Using Electronic Health Records Data for Causal Inferences about the HIV Care Cascade

Joseph Hogan

Department of Biostatistics
School of Public Health
Brown University

June 9, 2016
Acknowledgments

- Co-authors and collaborators
  - Paula Braitstein, U Toronto
  - Becky Genberg, Brown U
  - Rami Kantor, Brown U
  - Kirwa Kipruto, Brown U and Moi University
  - *Hana Lee, Brown U
  - Tao Liu, Brown U
  - Beverly Musick, Indiana U
  - Ann Mwangi, Moi University (Kenya)
  - Fatma Some, Moi University (Kenya)
  - Yizhen Xu, Brown U

- Funding
  - NIH grants R01-AI-108441, P30-AI-42853
  - USAID Contract 623-A-00-0-08-00003-00
Outline

- The HIV Care Cascade
  - Description
  - Care cascade as a way to frame health outcome targets

- Summarizing and modeling the HIV care cascade
  - Simple summaries
  - Mathematical models

- Big data opportunity: statistical modeling using EHR
  - Opportunities and challenges
  - Specific issues with EHR data
Outline

- Case study: Estimate efficacy of ‘test and treat’
  - Use EHR data from AMPATH
  - Statistical methods for causal inference

- Challenges and opportunities
  - Statistical modeling and mathematical modeling for complex systems
  - Opportunities provided by EHR
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
HIV care in LMIC: Evolution of treatment recommendations

**SCENARIOS OF ANTIRETROVIRAL TREATMENT ELIGIBILITY: WHO VISION**

Estimated millions of people eligible for ARV in LMIC in 2012

1. CD4 ≤ 200
   - Recommended since 2003

2. CD4 ≤ 350
   - Recommended since 2010

3. CD4 ≤ 350
   - + TB/HIV
   - + TB/HBV

4. CD4 ≤ 500
   - + TB/HIV
   - + TB/HBV

5. All HIV+
   - "Test and treat"

**Scenarios of ARV eligibility**

- ART regardless of CD4 count for:
  - Serodiscordant couples
  - Pregnant women
  - Children < 5 years

Source: WHO 2014
Cascade-based targets: 90-90-90

UNAIDS Report, 2014

90-90-90

An ambitious treatment target to help end the AIDS epidemic

UNAIDS
90-90-90: A schematic

- Find new cases
- 90% linked to care
- 90% initiated on treatment
- 90% with viral suppression
Challenges: Rigor vs clarity

- Programs / funding agencies need digestible summaries
  - Track progress
  - Evaluate new policies / interventions

- Data are typically highly complex, come from multiple sources
  - EHR – experiential data from clinical care
  - Country-level summaries from ministries of health
  - Others ...

- The care cascade is a complex process
  - How to represent using a model?
  - How to draw principled inferences from the model?
  - How to translate model outputs into simple summaries?
Summarizing and modeling the cascade

- Summaries using aggregated data
  - Proportion falling in each cascade category
  - Rates of transition or progression through the cascade
  - Can be for a specific care program, across country or region

- Analyses of specific aspects of the cascade
  - Identify correlates of transitions through cascade
  - Assess effect of specific intervention or policy

- Model-based representation of entire cascade
  - Mathematical models
  - Statistical models
FIGURE 1. Estimated percentage of persons living with HIV infection,* by outcome along the HIV care continuum — United States, 2011

Abbreviations: HIV = human immunodeficiency virus; ART = antiretroviral therapy.
* N = 1,201,100.

