Data Adaptive Pre-Specification for Experimental and Observational Data

Maya Petersen
with
Laura Balzer, Linh Tran, Mark van der Laan
Div. of Biostatistics, School of Public Health,
University of California, Berkeley

BITSS Workshop, Berkeley CA, December 11, 2014

Pre-specification is needed

- Statistical Inference relies on having a <u>well-defined experiment</u>
  - Population, sampling, data collection, analysis
  - <u>An estimator is an algorithm</u>
    - ie. a computer program
- If we do not have a pre-specified analysis plan (estimator), we no longer have a well-defined experiment
  - Estimator includes any decisions about
    - Which covariates we will adjust for
    - Model specification used to adjust
    - Many more…

Dangers of *ad hoc* analytic decisions

- Run a bunch of regressions and choose the one with
  1. Smallest p value?
  2. Results that make the most sense?
- ➢ Misleading (under) estimate of uncertainty
- ➢ Bias
  - Humans are good at creating narratives from complexity
  - Tendency to confirm what we expect to find
- As long there is "art" in statistics, we will continue to make a lot of wrong inferences

Pre-specification also has dangers

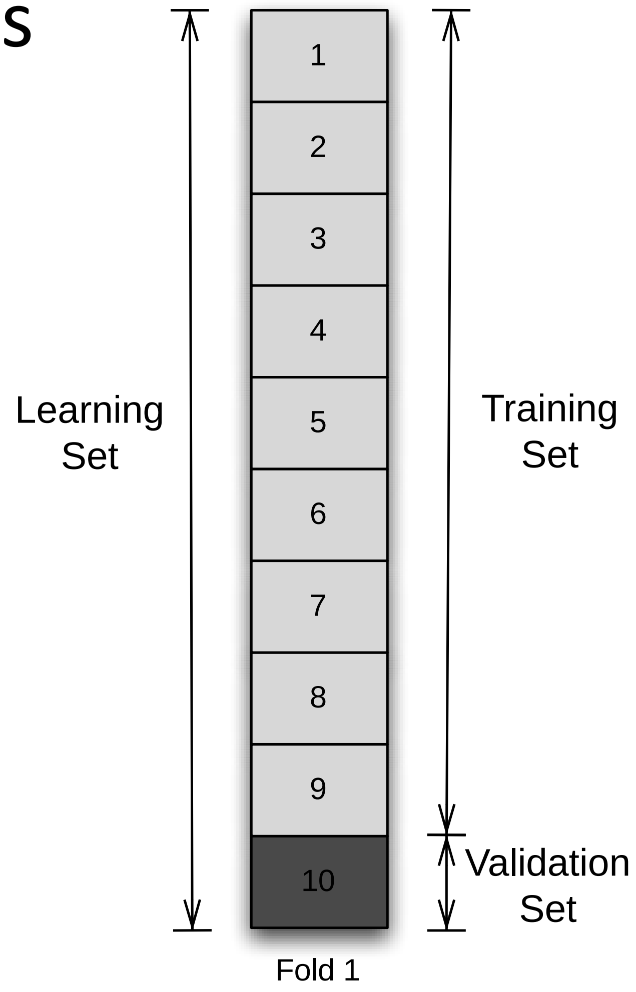
- Ex. Randomized Trials
  - Adjustment can reduce variance/improve power
  - Which covariate(s) to adjust for?
  - Pre-specify a poor choice -> **Less Power/Precision**
- Ex. Observational Data
  - Range of identification/adjustment strategies
  - Which variables to adjust for? Specification?
  - Pre-specify a poor choice -> **Bias**
- We <u>must</u> look at and learn from our data to make good decisions

Data-Adaptive Pre-Specification….

