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

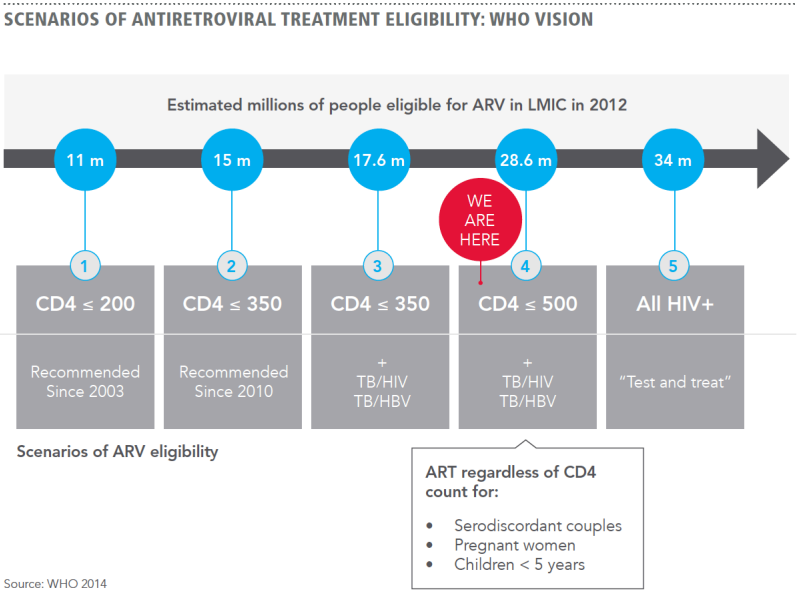
HIV care cascade

- 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



Cascade-based targets: 90-90-90

UNAIDS Report, 2014

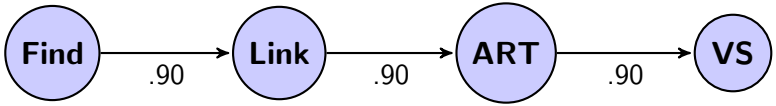
90-90-90

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



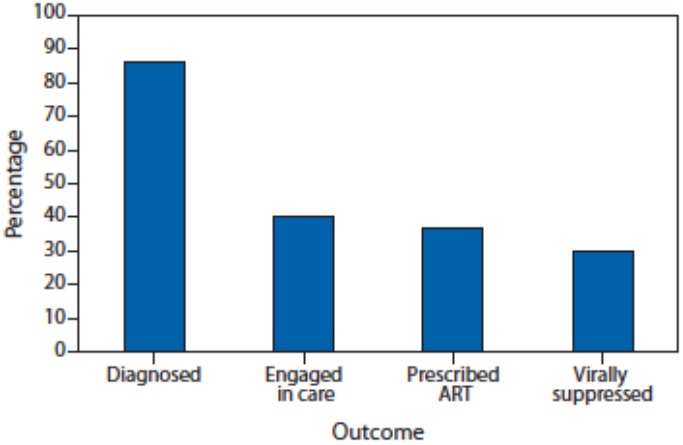
Challenges: Rigor vs clarity

- 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

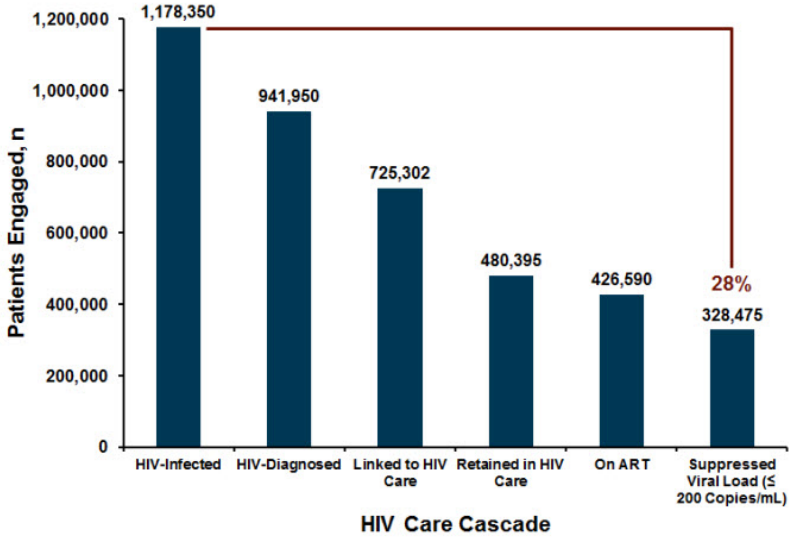
- Summaries using aggregated data
  - Proportion falling in each cascade category
  - Rates of transition or progression through the cascade
  - Can be for a specific care program, across country or region
- Analyses of specific aspects of the cascade
  - Identify correlates of transitions through cascade
  - 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 Morbidity and mortality weekly report 63.47 (2014): 1114.

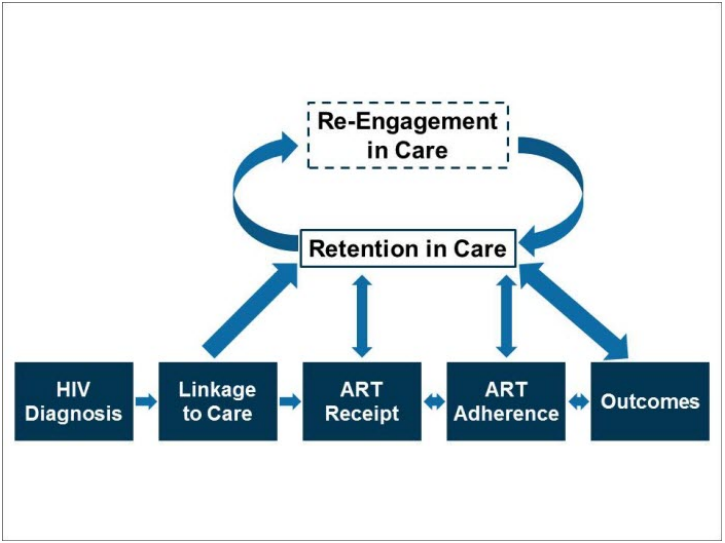


CDC MMWR 2011;60:1618-1623.