CDC MMWR 2011;60:1618-1623.
Challenges in modeling the full cascade

- How to write down the model?
  - Data-generating model is complex
  - Multiple states
  - Progression not ‘linear’

- Prevailing mode of analysis: microsimulation from mathematical models
Mathematical models of the care cascade

- Typically assume underlying parametric model
  - Stage specific components
  - Linked together to form model for entire cascade

- Components of the model are informed by different data sources
  - Data from individual programs
  - Surveillance data
  - National registries
  - Others

- Model is calibrated against target outcomes (e.g. national prevalence rates)

- Intervention effects calculated via simulation
Example: Compare home-testing and treatment strategies
Ying et al., Lancet HIV, 2016

- Model of individual-level progression through 9 HIV disease states
- Simulates HIV incidence and prevalence over 45 year period
- Uses model to capture effect of specific interventions

---

Home testing and counselling to reduce HIV incidence in a generalised epidemic setting: a mathematical modelling analysis

Roger Ying, Monisha Sharma, Connie Celum, Jared M Baeten, Heidi van Rooyen, James P Hughes, Geoff Garnett, Ruanne V Barnabas

Summary
Background  Home HIV testing and counselling (HTC) achieves high levels of HIV testing and linkage to care. Periodic home HTC, particularly targeted to those with high HIV viral load, might facilitate expansion of antiretroviral therapy (ART) coverage. We used a mathematical model to assess the effect of periodic home HTC programmes on HIV incidence in KwaZulu-Natal, South Africa.
HIV Natural History:
The natural history of HIV infection is modeled in stages defined by CD4 count and viral load as shown in Figure S1.

Figure S1. Model transition diagram. A diagram of the natural history of HIV infection. All movement is in one direction except for enrollment in and dropout from interventions from ART.
The ODEs for the nine disease states are:

\[
\frac{dX_{a,r}^{g,0,0}(t)}{dt} = b_{r,0}^{g,0}(t) + \sigma_{a,r}^{g,0} X_{a,r}^{g,0,0}(t) - \left( \mu_{a}^{g} + \lambda_{a,r}^{0}(t) + \pi_{a,r}^{0,0}(t) \right) X_{a,r}^{g,0,0}(t)
\]

\[
\frac{dX_{a,r}^{g,1,0}(t)}{dt} = b_{r,0}^{g,1}(t) + \lambda_{a,r}^{0} X_{a,r}^{g,0,0}(t) + \psi_{0}^{1,0}(t) X_{a,r}^{g,1,0}(t) + \psi_{0}^{1,1}(t) X_{a,r}^{g,1,0}(t) + \psi_{1}^{1,0}(t) X_{a,r}^{g,1,0}(t) + \psi_{1}^{1,1}(t) X_{a,r}^{g,1,0}(t) + \sigma_{a,r}^{g,1}(t) X_{a,r}^{g,9,6}(t) - \left( \mu_{a}^{g} + \alpha_{a}^{g,1} + \nu_{1} + \pi_{a,r}^{1,0}(t) \right) X_{a,r}^{g,1,0}(t)
\]

\[
\frac{dX_{a,r}^{g,2,0}(t)}{dt} = (\nu_{1} + \omega_{y-1}) X_{a,r}^{g,1,0}(t) + \sigma_{a,r}^{g,2} X_{a,r}^{g,9,6}(t) - \left( \mu_{a}^{g} + \alpha_{a}^{g,2} + \nu_{2} + \omega_{v} + \pi_{a,r}^{2,0}(t) \right) X_{a,r}^{g,2,0}(t)
\]

\[
\frac{dX_{a,r}^{g,3,0}(t)}{dt} = (\nu_{2} + \omega_{y-1}) X_{a,r}^{g,2,0}(t) + \sigma_{a,r}^{g,3} X_{a,r}^{g,9,6}(t) - \left( \mu_{a}^{g} + \alpha_{a}^{g,3} + \nu_{3} + \omega_{v} + \pi_{a,r}^{3,0}(t) \right) X_{a,r}^{g,3,0}(t)
\]

\[
\frac{dX_{a,r}^{g,4,0}(t)}{dt} = (\nu_{3} + \omega_{y-1}) X_{a,r}^{g,3,0}(t) + \sigma_{a,r}^{g,4} X_{a,r}^{g,9,6}(t) - \left( \mu_{a}^{g} + \alpha_{a}^{g,4} + \nu_{4} + \omega_{v} + \pi_{a,r}^{4,0}(t) \right) X_{a,r}^{g,4,0}(t)
\]

