Record linkage of national laboratory data in South Africa: a novel platform for HIV policy evaluation

Jacob Bor

with William MacLeod, Katia Oleinik, Sue Candy, Mhairi Maskew, Matthew Fox, Cornellius Nattey, Brendan Maughan-Brown, James Potter, Wendy Stevens, Ian Sanne, Sergio Carmona

February 2, 2018

NSF Big Data Hubs

"Data Sharing and Cyberinfrastructure Working Group"

Health Economics and Epidemiology Research Office HE²RO Wits Health Consortium University of the Witwatersrand

NATIONAL HEALTH LABORATORY SERVICE

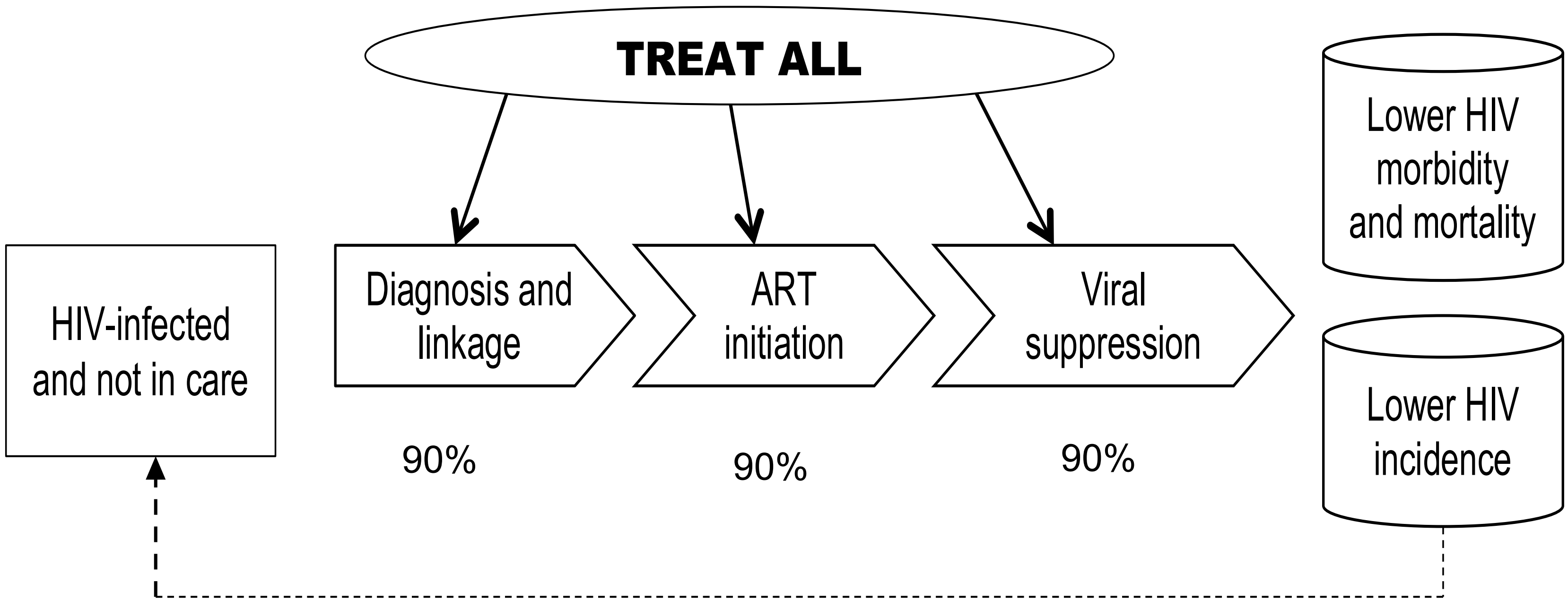
BOSTON UNIVERSITY

- **Large chronic disease epidemics worldwide**
- **Health systems challenge**
- **Key role of data to manage care and inform policy**

- **HIV is a manageable chronic disease**
- **37M people with HIV globally; 7M in South Africa**
- **Lifetime daily antiretroviral therapy (ART)**
  - Near-normal life expectancy
  - Treatment-as-prevention
- **New and ambitious paradigm: <u>‘treat all’ to end AIDS</u>**
  - South Africa moved to ‘treat all’ in Sept 2016

BU

A major challenge



Currently, no dataset provides a system-wide, longitudinal perspective on the HIV care cascade

- National Health Laboratory Service (NHLS) is the sole provider for South Africa’s national HIV program
- Longstanding BU-HE^{2}RO-NHLS collaboration
- ~40 million CD4, VL results, 2004 – May 2015
- >300 million lab tests results in full database
- High quality data; continuously-updated; system-wide
- <u>No unique patient ID</u>…

NATIONAL HEALTH LABORATORY SERVICE

Can we build a National HIV Cohort from routine laboratory data?

