

A stylized graphic of a globe, composed of several intersecting blue and grey arcs, located in the bottom-left corner of the slide.

## Using Big Data to Improve HIV Treatment Program Outcomes in South Africa

# What is big data analytics

**Big data analytics** is the process of examining large and varied **data** sets -- i.e., **big data** -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions.

## Data Science VS Big Data VS Data Analytics

DATA IS GROWING FASTER THAN EVER BEFORE.



Each person-  
**1.7 megabytes**  
created



### WHAT ARE THEY?



**Data Science** is a field that comprises of everything that related to data cleansing, preparation, and analysis.



**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.



**Data Analytics** Involves automating insights into a certain dataset as well as supposes the usage of queries and data aggregation procedures.

# South Africa analysis is a Big Data analysis because:

- (a) large dataset;
- (b) varied dataset;
- (c) Was used to uncover previously-unknown trends in HIV treatment adherence and success; and
- (d) Improved supervision in the health sector



# Context





# FAST-TRACK

ENDING THE AIDS EPIDEMIC BY 2030

by 2020

**90-90-90**

Treatment

**500 000**

New infections among adults

**ZERO**

Discrimination

by 2030

**95-95-95**

Treatment

**200 000**

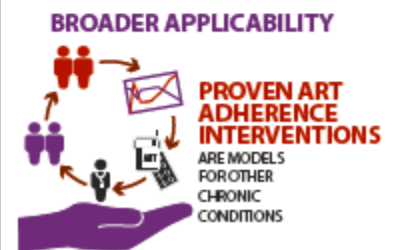
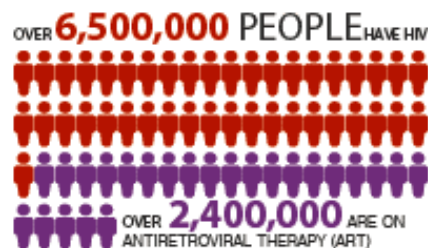
New infections among adults

**ZERO**

Discrimination

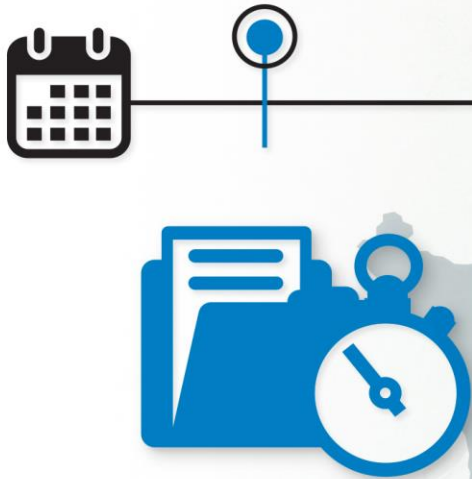
# SOUTH AFRICA's HIV Treatment program

- 1 in 5 people on HIV treatment live in South Africa
- HIV treatment is lifelong and adherence is essential



# Three-phased approach for WB support for HIV treatment program in South Africa

**Rapid  
management  
analysis** and  
“best” estimate  
in **3 months**



**Geospatial  
Intermediate “fuzzy  
data/big data”  
analysis** with  
proximate indicators  
in **1 year**



**Rigorous  
prospective  
evaluation** in  
**2 years**





# the CHALLENGE

Viral suppression in patients with HIV is the best indicator of success in an antiretroviral treatment (ART) program.

South Africa has the largest ART program but...  
**fragmented monitoring systems** with **large gaps in viral load (VL)** data entering the DHIS

Viral suppression in South African patients with HIV in the treatment program



## Four key questions:

1. Do people who are on HIV treatment, get their HIV viral load checked as per South Africa's HIV treatment guidelines?
2. Are people on HIV treatment in SA virally suppressed?
3. Does this viral suppression lead to improved health for HIV patients?
4. Are there spatial patterns to how data are distributed?



