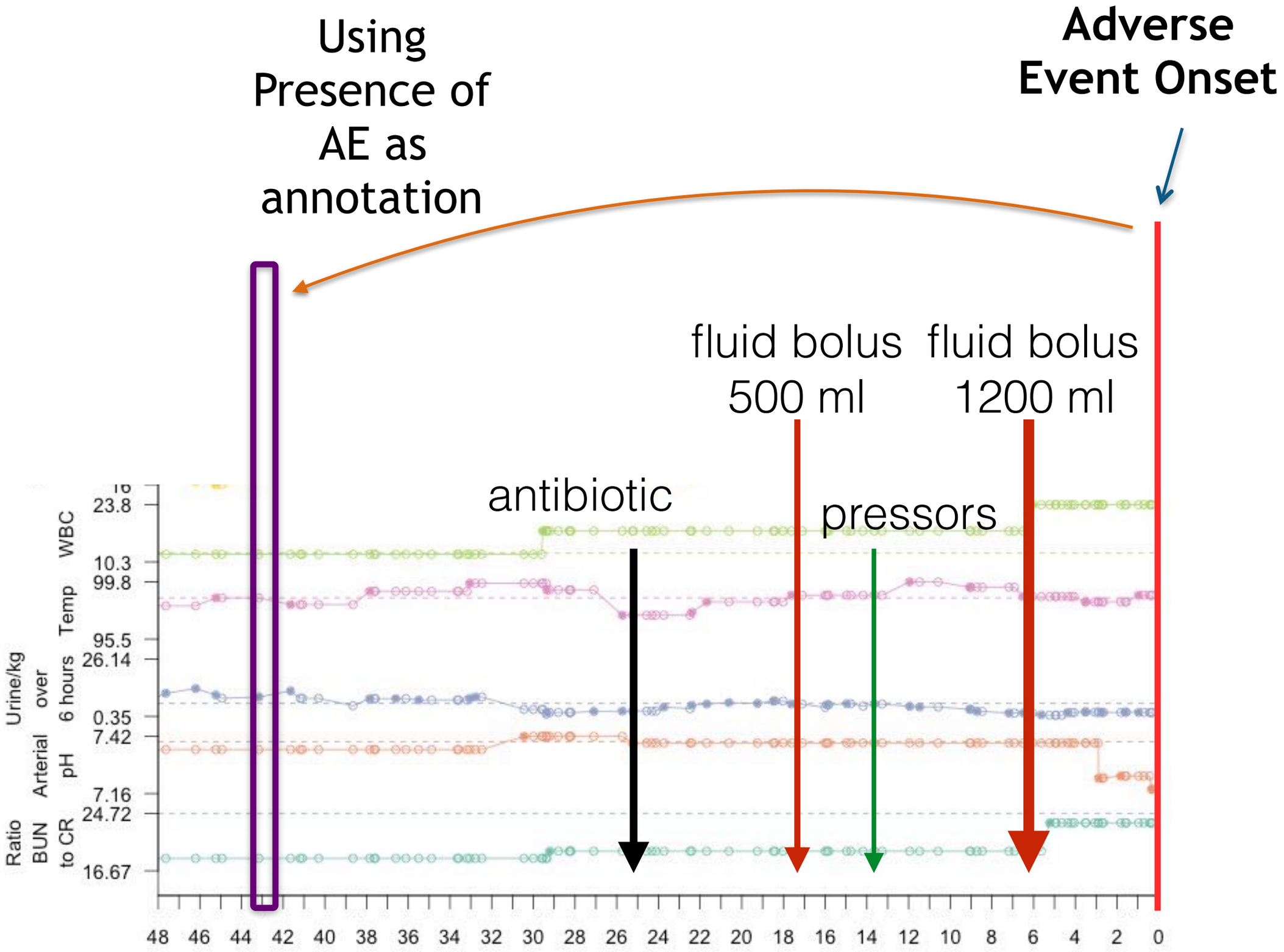
- Identify candidate risk factors
- Learn score and relative weights by regressing against observed mortality

Demographics	Co-morbidities	Physical exam / vital signs	Laboratory / imaging
▪ Age (1 point per year) Male Yr Female Yr -10 ▪ Nursing home residency +10	▪ Neoplasia +30 ▪ Liver disease +20 ▪ CHF +10 ▪ Cerebrovascular disease +10 ▪ Renal disease +10	▪ Mental confusion +20 ▪ Respiratory rate +20 ▪ SBP +20 ▪ Temperature +15 ▪ Tachycardia +15	▪ Arterial pH +30 ▪ BUN +20 ▪ Sodium +20 ▪ Glucose +10 ▪ Hematocrit +10 ▪ Pleural effusion +10 ▪ Oxygenation +10

Risk class (Points)	Mortality (%)	Recommended site of care
I (<50)	0.1	Outpatient
II (51–70)	0.6	Outpatient
III (71–90)	2.8	Outpatient or brief inpatient
IV (91–130)	8.2	Inpatient
V (>130)	29.2	Inpatient

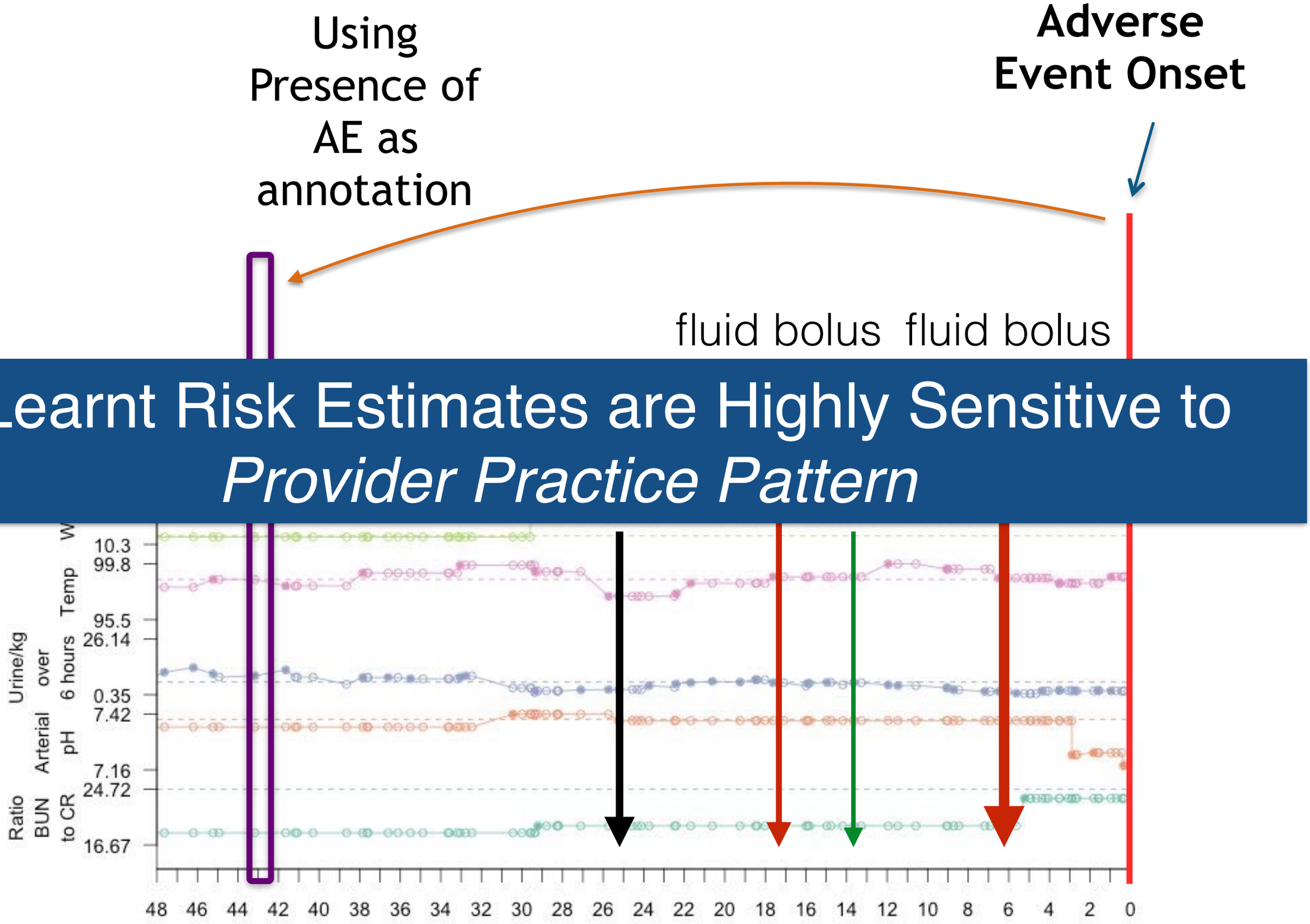
Fine et al., *N Engl J Med* 336 (4): 243–250, 1997

But, interventions *censor* the true label.



Paxton et al., 2013 **Dyagilev et al., 2015**

But, interventions *censor* the true label.



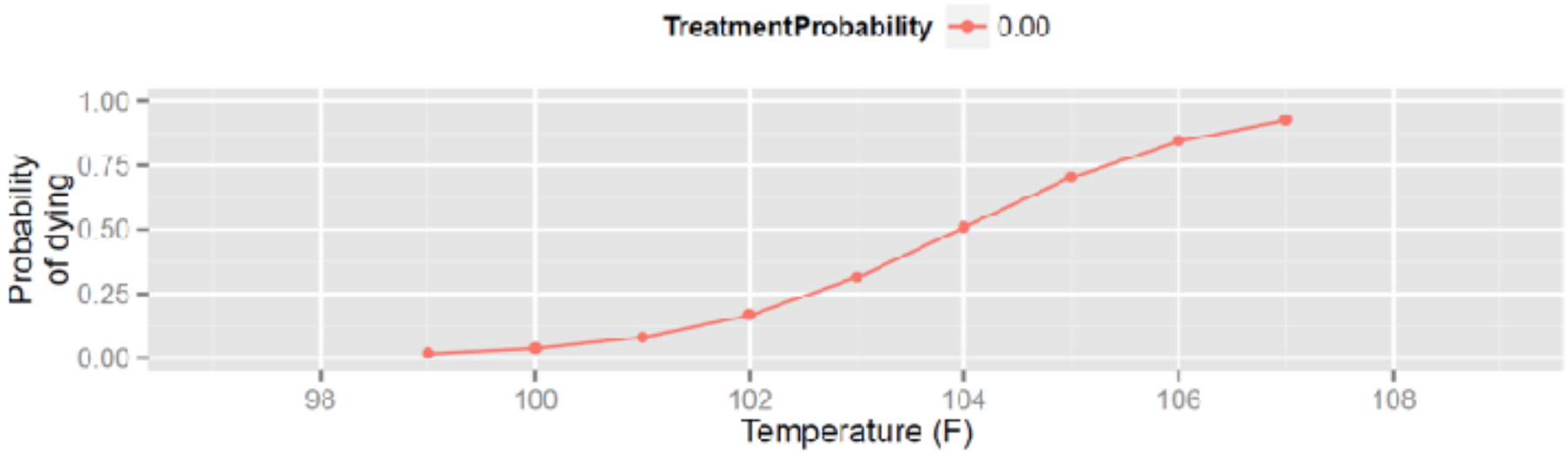
(!) Learnt Risk Estimates are Highly Sensitive to *Provider Practice Pattern*

Paxton et al., 2013 **Dyagilev et al., 2015**

Challenge: Learnt Risk Estimates Sensitive to Provider Practice Pattern

- Simple example (Flu)
  - Measure temperature
  - Measure WBC

- Increase in temperature or WBC increases risk of death

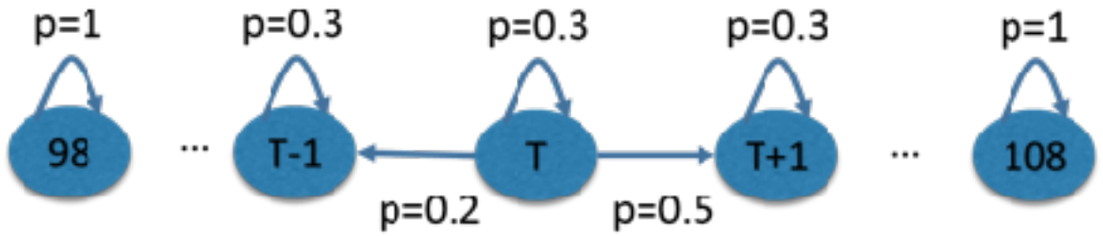


Bias Due to Interventional Confounds

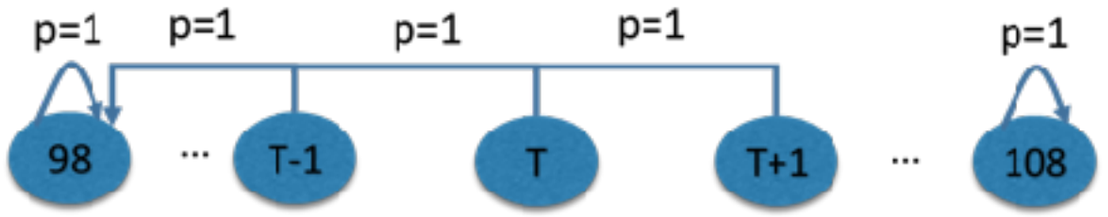
- Simulation
  - Patients with a flu get sicker and eventually die unless treated.
  - Vary treatment practice patterns: P(Treat | High temperature)

    vs. P(Treat | high WBC)

No antibiotics:



With antibiotics:



Treatment practice:
(1) no antibiotics for $T < 102$ deg F;
(2) administer antibiotics with probability ρ for $T \geq 102$ deg F

Dyagilev et al., 2015

Bias Due to Interventional Confounds

Vary provider practice patterns between train and test:

Scenario	$\rho_{\mathrm{T}}^{\mathrm{train}}$	$\rho_{\mathrm{WBC}}^{\mathrm{train}}$	$\rho_{\mathrm{T}}^{\mathrm{test}}$	$\rho_{\mathrm{WBC}}^{\mathrm{test}}$
#1	0	0	0	0
#2	0.1	0	0.1	0
#3	0.1	0	0	0
#4	0.3	0	0	0
#5	0.3	0	0	0.3

Increase probability of treating for rising temperature

Increasing discrepancy in physician prescription behavior in train vs. test environment

Dyagilev et al., 2015

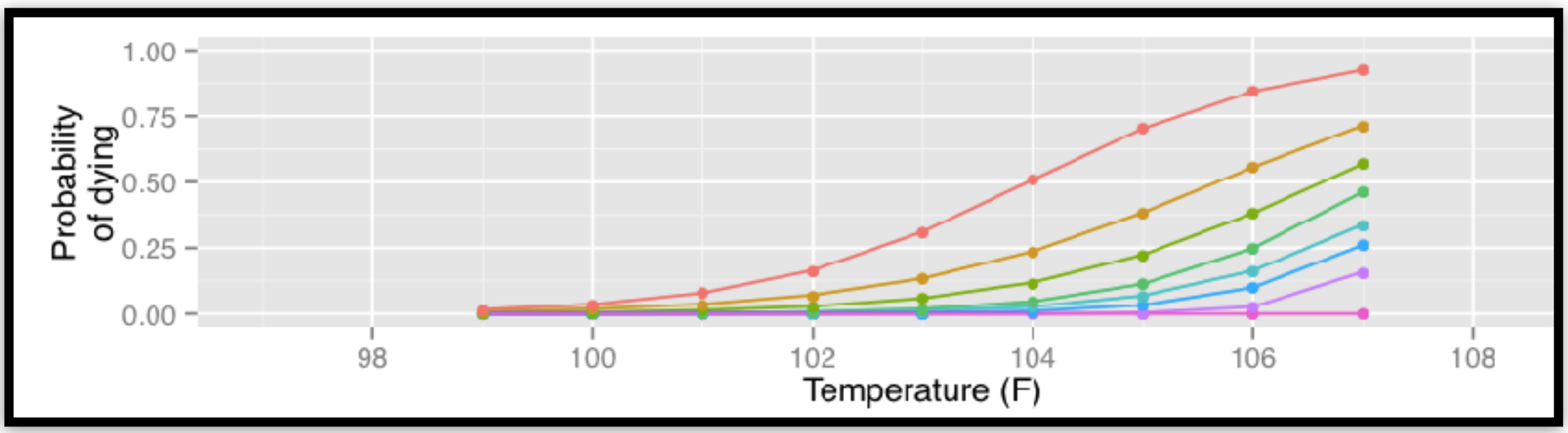
Bias Due to Interventional Confounds

Vary provider practice patterns between train and test:

Scenario	ρ_T^{train}	ρ_{WBC}^{train}	ρ_T^{test}	ρ_{WBC}^{test}	Logistic Regression
#1	0	0	0	0	0.974
#2	0.1	0	0.1	0	0.978
#3	0.1	0	0	0	0.963
#4	0.3	0	0	0	0.769
#5	0.3	0	0	0.3	0.510

Increase probability of treating for rising temperature

Increasing discrepancy in physician prescription behavior in train vs. test environment



Dyagilev et al., 2015

Bias Due to Interventional Confounds

Vary provider practice patterns between train and test:

Scenario	ρ_T^{train}	ρ_{WBC}^{train}	ρ_T^{test}	ρ_{WBC}^{test}	Logistic Regression
#1	0	0	0	0	0.974
#2	0.1	0	0.1	0	0.978
#3	0.1	0	0	0	0.963
#4	0.3	0	0	0	0.769
#5	0.3	0	0	0.3	0.510

Increase probability of treating for rising temperature

Increasing discrepancy in physician prescription behavior in train vs. test environment

Learned risk scores are high sensitive to changes in provider practice patterns:

- Resulting risk scores are also less interpretable
- They violate ***construct validity*** [Medsger et al., 2003]

Dyagilev et al., 2015

Naive application of predictive tools can give counterintuitive results

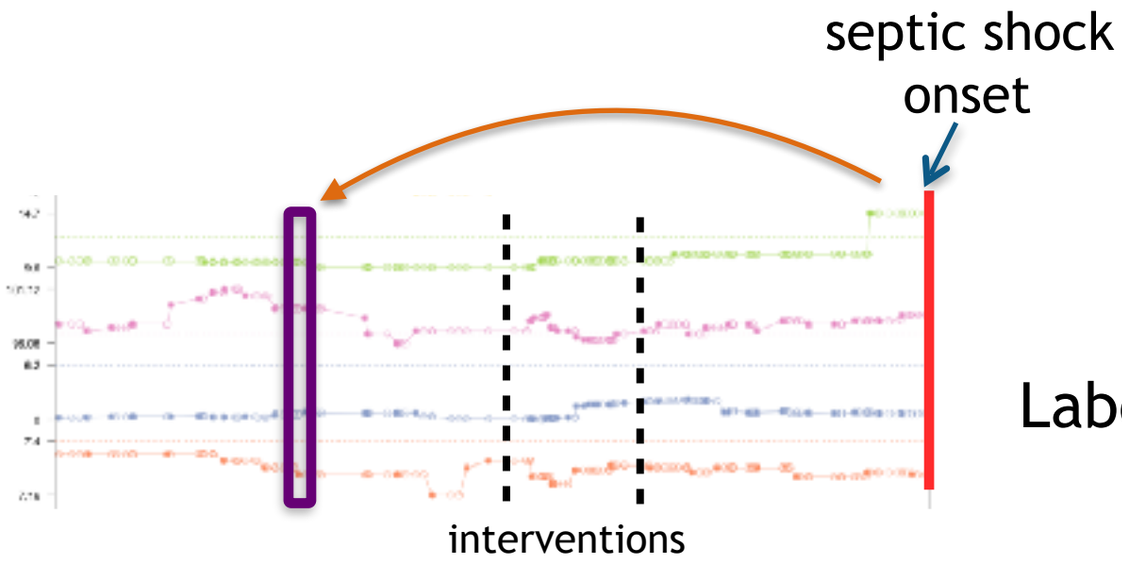
Example: (**Caruana et al., KDD, 2015**)

- ML method learned that patients with pneumonia with asthma history have lower mortality risk than the general population.
- This is counterintuitive — patients with asthma history have much higher risk *if not hospitalized*
- Pneumonia patients with asthma history were admitted to the ICU, and the intensive care lowered their risk of dying

If applied naively and without considering clinical context, machine learning methods may yield **counterintuitive predictions** and models with **unintended consequences.**

Caruana et al., 2015

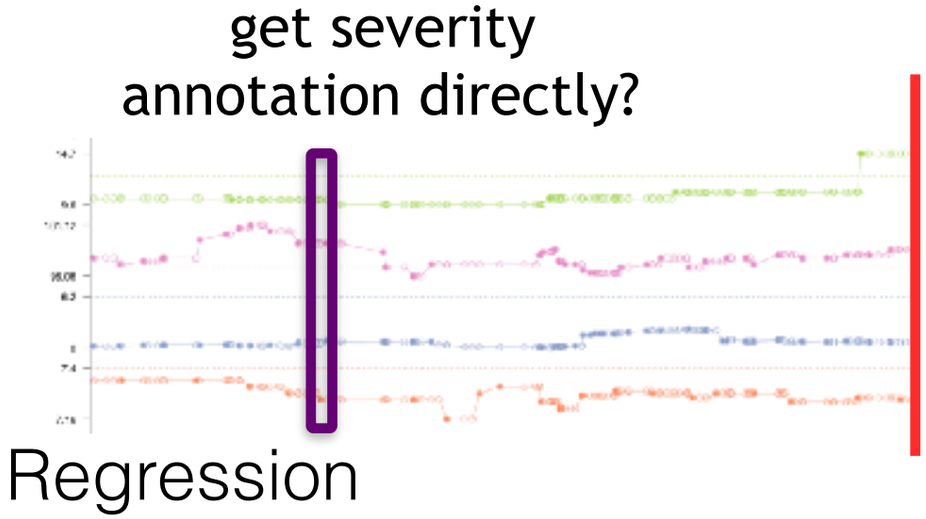
Need alternate forms of training and supervision



Labels are noisy or censored

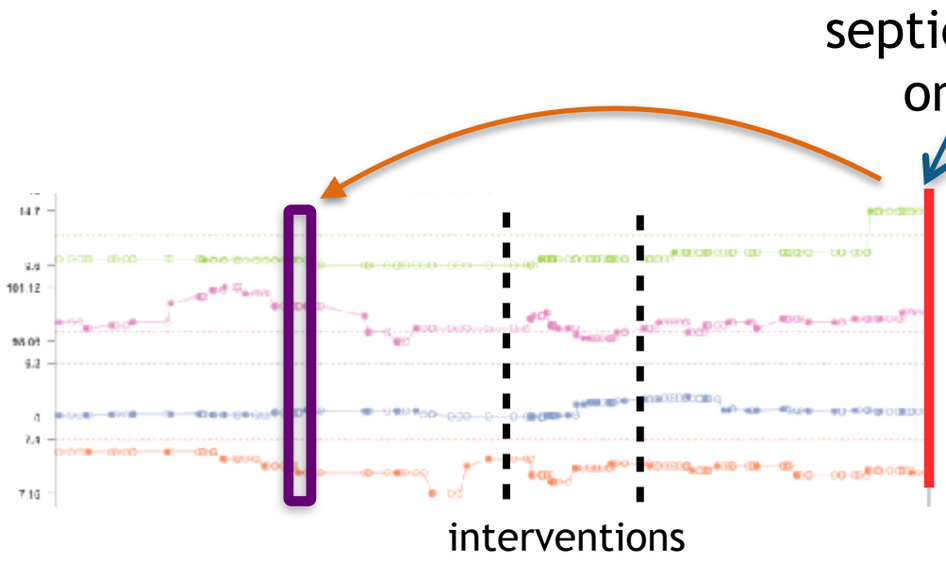
Henry et al., 2015

Alternate forms of labels:



May or may not be practical

Need alternate forms of training and supervision

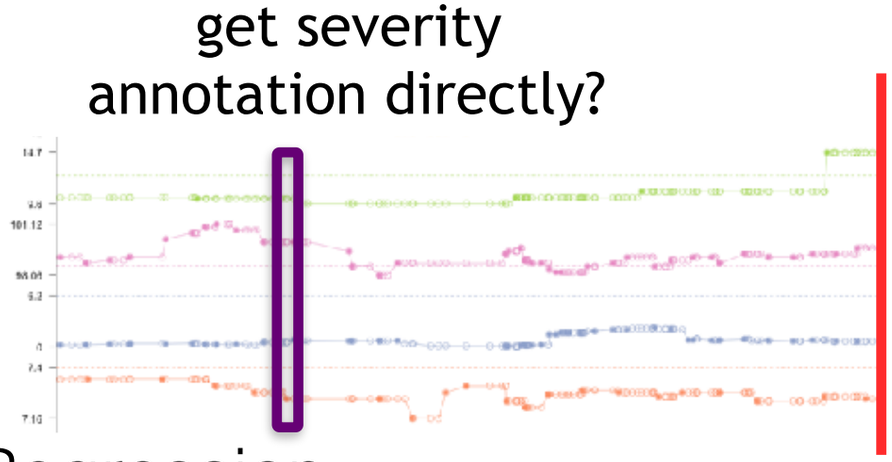


Labels are noisy or censored

Henry et al., 2015

Alternate forms of labels:

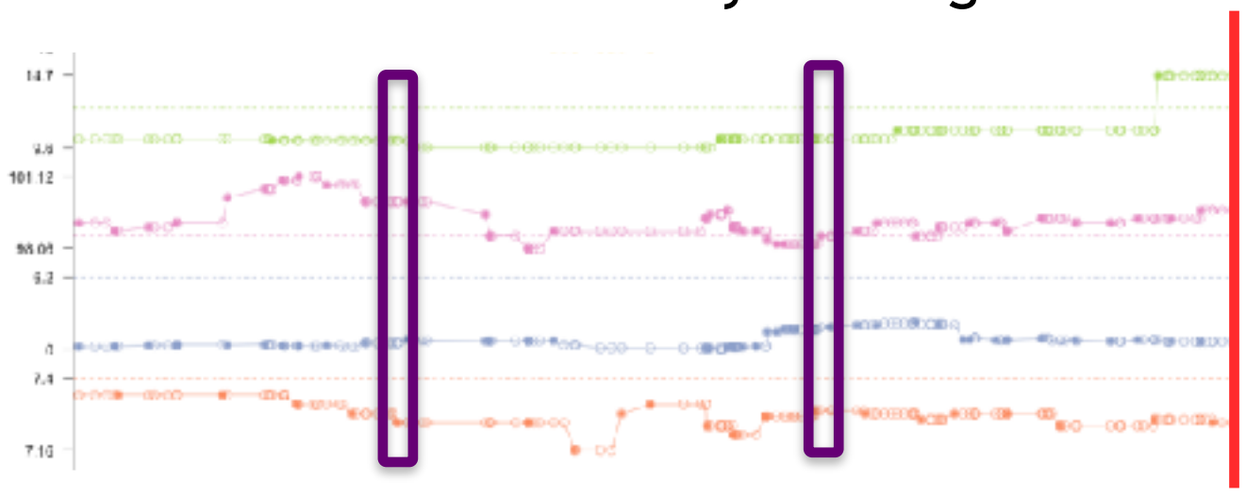
get severity
annotation directly?