\[
\frac{dX_{a,r}^{g,5,0}(t)}{dt} = (\nu_{4} + \omega_{y-1}) X_{a,r}^{g,4,0}(t) + \sigma_{a,r}^{g,5} X_{a,r}^{g,9,6}(t) - \left( \mu_{a}^{g} + \alpha_{a}^{g,5} + \nu_{5} + \omega_{v} + \pi_{a,r}^{5,0}(t) \right) X_{a,r}^{g,5,0}(t)
\]

\[
\frac{dX_{a,r}^{g,6,0}(t)}{dt} = (\mu_{a}^{g} + \psi_{0}^{g,0,0}(t) + \pi_{a,r}^{0,0}(t)) X_{a,r}^{g,6,0}(t)
\]

\[
\frac{dX_{a,r}^{g,7,0}(t)}{dt} = \pi_{a,r}^{g,0,0}(t) X_{a,r}^{g,5,0}(t) - \left( \sigma_{a,r}^{g,0} + \mu_{a}^{g} + \psi_{0}^{1,0} \lambda_{a,r}^{1}(t) \right) X_{a,r}^{g,7,0}(t)
\]

\[
\frac{dX_{a,r}^{g,8,0}(t)}{dt} = \pi_{a,r}^{g,0,0}(t) X_{a,r}^{g,0,0}(t) - \left( \sigma_{a,r}^{g,0} + \mu_{a}^{g} + \psi_{1}^{0,0} \lambda_{a,r}^{1}(t) \right) X_{a,r}^{g,8,0}(t)
\]

\[
\frac{dX_{a,r}^{g,9,6}(t)}{dt} = \sum_{v=1}^{5} \sum_{d=1}^{5} \left[ \pi_{a,r}^{g,d,v}(t) X_{a,r}^{g,d,v}(t) - \left( \sigma_{a,r}^{g,d} + \mu_{a}^{g} \right) X_{a,r}^{g,9,6}(t) \right]
\]
Some model input variables

<table>
<thead>
<tr>
<th>$\pi_{a,r}^{g,d,v}(t)$</th>
<th>The coverage of PrEP ($d = 0$), ART ($d = 1, \ldots, 5$), circumcision ($d = 6$), condom use among HIV-negative persons ($d = 7$), condom use among PrEP users ($d = 8$), and condom use among ART users ($d = 9$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{a}^{g,d}$</td>
<td>The HIV-associated mortality</td>
</tr>
<tr>
<td>$\nu_{d}$</td>
<td>The rate of progressing from CD4 state $d$ to $d + 1$</td>
</tr>
<tr>
<td>$\omega_{d}$</td>
<td>The rate of progressing from VL state $v$ to $v + 1$</td>
</tr>
<tr>
<td>$\psi_{d}$</td>
<td>The reduction in HIV transmission due to circumcision ($d = 0$), PrEP ($d = 1$), ART ($d = 2$), or condom use ($d = 3$)</td>
</tr>
</tbody>
</table>
### Source of selected inputs

<table>
<thead>
<tr>
<th>Duration of disease by CD4 count</th>
<th>Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute</td>
<td>0.25 year</td>
<td>Hontelez et al(^1^4)</td>
</tr>
<tr>
<td>&gt;500 cells per μL</td>
<td>1.88 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>500–350 cells per μL</td>
<td>1.22 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>350–200 cells per μL</td>
<td>5.90 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>&lt;200 cells per μL</td>
<td>1.96 years</td>
<td>Badri et al(^7)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Duration of disease by HIV viral load</th>
<th>Value</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute</td>
<td>0.25 years</td>
<td>Hontelez et al(^1^4)</td>
</tr>
<tr>
<td>&lt;1000 copies per mL</td>
<td>3.13 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>1000–10 000 copies per mL</td>
<td>1.99 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>10 000–50 000 copies per mL</td>
<td>4.40 years</td>
<td>Celum et al(^5), Baeten et al(^6)</td>
</tr>
<tr>
<td>&gt;50 000 copies per mL</td>
<td>1.44 years</td>
<td>Estimated</td>
</tr>
</tbody>
</table>

**Costs**

- Annual home HTC with community care workers: $28.06 per HIV-positive person
- $8.22 per HIV-negative person

Parameters based on the HTC study and other published work. For parameters with varying estimates, we chose values that best fit our data. HTC = home HIV testing and counselling. *2014 US dollars.