Collaboration between:

National Health Laboratory Services, South Africa

Health Economics and Epidemiology Research Office, University of Witwatersrand, South Africa

Boston University

- Departments of Global Health and Epidemiology
- Research Computing Services, Shared Computing Cluster
- Hariri Institute for Computing and Computational Science

INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

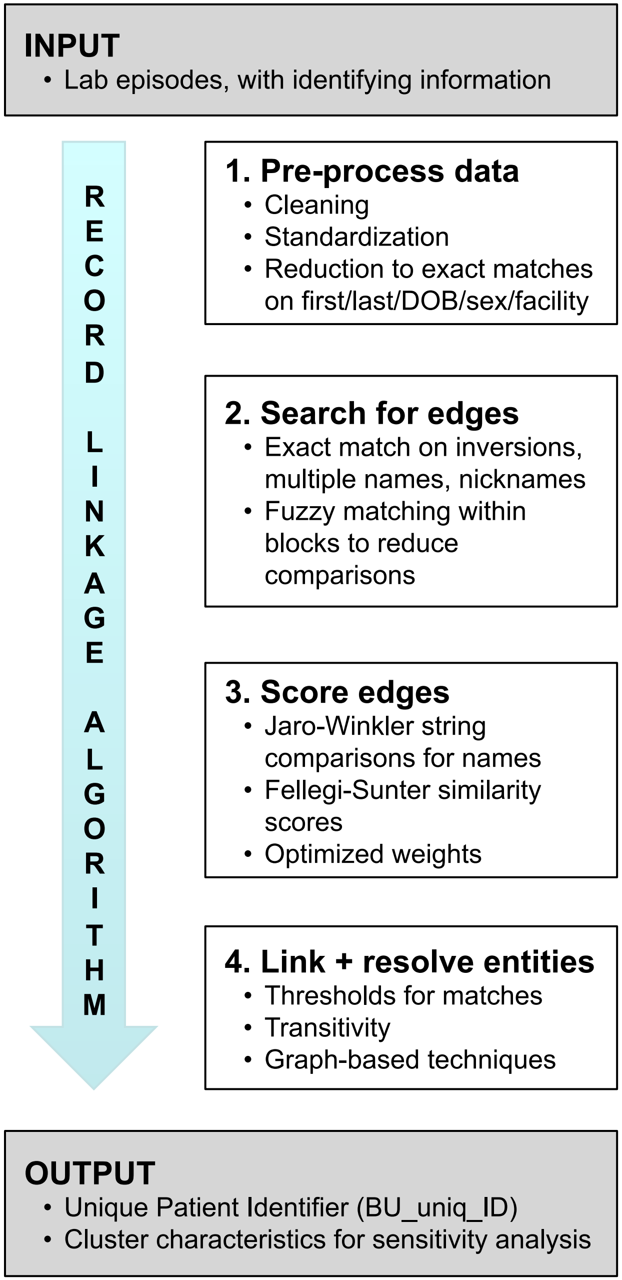
2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility



2. Search for edges

- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce number of comparisons
- Multiple blocking passes

INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

3. Score edges

- Jaro-Winkler string comparisons
- Fellegi-Sunter similarity scores

$$\begin{aligned} sim_k &= \log_2(m_k/u_k) \text{ if match} \\ &= \log_2((1-m_k)/(1-u_k)) \text{ if not match} \end{aligned}$$

$$totsim = \Sigma\, w_k * sim_k$$

- w_k optimized using training data

INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

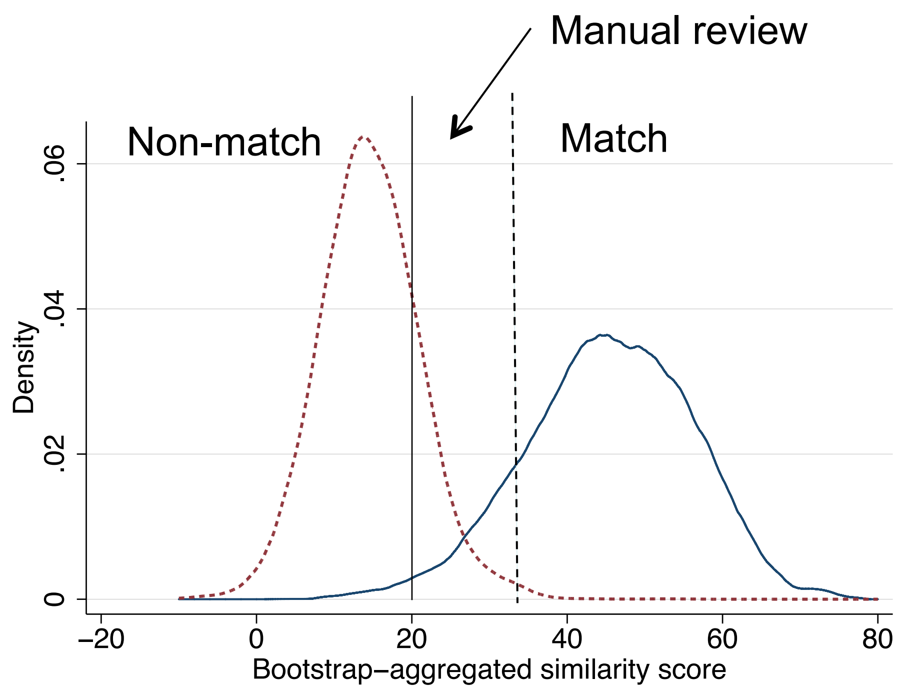
3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