# What routine (big) data were available in South Africa to answer these 4 questions?

## TIER.net

HIV Electronic Register

- Three Interlinked Electronic Registers (TIERs)
- Since 2011
- 3-tiered electronic patient management system
- Captures **patient-level data** on [HIV counselling and testing, pre-HIV-treatment and HIV-treatment services](#)





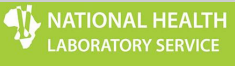
## NATIONAL HEALTH LABORATORY SERVICE

- NHLS is the largest diagnostic pathology service in South Africa
- Supports national and provincial health departments
- Provides laboratory and related public health services to over 80% of the population through a national network of laboratories
- Samples to NHLS laboratory, test performed and results via SMS printer to facility
- Manual transcription to patient file
- Houses a Corporate Data Warehouse (CDW) on all **laboratory tests and their results**
  - For HIV: viral load and CD4 test results
  - NO unique client identifiers

## dhis2

- District Health Information System
- South Africa's health management information system
- Summarises data from 'tick registers' and patient that are completed daily
- Data in DHIS based on national indicator set for health service monitoring
- NOT patient level monitoring
- Includes [aggregate](#) HIV data (number of patients and types of services, in aggregated form) on [HIV testing, HIV treatment and other HIV services](#)

# Big data approaches to answering the Government's 3 questions

	Data Science Analytical approach	Databases used
<b>VLD:</b> Do people who are on HIV treatment, get their HIV viral load detected as per South Africa's HIV treatment guidelines?	<ul style="list-style-type: none"> <li>Create a temporal patient database with consecutive lab results, per facility</li> <li>Compare VL tests performed at specific time intervals against the number of HIV treatment clients at facility</li> </ul>	   Harmonised master list of health facilities
<b>VLS:</b> Are people on HIV treatment in SA virally suppressed?	<ul style="list-style-type: none"> <li>Use temporal patient database with consecutive VL lab results, per facility</li> <li>Check VLS status disaggregated by sub population</li> </ul>	Temporal set of patient data Harmonised master list of health facilities
<b>CD4 recovery:</b> Does this viral suppression lead to improved health for HIV patients?	<ul style="list-style-type: none"> <li>Use temporal patient database with consecutive CD4 lab results, per facility</li> <li>Check CD4 status disaggregated by sub population</li> <li>Determine temporal change</li> </ul>	Temporal set of patient data Harmonised master list of health facilities
<b>Spatial distribution:</b> Are there spatial patterns?	2 types of spatial correlation analyses: <ul style="list-style-type: none"> <li>Moran's I</li> <li>Geary's c</li> </ul>	VLD and VLS results from above Harmonised master list of health facilities

# DATA SCIENTIST

Math  
Statistics  
Programming  
Database  
Domain Knowledge  
Soft Skills  
Communication  
Visualization