**Table 1**: Key parameters used in model
Comparison of treatment initiation policies

HIV Incidence with ART Expansion

- Current
- CD4<500
- Test and Treat
- No Intervention

Figure S6. HIV incidence with optimistic ART coverage scenarios.
Mathematical modeling: Summary

**Strengths**
- Can represent highly complex system in unified model
- Facilitates cost effectiveness analysis
- Calibration keeps model tied to observed data
- **Method of evidence synthesis**

**Limitations**
- Representative of specific population?
- Calibration identifies *one* model consistent with observed data
- Generating causal effects?
- Observed data may be heavily leveraged
- Issues with uncertainty quantification
Big data opportunity: EHR data

Large-scale EHR data can enable statistical models of cascade

Opportunities

- Longitudinal follow up on 1000’s of individuals
- Sample from well-defined population
- Reflects actual care setting
- Possible to develop coherent statistical models of full cascade

Challenges

- Irregular observation times
- How to operationalize states of the cascade
- Dropout, loss to follow up, misclassification
- Data are observational
Example: AMPATH Program in western Kenya

- AMPATH: Academic Model Providing Access to Healthcare
- PEPFAR-funded HIV care program based in Eldoret, Kenya
- Over 250,000 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
Goal: Compare two treatment policies
- Treat upon enrollment (test and treat)
- Treat when CD4 drops below 350

Outcome: Care state following enrollment
- Engaged in care (initial, and at each visit)
- Disengaged from care (one visit to the next)
- Deceased
- Lost to follow up
Statistical model of HIV cascade using EHR data

- Specify statistical model of transitions between states
  - Discrete-time state-space model
  - State membership depends on covariates

- Define state membership from messy data

- Causal structural model to compare treatment policies
  - Use G computation to estimate causal effects
  - Compare / contrast to mathematical modeling
Operationalize progression through care cascade:
State transitions between $t_{j-1}$ and $t_j$
Transition matrix representation

\[ S_j = \text{state at time } t_j \]
\[ p_{jk\ell} = P(S_j = \ell \mid S_{j-1} = k) \]

<table>
<thead>
<tr>
<th>State at ( t_{j-1} )</th>
<th>Engaged</th>
<th>Disengaged</th>
<th>LTFU</th>
<th>Death</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaged</td>
<td>( p_{j11} )</td>
<td>( p_{j12} )</td>
<td>0</td>
<td>( p_{j14} )</td>
<td>( p_{j15} )</td>
</tr>
<tr>
<td>Disengaged</td>
<td>( p_{j21} )</td>
<td>( p_{j22} )</td>
<td>( p_{j23} )</td>
<td>( p_{j24} )</td>
<td>0</td>
</tr>
<tr>
<td>LTFU</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Death</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Transfer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Can incorporate covariate effects via regression

- Use multinomial regression for longitudinal data

\[
\log \{ p_{jk\ell}(x_j) / p_{jkL}(x_j) \} = x_j^T \beta_{jk\ell} \quad \ell = 1, \ldots, L - 1
\]

- \( x_j \) = vector of covariates observed just prior to \( t_j \)
  - CD4 count
  - age, gender
  - treatment
  - enrollment year

- Coefficient \( \beta_{jk\ell} \) is a log relative rate ratio
Defining state membership

Start at enrollment, ascertain status every 200 days

- **Engaged**
  - Everyone engaged at enrollment
  - Remain engaged if visit within 200 day window

