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

$$g(\cdot) = \text{some function like (inverse) logit}$$

$$\beta = \text{vector of parameters estimated from data}$$

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?
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: \textit{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

HIV CARE CONTINUUM:
THE SERIES OF STEPS A PERSON WITH HIV TAKES FROM INITIAL DIAGNOSIS THROUGH THEIR SUCCESSFUL TREATMENT WITH HIV MEDICATION

DIAGNOSED WITH HIV
ENGAGED OR RETAINED IN CARE
LINKED TO CARE
PRESCRIBED ANTIRETROVIRAL THERAPY

ACHIEVED VIRAL SUPPRESSION

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

Model of state transitions over time

Each arrow represents transition over one time period

Engagement → Transfer-Out → Reported Death → Disengagement → Engagement

Lee et al., Stat Med (2017)
Challenge: Translate patient-level data into states
Challenge: Translate patient-level data into states

Days since enrollment

CD4 count

State
- Engaged
- Disengaged
- Transferred-out
- Deceased

Hogan (JWH @ Brown.edu)
Challenge: Translate patient-level data into states
Summary of available data

![Graph showing data summary]

- **Engaged**
- **Disengaged 1**
- **Disengaged 2**
- **Disengaged 3+**
- **Transferred-out**
- **Deceased**

**State**

- **Engaged**
- **Disengaged 1**
- **Disengaged 2**
- **Disengaged 3+**
- **Transferred-out**
- **Deceased**

**Days since enrollment**

**Number of observations**

Hogan (JWH @ Brown.edu)
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

<table>
<thead>
<tr>
<th>State at $t_{j-1}$</th>
<th>engaged</th>
<th>disengaged</th>
<th>transferred</th>
<th>out</th>
<th>died</th>
</tr>
</thead>
<tbody>
<tr>
<td>engaged</td>
<td>.85</td>
<td>.13</td>
<td>.01</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>disengaged</td>
<td>.06</td>
<td>.94</td>
<td></td>
<td></td>
<td>.01</td>
</tr>
<tr>
<td>transferred</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>died</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

- Assumes constant rate over time
- Death and transfer-out rates under-estimated (need tracing data)
Aggregated Transition Rates from AMPATH Data

![Diagram showing transition rates between engagement, disengagement, transfer-out, and reported death with probabilities and significance levels indicated.]
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)
- $x_j$: vector of covariates (some time-varying)
- Multinomial probabilities for transition $k \rightarrow \ell$ at each time interval

\[ p_{jkl}(x_j) = P(S_j = \ell \mid S_{j-1} = k, x_j) \]
Prediction model: Observed-data regressions

Multinomial regression for longitudinal data

$$\log \left\{ \frac{p_{jk\ell}(x_j)}{p_{jkL}(x_j)} \right\} = x_j^T \beta_{jk\ell} \quad \ell = 1, \ldots, L - 1$$

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

$$x_j = \text{covariates at time } j$$
Prediction model: Observed-data regressions

**Covariates:** $x_j = \text{vector of covariates observed just prior to } t_j$

- 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)

<table>
<thead>
<tr>
<th>State at $t_{j-1}$</th>
<th>Engaged</th>
</tr>
</thead>
<tbody>
<tr>
<td>State at $t_j$</td>
<td>Disengaged</td>
</tr>
<tr>
<td>age</td>
<td>$-0.02^*$</td>
</tr>
<tr>
<td>male</td>
<td>$0.18^*$</td>
</tr>
<tr>
<td>Enrollment Year</td>
<td>$0.011$</td>
</tr>
<tr>
<td>TravelTime</td>
<td>$-0.01$</td>
</tr>
<tr>
<td>WHO stage</td>
<td>$0.05^*$</td>
</tr>
<tr>
<td>Married</td>
<td>$-0.15^*$</td>
</tr>
<tr>
<td>Height</td>
<td>$-0.002$</td>
</tr>
<tr>
<td>log Weight</td>
<td>$-0.26^*$</td>
</tr>
<tr>
<td>undetectable VL</td>
<td>$-0.62$</td>
</tr>
<tr>
<td>Previous ARV</td>
<td>$-0.38^*$</td>
</tr>
<tr>
<td>CD4 Update</td>
<td>$-2.20^*$</td>
</tr>
<tr>
<td>latest log CD4+1</td>
<td>$-0.20^*$</td>
</tr>
</tbody>
</table>
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

\[ S_j = \text{state membership at time } t_j \]
\[ a_j = \text{treatment assigned at time } t_j \]
\[ \bar{a}_j = (a_0, \ldots, a_j) \]
\[ P_{\bar{a}_j}(S_j) = \text{distribution of } S_j \text{ under regime } \bar{a}_j \]

