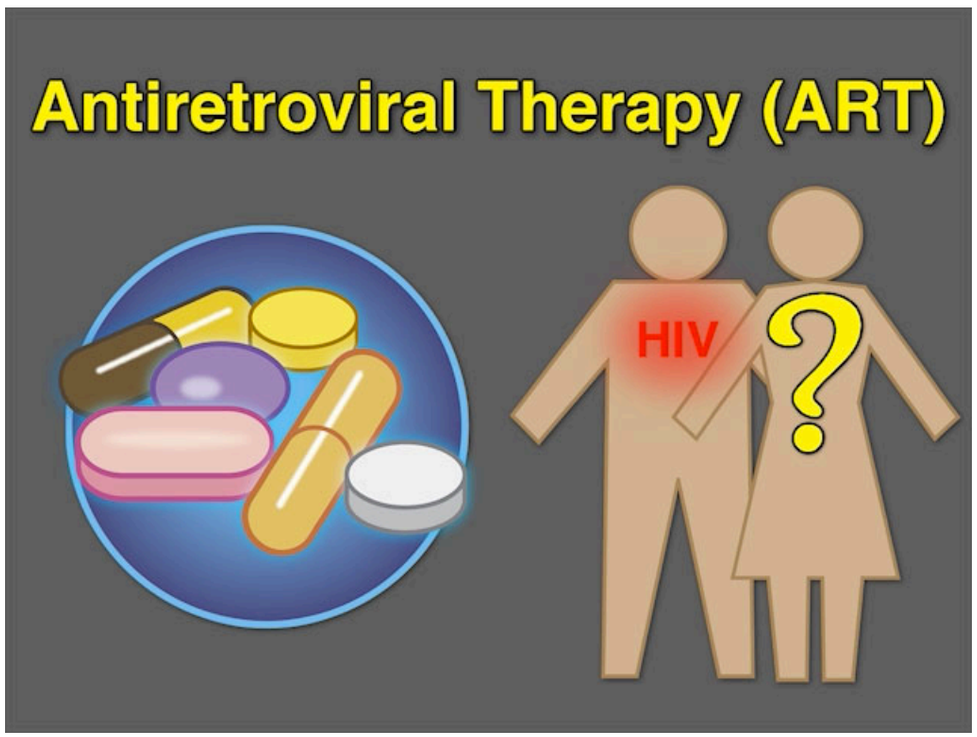
A Bayesian Reinforcement Learning Framework for Optimizing Sequential Combination Antiretroviral Therapy in People with HIV

Yanxun Xu
Department of Applied Math and Statistics
Mathematics Institute for Data Science
Division of Biostatistics and Bioinformatics
The Sidney Kimmel Comprehensive Cancer Center
Johns Hopkins University

September 17, 2022
PCIC

Sequential Decision Making in HIV

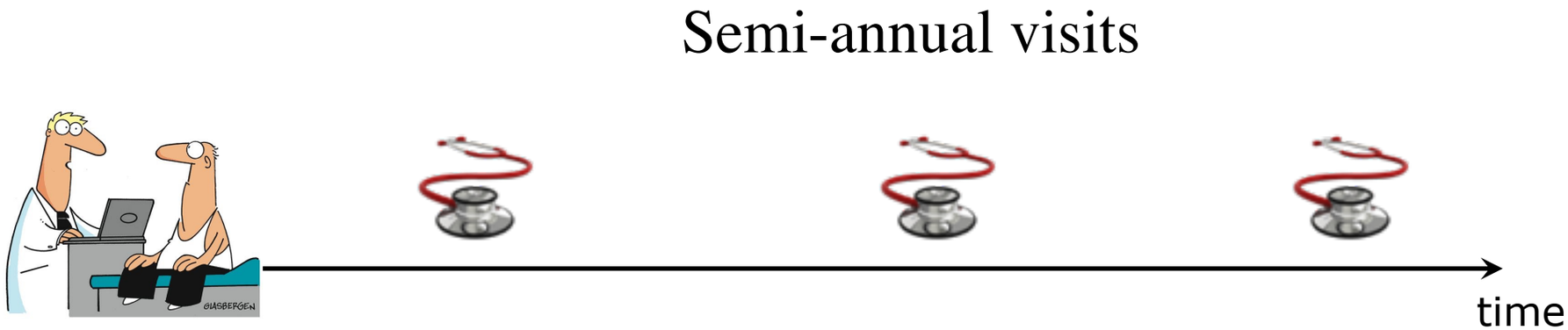


Drug classes:

- Nucleotide reverse transcriptase inhibitor (NRTI)
- Non-nucleotide reverse transcriptase inhibitor (NNRTI)
- Protease inhibitor (PI)
- integrase inhibitor (INSTI)
- Entry inhibitor (EI)
- Pharmacokinetic enhancer (Booster)

