Machine Learning Foundations & Methods for Precision (Medicine and Healthcare)

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

Peter Schulam
PhD Candidate
Computer Science

Introduction

A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery

Suchi Saria, *Johns Hopkins University*

Healthcare spending in the US is nearing $3 trillion per year, but in spite of this expenditure, the US is outpaced by most developed countries in terms of health and quality of life outcomes—for example, it ranks 36th internationally in life expectancy.[1] The share of health spending in its gross domestic product has increased sharply, from 5 percent of GDP in 1960 to more than 17 percent today,[2] a rate of increase that's widely believed to be unsustainable.[3]

Policy and regulatory reform have important roles to play in addressing these challenges. Yet one of the largest underexplored avenues is the better use of information derived from the vast amount of health data now being collected in digital format.[4] I believe that one of the most significant open fron-

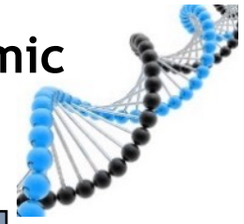
paper records that weren't amenable to retrospective, automated analyses. The Health Information Technology for Economic and Clinical Health (HITECH) Act, a program that was part of the American Recovery and Reinvestment Act of 2009, incentivized the adoption of Electronic Health Records (EHRs) to encourage the shift from paper to digital records. That program has made more than $15.5 billion available to hospitals and healthcare professionals conditioned on their meeting certain EHR benchmarks for so-called "meaningful use." It's one of the largest investments in healthcare infrastructure ever made by the federal government.

A survey by the American Hospital Association showed that adoption of EHRs has doubled from 2009 to 2011. Today, much of an individual's health data—demograph

Saria, IEEE Intelligent Systems, 2014

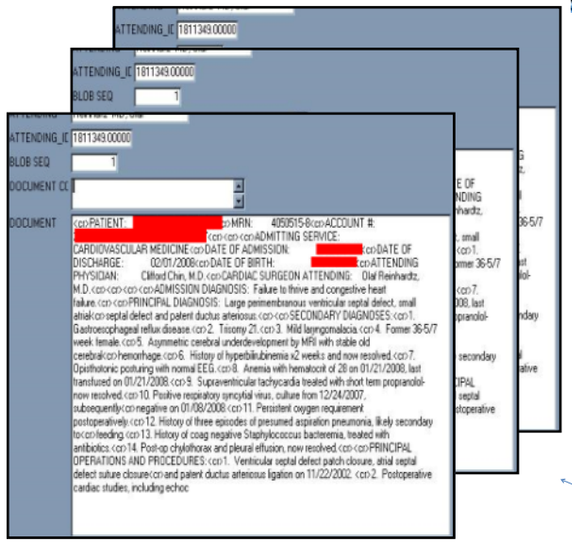
Electronic Health Records

Genomic data



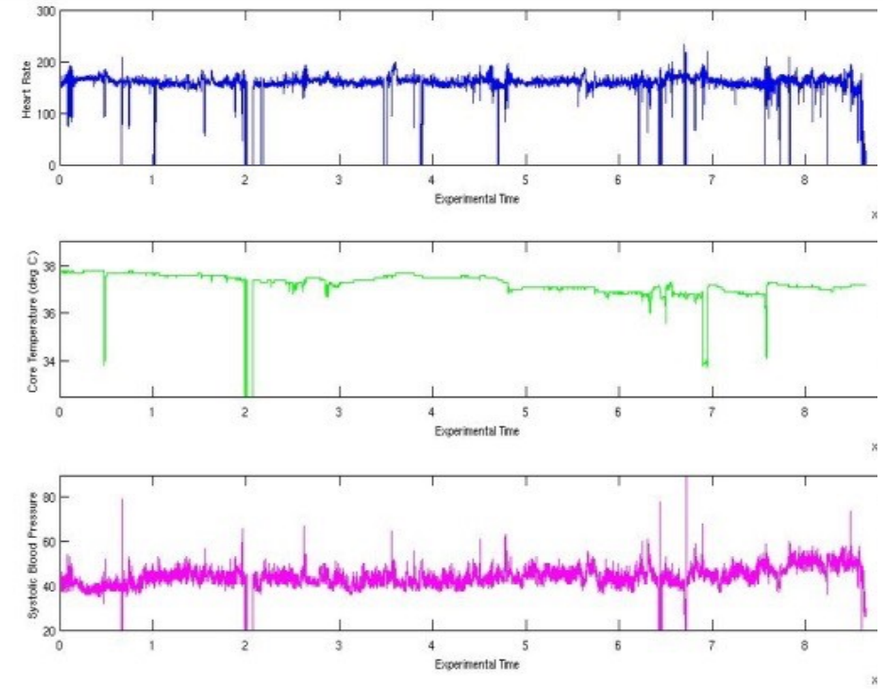
Sensors & Devices

Administrative Claims

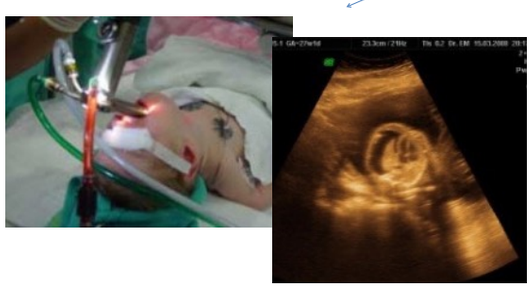


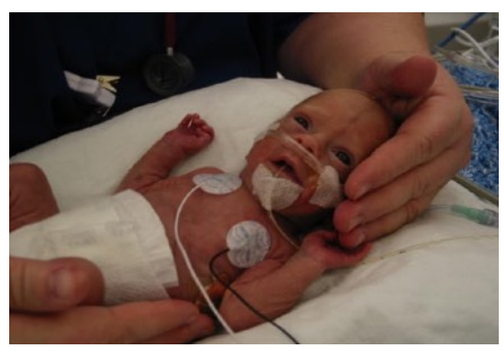
Progress notes

Continuous physiologic measurements

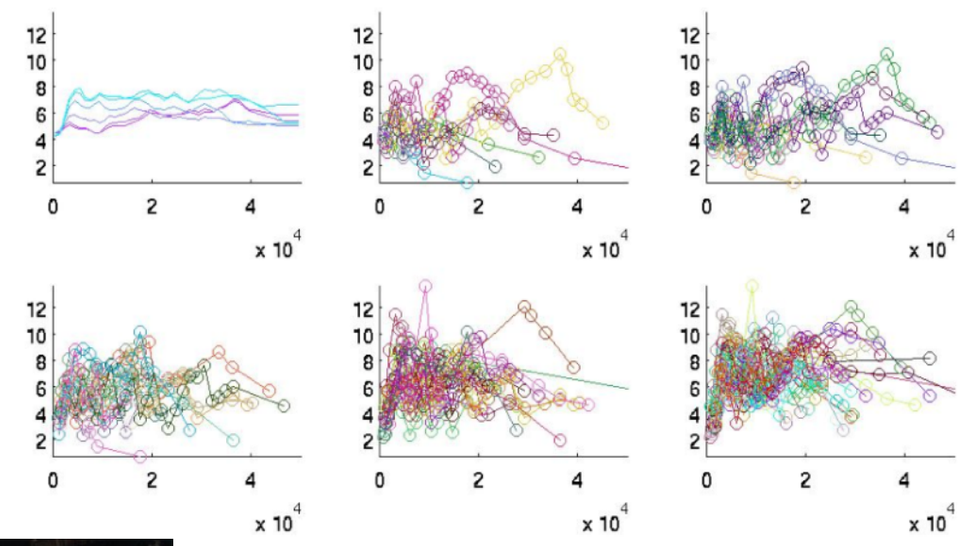


Imaging Data

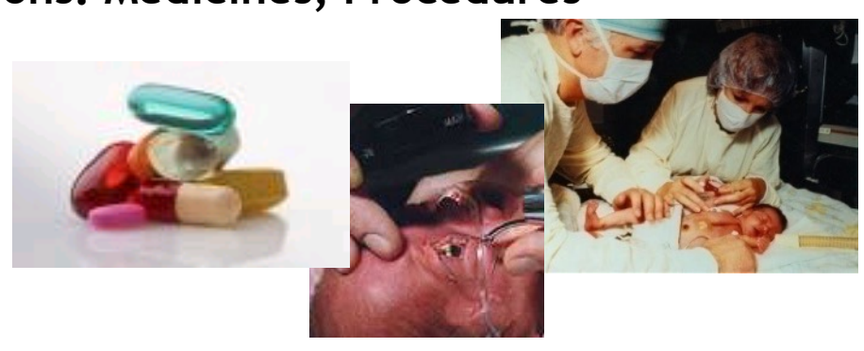




Discrete Events: Laboratory



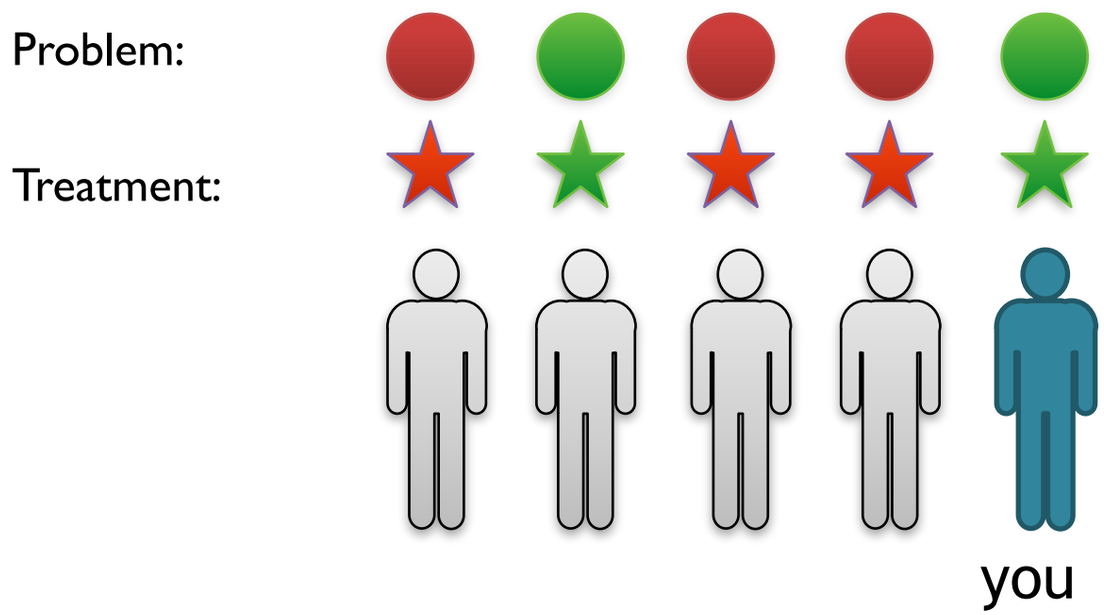
Interventions: Medicines, Procedures



Scope

- Focus of this talk is on **Precision and Personalized Medicine**
- Intended audience: Machine learners
  - Relevant to anyone with interest in personalization
  - Domains: education, recommender systems, retail

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*

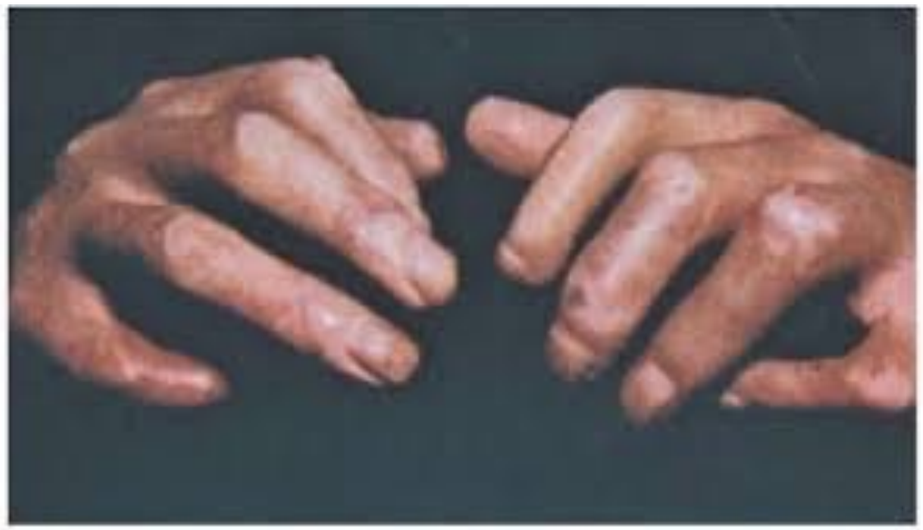
- **Example**: managing high blood pressure in adults

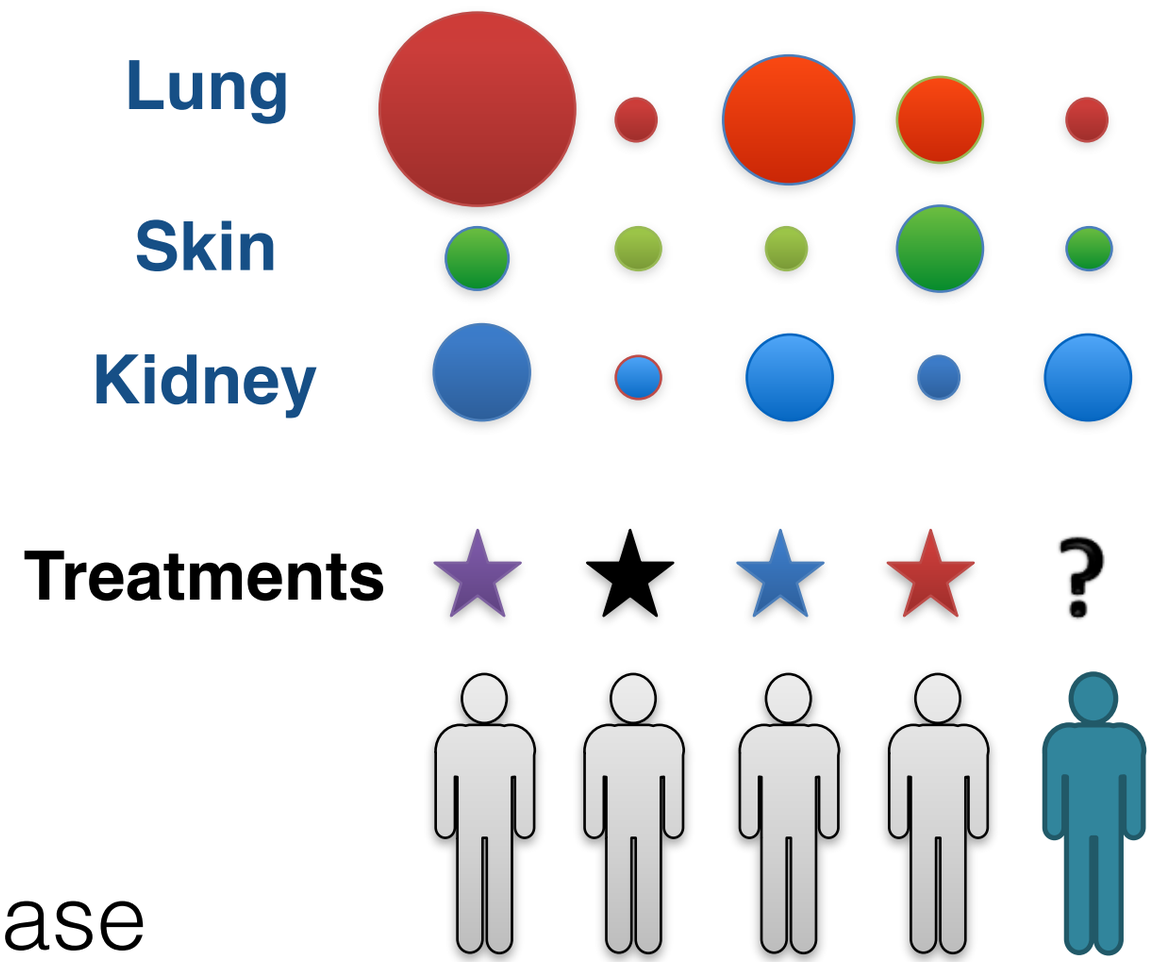
James, Oparil, Carter, et al. 2014

- "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





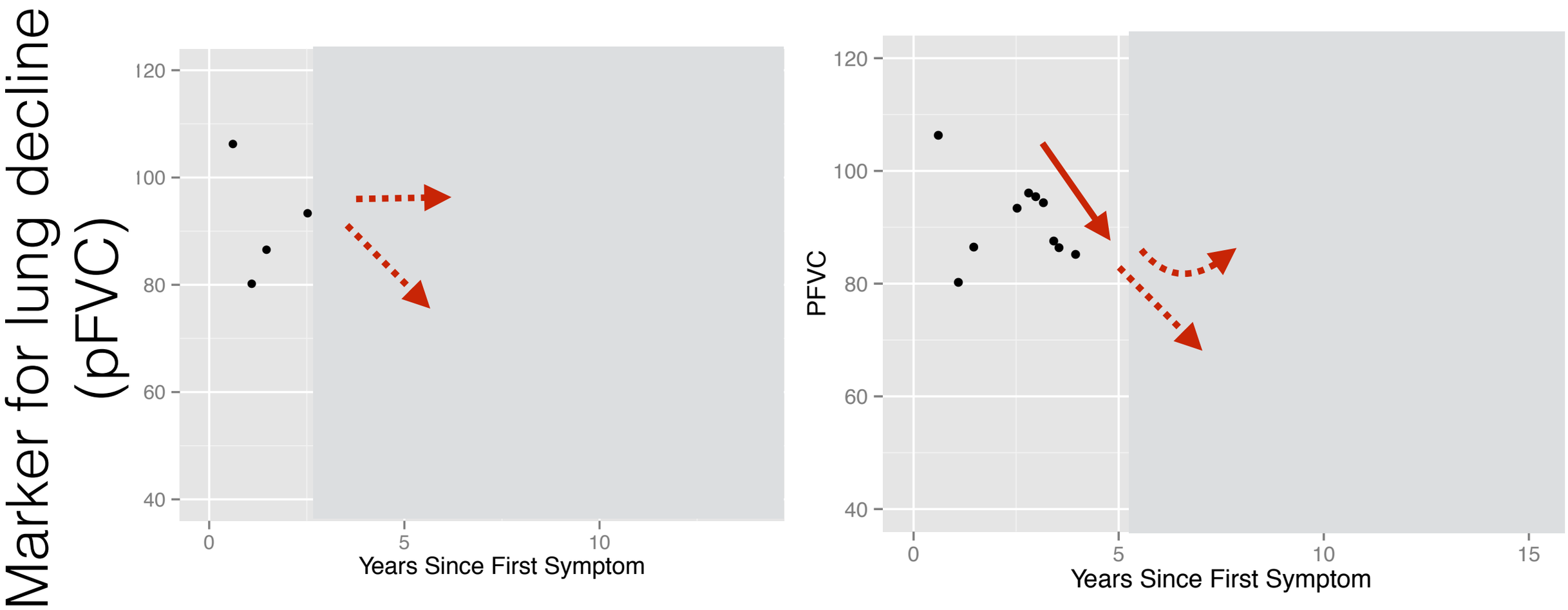
- Systemic autoimmune disease

  Main symptom: **skin** fibrosis

  Affects **many visceral organs**—lungs, heart, GI tract, kidney, vasculature, and muscles

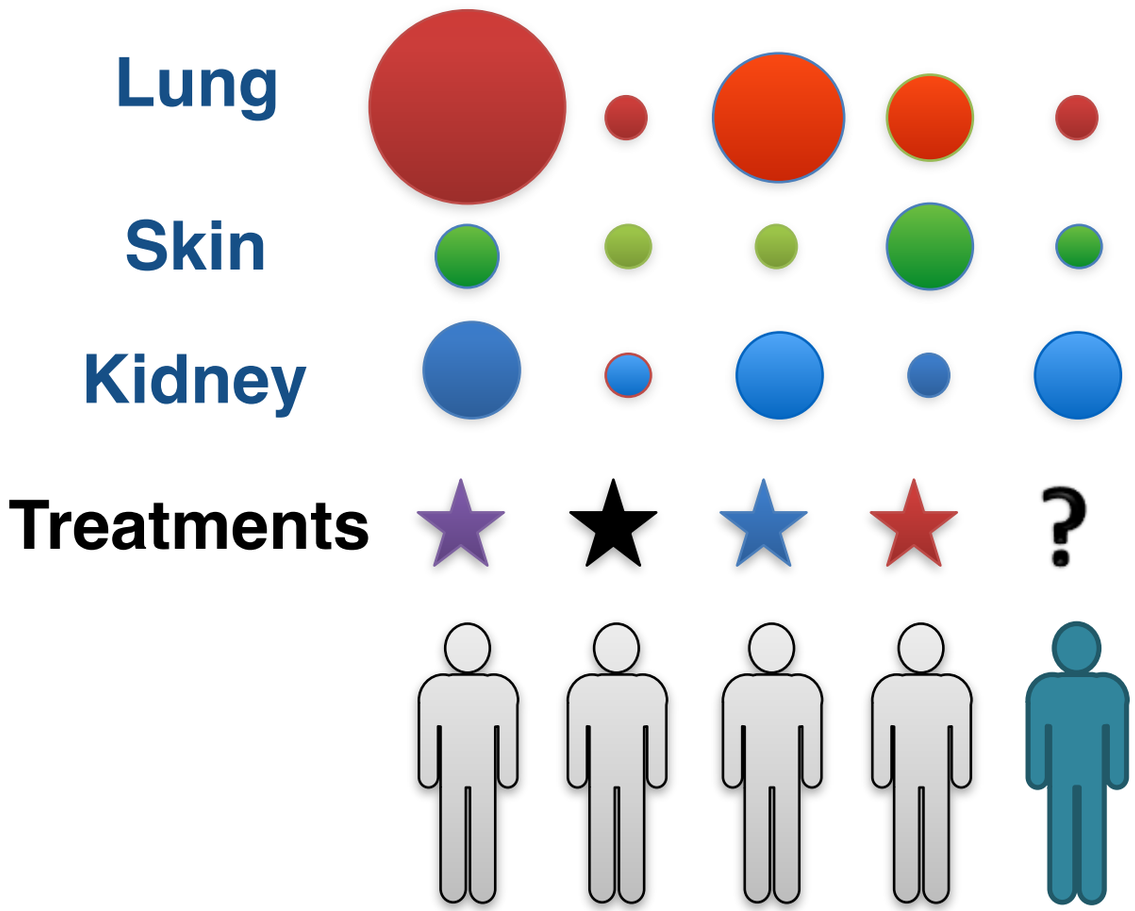
Affects 300K individuals; 80 other autoimmune diseases — lupus, multiple sclerosis, diabetes, Crohn's — many of which are systemic & highly multiphenotypic.

Targeted Treatment Plans



- Will this individual continue to decline?
- Should we **administer immunosuppressants, which can be toxic**?

The need for “precision”/“personalization”



Individual organs affected to varying extents (size) in a variety of ways (color).

Sources of variation:

- The profile of symptoms over time can vary
- Response to treatments can vary

(1) Characterize diseases more precisely? Is diabetes one disease or many diseases?
(2) Moving away from coarse rules to **algorithms for generating targeted treatment plans**.

Problem Setting

Sequential Data: No Control over Data Collection Process

- (1) Off-line learning:
  - Learn from data about other individuals to generalize to a given individual

- (2) Online learning:
  - Learn as we collect new data about a given individual from repeated measurements

Learning with Control over Data Collection

- (3) Reinforcement Learning:
  - Explore to improve model learning

Problem Setting

Sequential Data: No Control over Data Collection Process

- (1) Off-line learning:
  - Learn from data about other individuals to generalize to a given individual

- (2) Online learning:
  - Learn as we collect new data about a given individual from repeated measurements

Learning with Control over Data Collection

- (3) Reinforcement Learning:
  - Explore to improve model learning

Scope

- Focus of this talk is on **Precision and Personalized Medicine**
- Intended audience:
  - Machine Learners
    - Relevant to anyone with interest in personalization
- Key takeaways:
  - Provide computational strategies for personalization
    - Describe example data
    - Introduce concrete applications
    - Give intuition into why approach it one way vs another
  - Throughout make connections to literature in sub-areas of machine learning, reinforcement learning, causal inference, and informatics

Overview

- **Part 1—Setting up the problem of Individualization**
  - Example using a chronic disease
  - Simple setting: No Treatment Effects
  - **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Key ideas: Transfer learning, Multilevel modeling

- **Part 2—Estimating Treatment Effects & Individualized Treatment Effects**
  - Example using inpatient data
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP

- **Part 3—Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

No Control over Data Collection Process

- **Part 4—From Predictions to Treatment Rules**
  - Key ideas: Q-learning, Dynamic Treatment Regimes
  - Connections to Reinforcement Learning

Control over Data Collection Process

Overview

- **Part 1—Setting up the problem of Individualization**
  - Example using a chronic disease
  - Simple setting: No Treatment Effects
  - **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Key ideas: Transfer learning, Multilevel modeling

- **Part 2—Estimating Treatment Effects & Individualized Treatment Effects**
  - Example using inpatient data
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP

- **Part 3—Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

No Control over Data Collection Process

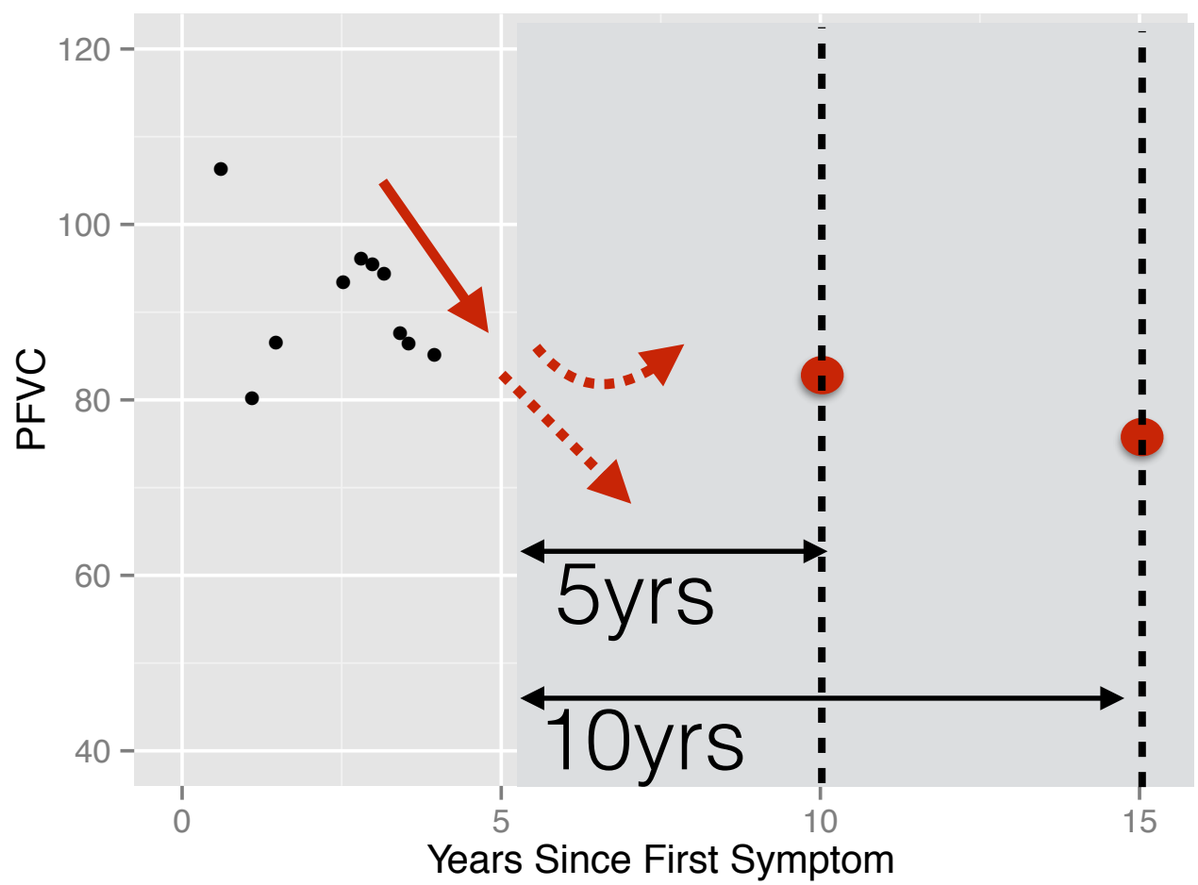
- **Part 4—From Predictions to Treatment Rules**
  - Key ideas: Q-learning, Dynamic Treatment Regimes
  - Connections to Reinforcement Learning

Control over Data Collection Process

Individualization and why do we need it?

Develop a predictive model by using regression on the observed risk factors

y = f (age, gender, baseline test values)

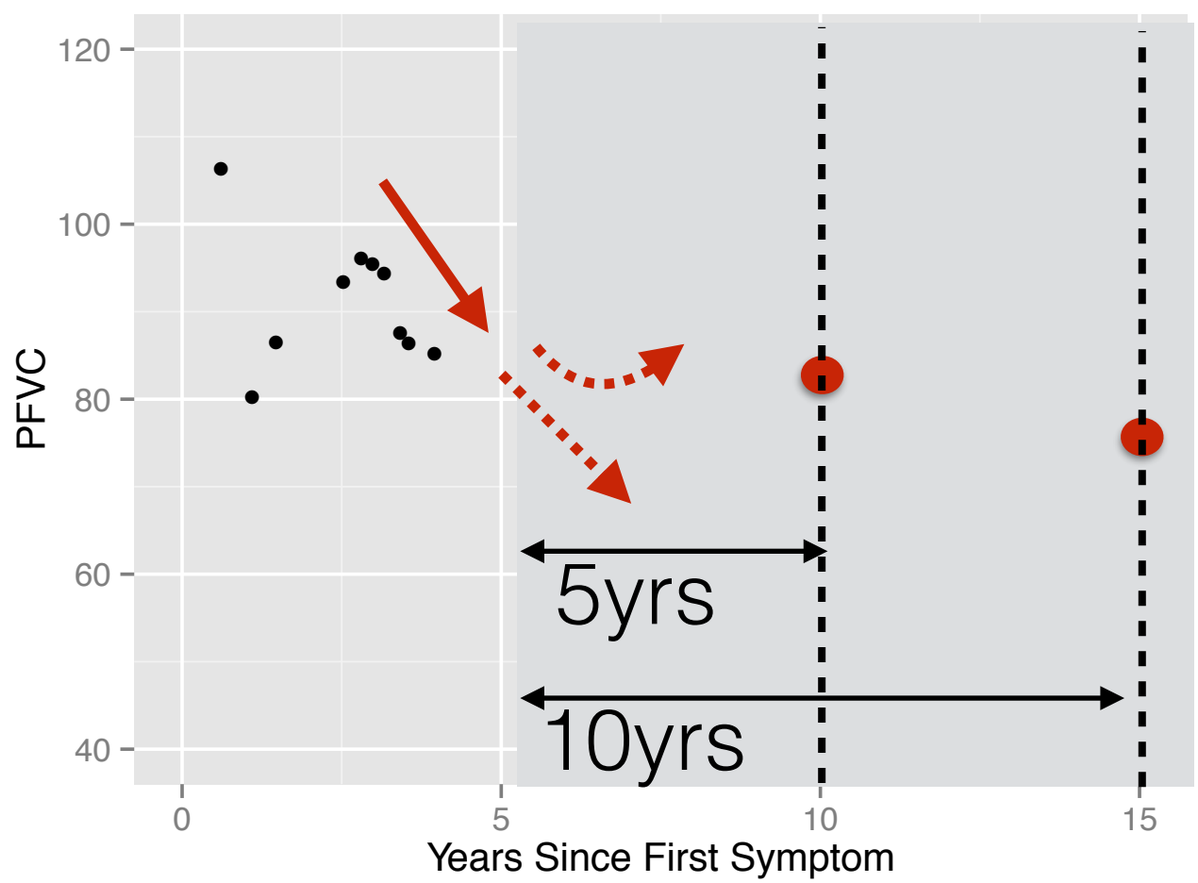


~~Population~~ Precision medicine

Develop a predictive model by using regression on the observed risk factors

y = f (age, gender, baseline test values, ….)

Expand the set of covariates to include high-dimensional molecular measurements



Shipp et al. 2002

Ziegler et al. 2012

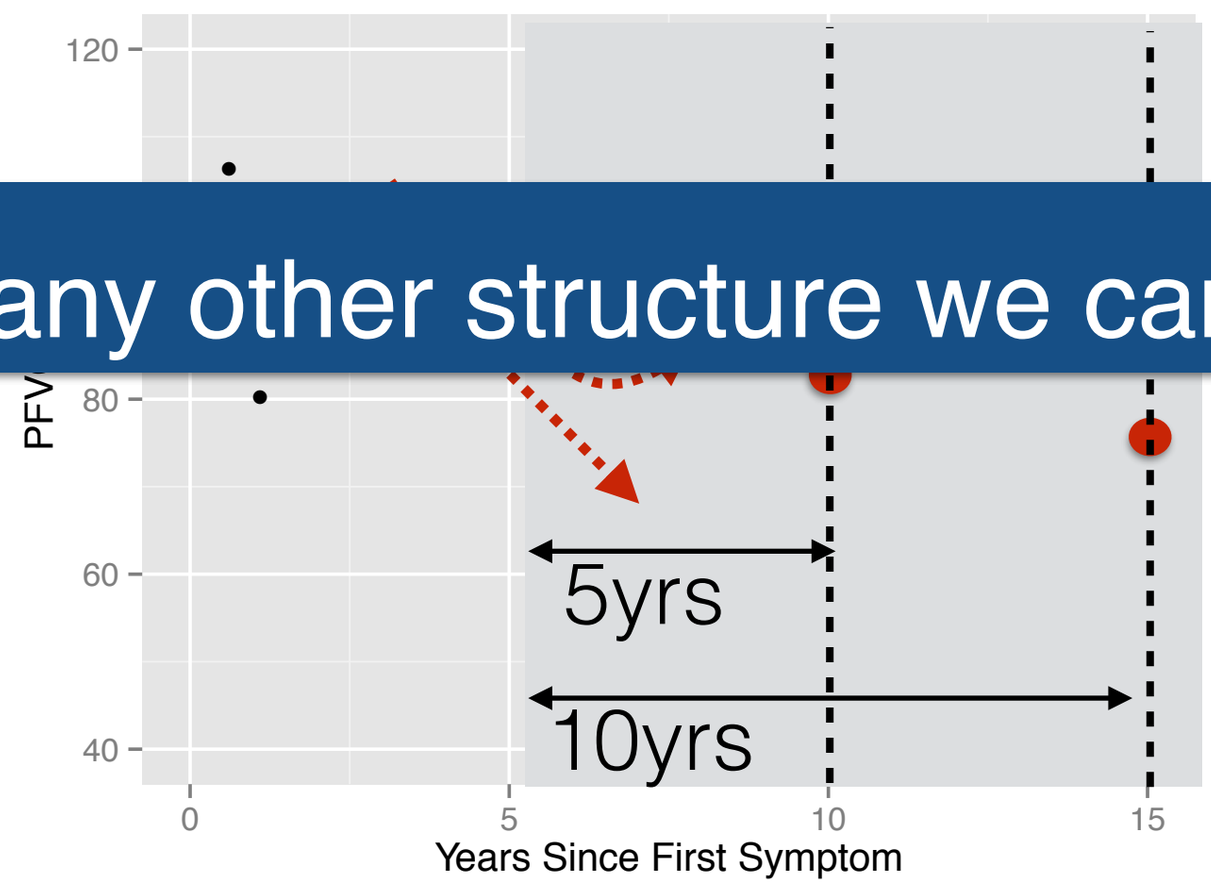
Collins and Varmus, 2015

~~Population~~ Precision medicine

Develop a predictive model by using regression on the observed risk factors

y = f (age, gender, baseline test values, ….)

Expand the set of covariates to include high-dimensional molecular measurements



Is there any other structure we can capture?