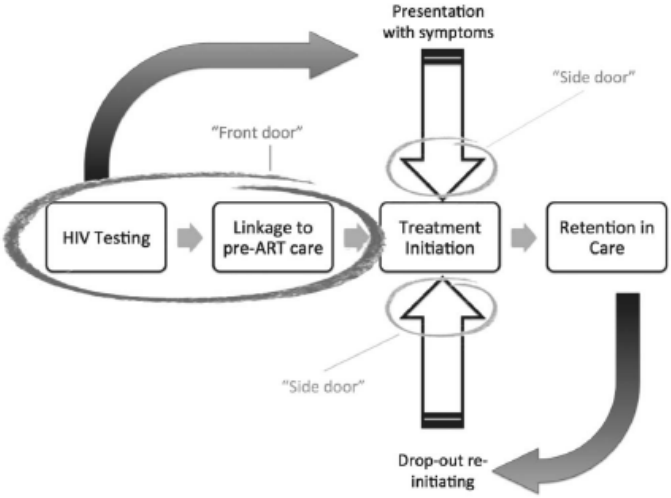
Challenges in modeling the full cascade

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

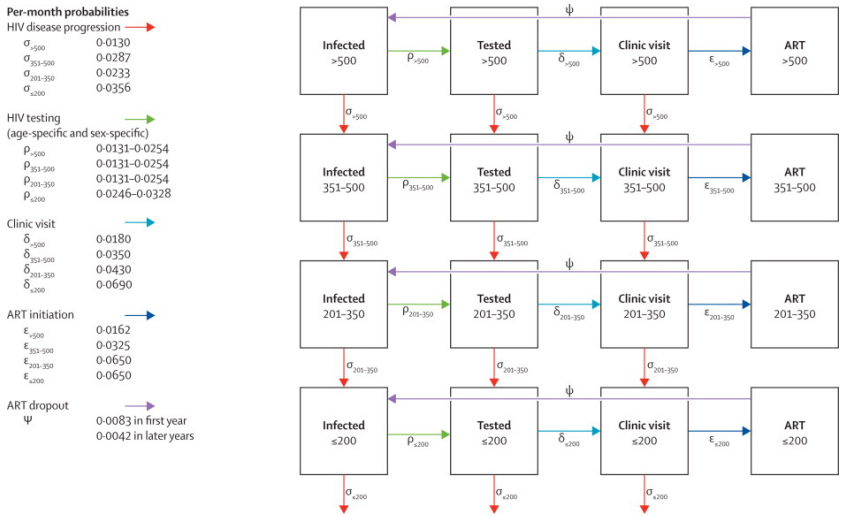
- Prevailing mode of analysis: microsimulation from mathematical models



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



Hallett TB, Eaton JW. "A side door into care cascade for HIV-infected patients?" JAIDS 63 (2013): S228-S232.



Smith et al. "Cost-effectiveness of community-based strategies to strengthen the continuum of HIV care in rural South Africa: a health economic modelling analysis." The Lancet HIV (2015): e159-e168.

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
Genberg, Hogan, Braitstein, 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.

Lancet HIV 2016; 3: e275–82
Published Online
May 11, 2016
http://dx.doi.org/10.1016/S2352-3018(16)30009-1

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.

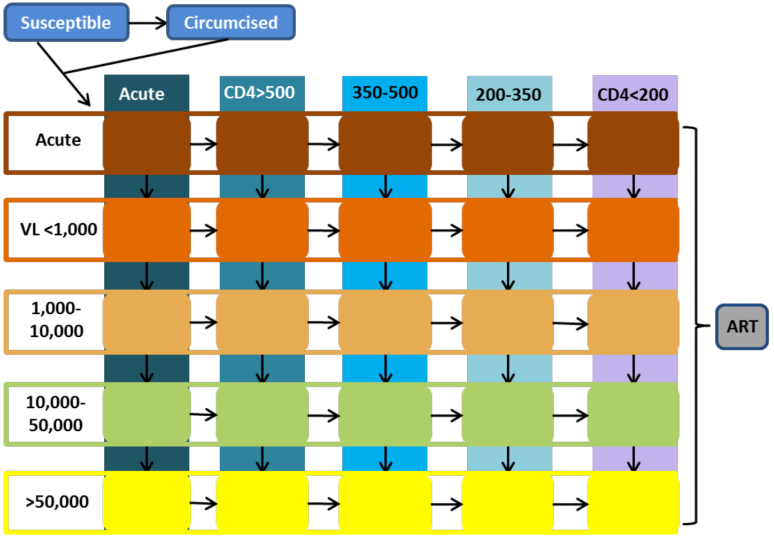


Figure S1. Model transition diagram. A diagram of the natural history of HIV infection. All movement is in one direction except for enrollment in and dropout from interventions from ART.

The ODEs for the nine disease states are:

$$\frac{dX_{a,r}^{g,0,0}(t)}{dt} = b_{r,0}^{g,0}(t) + \sigma_{a,r}^{g,0} X_{a,r}^{g,7,0}(t) - \left(\mu_a^g + \lambda_{a,r}^{g,0}(t) + \pi_{a,r}^{g,0,0}(t)\right) X_{a,r}^{g,0,0}(t)$$

$$\frac{dX_{a,r}^{g,1,\nu}(t)}{dt} = b_{r,0}^{g,1}(t) + \lambda_{a,r}^{g,0} X_{a,r}^{g,0,0}(t) + \psi_0 \lambda_{a,r}^{1,0}(t) X_{a,r}^{1,6,\nu}(t) + \psi_0 \psi_1 \lambda_{a,r}^{1,1}(t) X_{a,r}^{1,7,0}(t) + \psi_1 \lambda_{a,r}^{g,1}(t) X_{a,r}^{g,9,6}(t) + \sigma_{a,r}^{g,1}(t) X_{a,r}^{g,9,6} - \left(\mu_a^g + \alpha_a^{g,1} + \nu_1 + \pi_{a,r}^{g,1,\nu}(t)\right) X_{a,r}^{g,1,\nu}(t)$$