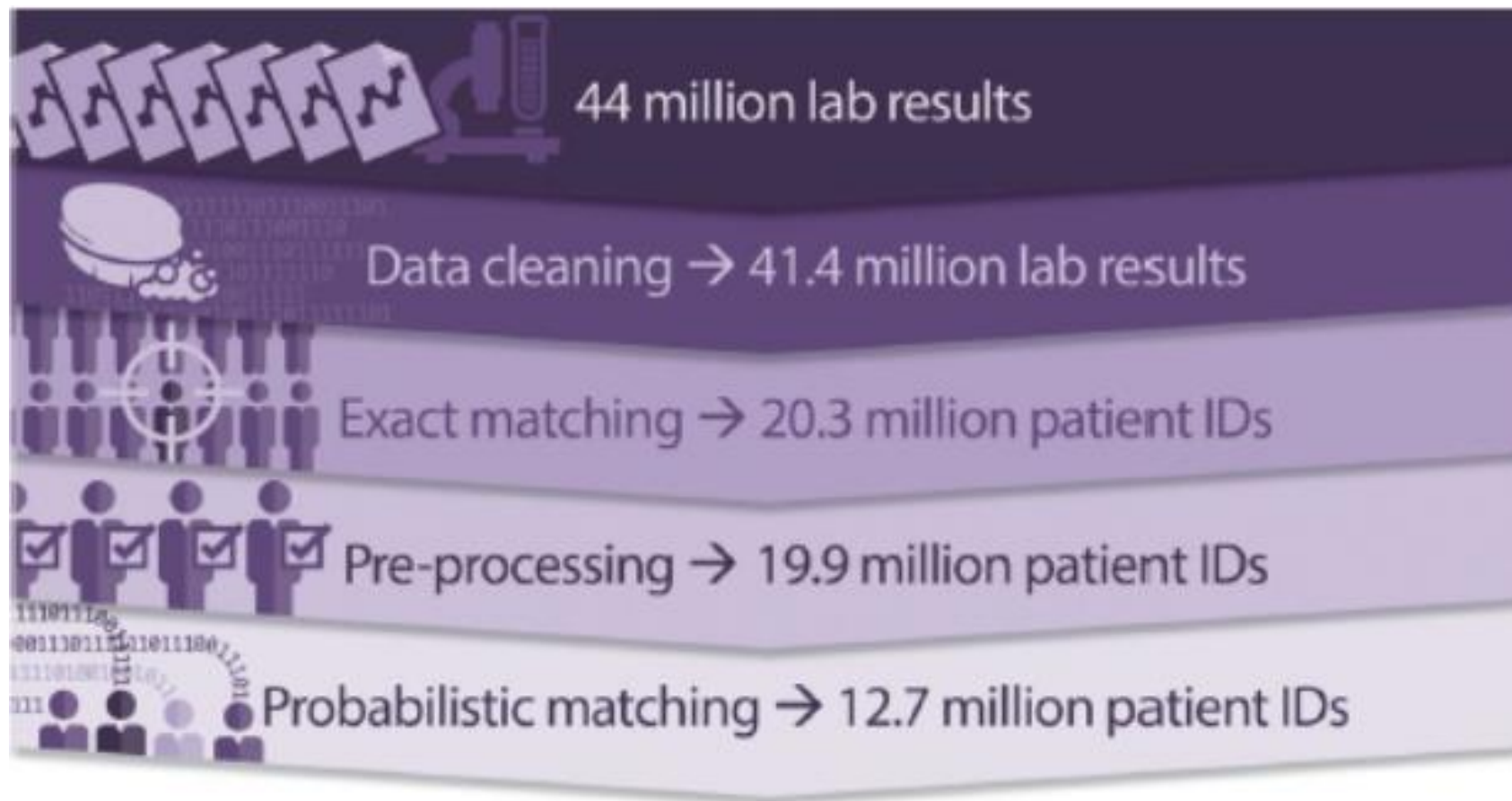
# Data Science Aspects



# Data Science Aspects

- Obtained harmonised list of health facilities
  - 4 different lists, curated by different persons
- Crowd sourcing to obtain some health facility names and locations
- 3,642 of 3,775 DHIS-facilities could be linked to NHLS data

# Fuzzy Matching Algorithm developed for the purpose of these analyses



# After patient-linked cohort established ....

- Estimated proportion of clients receiving a VL test in a 12-month period at the facility level.
- Grouped VL test results in four categories (<400, 400–1000, >1000, and >10,000 copies/mL), as per the VL-based client management guidance in the National ART guideline.
- Estimated the proportion of viral load tests done (VLD) and proportion of ART clients virally suppressed (VLS) by province, district, subdistrict and health facility.
- Assessed if there is any relationship between facility size (determined using the number of clients on ART at each facility) and viral suppression levels.
- Determined if poorer-performing facilities were spatially grouped (i.e. in one district).



# ▶ South Africa Big Data and Geospatial Analysis



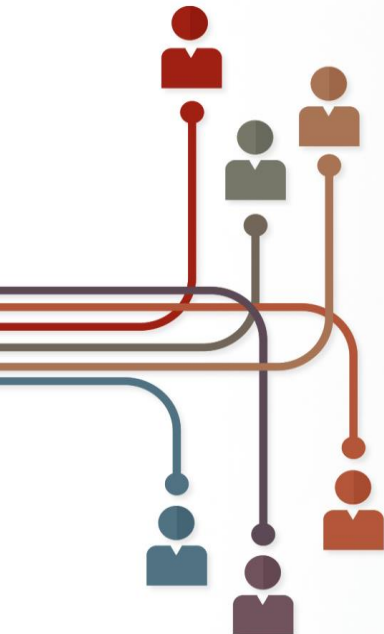
VL and CD4 data are **underused** for decision-making in SA