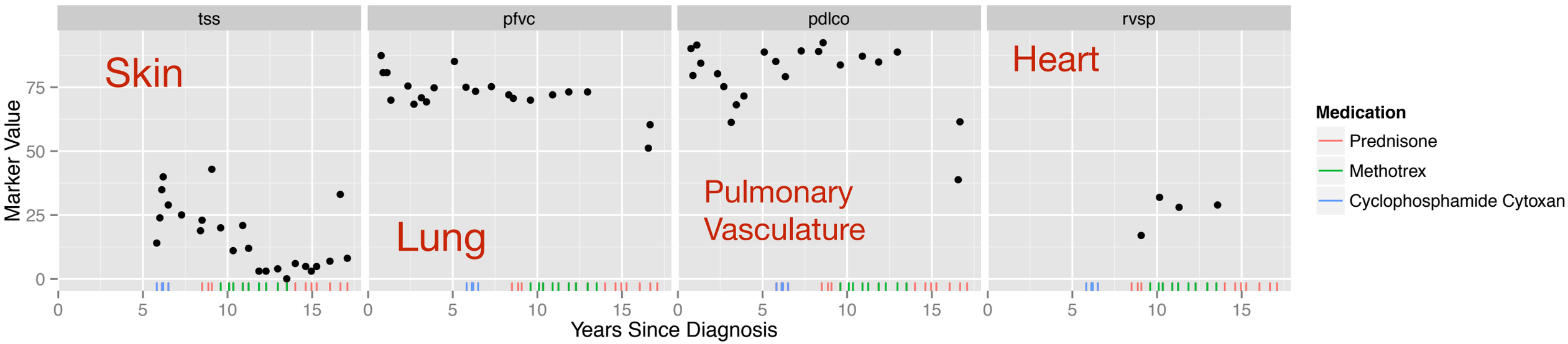
Shipp et al. 2002

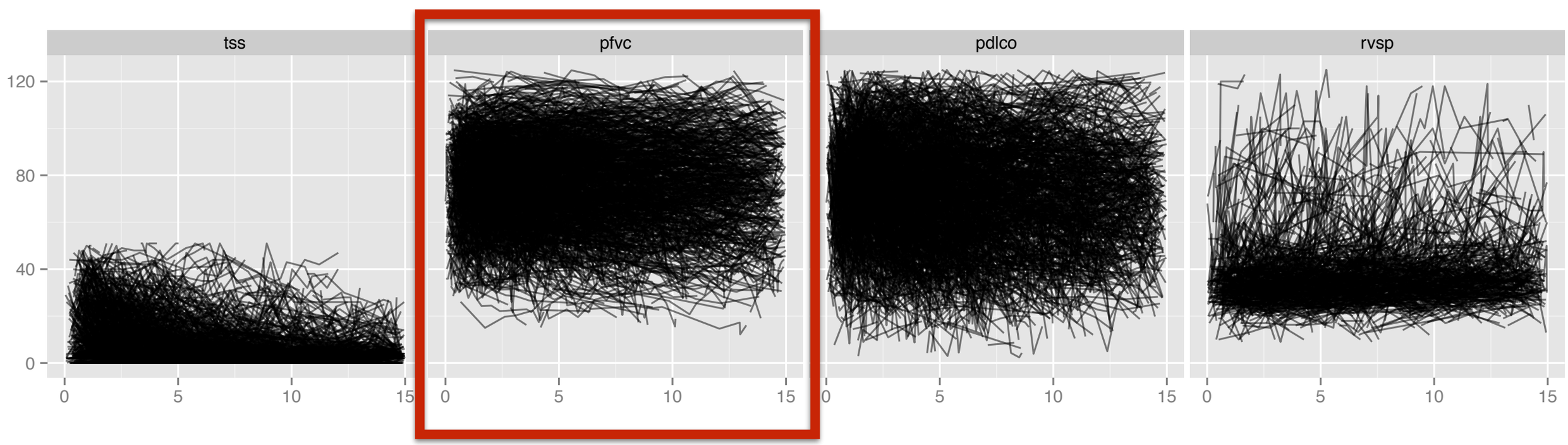
Ziegler et al. 2012

Collins and Varmus, 2015

Data & Problem Motivation

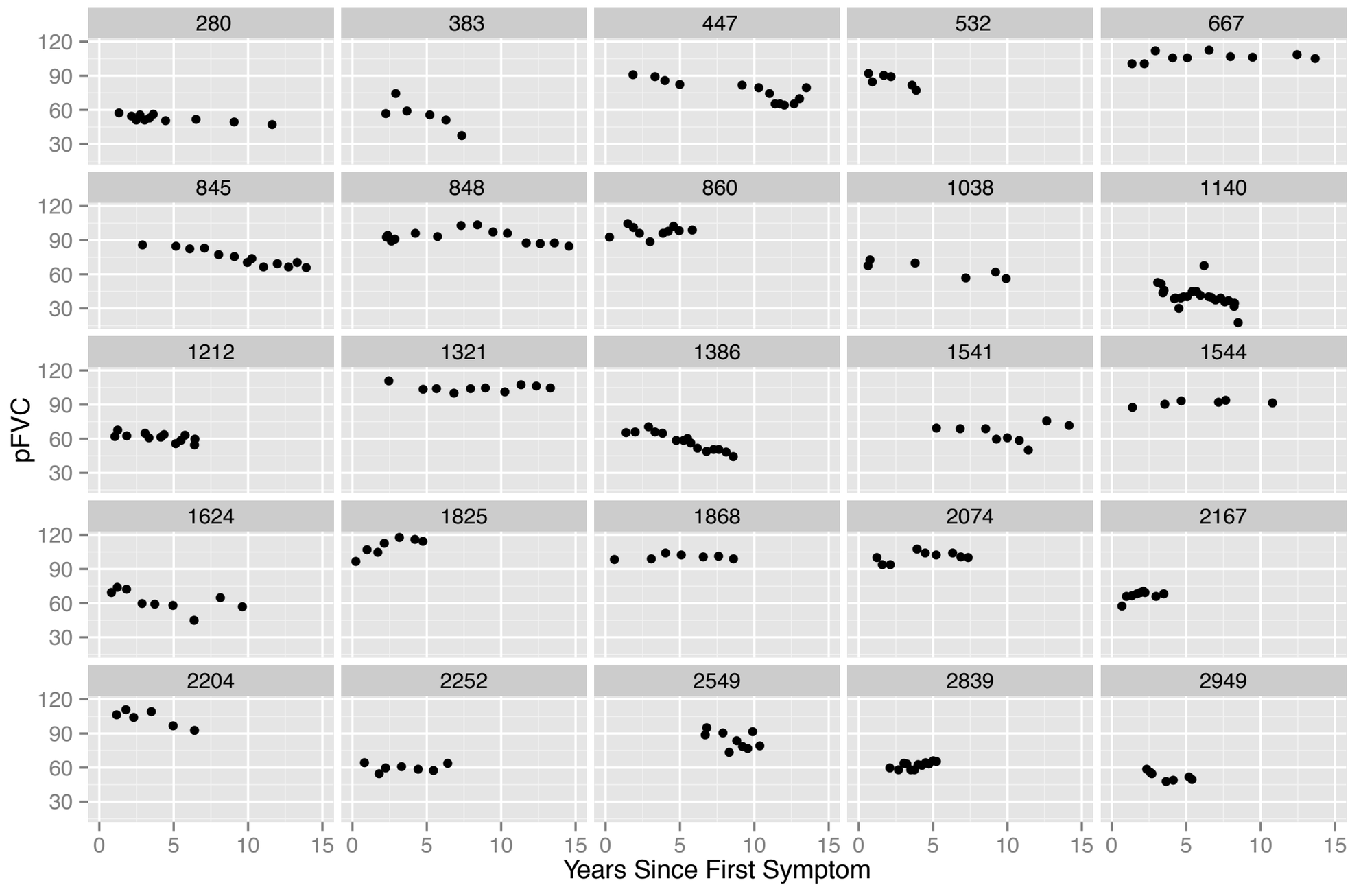
- **Functional markers collected to track organ health**



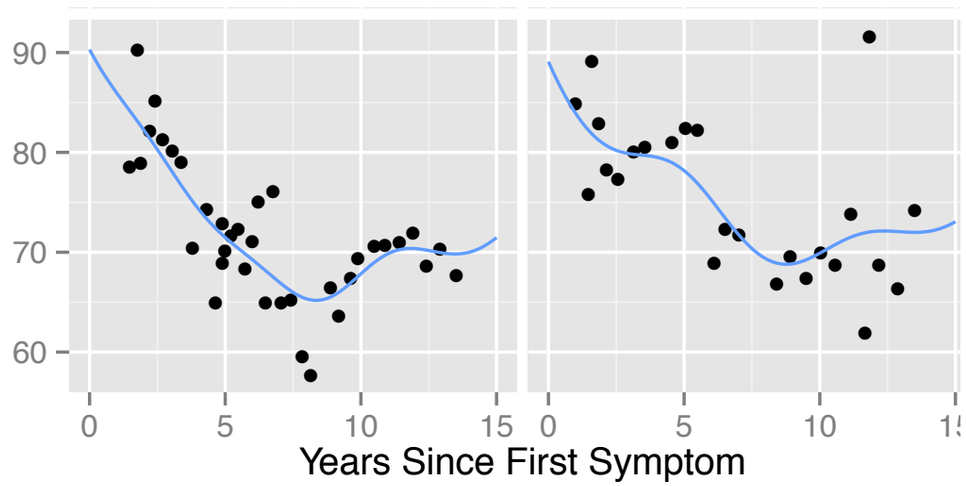


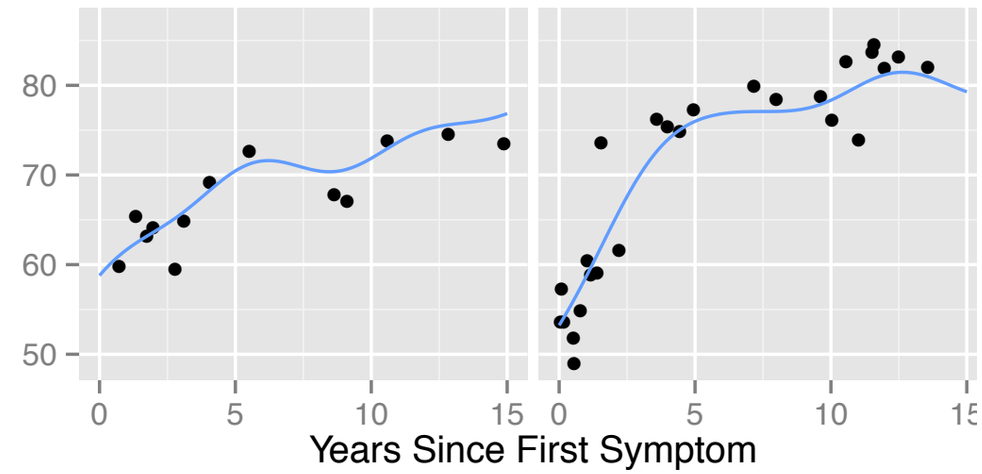
Data & Problem Motivation

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



Predicting Disease Trajectories

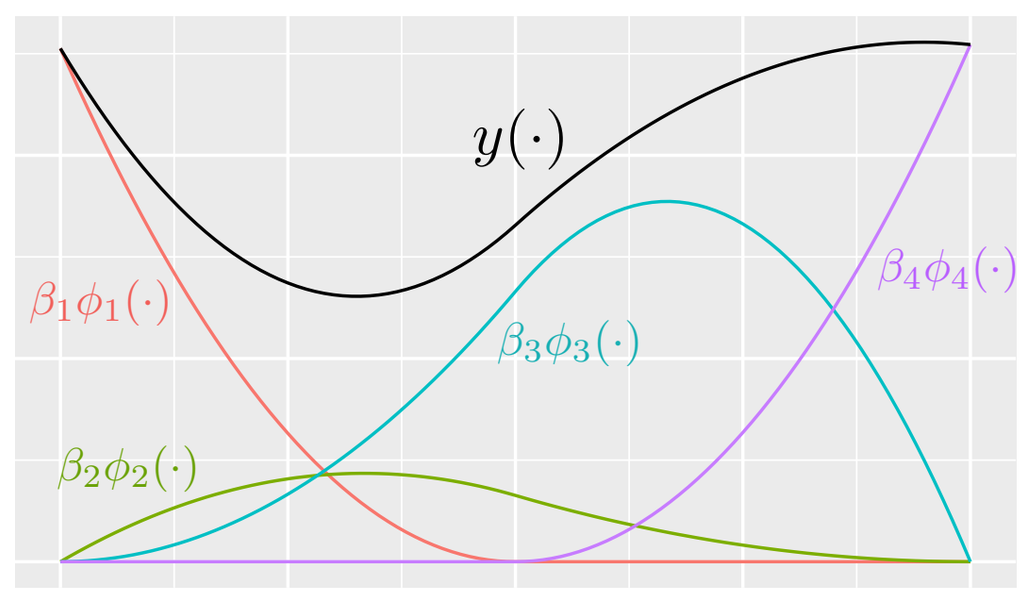




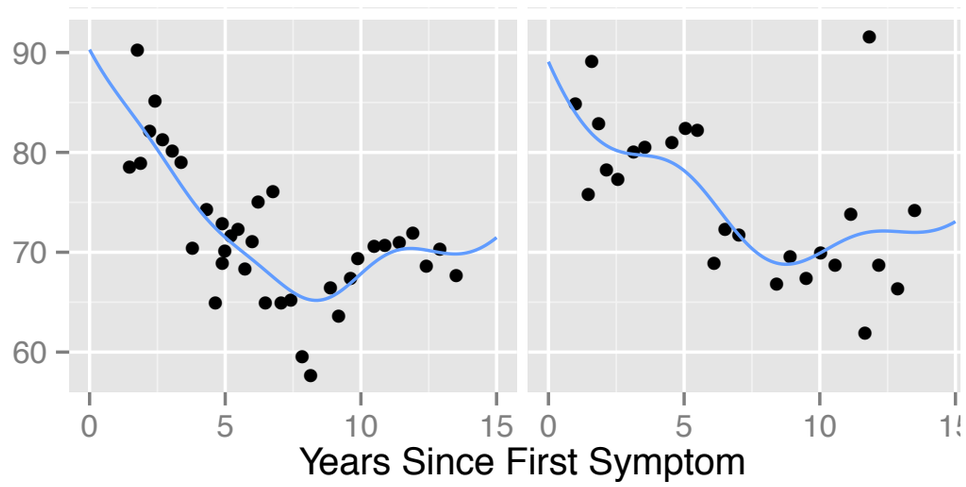
Function valued-regression

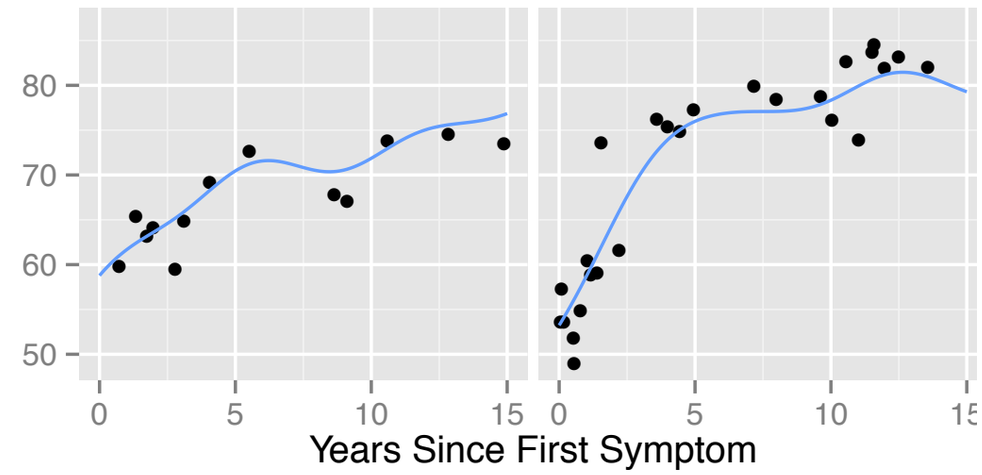
y = f (age, gender, baseline test values, $[\phi_1(t), \ldots, \phi_d(t)]$)

Expand the set of covariates to include non-linear functions of time



Predicting Disease Trajectories





Function valued-regression

y = f (age, gender, baseline test values, ….)

BUT

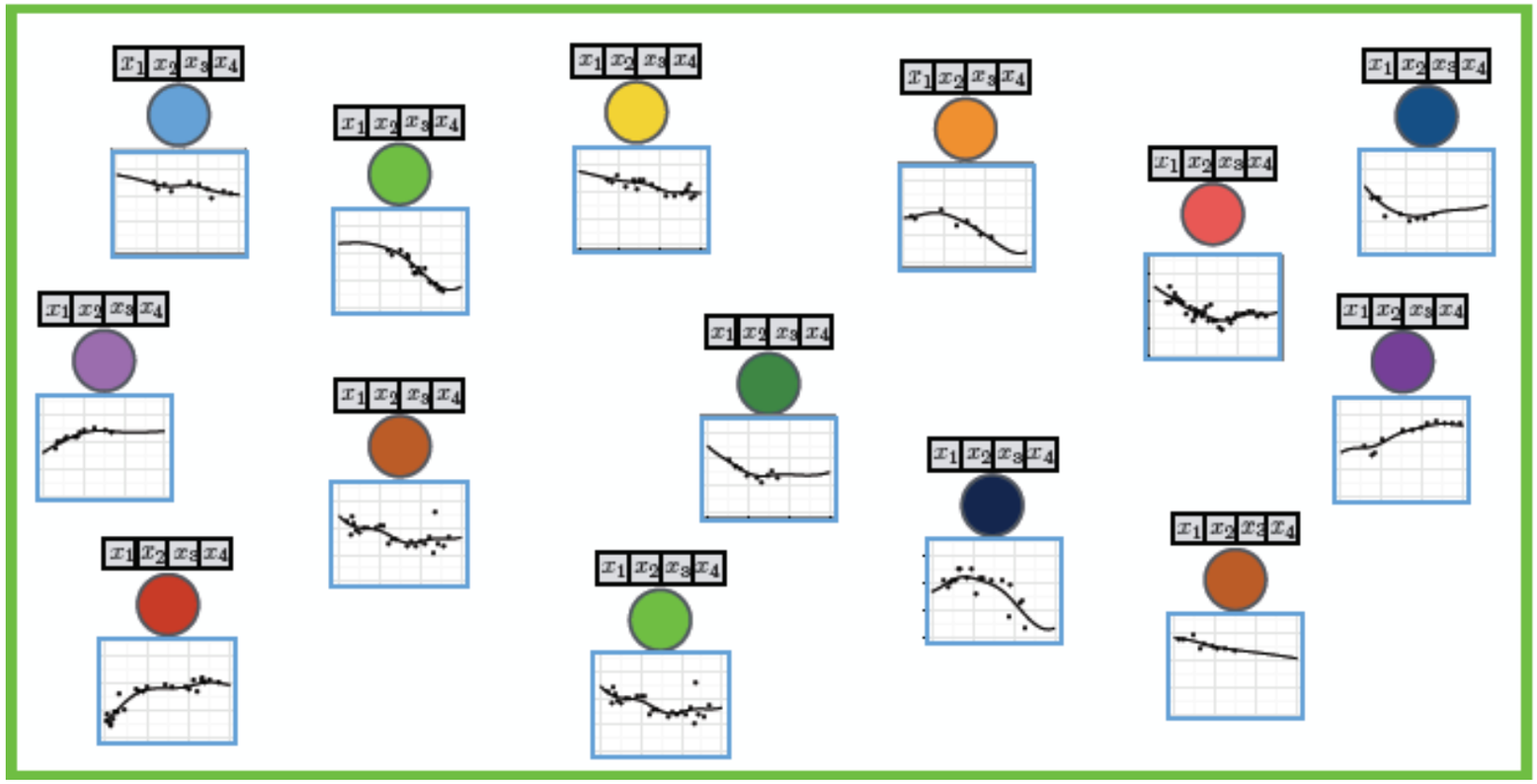
this assumes that sources of heterogeneity across individuals entirely explained away by observed factors.

Many factors leading to differences in trajectory may be unobserved (e.g., difference in genetic mutations, athleticism, lifestyle)

- **Account for heterogeneity in disease course due to both observed and latent factors**

Schulam and Saria, 2015

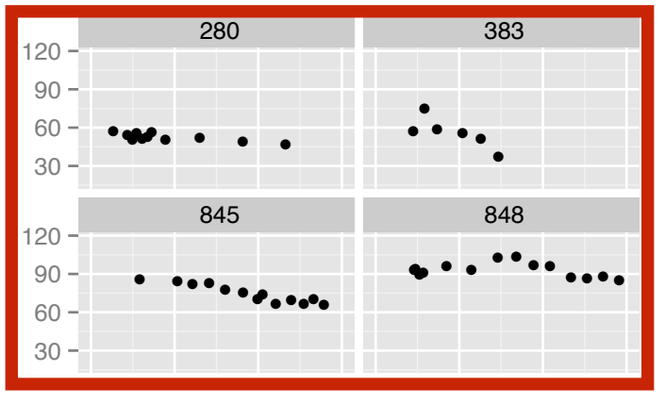
Transfer information from others to refine estimates for a given individual.



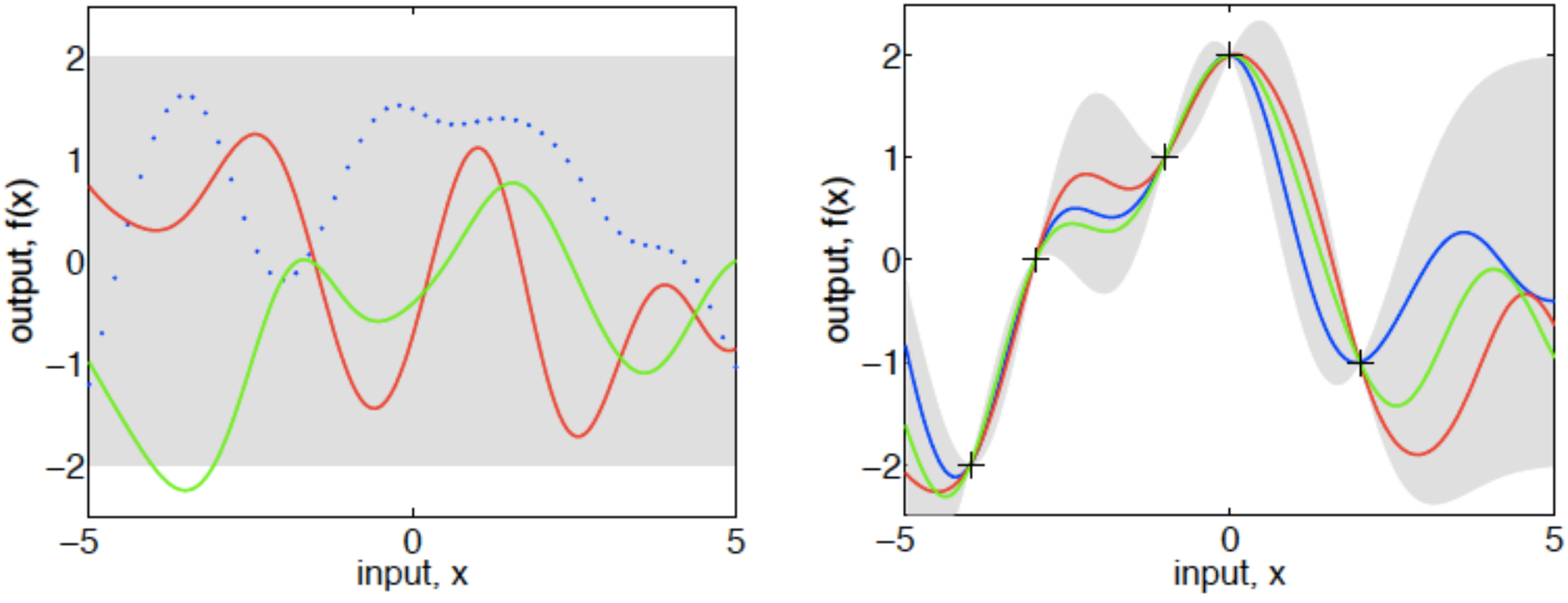
- #1 Specify Latent Variable Models to make inferences about latent (individual-specific) sources of heterogeneity
- #2 Learn the transfer hierarchy — i.e. whom to transfer from and what to transfer?
- #3 Bayesian formulation to prevent overfitting and learn as new data are collected on the individual

Background: Gaussian Processes

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



$$
\begin{aligned}
f(\mathbf{x}) &\sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \\
m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\
k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]
\end{aligned}
$$

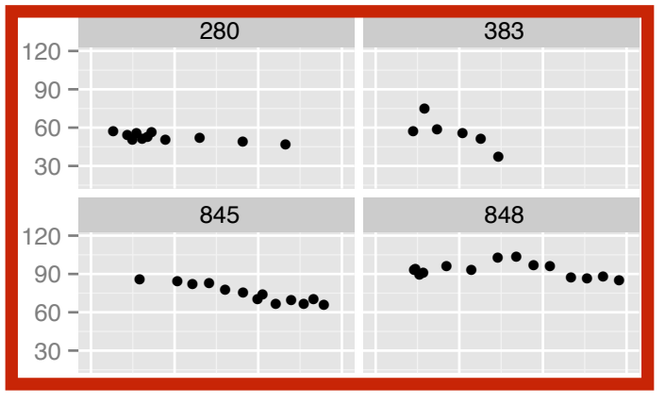


$$
\begin{aligned}
\mathbf{f}_* | X_*, X, \mathbf{f} \sim \mathcal{N}\big(&K(X_*, X)K(X, X)^{-1}\mathbf{f}, \\
&K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)\big)
\end{aligned}
$$

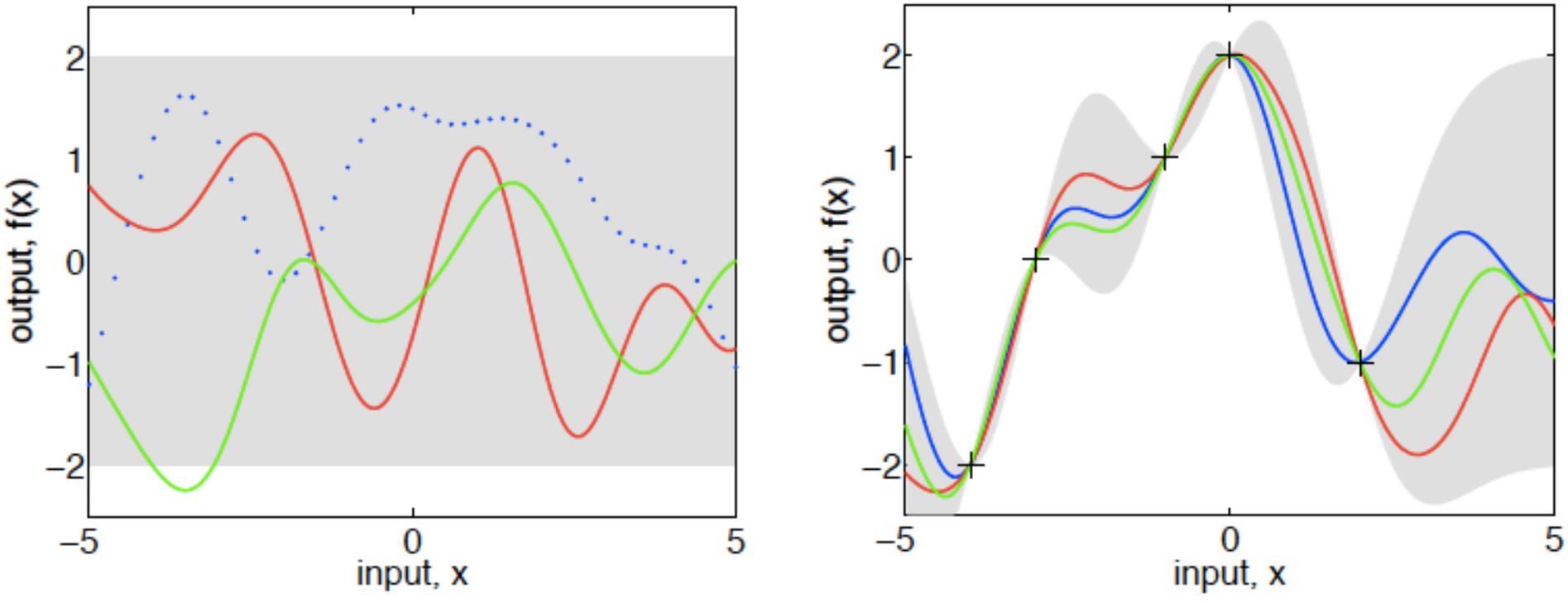
Rasmussen and Williams, 2006

Background: Gaussian Processes

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



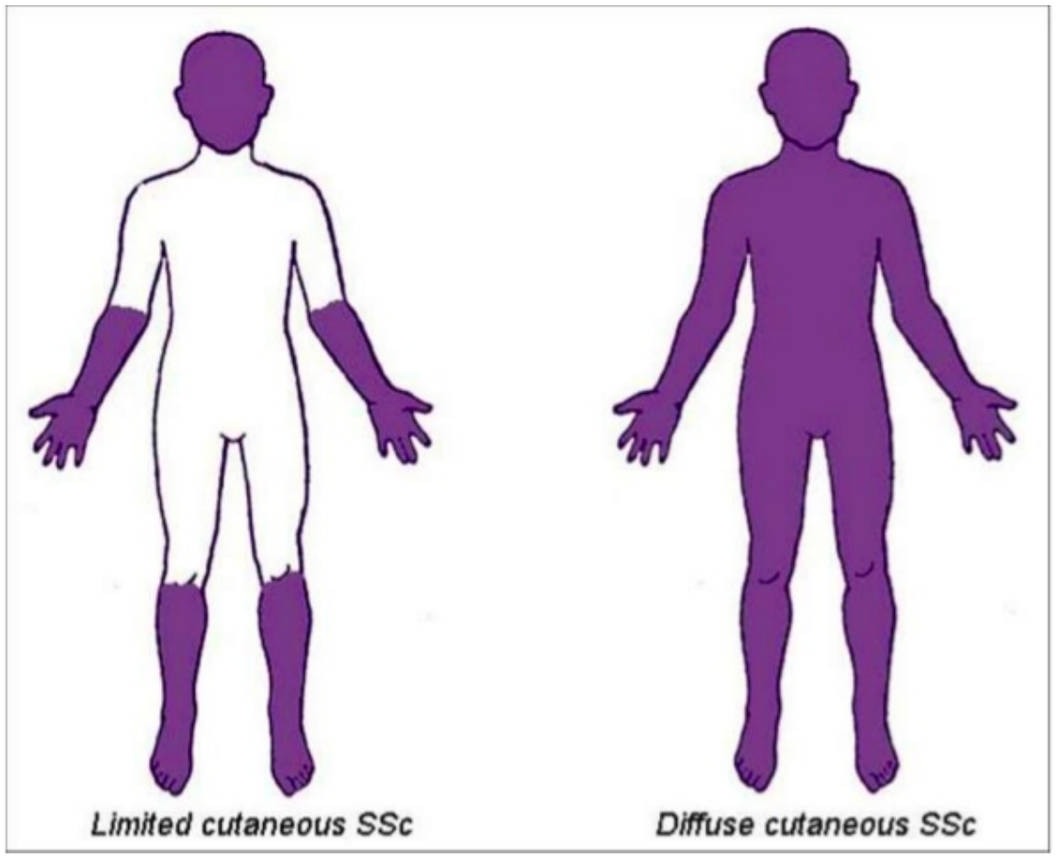
$$
\begin{aligned}
f(\mathbf{x}) &\sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \\
m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\
k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]
\end{aligned}
$$

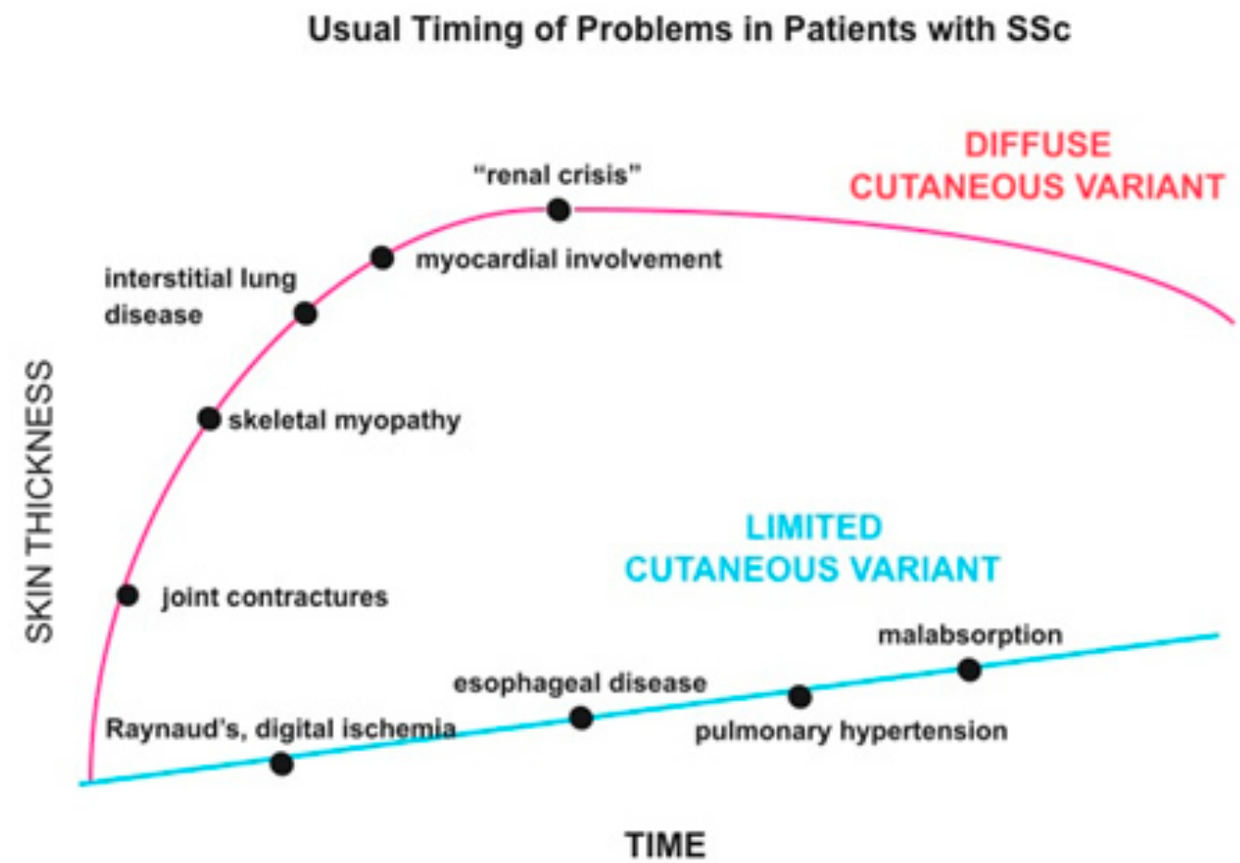


$$
\begin{aligned}
\mathbf{f}_* | X_*, X, \mathbf{f} \sim \mathcal{N}\big(&K(X_*, X)K(X, X)^{-1}\mathbf{f}, \\
&K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)\big)
\end{aligned}
$$

Rasmussen and Williams, 2006

Disease Subtypes and Latent Mechanism Driving Subtypes





. J. Varga, C.P. Denton, and F.M. Wigley. *Scleroderma: From Pathogenesis to Comprehensive Management*. Springer Science & Business Media, 2012.

http://www.hopkinsarthritis.org/wp-content/uploads/2011/04/image-11.jpg

http://www.slideshare.net/maushard/skin-manifestations-of-scleroderma-by-dr-lorinda-chung-md

---

Subtyping research in other diseases:

Autism: State and Sestan, 2012 | Doshi-Velez et al., 2014 **Parkinson's:** Lewis et al. 2005

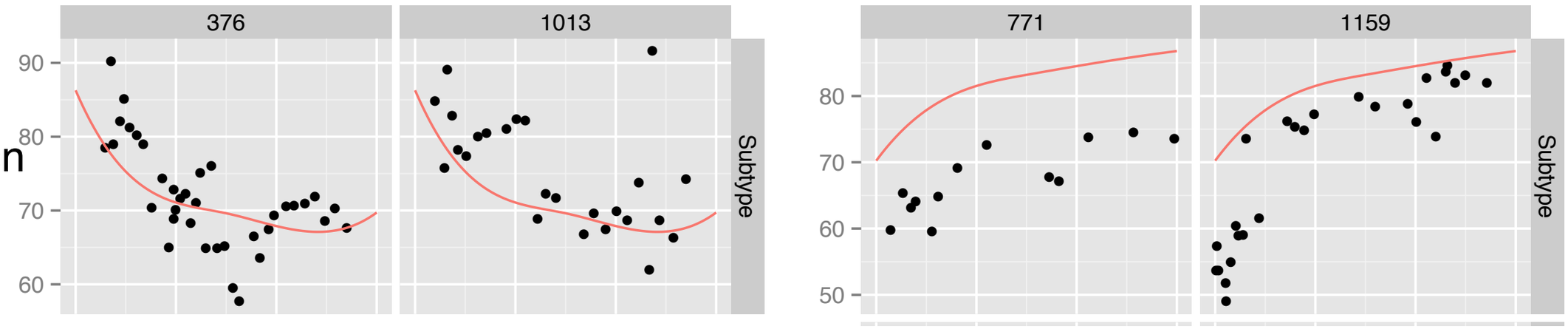
Cardiovascular disease: De Keulenaer and Brutsaert, 2009 **Asthma:** Anderson 2008

Latent Subpopulation Structure

Subtype 1

Subtype 2

Subpopulation



Two different subpopulations:
Subtype 1: Decliners who stabilize
Subtype 2: Those who improve over time

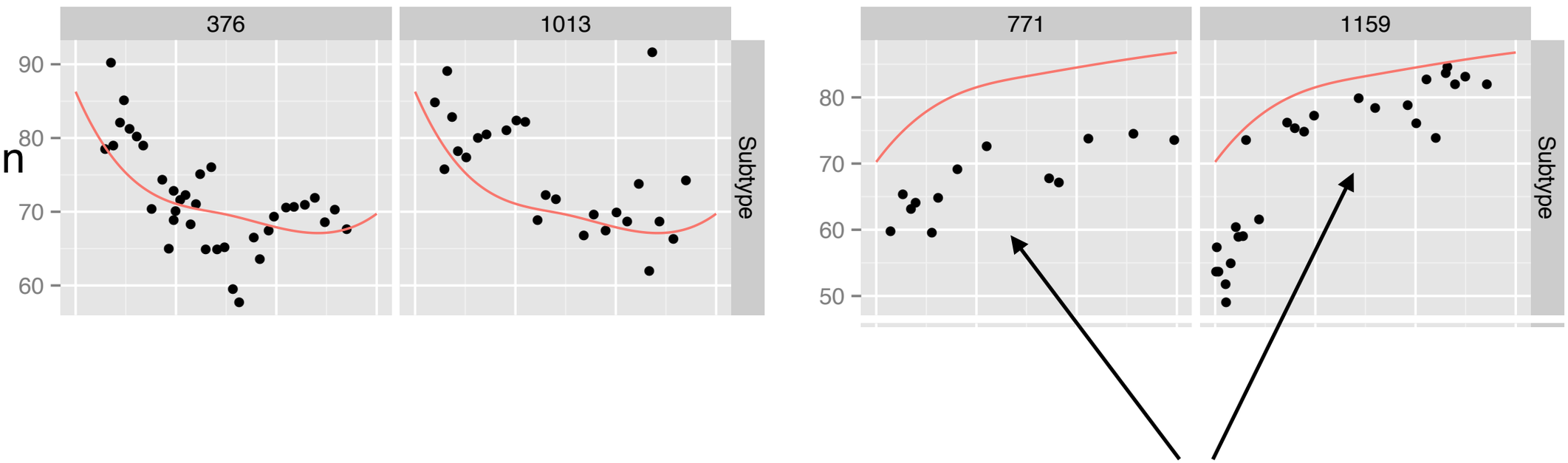
- Can we learn or make inferences about this systematic deviation as we observe more data about this individual?