- Machine-learning to the rescue?
- Wide range of data-adaptive or machine learning methods for prediction/classification
- Look at and learn from data in an *a priori* specified way

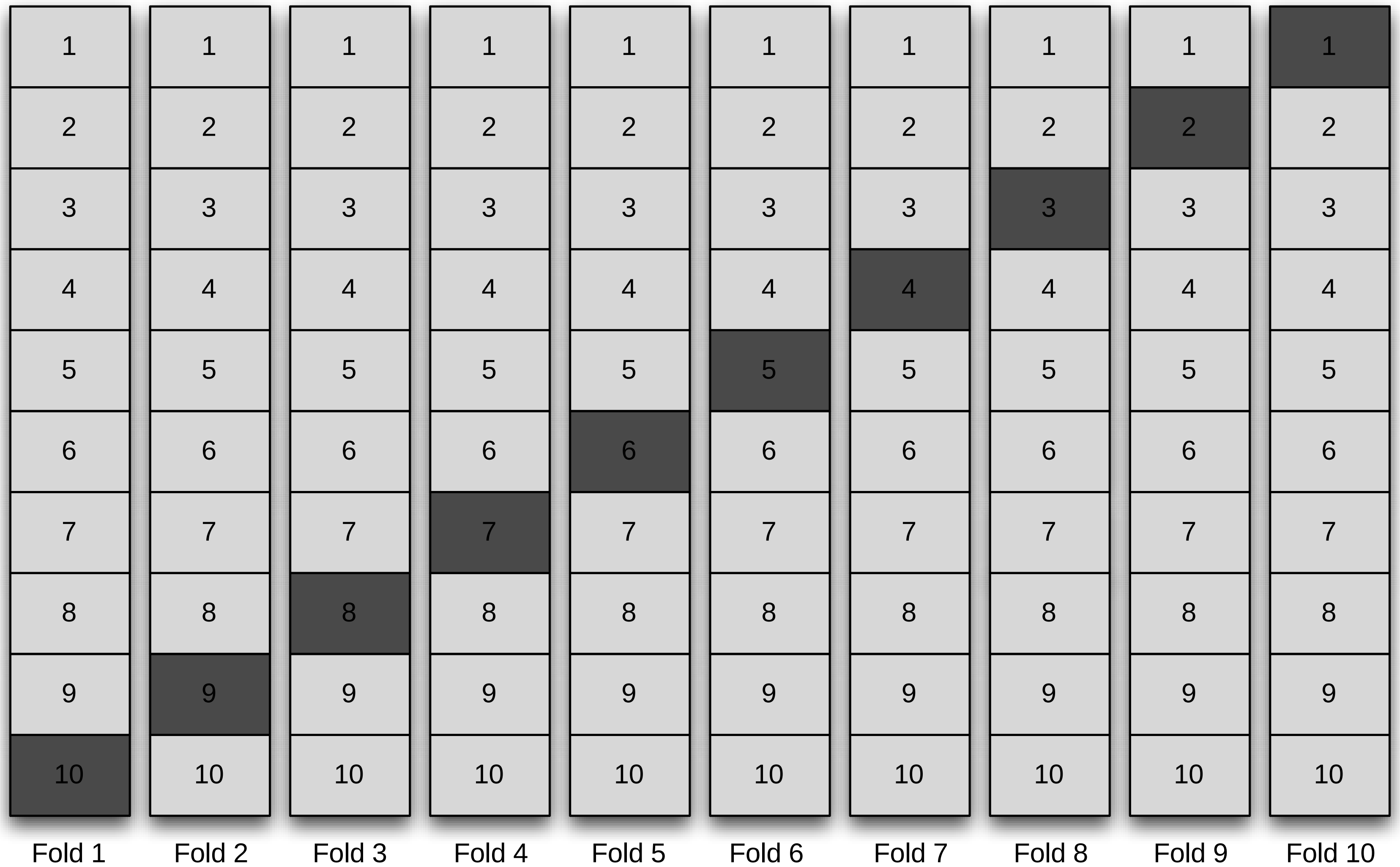
Example: "Super Learner"

- "Competition" of algorithms
  - Parametric models
  - Data-adaptive (ex. Random forest, Neural nets)
- Best "team" wins
  - Convex combination of algorithms
- Performance judged on independent data
  - Internal data splits



Van der Laan, Polley, 2007

Example: “Super Learner”



Problem solved?