Regression

May or may not be practical

Severity ranking



Comparison Pairs: **Dyagilev et al., 2015**

Today: Joint modeling of states and actions

Transportability not always possible: Bareinboim and Pearl, 2013

”Causal Predictions”

- Statistical model’s predictions captures correlations that depend on **provider practice**
  E.g. “treat when temperature rises above 100”

- What we learn is “*what is likely to happen* ***if they receive the treatments they did receive***”

- The desired target is: “what is likely to happen to this patient given their history if we **do not treat vs treat**”

- In order to ask, what will happen if we **do** X vs Y, we draw ideas from causal inference and develop models for reasoning about counterfactual outcomes.

Bottou et al., 2012

Outline

#1 Challenges with naive application of off-the-shelf predictive methods.

#2 The use of counterfactual reasoning for personalization
- BG: Potential Outcomes Framework
- BG: SWIGs

#3 Learning from noisy, observational traces
- Classical approaches that treat as discrete time data work poorly
- Treat as functional data
- BG: Gaussian Processes

#4 CGPs — Counterfactual Reasoning from Traces
- Define framework
- Example applications

Example: Exercise and Blood Pressure

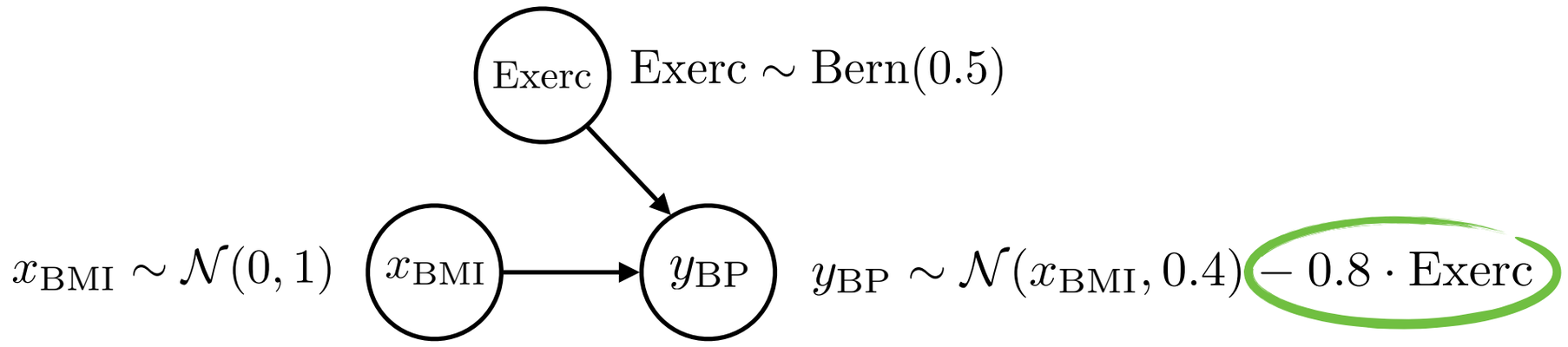
- Hypothesis: exercise lowers blood pressure
- In this example, we have:
  - (a) A treatment (exercise)
  - (b) An outcome (blood pressure)
- How can we use data to estimate whether exercise will lower blood pressure?

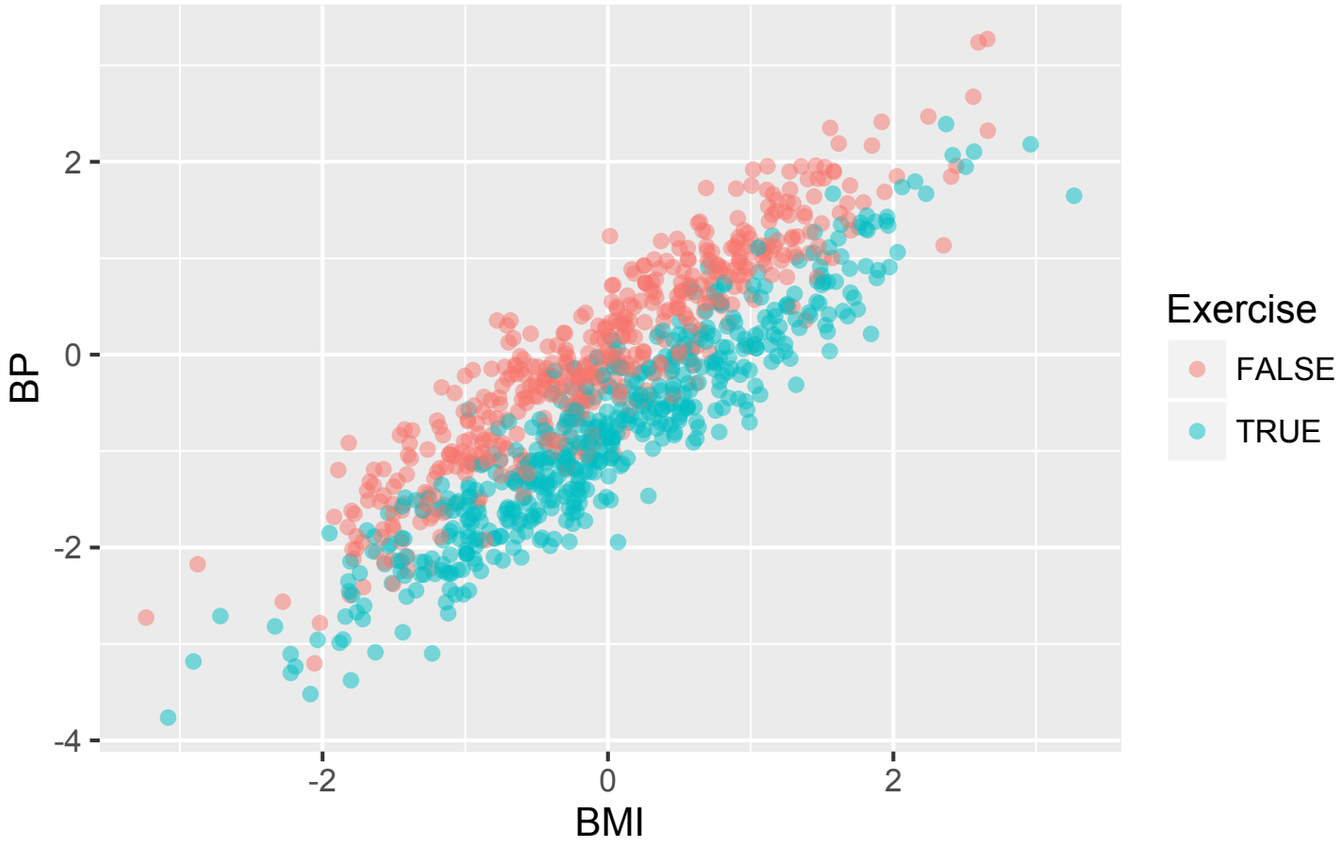
Example: Exercise and Blood Pressure

- Grab an existing dataset containing people who did and did not exercise and have measurements of blood pressure
  - Average the change in blood pressure among people who exercise and among those who don't
- Will this work?

Randomized Controlled Trial (RCT)

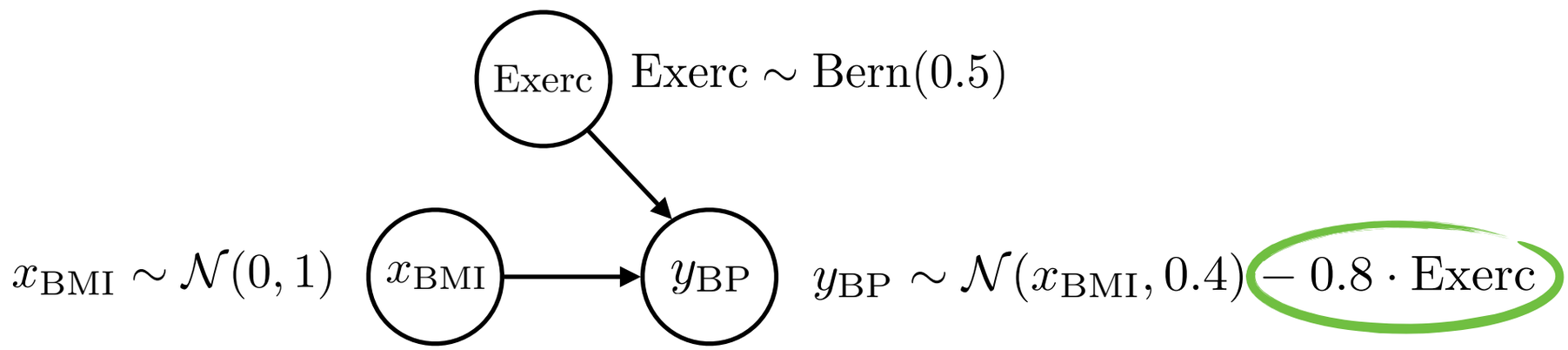
- Dataset generative model:



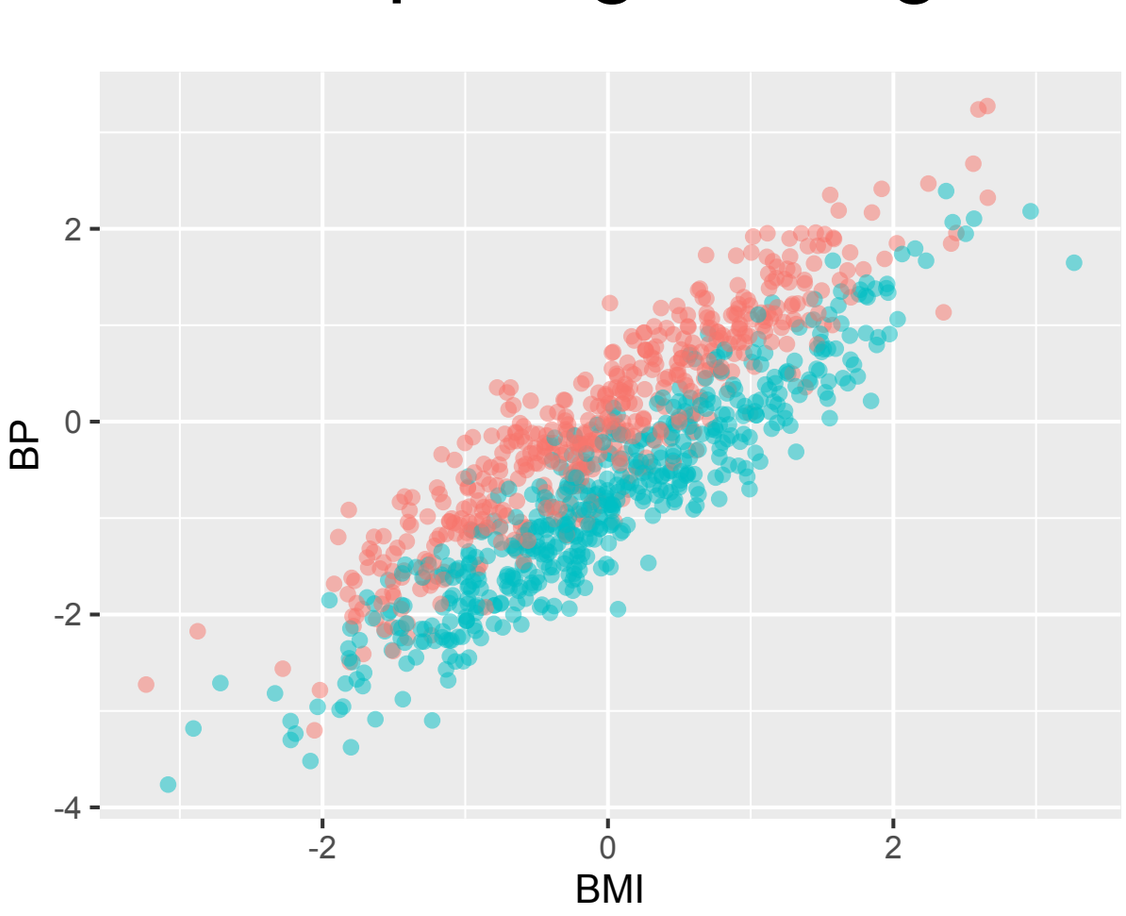


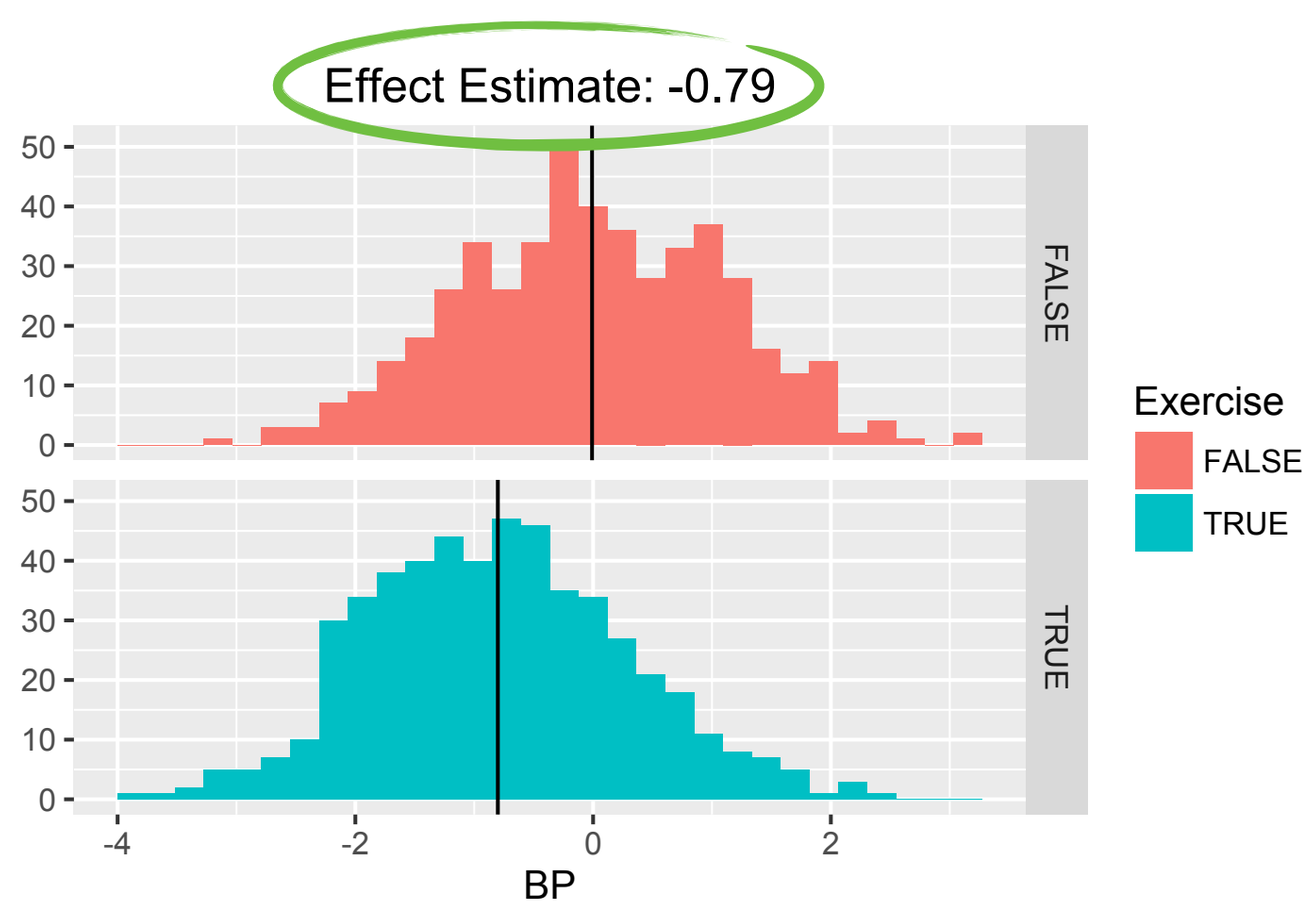
Randomized Controlled Trial (RCT)

- Dataset generative model:



- Comparing averages will work!



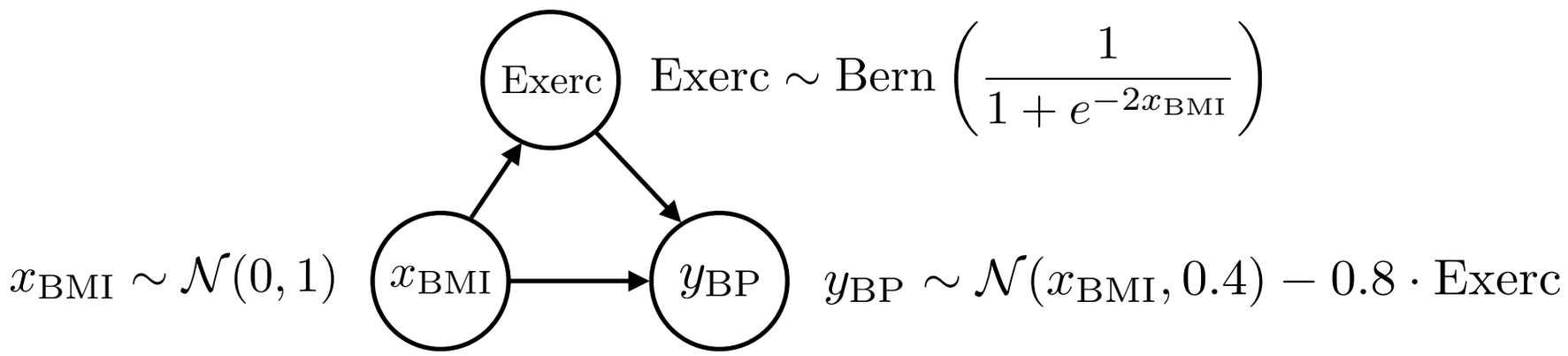


Observational Data

- Instead of running an expensive trial, suppose we simply collect information on 1000 individuals from general clinics around the country

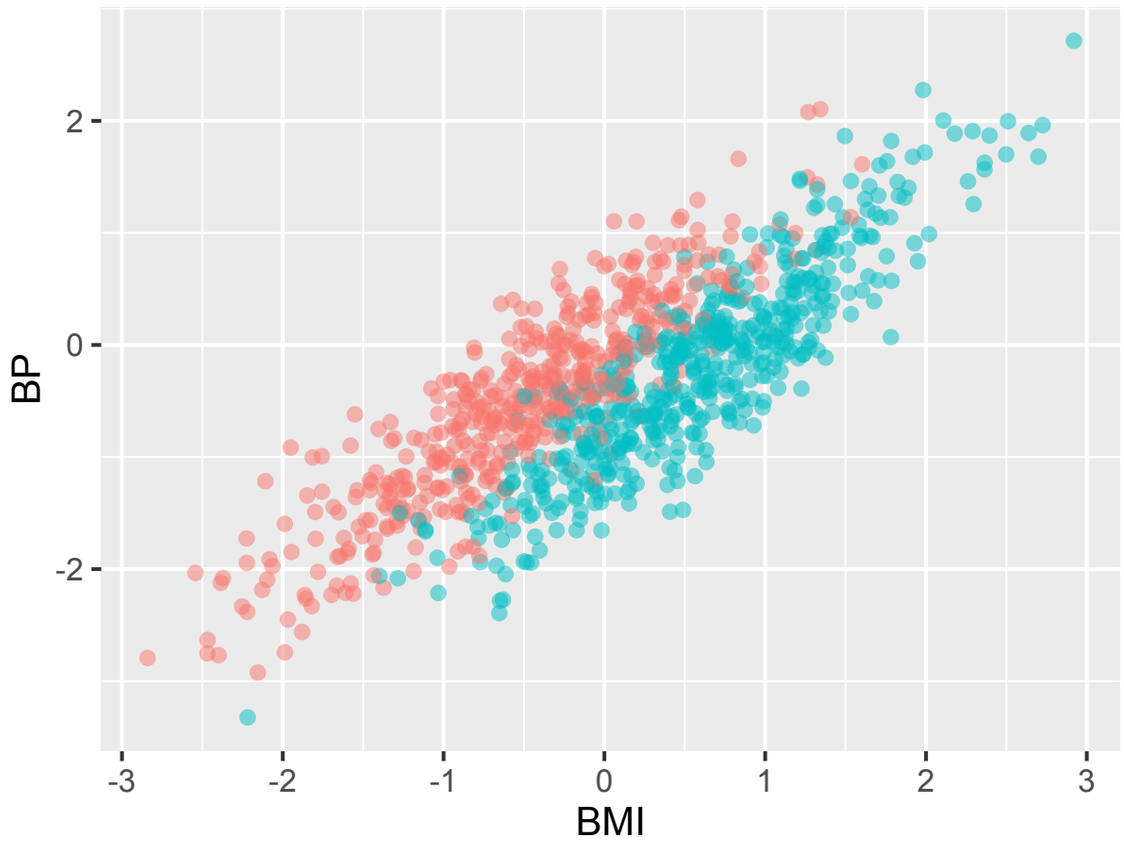
- In the observational data, **exercise is *assigned by the clinicians* caring for the individuals**

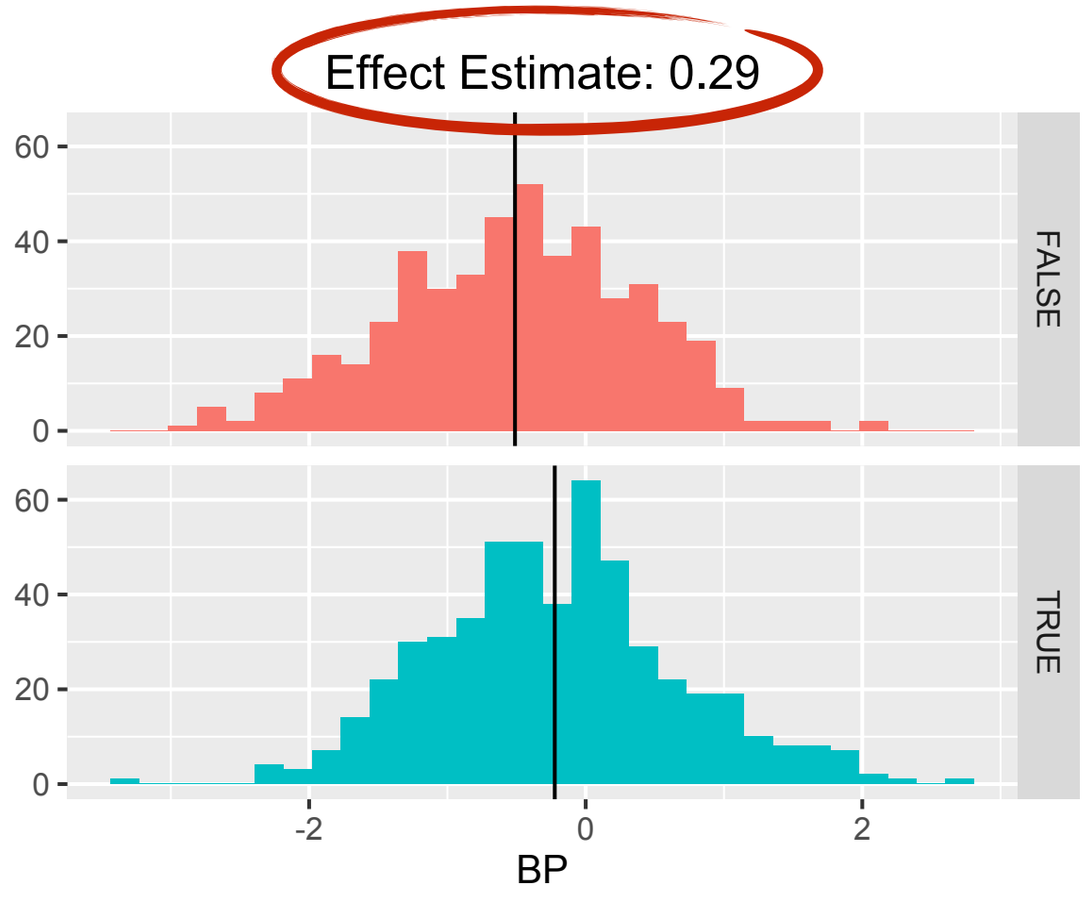
- In particular, we assume that a higher BMI makes prescription of exercise more likely:



Observational Data

- Simply comparing averages no longer works!
- What's going on? How can we adjust for this bias?

