Using Electronic Health Records Data for Predictive and Causal Inference About the HIV Care Cascade

Joseph Hogan

Department of Biostatistics
School of Public Health
Brown University

University of South Carolina
National Big Data Health Science Conference
February 10, 2020

Predictive vs Causal Inference

Predictive inference concerned with predicting outcome Y from set of inputs X

- Predict failure to show up at next clinical appointment
- Predict 5-year mortality among those with newly diagnosed disease
- Predict presence of fetal heart defect using image data

Predictive vs Causal Inference

Predictive rules are usually built using statistical or machine learning models of the form

$$\Pr(Y \mid X) \;=\; g(X; \beta)$$

where

$$\begin{aligned} g(\cdot) &= \text{some function like (inverse) logit} \\ \beta &= \text{vector of parameters estimated from data} \end{aligned}$$

Here the function $g(\cdot)$ is generic; could be

- regression function
- random forest
- classification tree
- etc.

Predictive vs Causal Inference

Causal inference concerned with answering 'what if' kinds of questions

- I have a coronary artery blockage. Should I get surgery or pursue a more conservative course of treatment?
- Among individuals newly diagnosed with HIV, is it better to start treatment immediately or wait until symptoms develop?

Contrast these with *predictive* questions, answered (for example) from an EHR database

- Did those who received surgery have better 30-day mortality?
- Did those who received HIV treatment early have longer expected survival?

Predictive vs Causal Inference

What is needed to generate *predictive* inferences?

- Temporal ordering: X comes before Y
- Lots of replicates of (X, Y) from a representative population
- Methods to train and validate statistical models or algorithms

What is needed to generate *causal* inferences?

Predictive vs Causal Inference

Consider the following two statements. Both can be simultaneously true

- (Predictive) Those who receive HIV treatment immediately upon diagnosis have shorter survival time, on average, than those who wait.
- (Causal) Given the choice to treat immediately or wait until symptoms develop, treating immediately will lead to longer survival on average

Predictive vs Causal Inference

This brings us back to the question: *what is needed to make causal inferences?*

Causal inference from observed data can be complicated, but two things are essential:

- A plausible model of the causal effect of exposure or treatment on outcome
- Randomized assignment to the exposure of interest; OR, assumptions that allow us to mimic randomization

With observational data such as EHR, the assumptions underlie common methods

- Matching
- Inverse probability weighting
- Covariate adjustment
- Standardization

AMPATH Program in western Kenya

- AMPATH: Academic Model Providing Access to Healthcare
- PEPFAR-funded HIV care program based in Eldoret, Kenya
- Over 150,000 individuals in care at over 100 clinical sites
- Electronic health record: AMPATH Medical Record System
  - data from several million clinical encounters
  - augmented with lab data (CD4, others where available)
  - stored on a central server
  - expanding to mobile data entry

HIV care cascade

- Conceptual model describing progression through stages of HIV care
- Key stages
  - Identify new cases
  - Link to care
  - Initiate treatment
  - Positive treatment outcomes (e.g., viral suppression)
  - Retain in care
- More recently: used to frame policy goals

HIV care cascade

Source: `aids.gov`

Goals for understanding cascade

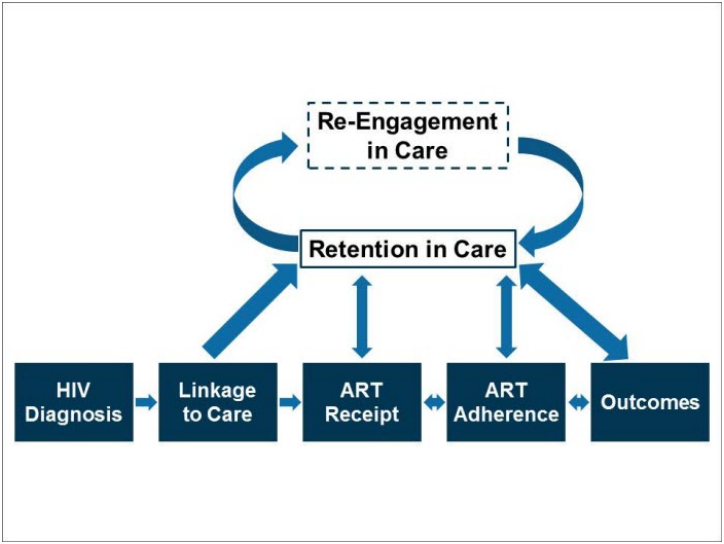
Prediction

- Generate predictive models of transition between states
- E.g., flag those who are at risk for negative outcomes
- Regression, Machine learning

Evaluation

- Causal inference about a policy, treatment, exposure
- E.g., what is the effect of immediate treatment initiation, compared to marker-based initiation?
- Causal structural models