- Not without some additional help…
  - Sophisticated machine learning methods available
  - Powerful tools for <u>Prediction</u>
- However, if used isolation <u>don't let us make reliable inferences about causally motivated parameters</u>
  - Not targeting the question of interest
  - ➢ Too much bias and misleading confidence intervals/hypothesis tests

Targeted Learning

- Targeted Maximum Likelihood Estimation
  - General statistical methodology
    - For a range of causally and non-causally motivated statistical quantities
  - Uses state-of-the art machine learning
  - Updates output in a targeted way
    - Reduce bias
    - Regain statistical properties for reliable inference
- Efficient (minimal asymptotic variance)
  - If nuisance parameters estimated well
- Often nice robustness properties

Van der Laan, Rose, Springer 2011

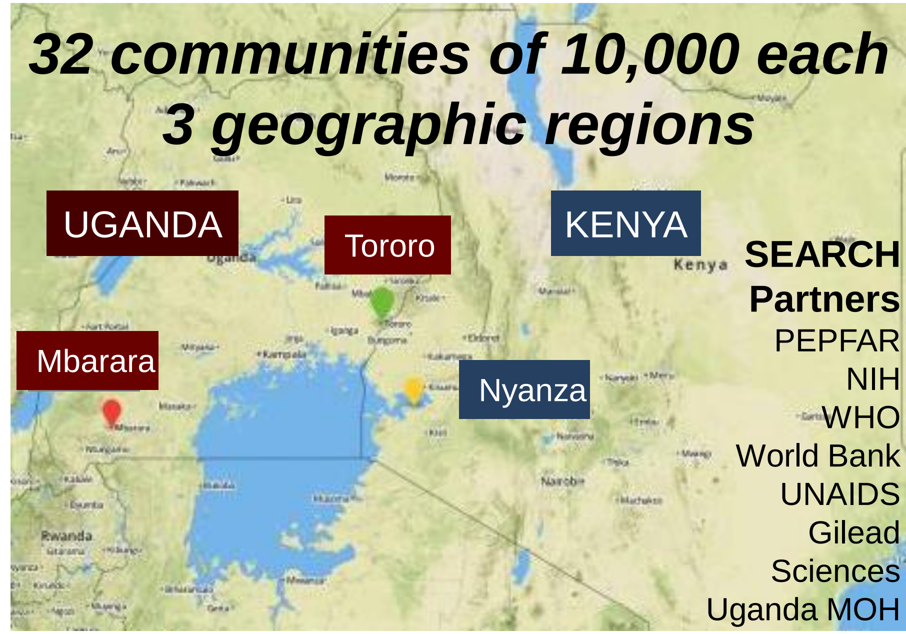
Adaptive Pre-Specification: Randomized Trials

SEARCH
SUSTAINABLE EAST AFRICA RESEARCH IN COMMUNITY HEALTH

SEARCH Consortium:
Sustainable East Africa Research in Community Health

SEARCH: Study Questions and Design

- **Can a population-based ART strategy “shut down” new HIV infections?**
  - What are the additional gains?(maternal child health, TB, education, household earning power)
  - What is the best way to do it? Cost?
  - Can efficient HIV chronic care models be adapted to establish care for other chronic diseases (hypertension and diabetes)?



SEARCH Partners
PEPFAR
NIH
WHO
World Bank
UNAIDS
Gilead Sciences
Uganda MOH
Kenya MOH

SEARCH: Cluster randomized trial of universal vs. standard ART

Intervention

ART at all CD4+
Annual & targeted testing
Enhanced linkage & retention

16 communities
n = 10,000 each

Standard of

Country-guided ART

16 communities
n = 10,000 each

Community Health Campaign

HIV Screening/Diagnosis
Malaria testing & care
HTN and Diabetes testing
Maternal/child health

Outcomes Year 3 and

Community Health
- HIV incidence
- HIV population viral metrics
- AIDS
- Maternal and child health
- TB
- NCD (HT, DM)

Community Productivity/Costs
- Workforce participation
- Child labor prevalence
- Agricultural output
- Household income
- Educational attainment
- Healthcare utilization

SEARCH: Pre-Specified Analysis Plan

- Primary study outcome: Impact on Incident HIV

1. Estimate community-level outcome: 5 year HIV cumulative incidence
   - Probability of becoming infected over 5 years given uninfected at baseline
2. Compare average cumulative incidence between control and intervention communities
   - 32 matched pairs-> limited ability to adjust
   - Many candidate adjustment variables…
   - Which (if any) community covariate to adjust for?