4. Link + resolve entities
- Decision rule for matches
- Transitivity

Traditional approach



INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

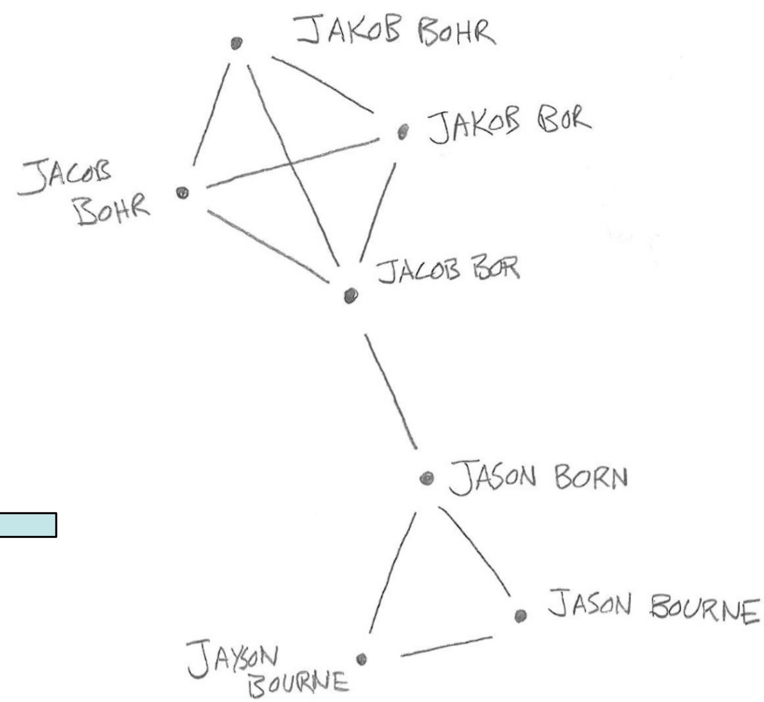
3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

4. Link + resolve entities

Graph-based entity resolution



INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

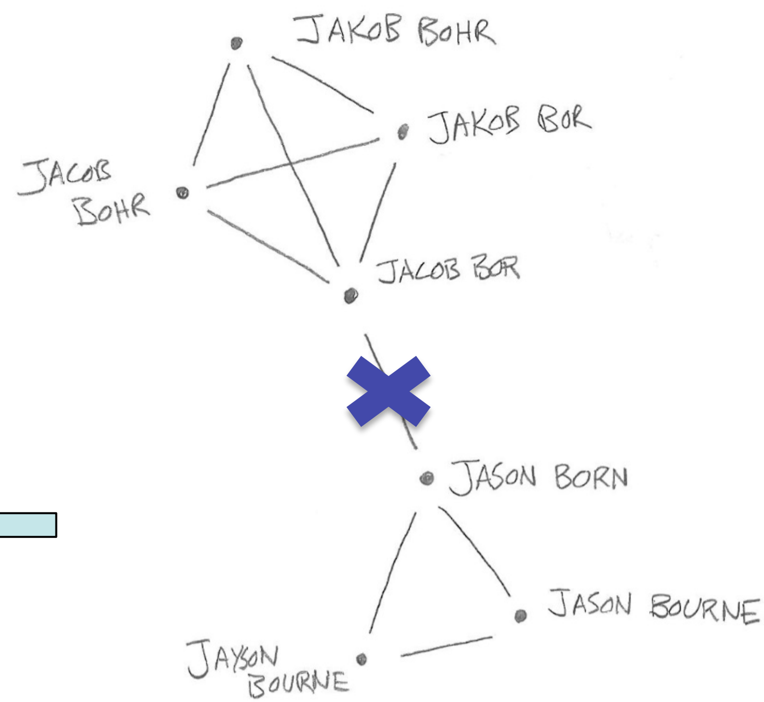
3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

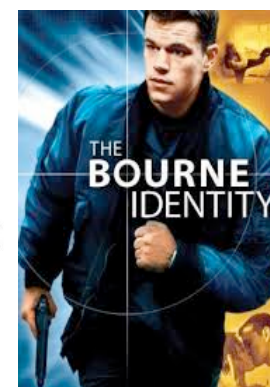
4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

4. Link + resolve entities

Graph-based entity resolution





4. Link + resolve entities

Graph-based entity resolution

INPUT
- Lab episodes, with identifying information

RECORD LINKAGE ALGORITHM

1. Pre-process data
- Cleaning
- Standardization
- Reduction to exact matches on first/last/DOB/sex/facility

2. Search for edges
- Exact match on inversions, multiple names, nicknames
- Fuzzy matching within blocks to reduce comparisons

3. Score edges
- Jaro-Winkler string comparisons for names
- Fellegi-Sunter similarity scores
- Optimized weights

4. Link + resolve entities
- Thresholds for matches
- Transitivity
- Graph-based techniques

OUTPUT
- Unique Patient Identifier (BU_uniq_ID)
- Cluster characteristics for sensitivity analysis