With **>3 million ART patients**, targeting of ART adherence efforts **must be guided by data**



**Linking databases** from National Health Laboratory and National Health System **through complex matching procedure** and newly developed algorithm



**44 million lab results** matched to 12.68 million new, unique patient IDs



Provides **spatial and demographic pattern** of viral suppression levels

**Mitigates lack of working unique patient identifier** in South Africa





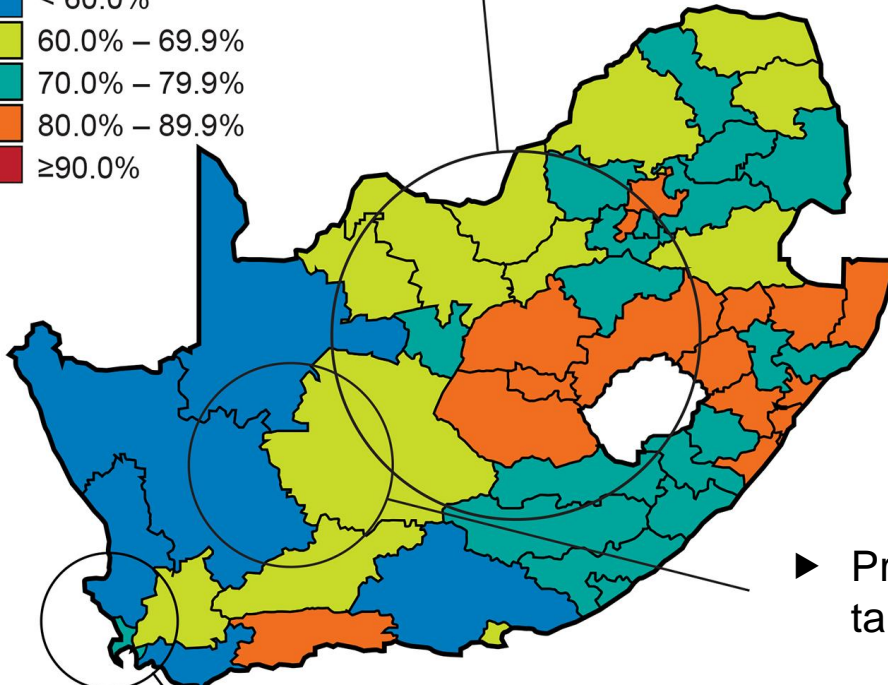
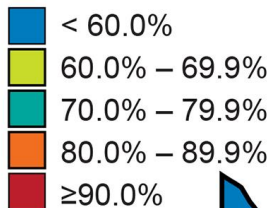


# Big Data Analytics Aspects

# ► Summary big data analysis

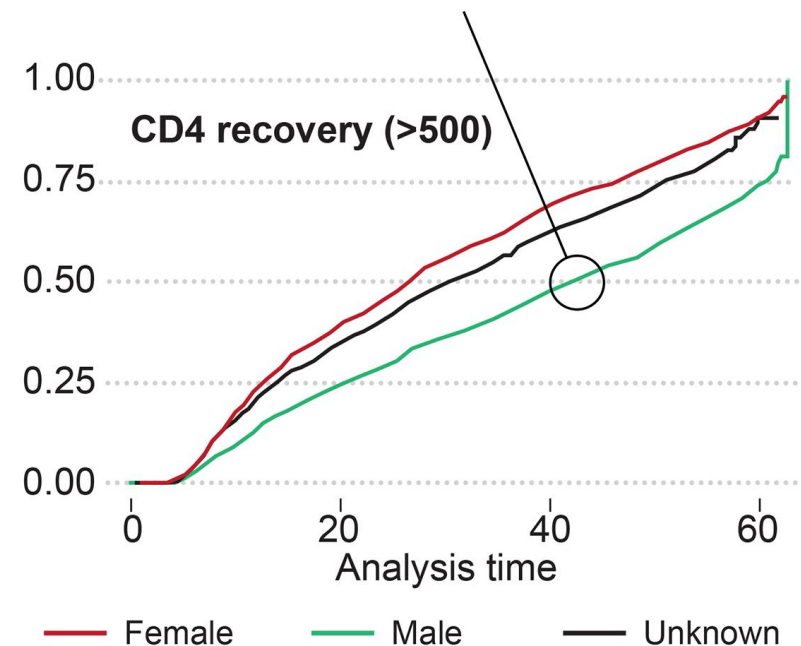
- 13% of HIV treatment patients highly infectious (VLS >10,000 copies)
- **Best facilities** had VLS **60%** higher than worst
- **Best districts** had VLS **40%** higher than worst
- Largest quartile facilities **15%** above smallest quartile facilities