Data-adaptive pre-specification

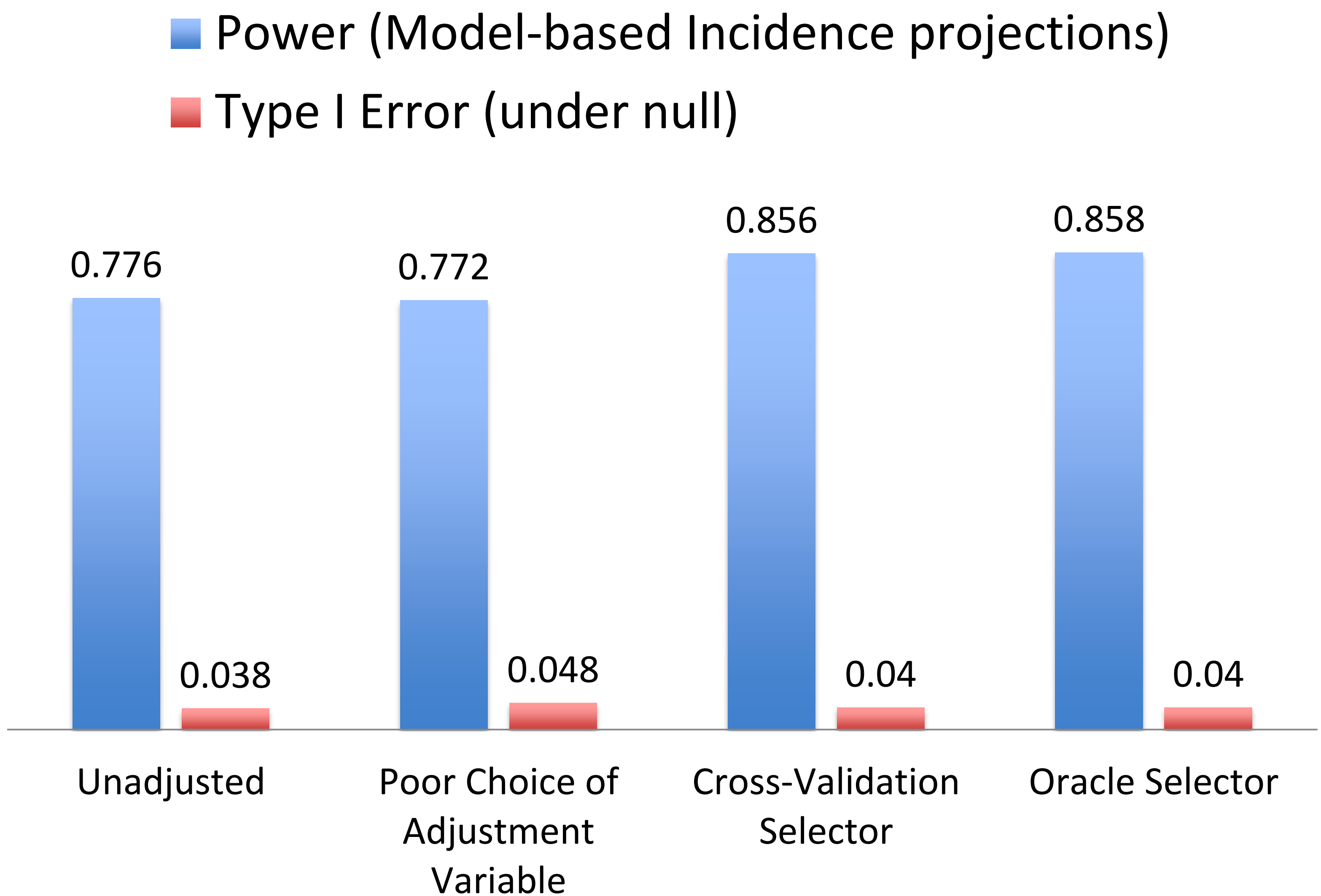
- Pre-specify:

1. Candidate adjustment variables
    - Baseline HIV prevalence
    - % population with HIV viral load<400 copies/ml
    - Median HIV viral load
    - None (no adjustment)
2. Final estimator
    - Method of adjustment: Main term logistic regression of outcome on intervention and a single covariate
    - Algorithm for selecting between candidate regressions: Leave-one-out cross validation