Ongoing: Hariri Research Award to develop new graph-based linkage techniques with George Kollios and Lorenzo Orecchia

Linkage Results

- 38.5 million lab test results (through 2015q1)
- 18.7 million exact matches on first name, last name, date of birth, gender, and facility
- 9.2 million unique patients identified through probabilistic matching techniques

→ "NHLS National Patient Cohort"

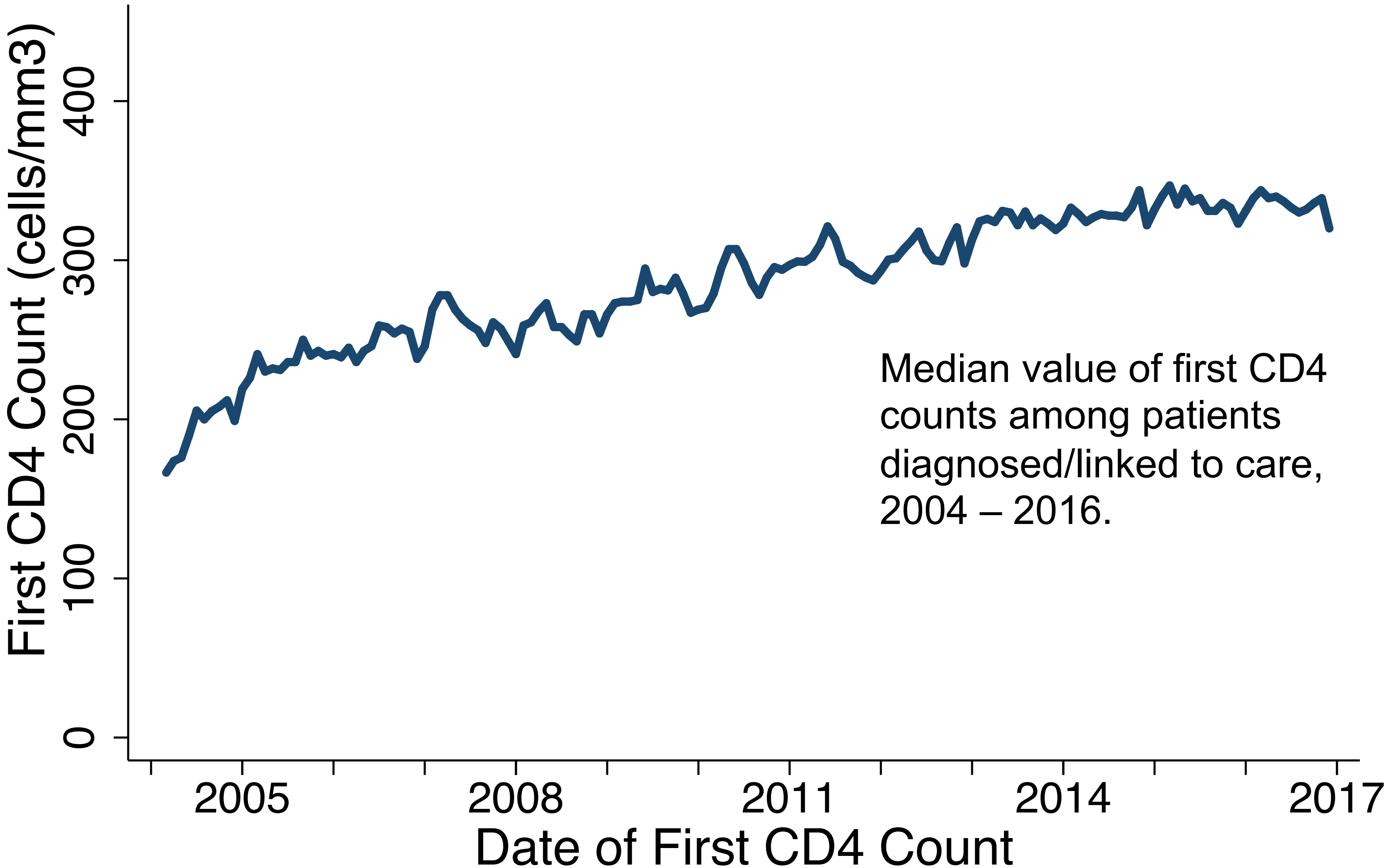
Cohort Profile

- **9.2 million people** have ever sought care for HIV. About **40% of these are single CD4 counts**. Many who test positive never return to care.
- **3.1 million patients were on ART** and virologically monitored during 2013-2014. Compares to 3 million reported to be on ART by NDOH.