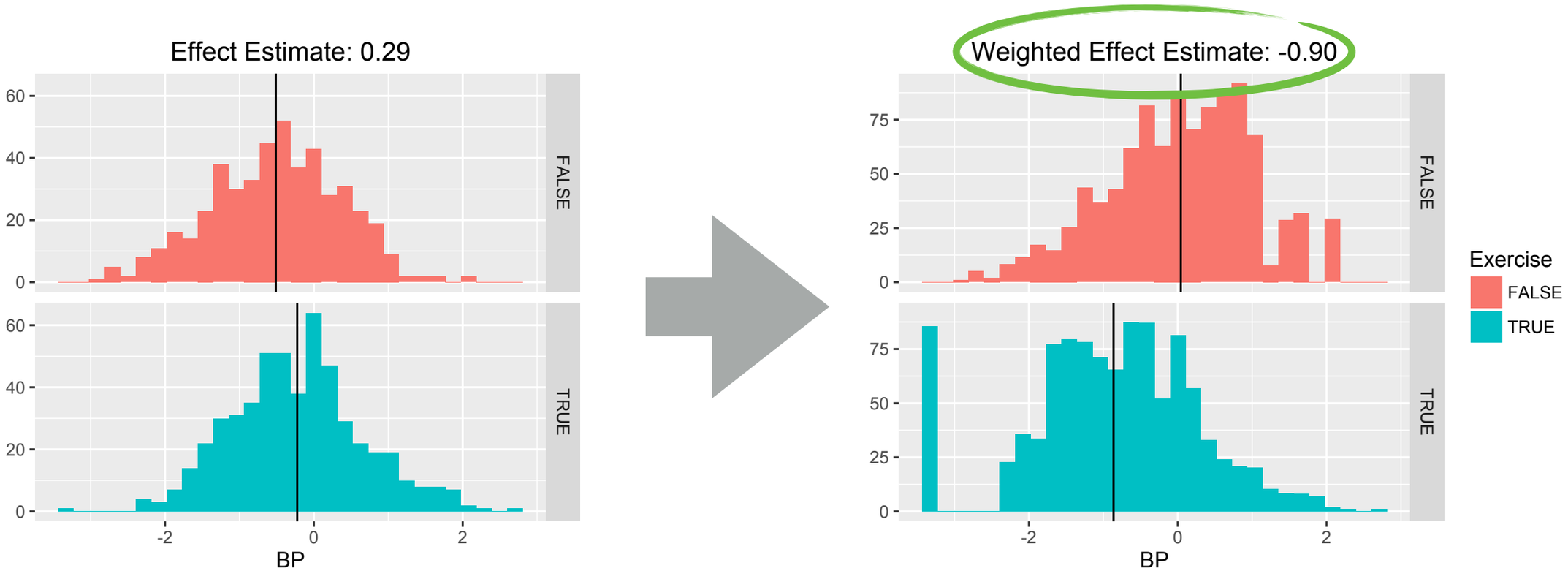




Approach 1: Weighting

- If we know (or can estimate) a model of treatment assignment, then a common approach is to use *inverse probability of treatment weights*

- Intuitive idea: when computing averages, count an individual more if she was unlikely to receive treatment (probability is low —> weight is high) and vice versa



Horvitz and Thompson, 1952 **Robins et al. 2000**

Approach 1: Weighting

- For each individual, compute weight:

$$w_i = \frac{1}{p(A_i = a_i \mid \mathbf{X}_i = \mathbf{x}_i)}$$

Must know or estimate the treatment assignment model

- Compute weighted averages among treated/not treated

$$\bar{y}_{\text{Exerc}} = \frac{\sum_{i=1}^{n} w_i \cdot y_i \cdot \mathbb{I}[\text{Exerc} = 1]}{\sum_{i=1}^{n} w_i \cdot \mathbb{I}[\text{Exerc} = 1]} \qquad \bar{y}_{\text{No Exerc}} = \frac{\sum_{i=1}^{n} w_i \cdot y_i \cdot \mathbb{I}[\text{Exerc} = 0]}{\sum_{i=1}^{n} w_i \cdot \mathbb{I}[\text{Exerc} = 0]}$$

- Other approaches: matching, propensity scores

Rosenbaum and Rubin, 1983 **Shalit and Sontag Tutorial, ICML 2016**

Hernán and Robins, Forthcoming Textbook

- Off-policy evaluation:

Dudik et al., 2011 **Jiang and Li, 2016** **Paduraru et al. 2013**

Alternative Framework: Potential Outcomes

- We will approach this problem using the framework of *potential outcomes*

Rubin, 1974 **Neyman et al., 1923** **Rubin, 2005**

- For an individual, conceptualize two "alternate realities"
  - (1) They exercise
  - (2) They do not exercise
- In each reality, we can measure blood pressure and measure the *potential outcome*
- **If we know both potential outcomes, we can answer the question of whether exercise lowers blood pressure**

Potential Outcomes

- To formalize, define two distinct random variables:
  - Y(a) : blood pressure *with* exercise
  - Y(b) : blood pressure *without* exercise
- More generally, we can index a set of random variables using a set of actions/treatments:

$$\{Y(a) : a \in \mathcal{A}\}$$

- Offers a way to reason about *counterfactuals*.
- **Goal**: learn statistical models to estimate potential outcomes

Critical Assumptions

- To learn the potential outcome models, we will use three important assumptions:

- (1) Consistency

  - Links observed outcomes to potential outcomes

- (2) Treatment Positivity

  - Ensures that we can learn potential outcome models

- (3) No unmeasured confounders (NUC)

  - Ensures that we do not learn biased models

(1) Consistency

- Consider a dataset containing observed outcomes, observed treatments, and covariates:

$$\{y_i, a_i, \mathbf{x}_i\}_{i=1}^n$$

  - E.g.: blood pressure, exercise, BMI

- Consistency allows us to replace the observed response with the potential outcome of the observed treatment

$$Y \triangleq Y(a) \mid A = a$$

- Under consistency our dataset satisfies

$$\{y_i, a_i, \mathbf{x}_i\}_{i=1}^n \triangleq \{y_i(a_i), a_i, \mathbf{x}_i\}_{i=1}^n$$

(2) Positivity

- When working with observational data, for any set of covariates $\mathbf{x}$ we need to **assume a non-zero probability of seeing each treatment**

  - Otherwise, in general, cannot learn a conditional model of the potential outcomes given those covariates

- Formally, we assume that

$$\mathrm{P}_{\mathrm{Obs}}(A = a \mid \mathbf{X} = \mathbf{x}) > 0 \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

(3) No Unmeasured Confounders (NUC)

- In our exercise example, BMI is a *confounder*
  - BMI induces a statistical dependency between the observed treatment and observed outcome
- In general, unless we observe all confounders, we cannot learn unbiased models of potential outcomes from observational data
- Formally, NUC is an statistical independence assertion:

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

Making NUC intuitive using Single-World Intervention Graphs

- **SWIGs extend graphical models to explicitly represent potential outcomes**

- To obtain a SWIG, we define a causal graphical model and specify the set of treatment variables

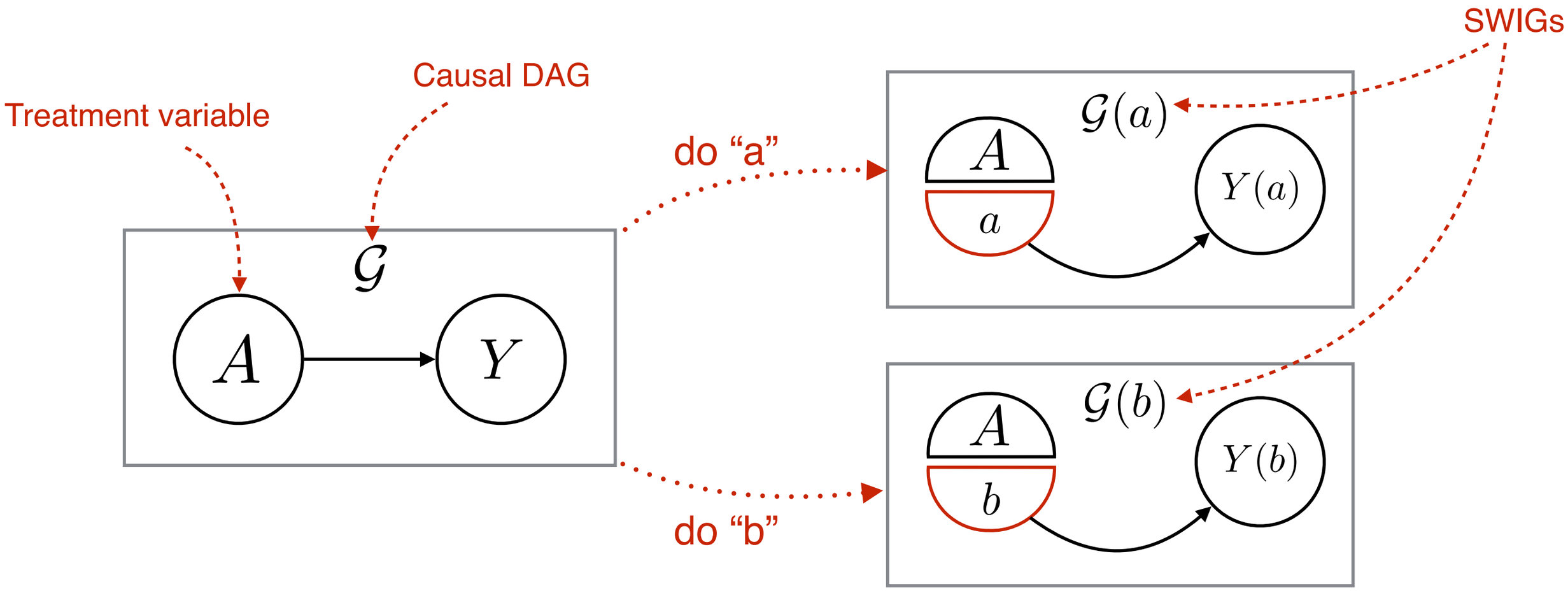
- We apply *node-splitting* operations to treatment variables to represent interventions

- Useful tool to determine which conditional distributions you need, and how to simulate trial

NIPS tutorial: **Richardson, 2014** **Richardson and Robins, 2014**

Example SWIG

- We apply ***node-splitting*** **operations to treatment variables to represent interventions**
- A simple “a” vs “b” example:

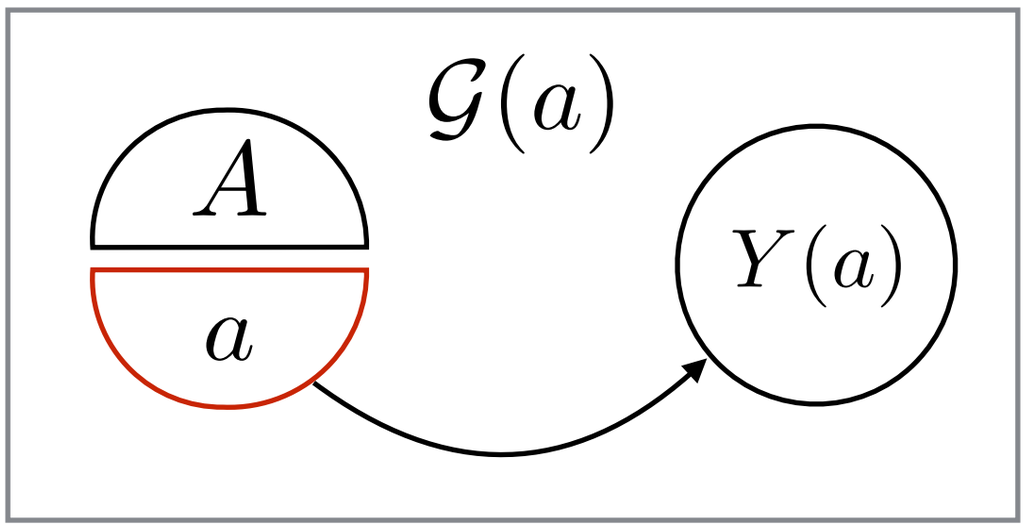


Richardson, 2014

Richardson and Robins, 2014

Interpreting SWIGs

- Treat SWIGs as standard causal graphs
  - Semi-circle nodes are just reminders that we have applied a node-splitting operation
- From this graph, can read that Y(a) is independent of the observed treatment A



Richardson, 2014

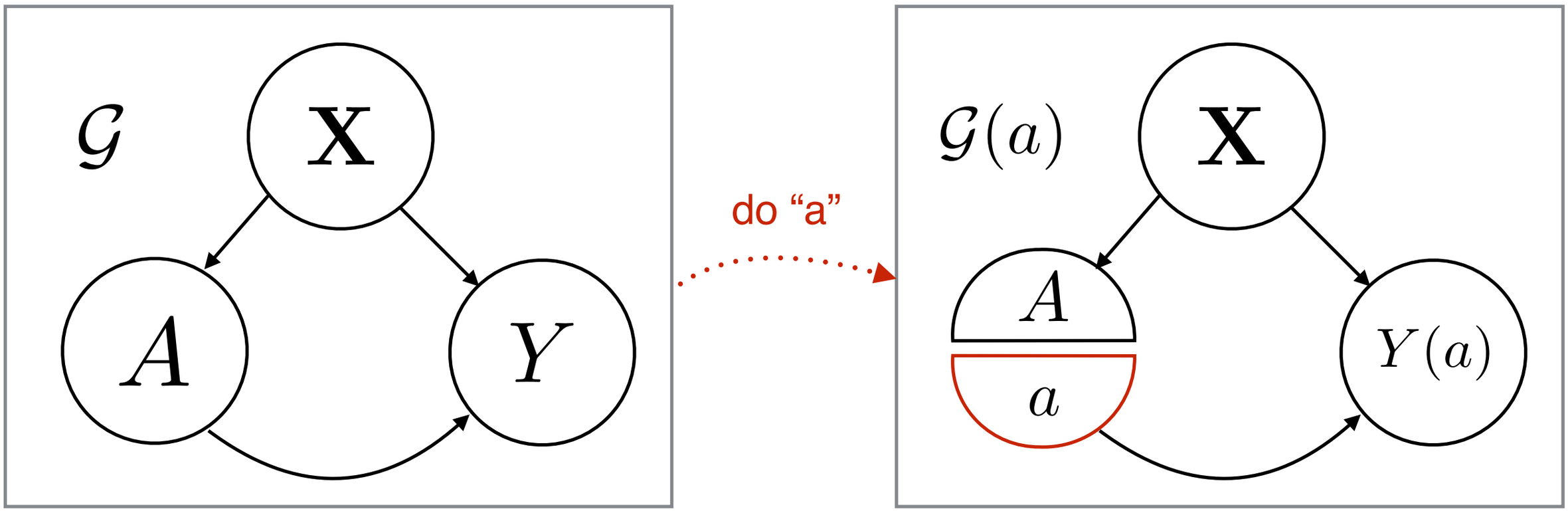
Richardson and Robins, 2014

NUC in SWIG Language

- SWIGs make NUC assumption easy to express

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

- Confounders X d-separate potential outcomes from observed treatment random variable when intervening on treatment



Richardson, 2014

Richardson and Robins, 2014

Using Models to Adjust for Bias

- Assume models of potential outcomes given covariates

$$\{\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) : a \in \mathcal{A}\}$$

- We can use them to adjust for bias in observational data

- Key idea: use models to “simulate” an RCT

Rubin 1977 **Robins 1986**

Learning Potential Outcome Models

- To simulate data from a new policy, we need to learn the potential outcome models
  - If we have an observational dataset where assumptions 1-3 hold, then this is possible!
- Assumptions allow estimation of potential outcomes from (observational) data:

$$\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) = \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}, A = a) \quad \text{(A3)}$$

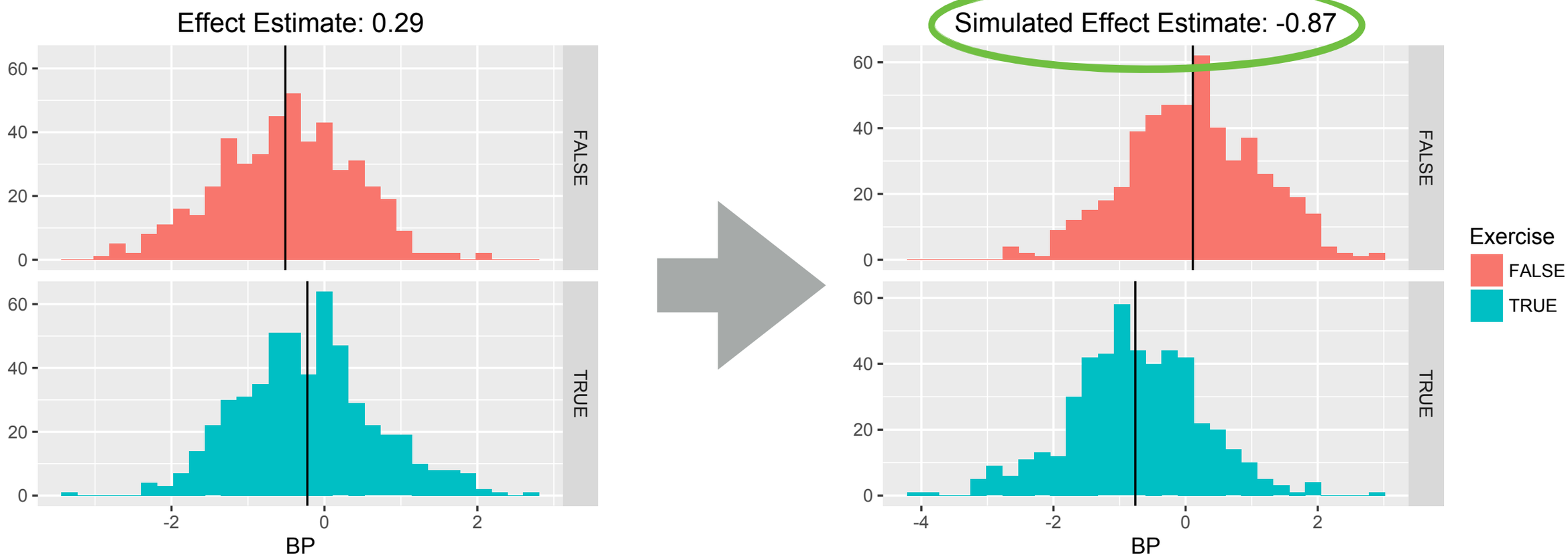
$$= \boxed{\mathrm{P}(Y \mid \mathbf{X} = \mathbf{x}, A = a)} \quad \text{(A1)}$$

Estimation requires a statistical model for estimating conditionals

Exercise and Blood Pressure

- Returning to our exercise and blood pressure example
- We fit a model for blood pressure given exercise and BMI
- With estimated models, treatment effects are estimated as:

$$\mathbb{E}[Y(1)-Y(0)] = \frac{1}{N}\sum_{n=1}^{N}(Y_n(1)-Y_n(0))$$



Going beyond PATE

PATE: Population Average Treatment Effect:

$$\mathbb{E}[Y(1)-Y(0)] = \frac{1}{N}\sum_{n=1}^{N}(Y_n(1)-Y_n(0))$$

To account for the heterogeneous treatment effect among patients, it is more of interest to look at **CATE**, the conditional average treatment effect:

$$\mathbb{E}[Y(1) - Y(0) \mid C_1 = c_1]$$

See e.g.: **Foster et al., 2011** **Imai et al., 2013** **Tian et al., 2014** **Athey and Imbens, 2016**

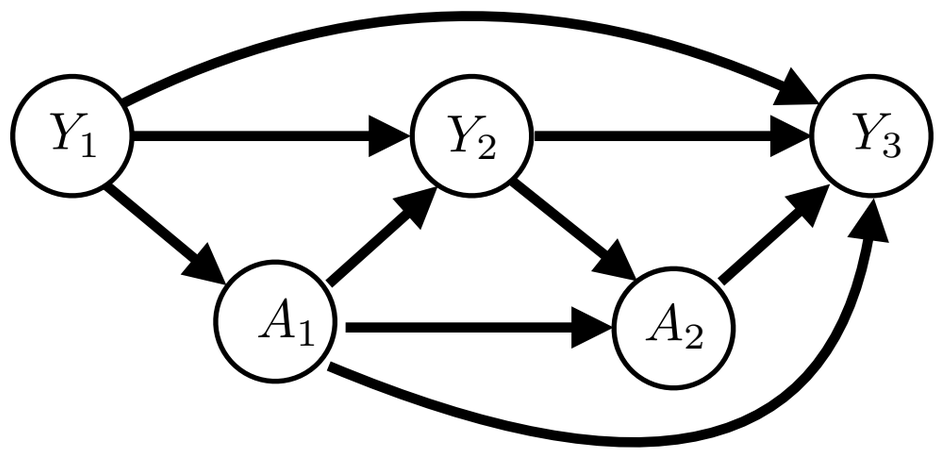
Sequential Treatment Assignment and Time-Varying Confounding

- Interventions and observations are interleaved

  - Intervention effects future observations
    Those observations affect future interventions
    And so on…

  - When can we disentangle to learn unbiased models of potential outcomes?

- Also called time-varying confounding.