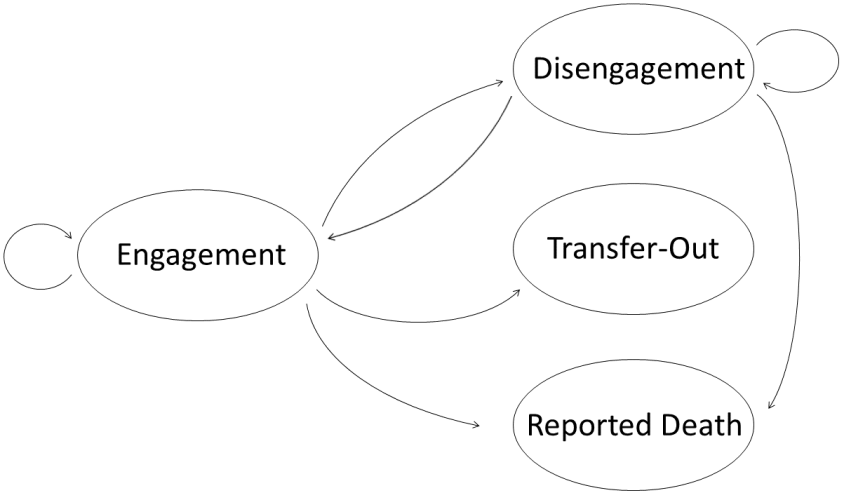
Can be complex to model progression through care



Mugavero MJ, Norton WE, Saag MS. Health care system and policy factors influencing engagement in HIV medical care: piecing together the fragments of a fractured health care delivery system. Clin Infect Dis. 2011;52:S238-S246

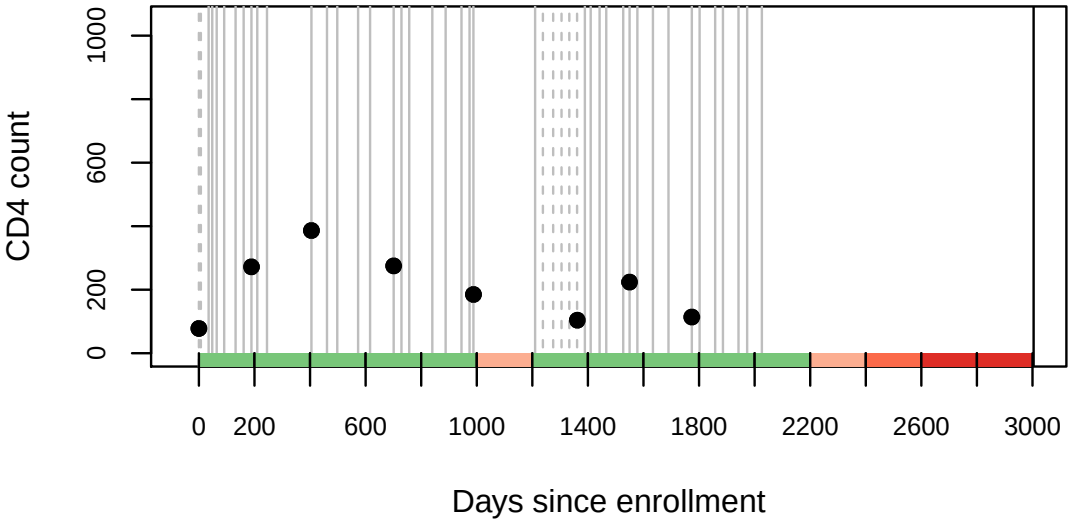
Model of state transitions over time

Each arrow represents transition over one time period

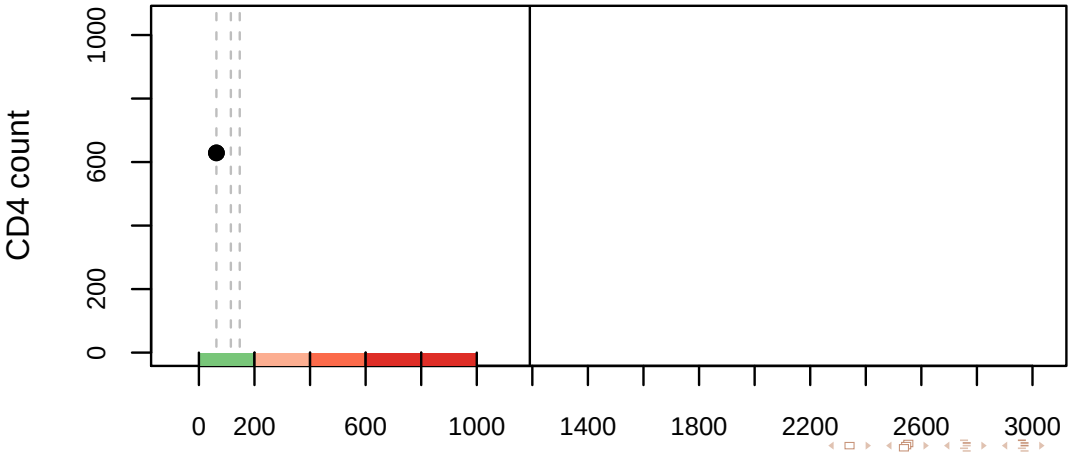


Lee et al., Stat Med (2017)

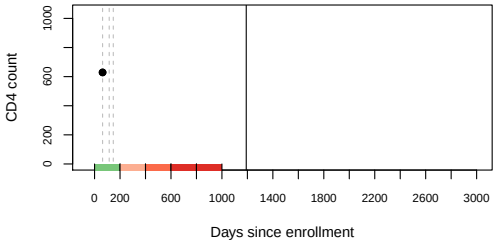
Challenge: Translate patient-level data into states