Latent Individual-specific Structure

Subtype 1

Subtype 2

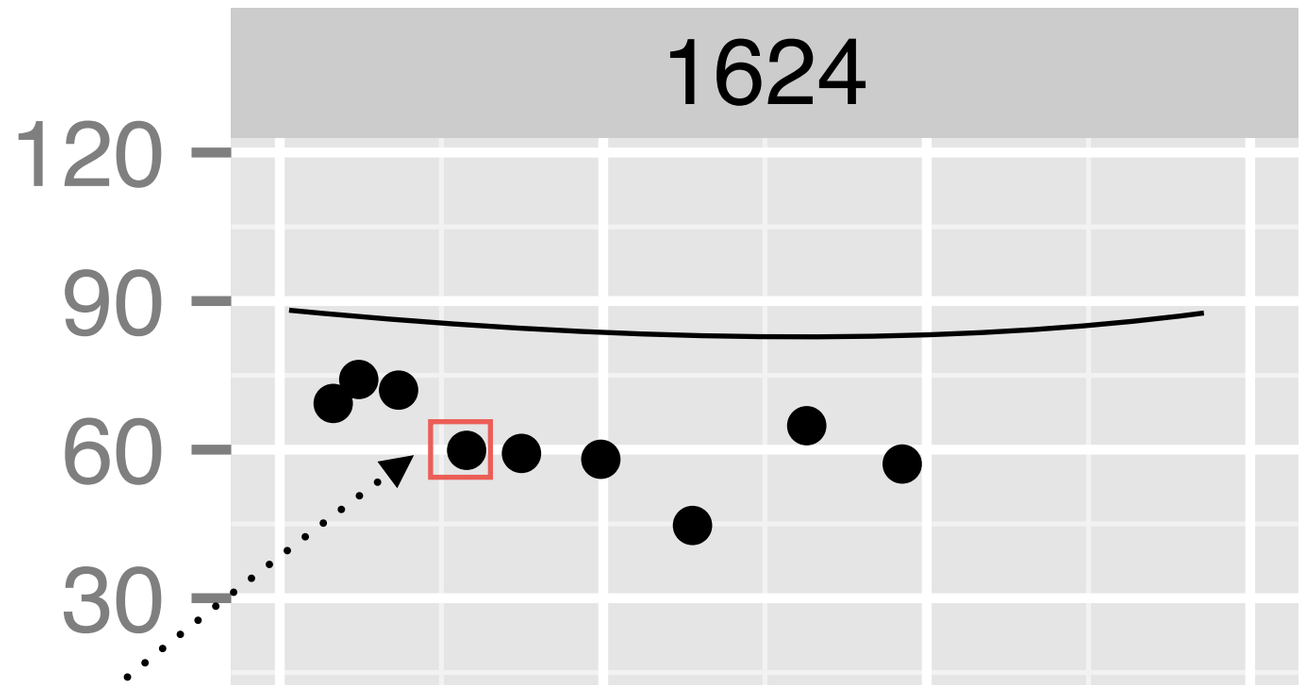
Subpopulation



Individuals vary within subgroups

- Can we learn or make inferences about this systematic deviation as we observe more data about this individual?

Bayesian Formulation for Disease Trajectories



Population trajectory

Function valued-regression
f (age, gender, baseline test values, ….)

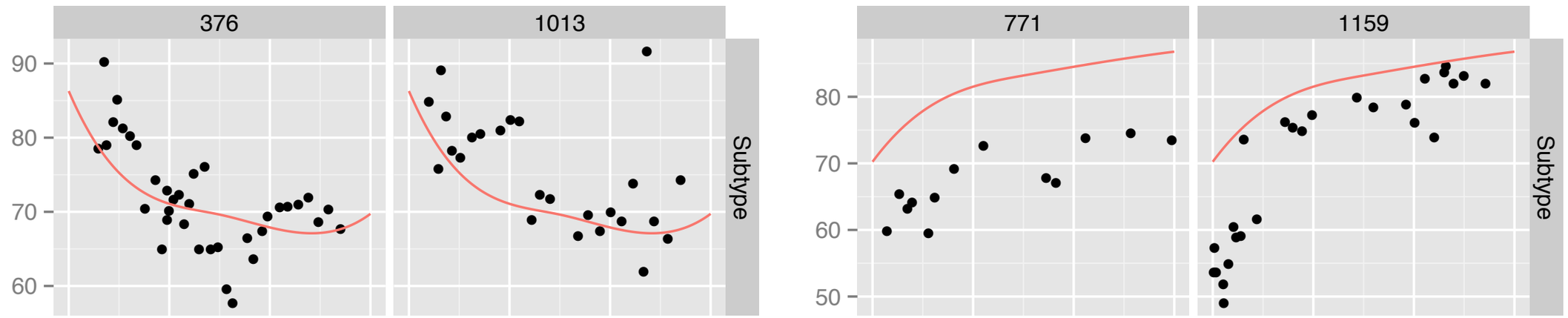
$$\boxed{y_{ij}} \mid z_i, \vec{b}_i, f_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda \; \vec{x}_i}_{\text{(A) population}} \; , \sigma^2 \right)$$

Ramsay and Silverman 2005

Accounting for Latent Sources of Heterogeneity

Subtype 1 | Subtype 2

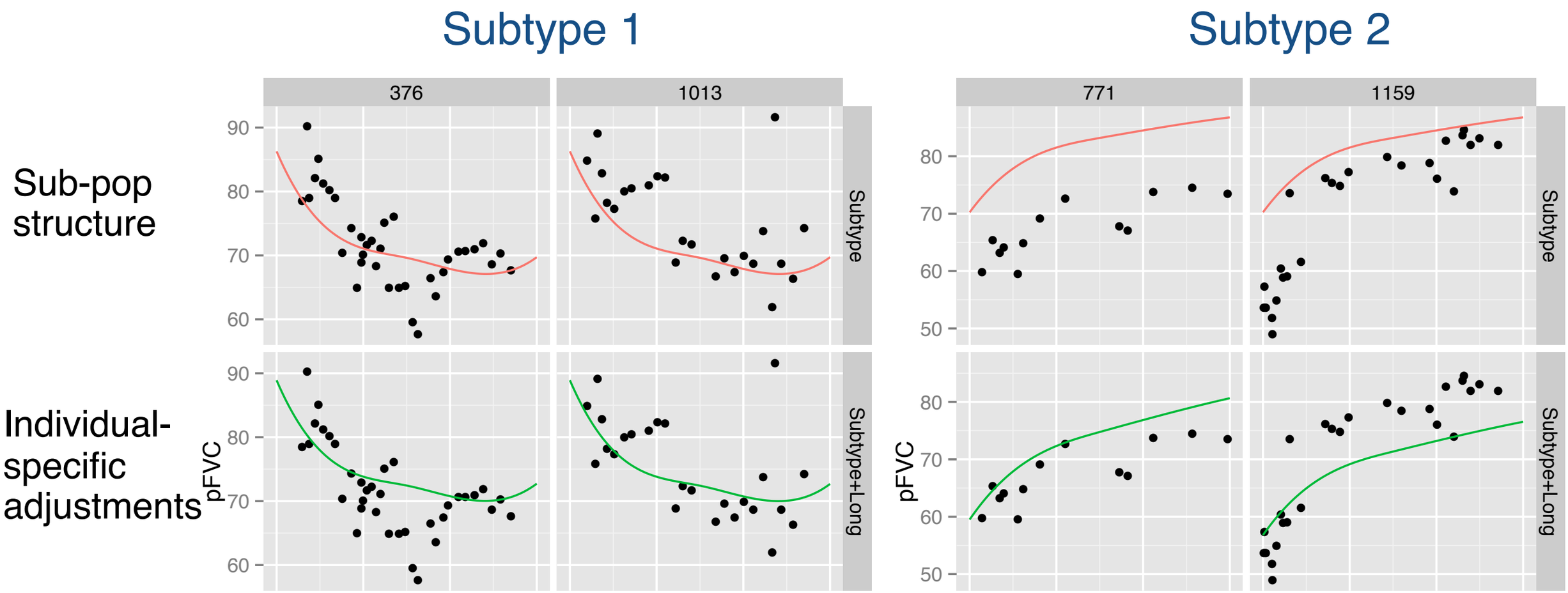
Sub-pop structure



z_i indexes a given subpopulation

$$\boxed{y_{ij}} \mid z_i, \vec{b}_i, f_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\ \vec{x}_i}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} \cdot \qquad , \sigma^2 \right)$$

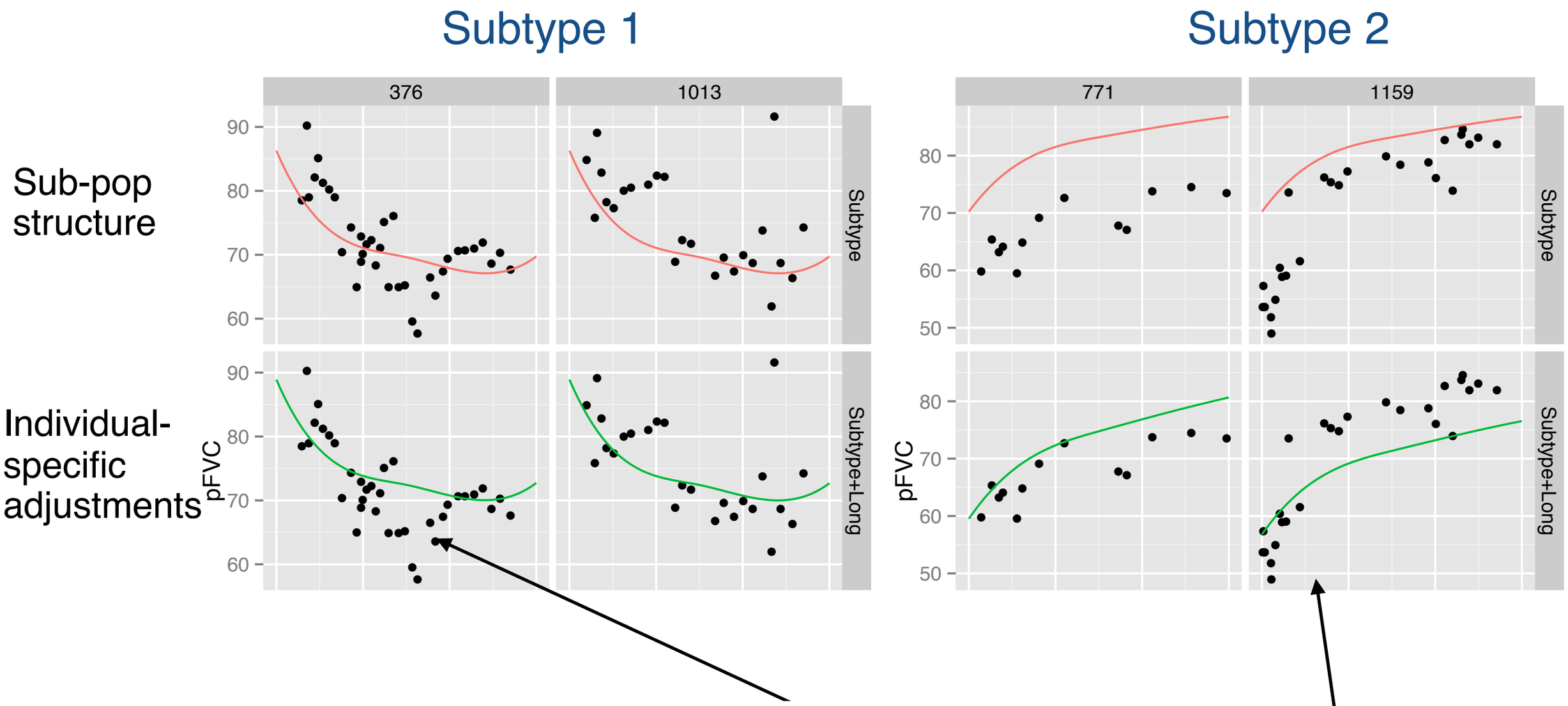
Accounting for Latent Sources of Heterogeneity



b_i parameters specifying individual-specific adjustments
Treated as random effects

$$\boxed{y_{ij}} \mid z_i, \vec{b}_i, f_i \sim \mathcal{N}\Big(\underbrace{\Phi_p(t_{ij})^\top \Lambda\, \vec{x}_i}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) ind. long-term}} \quad , \sigma^2 \Big)$$

Accounting for Latent Sources of Heterogeneity



explains remaining noise sources

$$\boxed{y_{ij}} \mid \vec{x}_{ip}, z_i, b_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\, \vec{x}_{ip}}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) individual}} + \underbrace{f_i(t_{ij})}_{\text{(D) structured noise}}, \sigma^2 \right)$$

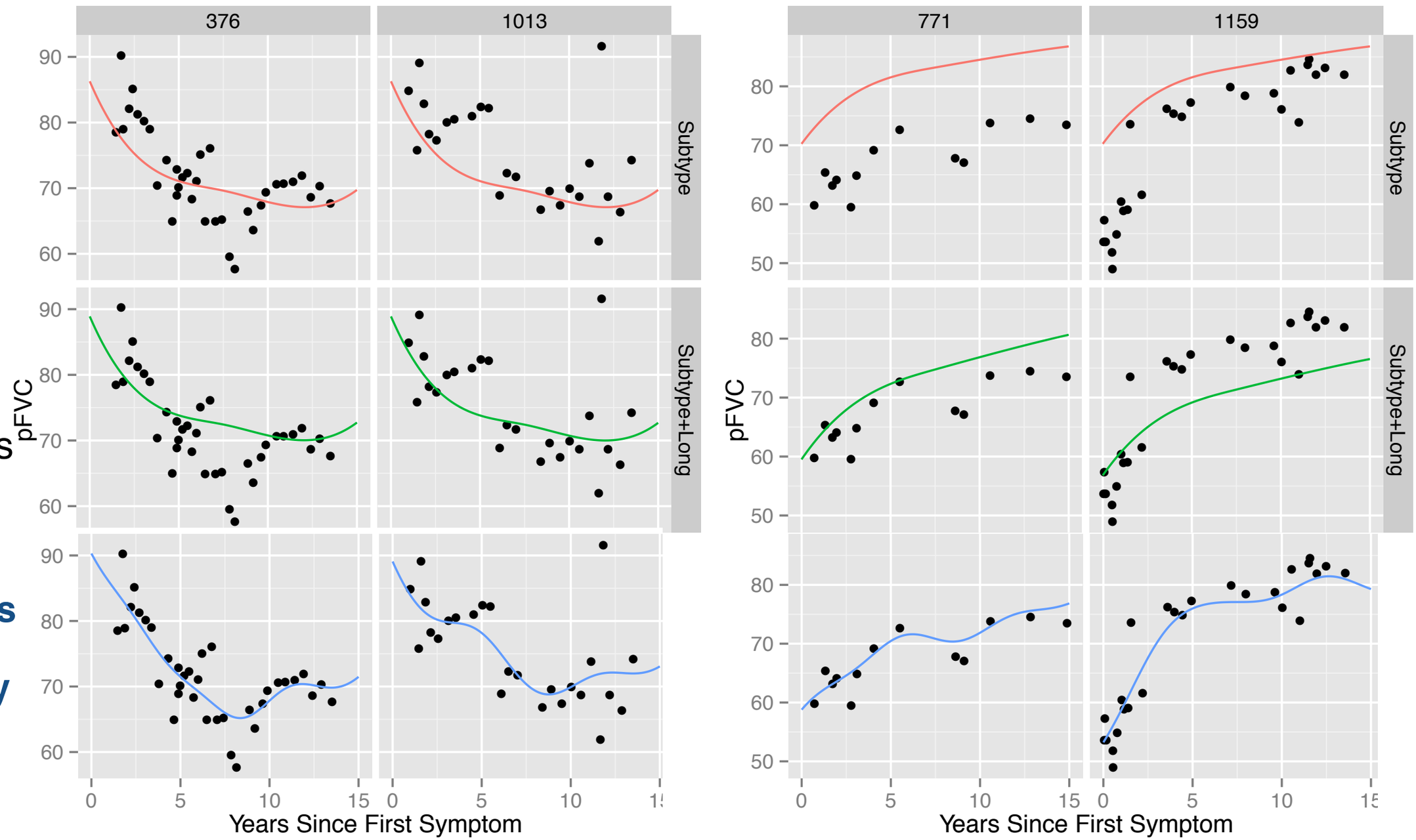
Schulam, Saria, 2015

Accounting for Latent Sources of Heterogeneity

Subtype 1 | Subtype 2

Sub-pop structure

Individual-specific adjustments



Reminder:
Only means are shown. Uncertainty bands not shown

$$\boxed{y_{ij}} \mid \vec{x}_{ip}, z_i, b_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\, \vec{x}_{ip}}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) individual}} + \underbrace{f_i(t_{ij})}_{\text{(D) structured noise}}, \sigma^2 \right)$$

Lawrence, 2004

Schulam, Saria, 2015

Sharing occurs at multiple resolutions

- Use hierarchical Bayes to allow transfer at multiple resolutions. Parameters use different subsets of the data:
  - Population trajectory: data from all individuals
  - Subtype mean trajectories: data from subgroups of similar individuals
  - Individual adjustments: repeated measurements on the given individual
  - Transient adjustments: trends over short periods of time

$$y_{ij}|\vec{x}_{ip}, z_i, b_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda \, \vec{x}_{ip}}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) individual}} + \underbrace{f_i(t_{ij})}_{\text{(D) structured noise}}, \sigma^2 \right)$$

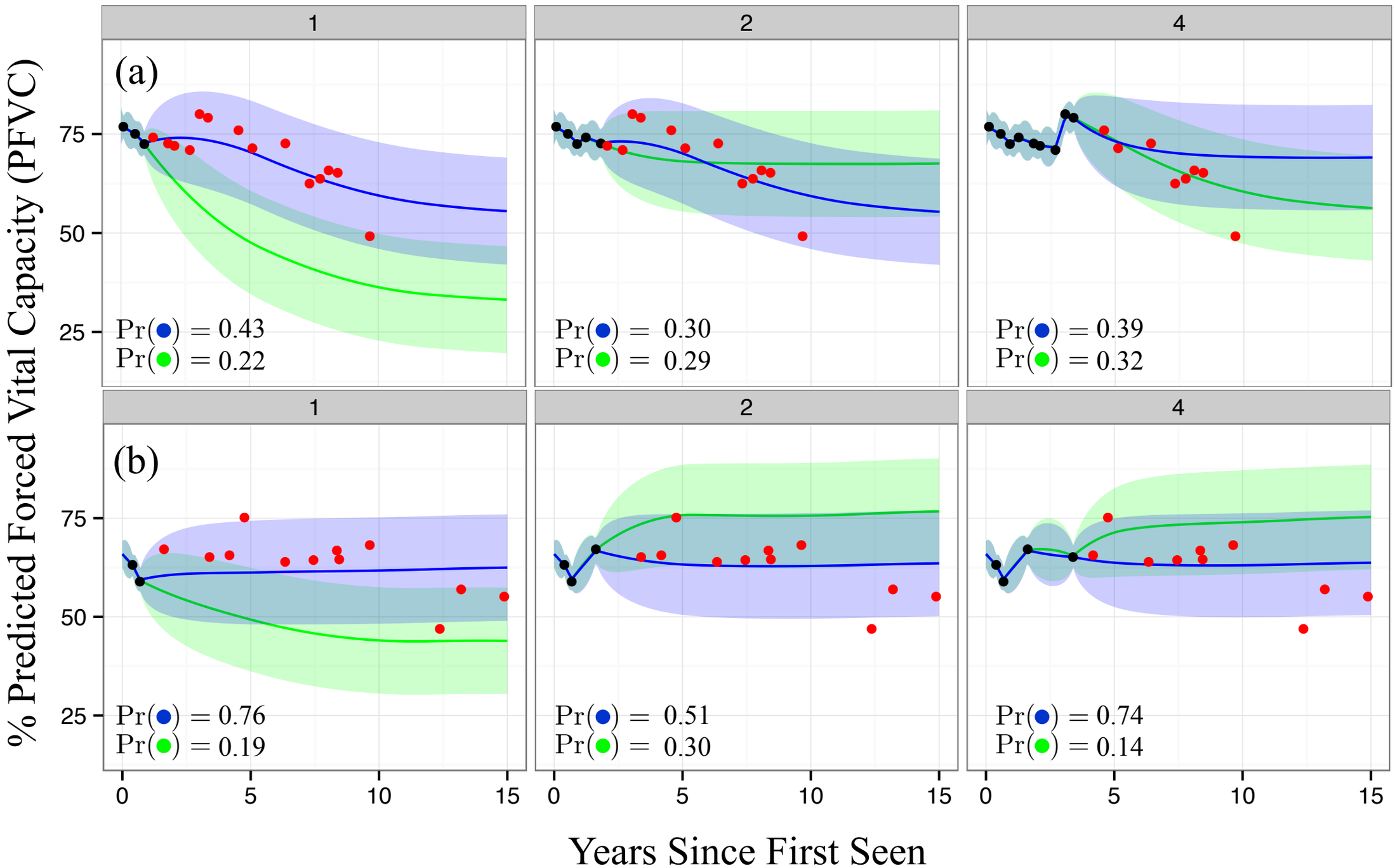
Posterior Predictive Distribution and Dynamic Personalization

$$\hat{y}(t_i') = \underbrace{\Phi_p(t_i')^\top \Lambda \vec{x}_{ip}}_{\text{Population Prediction}} + \underbrace{\Phi_z(t_i')^\top \mathbb{E}^*_{z_i}\left[\vec{\beta}_{z_i}\right]}_{\text{History-dependent Subpopulation Prediction}} + \underbrace{\Phi_\ell(t_i')^\top \mathbb{E}^*_{\vec{b}_i}\left[\vec{b}_i\right]}_{\text{History-dependent Individual Long-Term Prediction}} + \underbrace{\mathbb{E}^*_{f_i}\left[f_i(t_i')\right]}_{\text{History-dependent Individual Short-Term Prediction}}$$

- Use the posterior predictive for *online* predictions as new data are collected.

- Mean of posterior predictive has an intuitive form: replace unobserved individual-specific parameters with their expectations given the clinical history.

Qualitative Analysis



Schulam, Saria, 2015

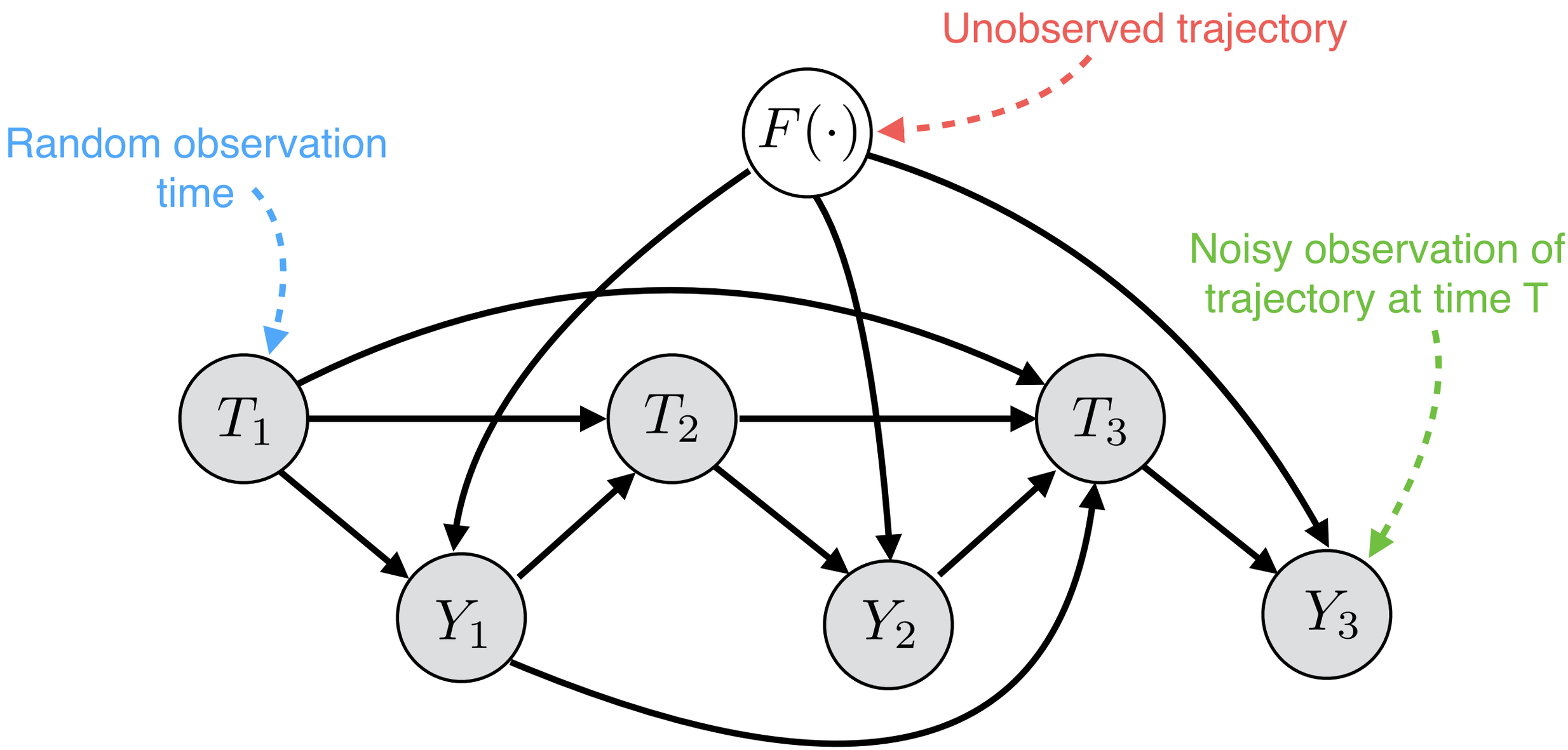
Missing Data

- We observe the trajectory at a finite number of times
  - Do we need to worry about bias due to when the measurements were made?
- When we want to model a *trajectory* there is always going to be missing data
- Is there bias due to when the data are missing?
- When can we use likelihood-based learning?

Little and Rubin, 2014

Missing Data Model

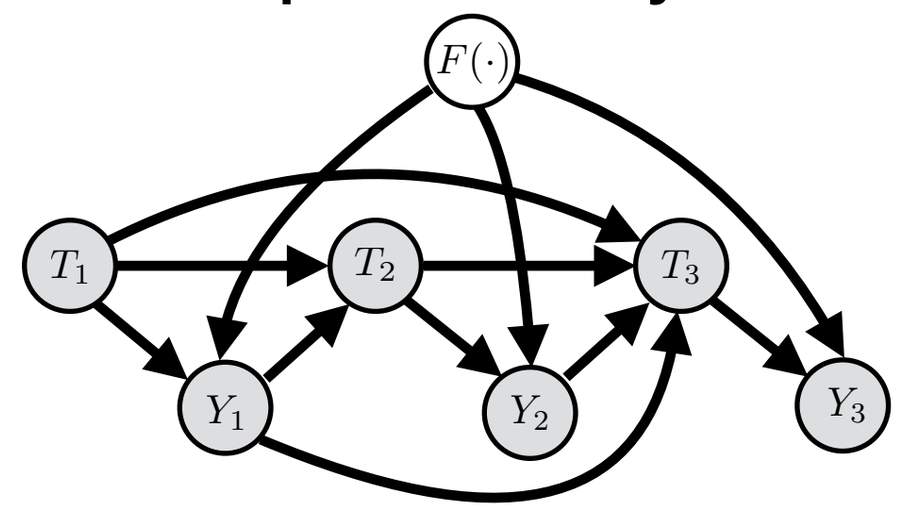
- Consider the three-observation example



Schulam, Saria, 2016

Missing Data Model

- For an arbitrary number of observations, the probability of the observed data can be factored



$$\int p(F = f) p(T_{1:n}, Y_{1:n} \mid F = f) df$$

$$= \int p(F = f) \prod_{i=1}^{n} p(T_i \mid \bar{\mathbf{t}}_i, \bar{\mathbf{y}}_i) p(Y_i \mid t_i, f) df$$

$$= \left[\prod_{i=1}^{n} p(T_i \mid \bar{\mathbf{t}}_i, \bar{\mathbf{y}}_i) \right] \left[\int p(F = f) \prod_{i=1}^{n} p(Y_i \mid t_i, f) df \right]$$

Sampling Model

Trajectory Model

Times of measurements are independent of the latent trajectory

- Allows fitting of likelihood without modeling measurement times

Schulam, Saria, 2016

Missing Data Assumptions

- Note: this is **Missing at Random (**MAR) **Little and Rubin, 2014**
  - The choice of when to measure is based on observed data only.
- Common to decide when to measure based on past observed data
- For example:
  - If there are no recent tests, then clinician is more likely to order a new test.
  - If the past few tests suggest results getting worse, clinician may increase frequency of measurement.
- <u>More explicitly, we made the following assumption</u>
  - The times at which the trajectory is observed depend on (a) observed baseline covariates, and (b) the previous measurement times and values of observed time-dependent variables

Missing Not at Random (MNAR)

- These assumptions do not always hold
- When the observation times depend on unobserved variables, the missing data is **Missing Not at Random**

Little and Rubin, 2014

- For example:
  - If individuals schedule their own visits, they may only have measurements when they feel sick
  - If observation times are determined by other time-dependent variables (e.g. other lab tests) that are **not** in the data

General Ideas vs. Domain Specific

$$y_{ij}|\vec{x}_{ip}, z_i, b_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\, \vec{x}_{ip}}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) individual}} + \underbrace{f_i(t_{ij})}_{\text{(D) structured noise}}, \sigma^2\right)$$

What to take away to new problems?

- #1 Latent Variable model to account for latent sources of heterogeneity
- #2 Posterior Predictive distribution to prevent overfitting and learn as new data are collected on the individual
- #3 Transfer at multiple resolutions

- Choice of hierarchy potentially introduces bias. Generates intermediate quantities that are interpretable by clinicians.

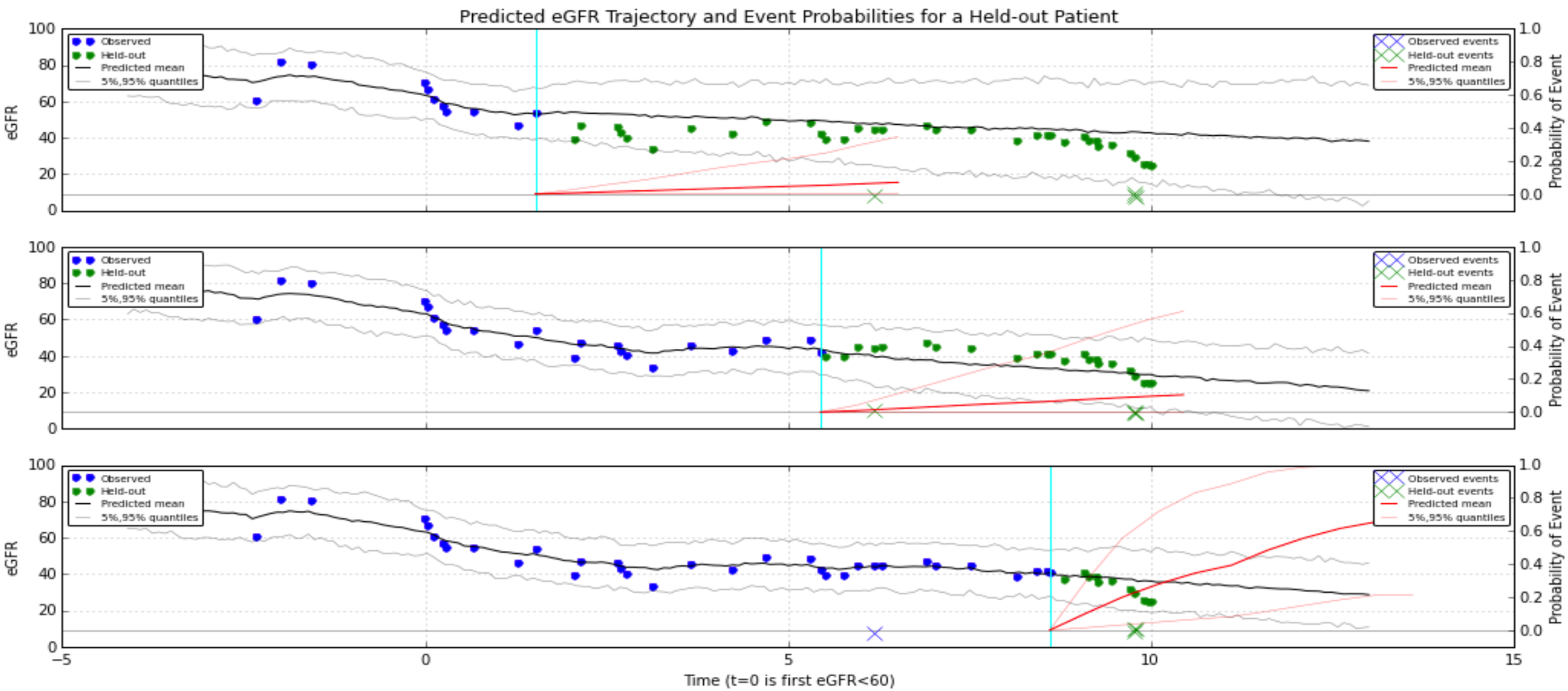
Two useful by-products: (1) Couple models, (2) Subtyping

Which modeling decisions were specific to this app?

- #1 No treatment effects
- #2 Choice of basis for the trajectories and noise models should reflect properties of the disease data.