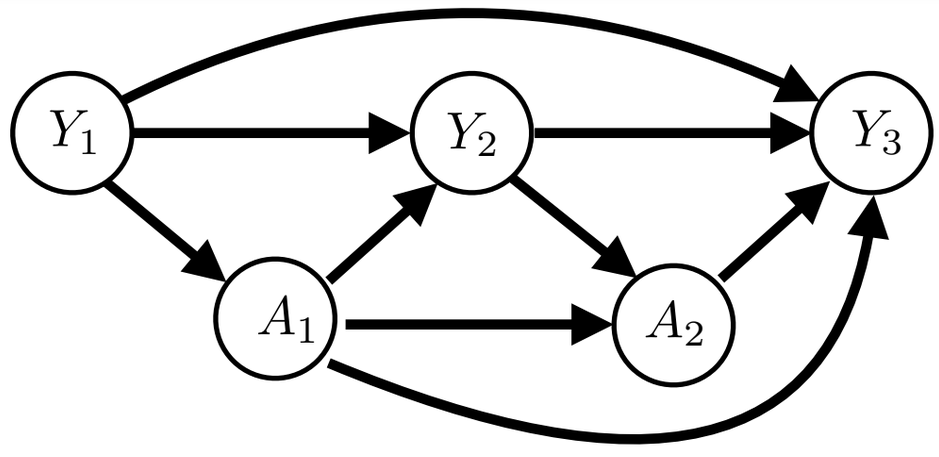
Robins 1986

Sequential Treatment Assignment and Time-Varying Confounding

- Interventions and observations are interleaved

  - Intervention effects future observations
    Those observations affect future interventions
    And so on…

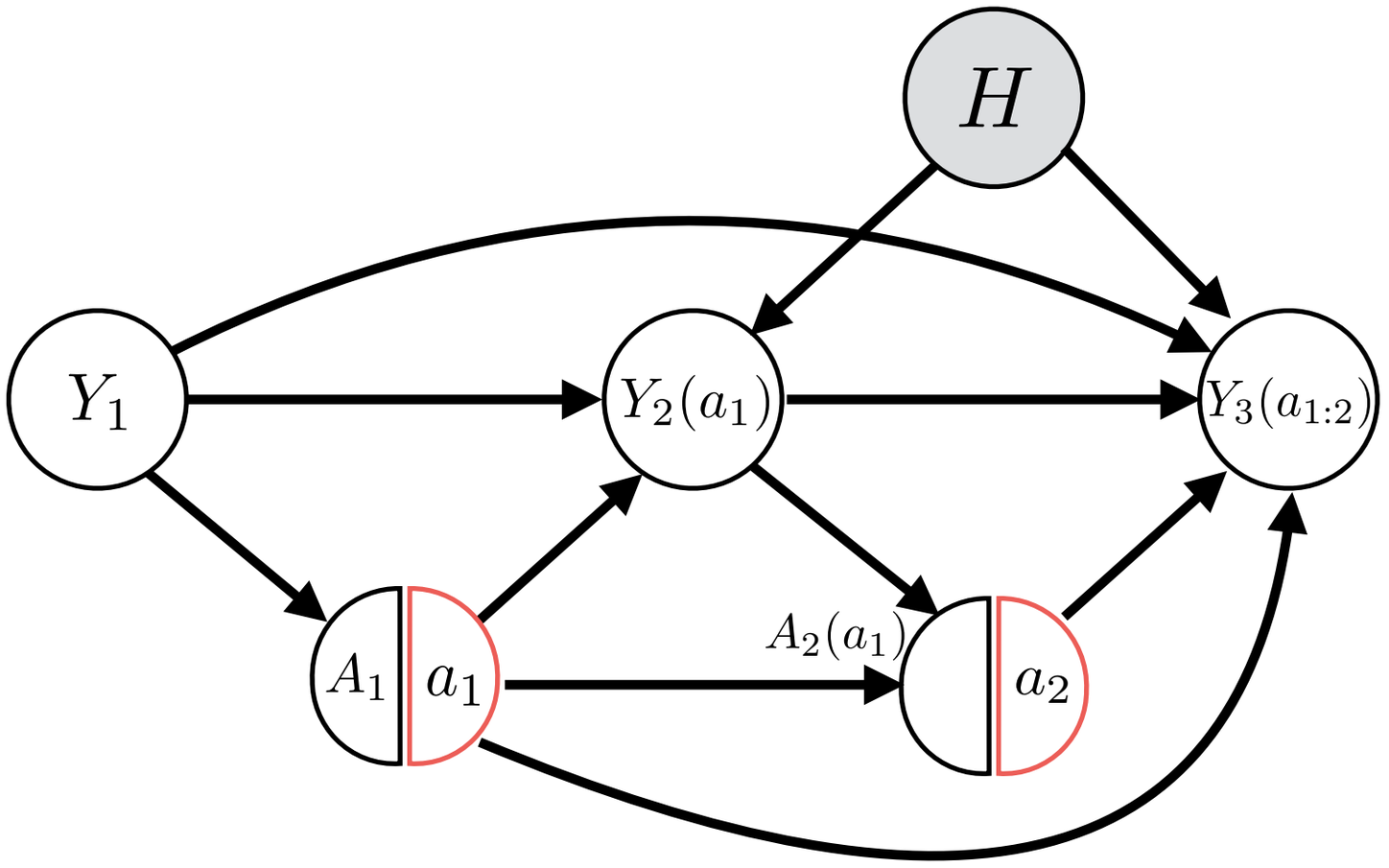
- As in single-treatment, single-outcome examples, we need assumptions that allow us to link **conditional distributions** to the target **potential outcome models**



Robins 1986

SWIG for Sequential Setting

- The SWIG is:



- The SWIG shows us that for each outcome, conditioning on previous outcomes d-separates from observed treatments

$$P(Y_1 = y_1)P(Y_2(a_1) = y_2 \mid Y_1 = y_1)P(Y_3(a_1, a_2) = y_3 \mid Y_1 = y_1, Y_2(a_1) = y_2)$$
$$= P(Y_1 = y_1)P(Y_2 = y_2 \mid Y_1 = y_1, A_1 = a_1)P(Y_3 = y_3 \mid Y_1 = y_1, Y_2 = y_2, A_1 = a_1, A_2 = a_2)$$

Robins 1986

Using Potential Outcomes Framework to Simulate RCT

- Our observational data is drawn from

$$Q \triangleq \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Obs}}(A \mid \mathbf{x})\mathrm{P}(Y \mid a, \mathbf{x}) = \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Obs}}(A \mid \mathbf{x})\mathrm{P}(Y(a) \mid \mathbf{x})$$

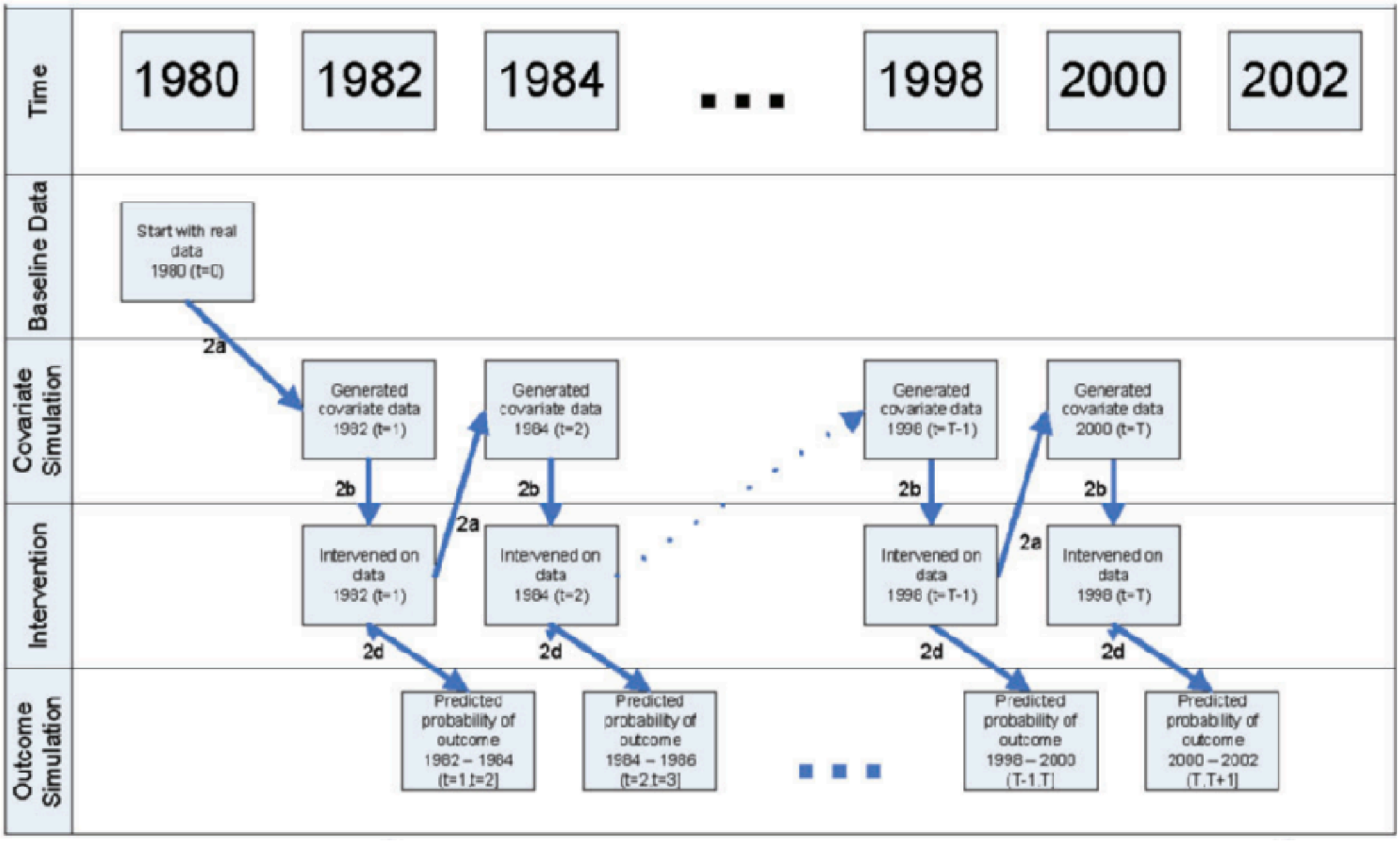
- We want experimental data drawn from

$$P \triangleq \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Exp}}(A)\mathrm{P}(Y \mid a, \mathbf{x}) = \mathrm{P}(\mathbf{X})\mathrm{P}_{\mathrm{Exp}}(A)\mathrm{P}(Y(a) \mid \mathbf{x})$$

- If we know potential outcome models:
  - Draw from empirical covariate distribution: $\mathbf{X} \sim \{\mathbf{x}_i\}_{i=1}^{n}$
  - Flip fair coin to assign treatment: $A \sim \mathrm{Bern}(0.5)$
  - Simulate outcome from model: $\mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x})$

Example: Intervening on Coronary Heart Disease

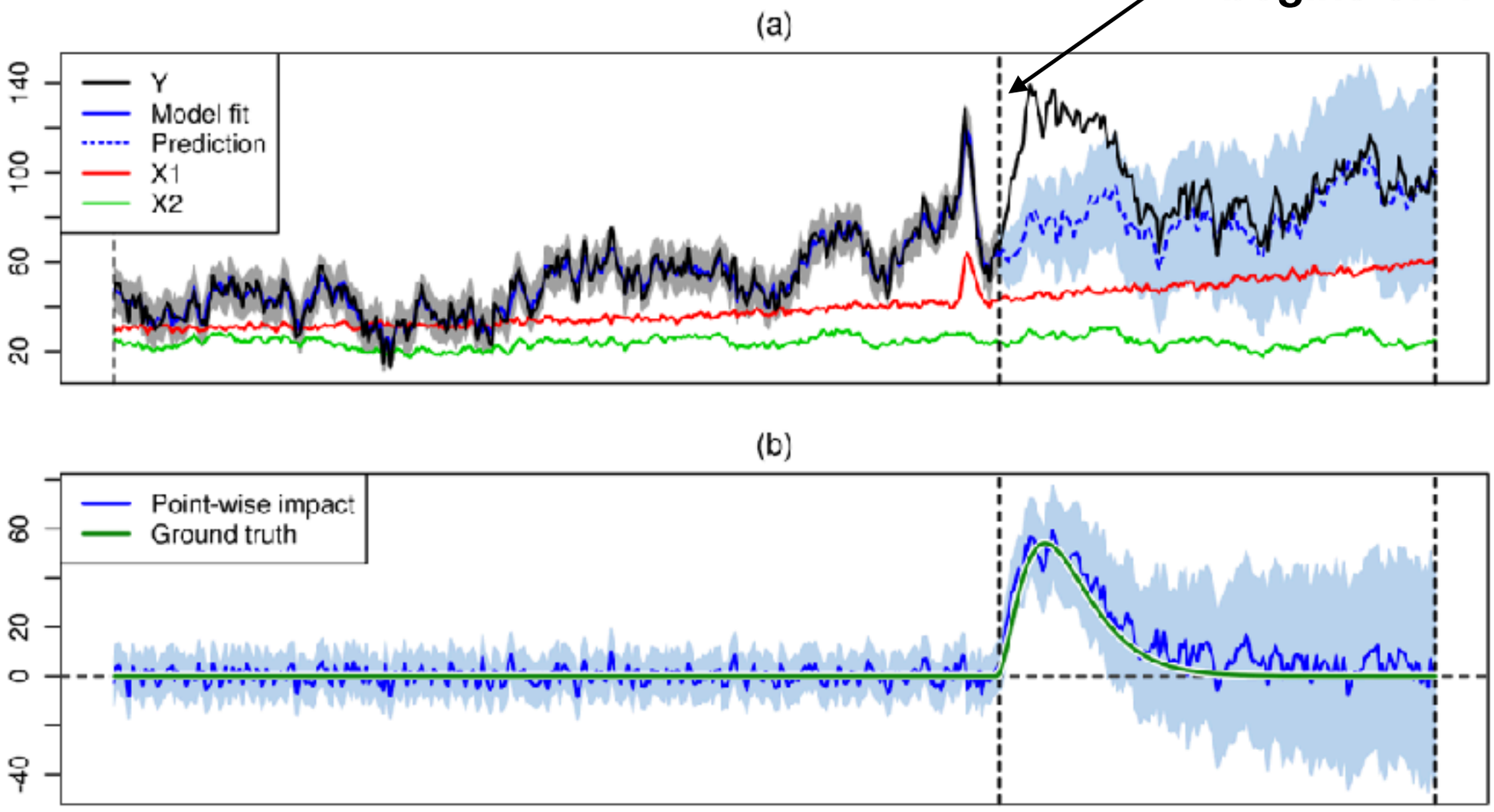
Estimate the population risk of coronary heart disease (CHD) under interventions such as quit smoking, maintain BMI < 25.



Taubman et al. 2009

Single Action on Discrete Time Series

- Google's "Causal Impact"
  - Target time series Y: receives intervention
  - Control time series X1, X2. (Do not receive intervention.)
    - These are predictive of Y.
  - The relation between Y and (X1, X2) remains the same pre and post intervention.
  - Predict the counterfactual of Y (under no treatment) using X1 and X2.



Brodersen et al. 2014

Many examples of using potential outcome

Albert, 2007

Bottou et al., 2013

Mithas and Krishnan, 2008

West et al., 2011

and many others …

Johansson et al., 2016

Chakraborty and Murphy, 2014

Athey and Imberns, 2016

Outline

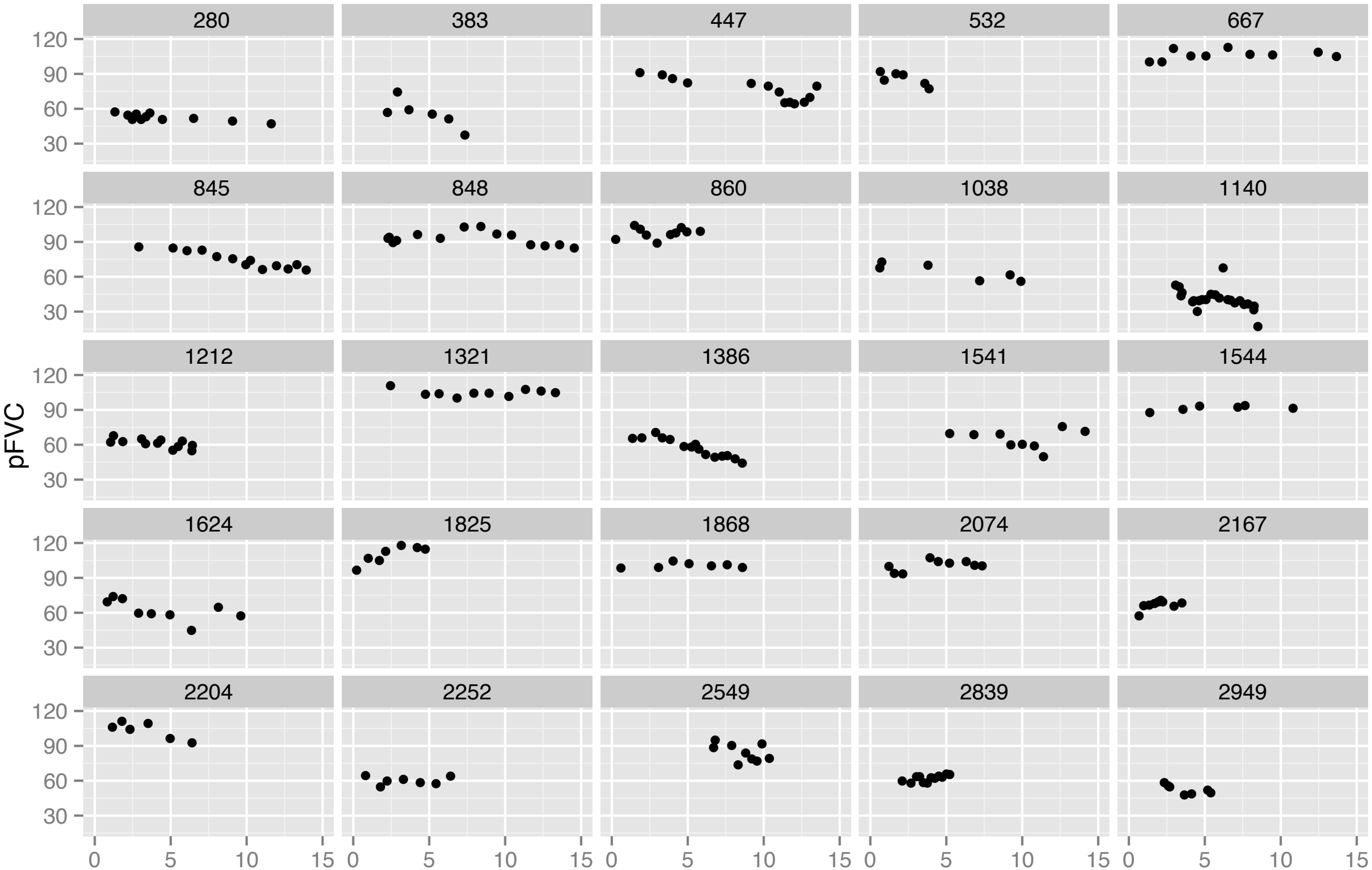
#1 Challenges with naive application of off-the-shelf predictive methods.

#2 The use of counterfactual reasoning for personalization
- BG: Potential Outcomes Framework
- BG: SWIGs

#3 Learning from noisy, observational traces
- Classical approaches that treat as discrete time data work poorly
- Treat as functional data
- BG: Gaussian Processes

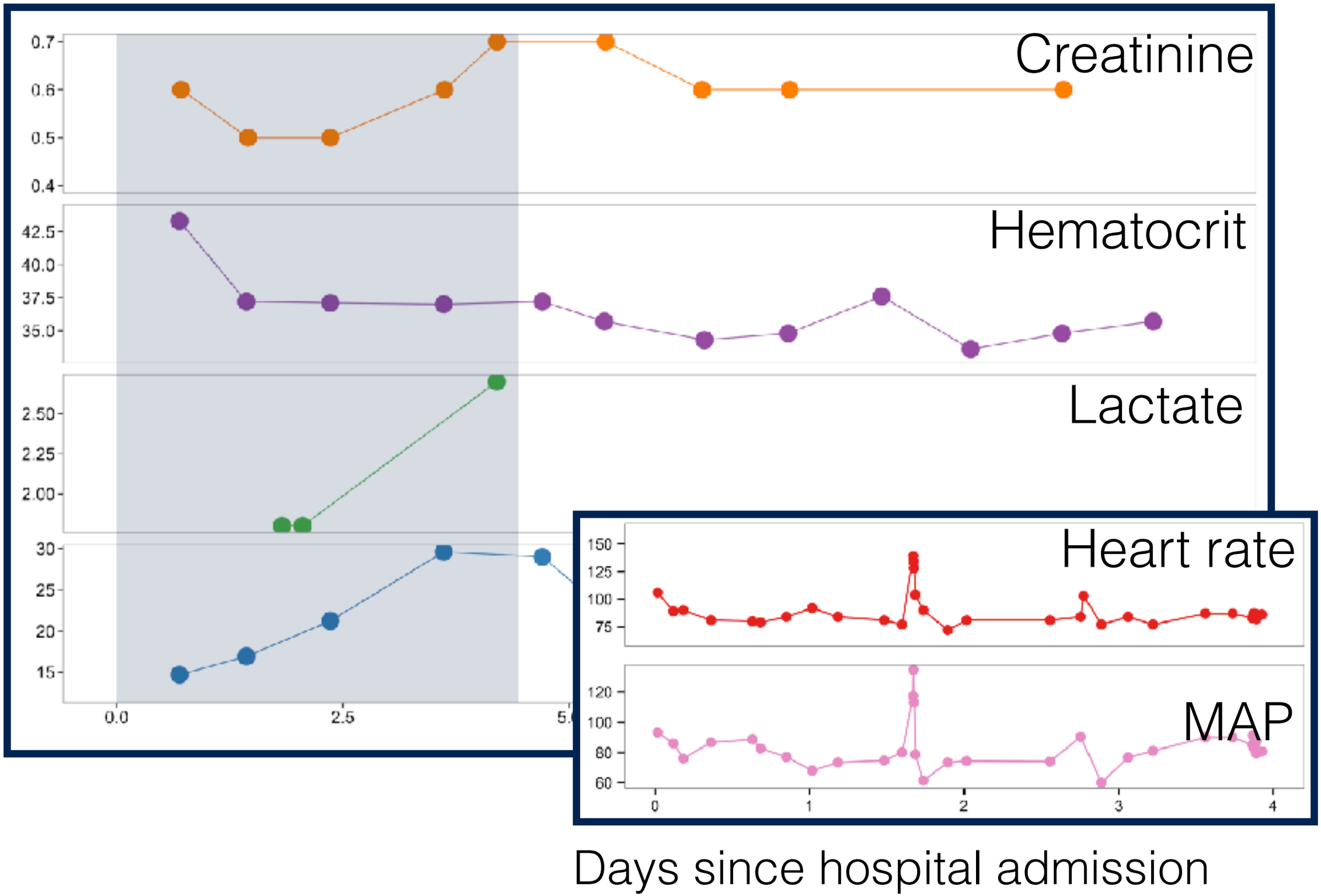
#4 CGPs — Counterfactual Reasoning from Traces
- Define framework
- Example applications

Imputation is not a solution



Lung marker data from 25 individuals

Very Different Sampling Granularities

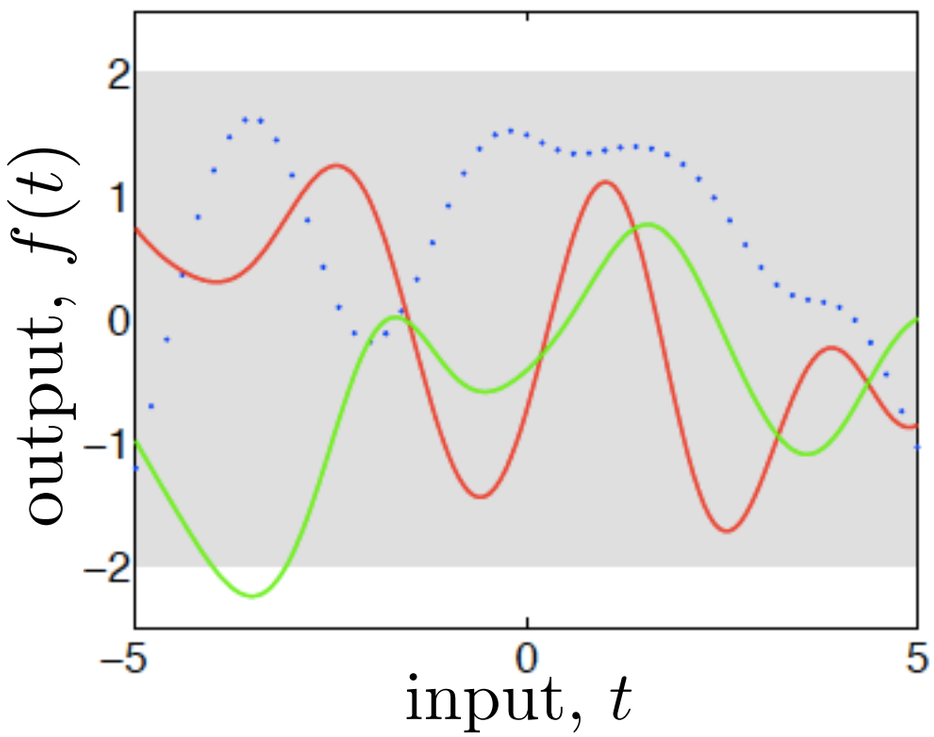


Background: Gaussian Processes

A Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A GP is fully defined using a mean and a covariance (kernel) function.

$$
\begin{aligned}
f(\mathbf{t}) &\sim \mathcal{GP}(m(\mathbf{t}), k(\mathbf{t}, \mathbf{t}')) \\
m(\mathbf{t}) &= E[f(\mathbf{t})], \\
k(\mathbf{t}, \mathbf{t}') &= E[(f(\mathbf{t}) - m(\mathbf{t}))(f(\mathbf{t}') - m(\mathbf{t}'))]
\end{aligned}
$$