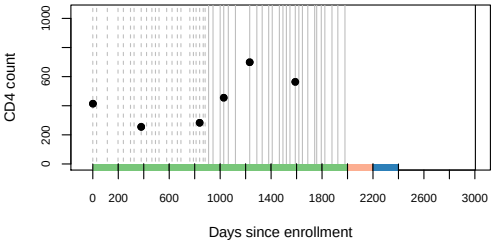


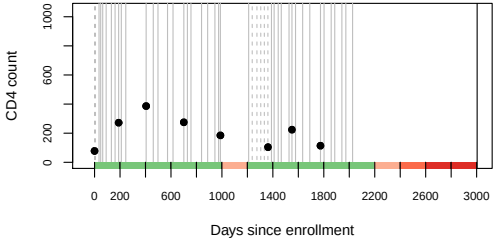
Challenge: Translate patient-level data into states

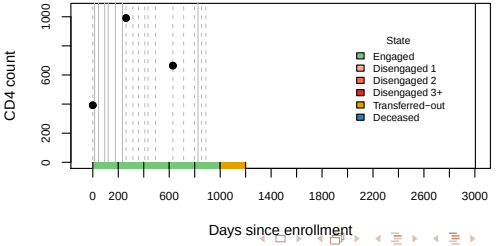


Challenge: Translate patient-level data into states

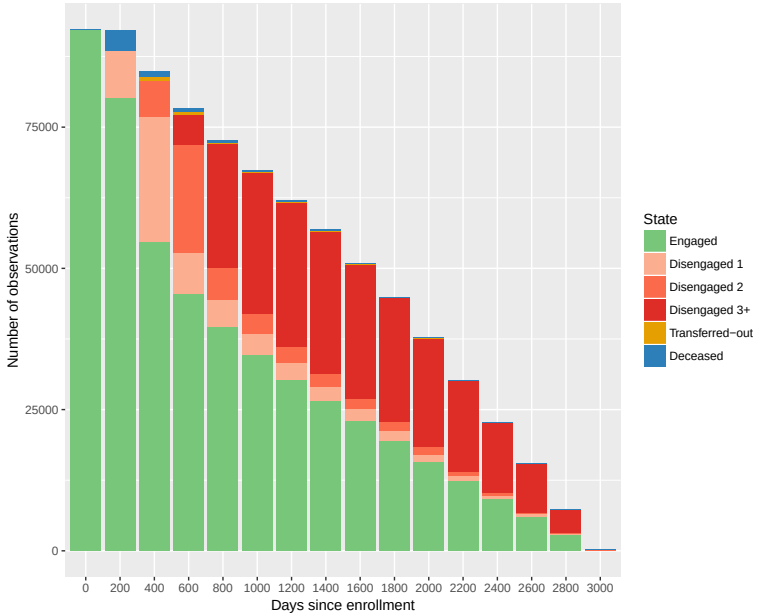








Summary of available data



Analytic approach

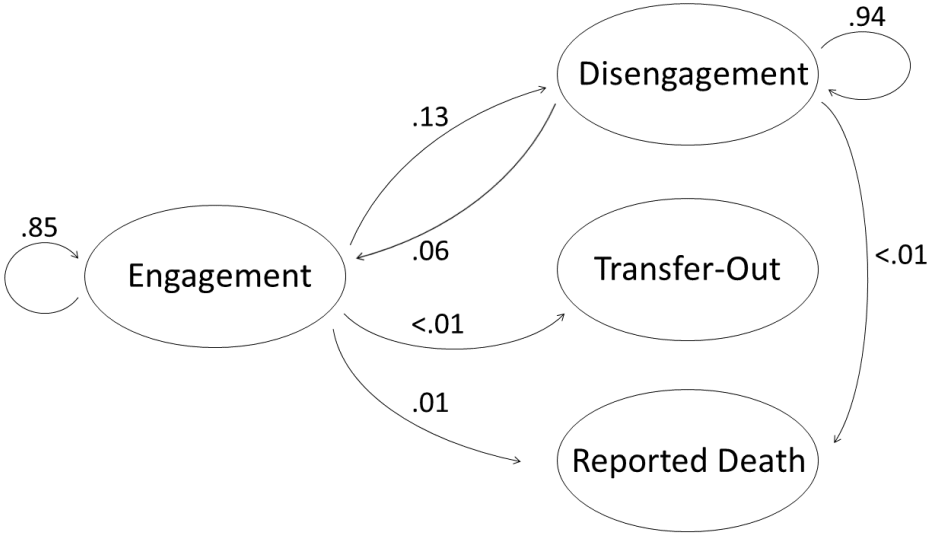
1. Organize data into states
2. Specify model for observed data
   - Transition between states
   - Dependence of transitions on covariates
   - Longitudinal model for covariates
3. How to use fitted observed-data models
   - Summary of transition rates (fill in numbers on graph)
   - Individual-level predictions
   - Causal policy comparisons

Aggregated Transition Rates from AMPATH Data

	State at t_j			
State at t_{j-1}	engaged	disengaged	transferred out	died
engaged	.85	.13	¡.01	.01
disengaged	.06	.94		¡.01
transferred			1	
died				1

- Assumes constant rate over time
- Death and transfer-out rates under-estimated (need tracing data)