Another example: Chronic Kidney Disease Prediction

- Use clinical markers measured over time (eGFR) to dynamically predict the probability of stroke



Futoma et al. 2016

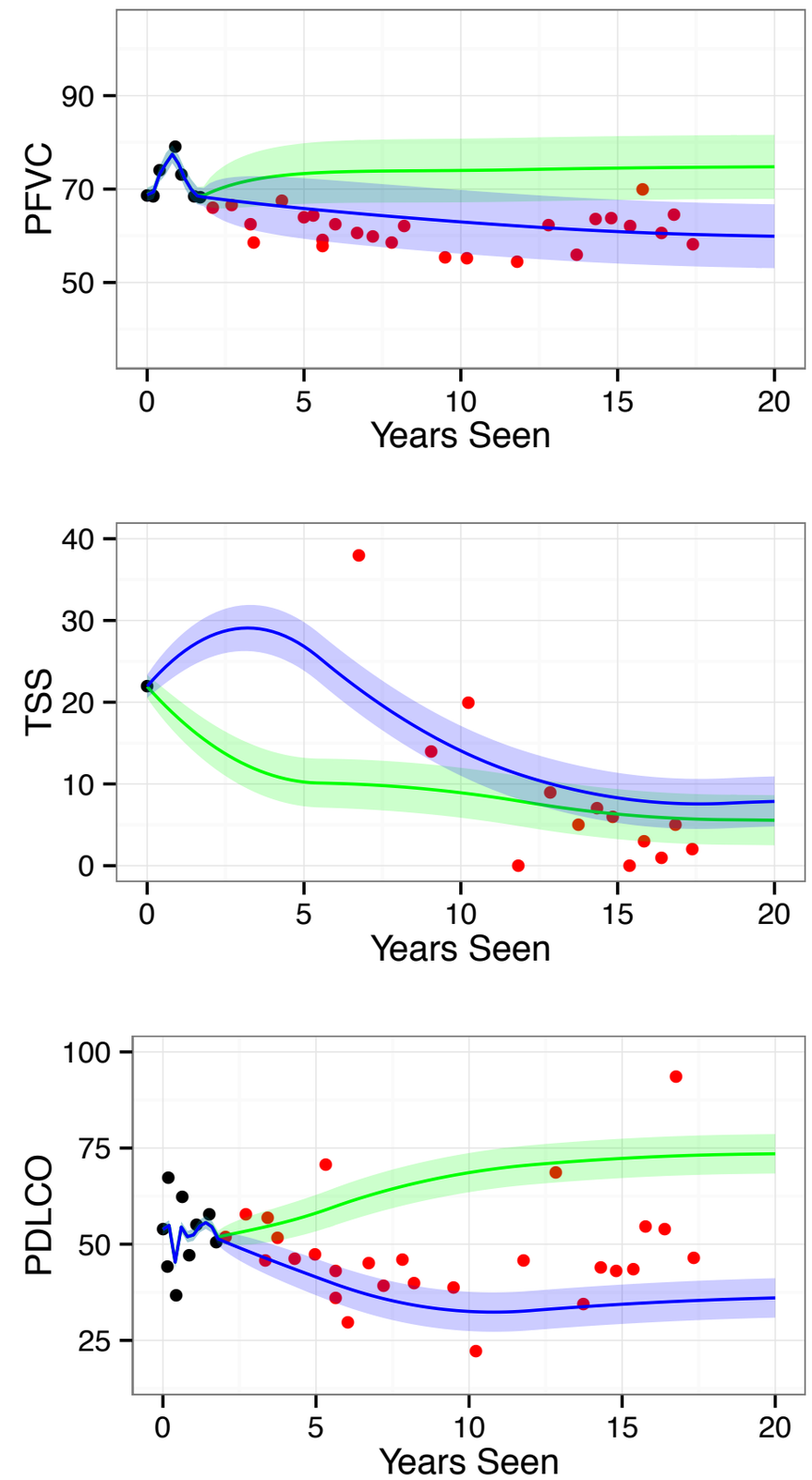
Extending to Multivariate Trajectory Data

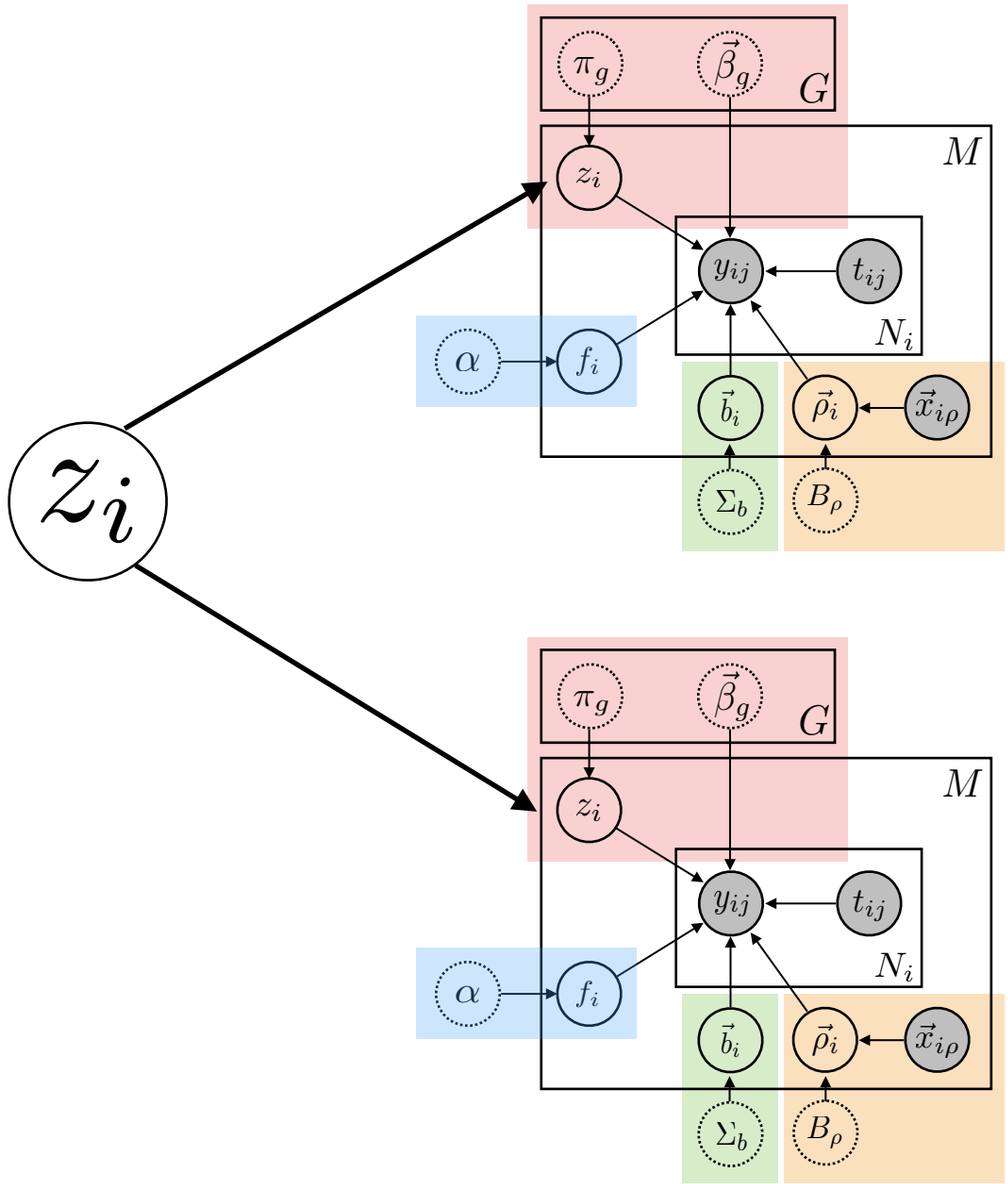
Motivation:

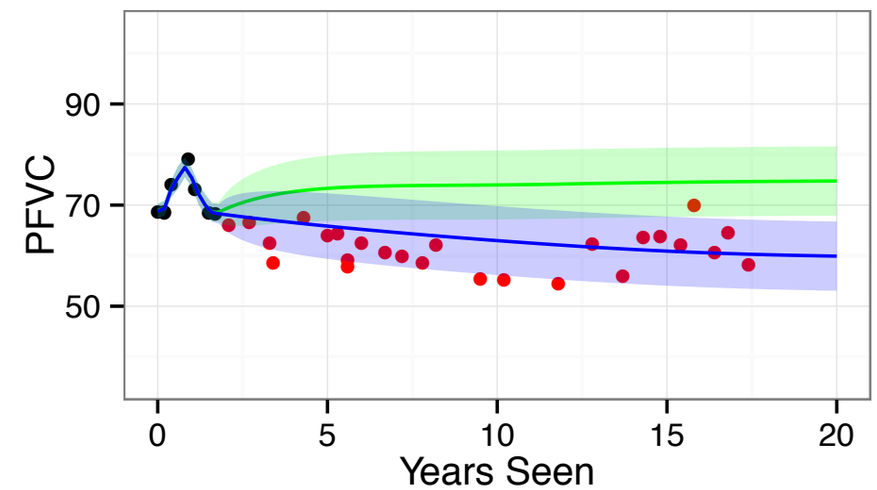
- Lung subtypes likely related to skin subtype.
- In **systemic** diseases, many clinical markers are measured to monitor different organ systems

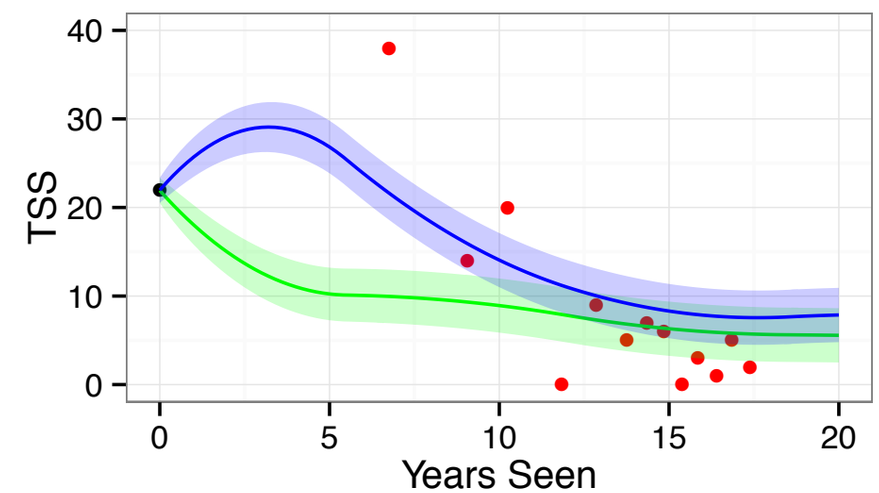
Challenges:

- Measurement times are not aligned
- Some measurements may never be made on an individual
- Rate of measurement varies across individuals





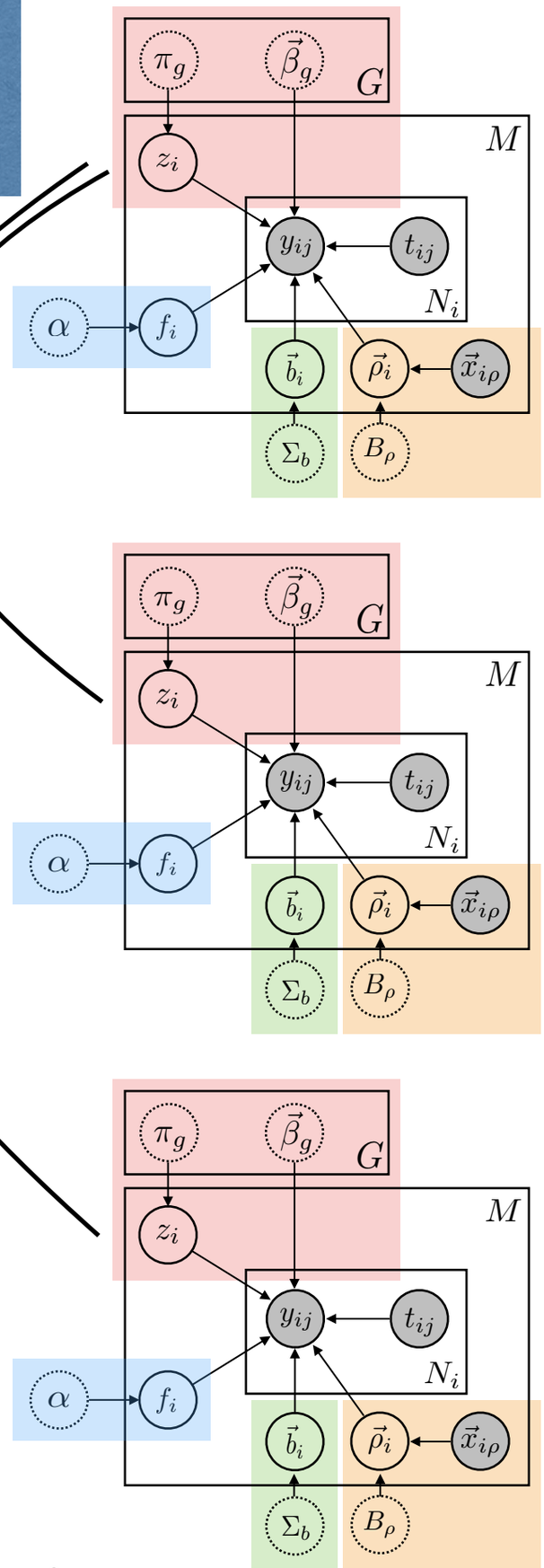




- Need joint models that can flexibly encode complex dependencies across markers
  - Classic assumption of Naive Bayes structure is incorrect. In general, hard to specify generative model.

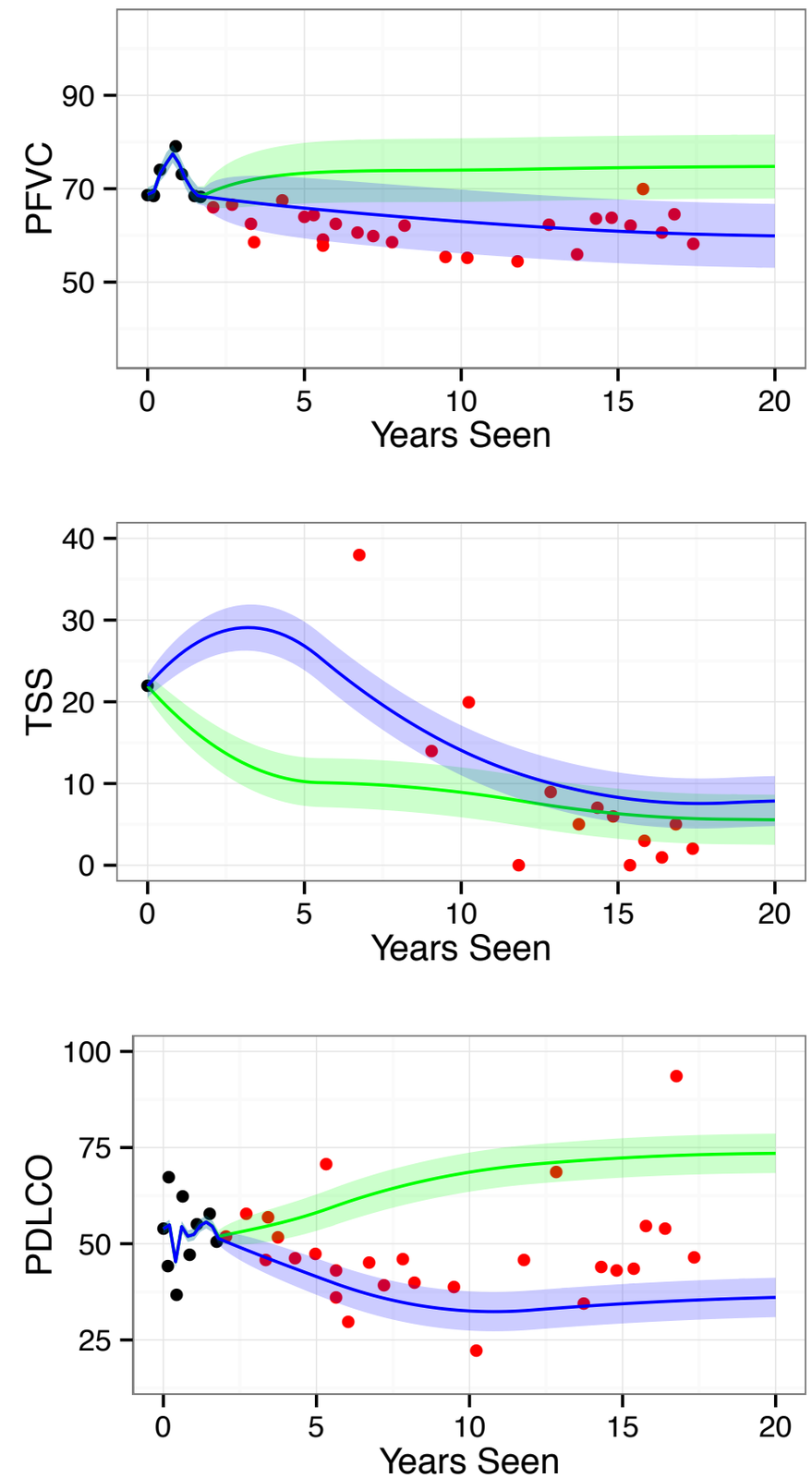
Coupled Latent Variable Models

Schulam, Saria, 2016



Conditional random field (CRF) to model pairwise dependencies

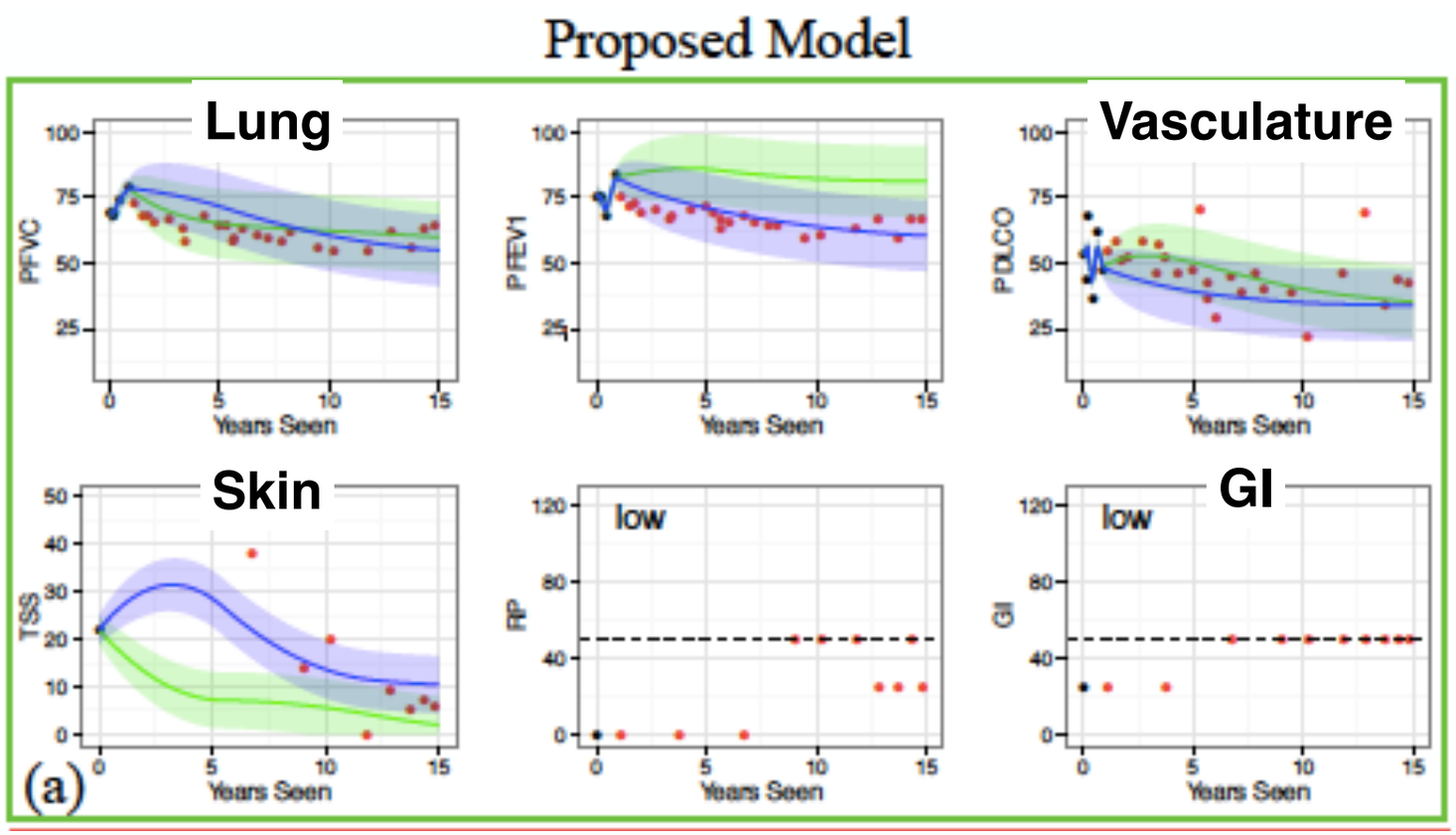
Model target marker conditioned on auxiliary markers.

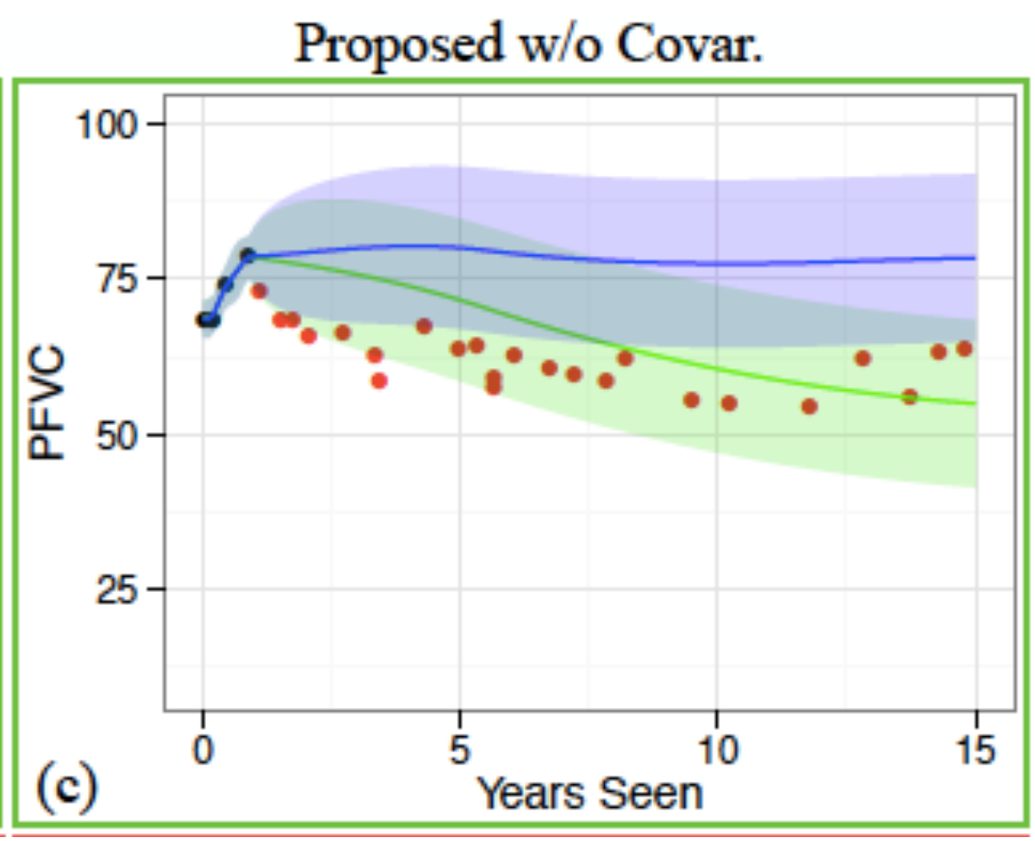


$$\hat{y}\left(t_i^*\right) = \sum_{z_i=1}^{G} \int_{R^{d_\ell}} \int_{R^{N_i}} \mathbb{E}\left[y_i^* | z_i, \vec{b}_i, f_i\right] P\left(z_i, \vec{b}_i, f_i | \vec{y}_{i,\leq t}, \vec{y}_{1:C,i,\leq t}, x_{ip}, \Theta\right) df_i, d\vec{b}_i$$

Discussion

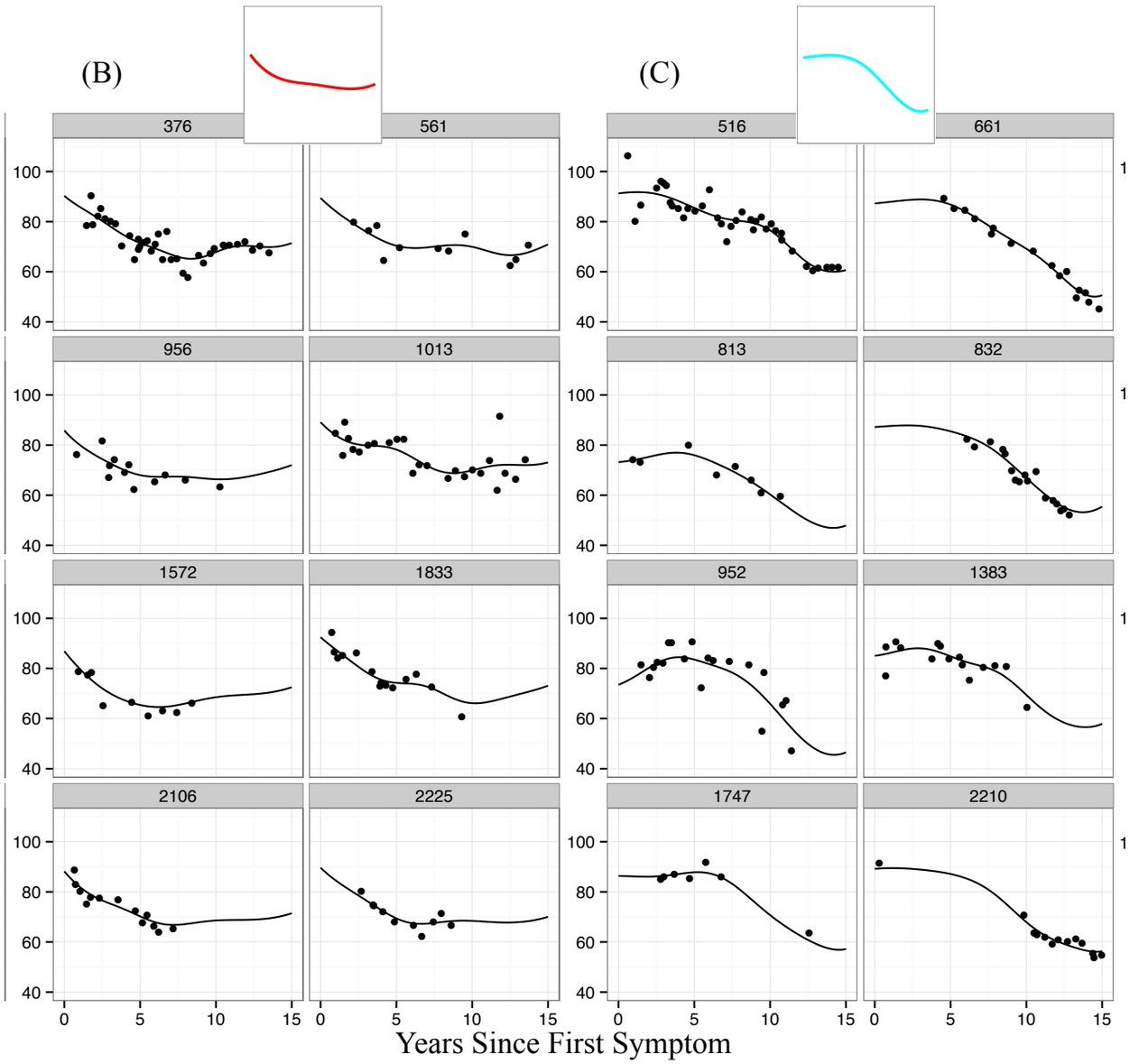
Schulam, Saria, 2016





- Allows what-if reasoning.
- Ways to incorporate domain knowledge into individual marker-level sub-models.
- Plug-in / replace better models as they become available.
- Open question: Calibrated posteriors for specific clinical tasks

Subtypes and Precision Medicine



Eg: Subtypes based on disease trajectories **Schulam et al., 2015**

Other e.g. of sub-grouping patients:

Doshi-Velez et al., 2014 **Yuen et al., 2002** **Wang et al.,**

- **Desired**: Identify subgroups with distinct underlying biological mechanism driving disease.
- **Current Approach**: Identify candidate subtypes via clustering and associate with molecular determinants. See brief introductory review: **Saria, Goldenberg 2015**
- Open question: How do we increase the efficiency of subtype discovery experiments?
  - Combine high-dimensional multivariate data to identify subtypes?
  - Current approaches (e.g., k-means with a pre-specified distance metric).
  - Learning metrics: **Sun et al., 2012**

Related Ideas

- Functional Data Analysis

  Ramsay and Silverman 2005 | Bahadori et al. 2015 | Schulam, Arora 2016

- Modeling Disease Trajectories

  Ross and Dy, 2013 | Wang et al., 2014 | Ghassemi et al. 2014

  Rizopoulos et al. 2015 | Liu and Hauskrecht, 2016 | Elibol et al. 2016 | Wang et al., 2015

- Dynamical Prediction:

  Yu et al. 2008 | Rizopoulos 2011 | Proust-Lima et al. 2014 | Yoon et al. 2016

- Personalization

  Berkovsky et al. 2008 | Salakhutdinov and Mnih 2008 | Adomavicius and Tuzhilin 2010

- Multi-resolution/hierarchical models

  Konstan and Riedl 2012

  Gelman and Hill, 2006

- Multivariate time-to-event

  Rizopoulous and Ghosh, 2011 | Andrinopoulou et al. 2014 | Futoma et al. 2016

Overview

- **Part 1—Setting up the problem of Individualization**
  - Example using a chronic disease
  - Simple setting: No Treatment Effects
  - **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Key ideas: Transfer learning, Multilevel modeling
- **Part 2—Estimating Treatment Effects & Individualized Treatment Effects**
  - Example using inpatient data
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP
- **Part 3—Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

No Control over Data Collection Process

- **Part 4—From Predictions to Treatment Rules**
  - Key ideas: Q-learning, Dynamic Treatment Regimes
  - Connections to Reinforcement Learning

Control over Data Collection Process

Example: Exercise and Blood Pressure

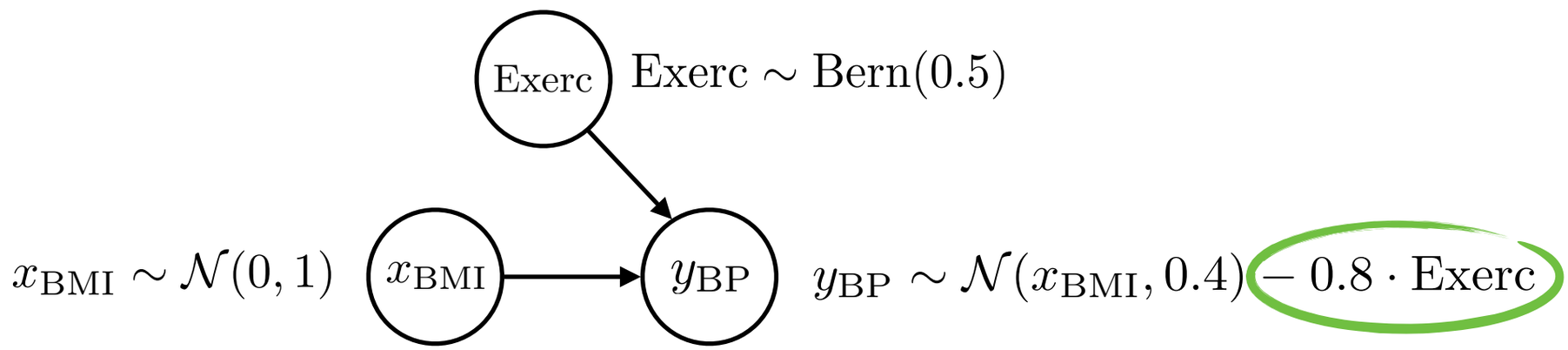
- Hypothesis: exercise lowers blood pressure
- In this example, we have:
  - (a) A treatment (exercise)
  - (b) An outcome (blood pressure)
- How can we use data to estimate whether exercise will lower blood pressure?

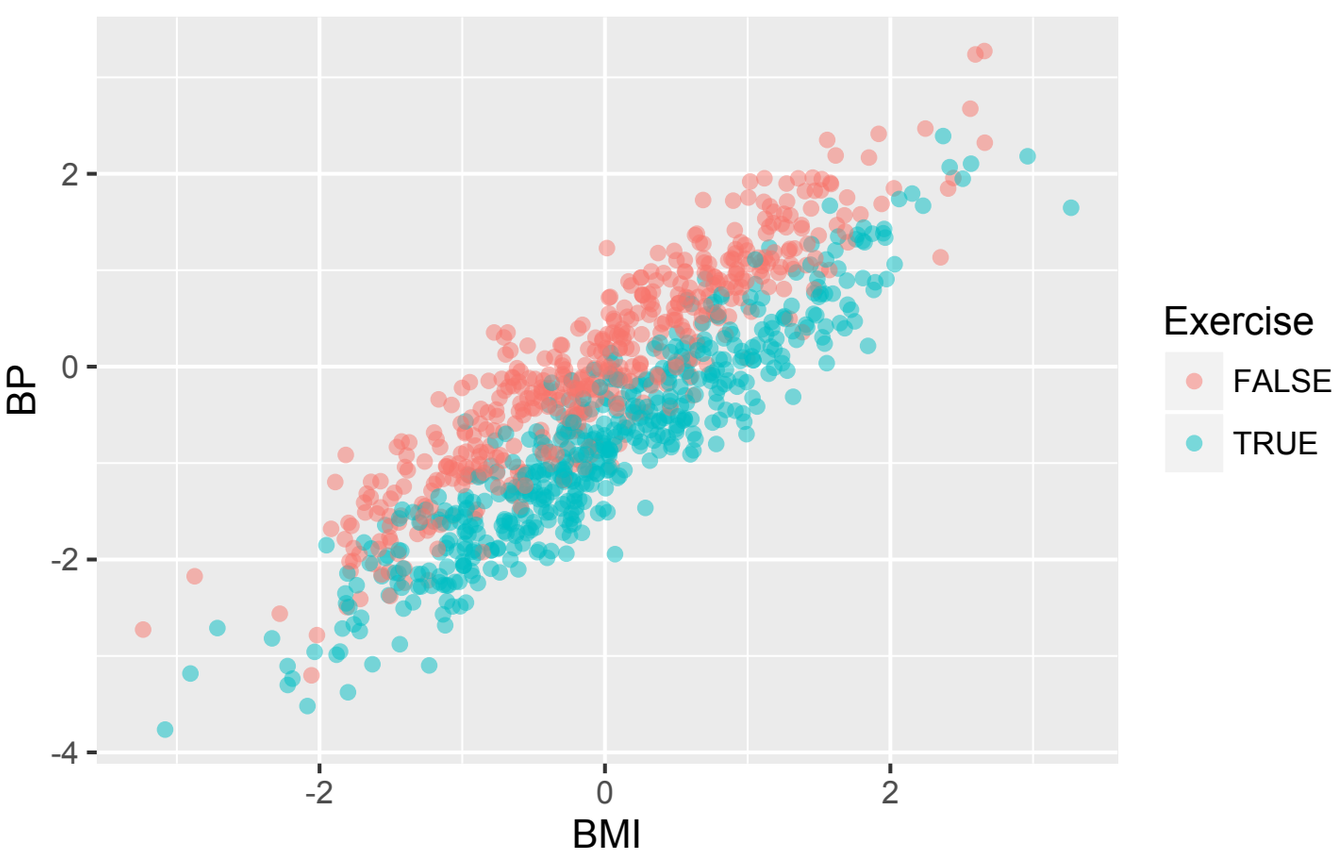
Example: Exercise and Blood Pressure

- 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)

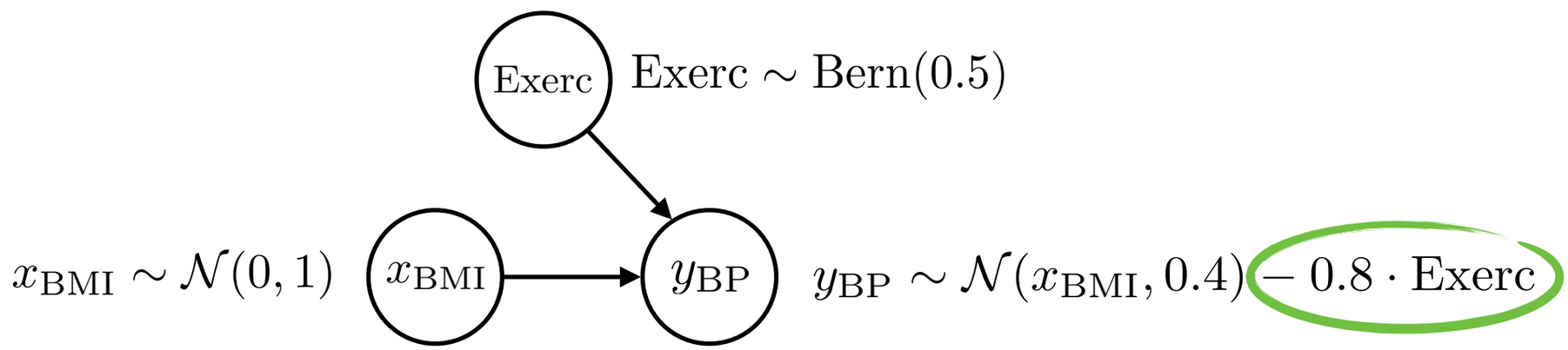
- Dataset generative model:



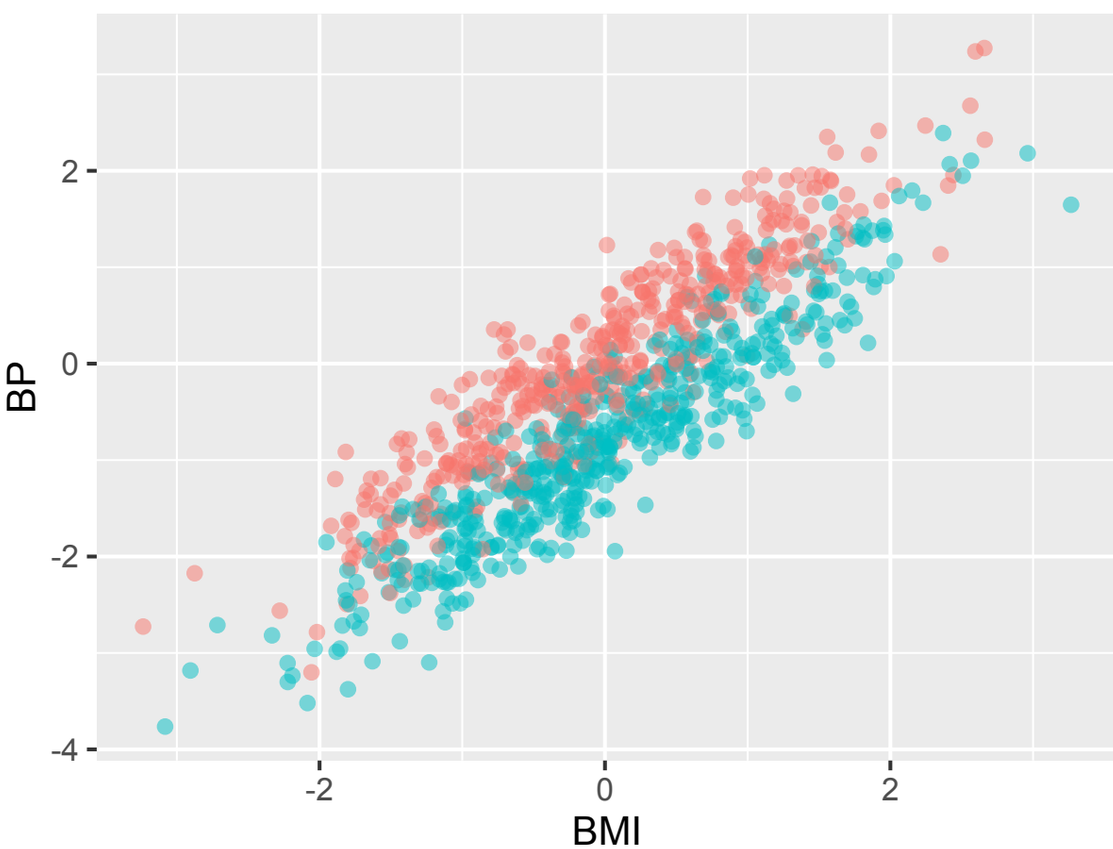


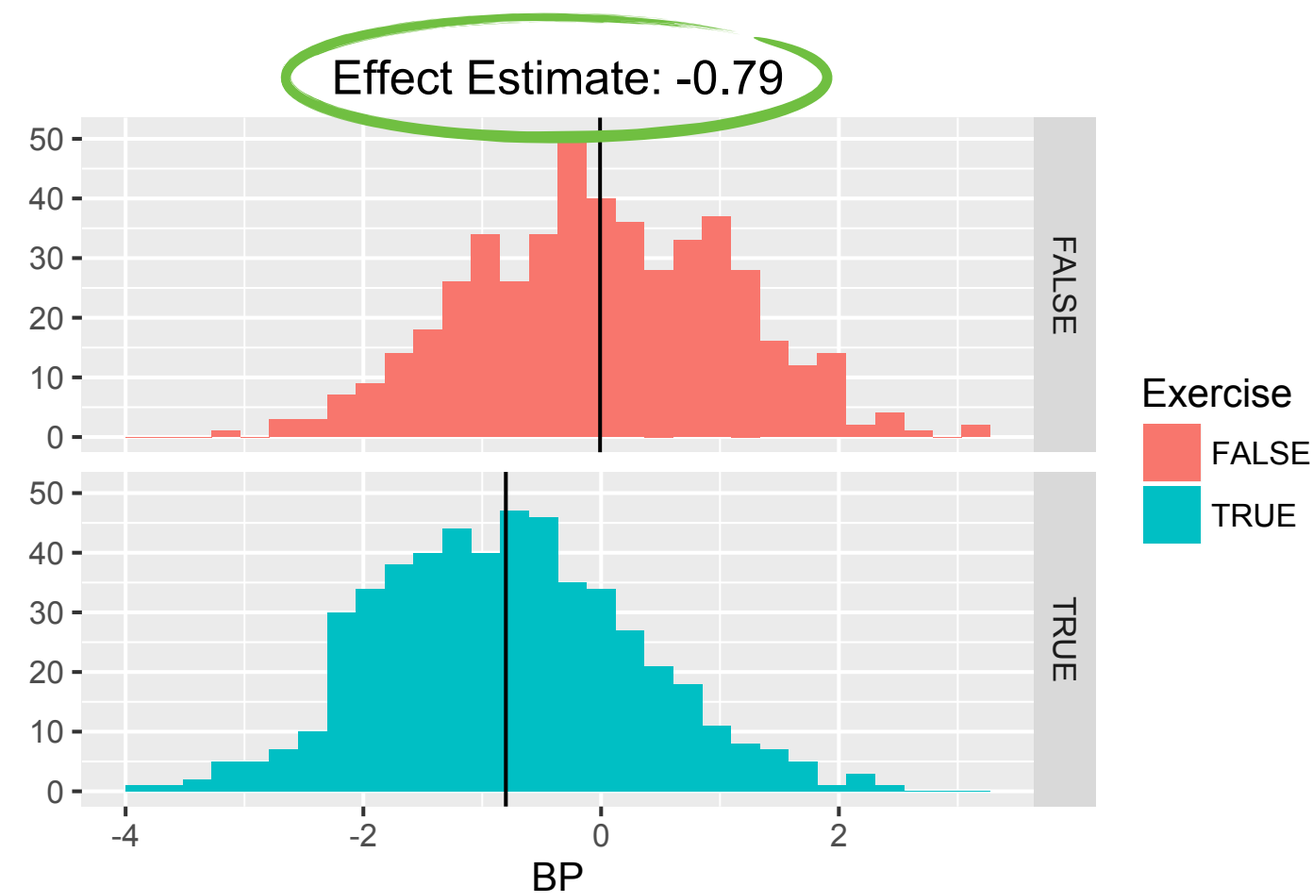
Randomized Controlled Trial (RCT)

- Dataset generative model:



- Comparing averages will work!