Proportion viral load suppression



- Cape Town positive outlier but rest of Western Cape negative outlier

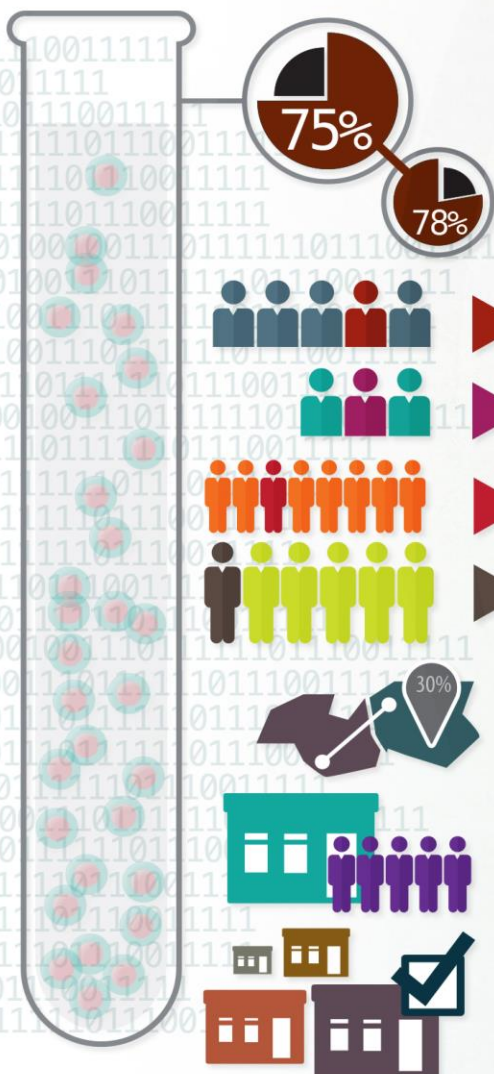
- CD4 immune reconstitution **50% lower among men** and even lower among older men



- Prospective impact evaluation targeting **lowest VLS districts**



## ► VLS results



**75%** had received at least one VL test in the 12 months HIS **only reports half of these**  
Of these, **78% clients virologically suppressed**, but:

- 1 in 5 not suppressed
- 1 in 3 of the under 25s not suppressed
- 1 in 8 had VL > 10,000 (high risk of transmission)
- 1 in 6 male patients > 10,000