Aggregated Transition Rates from AMPATH Data



Can Model State Transitions over Time

For interval $j \in \{1, \ldots, J\}$,

- S_j: multinomial state at time j
  - disengaged (0), engaged (1), transfer out of care program (2), died (3)
- $\boldsymbol{x}_j$: vector of covariates (some time-varying)
- Multinomial probabilities for transition $k \to \ell$ at each time interval

$$p_{jkl}(\boldsymbol{x}_j) = P(S_j = \ell \,|\, S_{j-1} = k, \boldsymbol{x}_j)$$

Prediction model: Observed-data regressions

Multinomial regression for longitudinal data

$$\log\{p_{jk\ell}(\boldsymbol{x}_j)/p_{jkL}(\boldsymbol{x}_j)\} = \boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta}_{jk\ell} \quad \ell = 1, \ldots, L-1$$

$$p_{jk\ell}(\boldsymbol{x}_j) = \text{probability of transition from } k \text{ to } \ell \text{ at time } j$$

$$\boldsymbol{x}_j = \text{covariates at time } j$$

Prediction model: Observed-data regressions

Covariates: $\boldsymbol{x}_j$ = vector of covariates observed just prior to t_j

- CD4 count (baseline and time-varying)
- baseline viral load
- height, weight
- HIV stage (graded 1-4)
- age, gender, marital status
- treatment status
- travel time to clinic
- enrollment year
- calendar year

Regression models

Multinomial regression for transition from 'engaged' at $j = 3$ (day 600)

State at t_{j-1}	Engaged		
State at t_j	Disengaged	Transfer	Death
age	–0.02*	–0.01*	0.01*
male	0.18*	–0.05	0.10
Enrollment Year	0.011	–0.04*	–0.06*
TravelTime	–0.01	0.01	–0.04
WHO stage	0.05*	0.06	0.09*
Married	–0.15*	–0.08	–0.16*
Height	–0.002	0.00	0.00
log Weight	–0.26*	–0.13	–0.29*
undetectable VL	–0.62	–0.05	–6.21
Previous ARV	–0.38*	0.21*	–0.12
CD4 Update	–2.20*	–1.49*	–0.52*
latest log CD4+1	–0.20*	–0.13*	–0.31*

Regression models

- Fit at each time and for each transition
- Can be used for prediction and/or variable selection
- This version has linear covariate effects
  - Can generalize to use machine learners for more flexibility
  - We use BART for multinomial outcomes in the application

From Predictive to Causal Models: The Big Picture

Causal inference using Bayesian G computation algorithm (GCA)

- GCA is a causal inference technique that simulates potential outcomes from predictive models fitted to observed data
- Assumes that within individuals having same covariate values, treatment is randomly allocated
- Validity of inference relies heavily on assumption that predictive models are correctly specified
- To minimize mis-specification, we use Bayesian additive regression trees (BART) for the predictive components

Causal modeling

Question:
How would 'treat immediately' impact progression through the care cascade?

Comparison regimes:

- Policy 1: Treat immediately upon enrollment
- Policy 2: Treat when CD4 falls below 350

Outcome:

- State membership probability at each time interval

Causal structural model to compare treatment policies

Structural model

$$
\begin{aligned}
\boldsymbol{S}_j &= \text{state membership at time } t_j \\
a_j &= \text{treatment assigned at time } t_j \\
\overline{a}_j &= (a_0, \ldots, a_j) \\
P_{\overline{a}_j}(\boldsymbol{S}_j) &= \text{distribution of } \boldsymbol{S}_j \text{ under regime } \overline{a}_j
\end{aligned}
$$

To compare two different regimes $\overline{a}$ and $\overline{a}^*$, want to compare

$$P_{\overline{a}}(\boldsymbol{S}_J) \qquad \text{and} \qquad P_{\overline{a}^*}(\boldsymbol{S}_J)$$

Example: 'treat immediately' is the regime

$$\overline{a}_J = (1, 1, 1, \ldots, 1)$$

What we need to estimate *causal* models

- Observed data

$$
\begin{aligned}
S_j &= \text{state at time } j \\
X_j &= \text{time-varying confounders (CD4)} \\
V &= \text{baseline confounders (age, gender, site, CD4)} \\
A_j &= \text{observed ART status at time } j
\end{aligned}
$$

- Collection of *predictive* models
  - $P(S_j \mid S_{j-1}, X_{j-1}, A_j, V)$
  - $P(X_j \mid S_{j-1}, X_{j-1}, A_j, V)$
- Assumptions
  - 'No unmeasured confounders'
  - First-order Markov dependence for S and X

G computation for estimating causal quantities

Target: $P_{a_0}(\boldsymbol{S}_1)$ when $a_0 = 1$
(state membership distribution if everyone receives treatment at baseline)

Confounders: $X_0 =$ baseline CD4 count, $V =$ (age, gender)

G computation:

$$P_1(\boldsymbol{S}_1) \;=\; \int P(\boldsymbol{S}_1 \,|\, A_0 = 1, X_0, V)\; P(X_0, V)\; d(X_0, V)$$

Implementation

$$\widehat{P}_1(\boldsymbol{S}_1) \;=\; (1/n) \sum_{i=1}^{n} \widehat{P}(\boldsymbol{S}_1 \,|\, A_0 = 1, X_{0i}, V_i)$$

G computation for estimating causal quantities

Target: $P_{a_0,a_1}(\boldsymbol{S}_2)$

Patient state probabilities

- at time $t = 2$
- under treatment regime a_0, a_1

Confounders:

- $X_j =$ CD4 count (could be other stuff)
- $V =$ (age, gender, other baseline covariates)

How to use observed-data models as plug-ins

Target: $P_{a_0,a_1}(\boldsymbol{S}_2)$

$$
\begin{aligned}
P_{a_0,a_1}(S_2) \;=\; & \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \\
& P(X_1 \mid A_0 = a_0, X_0, V, S_1) \\
& P(S_1 \mid A_0 = a_0, X_0, V) \\
& P(X_0, V) \\
& d(S_1, X_1, X_0, V)
\end{aligned}
$$

Plug in fitted models for state transitions

How to use observed-data models as plug-ins

Target: $P_{a_0,a_1}(\boldsymbol{S}_2)$

$$
\begin{aligned}
P_{a_0,a_1}(S_2) \;=\; & \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \\
& P(X_1 \mid A_0 = a_0, X_0, V, S_1) \\
& P(S_1 \mid A_0 = a_0, X_0, V) \\
& P(X_0, V) \\
& d(S_1, X_1, X_0, V)
\end{aligned}
$$

Plug in fitted models for time-varying covariates

How to use observed-data models as plug-ins

Target: $P_{a_0,a_1}(\boldsymbol{S}_2)$

$$
\begin{aligned}
P_{a_0,a_1}(S_2) \;=\; & \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \\
& P(X_1 \mid A_0 = a_0, X_0, V, S_1) \\
& P(S_1 \mid A_0 = a_0, X_0, V) \\
& P(X_0, V) \\
& d(S_1, X_1, X_0, V)
\end{aligned}
$$

Fix treatment regime or policy a_0, a_1

How to use observed-data models as plug-ins

Target: $P_{a_0,a_1}(\boldsymbol{S}_2)$

$$
\begin{aligned}
P_{a_0,a_1}(S_2) \;=\; & \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \\
& P(X_1 \mid A_0 = a_0, X_0, V, S_1) \\
& P(S_1 \mid A_0 = a_0, X_0, V) \\
& P(X_0, V) \\
& d(S_1, X_1, X_0, V)
\end{aligned}
$$

Average over the distribution of specific population of interest

Implementation on EHR Data

Step 1: Model learning on 50,000 subjects

Step 2: Model validation on 26,740 subjects

Step 3: Bayesian simulation on 10,000 subjects randomly sampled from all

Step 1: Fit predictive models

Outcome models $P(S_j \mid A_{j-1}, X_{j-1}, V)$

- Use multinomial BART

Time varying covariate models $P(X_j \mid A_{j-1}, X_{j-1}, S_{j-1}, V)$

- Use coutinuous-outcome BART

Next 3 slides:

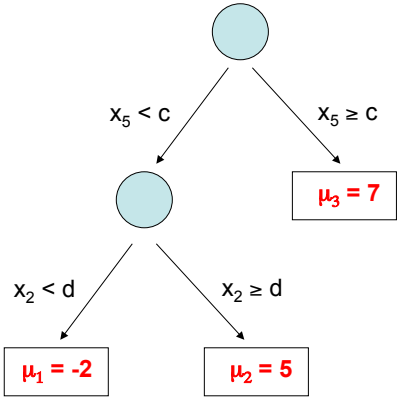
`http://www.rob-mcculloch.org/`

A Regression Tree Model

Let T denote the tree structure including the decision rules.

$M = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denotes the set of bottom node μ's.

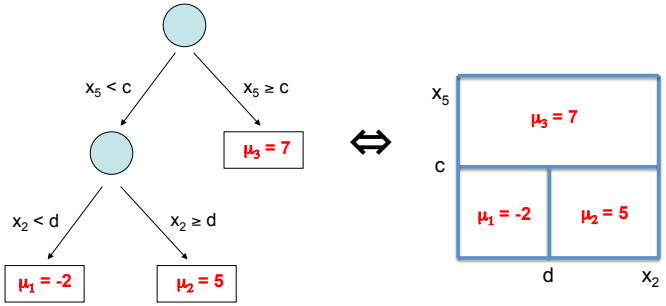
Let $g(x; T, M)$, be a regression tree function that assigns a μ value to x.



A single tree model:

$$y = g(x; T, M) + \epsilon.$$

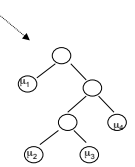
A coordinate view of $g(x; T, M)$



Easy to see that $g(x; T, M)$ is just a step function.

The BART Model

$$Y = g(x;T_1,M_1) + g(x;T_2,M_2) + \ldots + g(x;T_m,M_m) + \sigma z, \quad z \sim N(0,1)$$