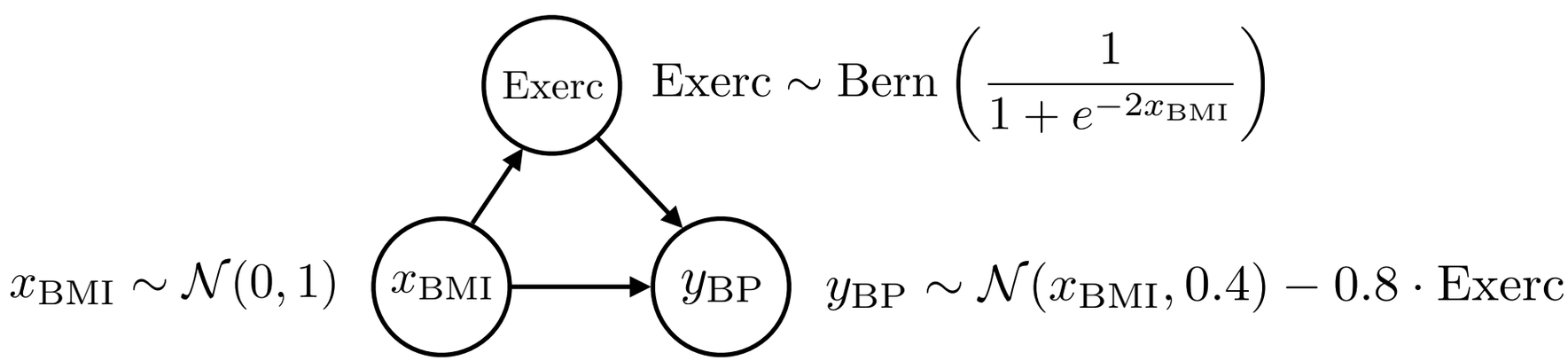


Observational Data

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

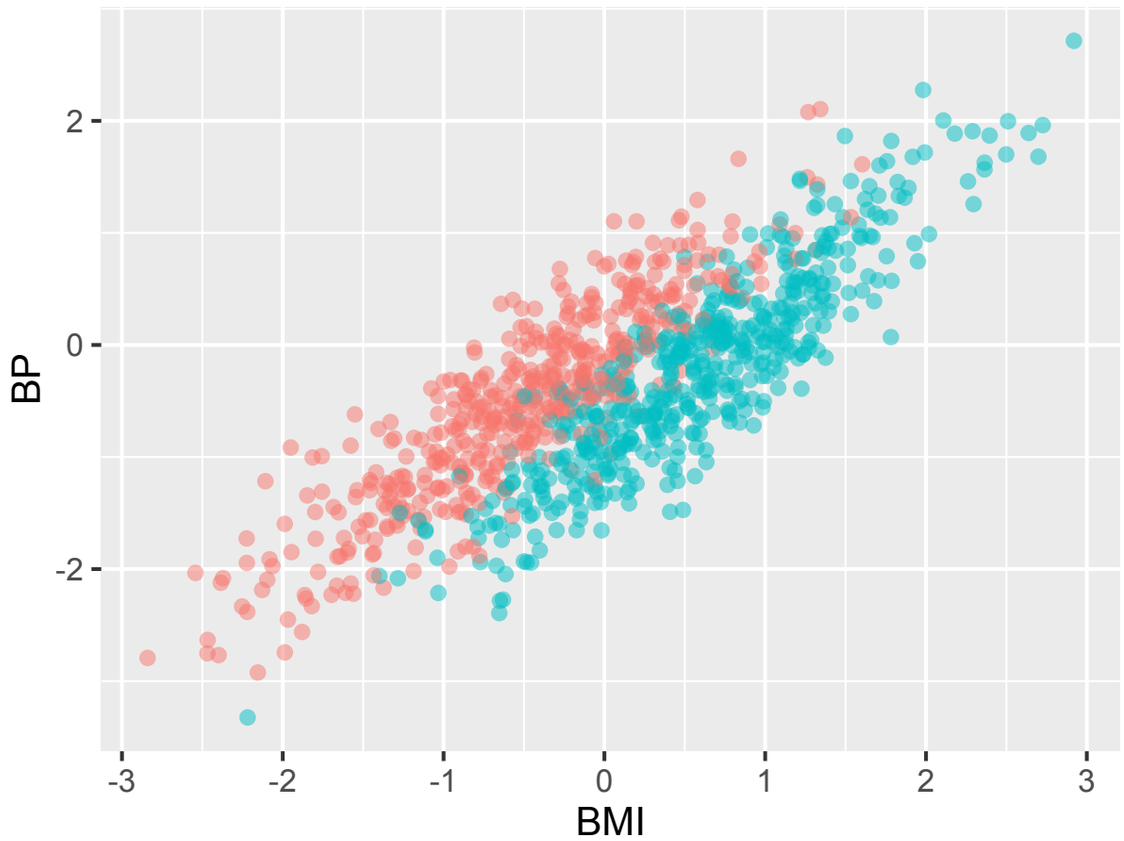
- In the observational data, **exercise is *assigned by the clinicians* caring for the individuals**

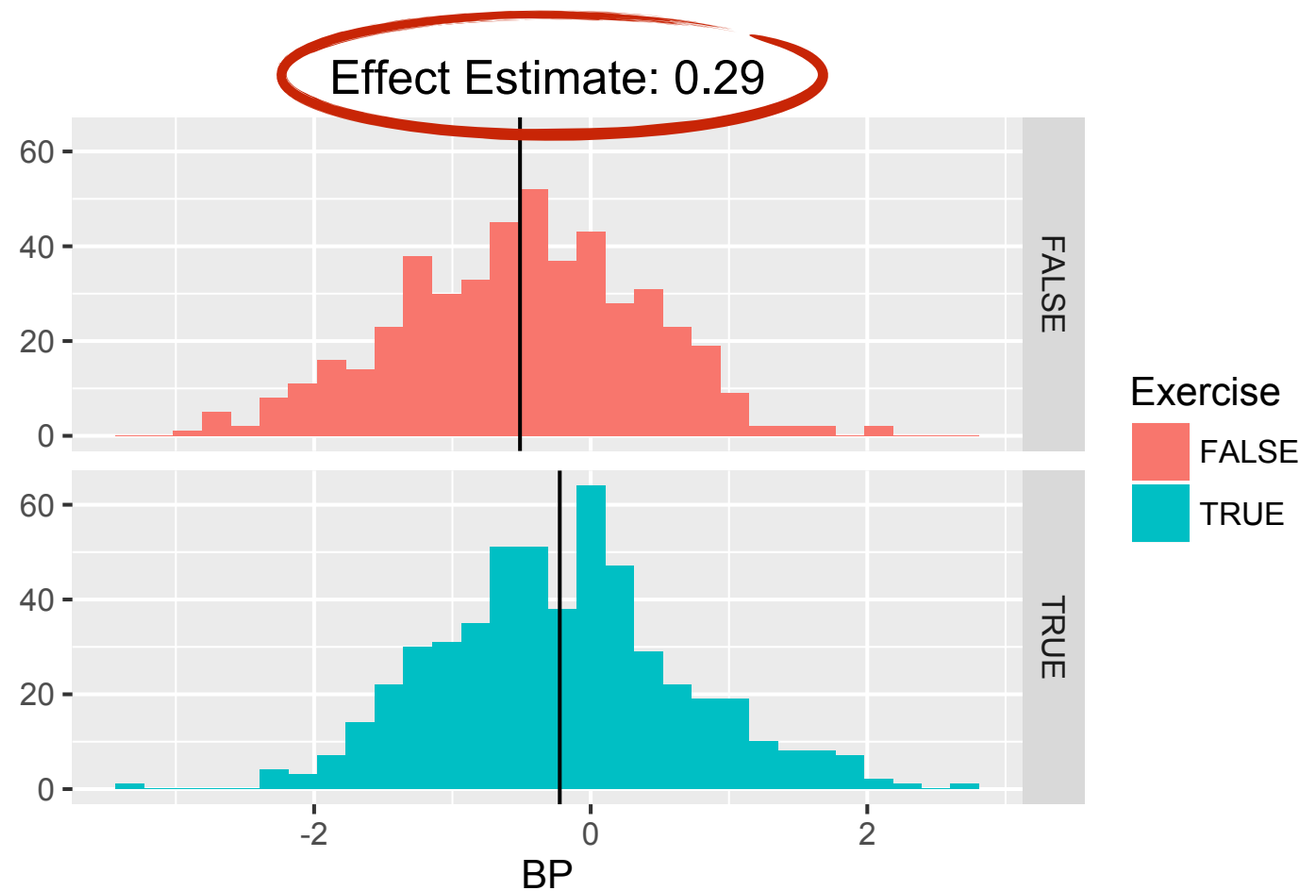
- In particular, we assume that a higher BMI makes prescription of exercise more likely:



Observational Data

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

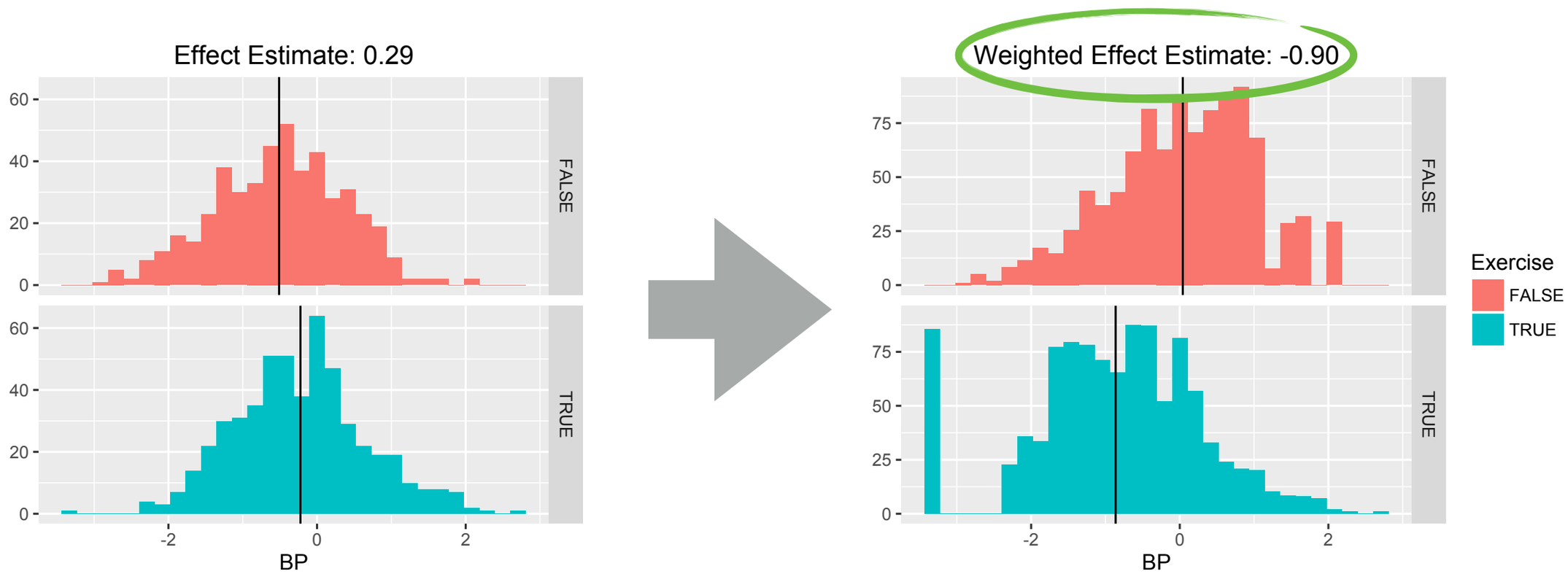




Approach 1: Weighting

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

- <u>Intuitive idea</u>: 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., 1990** **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*
  - It 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:

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

To explain NUC graphically, we introduce the graphical notation of SWIGs.

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

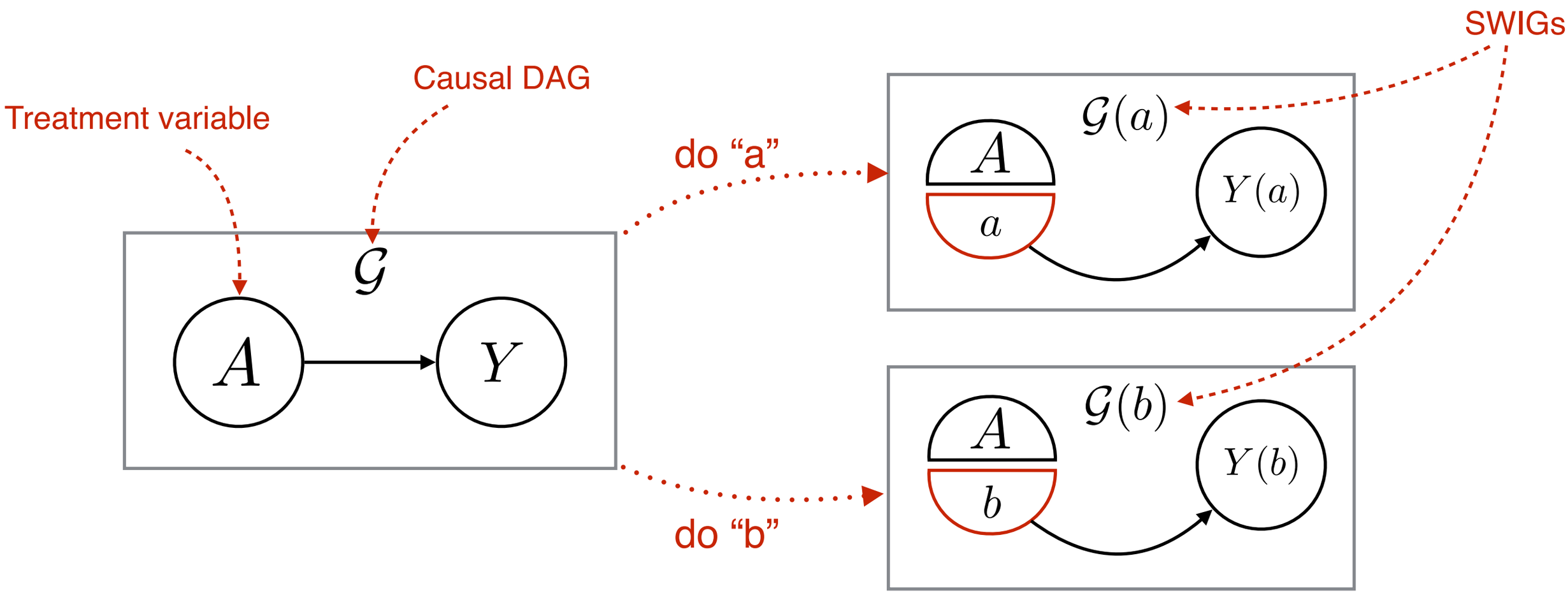
- We apply ***node-splitting*** **operations to treatment variables to represent interventions**

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:

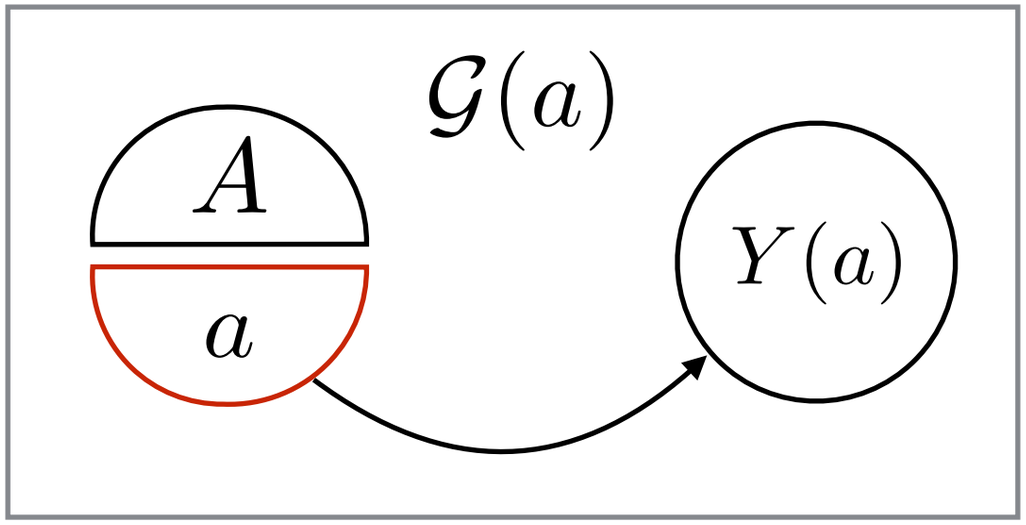


Richardson, 2014

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

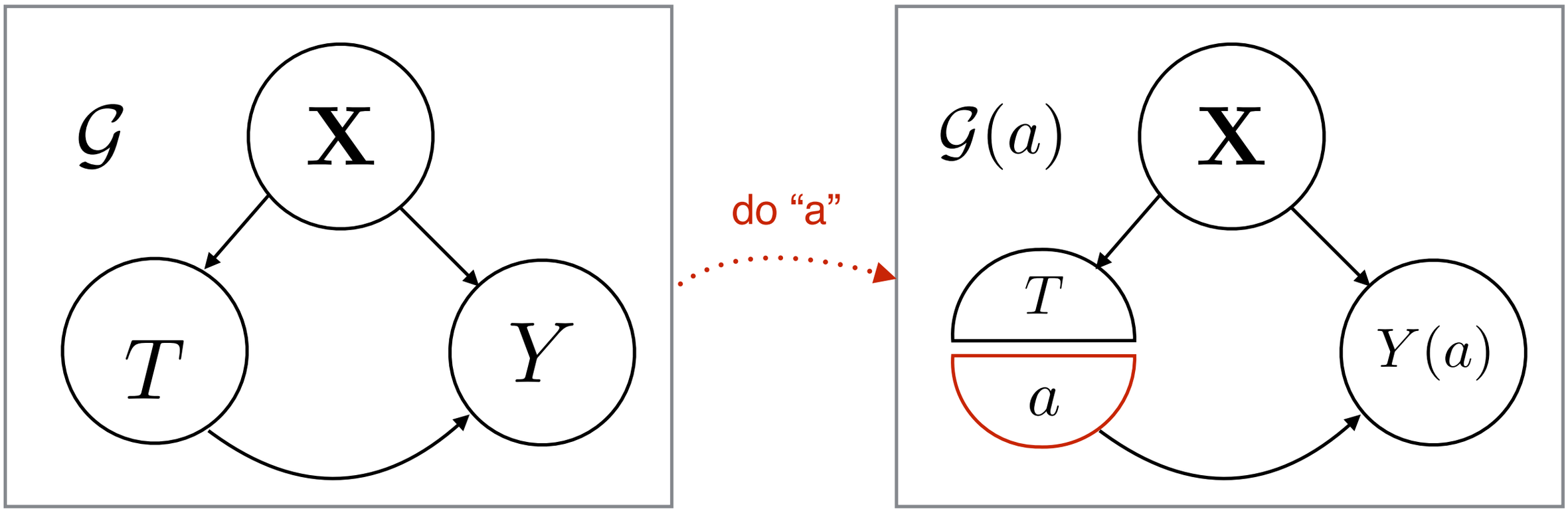
Richardson and Robins, 2014

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

Using Potential Outcomes Framework to Simulate RCT

- Our observational data is drawn from

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

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

Learning Potential Outcome Models

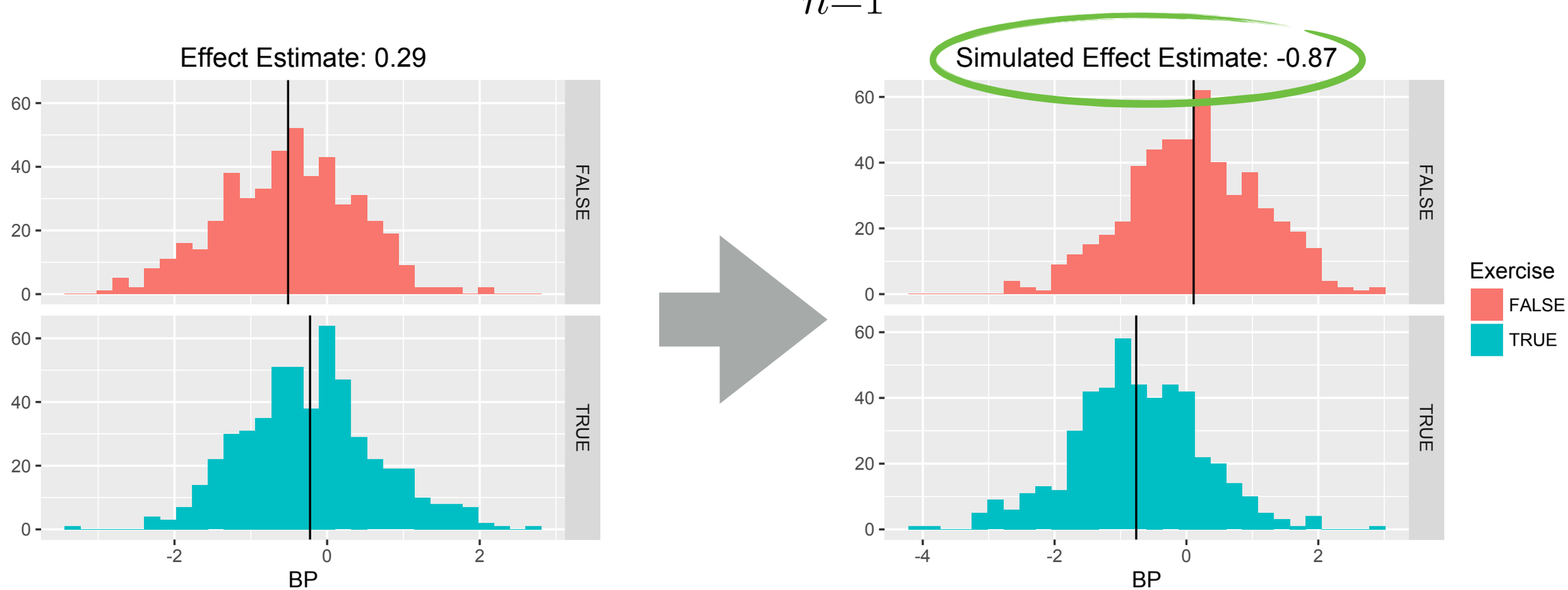
- To simulate data from a new policy, we need to learn the potential outcome models
  - If we have an observational dataset where assumptions 1-3 hold, then this is possible!
- Assumptions allow estimation of potential outcomes from (observational) data:

$$\begin{aligned} \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}) &= \mathrm{P}(Y(a) \mid \mathbf{X} = \mathbf{x}, A = a) && \text{(A3)} \\ &= \mathrm{P}(Y \mid \mathbf{X} = \mathbf{x}, A = a) && \text{(A1)} \end{aligned}$$

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:

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

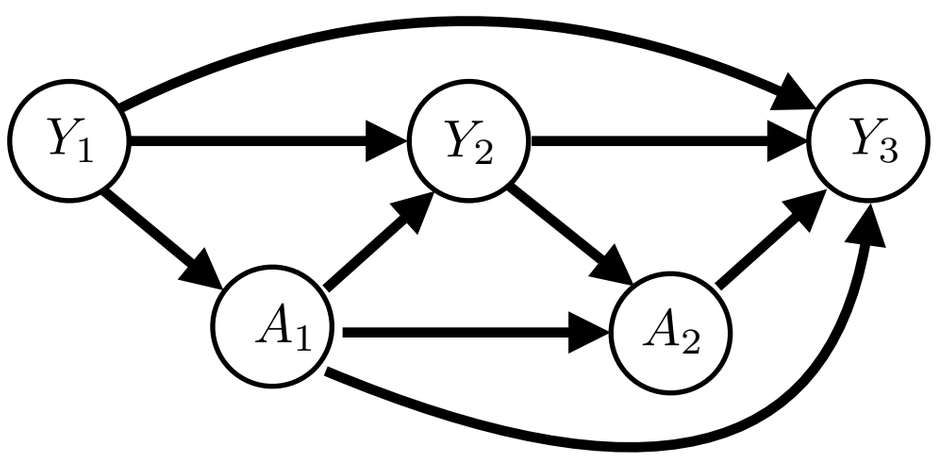
To account for the heterogeneous treatment effect among patients, it is more of interest to look at **CATE**, the conditional average treatment effect:

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

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

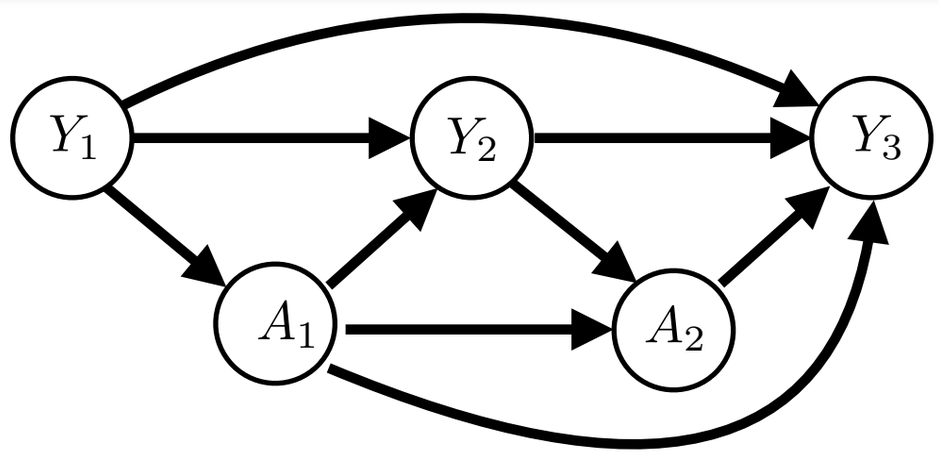
- Interventions and observations are interleaved
  - Intervention effects future observations
    Those observations affect future interventions
    And so on…
  - When can we disentangle to learn unbiased models of potential outcomes?
- Also called time-varying confounding.



Robins 1986

Sequential Treatment Assignment and Time-Varying Confounding

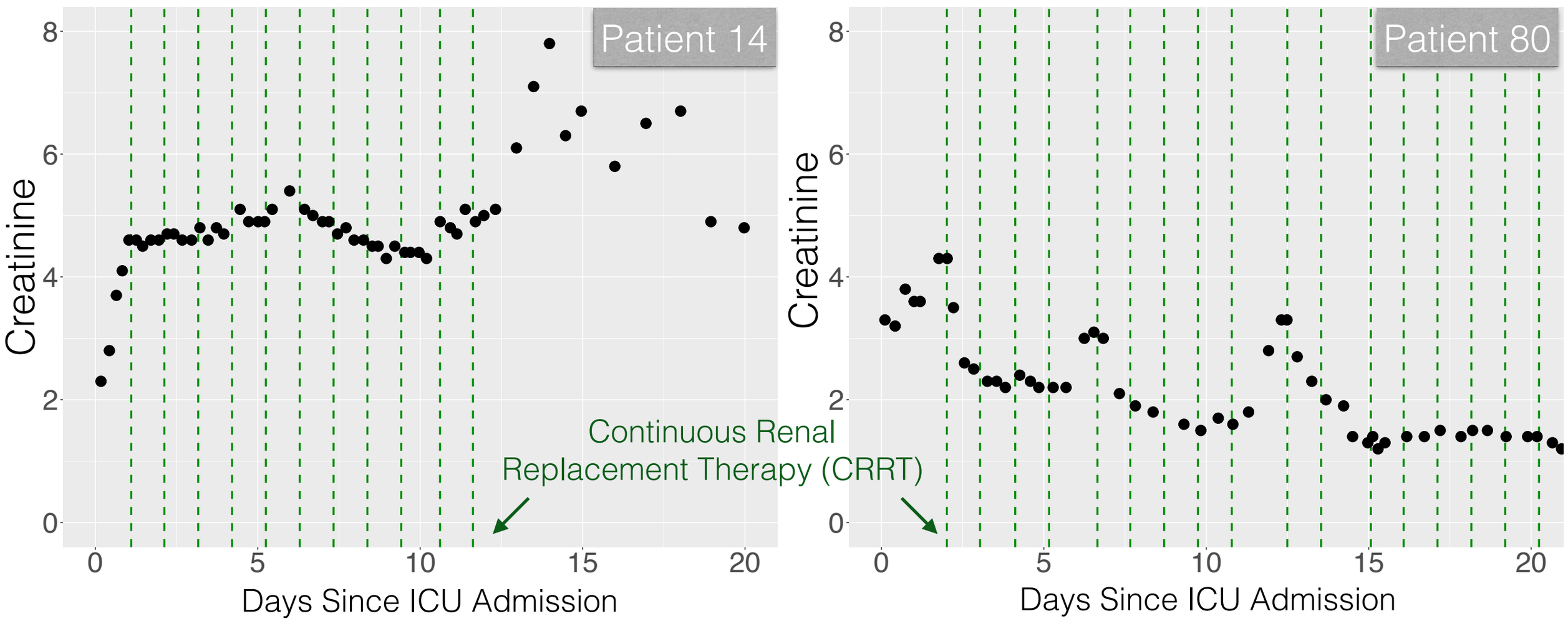
- Interventions and observations are interleaved
  - Intervention effects future observations
    Those observations affect future interventions
    And so on…
- As in single-treatment, single-outcome examples, we need assumptions that allow us to link **conditional distributions** to the target **potential outcome models**



Robins 1986

Estimating Individualized Treatments Effects From Clinical Records

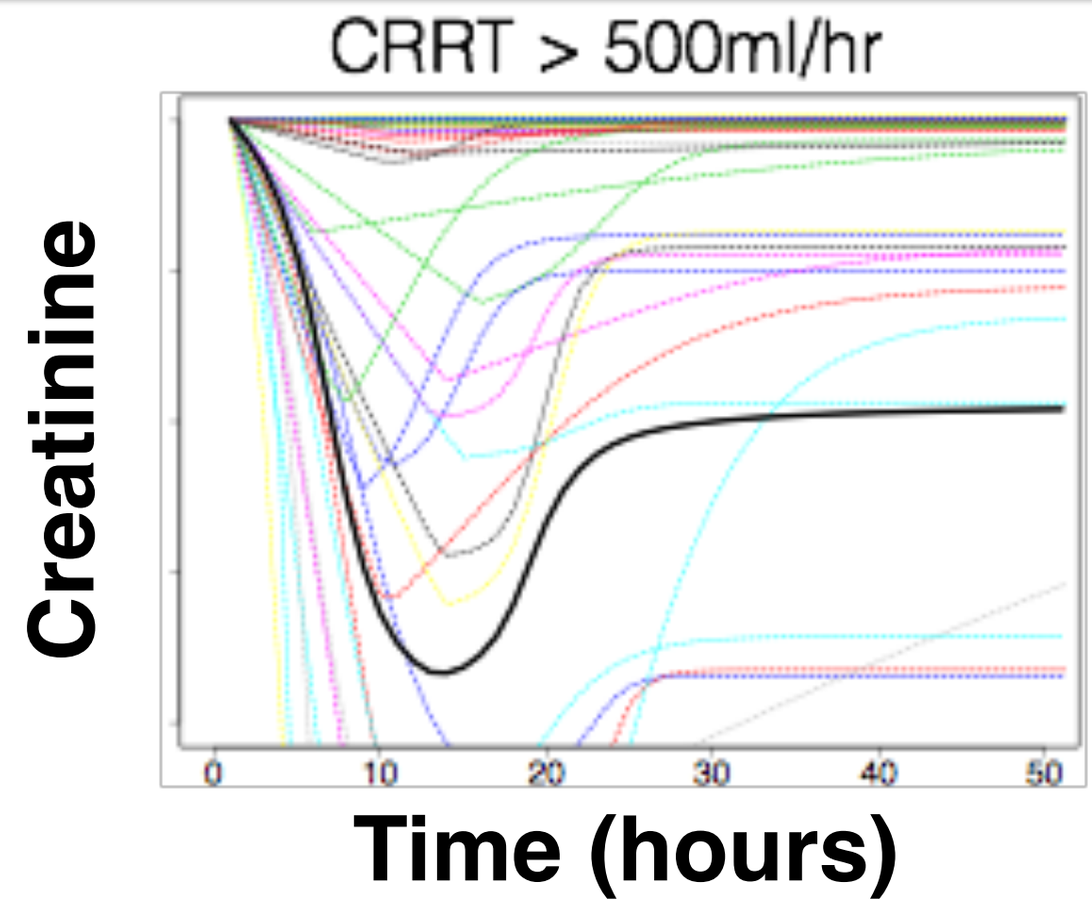
For many disease, response to therapy varies greatly across **individuals**. To **personalize therapy**, we need to estimate at the individual level their likely **response to treatment**.



Distribution over Individualized Treatment Response Curves

We wish to obtain **uncertainty** estimate over an **individual's treatment response over time**. And we want to estimate this from routinely collected data

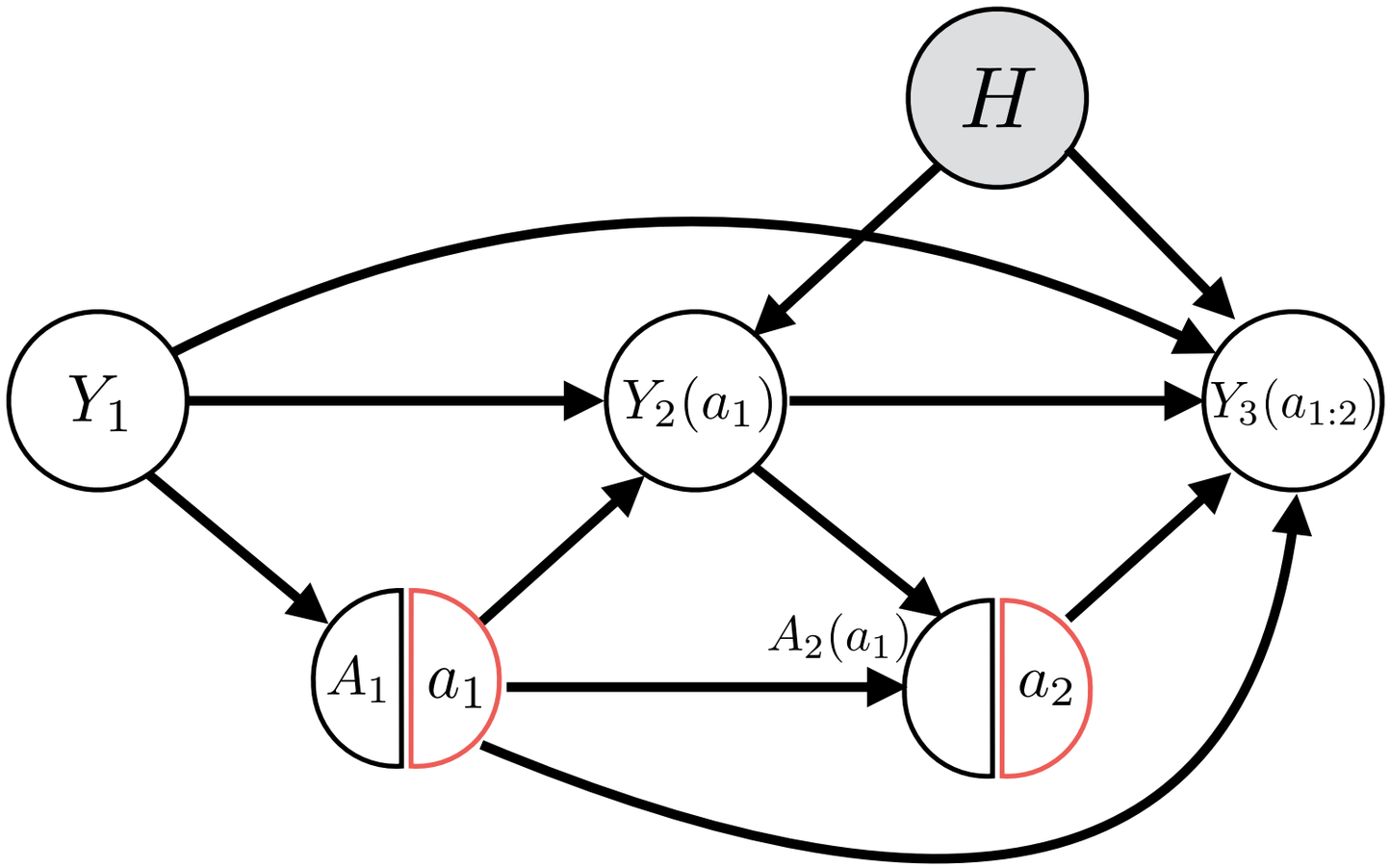
- sparse, irregularly sampled clinical time series



- **Population averages vs. Individualized Estimates**
  - Refined as new measurements are collected on the individual
- **Point-in-time vs. Treatment Response Curve**

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

Approach: g-formula

Robins 1986

For patient i:

Observations $\boldsymbol{Y}_i = \{Y_{i1}, ..., Y_{iJ_i}\}$ measured at times $\boldsymbol{t}_i = \{t_{i1}, ..., t_{iJ_i}\}$

Treatments $\boldsymbol{A_i} = \{A_{i1}, ..., A_{iL_i}\}$ prescribed at times $\boldsymbol{\tau_i} = \{\tau_{i1}, ..., \tau_{iL_i}\}$

A set of covariates $\boldsymbol{C}_{ij} \in \mathbb{R}^p$

Estimation requires a statistical model for estimating conditionals:

$$P(Y_{ij} | a_{i,j}, \mathbf{a}_{i,\leq j-1}, \mathbf{y}_{i,\leq j-1}, \mathbf{C}_{ij})$$

- Likelihood based approach; use flexible BNP to **reduce error due to model mis-specification**

Ferguson, 1973 | Müller and Mitra, 2013 | Müller and Rodriguez, 2013

- Other estimation techniques can be used.