- **Disengaged**
  - No visit within 200 day window
  - Can re-engage if new visit appears in next window

- **LTFU**
  - Two consecutive windows disengaged; and
  - No further record of visits
Organizing data into states

(t=0)

(enrollment)
Organizing data into states

(t=0)
(t=200)
(t=400)
(t=600)

(enrollment)
Organizing data into states

(t=0) (enrollment) (t=200) Engaged (t=400) (t=600)

(enrollment) Engaged
Organizing data into states

(t=0) (enrollment) (t=200) Engaged (t=400) Engaged (t=600)
Organizing data into states

(t=0) (enrollment) (t=200) Engaged (t=400) Engaged (t=600) Disengaged
Summary of available data

Unique individual enrollments: 57,596

Unique visits: 1,493,150

Observations used in analysis: 321,834
## Transition rate estimates

**Time-aggregated estimates**

57,000+ individuals  
Enrolled between 6/2008 – 9/2012

<table>
<thead>
<tr>
<th>State at $t_{j-1}$</th>
<th>Engaged</th>
<th>Disengaged</th>
<th>LTFU</th>
<th>Death</th>
<th>Transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engaged</td>
<td>.86</td>
<td>.11</td>
<td>0</td>
<td>.02</td>
<td>.01</td>
</tr>
<tr>
<td>Disengaged</td>
<td>.12</td>
<td>.54</td>
<td>.33</td>
<td>.01</td>
<td>0</td>
</tr>
<tr>
<td>LTFU</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Death</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Transfer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Calculating state probabilities under differential follow up

**Assumptions**
- First-order Markov structure
- Length of follow up is unrelated to outcome

**Procedure**
- Use data to estimate $P(S_1)$, $P(S_2 | S_1)$, $P(S_3 | S_2)$, ...
- Calculate marginal probabilities

\[
\hat{P}(S_j) = \sum_{S_1, \ldots, S_{j-1}} \hat{P}(S_1) \prod_{k} \hat{P}(S_k | S_{k-1})
\]
Causal structural model to compare treatment policies

**Question:** Relative to CD4-specific treatment rules, how does ‘treat immediately’ impact progression through the care cascade?

**Comparison regimes:**
- Policy 1: Treat immediately (‘test and treat’)
- 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

\[ \bar{a}_J = (1, 1, 1, \ldots, 1) \]
Statistical and computational issues

1. Differential lengths of follow up
   - Depends on enrollment time
   - Make MAR assumption to estimate state membership probabilities

2. Treatment not randomized
   - Adjust for age, gender, (time-varying) CD4
   - Use G computation algorithm

3. Requires high-dimensional integration over time-varying covariates
   - Need model for covariates
   - Use Monte Carlo integration
   - Similarities to mathematical modeling
Schematic: Evolution of longitudinal data

$S = \text{state membership}$

$X = \text{CD4 count}$

$A = \text{treatment}$

$V = \text{gender, age at enrollment}$
Assumptions needed

- Treatment is randomly allocated for individuals sharing same observed-data history
  - CD4, age, gender, treatment, state
  - (Keeping it simple for this example)

- Length of follow up depends only on observed-data history

- First-order Markov dependence
Implementation

- CD4 model has 3 categories
  - $< 350$
  - $\geq 350$
  - missing

- $A_j$ represents most recently observed treatment status

- Fit sequence of observed-data models for $j = 1, \ldots, J$

\[
P(S_j \mid A_{j-1}, X_{j-1}, V)
\]
\[
P(X_j \mid A_{j-1}, X_{j-1}, S_{j-1}, V)
\]

- Use G computation implemented with Monte-Carlo simulation
G computation for estimating causal quantities

Method of imputing ‘counterfactual’ outcomes under different treatment regimes

- Specify sequence of observed-data models
  - Outcome models
  - Covariate models
  - Models can be arbitrarily complex (machine learning)