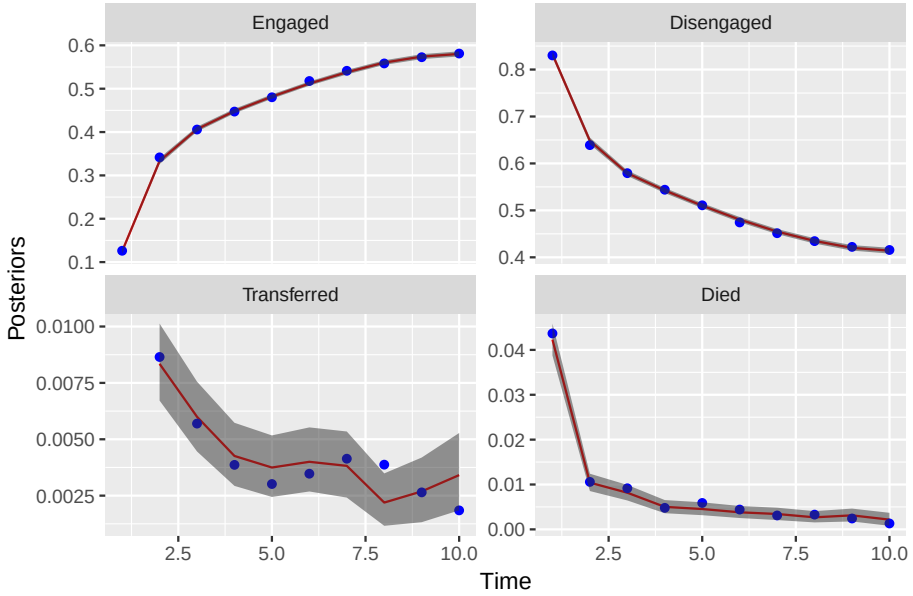
$m = 200, 1000, \ldots,$ big, $\ldots$.

$f(x \mid \cdot)$ is the sum of all the corresponding μ's at each bottom node.

Such a model combines additive and interaction effects.

Step 2: Validate fit of predictive models

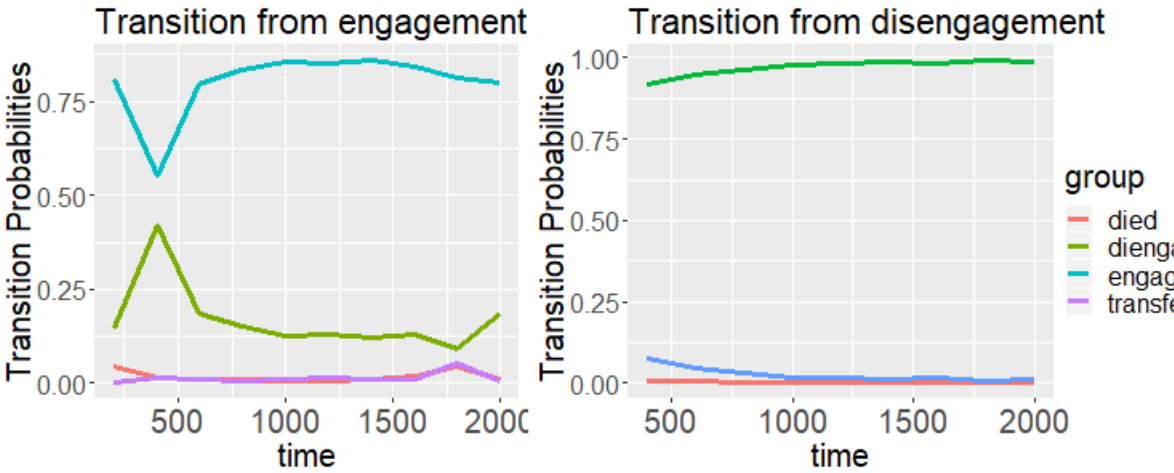
Use posterior predictive distribution for 10K out-of-sample observations



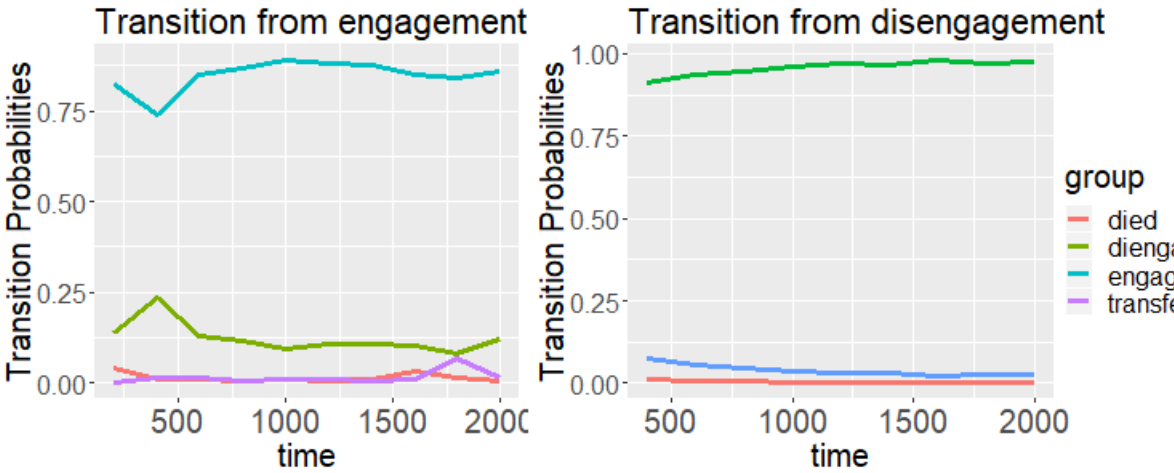
Step 3: Implement G computation to calculate causal effects

Use 10K observations to generate posterior predictive outcomes under different treatment policies using G computation algorithm