A combination of ART drugs

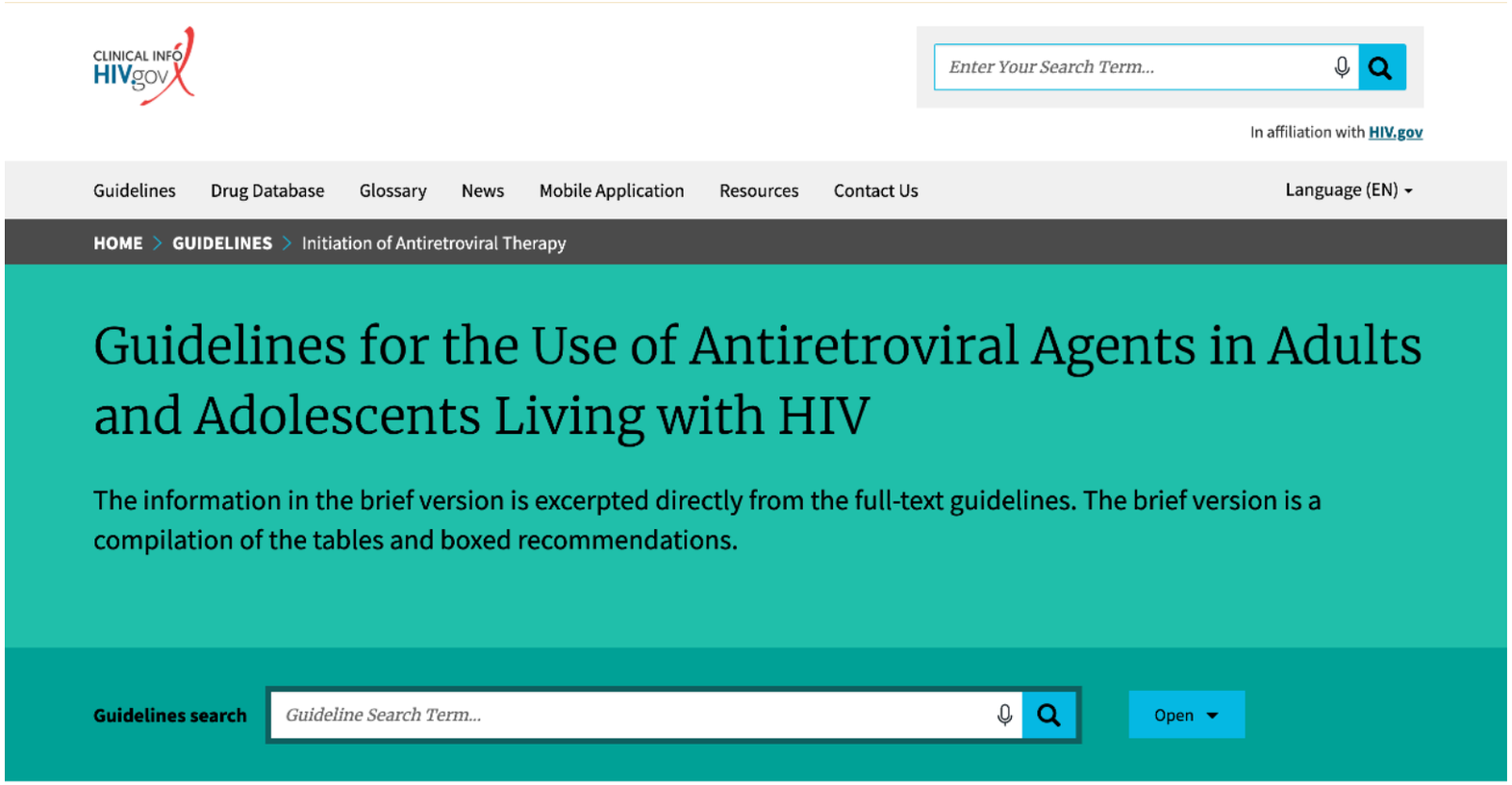
Precision Medicine in HIV



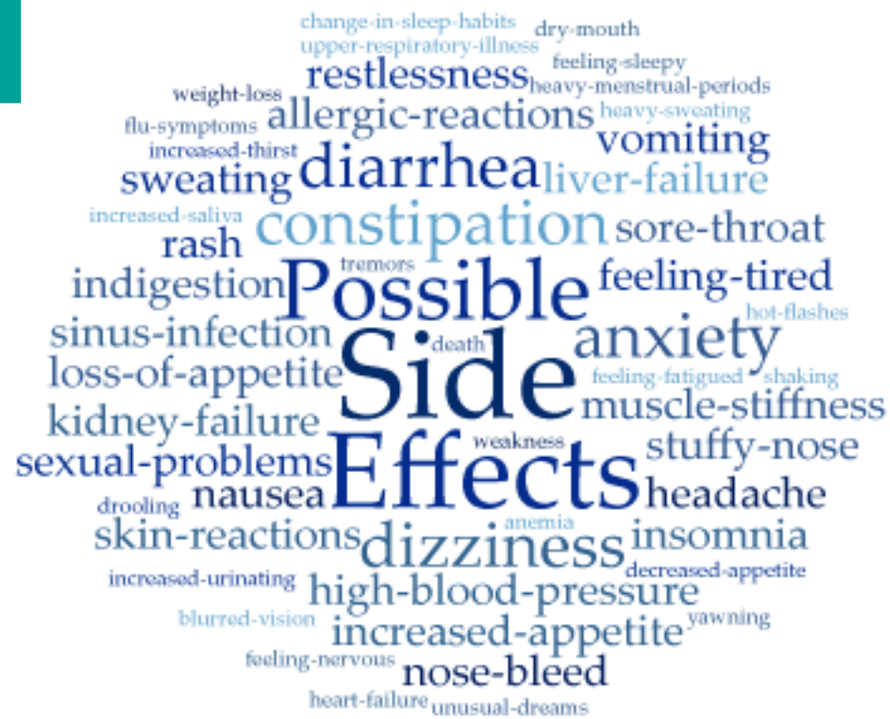
- Sociodemographics
- Laboratory tests
- Behavioral characteristics
- Clinical factors

Assign ART regimen

Precision Medicine in HIV



- Treatment-naive:
- Pre-treated:

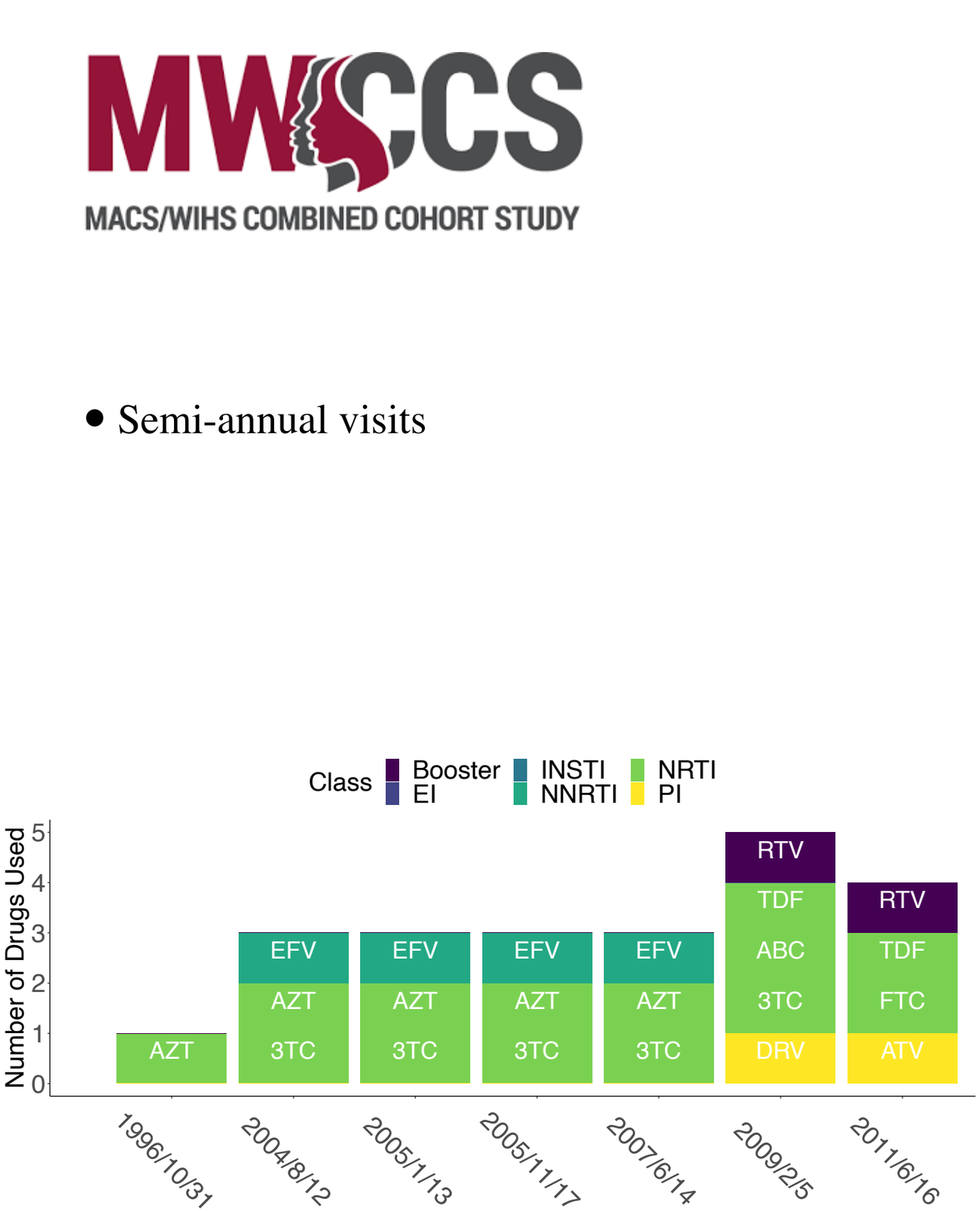


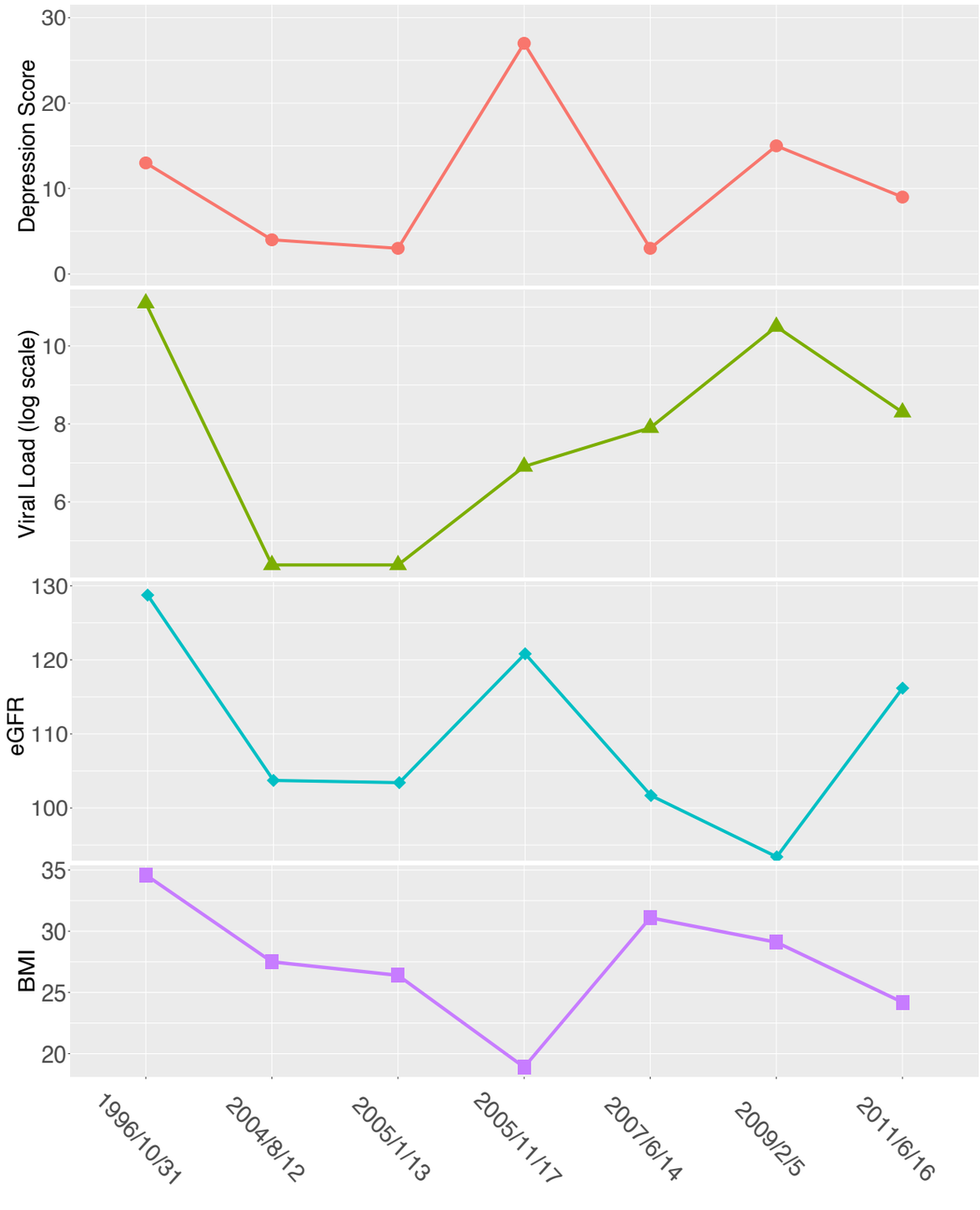
Goal: determine the personalized ART regimen to optimize the *long-term* health

Large-Scale HIV Studies

MWCCS
MACS/WIHS COMBINED COHORT STUDY

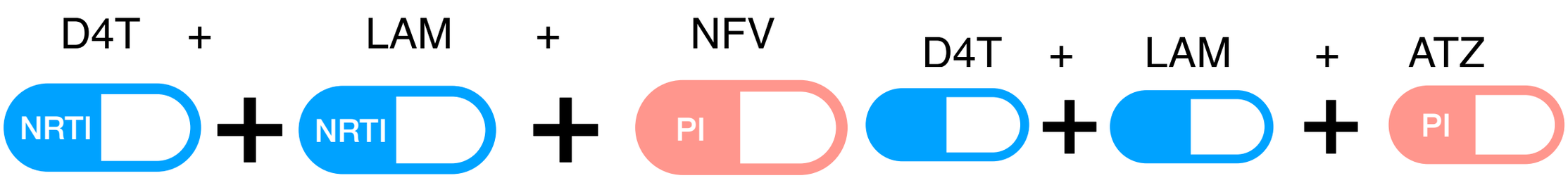
- Semi-annual visits





Challenges

- Estimate the effects of ART regimens from a high-dimensional and unbalanced space.
  - High-dimensional: with more than 30 ART drugs on the market, there are a large number of possible drug combinations.
  - Unbalanced: some drug combinations are frequent whereas others are rare.



- Generate a realistic ART regimen from a large discrete space.
  - How to represent an ART regimen?

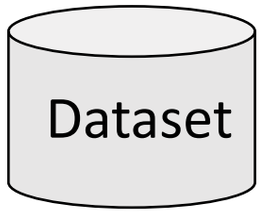
  A binary vector (2^N dimension)? **(Drug 1, Drug 2, …, Drug N)**

- Optimize treatments from observational data (*distribution shift*).

Approach Overview

Environment

MWCCS
MACS/WIHS COMBINED COHORT STUDY

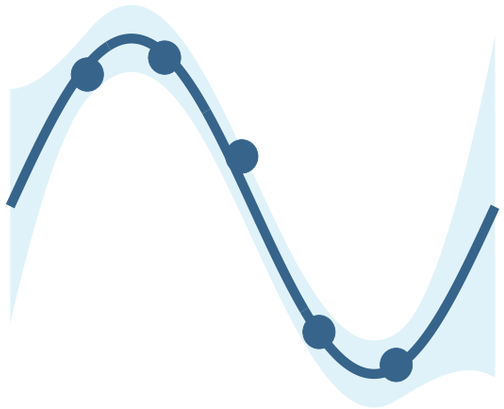


Diagnosis

Survey

Lab Test

Step I: Bayesian Dynamics Model

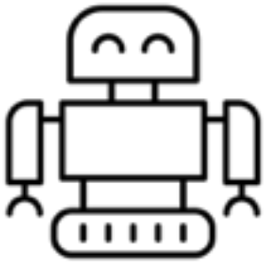


Learned Dynamics

Pessimistic Environment with Uncertainty Penalization

Step II: Offline Reinforcement Learning

Combination Antiretroviral Therapy



Policy Optimizer

Problem Formulation

- Baseline covariates (e.g., race): $\boldsymbol{X}_{i0}$.
- Visit times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{i,J_i})$.
- M time-varying health *state* variables: $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{i,J_i})$ with $\boldsymbol{Y}_{ij} \in \mathbb{R}^M$.
- $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{i,J_i})$ with Z_{ij} denoting the cART regimen used by individual i during the time period $(t_{i,j-1}, t_{ij}]$.

$$\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^I = \{\boldsymbol{X}_{i0}, \boldsymbol{t}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i\}_{i=1}^I$$

- State and treatment histories: $\overline{\boldsymbol{Y}_{ij}} = \{\boldsymbol{Y}_{ij'} : j' \leq j\}$, $\overline{Z_{ij}} = \{Z_{ij'} : j' \leq j\}$.
- Dynamic model: $\boldsymbol{Y}_{i,j+1} = f(\overline{\boldsymbol{Y}_{ij}}, \overline{Z_{i,j+1}}; \boldsymbol{\phi})$

Goal: optimize ART assignments to maximize the individual's long-term health outcomes

An Optimization Problem

- For any individual i, suppose that she already has J_i visits.
- Let $\boldsymbol{Y}_i^{\text{new}} = \{\boldsymbol{Y}_{ij} : j > J_i\}$ and $\boldsymbol{Z}_i^{\text{new}} = \{Z_{ij} : j > J_i\}$ denote her future longitudinal states and ART regimens.
- ART policy function: $\pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta})$.
- A stochastic reward function: $r_i(\boldsymbol{Y}_i^{\text{new}})$.

Denote the expected reward for any individual i to be:

$$R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} \mid \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})}[r_i(\boldsymbol{Y}_i^{\text{new}})] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi}.$$

Uncertainty

Goal: find the optimal policy function $\pi(\cdot, \cdot; \boldsymbol{\theta}_i^\star)$ such that

$$\boldsymbol{\theta}_i^\star = \arg\max_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}).$$

An Optimization Problem

Policy gradient via stochastic gradient descent (SGD)