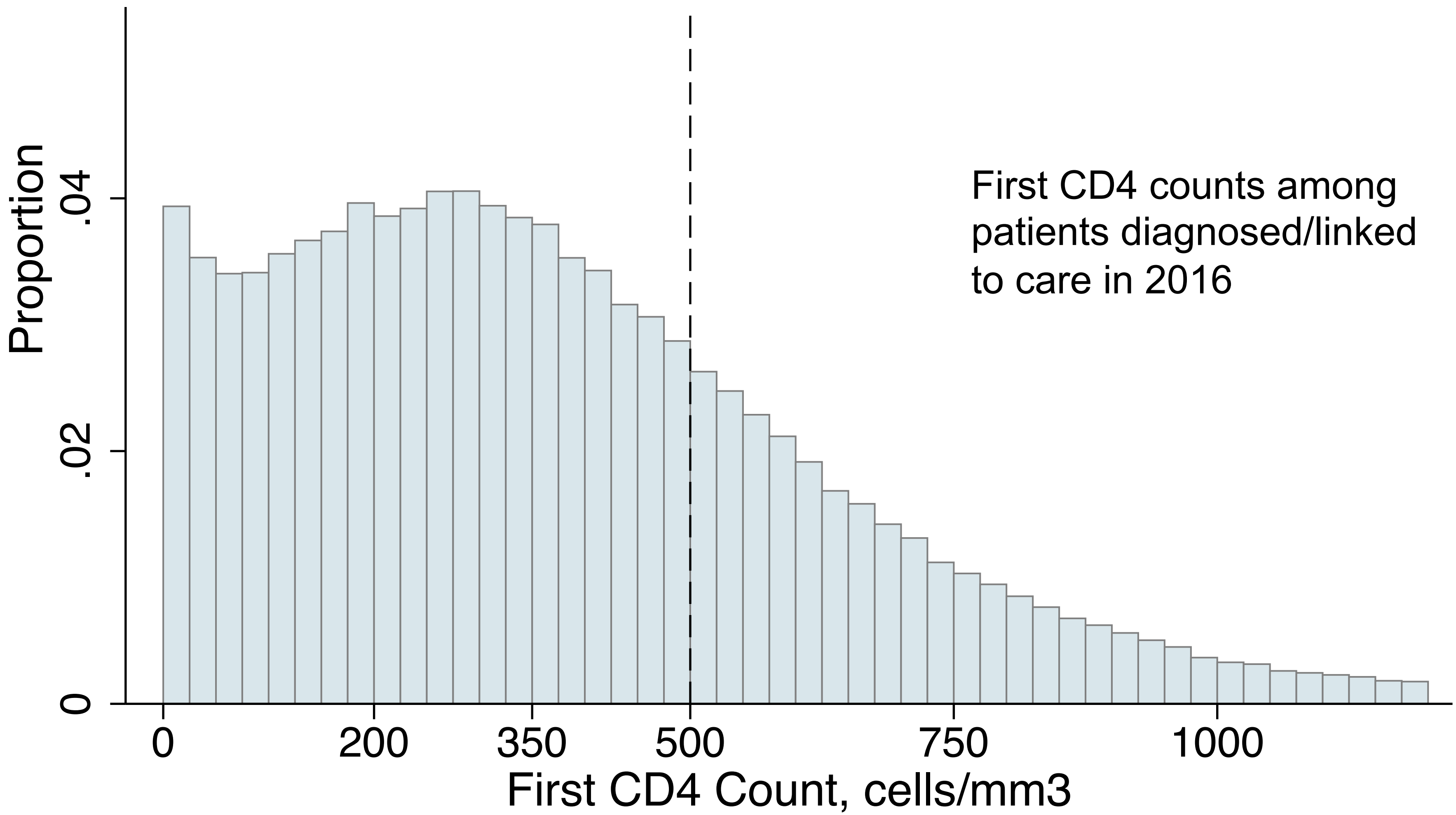
Can South Africa "treat all"?

Preliminary findings

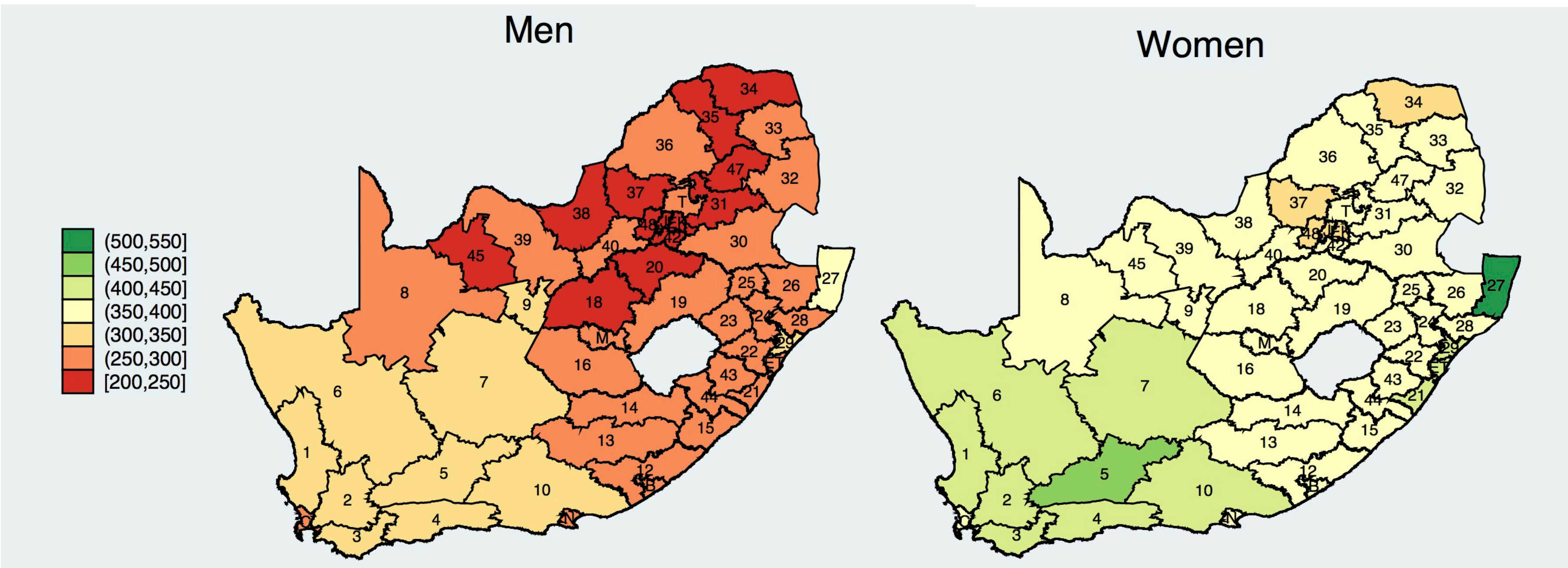
Patients are presenting for HIV care earlier in infection than ever before