Leave-one-out cross validation

1. Fit each candidate regression on 15/16 pairs
   - Evaluate squared prediction error on remaining pair
2. Repeat 16 times, leaving out each pair in turn
   - 32 squared prediction errors - one for each community
3. Average prediction errors across communities and select regression with the smallest
   - Best performance on independent data
4. Re-fit selected regression on all 32 communities and use to estimate treatment effect
   - In RCT with many classes of glm, no update needed

Data-Adaptive Adjustment: More power and good Type I error control

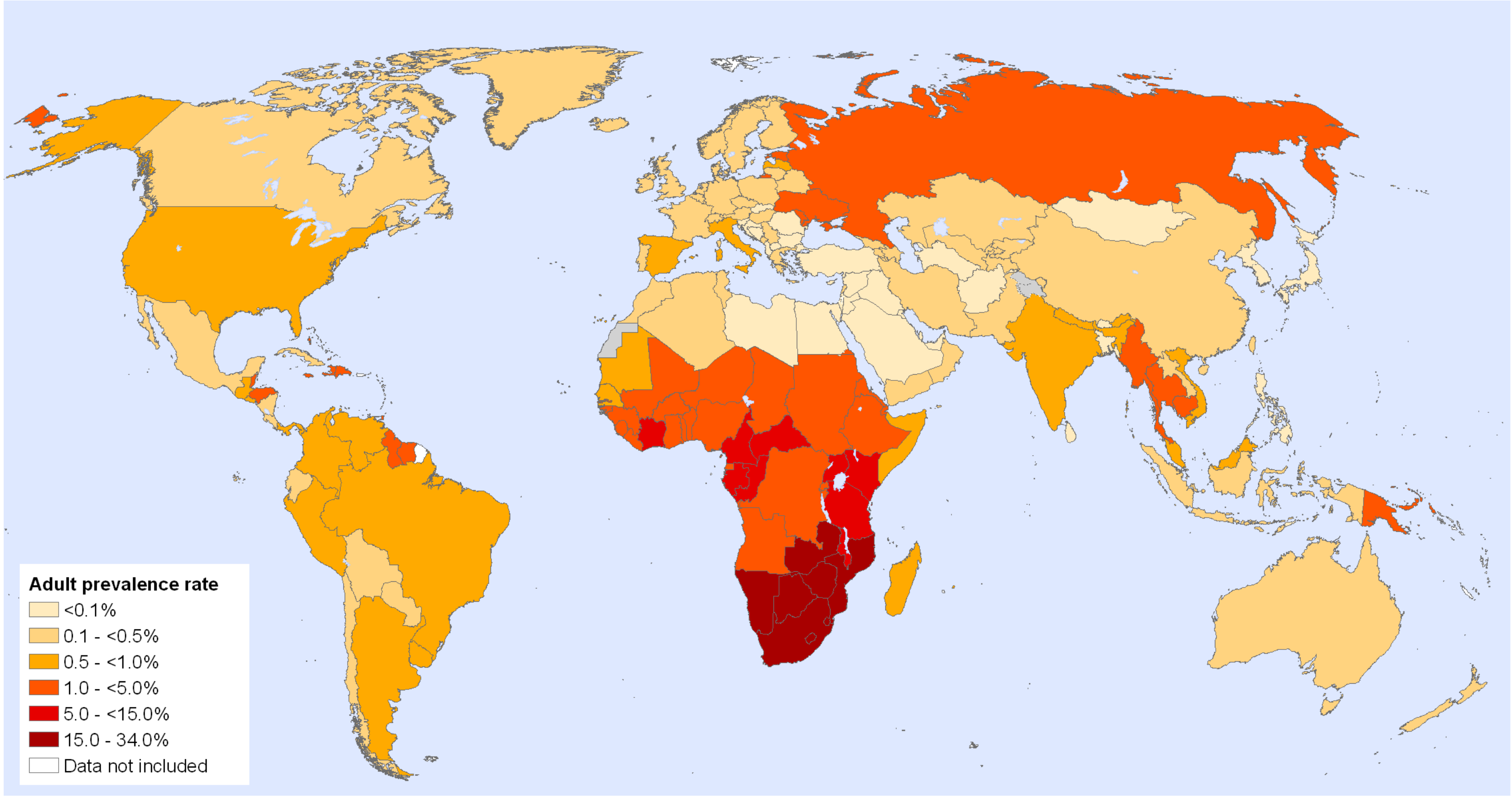