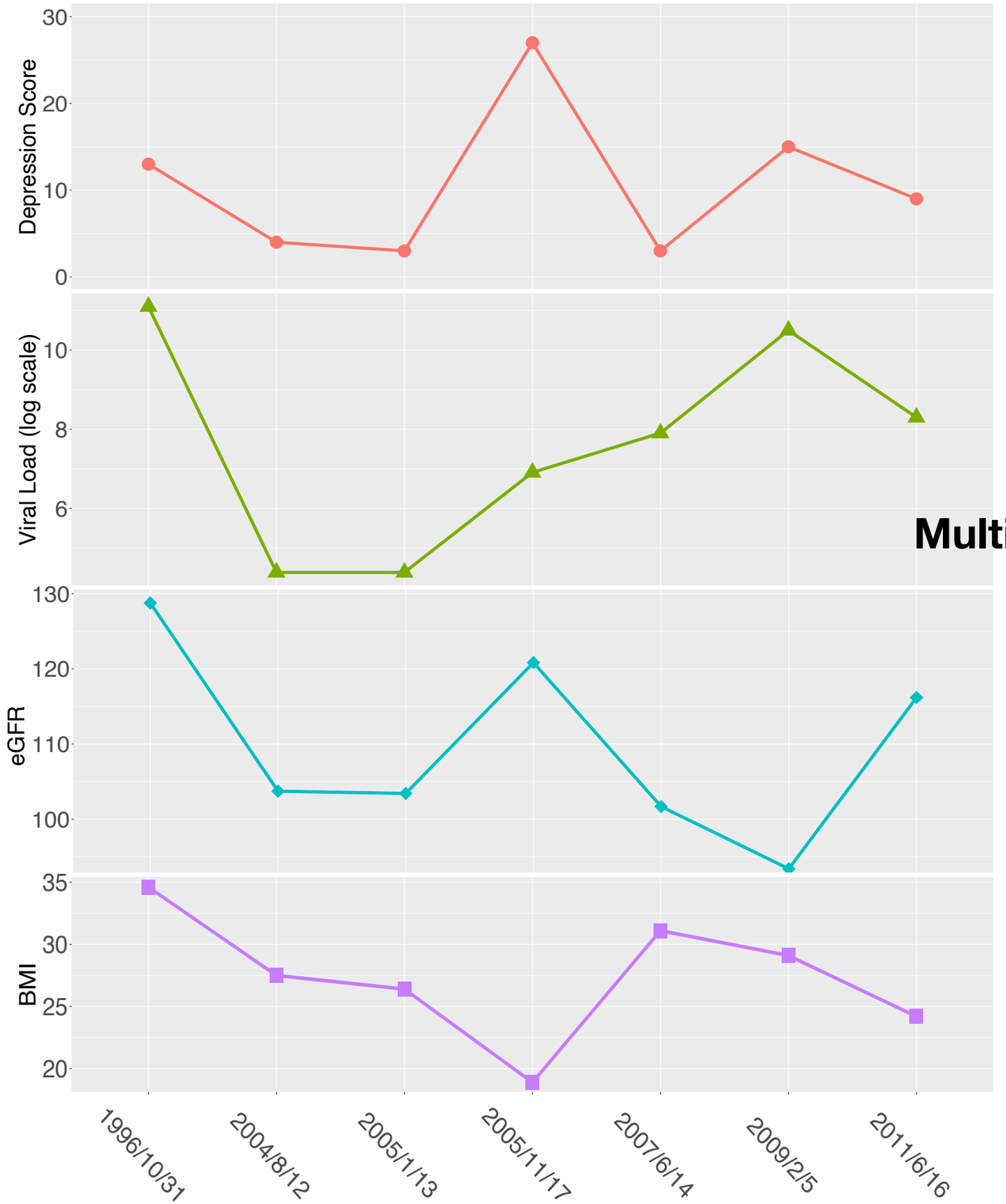
$$\boldsymbol{\theta}_{i,q+1} = \boldsymbol{\theta}_{i,q} + s_{i,q} \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{i,q}}$$

$$R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} | \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})} [r_i(\boldsymbol{Y}_i^{\text{new}})] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi}.$$

$$\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} | \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})} \left[r_i(\boldsymbol{Y}_i^{\text{new}}) \nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right) \right] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi}$$

- Sample future states
- Define a reward function
- Parameterize the policy function

Modeling Longitudinal States



Multivariate Gaussian Process (MGP)

MGP

$$Y_{im}(t) = f_{im}(t) + \epsilon_{im}$$

$(f_{i1}(t), \ldots, f_{iM}(t))$ are MGP-distributed

- Mean $(\mu_{i1}(t), \ldots, \mu_{iM}(t))$.
- Separable covariance function $cov(f_{im}(t), f_{im'}(t')) = C^M_{mm'} C^t(t, t')$.
- C^M: $M \times M$ covariance matrix.
- Ornstein-Uhlenbeck (OU) kernel $C^t(t, t') = \rho_t^{|t-t'|}$.

MGP

$$\mu_{im}(t) = \boldsymbol{X}_{i0}\boldsymbol{\beta}_m + \boldsymbol{V}(t)\boldsymbol{\alpha}_{im} + h_m(\overline{Z_i(t)})$$

Baseline

$$\boldsymbol{V}(t) = (1, t)$$

Random effects

ART Regimen Similarity

$$h_m(\overline{Z_i(t)}) = \frac{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)\gamma_{md}}{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)}$$

Representative ART Regimen

- Desired properties: (1) sharing-of-information; (2) reducing the high-dimensionality.

ART Regimen Similarity

ART Regimen Similarity

$$\frac{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)\gamma_{md}}{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)}$$

- Linear Kernel
  - Computes the similarity between regimens based on the proportion of **common drugs** that two regimens share.

D4T (NRTI) + LAM (NRTI) + NFV (PI)

D4T + LAM + ATZ (PI)

D4T + LAM + EFV (NNRTI)

Representative ART Regimen

FTC (NRTI) + TDF (NRTI) + RTV (PI)

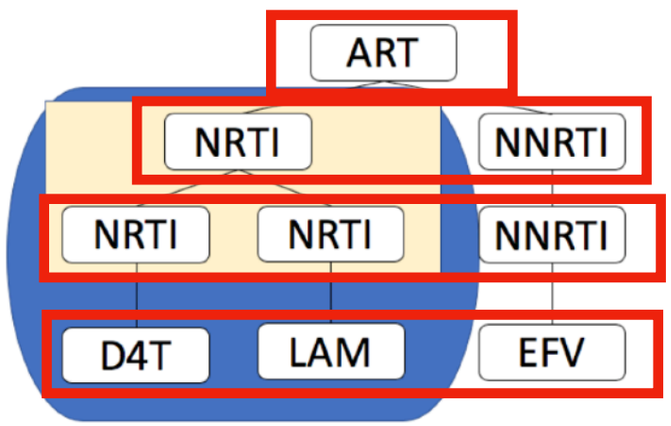
ART Regimen Similarity

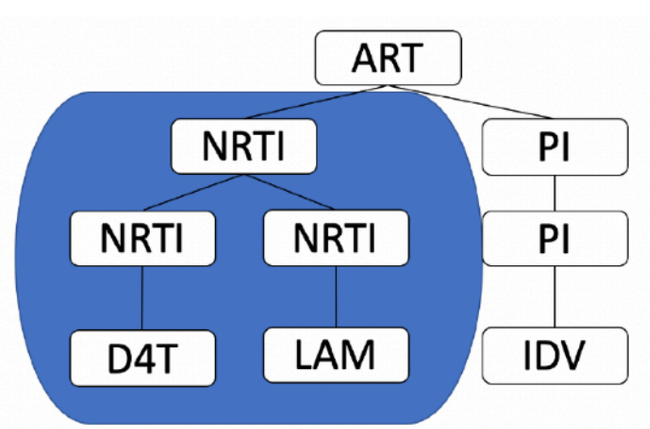
ART Regimen Similarity

$$\frac{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)\gamma_{md}}{\sum_{d=1}^{D} \kappa(Z_i(t), z_d)}$$

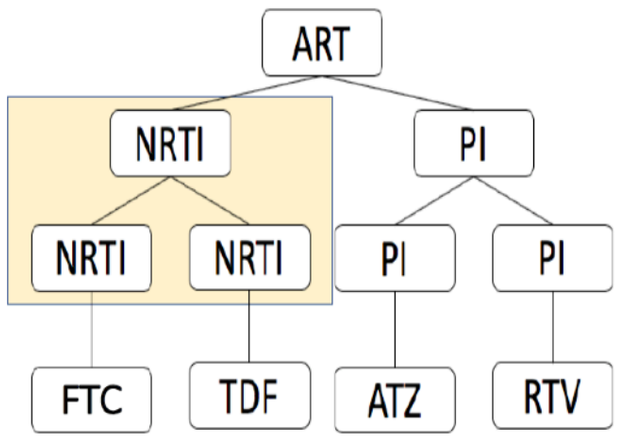
Representative ART Regimen

- Subset-tree (ST) Kernel
  - Calculates the similarity between regimens across **all levels of the tree representation.**





(a) Regimen A



(b) Regimen B

	A	B	C
A	5.23	2.13	0.50
B	2.13	5.23	1.25
C	0.50	1.25	6.51

(c) Regimen C

(d) Similarity score matrix

Modeling Longitudinal States

Posterior Inference:

- Assign priors to all unknown parameters
- Obtain posterior distributions from MCMC

- Sample future states 
- Define a reward function
- Parameterize the policy function

Reward

Define the reward function based on

viral load, kidney function, and depression in the next two years

$$r_i(\boldsymbol{Y}_i^{\text{new}}) = -\sum_{j=J_i+1}^{J_i+4} \Big\{ \underbrace{w_{i1} Y_{ij1}}_{\text{depression}} + \underbrace{w_{i2} |Y_{ij2} - T_V| \mathbb{I}(Y_{ij2} > T_V)}_{\text{viral load}} + \underbrace{w_{i3} |Y_{ij3} - T_E| \mathbb{I}(Y_{ij3} < T_E)}_{\text{eGFR}} \Big\}$$

abnormal threshold

Personalized weights: $\boldsymbol{w}_i = (w_{i1}, w_{i2}, w_{i3})$

An Uncertainty-Penalized Reward

Distribution shift: Model-based uncertainty-penalized policy optimization (Yu et al. 2020)

A pessimistic environment: $\widetilde{r}_i(\boldsymbol{Y}_i^{\text{new}}) = r_i(\boldsymbol{Y}_i^{\text{new}}) - \lambda u(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}})$

Tuning parameter

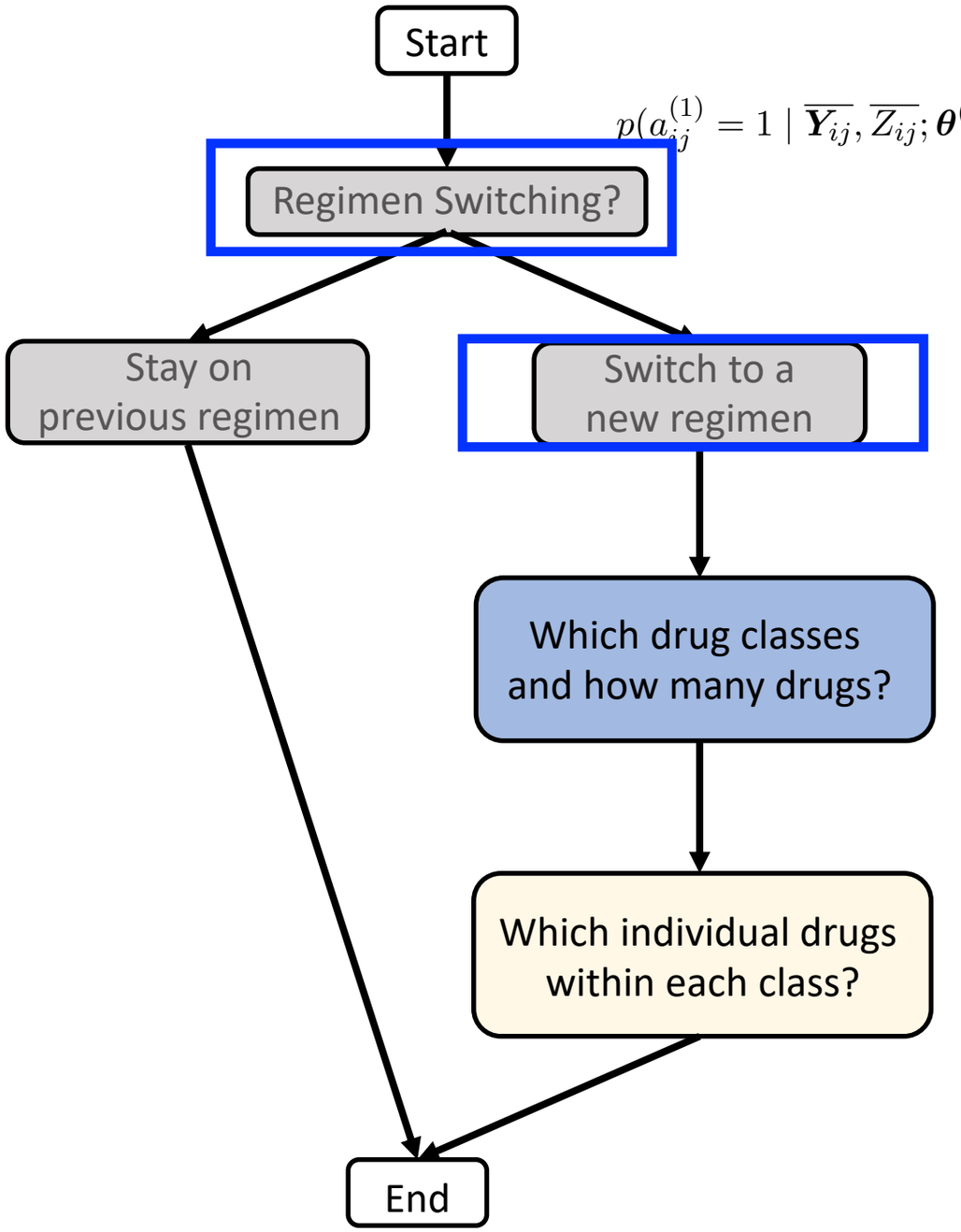
Uncertainty

$u(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) = \sum_{j=J_i+1}^{J_i+4} \sum_{m=1}^{M} \sqrt{\text{Var}(Y_{ijm} \mid Z_{ij}, \mathcal{D})}$ posterior predictive distribution

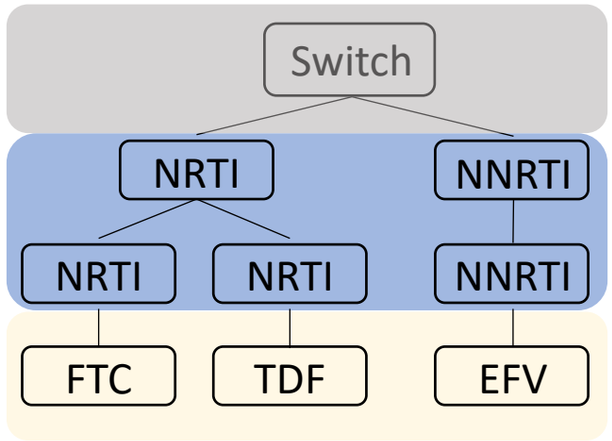
- Sample future states
- Define a reward function

- Parameterize the policy function

Decision Process for Assigning ART



$$p(a_{ij}^{(1)} = 1 \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(1)}) = \begin{cases} 1, & \text{if } Y_{ij1} > T_D \text{ or } Y_{ij2} > T_V \text{ or } Y_{ij3} < T_E, \\ 0, & \text{otherwise,} \end{cases}$$



Level I

Level II

multi-class logistic regression model

Level III

Wallenius' noncentral hypergeometric distribution

An Optimization Problem

Policy gradient via stochastic gradient descent (SGD)

$$\boldsymbol{\theta}_{i,q+1} = \boldsymbol{\theta}_{i,q} + s_{i,q} \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_{i,q}}$$

$$\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} | \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})} \left[r_i(\boldsymbol{Y}_i^{\text{new}}) \nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right) \right] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi}$$