$$\frac{dX_{a,r}^{g,2,\nu}(t)}{dt} = (\nu_1 + \omega_{\nu-1}) X_{a,r}^{g,1,\nu}(t) + \sigma_{a,r}^{g,2} X_{a,r}^{g,9,6}(t) - \left(\mu_a^g + \alpha_a^{g,2} + \nu_2 + \omega_\nu + \pi_{a,r}^{g,2,\nu}(t)\right) X_{a,r}^{g,2,\nu}(t)$$

$$\frac{dX_{a,r}^{g,3,\nu}(t)}{dt} = (\nu_2 + \omega_{\nu-1}) X_{a,r}^{g,2,\nu}(t) + \sigma_{a,r}^{g,3} X_{a,r}^{g,9,6}(t) - \left(\mu_a^g + \alpha_a^{g,3} + \nu_3 + \omega_\nu + \pi_{a,r}^{g,3,\nu}(t)\right) X_{a,r}^{g,3,\nu}(t)$$

$$\frac{dX_{a,r}^{g,4,\nu}(t)}{dt} = (\nu_3 + \omega_{\nu-1}) X_{a,r}^{g,3,\nu}(t) + \sigma_{a,r}^{g,4} X_{a,r}^{g,9,6}(t) - \left(\mu_a^g + \alpha_a^{g,4} + \nu_4 + \omega_\nu + \pi_{a,r}^{g,4,\nu}(t)\right) X_{a,r}^{g,4,\nu}(t)$$

$$\frac{dX_{a,r}^{g,5,\nu}(t)}{dt} = (\nu_4 + \omega_{\nu-1}) X_{a,r}^{g,4,\nu}(t) + \sigma_{a,r}^{g,5} X_{a,r}^{g,9,6}(t) - \left(\mu_a^g + \alpha_a^{g,5} + \nu_5 + \omega_\nu + \pi_{a,r}^{g,5,\nu}(t)\right) X_{a,r}^{g,5,\nu}(t)$$

$$\frac{dX_{a,r}^{g,6,0}(t)}{dt} = b_{r,1}^{g,0}(t) + \sigma_{a,r}^{g,0} X_{a,r}^{g,6,0}(t) - \left(\mu_a^g + \psi_0 \lambda_{a,r}^{g,0}(t) + \pi_{a,r}^{g,0,0}(t)\right) X_{a,r}^{g,6,0}(t)$$

$$\frac{dX_{a,r}^{g,7,0}(t)}{dt} = \pi_{a,r}^{g,0,0}(t) X_{a,r}^{g,5,0}(t) - \left(\sigma_{a,r}^{g,0} + \mu_a^g + \psi_0 \psi_1 \lambda_{a,r}^{g,1}(t)\right) X_{a,r}^{g,7,0}(t)$$

$$\frac{dX_{a,r}^{g,8,0}(t)}{dt} = \pi_{a,r}^{g,0,0}(t) X_{a,r}^{g,0,0}(t) - \left(\sigma_{a,r}^{g,0} + \mu_a^g + \psi_1 \lambda_{a,r}^{g,1}(t)\right) X_{a,r}^{g,8,0}(t)$$

$$\frac{dX_{a,r}^{g,9,6}(t)}{dt} = \sum_{\nu=1}^{5} \sum_{d=1}^{5} [\pi_{a,r}^{g,d,\nu}(t) X_{a,r}^{g,d,\nu}(t) - \left(\sigma_{a,r}^{g,d} + \mu_a^g\right) X_{a,r}^{g,9,6}(t)]$$

Some model input variables

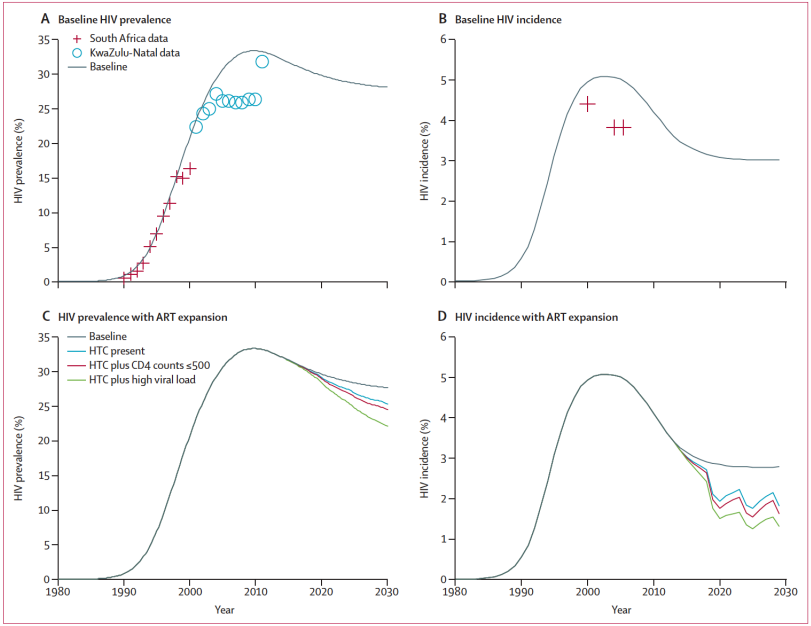
$\pi_{a,r}^{g,d,v}(t)$	The coverage of PrEP ($d = 0$), ART ($d = 1, \ldots, 5$), circumcision ($d = 6$), condom use among HIV-negative persons ($d = 7$), condom use among PrEP users ($d = 8$), and condom use among ART users ($d = 9$)
$\alpha_a^{g,d}$	The HIV-associated mortality
ν_d	The rate of progressing from CD4 state d to $d + 1$
ω_d	The rate of progressing from VL state v to $v + 1$
ψ_d	The reduction in HIV transmission due to circumcision ($d = 0$), PrEP ($d = 1$), ART ($d = 2$), or condom use ($d = 3$)

Source of selected inputs

	Value	Reference
Duration of disease by CD4 count		
Acute	0·25 year	Hontelez et al[14]
>500 cells per μL	1·88 years	Celum et al,[15] Baeten et al[16]
500–350 cells per μL	1·22 years	Celum et al,[15] Baeten et al[16]
350–200 cells per μL	5·90 years	Celum et al,[15] Baeten et al[16]
<200 cells per μL	1·96 years	Badri et al[17]
Duration of disease by HIV viral load		
Acute	0·25 years	Hontelez et al[14]
<1000 copies per mL	3·13 years	Celum et al,[15] Baeten et al[16]
1000–10 000 copies per mL	1·99 years	Celum et al,[15] Baeten et al[16]
10 000–50 000 copies per mL	4·40 years	Celum et al,[15] Baeten et al[16]
>50 000 copies per mL	1·44 years	Estimated
Costs*		
Annual home HTC with community care workers	$28·06 per HIV-positive person $8·22 per HIV-negative person	Smith et al[18]