Best performing districts had 30% higher VLS than worst performing districts

Facilities and districts with **higher ART patient** numbers do better on VLS

200 clinics with **VLS below 50%**

3.6% of clinics reach VLS of 90% or more

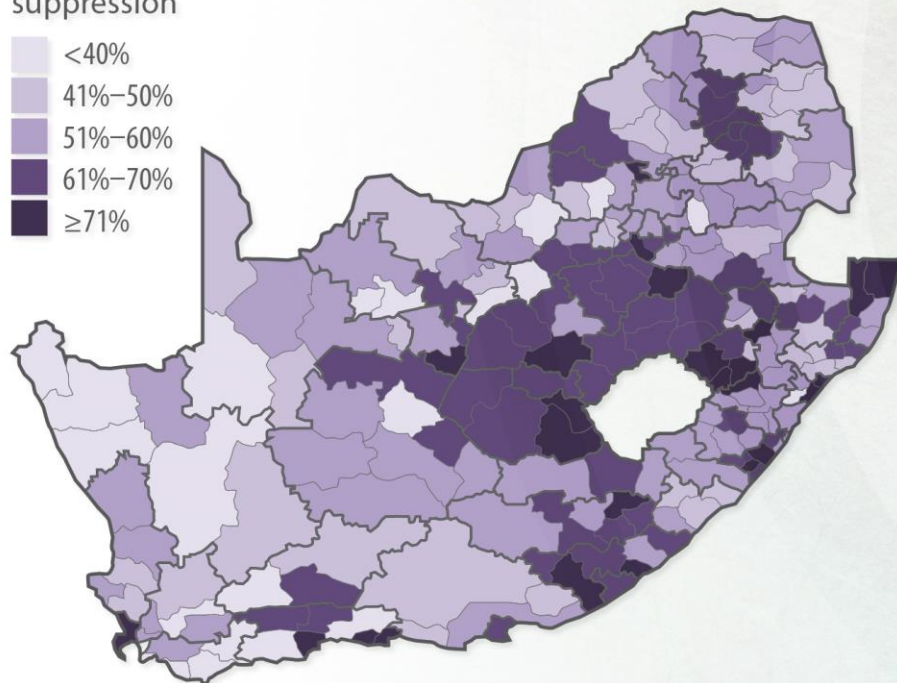
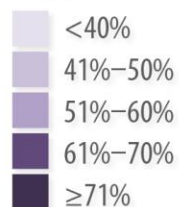


# ► Good and not-great VLS results

## Identifying Successes

Proportion of ART clients with known VL suppression (<400 cp/ml)

Proportion viral load suppression

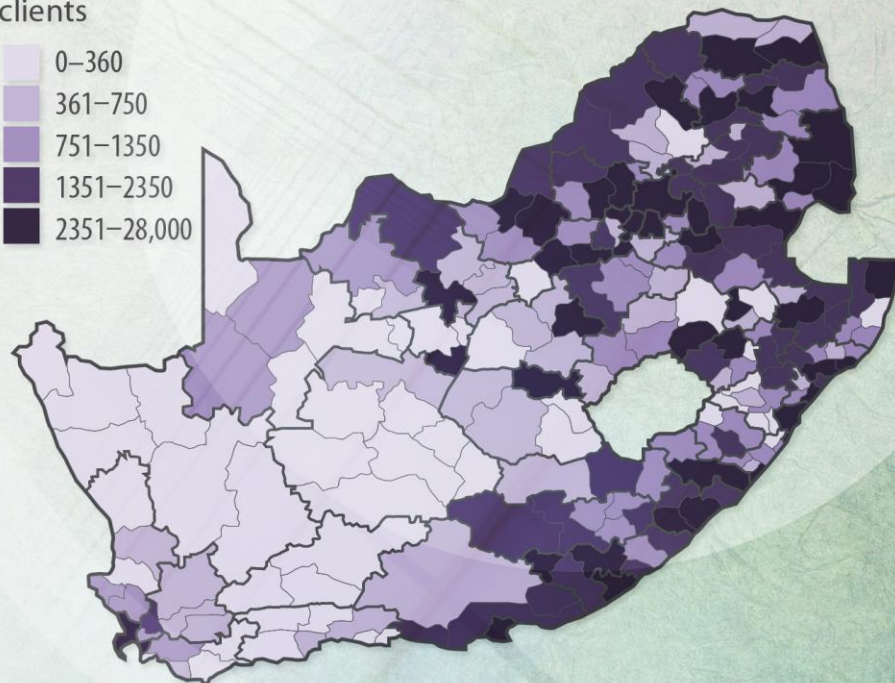
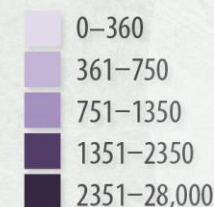


Can we learn from the dark-shaded sub-districts?