Xu et al., 2016

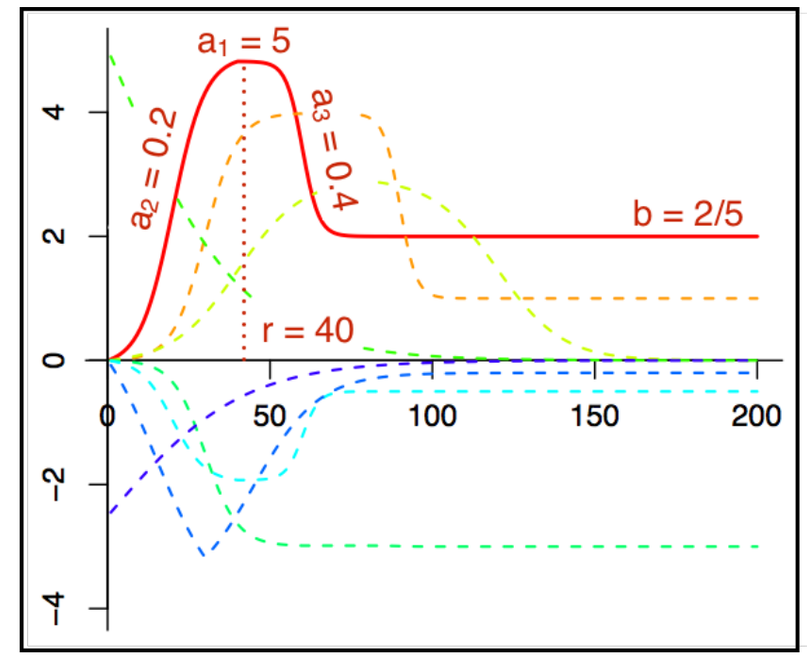
ITR: Additive Treatment Effects

Xu et al., 2016

$$\boldsymbol{y}_i \mid \boldsymbol{a}_i, \boldsymbol{c}_i = \underbrace{u_i(\boldsymbol{c}_i)}_{\text{baseline progression}} + \underbrace{f_i(\boldsymbol{a}_i)}_{\text{treatment responses}} + \underbrace{\boldsymbol{\epsilon}_i}_{\text{noise}},$$

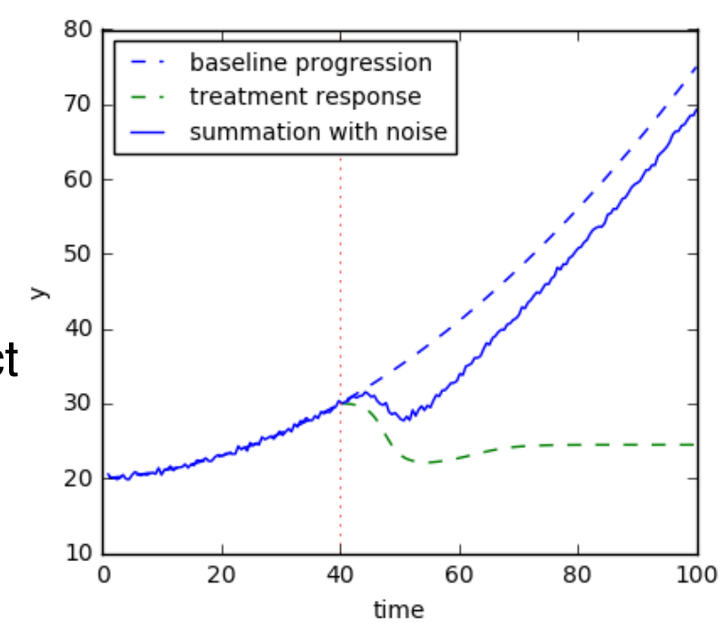
Parametrization for the treatment response curve:

$$g_{id}(t) = \begin{cases} b_0 + \frac{\alpha_{1_{id}}}{1+\exp(-\alpha_{2_{id}}(t-\gamma_{id}/2))}, & 0 \le t < \gamma_{id}, \\ b \cdot g_{\gamma_{id}} + \frac{\alpha_0}{1+\exp(\alpha_{3_{id}}(t-3\gamma_{id}/2))}, & t \ge \gamma_{id}, \end{cases}$$

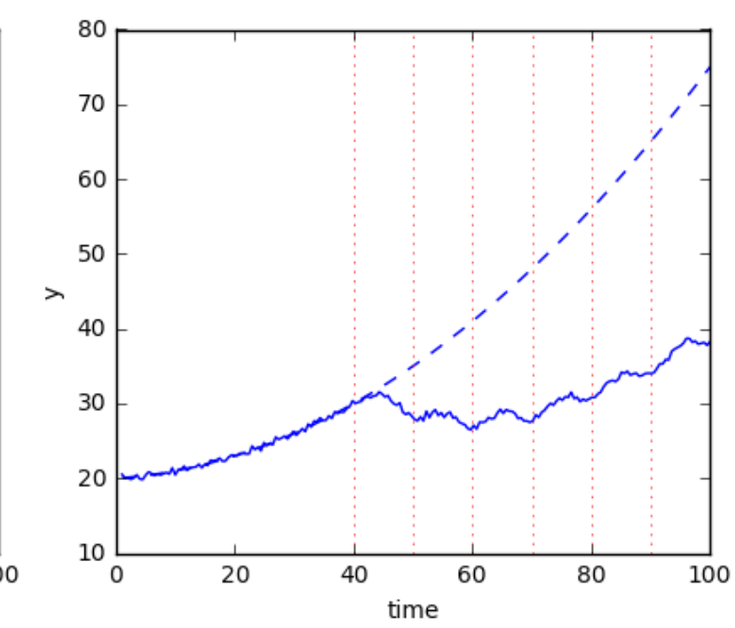


Key parameters:

- a1: peak effect
- a2: how quickly the effect reaches the peak
- a3: how quickly the effect diminish
- r: change point
- b: the ratio of the final effect to the peak effect



(a) A simulated trajectory with one treatment



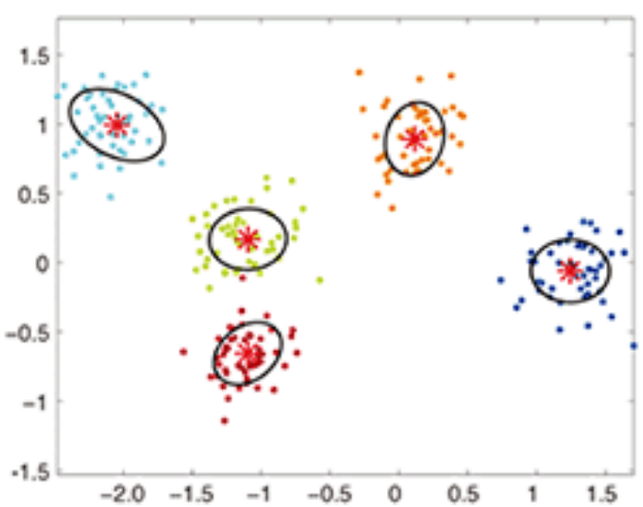
(b) A simulated trajectory with multiple treatments

Choices to reduce error due to model mis-specification

$$\boldsymbol{y}_i \mid \boldsymbol{a}_i, \boldsymbol{c}_i = \underbrace{u_i(\boldsymbol{c}_i)}_{\text{baseline progression}} + \underbrace{f_i(\boldsymbol{a}_i)}_{\text{treatment responses}} + \underbrace{\boldsymbol{\epsilon}_i}_{\text{noise}},$$

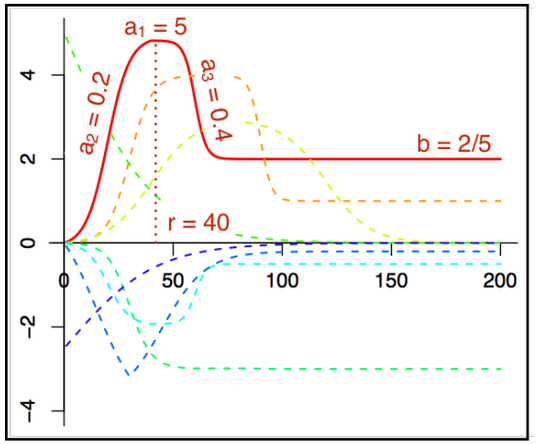
Gaussian Process to flexible model longitudinal traces

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

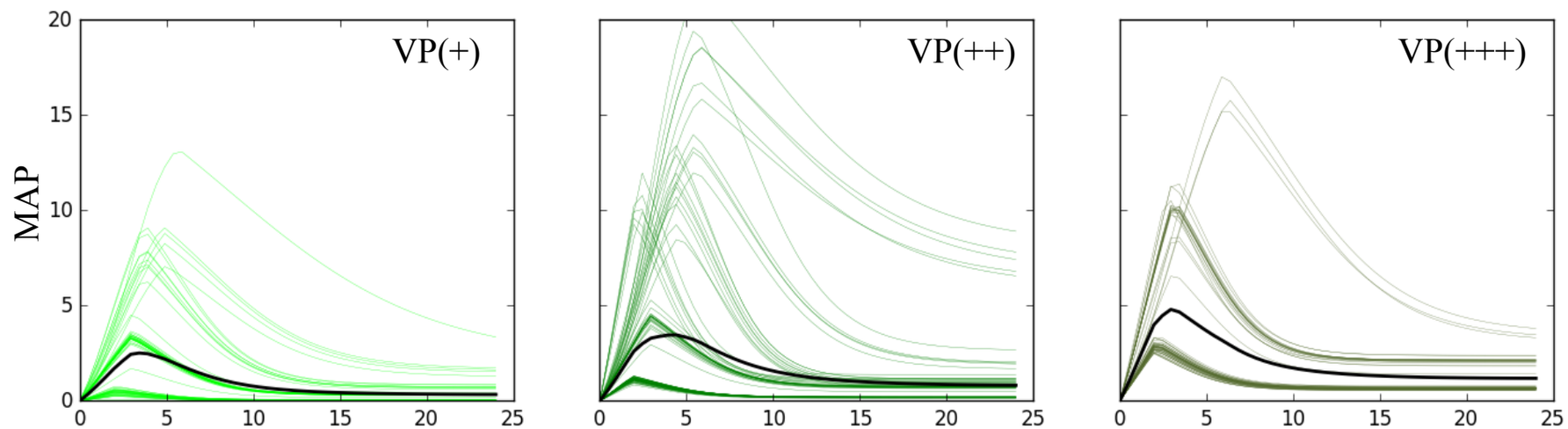


Xu et al., 2016

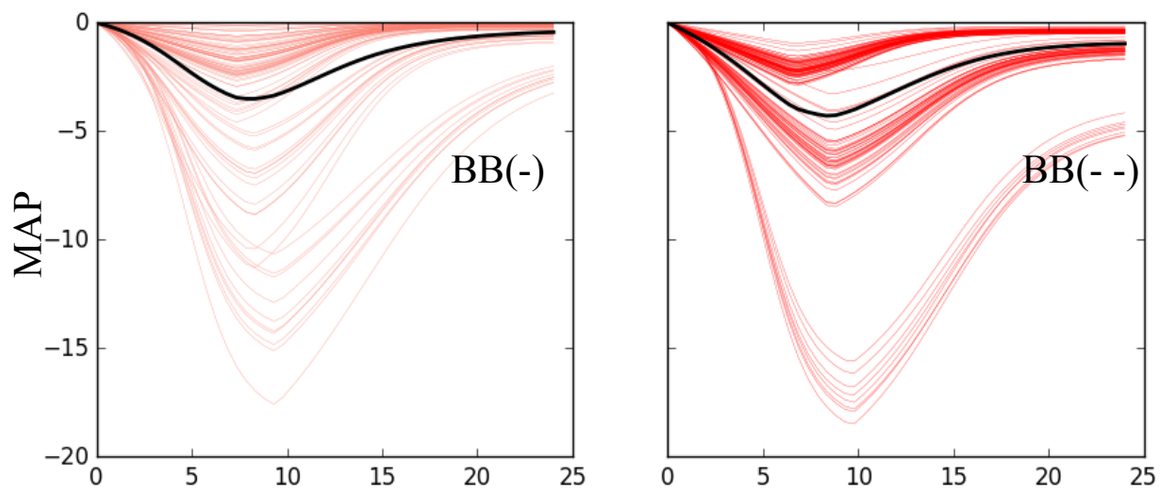
Heterogeneous Treatment Response

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.

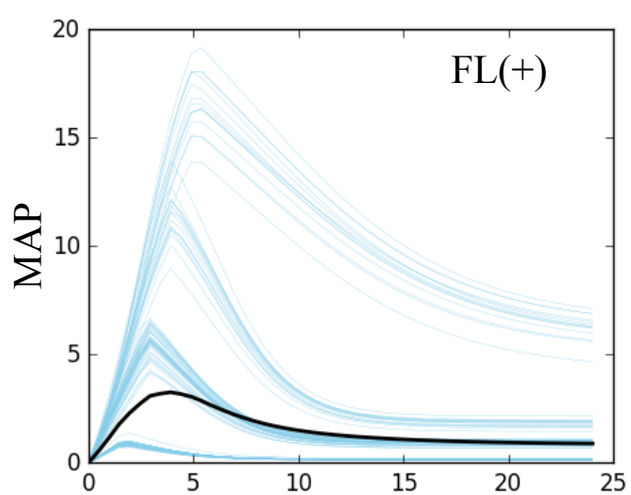
Vasopressor:



Beta-blocker:



Fluid_bolus:



Xu et al., 2016

Overview

- **Part 1—Setting up the problem of Individualization**
  - Example using a chronic disease
  - Simple setting: No Treatment Effects
  - **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Key ideas: Transfer learning, Multilevel modeling

- **Part 2—Estimating Treatment Effects & Individualized Treatment Effects**
  - Example using inpatient data
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP

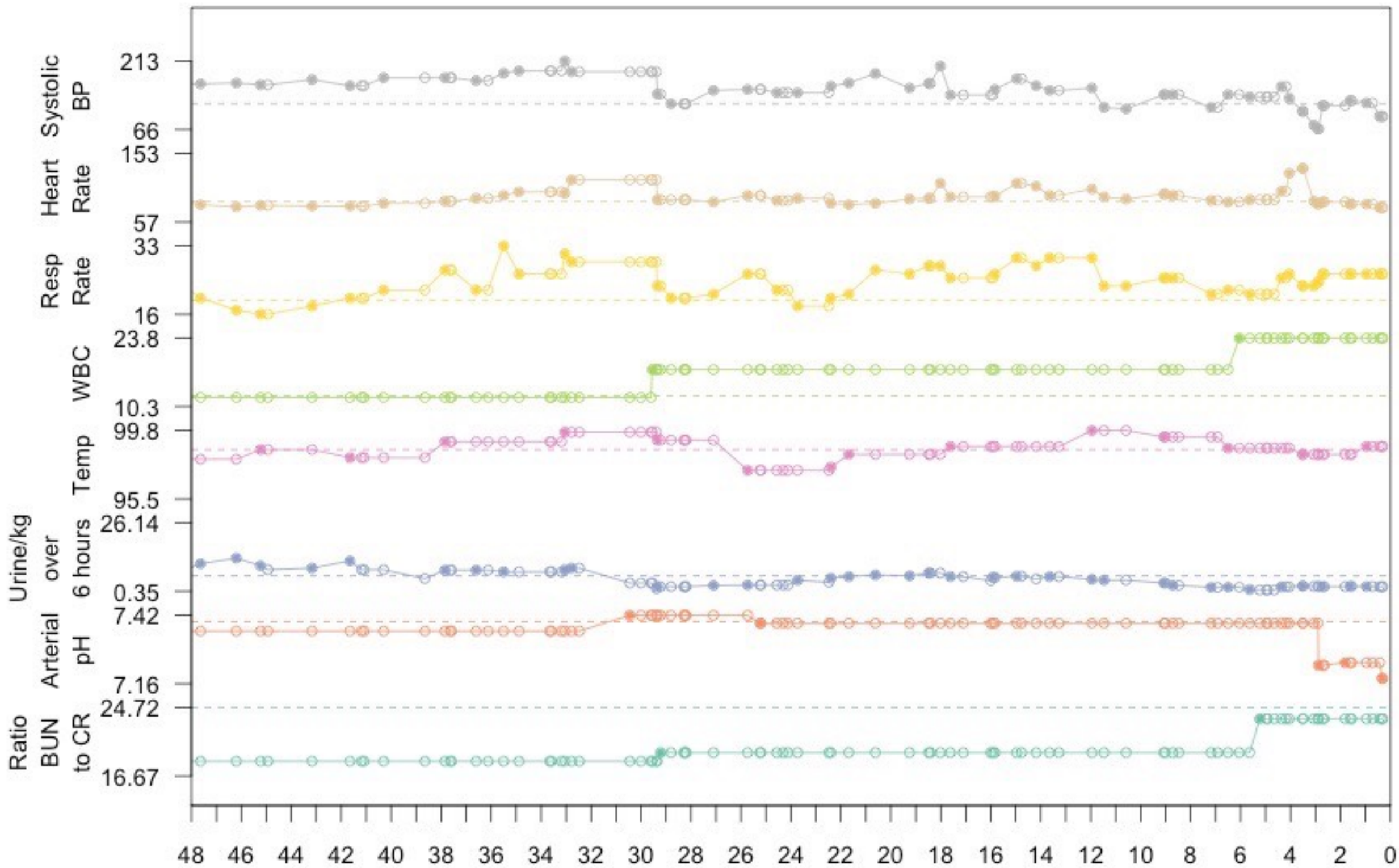
- **Part 3—Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

No Control over Data Collection Process

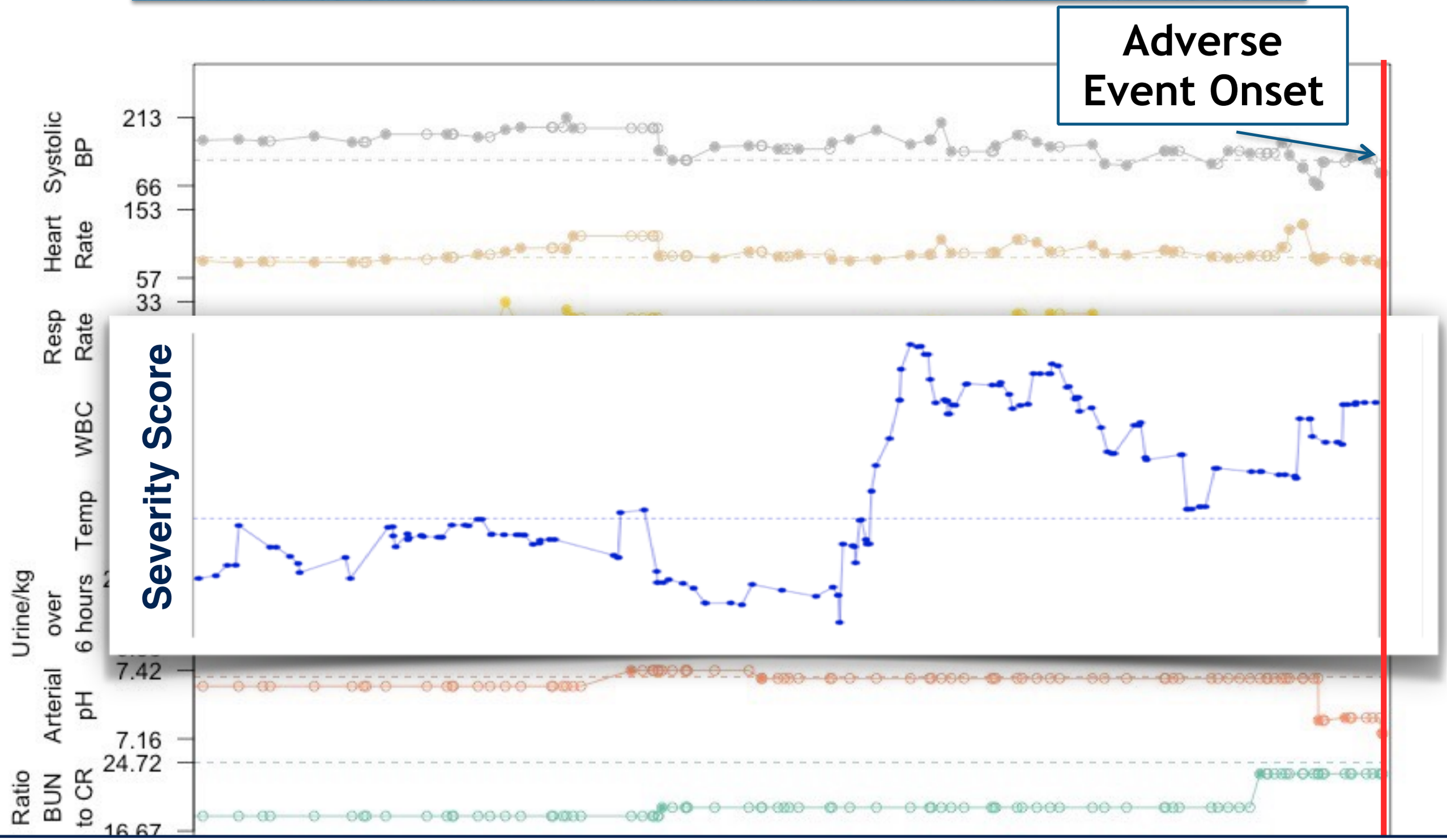
- **Part 4—From Predictions to Treatment Rules**
  - Key ideas: Q-learning, Dynamic Treatment Regimes
  - Connections to Reinforcement Learning

Control over Data Collection Process

Continuous Monitoring

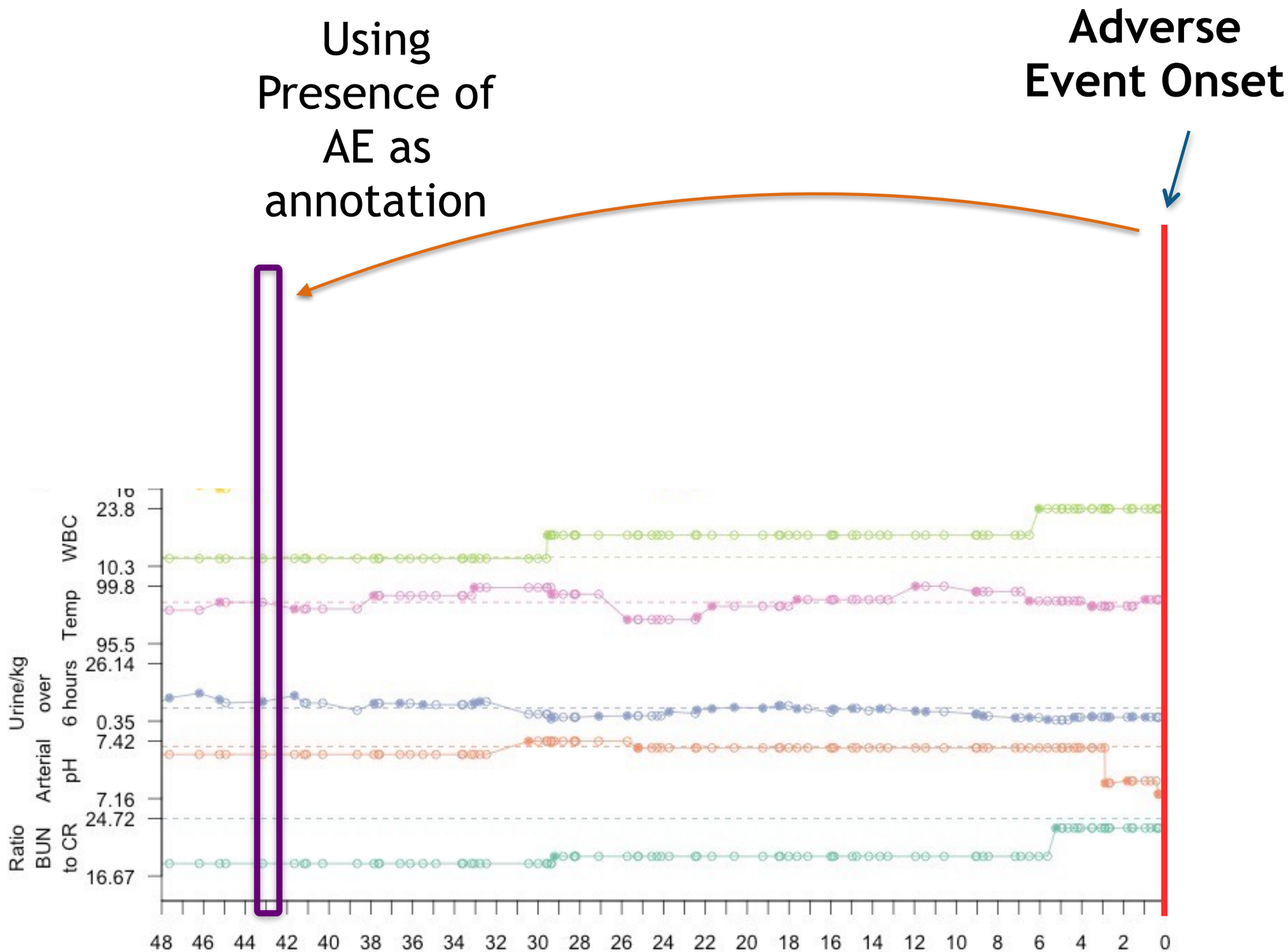


Continuous Monitoring



Predictive Model for Forecasting Downstream Adverse Event

Use supervised learning for distinguishing patients with AE from those without



Pneumonia Severity Index: Risk of Mortality

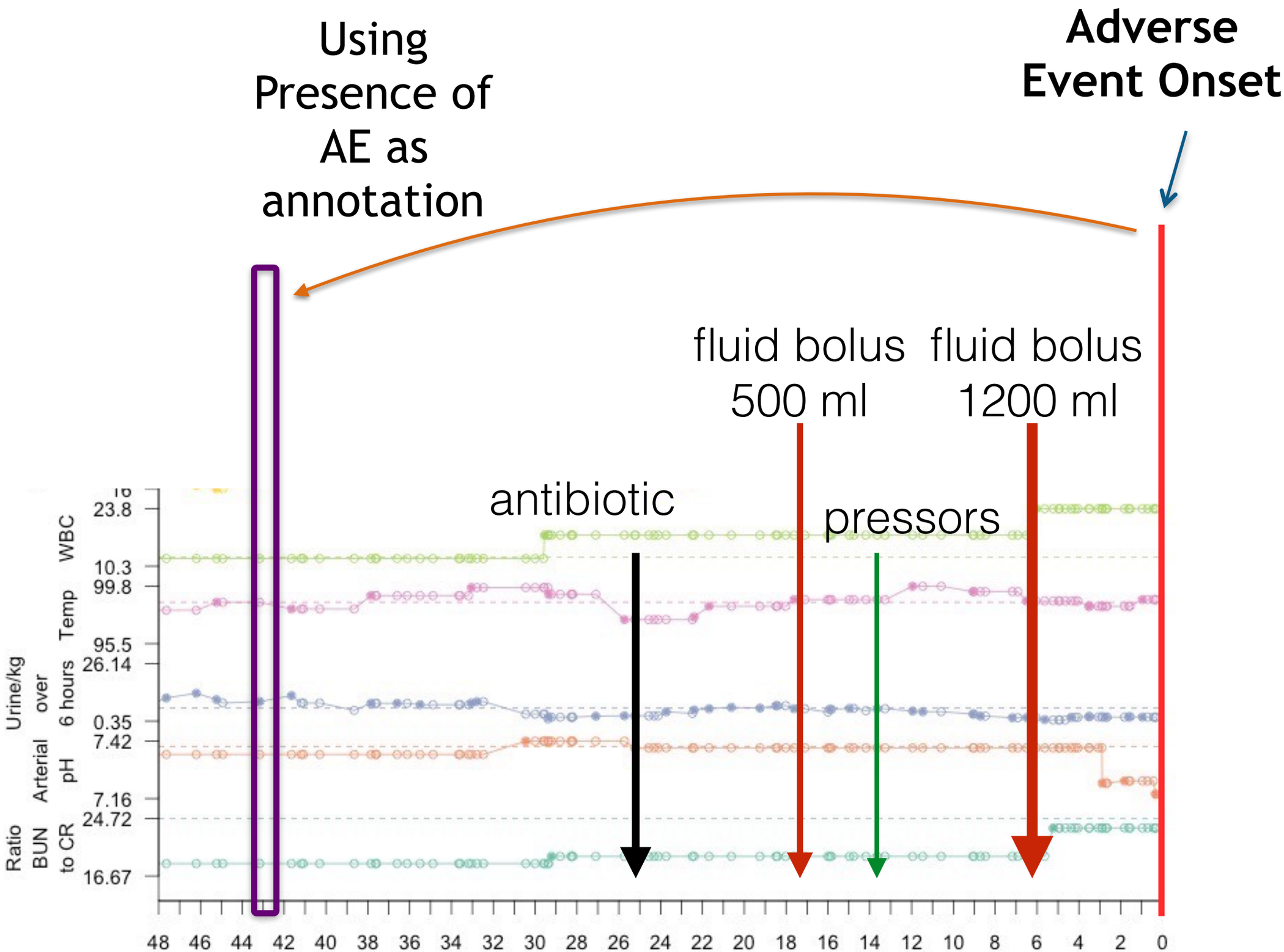
- Identify candidate risk factors
- Learn score and relative weights by regressing against observed mortality

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., 2016

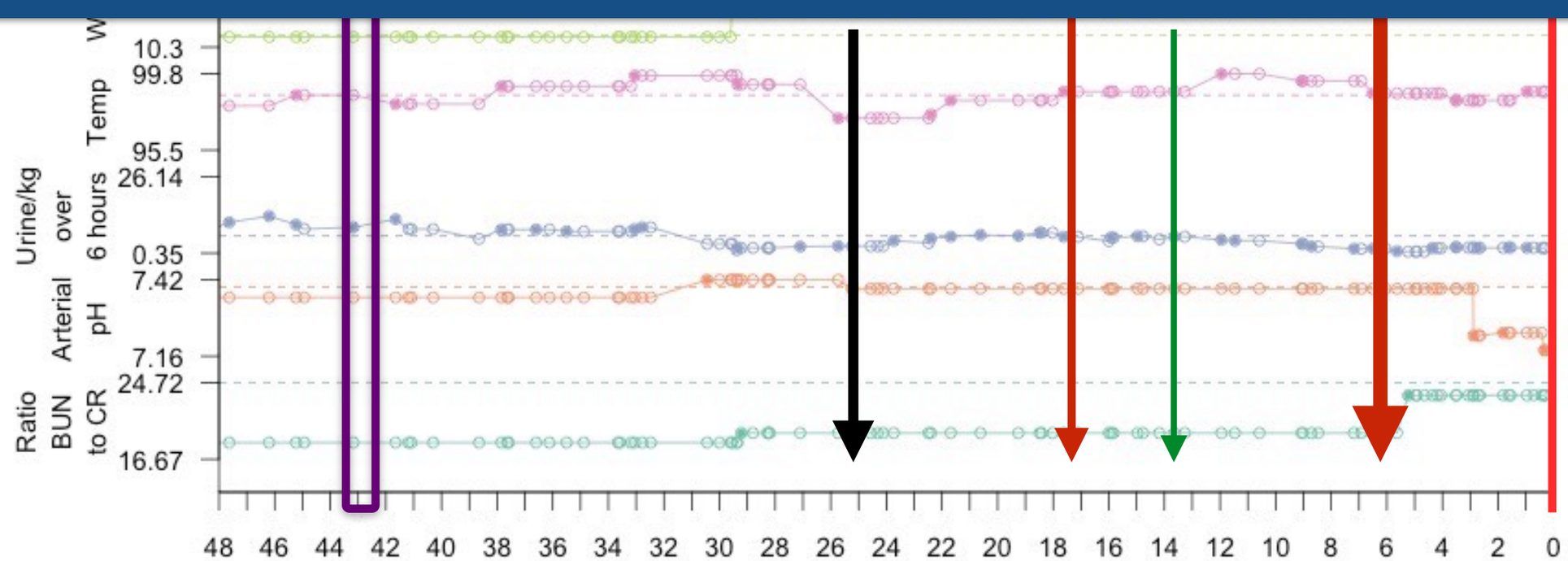
But, interventions *censor* the true label.