Adaptive Pre-Specification: Observational Analyses

International Epidemiologic Databases to Evaluate AIDS-

East Africa

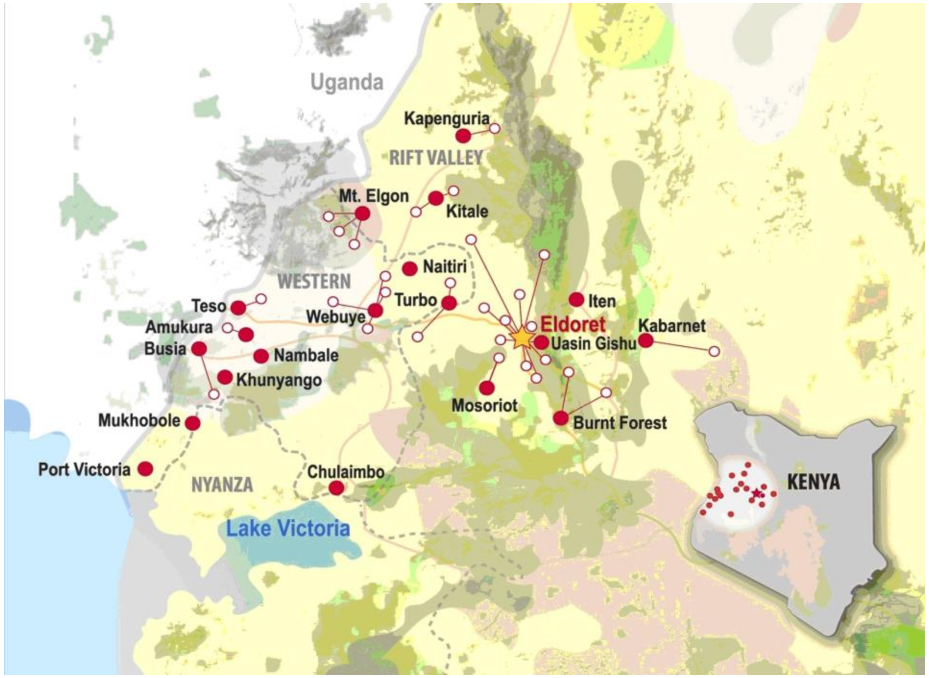
HIV treatment gap in resource-limited settings



- 4.5 on antiretroviral therapy, 9 million in need
- Shortage of financial and human resources

Low Risk Express Care (LREC)

- Task-shifting HIV care for stable patients from clinicians to nurses
- Implemented in 15 clinics in Kenya 2007-2008
  – USAID- AMPATH partnership



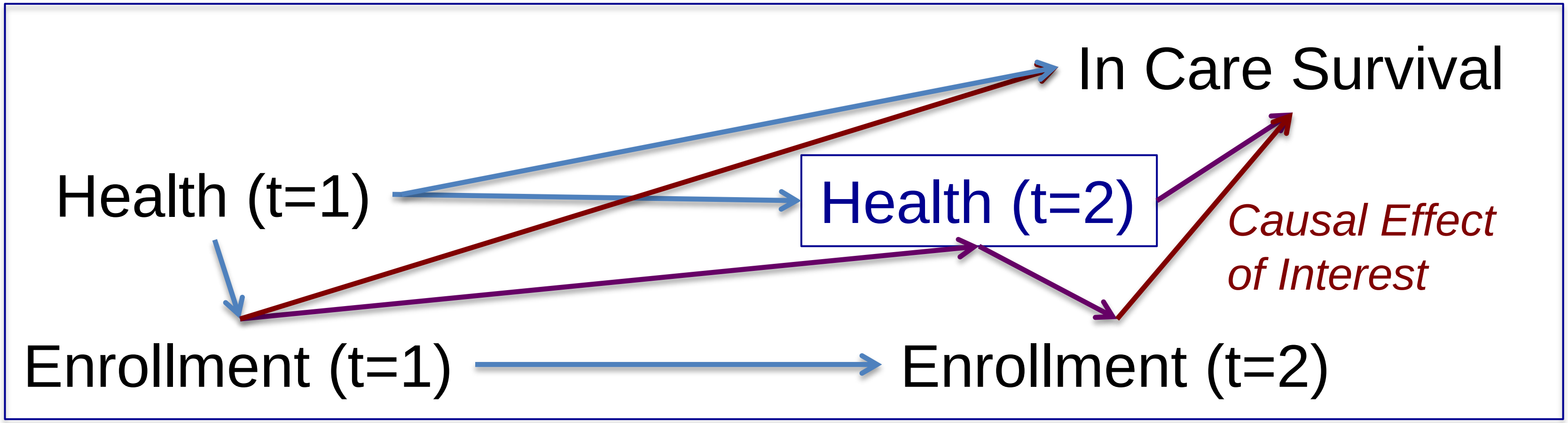
- Subset of eligible patients enrolled at varying times (Non-random)