But many still present quite late

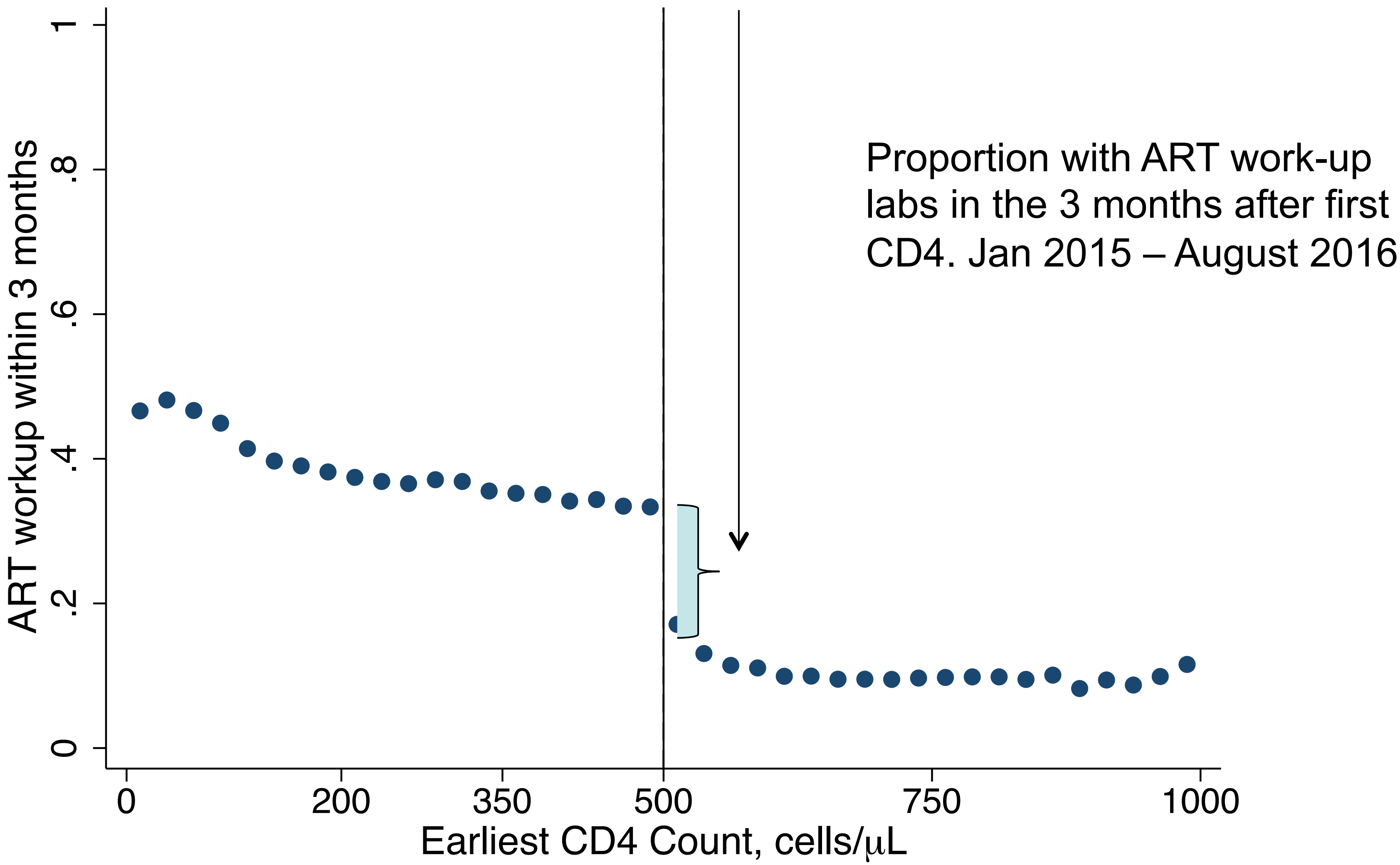


Heterogeneity by gender and district

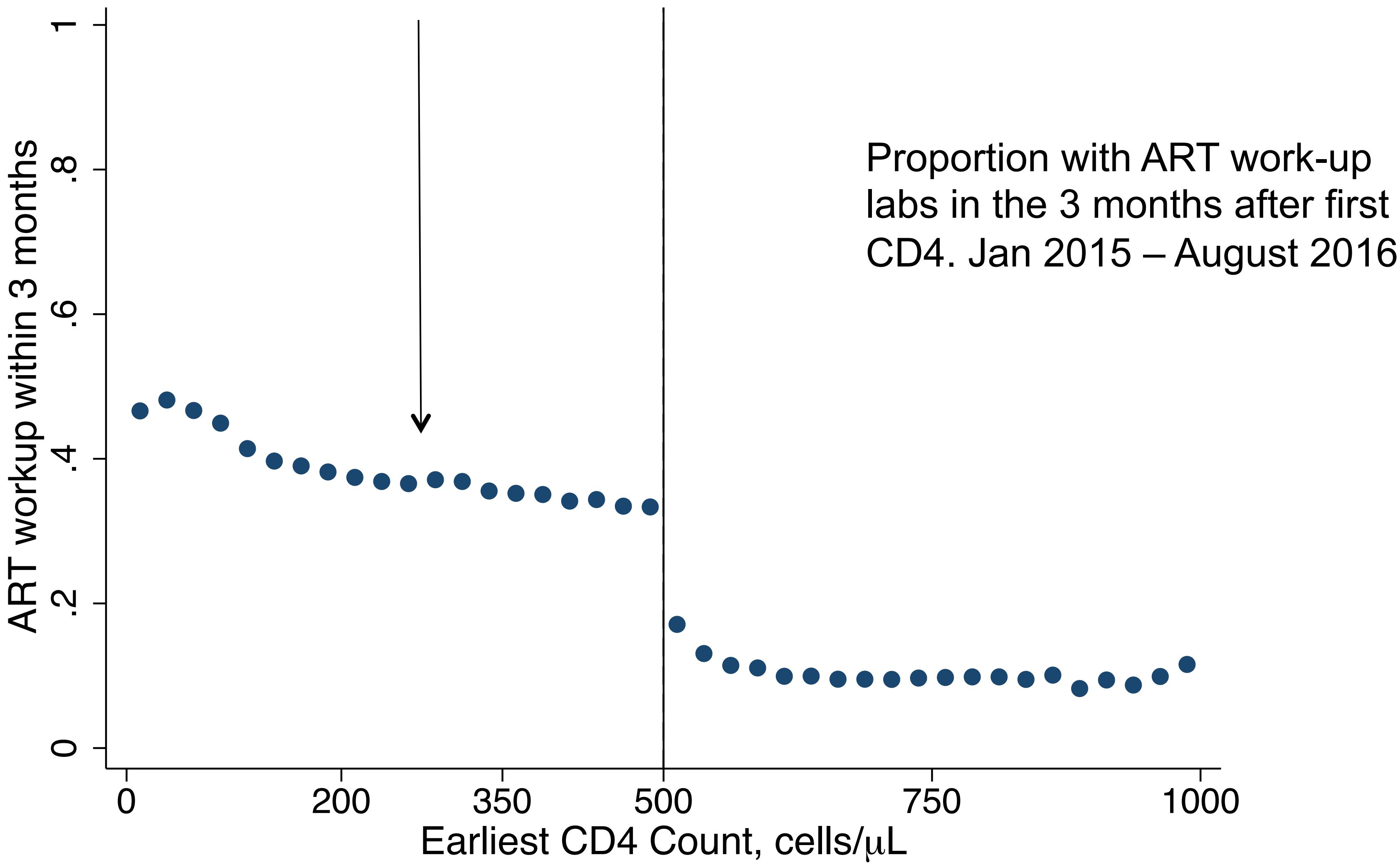


Median CD4 Counts at Presentation, 2014

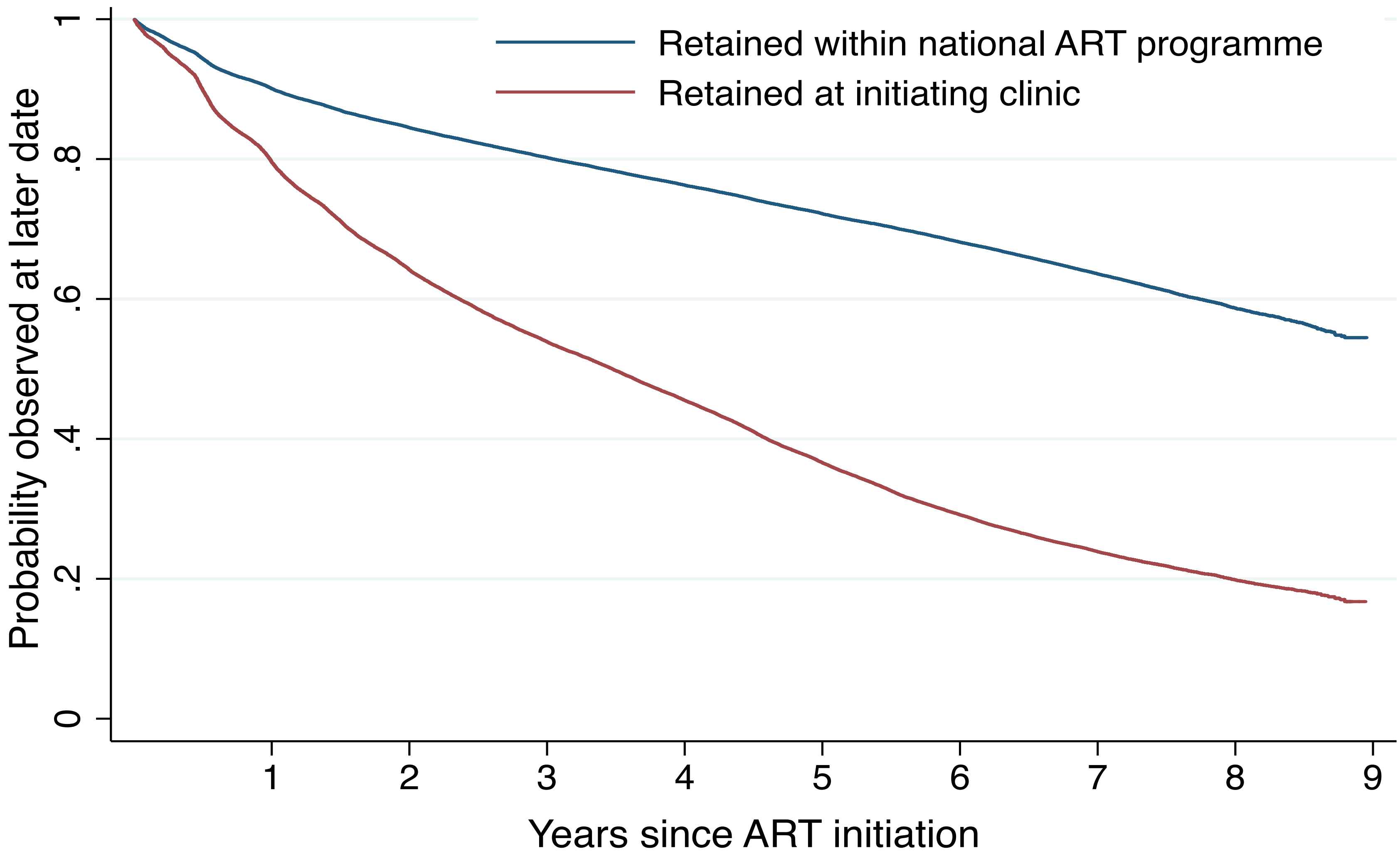
“Treat all” will increase ART uptake among patients with CD4>500



But many patients do not start ART **despite being eligible**



Retention on ART is higher than previously thought



Can South Africa "treat all"?

Perhaps, but further efforts are needed

- To increase early diagnosis and linkage, particularly among men and in some districts
- To increase ART uptake among those offered therapy

What’s next?

Building a "digital population health" ecosystem from routine laboratory data

Novel Linkage of Routine Laboratory Data for Complete Public Sector Health System