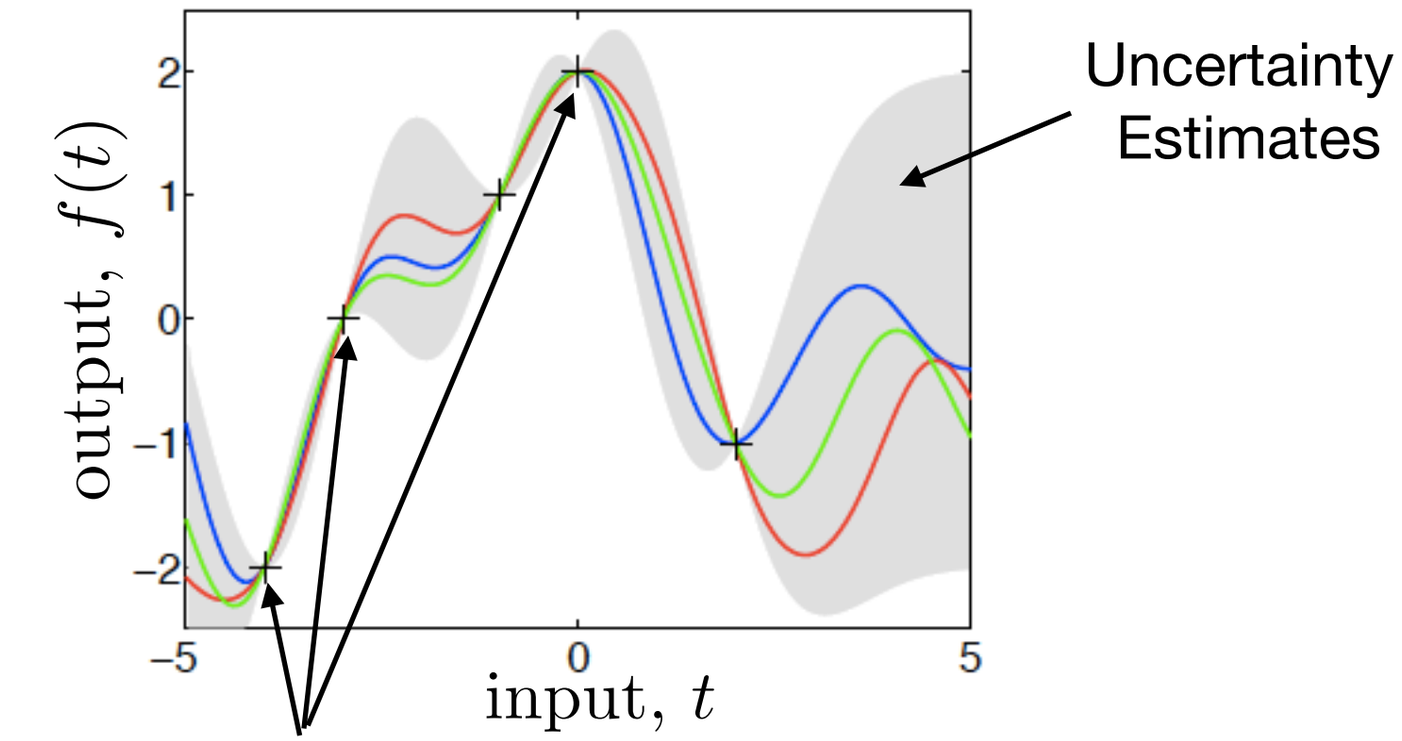
Rasmussen and Williams, 2006

Background: Gaussian Processes

Posterior f^* at t^* $(f^* = f(t^*))$ given the observations $(\mathbf{t}, \mathbf{f})$:

$$p(f_*, \mathbf{f}) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(t_*, t_*) & k(t_*, \mathbf{t}) \\ k(\mathbf{t}, t_*) & k(\mathbf{t}, \mathbf{t}') \end{bmatrix}\right)$$

$$f_* | t_*, \mathbf{t}, \mathbf{y} \sim \mathcal{N}(k(t_*, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}\mathbf{f},$$
$$k(t_*, t_*) - k(t_*, \mathbf{t})k(\mathbf{t}, \mathbf{t})^{-1}k(\mathbf{t}, t_*))$$



Rasmussen and Williams, 2006

Background: Gaussian Processes

Suppose we observe N points from an individual's EHR signals over time:

$$\{(t_n, y_n)\}, n = 1, 2, ..., N\,,$$

Denoted by $\mathbf{t} = \{t_n\}_{n=1}^{N}\,, \mathbf{y} = \{y_n\}_{n=1}^{N}$

How do we fit a GP to this data?

$$\mathbf{y} = f(\mathbf{t}) + \epsilon$$

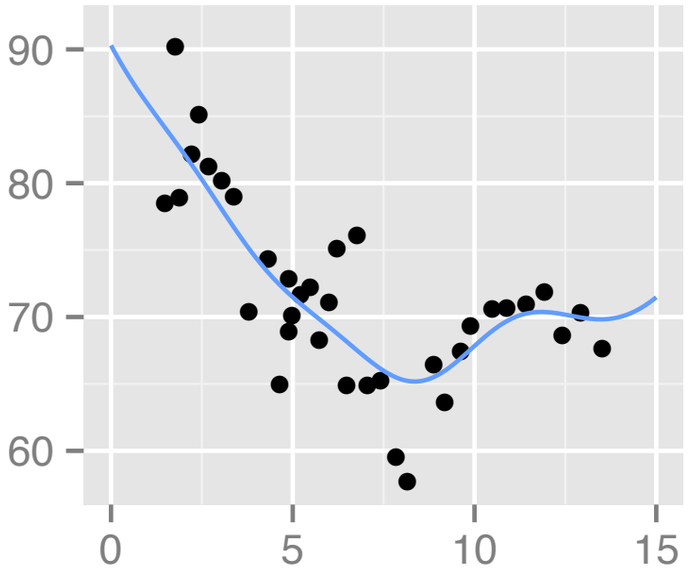
$$f(\mathbf{t}) \sim \mathcal{GP}(0, k(\mathbf{t}, \mathbf{t}'))\,, \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$$

1. Choose a kernel function: $k(t, t') = \exp(-\frac{1}{l^2}(t - t')^2)$
2. The mean function is usually set to 0: $m(\mathbf{t}) = 0$
3. Estimate l and σ^2 by maximizing the marginal likelihood function:

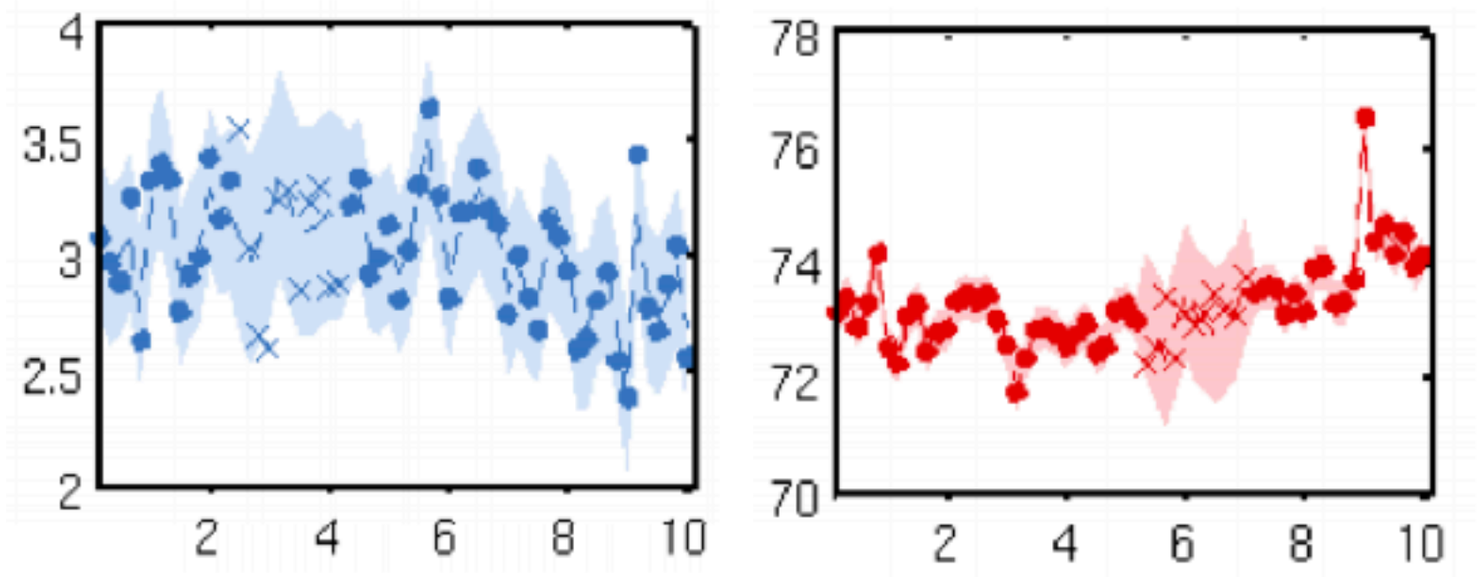
$$p(\mathbf{y}|\mathbf{t}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{t})\,\mathrm{d}\mathbf{f} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}(\mathbf{t}, \mathbf{t}') + \sigma^2\mathbf{I})$$

Examples: GPs for Modeling EHR Data

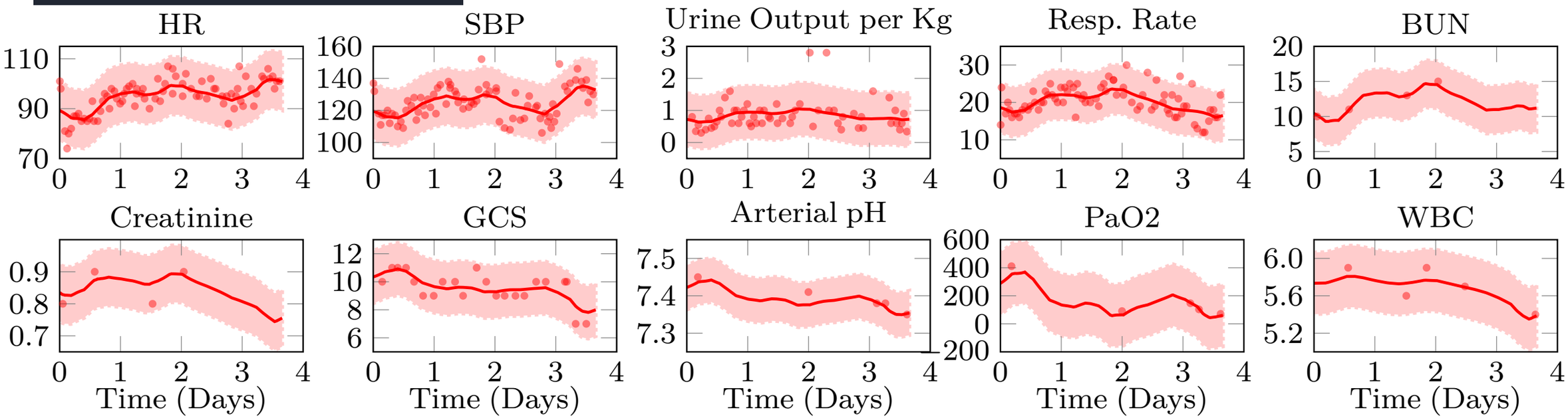
Schulam, Saria, 2015



Ghassemi et al., 2015



Soleimani et al., 2017



Ross and Dy, 2013

Futoma et al., 2016

Alaa et al., 2016

Outline

#1 Challenges with naive application of off-the-shelf predictive methods.

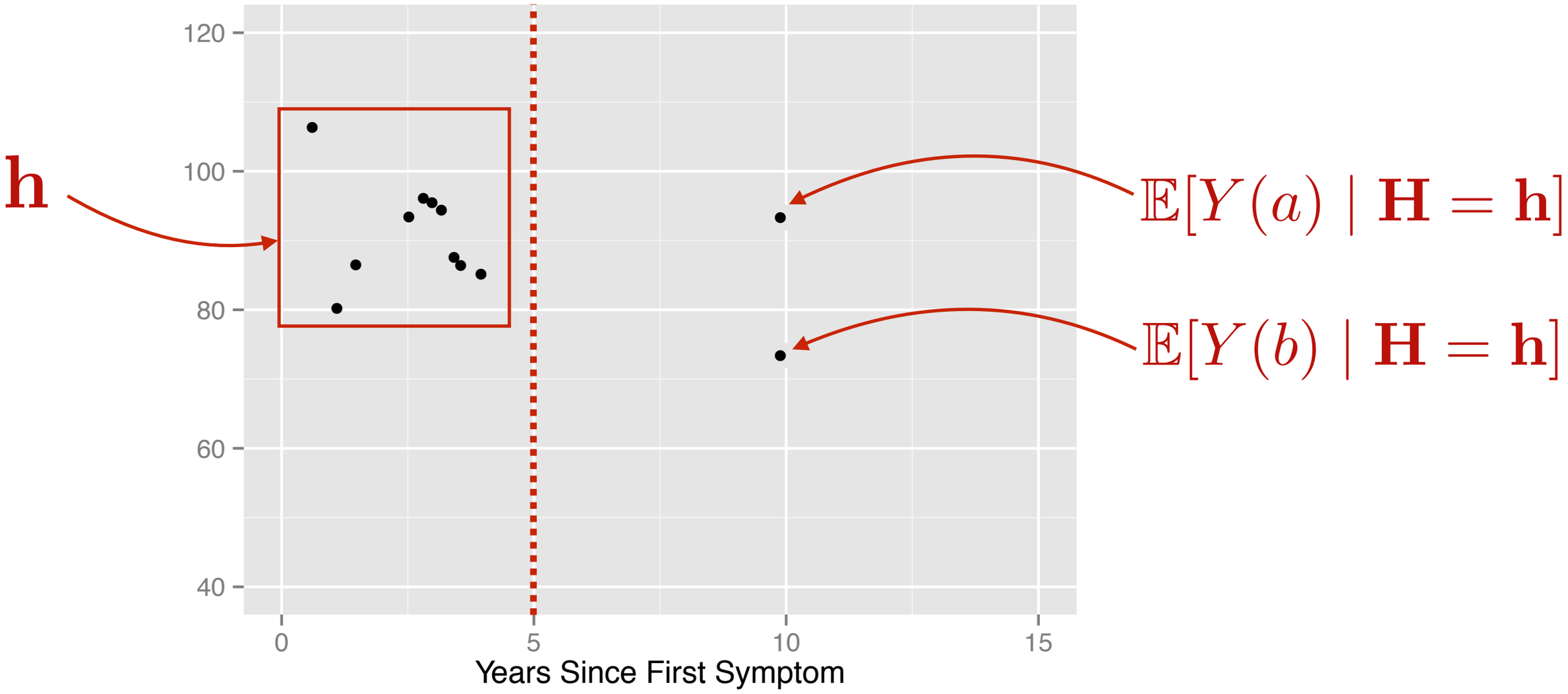
#2 The use of counterfactual reasoning for personalization
- BG: Potential Outcomes Framework
- BG: SWIGs

#3 Learning from noisy, observational traces
- Classical approaches that treat as discrete time data work poorly
- Treat as functional data
- BG: Gaussian Processes

#4 CGPs — Counterfactual Reasoning from Traces
- Define framework
- Example applications

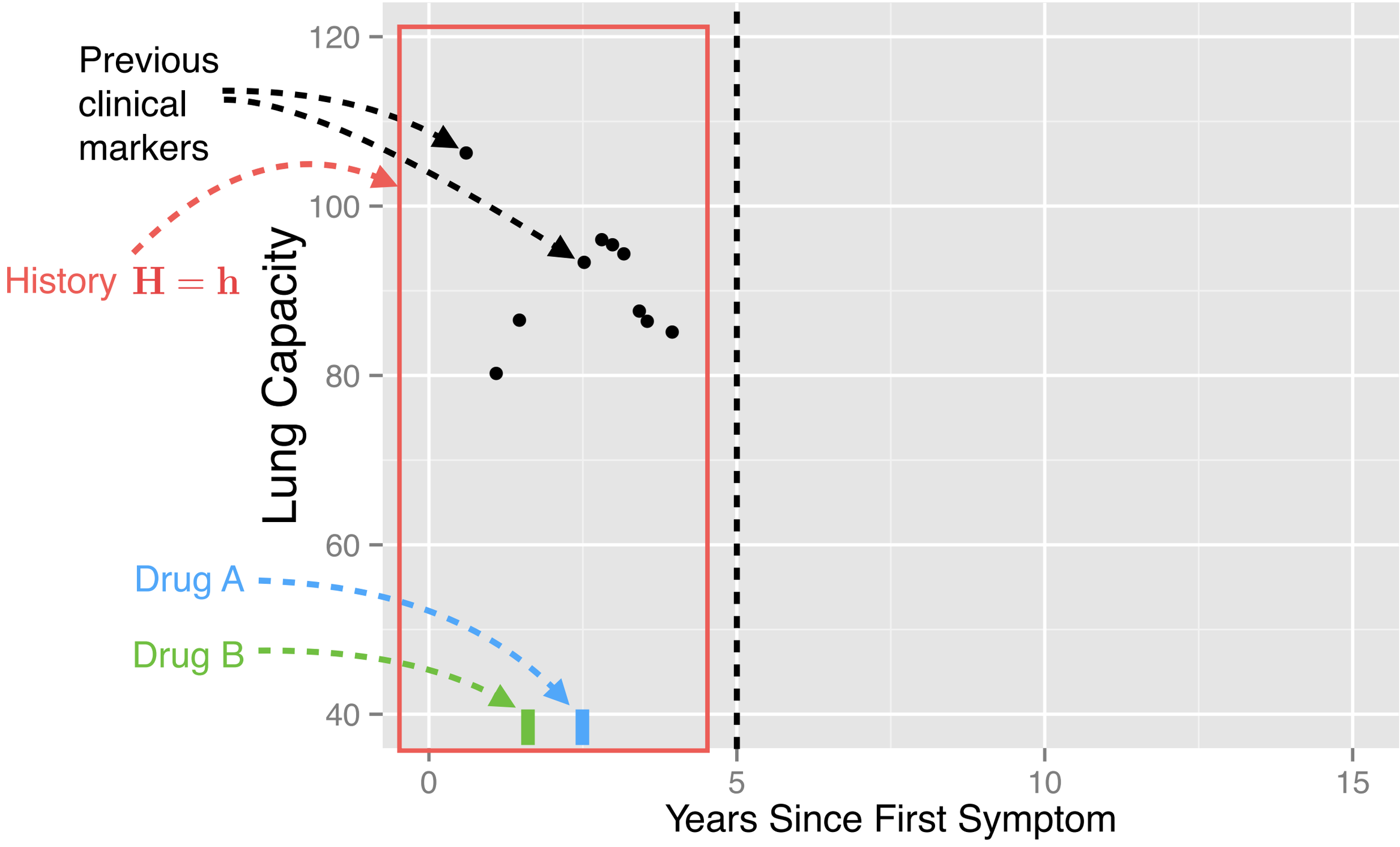
Personalized Treatment Planning

- Given what we know about a patient (their history), how should we choose a treatment plan?
- One solution: Estimate potential outcome at a future time under possible choices