- Use these models to generate predicted outcome under specific regimes
  - Requires averaging these over regime-specific covariate paths
  - High-dimensional integration over longitudinal data

- Integral calculated using Monte Carlo simulation
  - Similarities to microsimulation
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:

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

Implementation

$$\hat{P}_1(S_1) = (1/n) \sum_{i=1}^{n} \hat{P}(S_1 | A_0 = 1, X_{0i}, V_i)$$
Treat upon enrollment ('test and treat')
Next few slides:

- Compare proportions in each state over time

- Use rate difference, 95% confidence interval

- Based on 100 bootstrap samples (about 2 hrs on iMac)
Engaged in care
Mortality

![Chart showing mortality over days since enrollment.]
Loss to follow up
Substantive conclusions

- Inferences suggest strong benefit of treatment
  - Higher engagement in care
  - Lower loss to follow up

- Importance of LTFU finding
  - Many of those LTFU 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
Compare and contrast

- **G computation using statistical models**
  - Observed data typically sampled from target population
  - Based on a sequence of (simple) models fit to observed data
  - These models can be checked from data
  - Bigger data $\Rightarrow$ more complex models
  - Integration carried out using simulation

- **Microsimulation from mathematical model**
  - Data come from multiple sources
  - Based on a single model of a complex system
  - Questions about whether ‘fitted’ model corresponds to a data-generating mechanism for a target population
  - Simulation-based also, but does the simulation integrate in the right way to assess causal effects?
Mathematical and statistical modeling for causal inference

- Mathematical Modeling: Focus on the **model** (more model, less data)

- Statistical modeling: Focus on the **data** (more data, less model)

- Importance of EHR data
  - More opportunities to make data-driven statistical modeling the starting point
  - Yields decisions / inferences that are ‘closer to the data’
  - ‘Gold mine’: requires right tools to extract the gold
Complexity
Resolution (Data) Stat model
Math model

Joseph Hogan (JWH @ Brown.edu) HIV Care Cascade
June 9, 2016 62 / 78
Opportunities

- Data quality
- Methodology
- Representing and evaluating evidence
Opportunities: Data quality

- Good data $\Rightarrow$ good evidence

- AMPATH has developed high-functioning EHR
  - ‘This only works at AMPATH’
  - Why not similar systems for other LMIC?

- Opportunity for collaboration between statistics and informatics
  - e.g., embedding statistical analyses in EHR systems
  - e.g., refining EHR design to respond to statistical needs
  - real-time updates related to benchmarks
  - reinforcement learning
Opportunities: Methodology

**Currently:** Mathematical models influence trt recommendations for LMIC

- EHR presents important new opportunities
  - Larger databases $\Rightarrow$ more complex statistical models
- EHR data are **rich** but **messy**
- Lots of prospectors circling the gold mine
- **Statistical principles** could hardly be more important!
  - Distinguishing causal from predictive inference
  - Understanding what the data can and cannot tell us
Opportunities: Methodology

Can mathematical modeling techniques enrich statistical models?

- Import / represent missing information
  - Incomplete covariate histories
  - Unmeasured confounders

- Integrate outside data
  - Bias adjustment using tracing data (e.g., to correct mortality estimates)
  - Data on individual-level behavior (e.g., social networks)

- Bayesian platform provides formal mechanism for this
  - Weight model inputs according to strength of evidence

Goal: Bend the complexity/resolution curve
Opportunities: Generating and grading evidence

- EHR can be used for
  - Monitoring outcomes
    (Lee, Hogan, et al., 2016 CROI)
  - Evaluation of treatment strategies
    (Hu, Hogan, et al., JASA in revision)
  - Development of patient-level decision support
    (Liu et al., JASA 2013)

- Requires framework for grading evidence
  - Well-defined population?
  - Distinguish sources of uncertainty?
  - Check model fit?
  - External validation?
  - Reproducibility?
Thank you!
Temporal trends in transition from engaged
Temporal trends in transition from **disengaged**