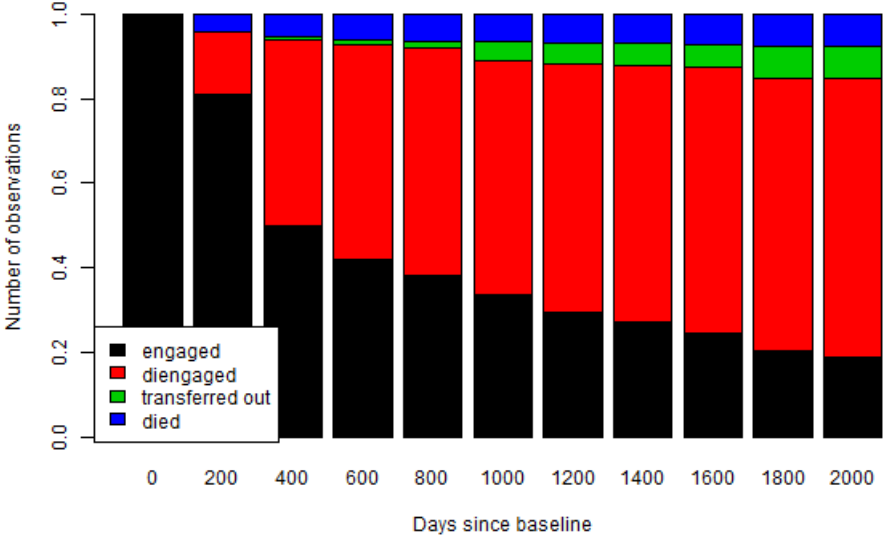
State transitions: Treat when CD4 < 350



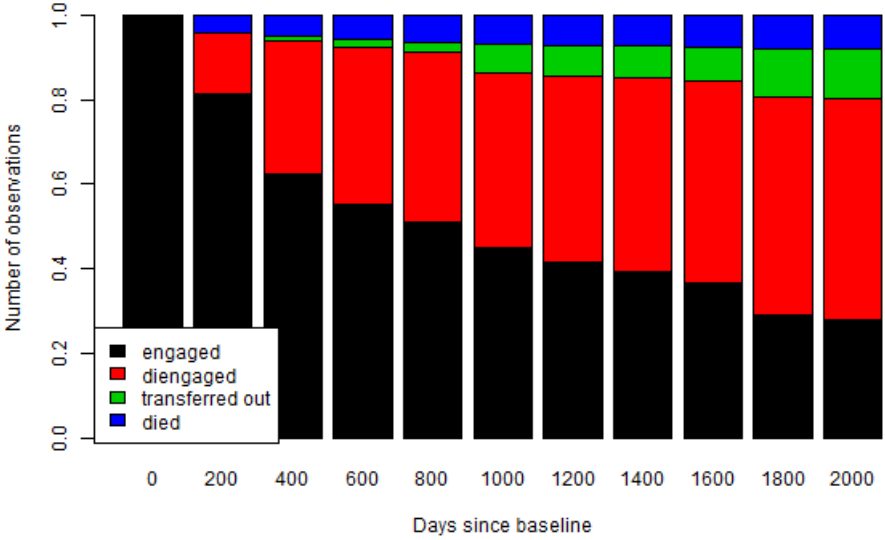
State transitions: Treat immediately



State membership: Treat if CD4<350



State Membership: Treat immediately

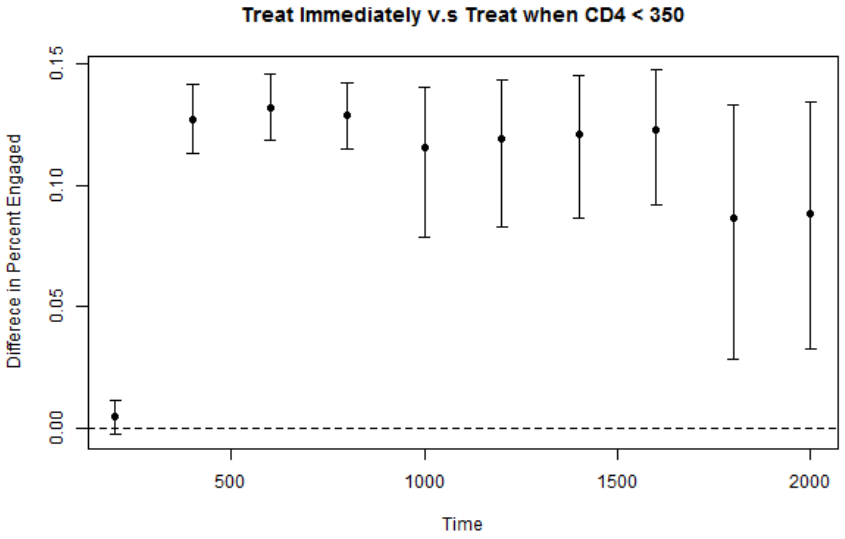


Inferences

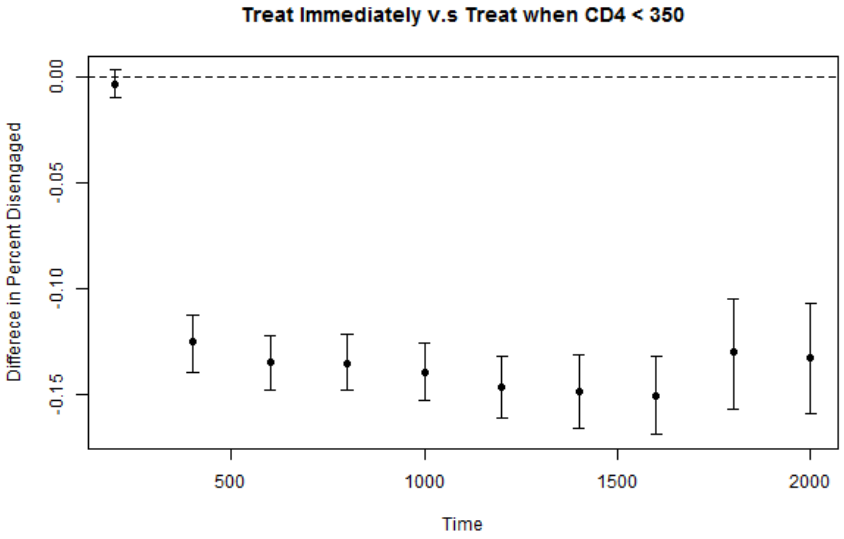
Next few slides:

- Compare proportions in each state over time
- Use rate difference, 95% posterior predictive interval

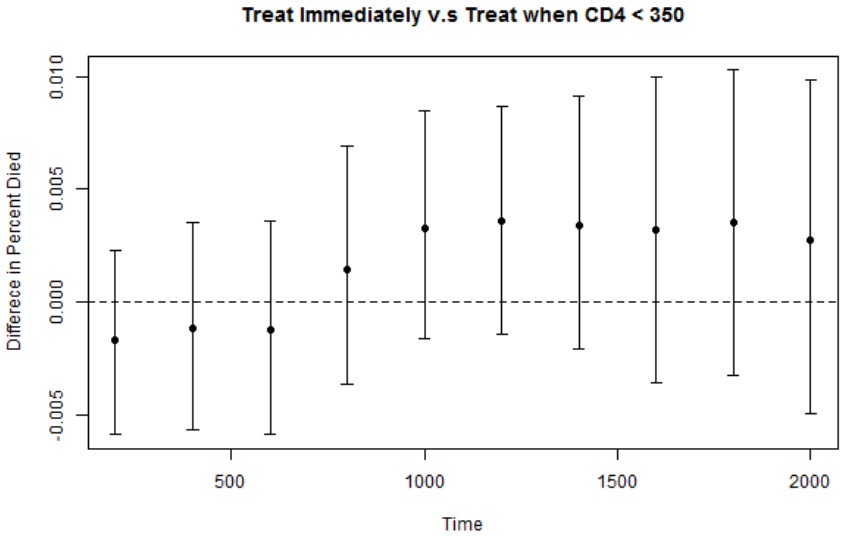
Engaged in care



Disengaged from care



Mortality


Treat Immediately v.s Treat when CD4 < 350
Differece in Percent Died
0.010
0.005
0.000
-0.005
500
1000
1500
2000
Time

Substantive conclusions

- Inferences suggest strong benefit of treatment
  - Higher engagement in care
  - Lower loss to follow up
- Importance of 'disengaged' finding
  - Many of those disengaged are likely to be deceased
  - Estimates available from 'tracing' data
  - Mortality can be as high as 20% (Yiannoutsos et al, 2016)
- Consequence: Preventing LTFU $\Rightarrow$ preventing mortality
  - Quantifying this = data integration problem

Summary

- EHR holds enormous promise for many kinds of inferences
- Illustrations here using HIV care cascade
  - Predictive inference for transitions between states
  - Causal inference for evaluating treatment policies
- Importance of distinguishing between predictive and causal inference