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**


A Baseline HIV prevalence
South Africa data
KwaZulu-Natal data
Baseline
HIV prevalence (%)
B Baseline HIV incidence
HIV incidence (%)
C HIV prevalence with ART expansion
Baseline
HTC present
HTC plus CD4 counts ≤500
HTC plus high viral load
HIV prevalence (%)
Year
D HIV incidence with ART expansion
HIV incidence (%)
Year

Comparison of treatment initiation policies

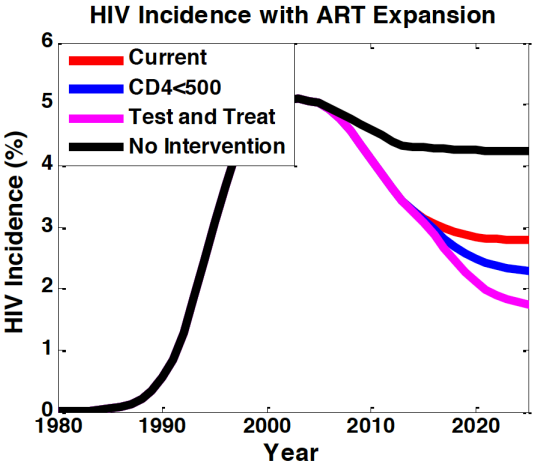


Figure S6. HIV incidence with optimistic ART coverage scenarios.

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

Statistical model of HIV cascade using EHR data

Goal: Compare two treatment policies

- Treat upon enrollment (test and treat)
- Treat when CD4 drops below 350

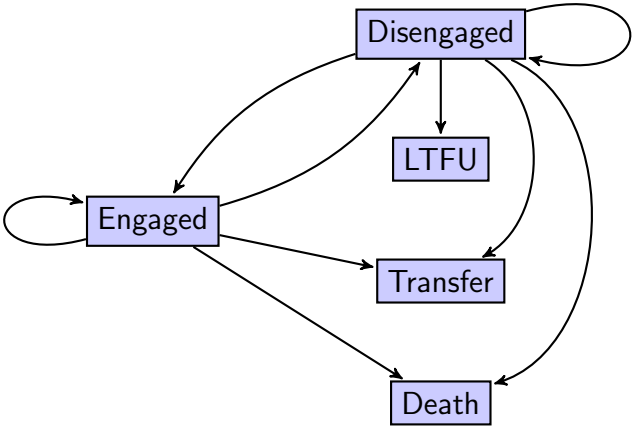
Outcome: Care state following enrollment

- 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

$$
\begin{aligned}
S_j &= \text{state at time } t_j \\
p_{jk\ell} &= P(S_j = \ell \,|\, S_{j-1} = k)
\end{aligned}
$$

	State at t_j				
State at t_{j-1}	Engaged	Disengaged	LTFU	Death	Transfer
Engaged	p_{j11}	p_{j12}	0	p_{j14}	p_{j15}
Disengaged	p_{j21}	p_{j22}	p_{j23}	p_{j24}	0
LTFU	0	0	1	0	0
Death	0	0	0	1	0
Transfer	0	0	0	0	1

Can incorporate covariate effects via regression

- Use multinomial regression for longitudinal data

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

- $\boldsymbol{x}_j$ = 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

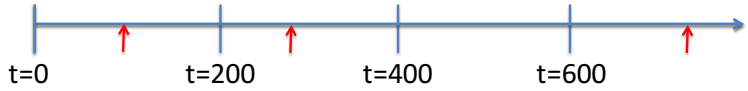
Start at enrollment, ascertain status every 200 days

- 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

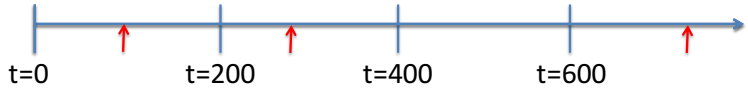
(enrollment)

Organizing data into states



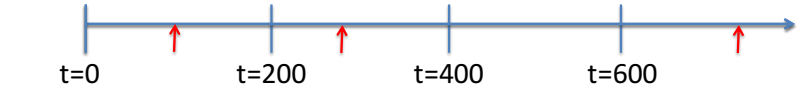
(enrollment)