Effect of LREC enrollment?

- Patient population: 15,225 Subjects eligible for LREC following program availability in a participating clinic
  - t=0: first date eligible for LREC after available in clinic
  - 5963 (39%) subsequently enroll
- Outcome: “In-Care” Survival
  - Failure = Death (any cause) or “Loss to follow up” (fail to return to clinic for 6.5 months)
- Longitudinal socio-demographic and clinical data
  - Age, sex
  - Disease severity, CD4 count, tuberculosis, pregnancy, antiretroviral use, adherence, etc…

Identification requires non-standard estimand

- All patients in analysis eligible (“low risk”)
- Enrollment at provider discretion
  - Sicker patients less likely to be enrolled
  - Drivers of enrollment *affected by prior treatment*



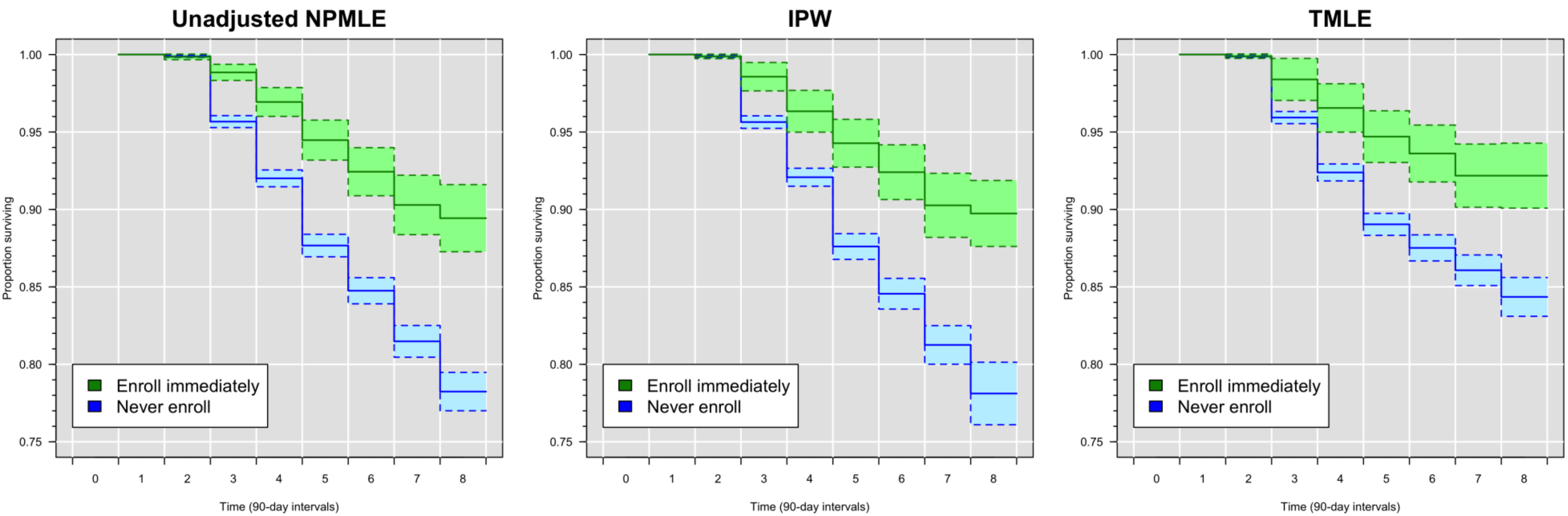
- Even with no unmeasured confounding, can’t identify using standard adjustment methods