  - Predictive: what will happen next?
  - Causal: what will happen if ... ?

Focus on capacity building in biostatistics

Collaborators on this project

Brown
Yizhen Xu
Rami Kantor, MD
Tao Liu, PhD
Allison DeLong, MS
Hana Lee, PhD (now at FDA)

Indiana U
Beverly Musick, MS

Moi / AMPATH
Ann Mwangi, PhD
Edwin Sang, MS
Victor Omodi, MS

U Florida
Mike Daniels

U Toronto
Paula Braitstein, PhD

Funding

- NIH Grant R01 AI 108441
- NIH Grant D43 TW 010050
- USAID Contract 623-A-00-0-08-00003-00
- Providence-Boston Center for AIDS Research

References to related work

1. Lee H, Wu X, Mugavero MJ et al. (2018). Beyond binary retention in HIV care: Predictors of cyclic process of engagement, disengagement and re-entry to care. AIDS 32, 2217-2225.
2. Lee H, Hogan JW, Genberg BL, et al. (2018). A state transition framework for patient-level modeling of engagement and retention in HIV care using longitudinal cohort data. Statistics in Medicine 37, 302–319.
3. Liu T, Hogan JW, Daniels, MJ et al. (2017). Improved HIV-1 viral load monitoring capacity using pooled testing with marker-assisted deconvolution. Journal of AIDS 75, 580-587.
4. Genberg BL, Hogan JW, Braitstein P (2016). Home testing and counselling with linkage to care. Lancet HIV 3(6):e244-6.
5. Hu L, Hogan JW, Mwangi AM, Siika A (2018). Modeling the causal effect of treatment initiation time on survival: Application to HIV/TB coinfection. Biometrics 74, 703-713.