Organizing data into states

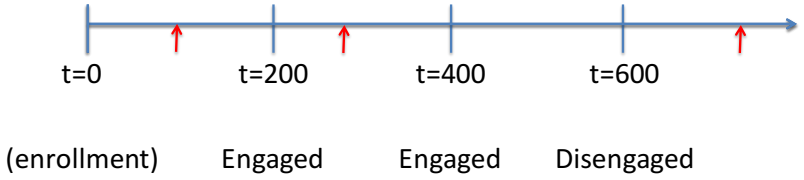


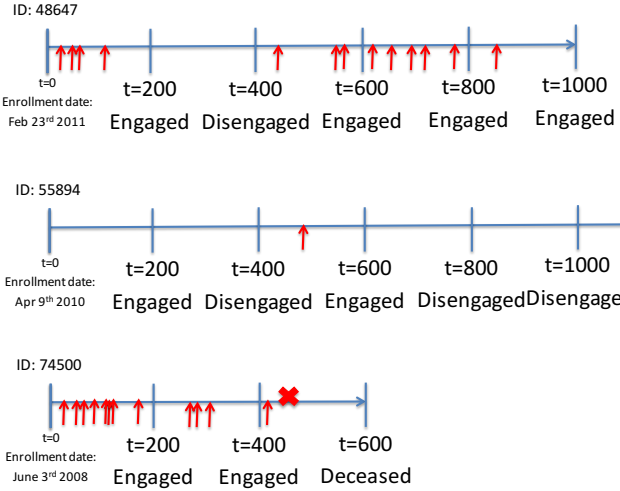
(enrollment) Engaged

Organizing data into states



Organizing data into states



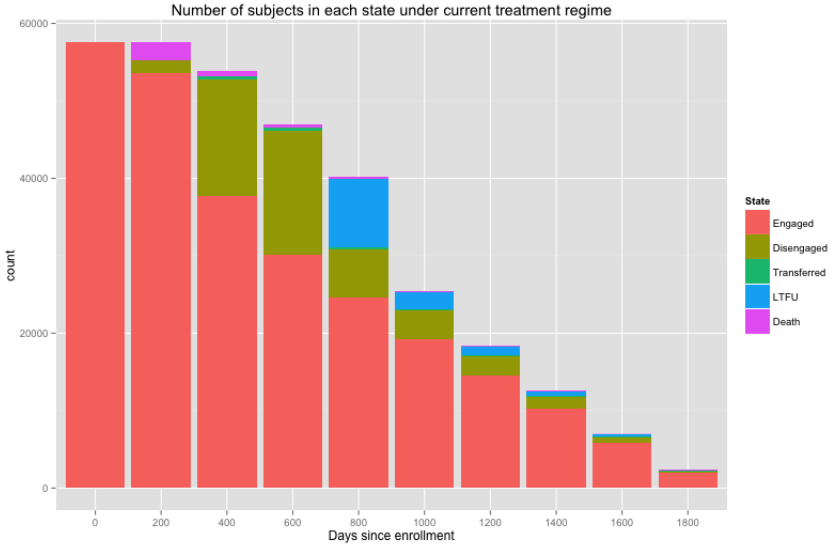

ID: 48647
t=0
Enrollment date:
Feb 23rd 2011
t=200
Engaged
t=400
Disengaged
t=600
Engaged
t=800
Engaged
t=1000
Engaged
ID: 55894
t=0
Enrollment date:
Apr 9th 2010
t=200
Engaged
t=400
Disengaged
t=600
Engaged
t=800
Disengaged
t=1000
Disengaged
t=1200
LTFU
ID: 74500
t=0
Enrollment date:
June 3rd 2008
t=200
Engaged
t=400
Engaged
t=600
Deceased

Summary of available data

Unique individual enrollments:
57,596

Unique visits:
1,493,150

Observations used in analysis:
321,834



Transition rate estimates

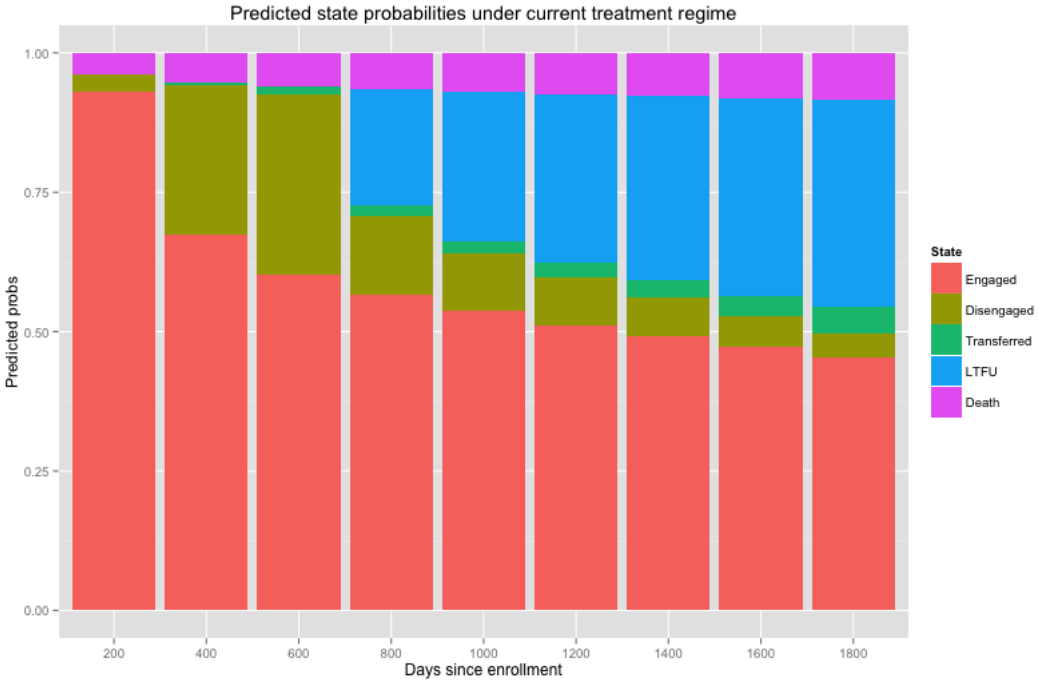
Time-aggregated estimates
57,000+ individuals
Enrolled between 6/2008 – 9/2012

	State at t_j				
State at t_{j-1}	Engaged	Disengaged	LTFU	Death	Transfer
Engaged	.86	.11	0	.02	.01
Disengaged	.12	.54	.33	.01	0
LTFU	0	0	1	0	0
Death	0	0	0	1	0
Transfer	0	0	0	0	1

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 $P(\boldsymbol{S}_1)$, $P(\boldsymbol{S}_2 \,|\, \boldsymbol{S}_1)$, $P(\boldsymbol{S}_3 \,|\, \boldsymbol{S}_2), \ldots$
  - Calculate marginal probabilities

$$\widehat{P}(\boldsymbol{S}_j) \;=\; \sum_{\boldsymbol{S}_1,\ldots,\boldsymbol{S}_{j-1}} \widehat{P}(\boldsymbol{S}_1) \prod_k \widehat{P}(\boldsymbol{S}_k \,|\, \boldsymbol{S}_{k-1})$$


Predicted state probabilities under current treatment regime
Predicted probs
1.00
0.75
0.50
0.25
0.00
200
400
600
800
1000
1200
1400
1600
1800
Days since enrollment
State
Engaged
Disengaged
Transferred
LTFU
Death