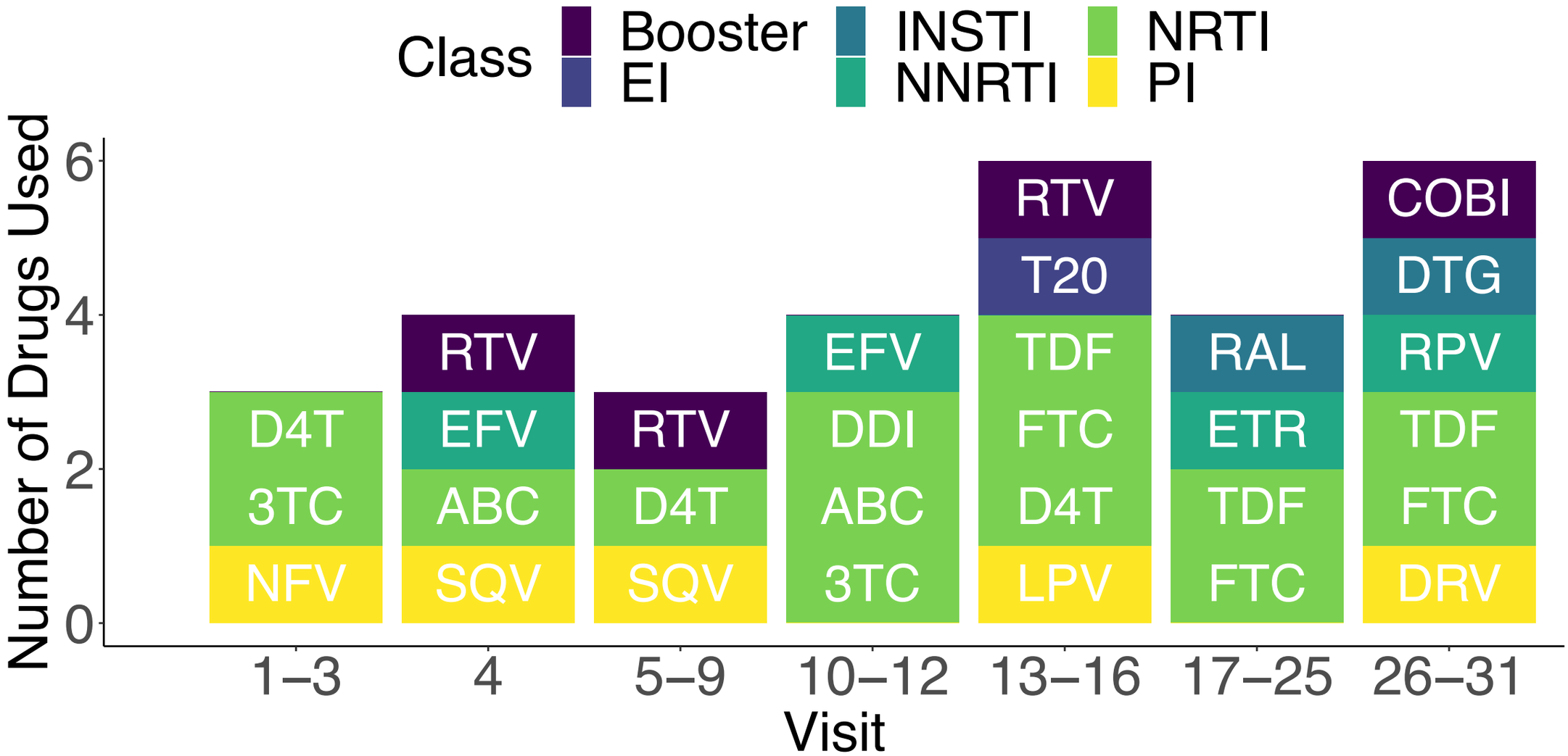
- Sample future states
- Define a reward function
- Parameterize the policy function

- Sample future states
- Define a reward function
- Parameterize the policy function

WIHS Data Analysis

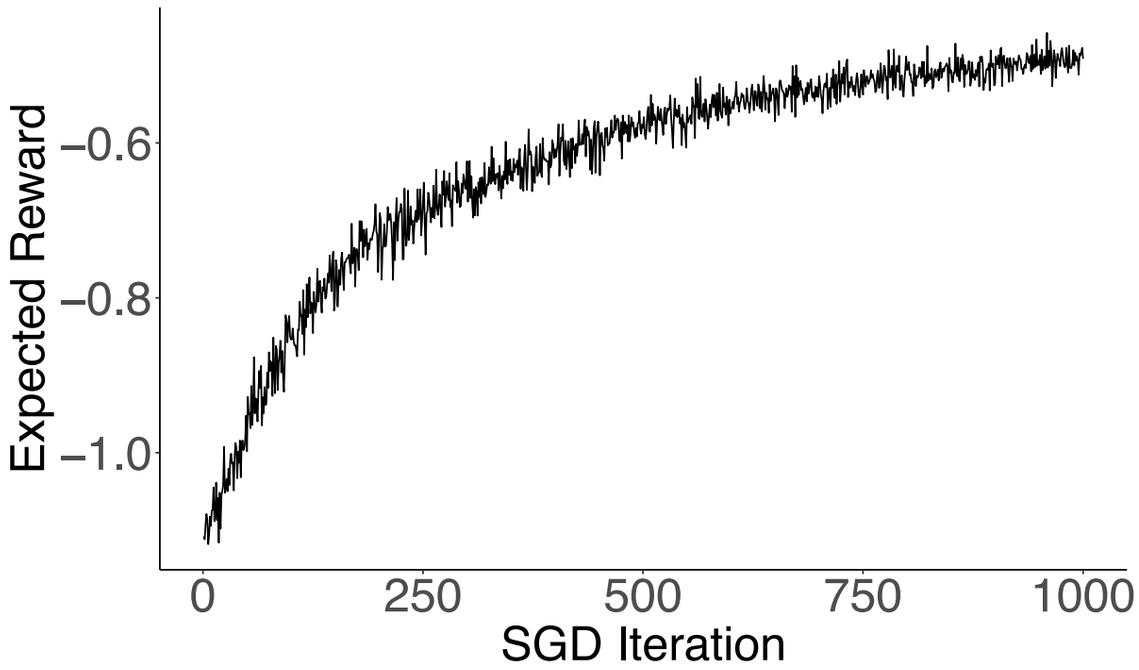
- I=339 women from the Washington DC site.
- M=4 state variables at each visit: depression, viral load, eGFR, and BMI.
- 8% missing rate.
- Baseline covariates: age, smoking status, substance use, employment status, hypertension, and diabetes.
- N=31 ART drugs, K=6 drug classes, D=105 representative ART regimens