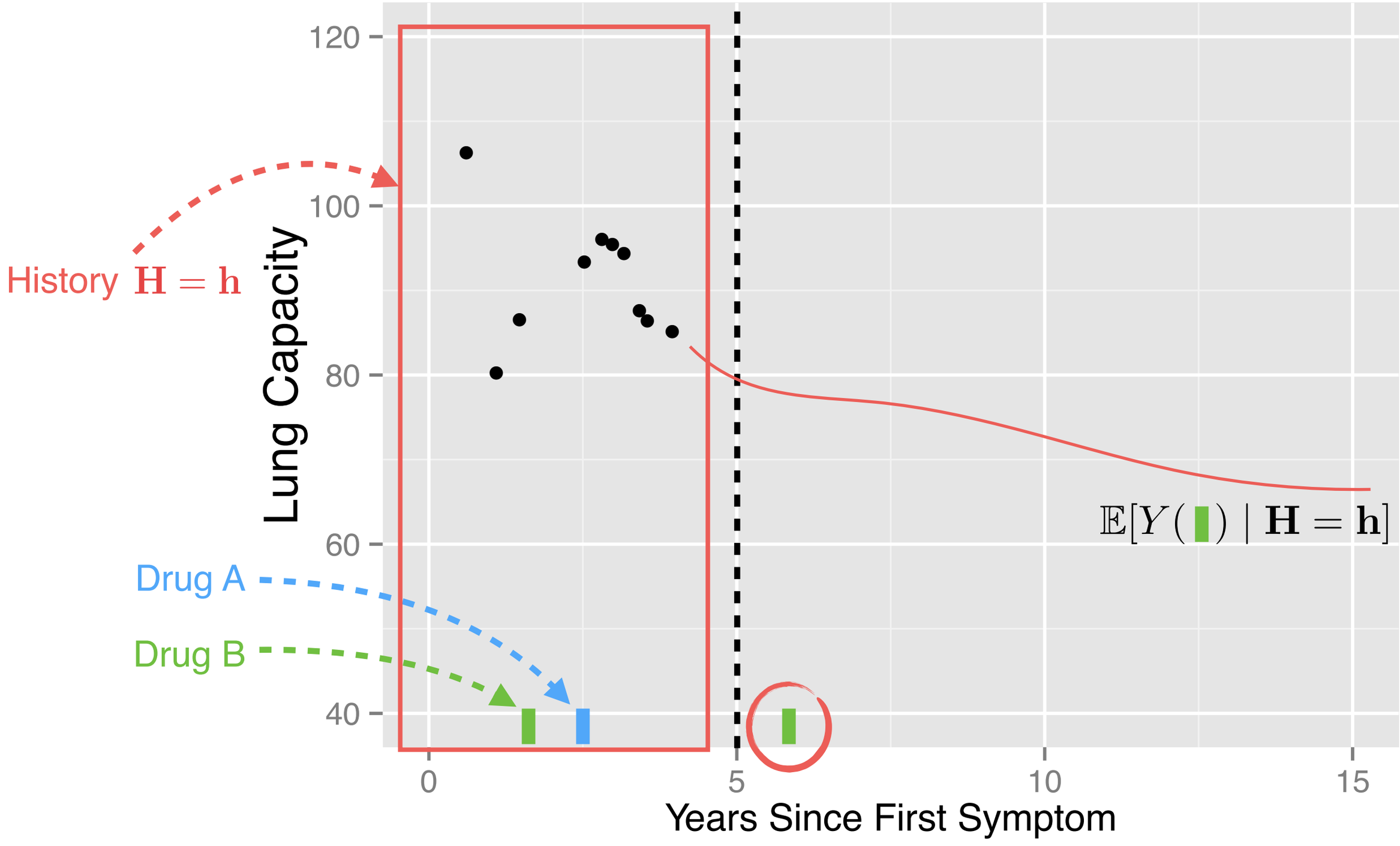
Personalized Treatment Planning

- Can we extend this idea to estimate an individual's future *trajectory?*



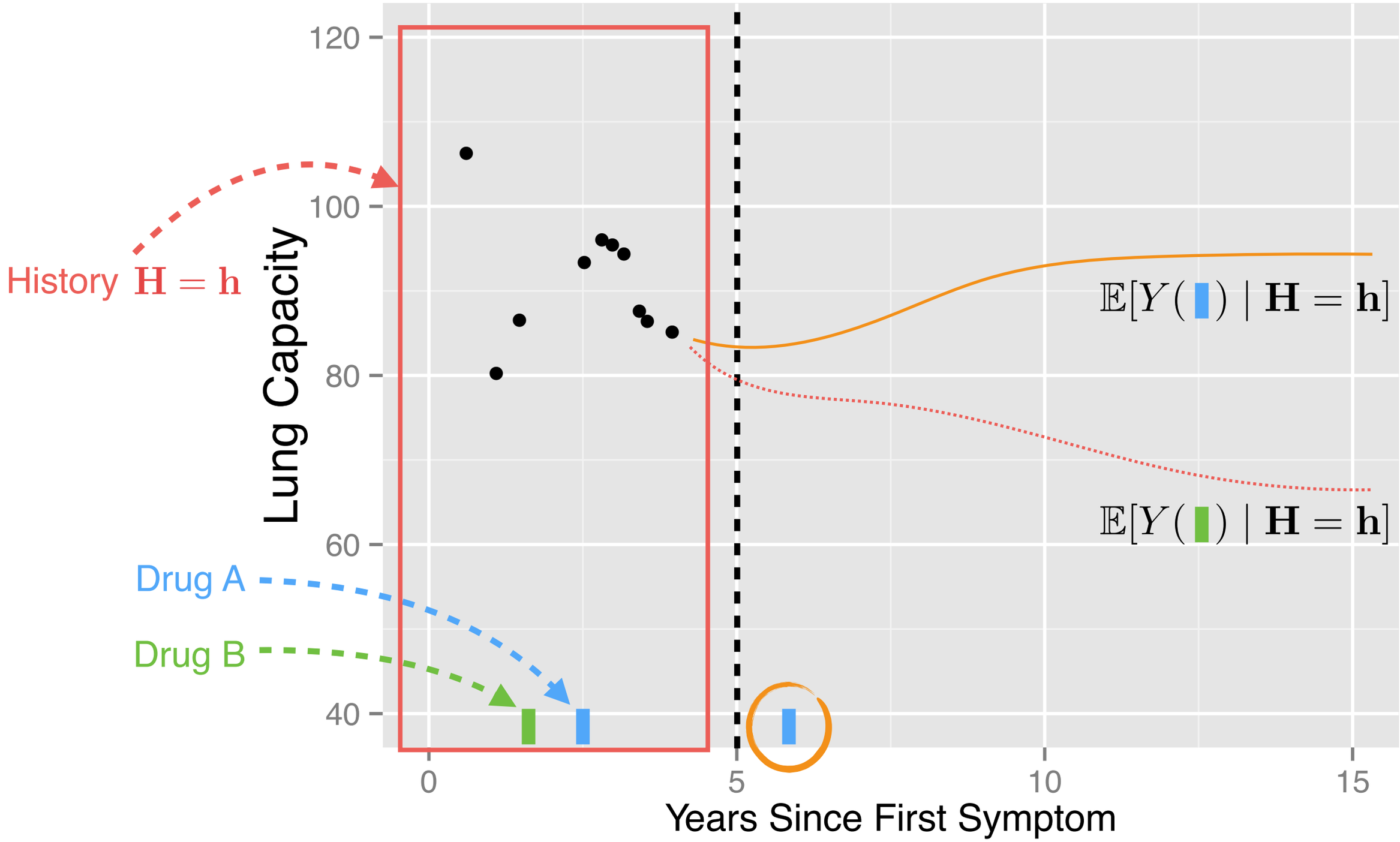
Personalized Treatment Planning

- Can we extend this idea to estimate an individual's future *trajectory?*



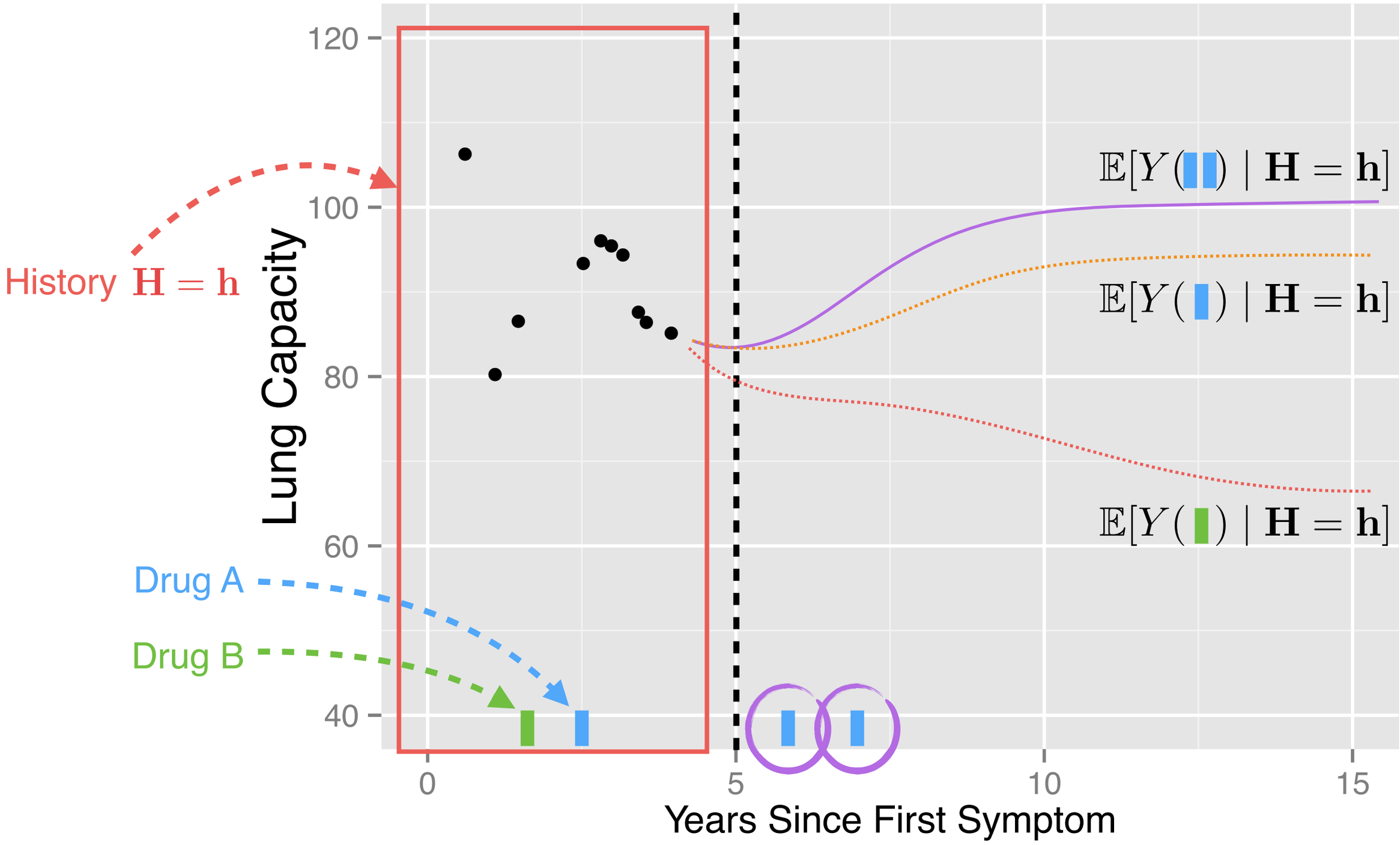
Personalized Treatment Planning

- Can we extend this idea to estimate an individual's future *trajectory?*



Personalized Treatment Planning

- Can we extend this idea to estimate an individual's future *trajectory?*



Trajectory-Valued Potential Outcomes

- We model our outcome as a stochastic process

$$\{Y_t : t \in [0, \tau]\}$$

- Our *target distribution* is the probability of the potential outcomes

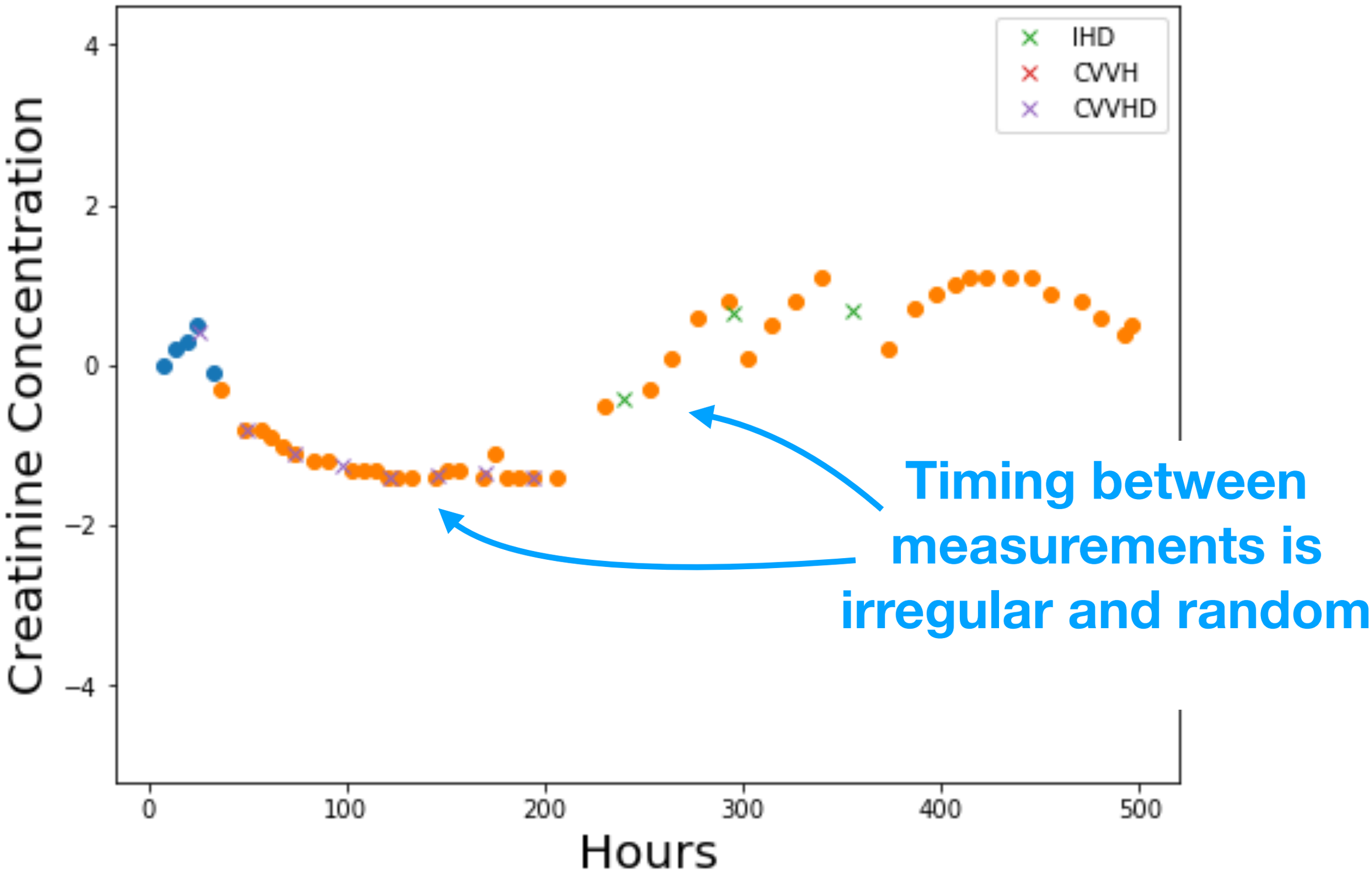
$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

Sequence of future interventions

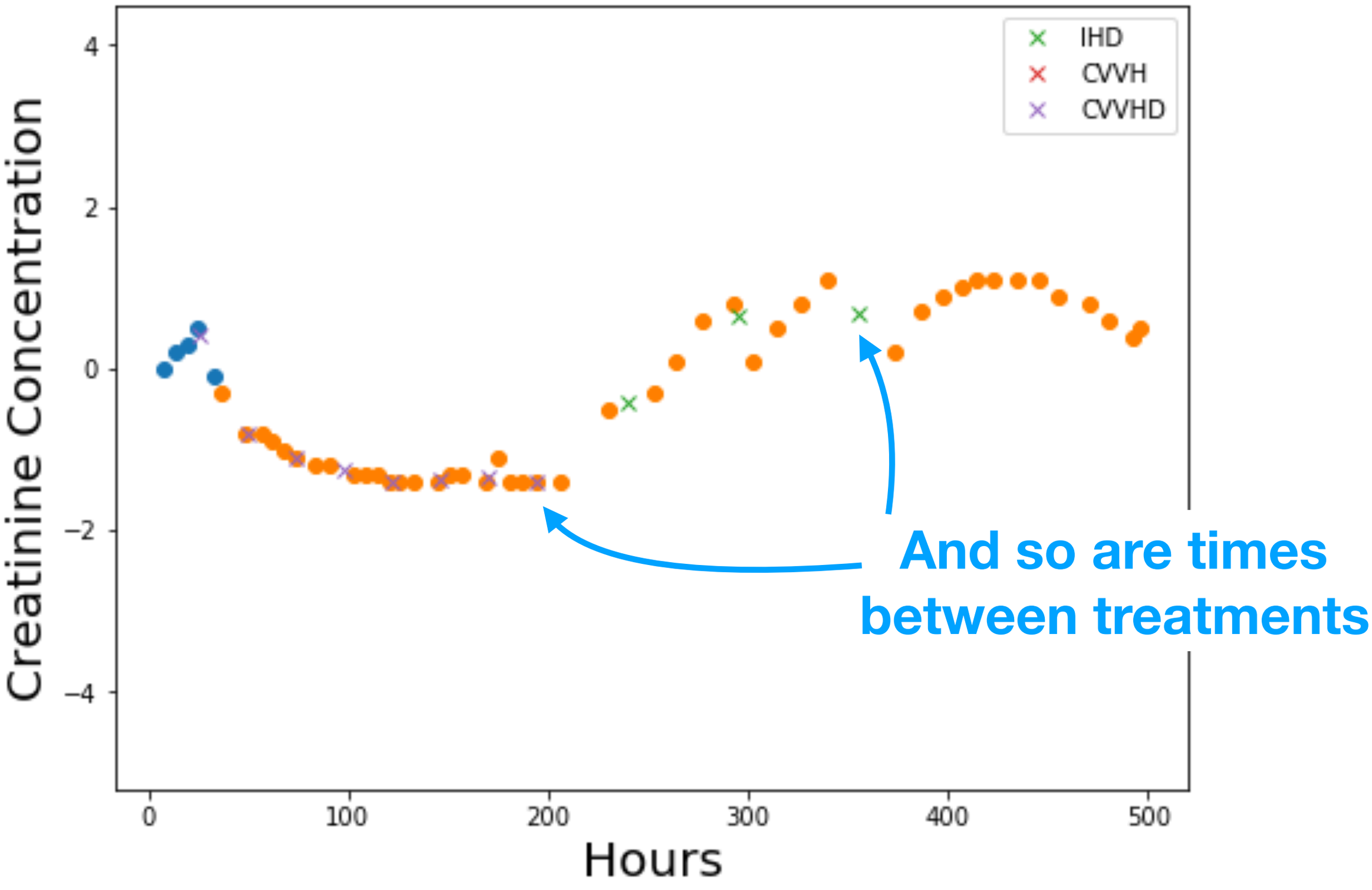
History at time t

Observational Traces

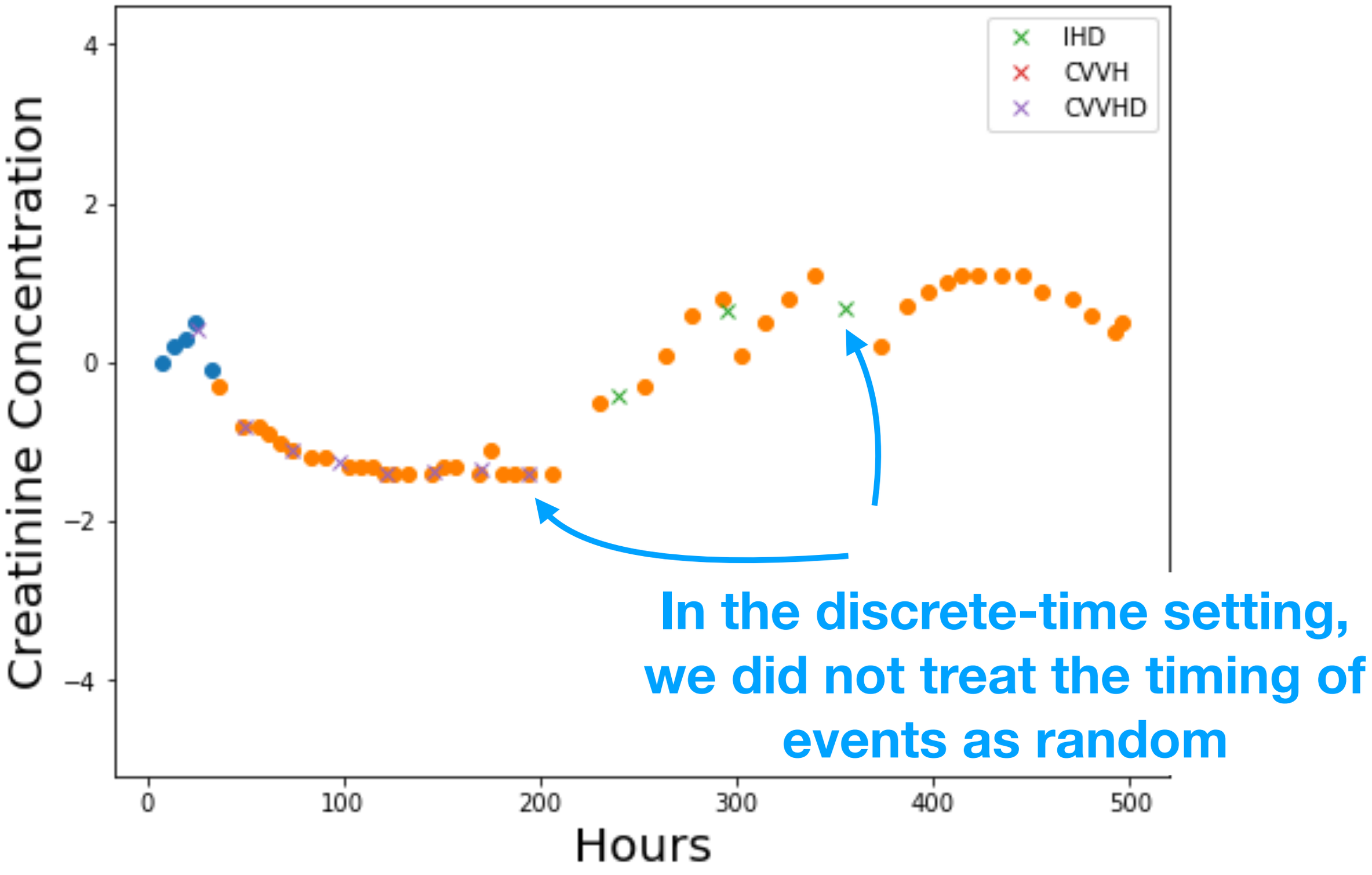
- Creatinine is a test used to measure kidney function.



Observational Traces



Challenges w/ Observational Traces



Observational Traces

- Our target distribution:

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

- We will learn using observational *traces*

Trace for individual i

$$\mathcal{D} \triangleq \left\{ \mathbf{h}_i = \{(t_{ij}, y_{ij}, a_{ij})\}_{j=1}^{n_i} \right\}_{i=1}^{m}$$

Timestamp

Observed outcome

Observed intervention

- The observed outcome or intervention can be “null” to account for when an outcome is measured, but no action is taken and vice versa

Learning Models from Observational Traces

- Road map:
  - (1) Posit probabilistic model of observational traces
  - (2) Derive maximum likelihood estimator
  - (3) Establish assumptions that connect probabilistic of observational traces to *target counterfactual model*

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

Counterfactual GPs for inference from traces

$$\mathcal{D} \triangleq \left\{ \mathbf{h}_i = \{(t_{ij}, y_{ij}, a_{ij})\}_{j=1}^{n_i} \right\}_{i=1}^{m}$$

- Posit a model for *when* a measurement is made or actions are taken and *what* the value of the measurements and actions are.
  - *Marked point processes (MPP) — posit* distribution over specific sequences of events
  - Today's talk: *Gaussian processes (GP)—relationship between the observed values*

See also:

Lok 2008

Arjas and Parner, 2004 **Cunningham et al., 2012** **Schulam and Saria, 2017**

Background: Marked Point Processes

- Point process: distribution over timestamps
  - Equivalent to a counting process $N_t = \sum_{i=1}^{N} \mathbb{I}_{(T_i \leq t)}$
  - To model MPP, define conditional hazard:

$$\lambda^*(t)\,\mathrm{d}t \triangleq \Delta\Lambda_t(\mathcal{H}_{t-}) = P(\Delta N_t = 1 \mid \mathcal{H}_{t-})$$

Probability of an increment in the counting process at time t given the history

- Marked point process: add a “mark” to each timestamp

$$\lambda^*(t, x) = \lambda^*(t) p^*(x \mid t)$$

Mark

Conditional density of mark

Defining the Mark Space

- We define the following mark space

$$\mathcal{X} = (\mathcal{Y} \cup \{\varnothing\}) \times (\mathcal{C} \cup \{\varnothing\}) \times \{0,1\} \times \{0,1\}$$

y — Outcome y (possibly null)

a — Action a (possibly null)

z_y — Is the outcome non-null?

z_a — Is the action non-null?

- And the corresponding conditional density

$$\begin{aligned} p^*(x \mid t) &= p^*(y, a, z_y, z_a \mid t) \\ &= p^*(y \mid t, z_y) p^*(a, z_y, z_a \mid t) \end{aligned}$$

Background: MPP Likelihood Function

- For any parameterization of the MPP, we can learn the parameters from traces using maximum likelihood

- For a single trace, maximize

$$\ell(\boldsymbol{\theta}, \{(t_j, x_j)\}_{j=1}^{n}) = \sum_{j=1}^{n+1} \log p^*(t_j, x_j)$$

- where: $x_j = (y_j, a_j, z_{y,j}, z_{a,j})$ $\qquad (t_{n+1}, x_{n+1}) = (\tau, \square)$

Censoring time

$$p^*(t_j, x_j) = \begin{cases} \lambda^*(t_j) p^*(x_j \mid t_j) \exp\{-\Lambda_j\} & \text{if } j \in \{1, \ldots, n\} \\ \exp\{-\Lambda_j\} & \text{if } j = n+1, \end{cases}$$

Counterfactual Likelihood

- We need to define the outcome model to finish defining the likelihood

- To learn parameters, using maximum likelihood estimation.

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log p^*_\theta(y_j \mid t_j, z_{yj}) + \sum_{j=1}^{n} \log \lambda^*_\theta(t_j) p^*_\theta(a_j, z_{yj}, z_{aj} \mid t_j) - \int_0^\tau \lambda^*_\theta(s)\, \mathrm{d}s.$$

Outcome model

Event time intensity

Event type model

* superscript denotes dependence on full history (past outcomes and actions)

Counterfactual Likelihood

- To learn the counterfactual GP, we maximize

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log p^*_\theta(y_j \mid t_j, z_{yj}) + \sum_{j=1}^{n} \log \lambda^*_\theta(t_j) p^*_\theta(a_j, z_{yj}, z_{aj} \mid t_j) - \int_0^\tau \lambda^*_\theta(s)\, \mathrm{d}s.$$

Outcome model

Event time intensity

Event type model

* superscript denotes dependence on full history (past outcomes and actions)

What assumptions are needed to connect this probabilistic model to the target counterfactual model?

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

Assumptions for Continuous-Time Traces

- The maximum likelihood estimate of the CGP learns our target distribution

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

- **<u>If</u>** we make assumptions:
  - (1) Consistency (as before)
  - (2) Continuous-time NUC
  - (3) Non-informative measurement times

Continuous-Time NUC

- Discrete-time setting, NUC:

$$Y(a) \perp A \mid \mathbf{X} = \mathbf{x} \quad : \quad \forall a \in \mathcal{A}, \forall \mathbf{x} \in \mathcal{X}$$

- Sufficient to assume that:

$$t \perp \{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_{t-}$$

Future potential trajectory for any sequence of actions ‘a’

- and that:

$$\{a, z_y, z_a\} \perp \{Y_s[\mathbf{a}] : s > t\} \mid t, \mathcal{H}_{t-}$$

Non-Informative Measurement Times

- Sufficient to assume:

$$p^*(y \mid t, z_y = 1)\,\mathrm{d}y = P(Y_t \in \mathrm{d}y \mid \mathcal{H}_{t-})$$

- Intuitively:
  - A measured outcome at time t is an unbiased observation of the trajectory at that time

- E.g. of a violation: measurements above a threshold are ignored and not recorded in the trace

Implications

- **If** the stated assumptions hold
  - (1) Consistency (as before)
  - (2) Continuous-time NUC
  - (3) Non-informative measurement times
- The maximum likelihood estimate of the CGP learns our target distribution

$$P(\{Y_s[\mathbf{a}] : s > t\} \mid \mathcal{H}_t)$$

Schulam and Saria, 2017

GPs in the context of CGP

Recall: CGP Likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n} \log p^*_\theta(y_j \mid t_j, z_{yj}) + \sum_{j=1}^{n} \log \lambda^*_\theta(t_j) p^*_\theta(a_j, z_{yj}, z_{aj} \mid t_j) - \int_0^\tau \lambda^*_\theta(s)\, \mathrm{d}s.$$

Outcome model parameterized using a GP

Event time intensity

Event type model

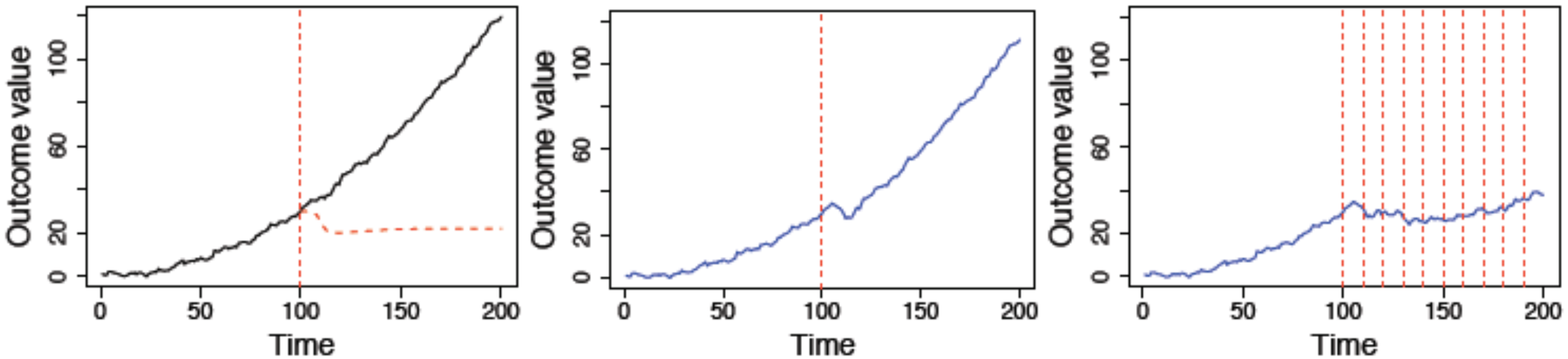
* superscript denotes dependence on full history (past outcomes and actions)

When estimating the outcome model, the event and action models can remain unspecified if we assume they have separate parameters.