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

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

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

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

Example: 'treat immediately' is the regime

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

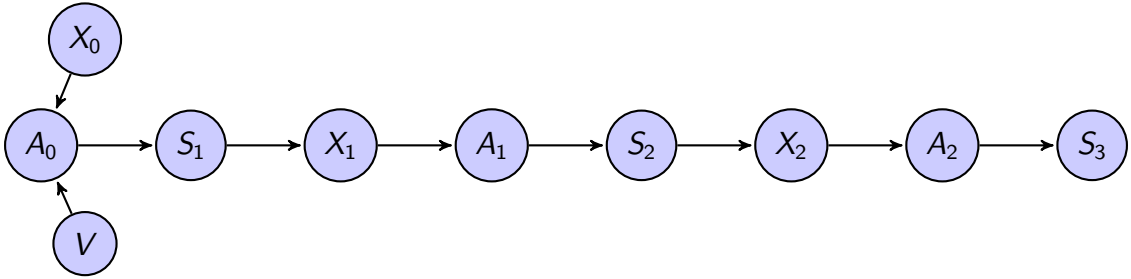
Statistical and computational issues

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



$$
\begin{aligned}
S &= \text{state membership} \\
X &= \text{CD4 count} \\
A &= \text{treatment} \\
V &= \text{gender, age at enrollment}
\end{aligned}
$$

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

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

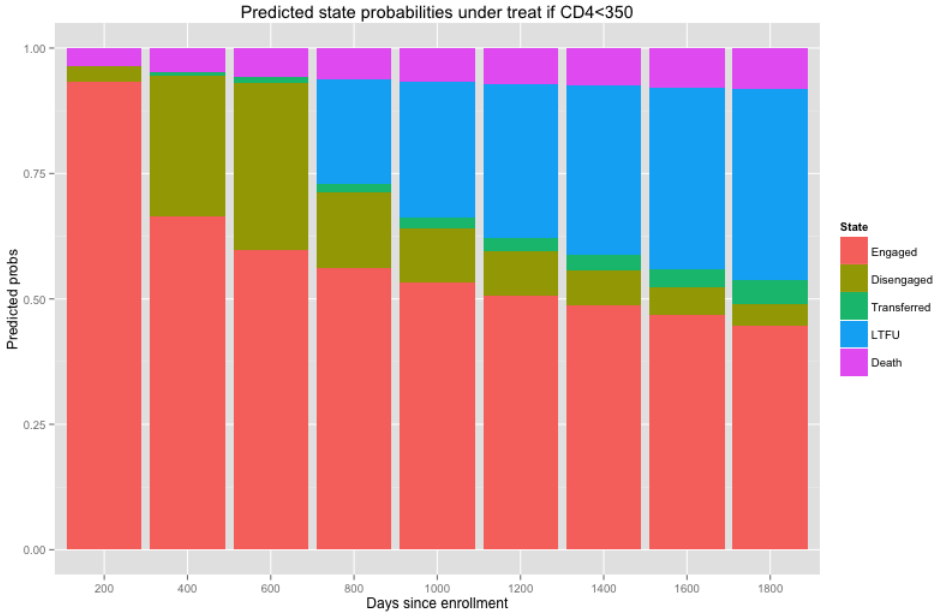
G computation:

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

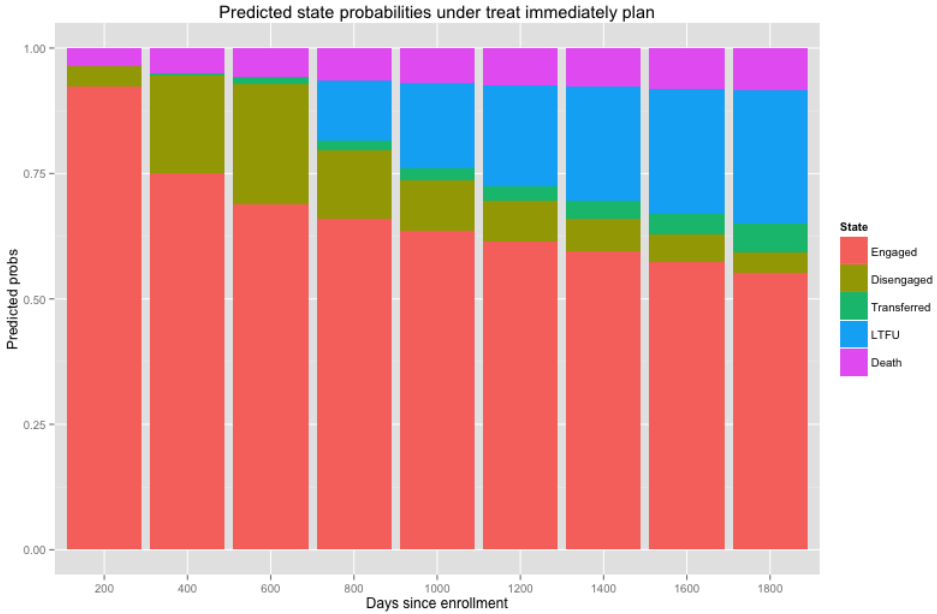
Implementation

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

Treat if CD4<350



Treat upon enrollment ('test and treat')