Using Presence of AE as annotation

Adverse Event Onset

fluid bolus fluid bolus

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

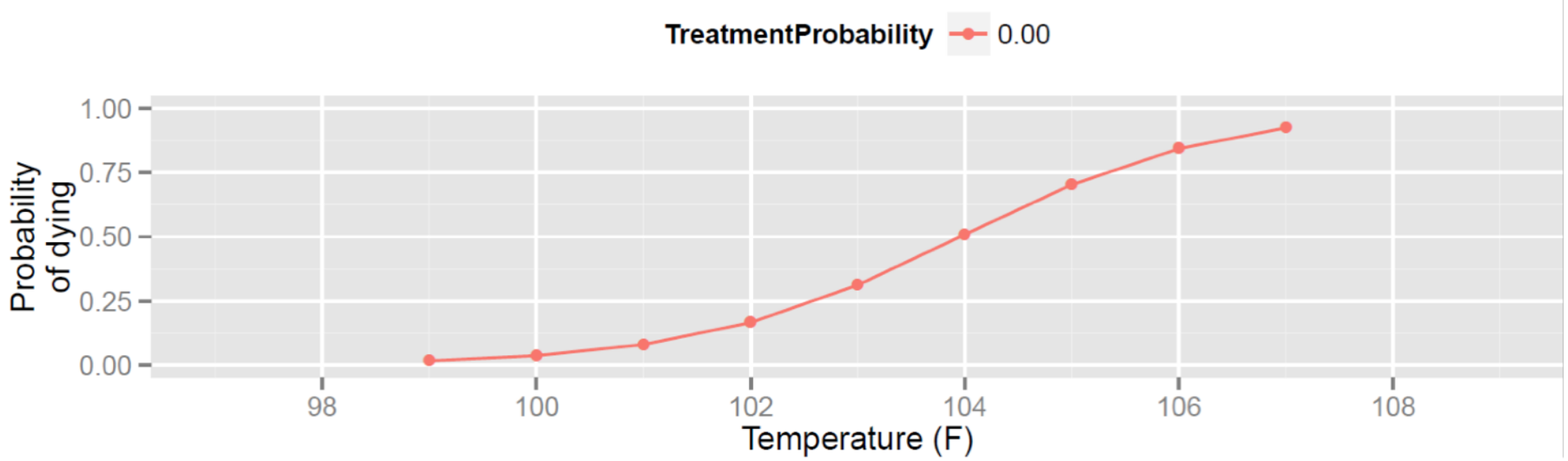


Paxton et al., 2013 **Dyagilev et al., 2016**

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



Dyagilev et al., 2016

Challenge: Learnt Risk Estimates Sensitive to Provider Practice Pattern

Key idea:

- Consider a unit where patients get treated as temperature increases above say, 102 degrees
  - Therefore, *fewer deaths due to rising temperature*
  - As fewer individuals experience death, the **algorithm no longer associates rise in temperature with risk.**

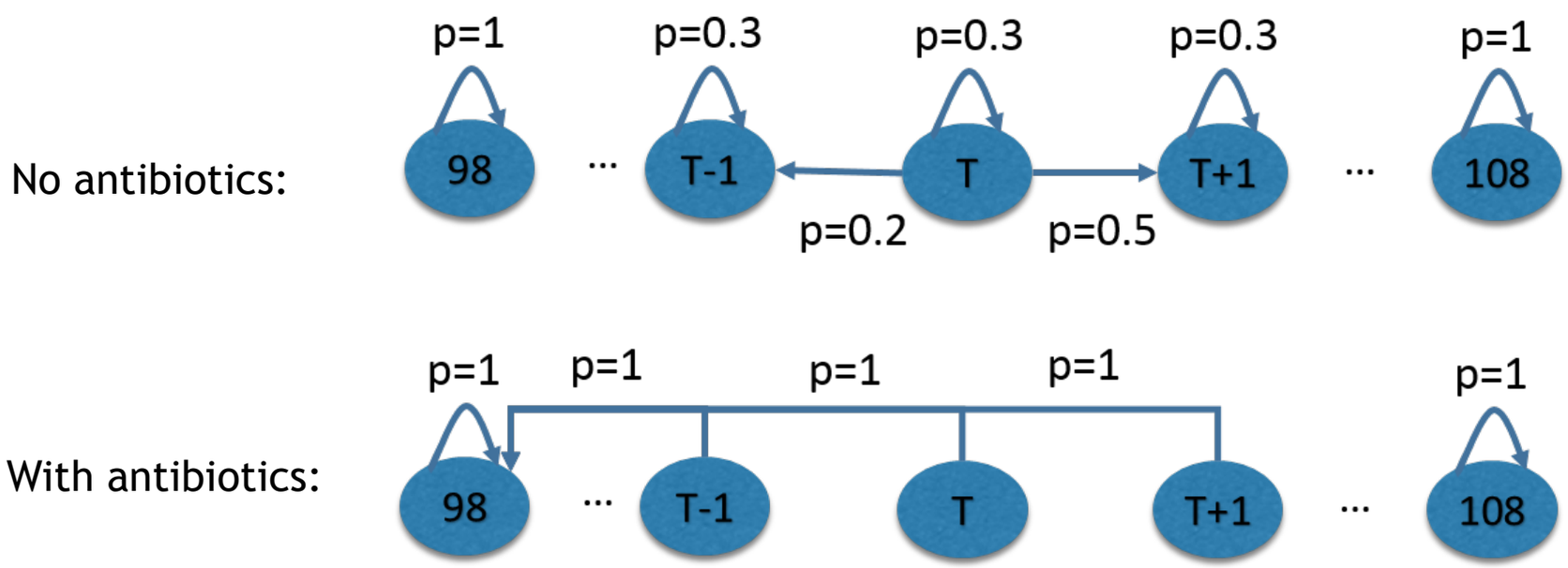
Dyagilev et al., 2016

Bias Due to Interventional Confounds

- Model flu severity; temperature is observed

- Example: Synthetic-Pneumonia
  - If flu, temperature increases unless medicated
  - When medicated, temperature returns to normal
  - At 108 deg F, subject dies

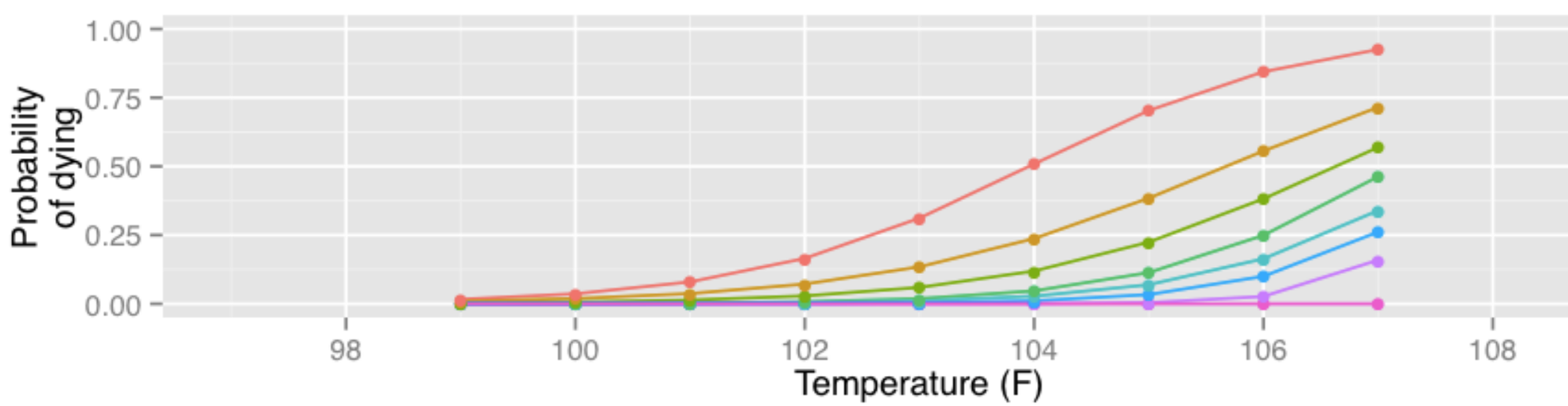
- Consider hospitals with different practice patterns:
  P(med | temperature)



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

Bias Due to Interventional Confounds

- Model flu severity; temperature is observed
- Simulate using Synthetic-Pneumonia model:
  - If flu, temperature increases unless medicated
  - When medicated, temperature returns to normal
  - At 108 deg F, subject dies
- Consider hospitals with different practice patterns: P(med | temperature)



Dyagilev et al., 2016

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	L-DSS
#1	0	0	0	0	0.974	0.973
#2	0.1	0	0.1	0	0.978	0.990
#3	0.1	0	0	0	0.963	0.974
#4	0.3	0	0	0	0.769	0.973
#5	0.3	0	0	0.3	0.510	0.978

Increase probability of treating for rising temperature

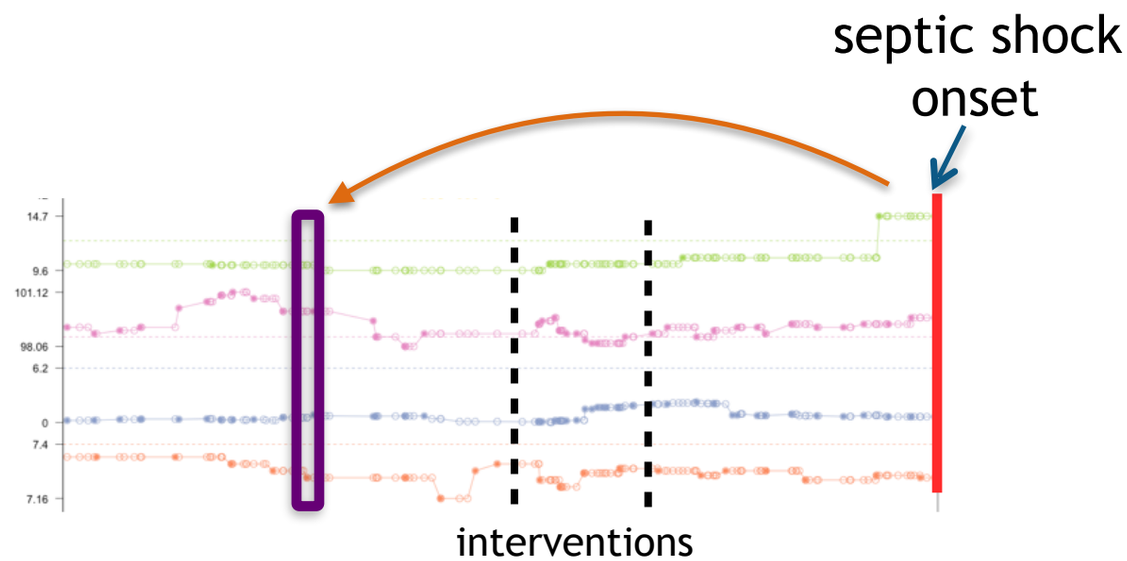
Increasing discrepancy in physician prescription behavior in train vs. test environment

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]

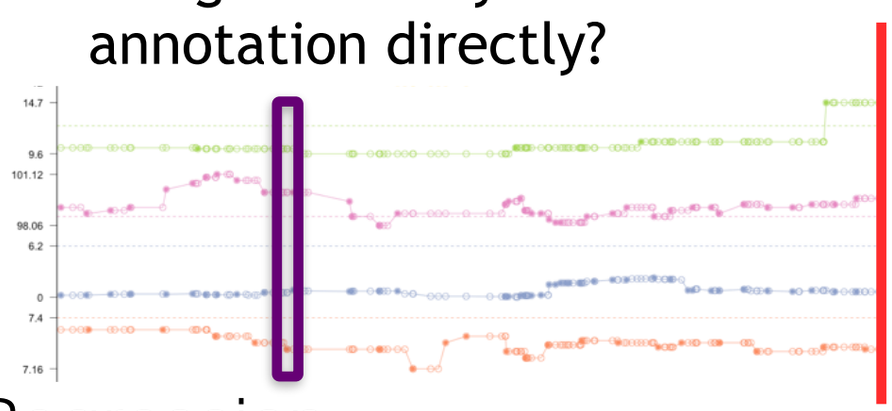
Dyagilev et al., 2016

Alternate forms of training and supervision?



Instead:

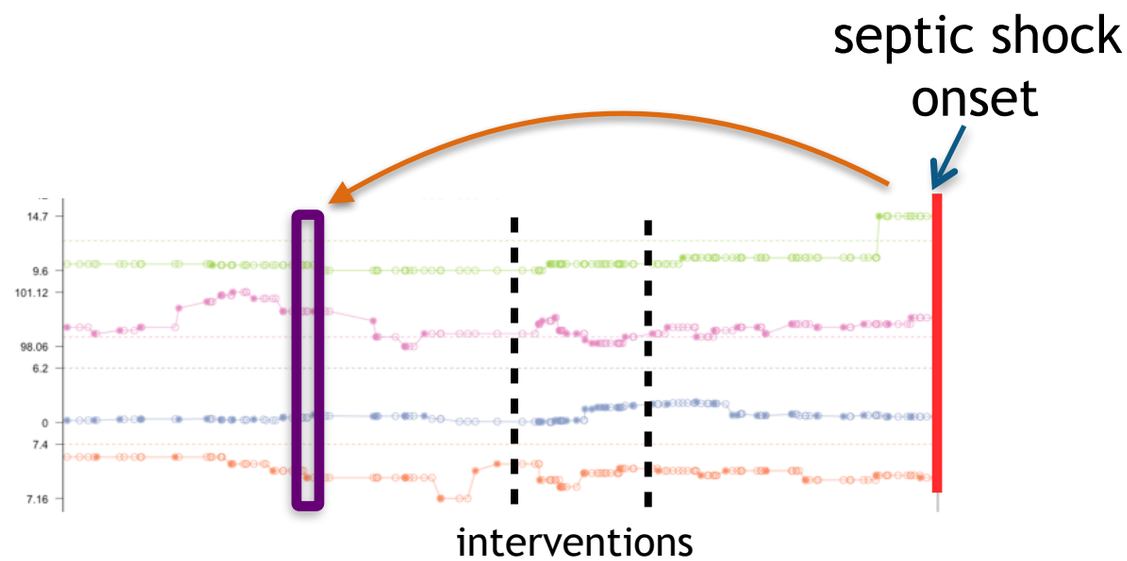
get severity
annotation directly?



Regression

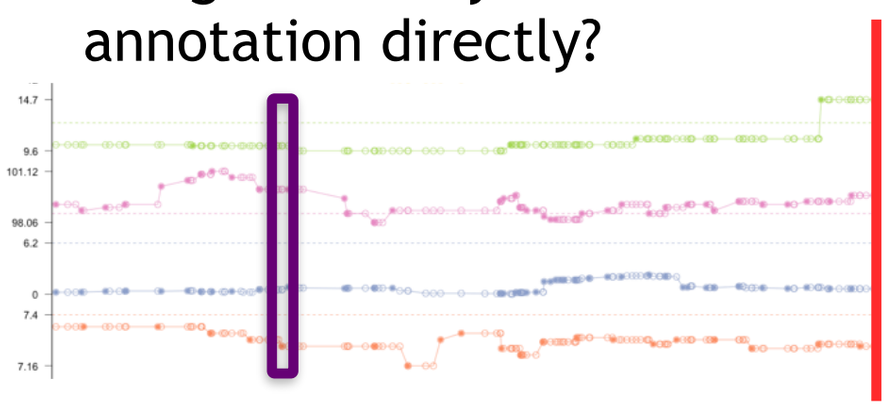
Often not practical because getting these annotations are challenging.

Alternate forms of training and supervision?



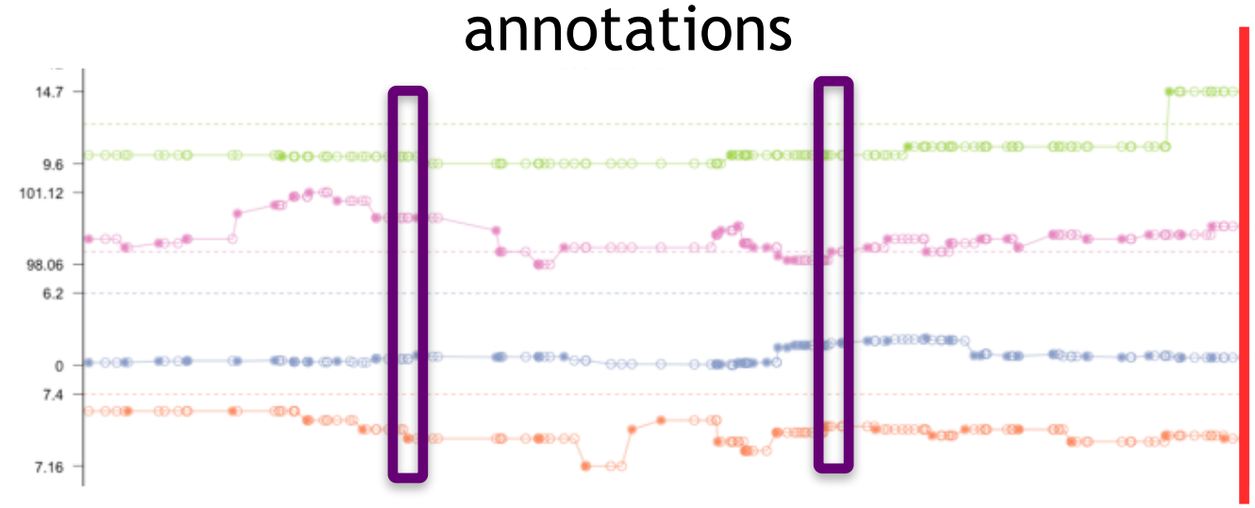
Instead:

get severity
annotation directly?



Regression

compare
severity
annotations



Comparison Pairs: **Dyagilev et al., 2016**

Today: Joint modeling of states and actions

Transportability not always possible: **Bareinboim and Pearl, 2013**

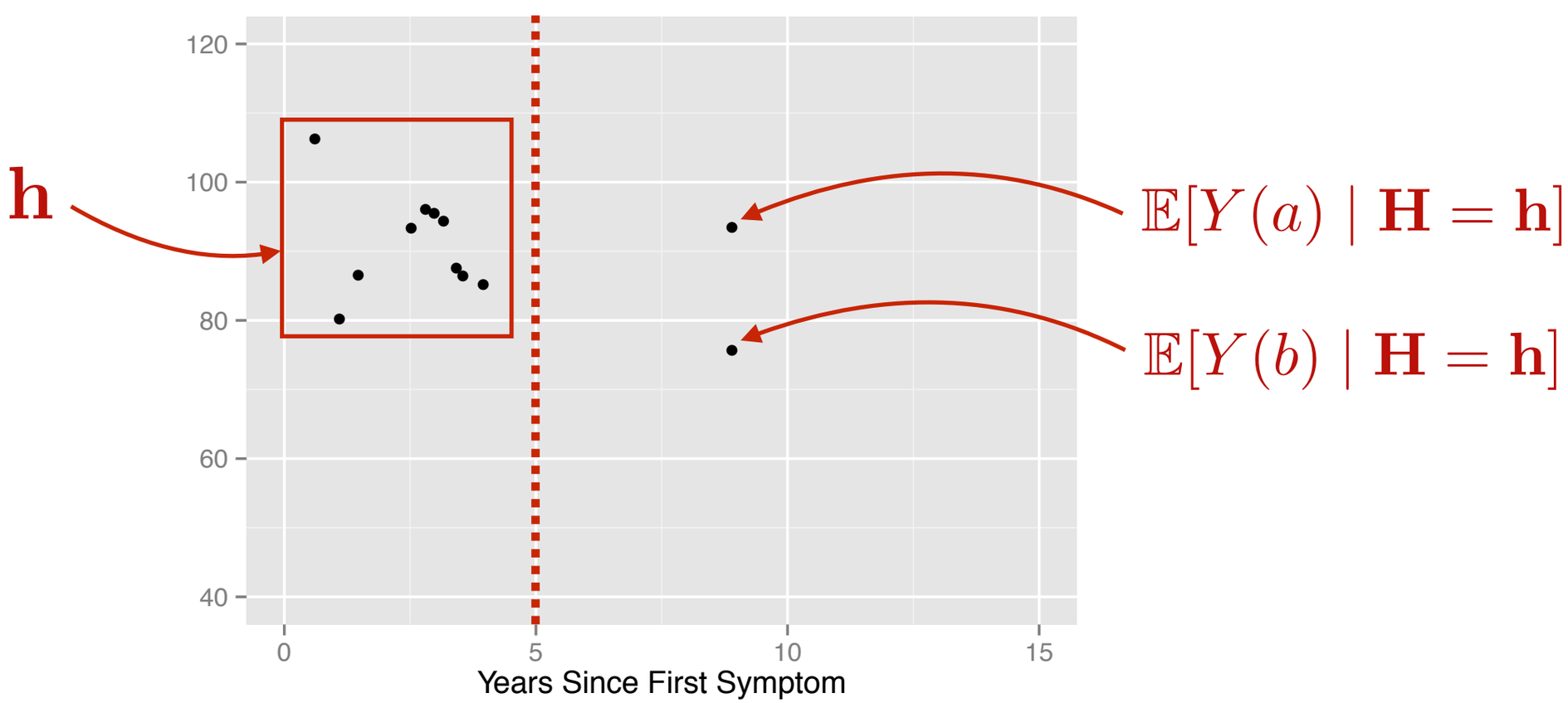
Causal Predictions

- Learnt risk is **conditional** on prescription patterns.
  - Statistical model's predictions may capture correlations that depend on **provider practice**
  - E.g. "treat when temperature rises above 100"
- What we observe is "**what happens 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**" We will refer to this idea as estimating the causal risk.

Bottou et al., 2012

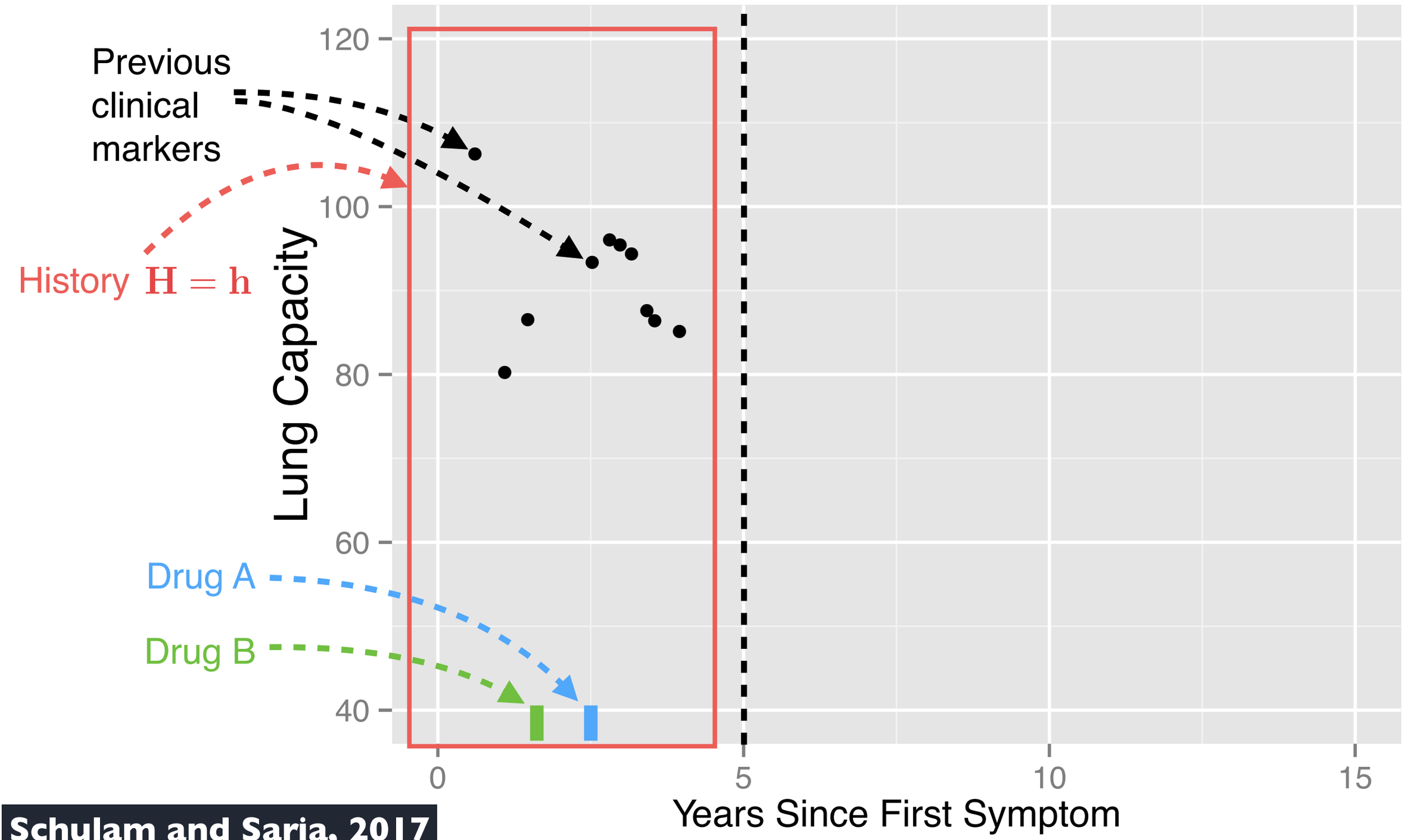
Personalization and Potential Outcomes

- Recall example application from Section 1
- Potential outcomes allow “what if?” reasoning
- To select best treatment for an individual, we can examine expected outcomes under each choice



Personalization and Potential Outcomes

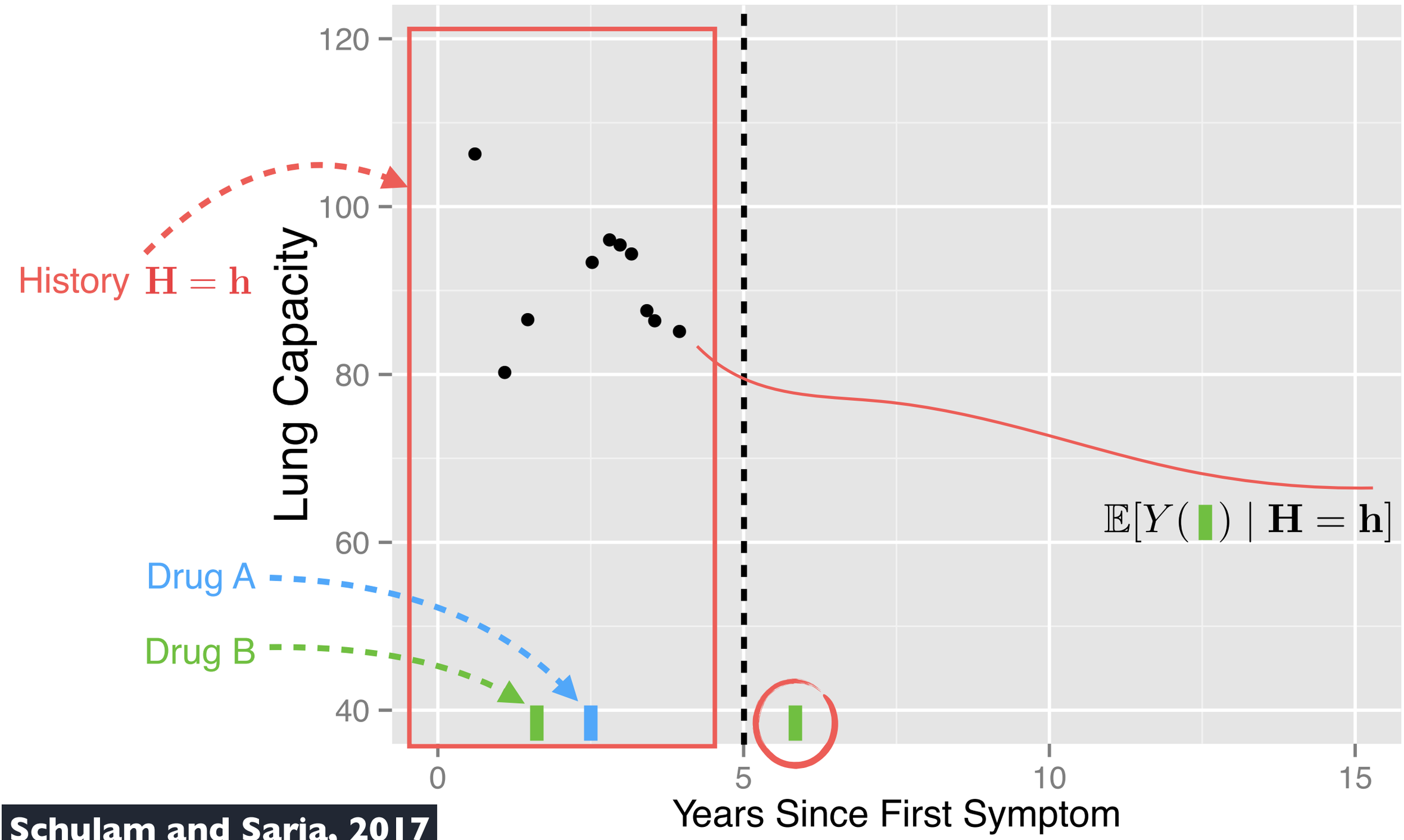
- What is the future **trajectory** under different **sequences of interventions**?



Schulam and Saria, 2017

Personalization and Potential Outcomes

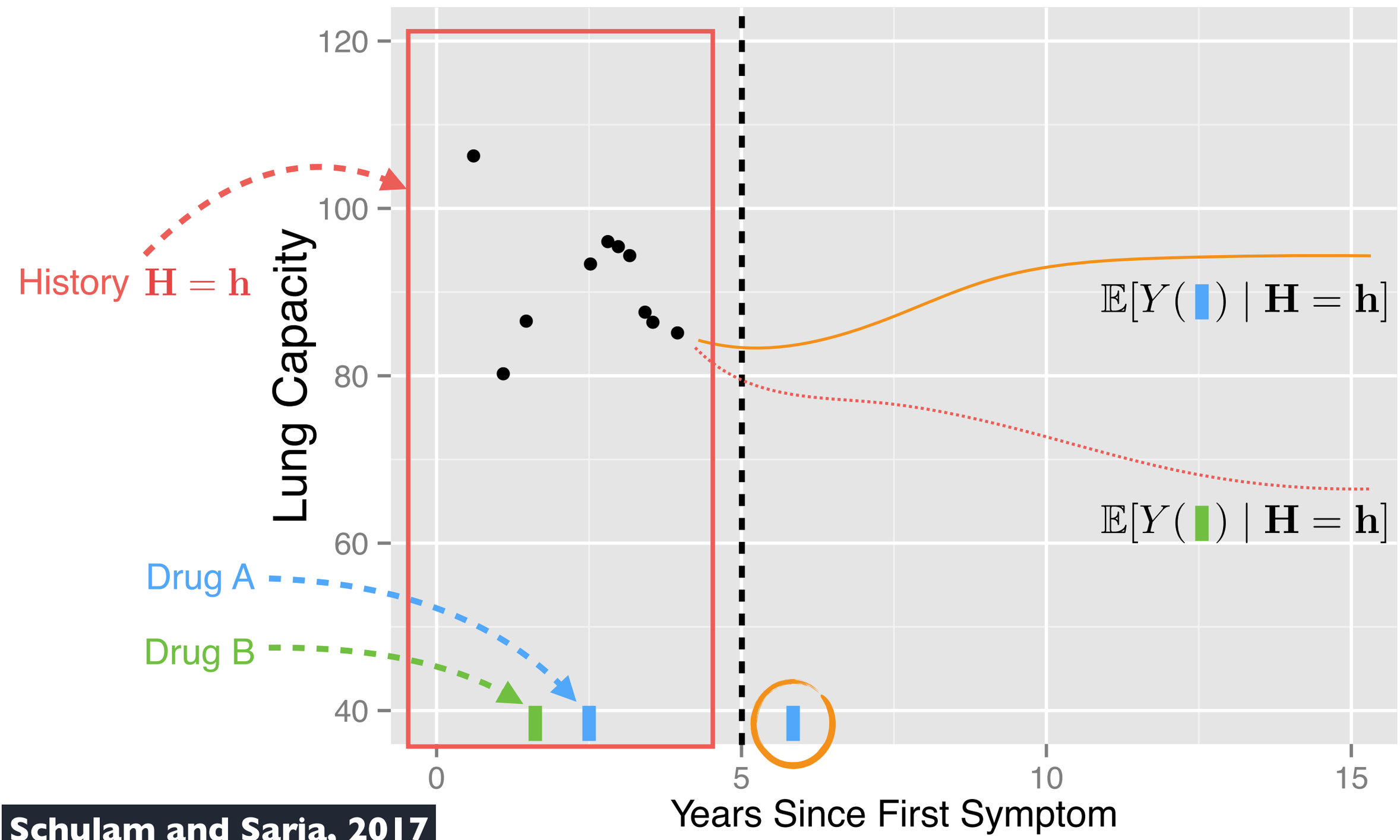
- What if we administer another dose of Drug B?



Schulam and Saria, 2017

Personalization and Potential Outcomes

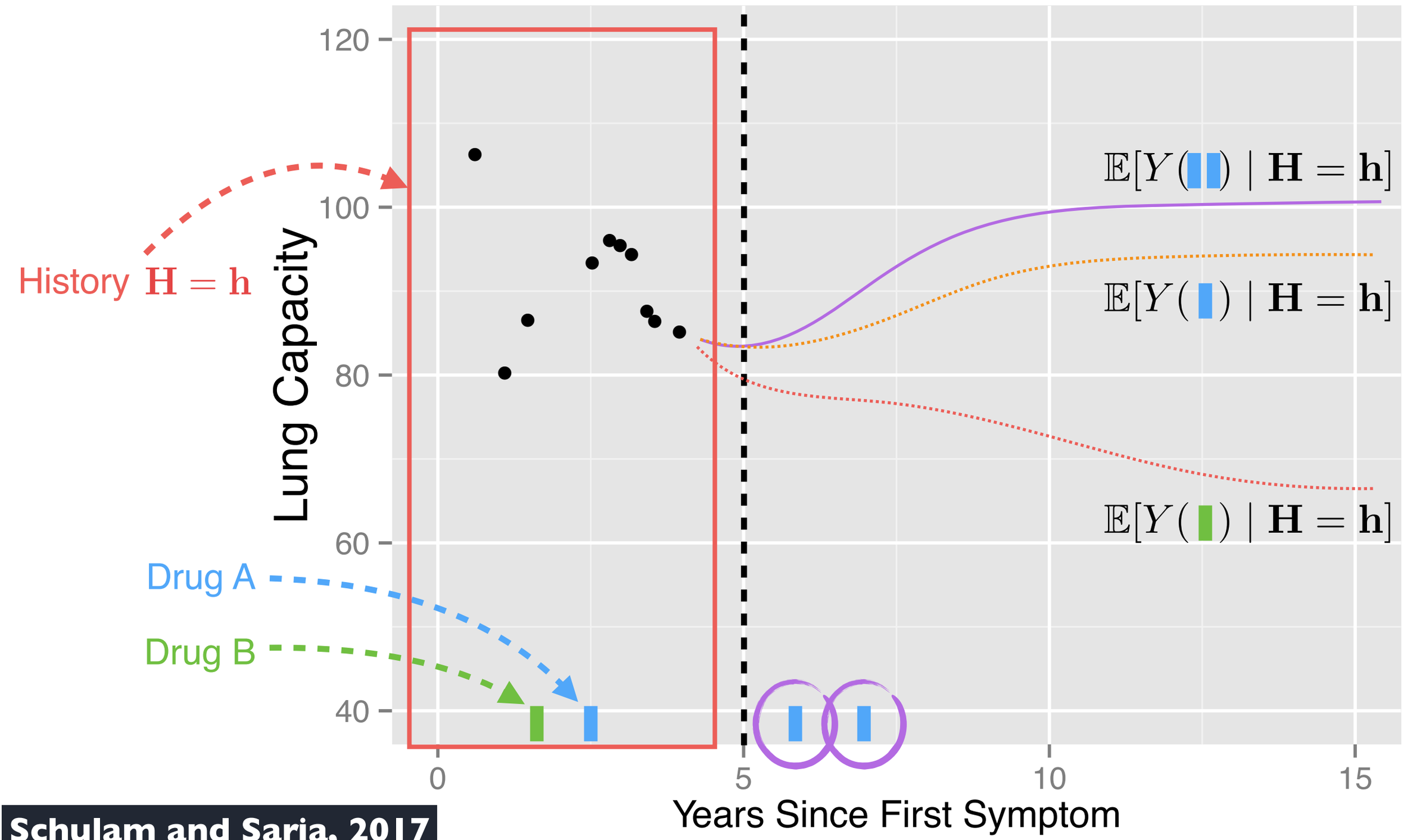
- What about another dose of Drug A?



Schulam and Saria, 2017

Personalization and Potential Outcomes

- What about two sequential doses of Drug A?



Schulam and Saria, 2017

Trajectory-Valued Potential Outcomes

- In the single-treatment, single-outcome case we learned models of the potential outcomes and used them to simulate experimental results

- We want to transplant this idea to the individual level:

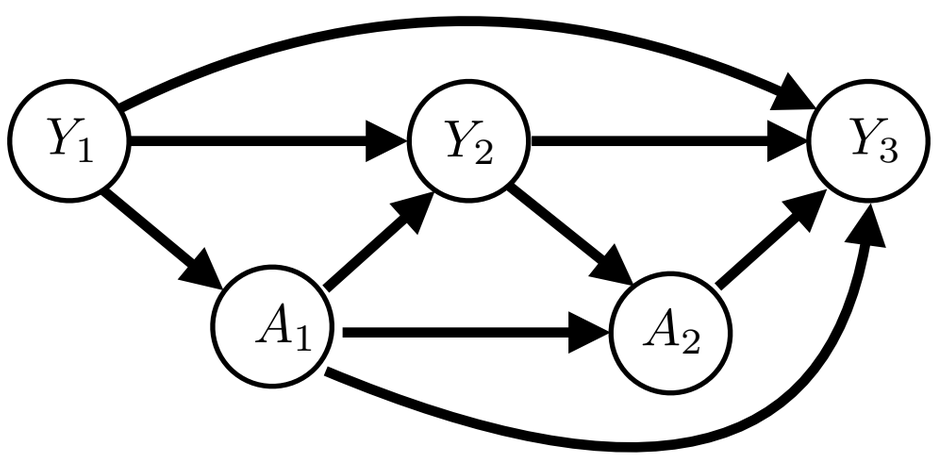
  - Can we learn personalized trajectory-valued potential outcome models?

  - If so, can we use those models to **simulate experiments that investigate the effect of different treatment decisions *for this person***?

Schulam and Saria, 2017

Recall: Sequential Treatment Assignment and Time-Varying Confounding

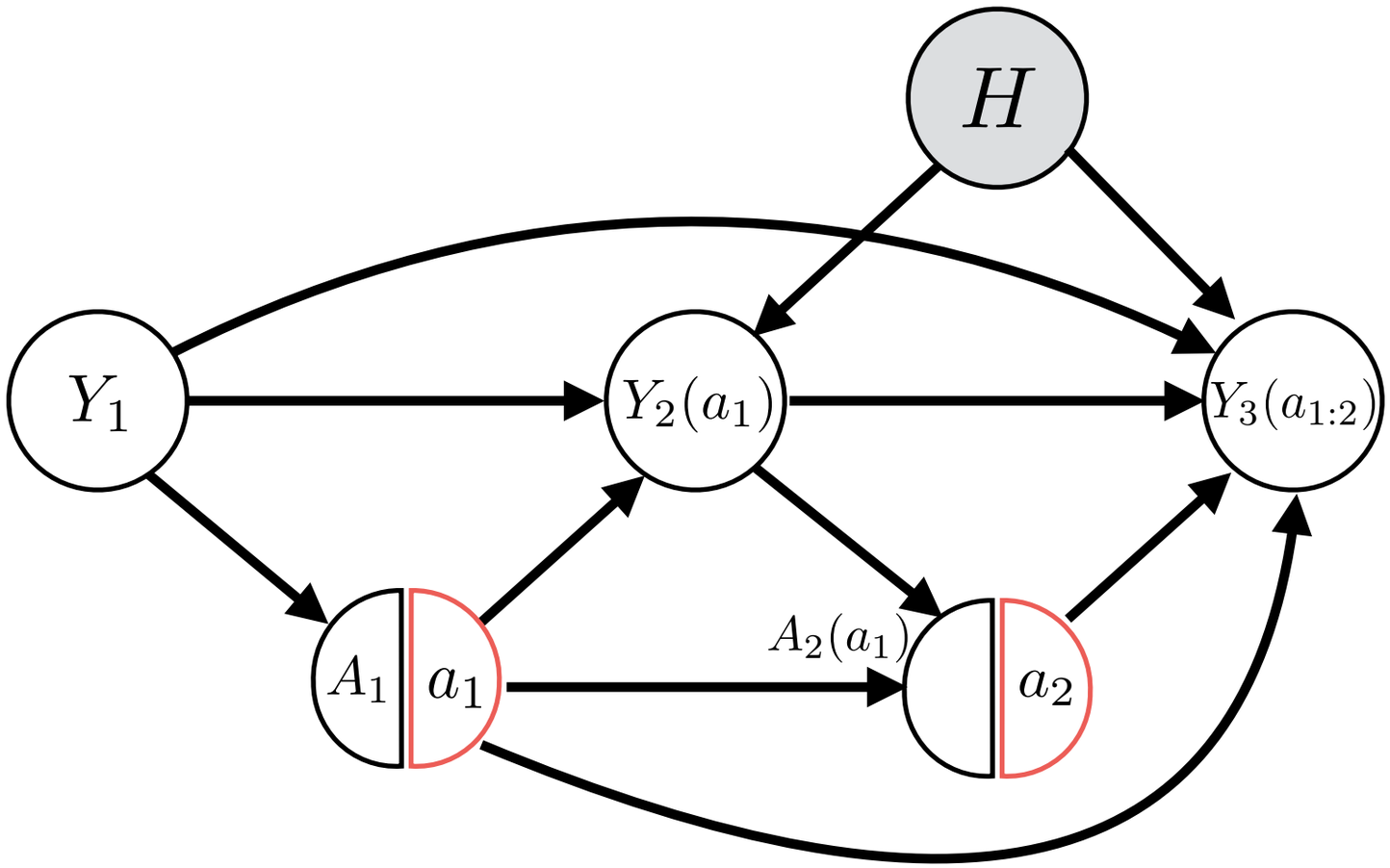
- Interventions and observations are interleaved
  - Intervention effects future observations
    Those observations affect future interventions
    And so on…
  - When can we disentangle to learn unbiased models of potential outcomes?
- Also called time-varying confounding.