Estimators

1. Inverse Probability Weighted Estimator
   – Current "Best Practice"
   – Propensity score based weights
     • Ex: Sicker patients that enroll/ healthier patients that don't enroll get up-weighted
   – Propensity score estimated with **pre-specified parametric model** (main-term logistic regression)
2. Targeted Maximum Likelihood Estimation
   – **Super Learner** to estimate
     • Series of iterated conditional expectations
     • Propensity score (for update)

Petersen et al, JCI 2014

TMLE-Super Learner: Improved control for measured confounders



- Estimated reduction in probability of death/drop-out by month 21 if enrolled immediately in LREC vs. never enrolled

Unadjusted NPMLE	IPW (Parametric Propensity score)	TMLE (Super Learner)
11% (9%, 14%)	12% (9%, 15%)	8% (5%, 10%)

Targeted Learning: Data-adaptive Pre-Specification

- Learn more…
  - Use flexible estimators that respond to the data
  - Data-adaptive or machine learning methods are not just for exploratory analysis
  - The problems we face are hard – if we don't respond to our data we will not get good answers
- But learn rigorously…
  - The estimator is an *a priori* specified algorithm
  - The algorithm itself is flexible- learns from data
  - Targeted to retain validity of statistical inference

Towards a General Learning System

User Input

- Question
  - Prediction versus causal
  - Point, longitudinal, static, dynamic, stochastic exposures
- Data
  - Longitudinal, Hierarchical
  - Missing data
- Model
  - Causal and statistical
  - Knowledge about data generating process

Output

- Target statistical parameter (estimand)
- Point estimate
- Statistical Inference
- Diagnostics
  - Suggested responses if insufficient support
- Guidance for interpretation
  - Ex: Assumptions for specific interpretations

Towards a General Learning System

User Input

- Question
  - Prediction versus causal
  - Point, longitudinal, static, dynamic, stochastic exposures
- Data
  - Longitudinal, Hierarchical
  - Missing data
- Model
  - Causal and statistical
  - Knowledge about data generating process

- Understanding and articulating the relevant questions
- Understanding the data
- Understanding (and optimizing) the experiment that generated it
  - Study design
  - Expert knowledge

School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

Laura Balzer
(SEARCH)

Mark van der Laan
Alan Hubbard

Linh Tran
(LREC)

IeDEA
International Epidemiologic Databases to Evaluate AIDS - East Africa

Constantin Yiannoutsos
Kara WoolsKaloustian
Beverly Musick
Yee Yee Kuhn
Abraham Siika
Sylvester Kimaiyo

DORIS DUKE
CHARITABLE FOUNDATION

Clinical Scientist Development Award

SEARCH
SUSTAINABLE EAST AFRICA RESEARCH IN COMMUNITY HEALTH

http://www.searchendaids.com/

World Health Organization
REPUBLIC OF KENYA
MINISTRY OF HEALTH
NIH
National Institutes of Health
UNAIDS
GILEAD
THE WORLD BANK
THE REPUBLIC OF UGANDA
MINISTRY OF HEALTH

PIs:	Diane Havlir, Moses Kamya
Statistician:	Maya Petersen
Vice-Chair:	Edwin Charlebois
Virologist:	Teri Liegler
KEMRI:	Elizabeth Bukusi
KEMRI:/UCSF:	Craig Cohen
UCSF:	Tamara Clark, Gabe Chamie, James Kahn, Vivek Jain, Elvin Geng, Carol Camlin
UC Berkeley:	Laura Balzer, Mark van der Laan

Software (Public R packages)

1. Super Learner: SuperLearner()
   - Ensemble Machine Learning for Prediction

2. Targeted Maximum Likelihood Estimation: ltmle()
   - Effect estimation of point treatment and longitudinal exposures
   - Super Learner + targeting for effect parameter
   - Dynamic Interventions
   - Mediation
   - Censoring, Missing Data