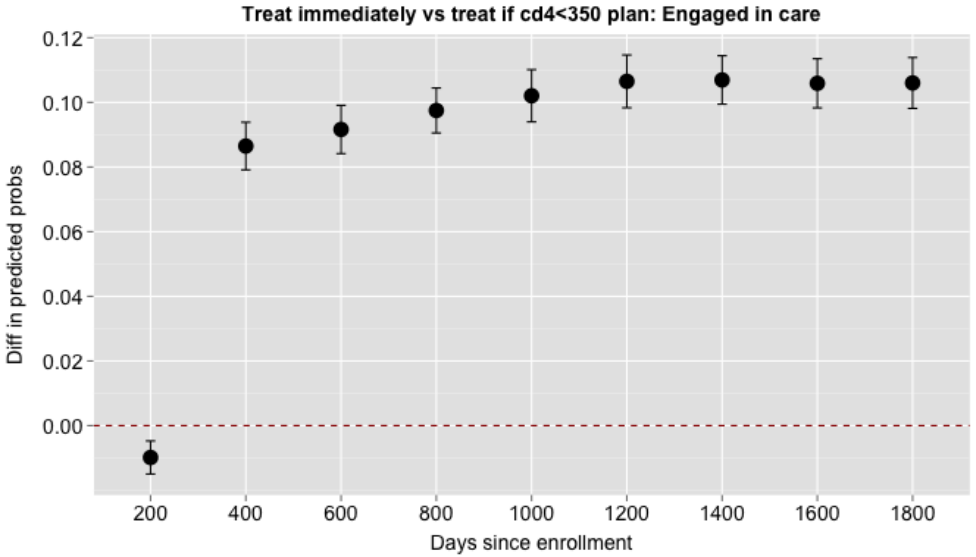


Inferences

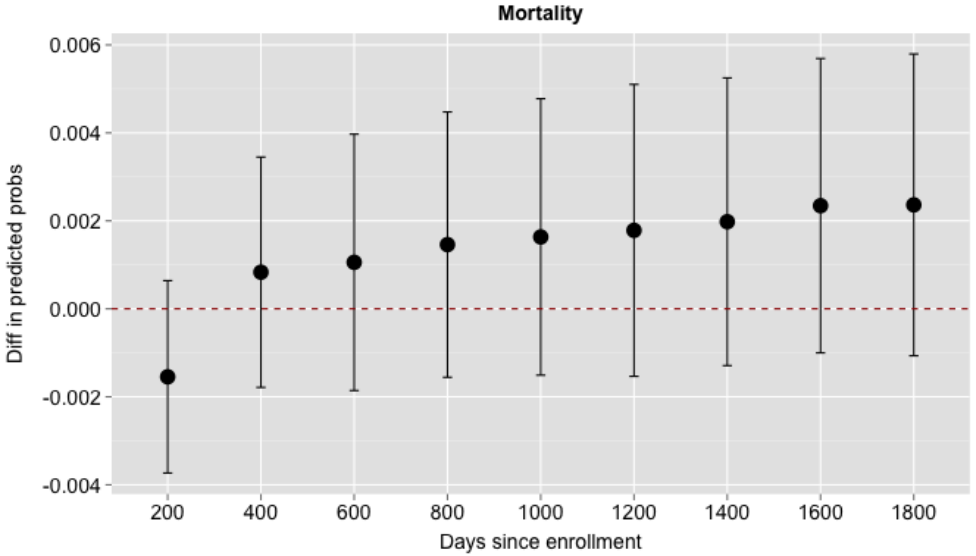
Next few slides:

- Compare proportions in each state over time
- Use rate difference, 95% confidence interval
- Based on 100 bootstrap samples (about 2 hrs on iMac)

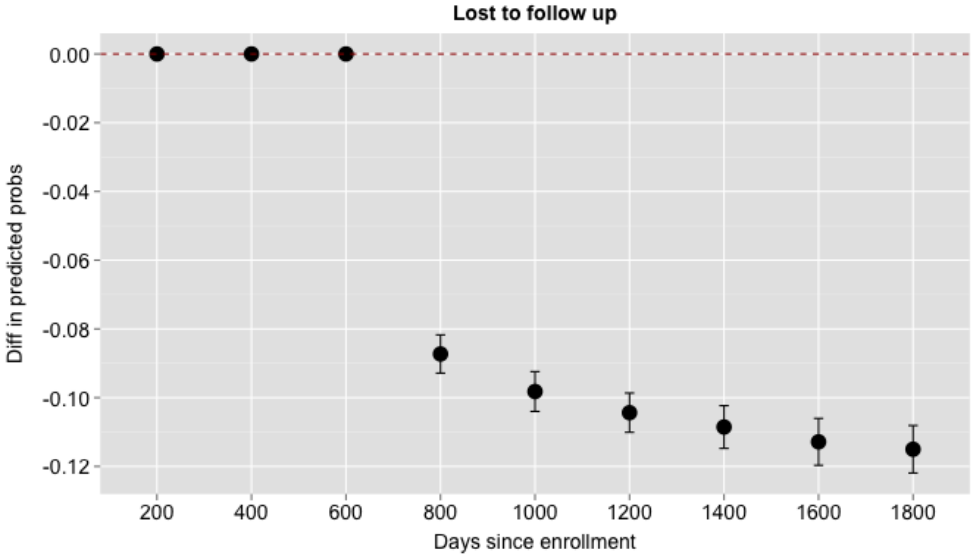
Engaged in care



Mortality


Mortality
Diff in predicted probs
0.006
0.004
0.002
0.000
-0.002
-0.004
200
400
600
800
1000
1200
1400
1600
1800
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

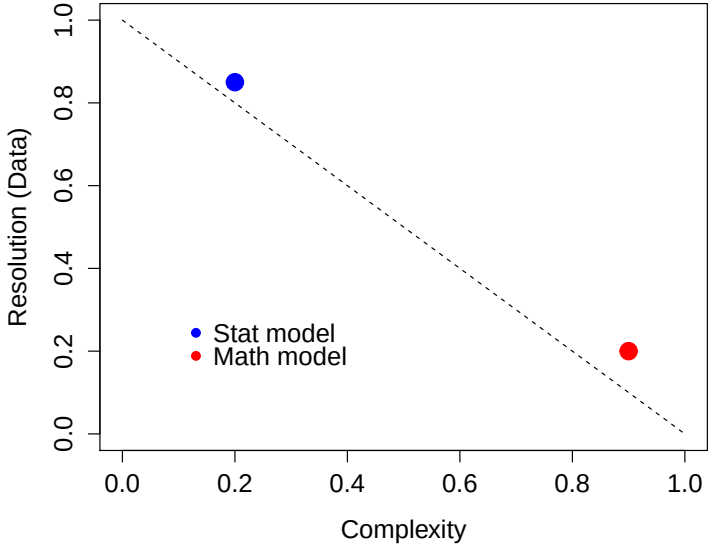
Compare and contrast

- 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


Resolution (Data)
0.0
0.2
0.4
0.6
0.8
1.0
Stat model
Math model
0.0
0.2
0.4
0.6
0.8
1.0
Complexity