Robins 1986

Recall: SWIG for Sequential Setting

- Assumptions: (1) Consistency, (2) Sequential Ignorability (NUC)



- 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

Handling Irregularity

- In an irregular trace (i.e. sequence of interleaved actions and observations), there can be multiple observations between actions:

$$\mathbf{h}_i = [(y_{i1}, t_{i1}), (a_{i1}, \tau_{i1}), (y_{i2}, t_{i2}), (y_{i3}, t_{i3})].$$

- We can handle irregularly sampled observations and treatments in a similar way [Part 1 and Part 2]

- We assume measurements are **missing at random** i.e. the choice of when to measure depends on the past observed data [Recall from Part 1]

Factoring Irregular Traces

- We can still factor these traces as we would regularly sampled traces (see paper for details)

- Define:

  - $\bar{\mathbf{y}}_k$ to be the observations prior to action k

  - $\bar{\mathbf{a}}_k$ to be the actions taken prior to action k

  - $\mathbf{y}_k$ to be observations after action k, but before k+1

- Then we can factor an arbitrary trace:

$$p(\mathbf{h} \mid \mathbf{x}) = p(\mathbf{y}_0 \mid \mathbf{x}) \prod_{k=1}^{m} p(a_k, \tau_k \mid \bar{\mathbf{y}}_k, \bar{\mathbf{a}}_k, \mathbf{x}) p(\mathbf{y}_k \mid \bar{\mathbf{y}}_k, a_k, \tau_k, \bar{\mathbf{a}}_k, \mathbf{x}),$$

Schulam and Saria, 2017

Irregular Traces and Functional Potential Outcomes

- Assuming Consistency and Sequential NUC (see paper for details)

$$p(\mathbf{y}_k \mid \bar{\mathbf{y}}_k, a_k, \tau_k, \bar{\mathbf{a}}_k, \mathbf{x}) = p(\mathbf{y}_k(a_k, \tau_k) \mid \bar{\mathbf{y}}_k, \bar{\mathbf{a}}_k, \mathbf{x})$$

- Therefore can maximize probability of irregular trace:

$$p(\mathbf{h} \mid \mathbf{x}) = p(\mathbf{y}_0 \mid \mathbf{x}) \prod_{k=1}^{m} p(a_k, \tau_k \mid \bar{\mathbf{y}}_k, \bar{\mathbf{a}}_k, \mathbf{x}) p(\mathbf{y}_k \mid \bar{\mathbf{y}}_k, a_k, \tau_k, \bar{\mathbf{a}}_k, \mathbf{x}),$$

- Policy is unknown, but assumed to be distinct so we can ignore the treatment policy terms when learning functional potential outcome models

Recall:

- 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

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

Schulam and Saria, 2017

Modeling Irregular Traces

- Many different ways to model conditional distributions over markers (green component in last slide)

- One example: Gaussian process

$$\mathrm{GP}(\ m(\cdot; \mathbf{a}, \mathbf{x}), k(\cdot, \cdot)\ )$$

Mean function depending on covariates and sequence of treatments

Covariance function independent of treatments

Schulam and Saria, 2017

Modeling Irregular Traces

- Many different ways to model conditional distributions over markers (green component in last slide)

- One example: Gaussian process

$$\mathrm{GP}(\ m_i(\cdot; \mathbf{a}, \mathbf{x}), k_i(\cdot, \cdot)\ )$$

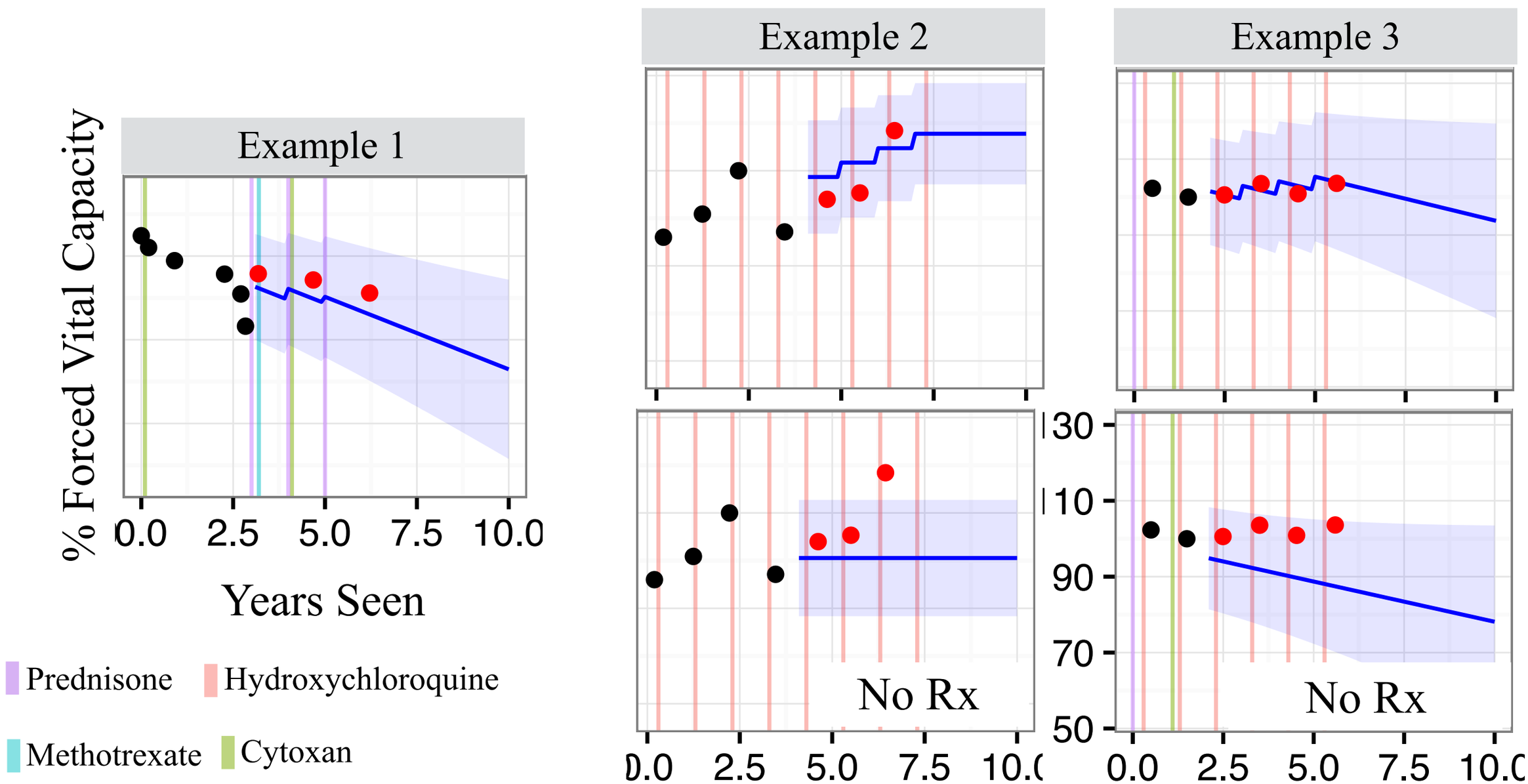
- Recall individualization approach from Part 1:

$$y_{ij} | \vec{x}_{ip}, z_i, b_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\ \vec{x}_{ip}}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) individual}} + \underbrace{f_i(t_{ij})}_{\text{(D) structured noise}}, \sigma^2 \right)$$

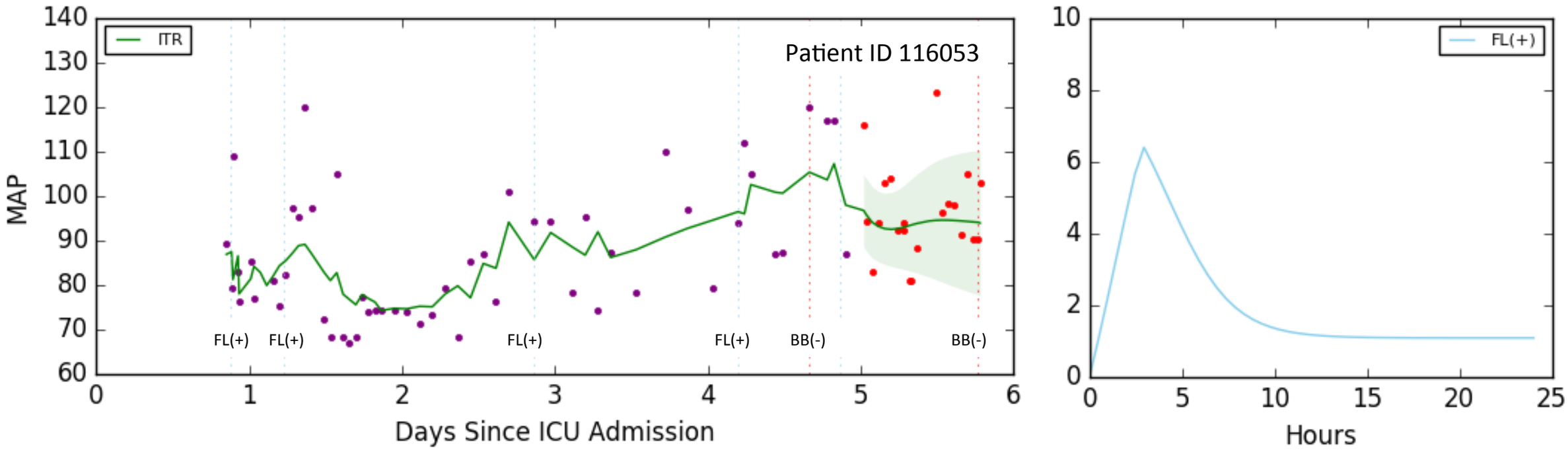
Schulam and Saria, 2017

Example: Lung Disease Trajectories

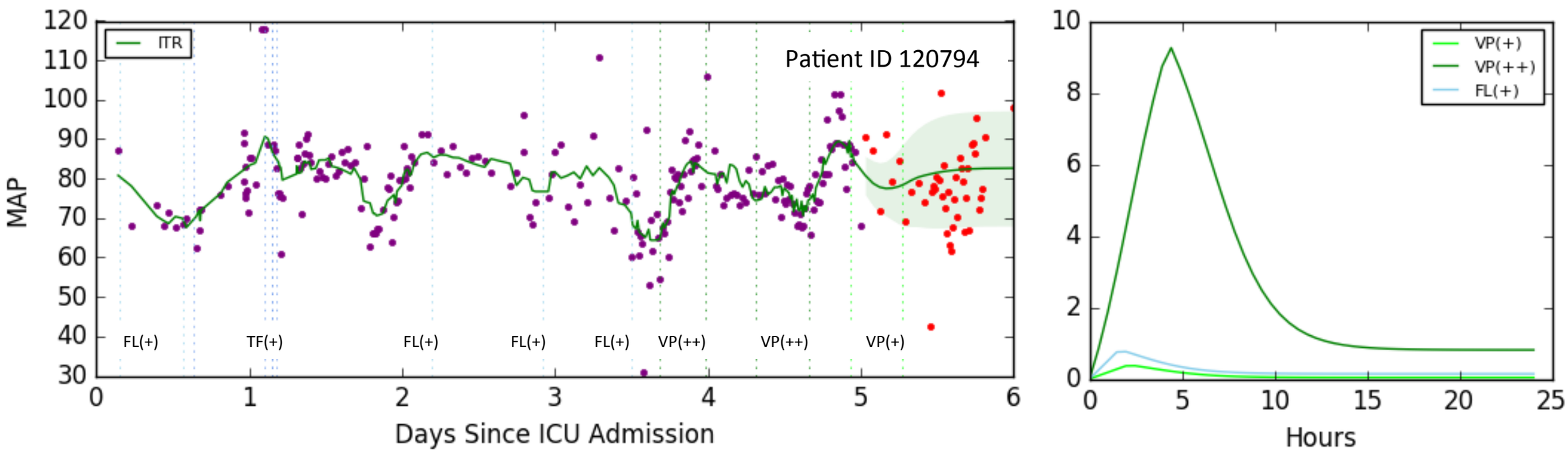
- Using previous lung disease progression patterns and learning from response to treatment, we can predict how individuals will respond to treatment and how they will progress when treatment is no longer given



Predicting trajectories for Targeting Treatments in Critically Ill Patients



(a) Example Trajectory of MAP and Treatment Response Curve of Fluid (FL)



(b) Example Trajectory of MAP and Treatment Response Curves of Vasopressor (VP) and Fluid (FL)

Xu et al., 2016 **Liu et al., 2016**

Caveats in Practice and Discussion

- Estimates of the individual components within the statistical model may not be good enough based on available data
  - Not enough data to train from.
  - Available measurements are not predictive.

- Inferences are correct assuming **no model mis-specification**.
  - Important aspect of causal modeling is getting your causal assumptions right.
    - Think hard about the problem—> avoids the chance of model mis-specification or making incorrect assumptions.
  - Semi-parametric or flexible nonparametric strategies are helpful here.
  - Methods to check sensitivity to assumption (e.g., posterior predictive checks)

- Driving modeling decisions based on practical utility
  - Decisions are made with a human in the loop.
    - Transparency does not have to be interpreted as the use of a linear model or a decision tree.
    - Estimating intermediate quantities that are interpretable or can serve as validation can be useful (e.g., subpopulation, individual-specific deviations)

- Need ways to monitor performance over time.

Overview

- **Part 1—Setting up the problem of Individualization**
  - Example using a chronic disease
  - Simple setting: No Treatment Effects
  - **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Key ideas: Transfer learning, Multilevel modeling

- **Part 2—Estimating Treatment Effects & Individualized Treatment Effects**
  - Example using inpatient data
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP

- **Part 3—Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

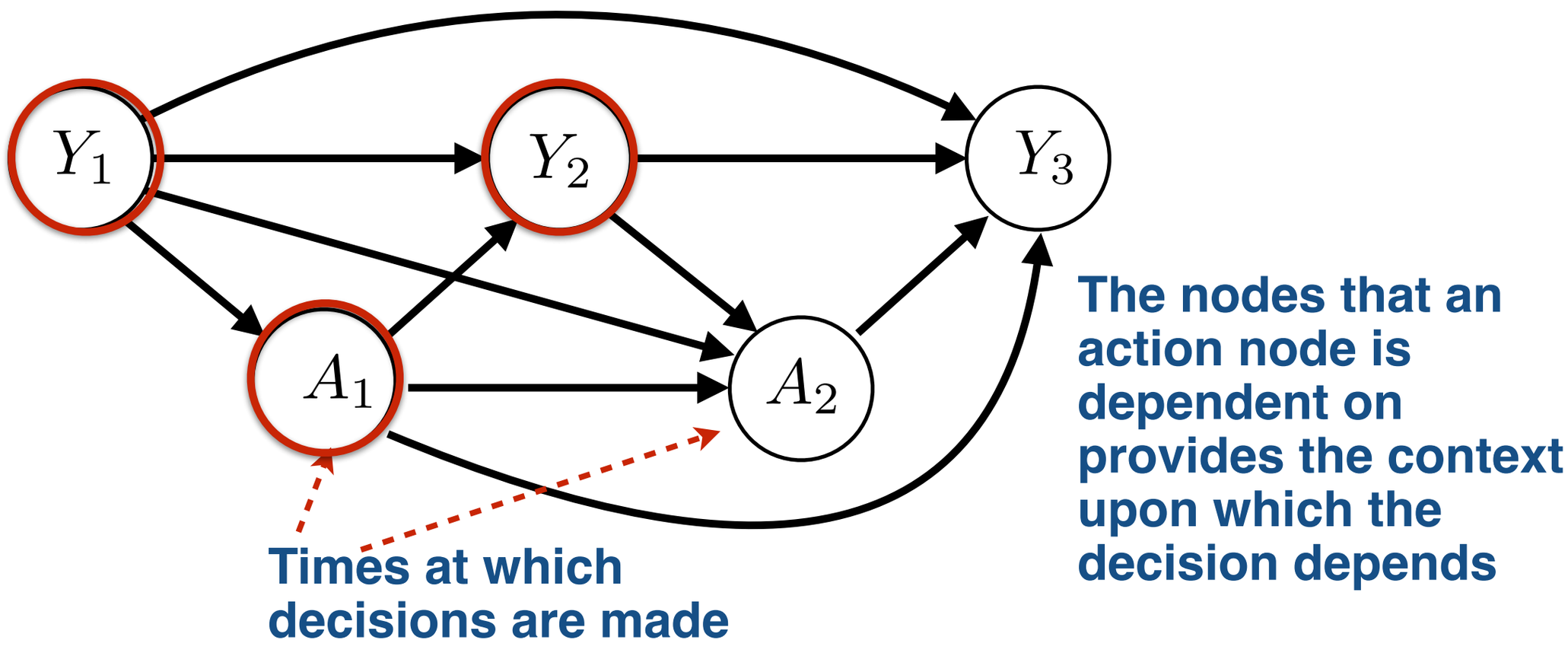
No Control over Data Collection Process

- **Part 4—From Predictions to Treatment Rules**
  - Key ideas: Q-learning, Dynamic Treatment Regimes
  - Connections to Reinforcement Learning

Control over Data Collection Process

Sequential Decision Making

- **A mapping of states to actions**
  - In reinforcement learning, this is called a *sequential policy*
  - In treatment planning, sequential policies called *dynamic treatment regime*
    - States are functions of an individual's clinical history, and the policy maps these histories to actions.



Sequential Treatments

- **A mapping of states (context) to actions**
  - In reinforcement learning, this is called a *sequential policy*
  - In statistics, it is called a *dynamic treatment regime*

- **To obtain such a policy,**
  - we can use model based or model-free methods
  - we use learn by either interacting with the world or learn from offline data.

- **Loosely speaking,**
  - model-based learns a dynamical model of the system (e.g., an MDP)—> as a by-product, also make predictions
  - for model-free methods, you evaluate the policy directly using traces

Review: **Paduraru et al., 2013**

Learning by Interacting with the World

- **Basic Q-learning algorithm**

Q-function or the **action-value function**

$$Q(s,a)=r(s,a)+\gamma max_{a'}(Q(s',a'))$$

$$r(s,a)=\text{Immediate reward}$$

$$\gamma=\text{relative value of delayed vs. immediate rewards (0 to 1)}$$

$$s'=\text{the new state after action a}$$

$$a,a':\text{actions in states } s \text{ and } s', \text{respectively}$$

Selected action:

$$\pi(s)=argmax_a Q(s,a)$$

Watkins 1989

Initialize Q-functions and update as you explore.

Learning by Interacting with the World

- **Basic Q-learning algorithm**

$$Q(s,a)=r(s,a)+\gamma max_{a'}(Q(s',a'))$$

$r(s,a)$ = Immediate reward

γ = relative value of delayed vs. immediate rewards (0 to 1)

s' = the new state after action a

a, a': actions in states s and s', respectively

Selected action:

$$\pi(s)=argmax_a Q(s,a)$$

Watkins 1989

$$P(a_i|s)=\frac{k^{\hat{Q}(s,a_i)}}{\sum_j k^{\hat{Q}(s,a_i)}}$$

Review:

Ghavamzadeh et al., 2015

Safe Reinforcement Learning

- Two broad approaches to safe RL

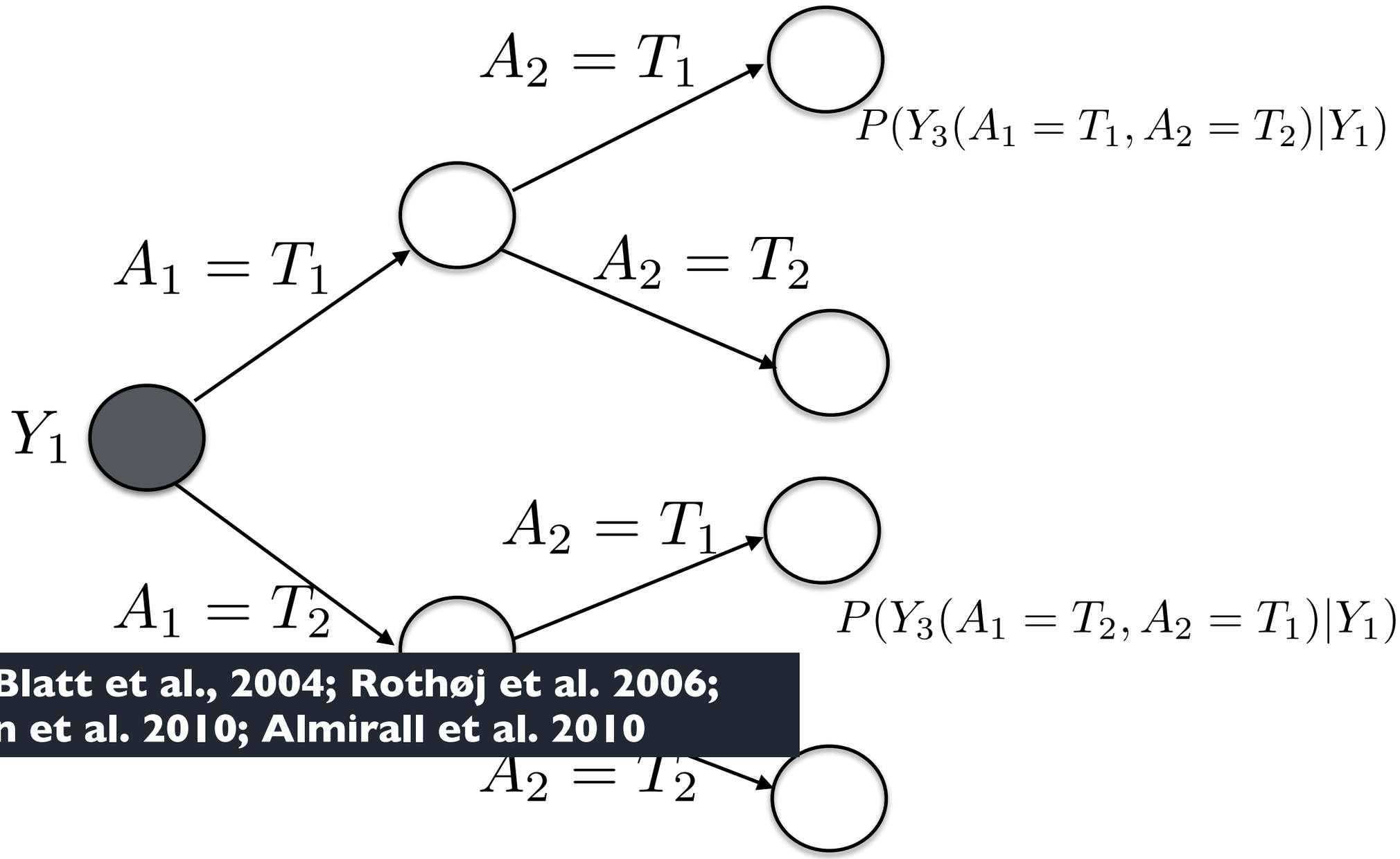
  Garciá and Fernández, 2015

- Modifying optimization criterion (notion of reward)
  - Penalize movement through “error states”

    Geibel and Wysotzki, 2005

- Modifying exploration strategies
  - Incorporate domain knowledge **Martín and Lope, 2009**
  - Apprenticeship: seed MDP parameters using a teachers demonstration **Abbeel and Ng, 2005**

Dynamic Treatment Regimes: Learning from Offline Data



Robins 2004; Blatt et al., 2004; Rothøj et al. 2006; Henderson et al. 2010; Almirall et al. 2010

Optimal decision at time 1

$$\text{argmax}_{A_1} \max_{A_2} f(P(Y_3(A_1, A_2)|Y_1))$$

Murphy 2003

Dudik et al., 2011

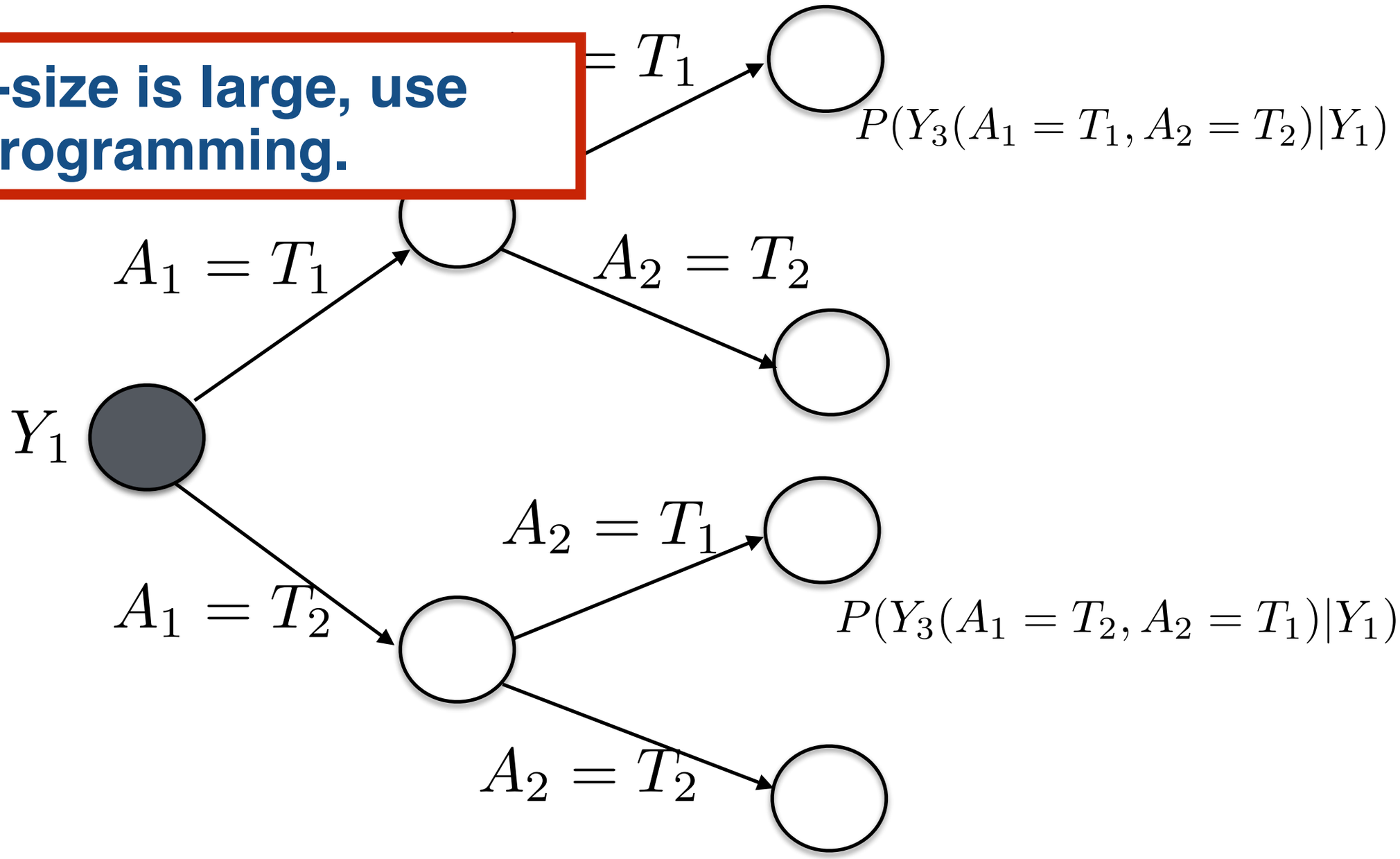
Jiang and Li, 2016

Robins 2004 **Blatt et al., 2004**

Rothøj et al., 2006 **Henderson et al., 2010** **Almirall et al., 2010**

Dynamic Treatment Regimes: Learning from Offline Data

When tree-size is large, use dynamic programming.



Optimal decision at time 1

$$\text{argmax}_{A_1} \max_{A_2} f(P(Y_3(A_1, A_2)|Y_1))$$

Murphy 2003

Dudik et al., 2011

Jiang and Li, 2016

Robins 2004 **Blatt et al., 2004**

Rothøj et al., 2006 **Henderson et al., 2010** **Almirall et al., 2010**

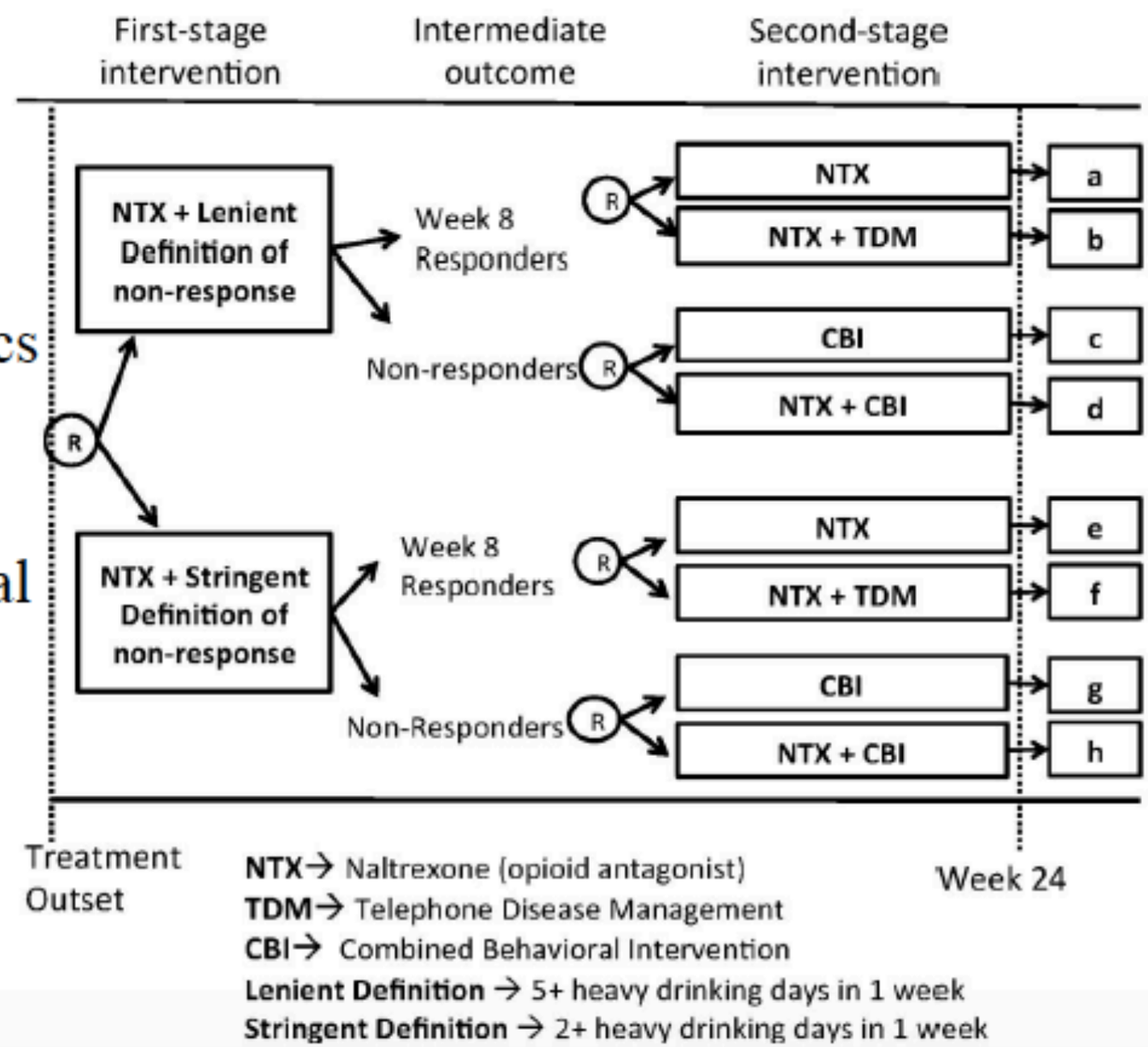
Sequential Multiple Assignment, Randomized Trial (SMART)

ExTENd(PI: Oslin): Treatment of Alcohol Dependence

Rationale:

Naltrexone (NTX, an opiate antagonist) is efficacious but

- Around 1/3 of patients relapse while on NTX,
- Hence, need to develop rescue tactics for non-responders
- And long-term maintenance tactics to for responders
- Because of various barriers: Physiological/social/psychological



- **Trials for evaluating sequential treatment strategies.**
- **Assignment is adaptive**

Slide from Inbal (Billie) Nahum-Shani, Nick Seewald, Susan Murphy

Conclusion & Discussion

- Need for individualization based on diverse data.

- Our practice of medicine will change radically in at least some areas in the next decade and there is an exciting opportunity for us to make a difference.

- **Bayesian Hierarchical Framework for Individualizing Predictions**
  - Motivated latent sources of variability that can be inferred to refine predictions
  - Discussed the problem of inferring disease trajectories

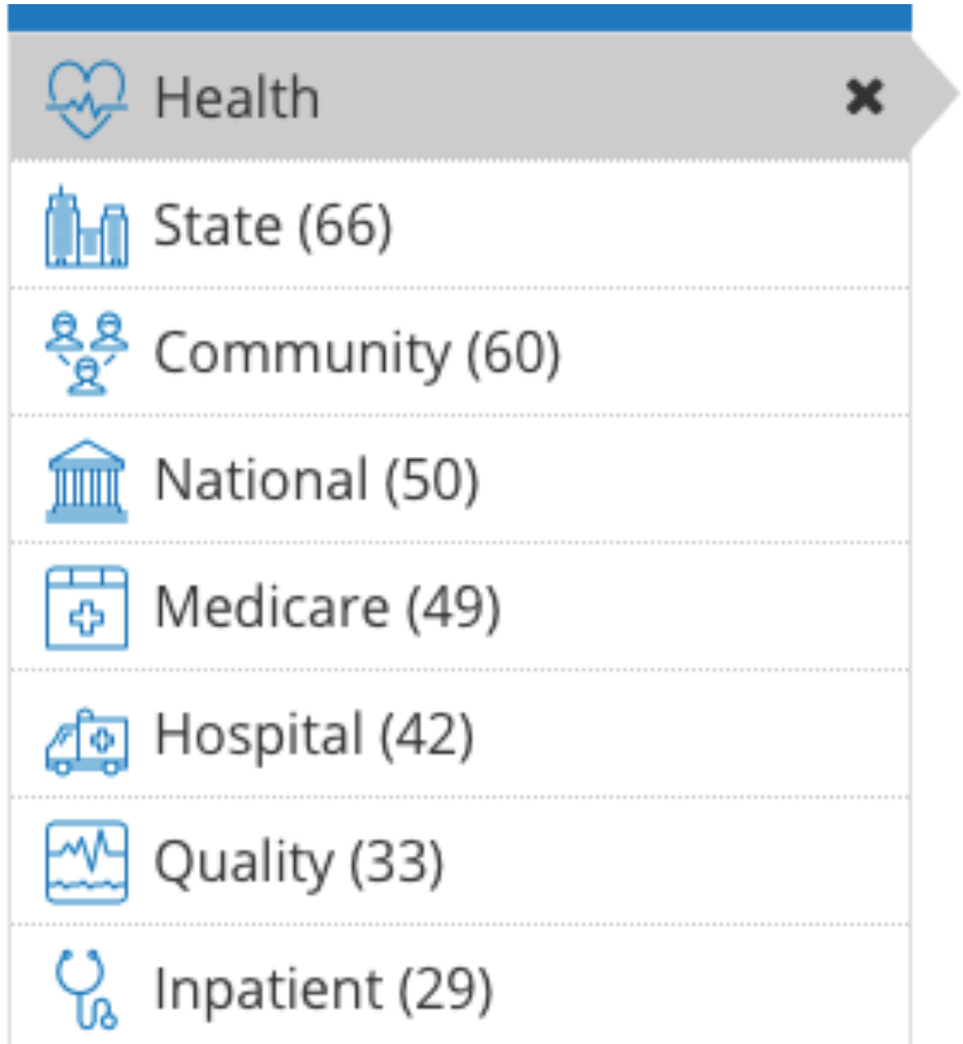
- **Estimating Treatment Effects & Individualized Treatment Effects**
  - Learning from observational data
  - Key ideas: Potential Outcomes, Causal Inference for Bias Adjustment, BNP

- **Causal Predictions**
  - Relax assumption from Part 1 about no treatment effects
  - Discuss predictions that are robust to changes in physician practice behavior

- **From Predictions to Treatment Rules**
  - Connections to Reinforcement Learning, Dynamic Treatment Regimes, SMART

Publicly available datasets

HealthData.gov



Thank you!

ssaria@cs.jhu.edu
www.suchisaria.com
@suchisaria

pschulam@cs.jhu.edu
www.pschulam.com
@pschulam