Schulam and Saria, 2017

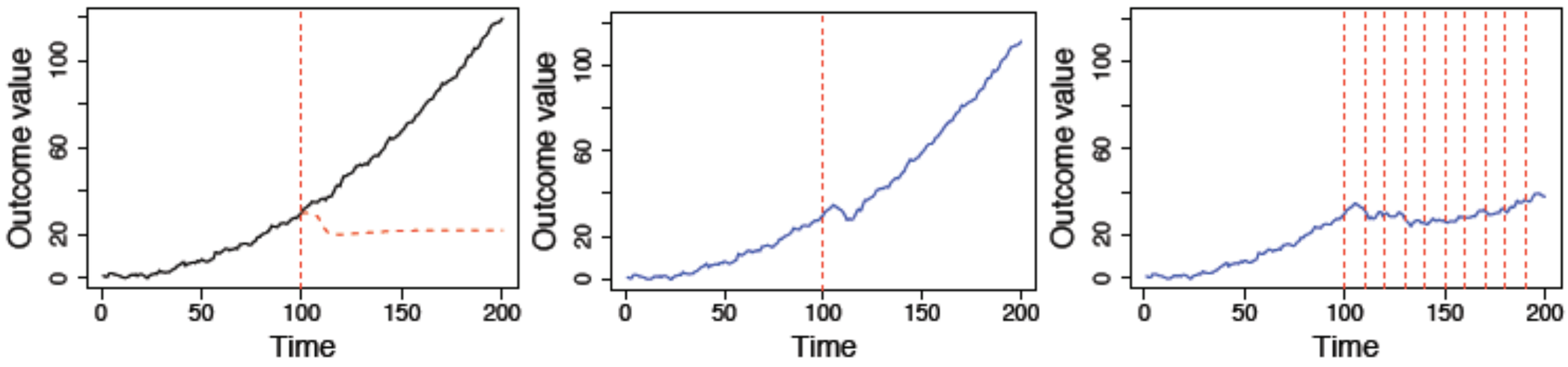
Additive Outcome Model

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$



Additive Outcome Model

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$



- Different choices for Baseline progression:

  Gaussian process, mixture of GPs, hierarchical GPs, etc.
- Treatment response: Pre-defined parametric functions,

  differential equations, etc.

Counterfactual GPs

Simulation Study: Given 12 hours of history, predict risk trajectory at hour 12.

Synthetically generated data:

- 200 trajectories for training and 200 for testing.

- Baseline Progression: GP with mean function parameterized using a 5-dimensional, order-3 B-spline
  - Class 1: declining mean
  - Class 2: first declining and then stabilizes
  - Class 3: Stable trajectory

    Covariance specified using Matern 3/2 kernels

- Additive Treatment: The intervention increases the mean function by a constant amount for 2 hours.

Schulam and Saria, 2017

GPs in the context of CGP

- **Baseline (RGP):** Identical to CGP model, but trained using classical GP maximum likelihood approach
  - RGP implicitly marginalizes over future interventions
  - RGP models $p(\{Y_t : t > 12\}|\mathcal{H}_{12})$

- **CGP** Model: Mixture of 3 GPs with unknown mean function coefficients, covariance function parameters, and treatment effect.
  - CGP controls for future interventions
  - CGP estimates $p(\{Y_t[\varnothing] : t > 12\}|\mathcal{H}_{12})$

Schulam and Saria, 2017

GPs in the context of CGP

Modified data generation that never produces actions after hour 12

Hours	RGP	CGP
$(12, 16]$	1.71	1.71
$(16, 20]$	1.86	1.86
$(20, 24]$	2.51	2.51

Mean absolute prediction error (MAE)

Schulam and Saria, 2017

GPs in the context of CGP

Modified data generation that never produces actions after hour 12

Hours	RGP	CGP
$(12, 16]$	1.71	1.71
$(16, 20]$	1.86	1.86
$(20, 24]$	2.51	2.51

With actions after hour 12

RGP	CGP
2.25	1.72
3.28	1.87
3.92	2.52

Mean absolute prediction error (MAE)

RGP's performance is degraded.

CGP's performance is approximately the same.

Schulam and Saria, 2017

Counterfactual Reasoning for Creatinine Trajectories

Real data Experiment: Predict creatinine level under no or alternative treatments

- Creatinine is a waste product in the blood.
- Patients with elevated creatinine levels and kidney injury receive dialysis, a procedure to filter the blood in place of kidneys

Schulam and Saria, 2017

Counterfactual Reasoning for Creatinine Trajectories

Real data Experiment: Predict creatinine level under no or alternative treatments

- Creatinine is a waste product in the blood.
- Patients with elevated creatinine levels and kidney injury receive dialysis, a procedure to filter the blood in place of kidneys

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$

$$K(t,t') = [1,t,t^2]\Sigma[1,t',t'^2]^T + \underbrace{\frac{\nu^2}{2\alpha^3}(2\alpha\min(t,t') + \mathrm{e}^{-\alpha t} + \mathrm{e}^{-\alpha t'} - 1 - \mathrm{e}^{-\alpha|t-t'|})}_{\text{integrated Ornstein-Uhlenbeck kernel}}$$

Schulam and Saria, 2017

GPs in the context of CGP

Real data Experiment: Predict creatinine level under no or alternative treatments

- Creatinine is a waste product in the blood.
- Patients with elevated creatinine levels and kidney injury receive dialysis, a procedure to filter the blood in place of kidneys

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$

$$K(t,t') = [1,t,t^2]\Sigma[1,t',t'^2]^T + \underbrace{\frac{\nu^2}{2\alpha^3}(2\alpha\min(t,t') + \mathrm{e}^{-\alpha t} + \mathrm{e}^{-\alpha t'} - 1 - \mathrm{e}^{-\alpha|t-t'|})}_{\text{integrated Ornstein-Uhlenbeck kernel}}$$

$$g(t;t_{j'}) = \underbrace{\frac{h_1 c}{c-d}(\mathrm{e}^{-d(t-t_{j'})} - \mathrm{e}^{-c(t-t_{j'})})}_{\text{short-term response}} + \underbrace{h_2(1-\mathrm{e}^{-r(t-t_{j'})})}_{\text{long-term response}}$$

$h_1, h_2, c, d, r, \alpha, \nu, \Sigma$ are free parameters.

Schulam and Saria, 2017

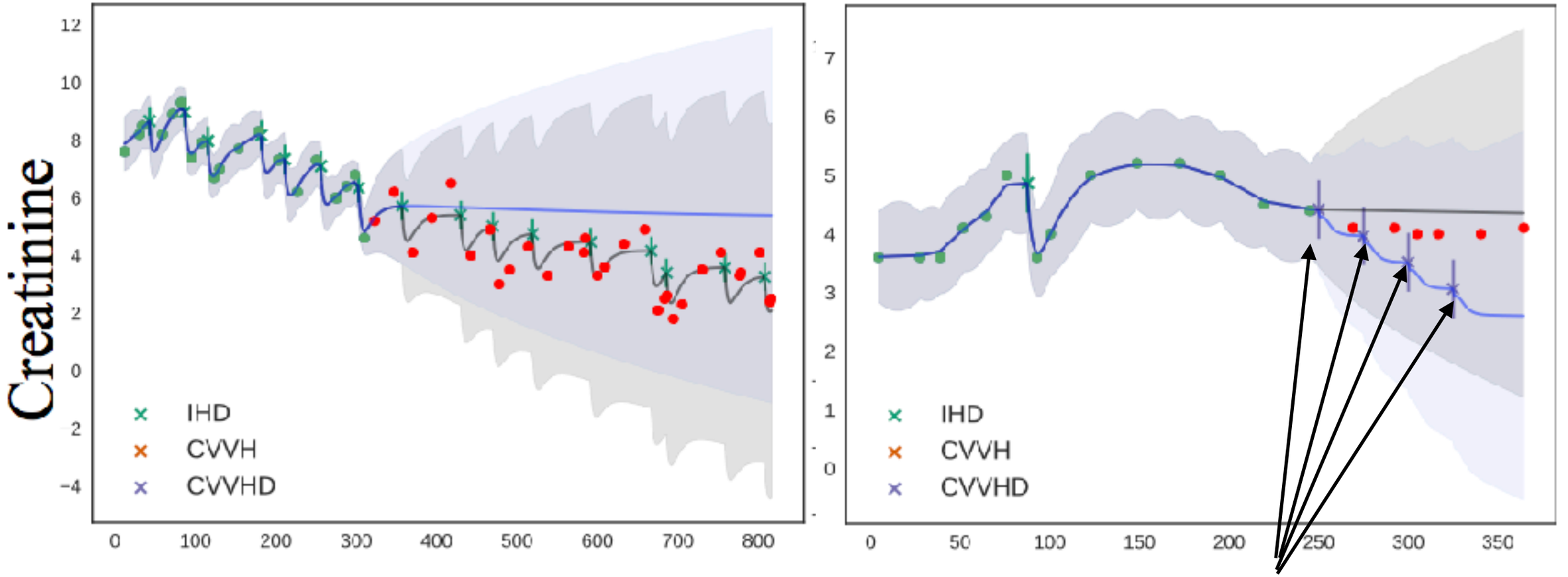
GPs in the context of CGP

Observed (blue dots) and held-out (red dots) data

black: Predictions under the factual sequence of treatments;

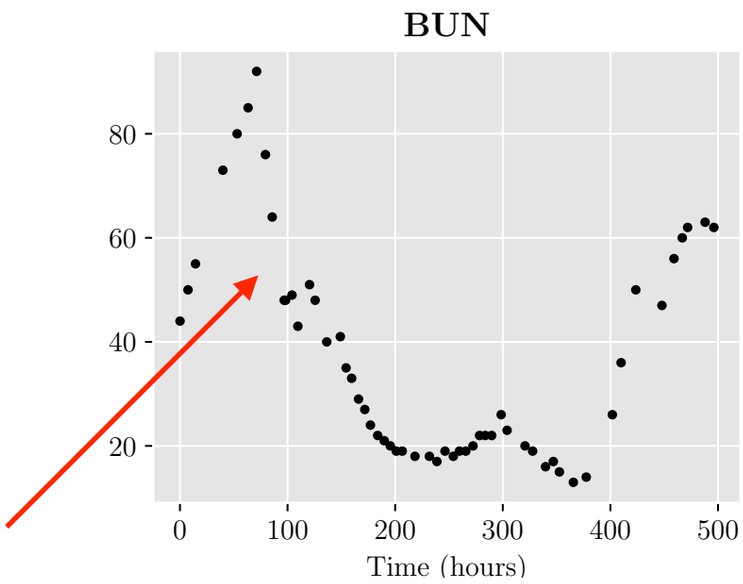
Blue: counterfactual predictions under no treatment

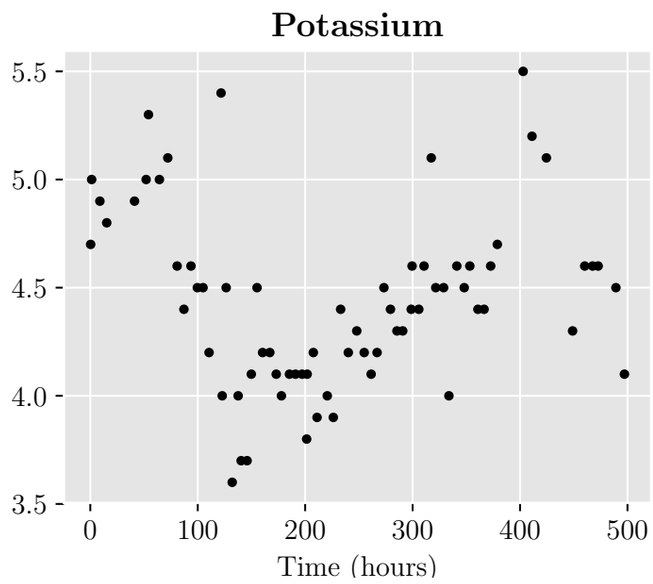
Blue: counterfactual predictions under CVVHD treatment

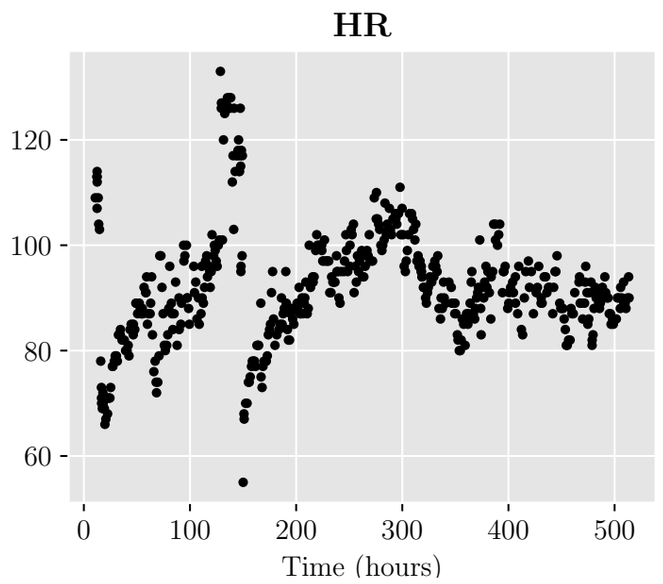


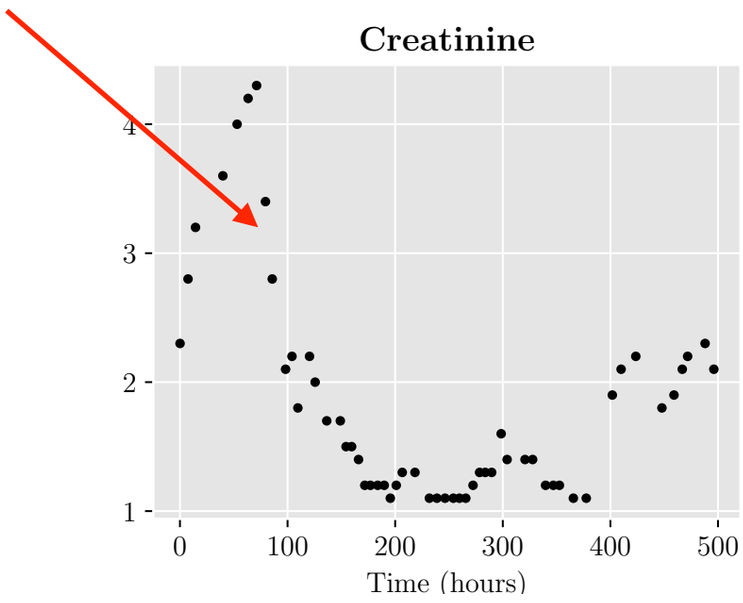
Schulam and Saria, 2017

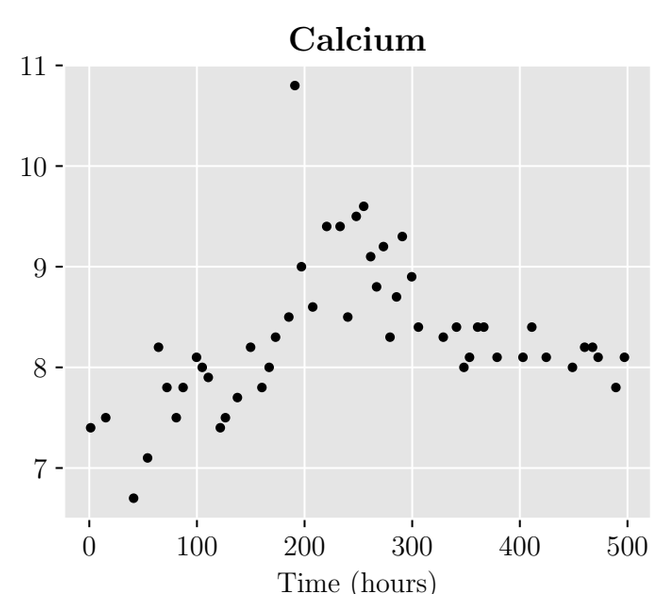
A Real ICU Patient with AKI

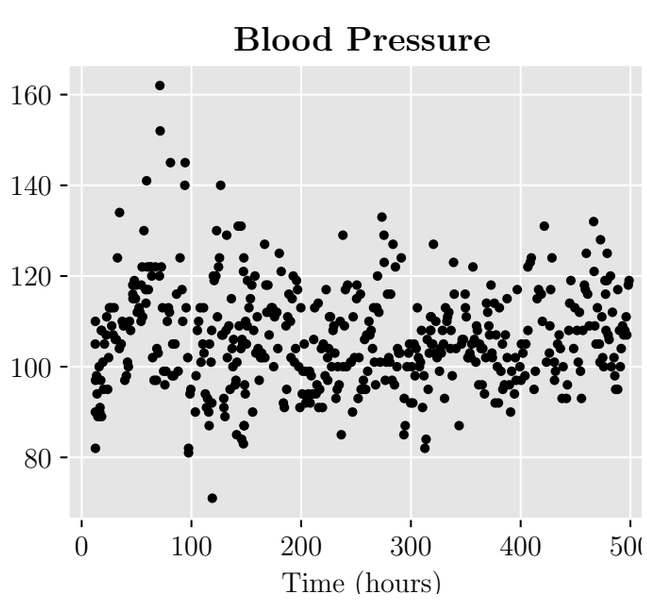






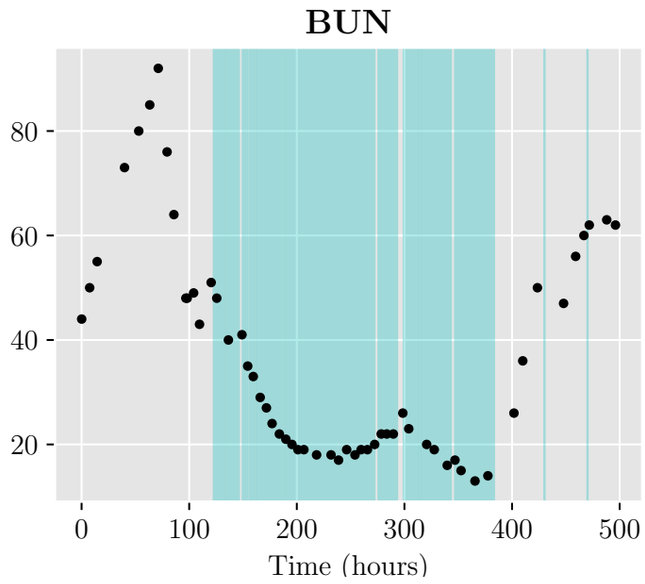


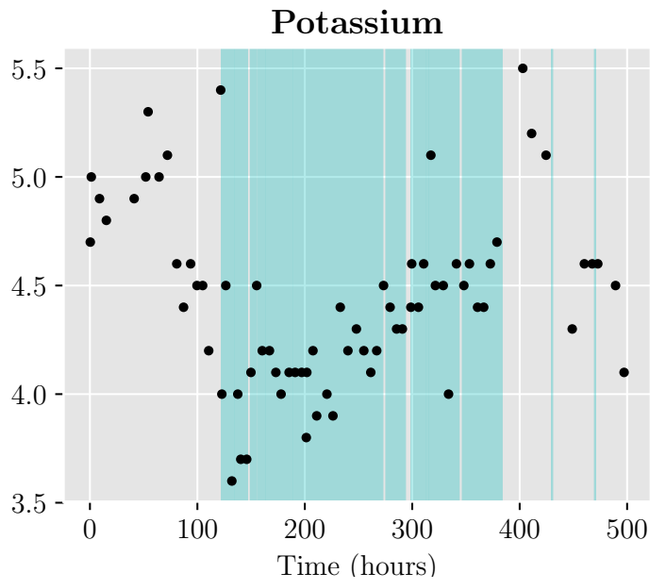


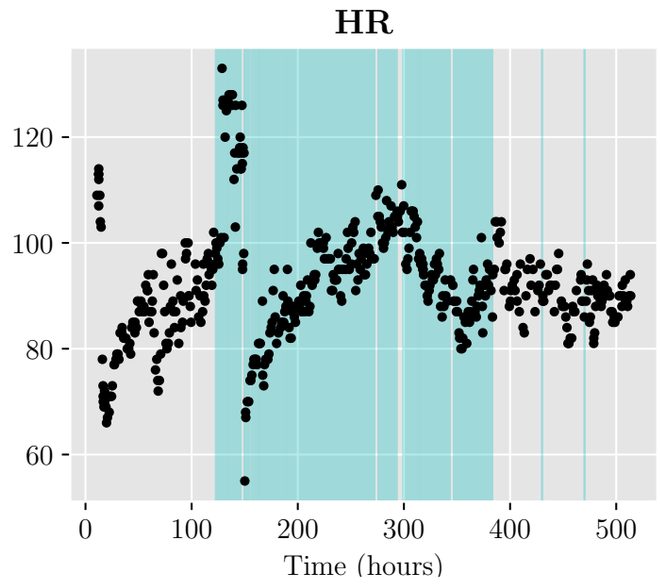


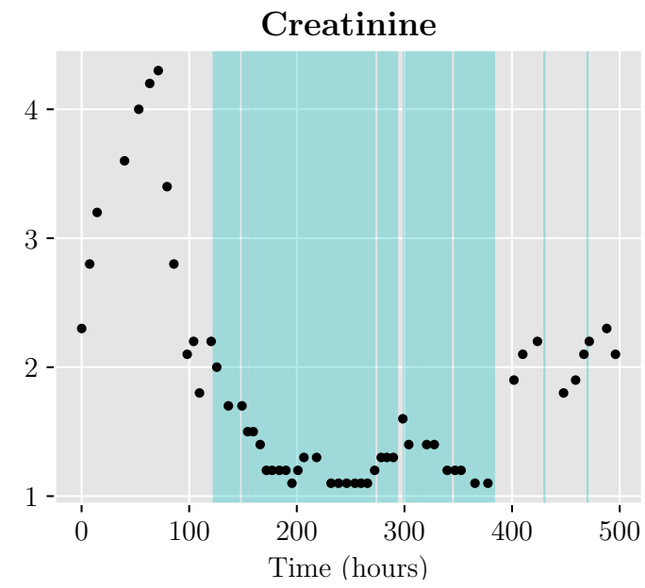
1. Irregularly sampled
2. Unaligned signals
3. Cross correlations

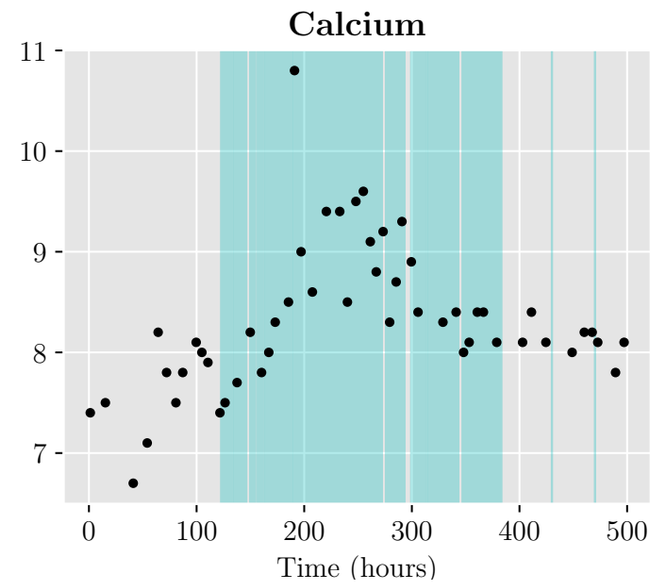
A Real ICU Patient with AKI

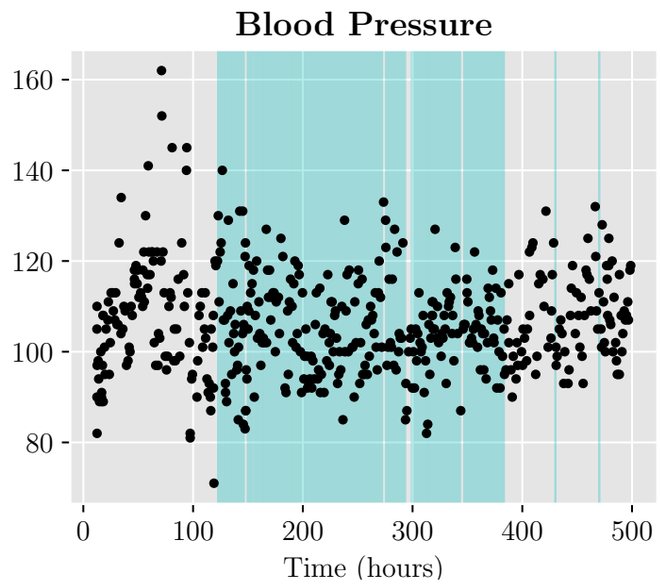












- Treatments (e.g., dialysis) is administered continuously

Extensions

- **Continuous-time actions, continuous-time multi-variate trajectories**

Soleimani, Subbaswamy, Saria, UAI 2017

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$

- Multi-output GP to capture correlations within and across signals
- Linear Time-Invariant System to describe treatment response to *arbitrary* treatment dosage and frequency.