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

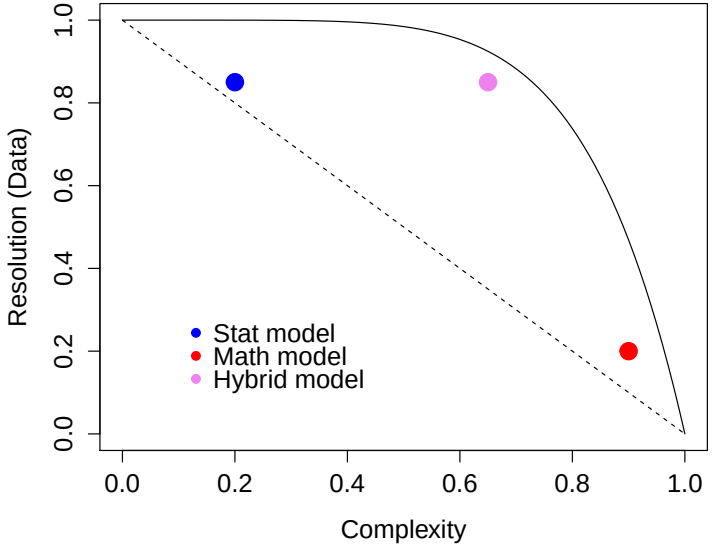
- EHR presents important new opportunities
  - Larger databases $\Rightarrow$ more complex statistical models
- EHR data are **rich** but **messy**
- Lots of prospectors circling the gold mine
- **Statistical principles** could hardly be more important!
  - Distinguishing causal from predictive inference
  - 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


Resolution (Data)
0.0
0.2
0.4
0.6
0.8
1.0
Complexity
Stat model
Math model
Hybrid model

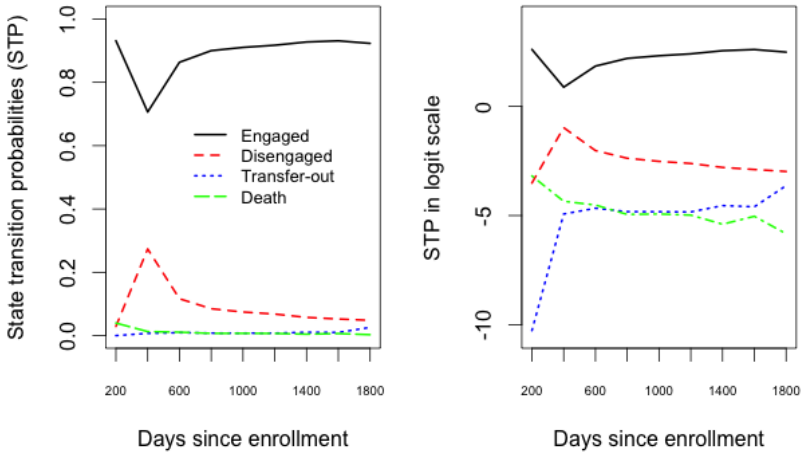
Opportunities: Generating and grading evidence

- 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)

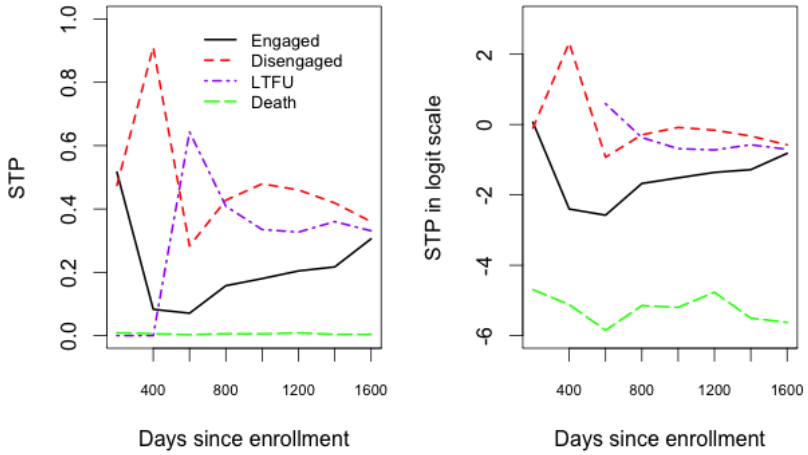
- Requires framework for grading evidence
  - Well-defined population?
  - Distinguish sources of uncertainty?
  - Check model fit?
  - External validation?
  - Reproducibility?

Thank you!

Temporal trends in transition from **engaged**



Temporal trends in transition from **disengaged**



Covariate effects: transition from **engaged**

(Adjusted for calendar year)

	Engaged	Disengaged	Death
Transition probability	.86	.12	.01
Male	—	1.19*	1.76*
Age > 35	—	0.67*	1.01
CD4$<$350, ART$-$	—	—	—
CD4$<$350, ART$+$	—	0.20*	0.42*
CD4$\geq$350, ART$-$	—	0.38*	0.10*
CD4$\geq$350, ART$+$	—	0.11*	0.09*

Covariate effects: transition from **disengaged**

(Adjusted for calendar year)

	Engaged	Disengaged	LTFU	Death
Transition probability	.12	.54	.33	.01
Male	—	0.96	0.92*	1.05
Age > 35	—	0.94*	0.98	1.23
CD4$<$350, ART$-$	—	—	—	—
CD4$<$350, ART$+$	—	0.40*	0.33*	0.67*
CD4$\geq$350, ART$-$	—	0.81*	0.71*	0.35*
CD4$\geq$350, ART$+$	—	0.34*	0.25*	0.12*

G computation for estimating causal quantities

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

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

G computation for estimating causal quantities

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

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

First-order Markov assumption

G computation for estimating causal quantities

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

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

Multinomial regression models for state transitions

G computation for estimating causal quantities

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

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

Need to specify new model for CD4 evolution

G computation for estimating causal quantities

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

$$
\begin{aligned}
\tilde{S}_{1i} &\sim \widehat{P}(S_1 \,|\, A_0 = a_0, X_{0i}, V_i) \\
\tilde{X}_{1i} &\sim \widehat{P}(X_1 \,|\, A_0 = a_0, X_{0i}, V_i, \tilde{S}_{1i}) \\
\\
\widehat{P}_{a_0,a_1}(S_2) &= (1/n) \sum_i \widehat{P}(S_2 \,|\, A_1 = a_1, \tilde{X}_{1i}, \tilde{S}_{1i}, V_i)
\end{aligned}
$$

Calculate via Monte Carlo simulation based on fitted models