To compare two different regimes \( \bar{a} \) and \( \bar{a}^* \), want to compare

\[ P_{\bar{a}}(S_J) \quad \text{and} \quad P_{\bar{a}^*}(S_J) \]

Example: ‘treat immediately’ is the regime

\[ \bar{a}_J = (1, 1, 1, \ldots, 1) \]
What we need to estimate *causal* models

- **Observed data**
  
  \[
  S_j = \text{state at time } j \\
  X_j = \text{time-varying confounders (CD4)} \\
  V = \text{baseline confounders (age, gender, site, CD4)} \\
  A_j = \text{observed ART status at time } j
  \]

- **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:** $P_{a_0}(S_1)$ when $a_0 = 1$
(state membership distribution if everyone receives treatment at baseline)

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

**G computation:**

$$P_1(S_1) = \int P(S_1 \mid A_0 = 1, X_0, V) \ P(X_0, V) \ d(X_0, V)$$

**Implementation**

$$\hat{P}_1(S_1) = (1/n) \sum_{i=1}^{n} \hat{P}(S_1 \mid A_0 = 1, X_{0i}, V_i)$$
G computation for estimating causal quantities

**Target:** \( P_{a_0,a_1}(S_2) \)

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:** \( P_{a_0,a_1}(S_2) \)

\[
P_{a_0,a_1}(S_2) = \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \, \text{d}(S_1, X_1, X_0, V)
\]

Plug in fitted models for state transitions
How to use observed-data models as plug-ins

Target: \( P_{a_0, a_1}(S_2) \)

\[
P_{a_0, a_1}(S_2) = \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \cdot P(X_1 \mid A_0 = a_0, X_0, V, S_1) \cdot P(S_1 \mid A_0 = a_0, X_0, V) \cdot P(X_0, V) \cdot d(S_1, X_1, X_0, V)
\]

Plug in fitted models for time-varying covariates
How to use observed-data models as plug-ins

Target: $P_{a_0, a_1}(S_2)$

$$P_{a_0, a_1}(S_2) = \int P(S_2 \mid A_1 = a_1, X_1, S_1, V) \cdot P(X_1 \mid A_0 = a_0, X_0, V, S_1) \cdot P(S_1 \mid A_0 = a_0, X_0, V) \cdot P(X_0, V) \cdot d(S_1, X_1, X_0, V)$$

Fix treatment regime or policy $a_0, a_1$
How to use observed-data models as plug-ins

**Target:** \( P_{a_0,a_1}(S_2) \)

\[
P_{a_0,a_1}(S_2) = \int P(S_2 | A_1 = a_1, X_1, S_1, V) \cdot P(X_1 | A_0 = a_0, X_0, V, S_1) \cdot P(S_1 | A_0 = a_0, X_0, V) \cdot P(X_0, V) \cdot d(S_1, X_1, X_0, V)
\]

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
- Use multinomial BART

Time varying covariate models
- Use continuous-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 $\mu$'s.

Let $g(x; T, M)$, be a regression tree function that assigns a $\mu$ 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

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

\[ m = 200, 1000, \ldots, \text{big}, \ldots \]

\( f(x | \cdot) \) is the sum of all the corresponding \( \mu \)'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

![Graph showing state membership over days since baseline]

- **engaged**
- **diengaged**
- **transferred out**
- **died**

Number of observations vs Days since baseline.
Inferences

Next few slides:

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

Treat Immediately v.s Treat when CD4 < 350

Difference in Percent Engaged

Time
Disengaged from care
Mortality

Treat Immediately v.s Treat when CD4 < 350

Difference in Percent Died

-0.005 -0.000 0.000 0.005 0.010

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)

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

**U Florida**
Mike Daniels

**Indiana U**
Beverly Musick, MS

**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