![Graph showing trends in disengaged status](image.png)
### Covariate effects: transition from engaged

(Adjusted for calendar year)

<table>
<thead>
<tr>
<th></th>
<th>Engaged</th>
<th>Disengaged</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transition probability</strong></td>
<td>.86</td>
<td>.12</td>
<td>.01</td>
</tr>
<tr>
<td>Male</td>
<td>—</td>
<td>1.19*</td>
<td>1.76*</td>
</tr>
<tr>
<td>Age &gt; 35</td>
<td>—</td>
<td>0.67*</td>
<td>1.01</td>
</tr>
<tr>
<td>CD4&lt;350, ART−</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CD4&lt;350, ART+</td>
<td>—</td>
<td>0.20*</td>
<td>0.42*</td>
</tr>
<tr>
<td>CD4≥350, ART−</td>
<td>—</td>
<td>0.38*</td>
<td>0.10*</td>
</tr>
<tr>
<td>CD4≥350, ART+</td>
<td>—</td>
<td>0.11*</td>
<td>0.09*</td>
</tr>
</tbody>
</table>
### Covariate effects: transition from disengaged

(Adjusted for calendar year)

<table>
<thead>
<tr>
<th></th>
<th>Engaged</th>
<th>Disengaged</th>
<th>LTFU</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transition probability</strong></td>
<td>.12</td>
<td>.54</td>
<td>.33</td>
<td>.01</td>
</tr>
<tr>
<td>Male</td>
<td>—</td>
<td>0.96</td>
<td>0.92*</td>
<td>1.05</td>
</tr>
<tr>
<td>Age &gt; 35</td>
<td>—</td>
<td>0.94*</td>
<td>0.98</td>
<td>1.23</td>
</tr>
<tr>
<td>CD4&lt;350, ART−</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CD4&lt;350, ART+</td>
<td>—</td>
<td>0.40*</td>
<td>0.33*</td>
<td>0.67*</td>
</tr>
<tr>
<td>CD4≥350, ART−</td>
<td>—</td>
<td>0.81*</td>
<td>0.71*</td>
<td>0.35*</td>
</tr>
<tr>
<td>CD4≥350, ART+</td>
<td>—</td>
<td>0.34*</td>
<td>0.25*</td>
<td>0.12*</td>
</tr>
</tbody>
</table>
G computation for estimating causal quantities

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

\[
P_{a_0,a_1}(S_2) = \int P(S_2 \mid A_0 = a_0, A_1 = a_1, X_0, X_1, S_1, V) \\
P(X_1 \mid A_0 = a_0, X_0, V, S_1) \\
P(S_1 \mid A_0 = a_0, X_0, V) \\
P(X_0, V) \\
d(S_1, X_1, X_0, V)
\]
G computation for estimating causal quantities

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

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

First-order Markov assumption

- $P(X_1 \mid A_0 = a_0, X_0, V, S_1)$
- $P(S_1 \mid A_0 = a_0, X_0, V)$
- $P(X_0, V)$
- $P(S_1, X_1, X_0, V)$
- $d(S_1, X_1, X_0, V)$
Target: \( P_{a_0,a_1}(S_2) \)

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

Multinomial regression models for state transitions
G computation for estimating causal quantities

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

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

Need to specify new model for CD4 evolution
G computation for estimating causal quantities

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

\[
\begin{align*}
\tilde{S}_{1i} & \sim \hat{P}(S_1 | A_0 = a_0, X_{0i}, V_i) \\
\tilde{X}_{1i} & \sim \hat{P}(X_1 | A_0 = a_0, X_{0i}, V_i, \tilde{S}_{1i})
\end{align*}
\]

\[
\hat{P}_{a_0, a_1}(S_2) = \frac{1}{n} \sum_{i} \hat{P}(S_2 | A_1 = a_1, \tilde{X}_{1i}, \tilde{S}_{1i}, V_i)
\]

Calculate via Monte Carlo simulation based on fitted models