Precision Medicine

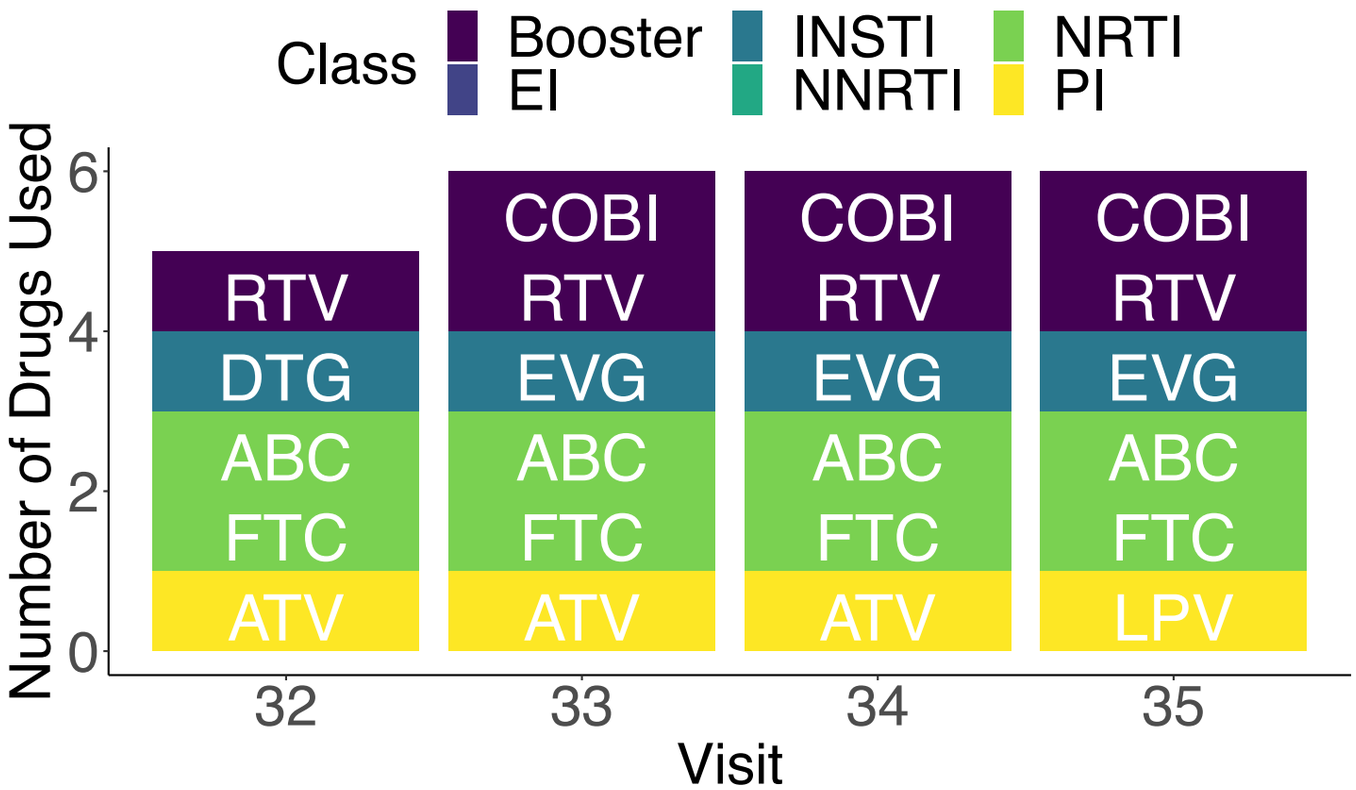


$$w = (1/3, 1/3, 1/3)$$

Precision Medicine

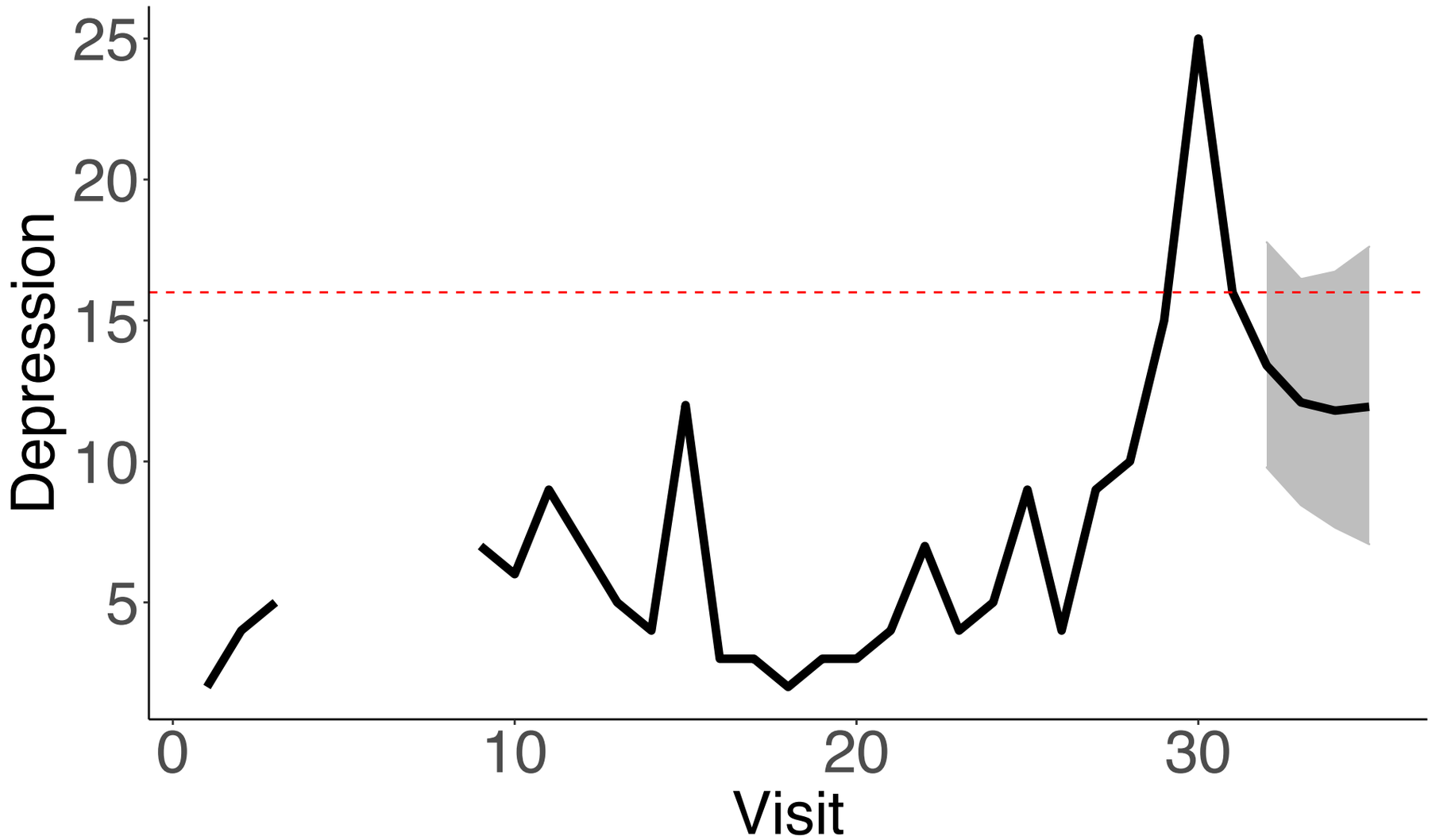


Optimal regimens for visits 32-35



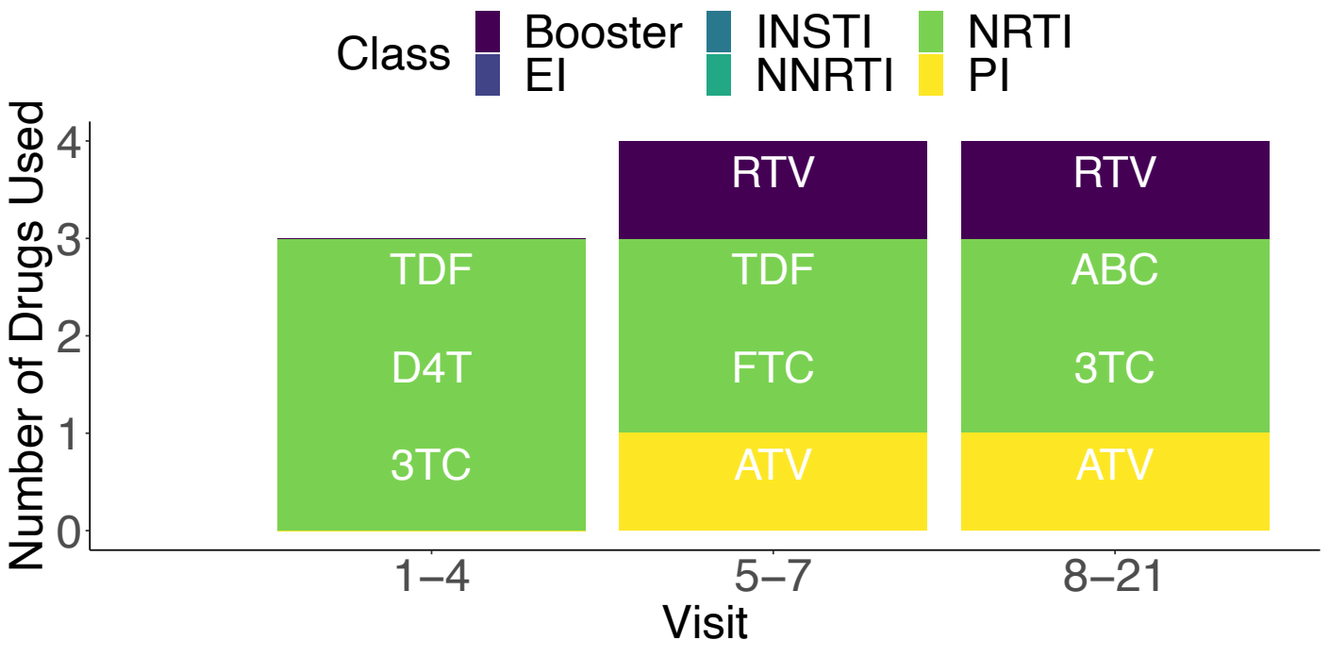
Precision Medicine

Predictive depression scores under the estimated optimal regimens

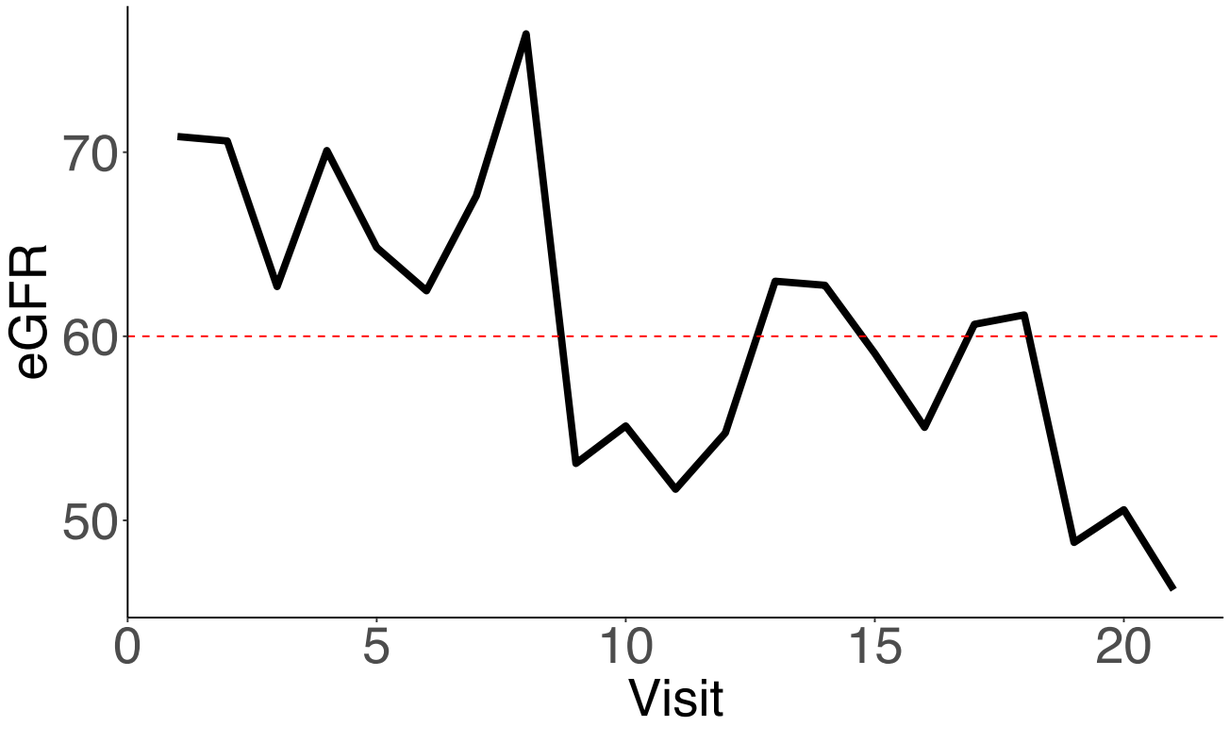


23% Improvement

Precision Medicine: uncertainty-penalized policy

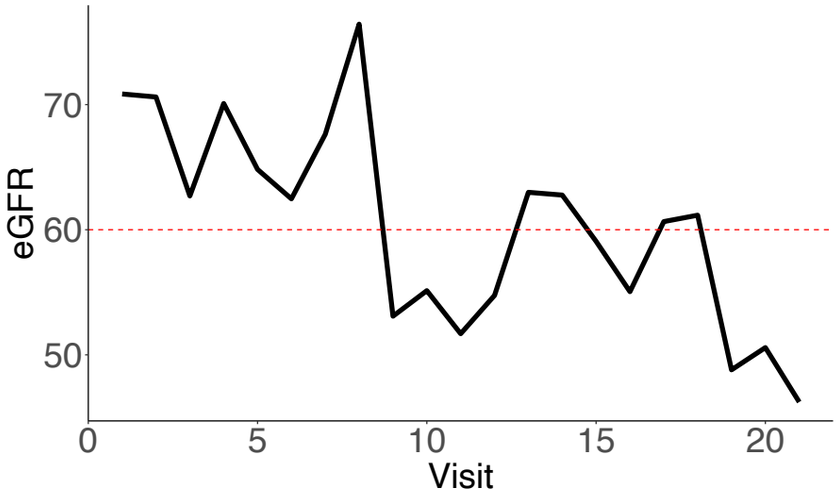


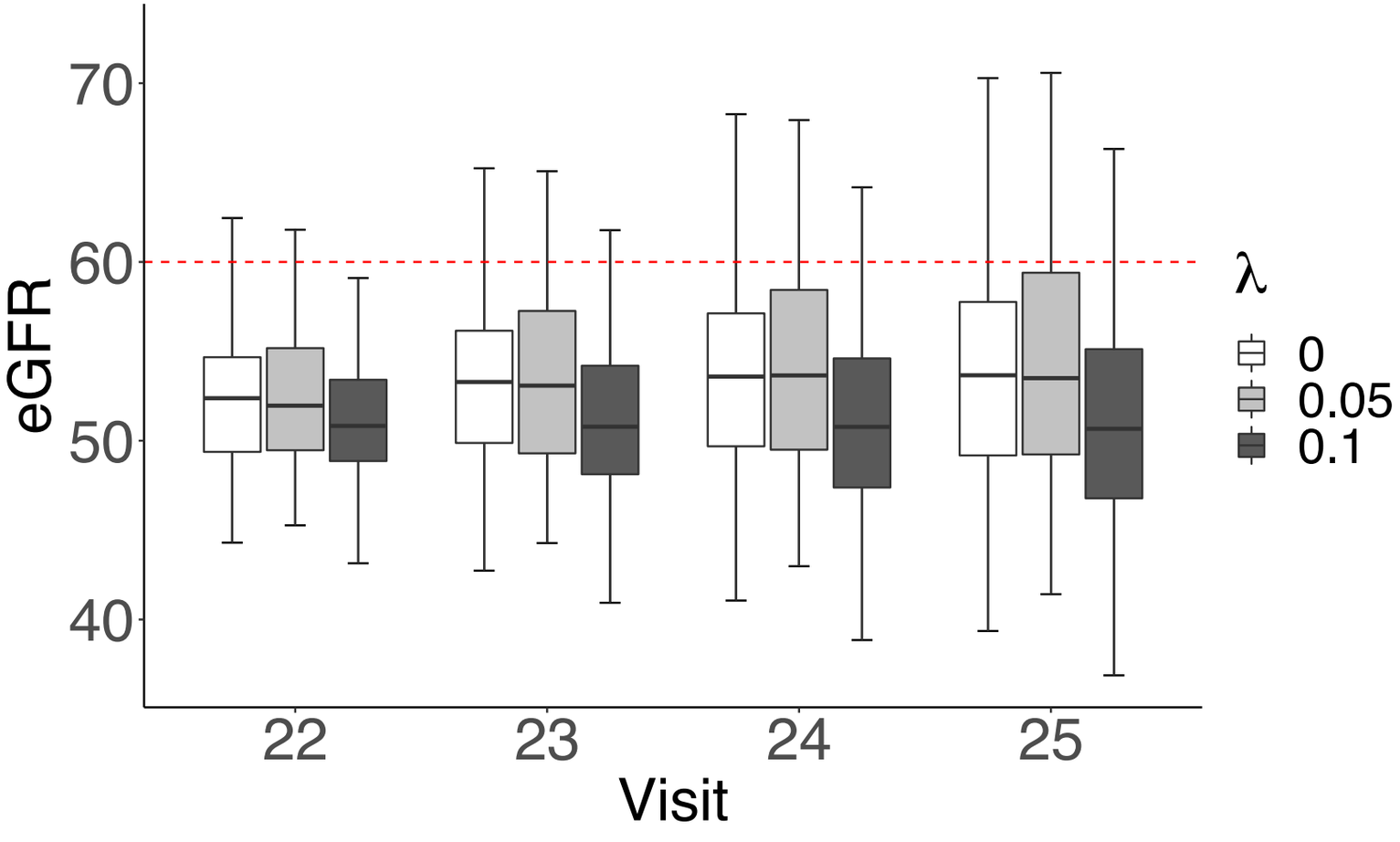
$$w = (0.1, 0.1, 0.8)$$



Precision Medicine: uncertainty-penalized policy

- When $\lambda = 0$, optimal sequence of regimens: FTC+ABC+EFV (two NRTIs + one NNRTI)
  0
- When $\lambda = 0.05$, optimal sequence of regimens: 3TC+ABC+EFV (two NRTIs + one NNRTI)
  71 times
- When $\lambda = 0.1$, optimal sequence of regimens: 3TC+ABC+ATV+RTV (two NRTIs + one PI + one Booster)
  96 times






Clinical Trials
Alzheimer
Data Science
Cancer
Nonparametrics
Reinforcement Learning
Dynamic Treatment Regimen
COVID-19
HIV
Machine Learning
Precision Medicine
Bayesian
Statistics
Mental Health
Electronic Health Record
Causal Inference
High Dimension
Genomics
Theory

Thank you!