Additional Patient Linkage
- Tier.net
- Vital statistics
- Other clinical cohorts

Methods development
- Graph-based record linkage
- Analytical approaches that account for uncertainty

Security & Ethics
- Data platform security, access
- Big Data bioethics

Integrate into Clinical Care
- Develop systems to integrate into patient care
- Inform chronic disease care

Population Impact Evaluation
- Facility catchment areas
- Linkage to external datasets
- Population health impact
- Spillover effects
- Behavioral responses
- Data to fit simulation models

Clinical Epidemiology
- Describe treatment cascades for different chronic diseases
- Integrated cohort

Continuous Quality Improvement
- Measure quality of care
- Use data to intervene to improve quality of care

Rapid Response Policy Evaluation
- Needs assessments
- Inform clinical guidelines
- Evaluation of policy changes

from patient to population

Extramural support

Awarded

- NIH R01 AI115979-01 (Fox/Maskew) – Analysis of National Lab Database to Evaluate the HIV treatment Rollout in South Africa

Submitted

- NIH R01 (Bor/Fox) – Big Data Methods for Real-Time Evaluation of “Treat All" in the Largest HIV Program in the World
- NIH R01 (Fox/Maskew) – Improving the Adolescent Transition to and Retention in Adult HIV Care in South Africa: a National View
- NIH DP2 (Bor) – Building a ‘Digital Population Health’ Ecosystem From Routine Laboratory Data
- NIH R21 (Jenkins) – Identifying TB transmission hot-spots from routinely-collected laboratory data

Thank you

jbor@bu.edu

sites.bu.edu/jbor