- Other approaches: continuously-administered treatments

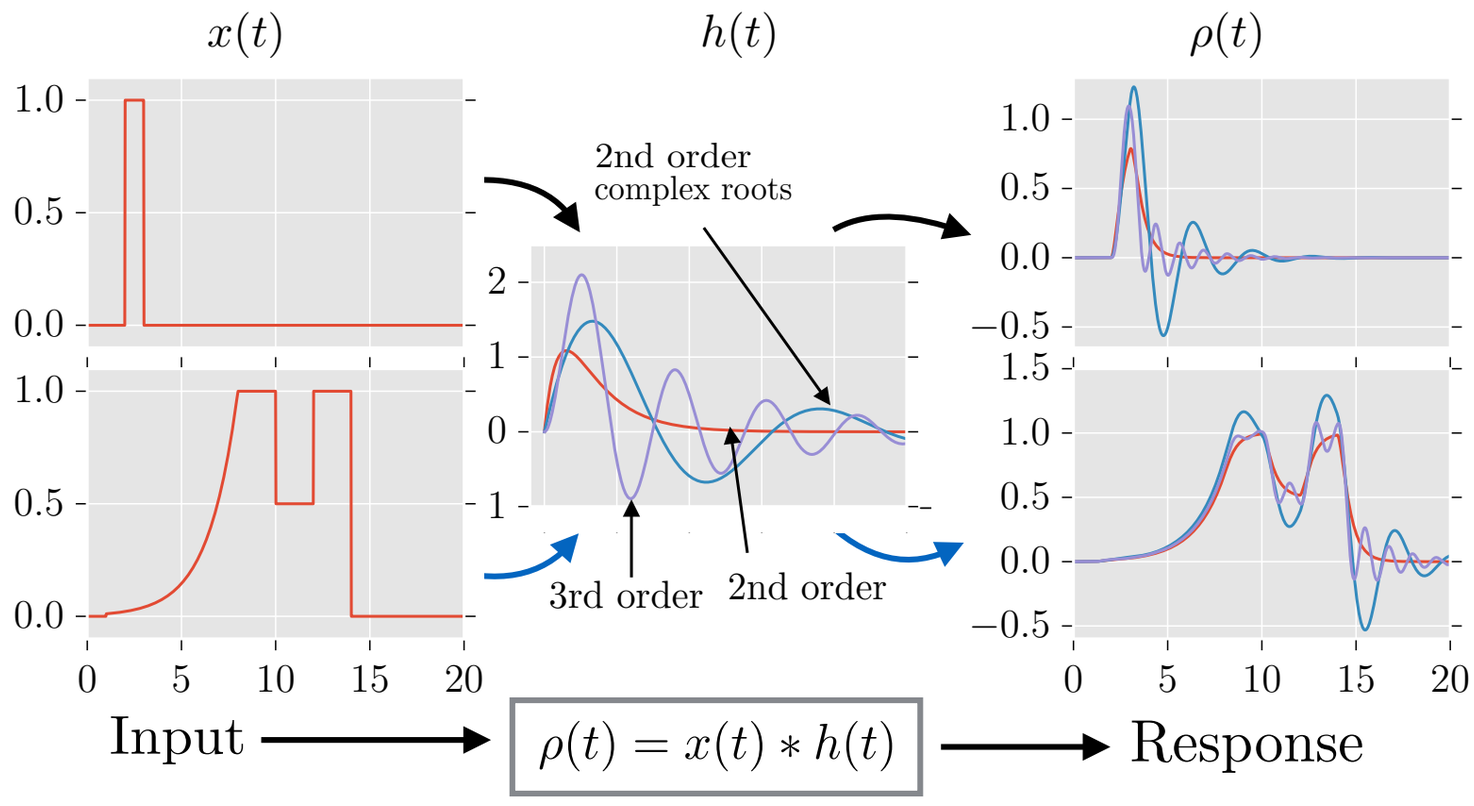
Johnson and Tsiatis, 2005

Tao, 2016

Continuous-time actions, continuous-time multi-variate trajectories

Input $x(t)$ convolved with *impulse-response* $h(t)$ to generate response $\rho(t)$

$$\rho(t) = x(t) * h(t) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)\mathrm{d}\tau$$



Similar ideas in pharmacokinetics:

Cutler, 1978

Rich et al., 2016

Shargel et al. 2005

Example: $h(t) = \dfrac{\alpha\beta}{\beta-\alpha}(\mathrm{e}^{-\alpha t} - \mathrm{e}^{-\beta t})1(t \geq 0)$

To allow sharing across signals: $g_d(t) = \psi \underbrace{\rho_0(t)}_{\text{shared}} + (1-\psi) \underbrace{\rho_d(t)}_{\text{signal-specific}}$

$\psi \in [0,1]$

Experiments: Data

- MIMIC II Clinical Database
- Continuous-time dialysis: CRRT and IHD
- Signals:
  - BUN, creatinine, potassium, calcium, BP, HR
- AKI patients with ≥ 10 observations in each signal
  - Dialysis only treatment for AKI
  - 67 patients
- First 70% of every patient's marker trajectory for training

Experiments: Baselines

- Bayesian Additive Regression Trees (BART)
  - Success in causal inference tasks
  - Typically for cross-sectional data
  - No natural representation of continuous-time treatments

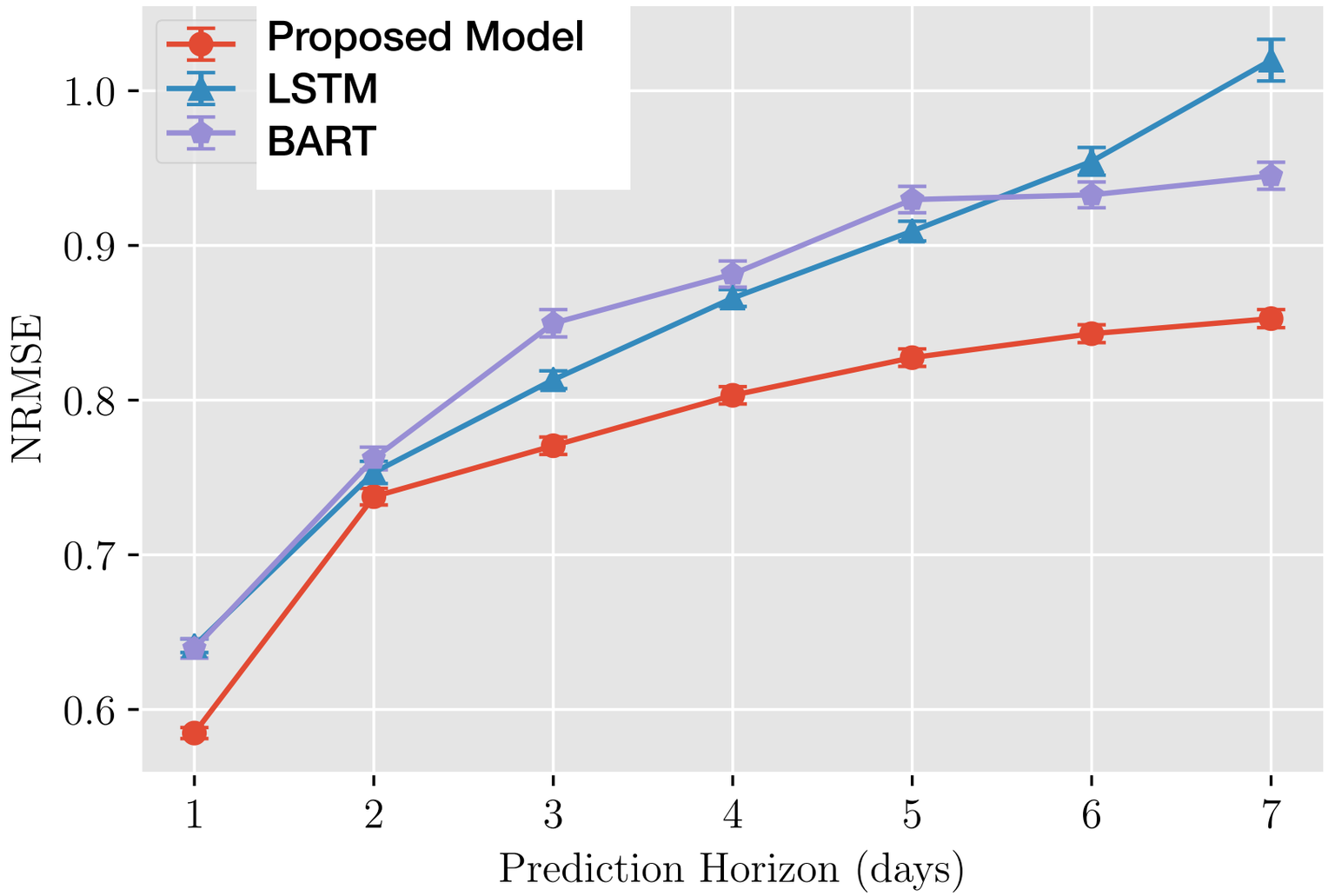
- Long Short Term Memory(LSTMs)
  - Neural network model for sequential data
  - Cannot naturally handle irregularly sampled data
  - No natural representation of continuous-time treatments

Chipman et al., 2010 **Hochreiter and Schmidhuber, 1997** **Dorie et al., 2016**

Experiments: Evaluation

- Baselines:
  - Separate model for each signal
  - Binning, LOCF imputation
  - Features: bin midpoint, time since last treatment, last treatment dose, marker value
  - Train using previous L bins
  - One step ahead prediction
- Metric: RMSE normalized by standard deviation of each signal, averaged across markers (NRMSE)

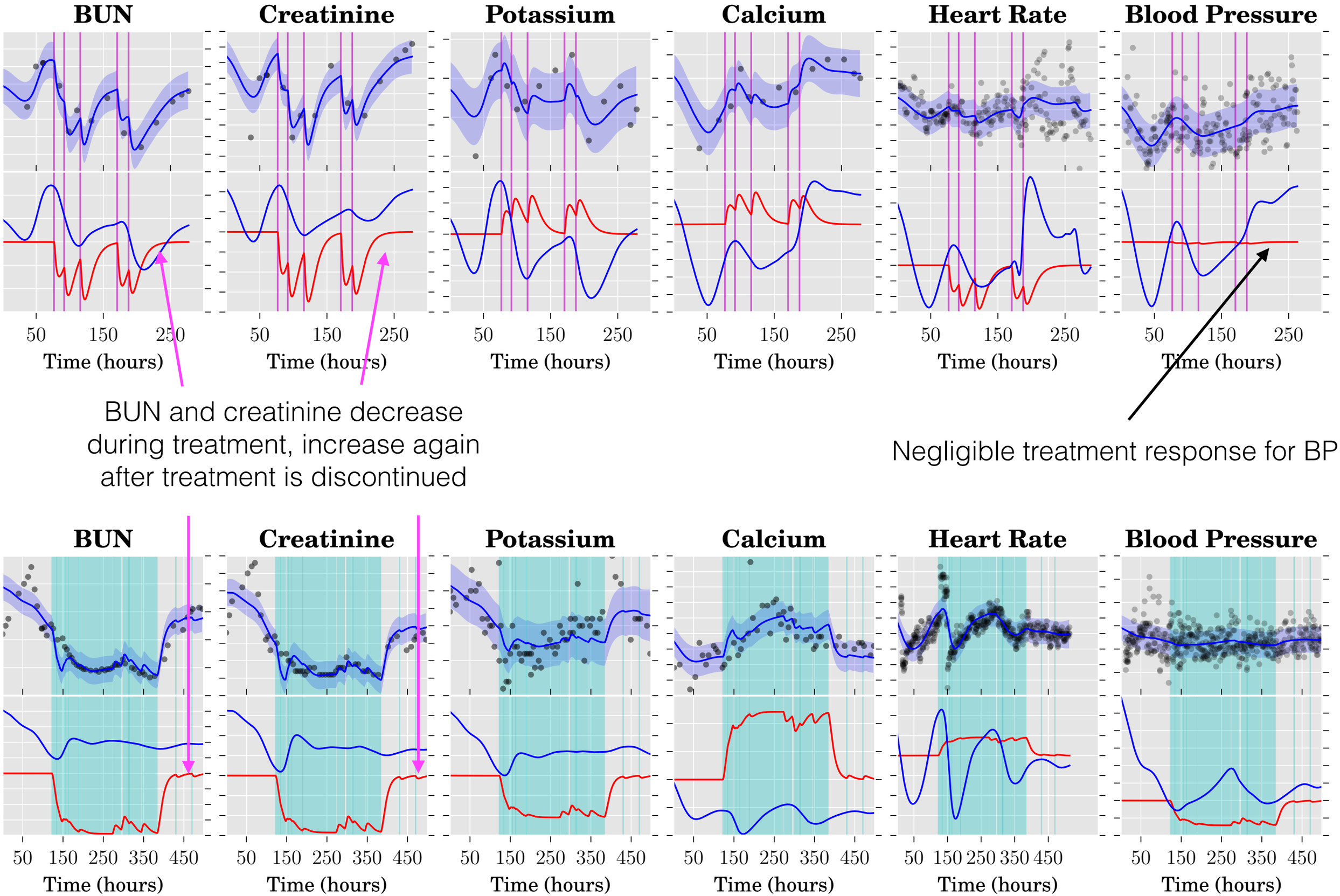
Quantitative Results



- Better relative performance at longer prediction horizons
- For horizon 7: on test regions with treatment, 15% than BART and 8% better than LSTM

Soleimani, Subbaswamy, Saria, UAI 2017

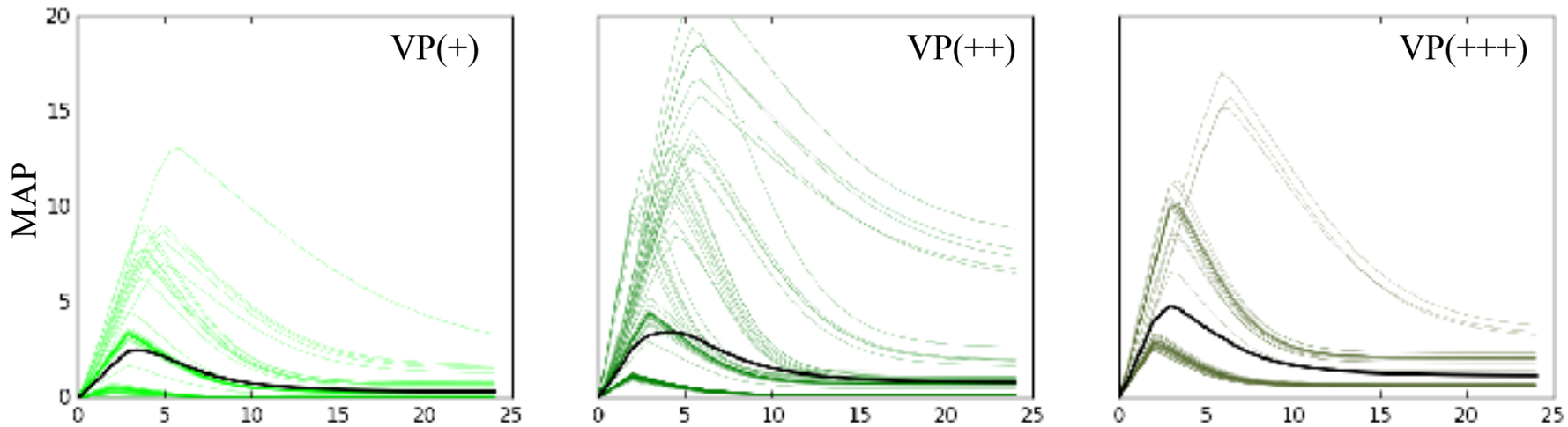
Qualitative Results



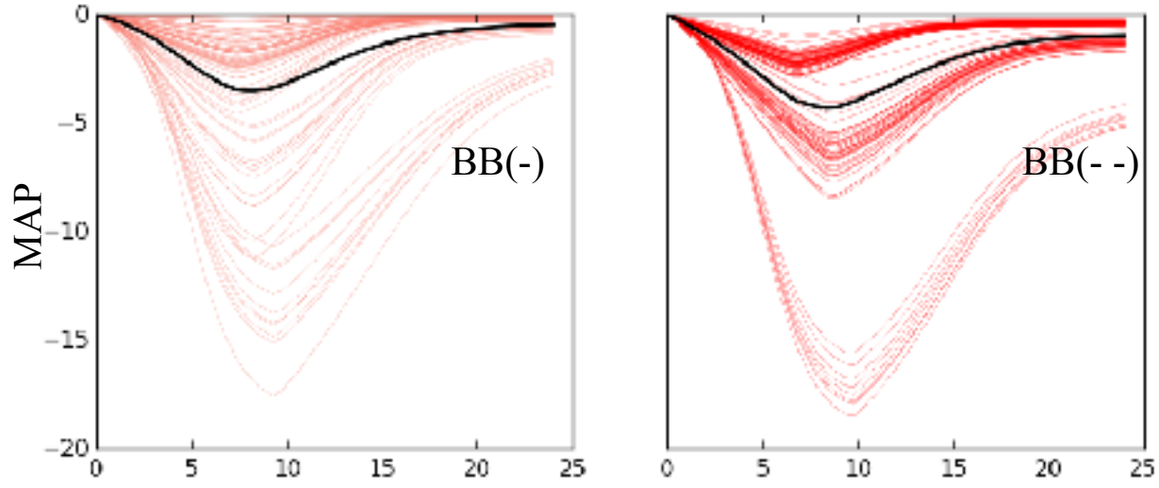
Bayesian Nonparametric for Estimation of Heterogeneous Treatment Response

Data: EHR collected over two years at Howard County General Hospital from 2013-2015. 300 ICU patients who were prescribed at least one of the treatments.

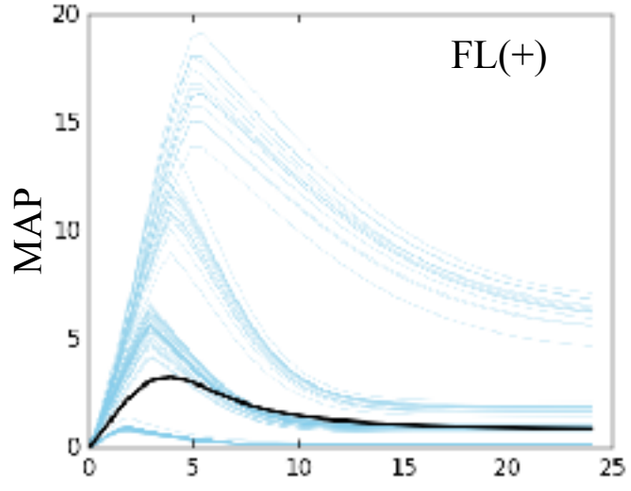
Vasopressor:



Beta-blocker:



Fluid_bolus:



Xu et al., 2016

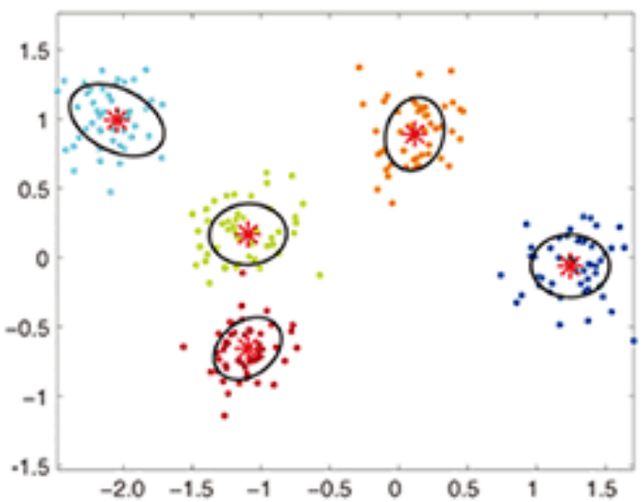
Liu, Henry et al., 2017

Bayesian Nonparametric for Estimation of Heterogeneous Treatment Response

$$y(t) = \underbrace{f^*(t)}_{\text{baseline progression (GP)}} + \underbrace{g^*(t;a)}_{\text{treatment response}} + \underbrace{\epsilon}_{\text{noise}}$$

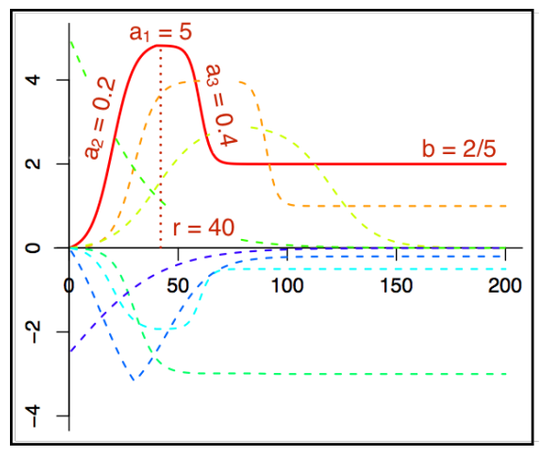
Gaussian Process to flexible model longitudinal traces

Dirichlet Process mixture prior to cluster treatment response and baseline progression parameters



- Each individual samples its parameters from a cluster mean
- No bias due to assuming that clusters are of equal size or a fixed number of clusters
- Posterior Predictive: Estimates refined with new data

Ferguson, 1973



Xu et al., 2016

Liu, Henry et al., 2017

Conclusion

- (1) Naive application of predictive models may lead to models that are counterintuitive and violate construct validity

Conclusion

- (1) Naive application of predictive models may lead to models that are counterintuitive and violate construct validity
- (2) Models for Counterfactual Reasoning from Observational Traces
- (3) Use understanding of mechanism to drive model development.
  - (4) Wrote down set of assumptions; approach mimics running a trial on this patient, *assuming all assumptions are satisfied.*

Conclusion

- (1) Naive application of predictive models may lead to models that are counterintuitive and violate construct validity
- (2) Models for Counterfactual Reasoning from Observational Traces
- (3) Use understanding of mechanism to drive model development.
  - (4) Wrote down set of assumptions; approach mimics running a trial on this patient, *assuming all assumptions are satisfied.*

- Open challenges:
  - A rigorous framework for when to trust the model: checking sensitivity to assumptions?
  - Flexible and richer models that more fully embrace the complexity of EHR data
  - Assumed missing at random in the talk today; extend to missing not at random
  - More easily incorporate known mechanisms into model building

Thank you!
ssaria@cs.jhu.edu
www.suchisaria.com
@suchisaria

hsoleimani@jhu.edu
www.hosseinsoleimani.com

All references throughout the slides are active links and clickable.