## Identifying Failure

Number of ART clients with high VL (>1,000 cp/ml)

Number of clients



Low hanging fruit for better adherence support



## ▶ CD4 recovery results

**900K CD4 tests included** of persons with estimated ART initiation in 2010–14 and evidence of VLS



47% with baseline CD4 <200



**70% tests are from females**



Males start ART later at median CD4 177 (females at 228)



Time to immune recovery:



longer among males, to 200, 350 and 500



increased with higher age of patient



decreased with the calendar year of ART initiation



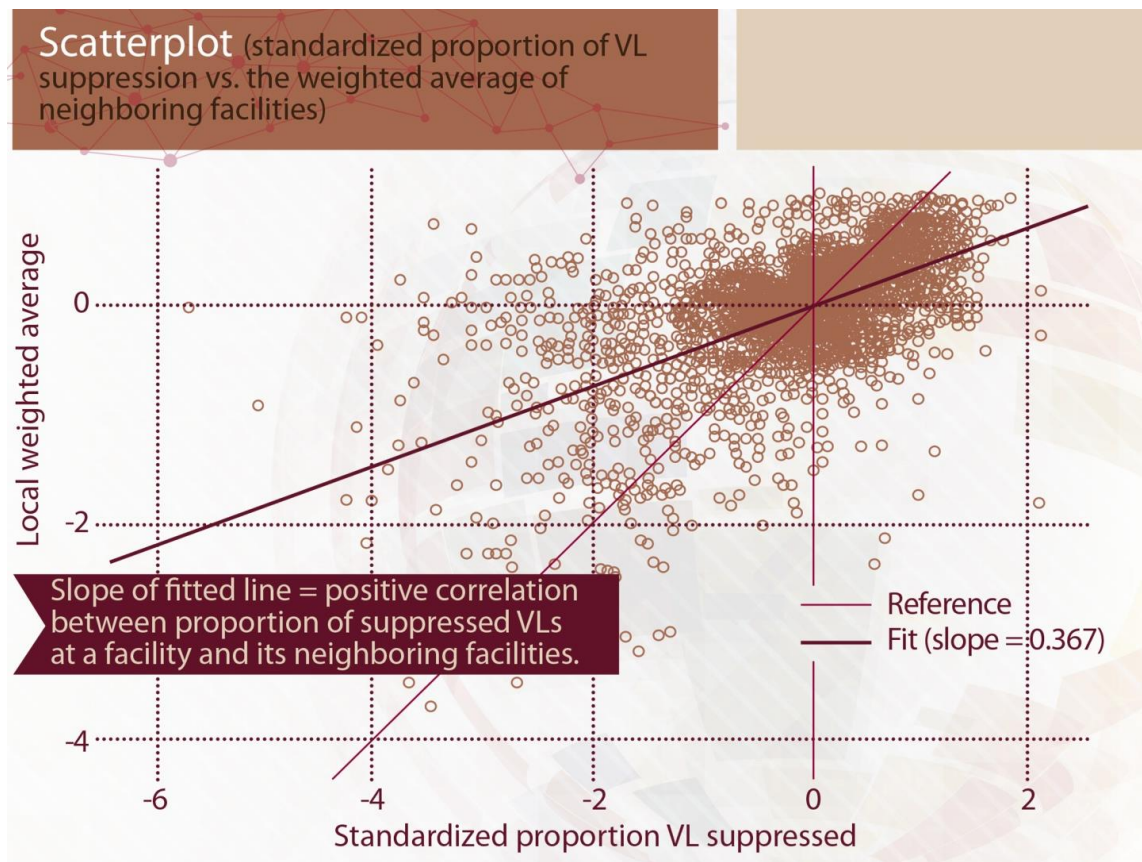


# **Geospatial Analytics Aspects**



**Moran's I:** 0.246 (0.241, 0.251 95% Confidence Interval)  
(0 = random spatial pattern)

**Geary's c:** 0.54 (0.48, 0.60 95% Confidence Interval)  
(0 = perfect spatial correlation; 2 = perfect dispersion)



**Because of results, decided to include district health management team-level efforts in facility improvements (part of prospective impact evaluation)**



*There has been considerable expansion in routine data generation over the last decade in health care (and beyond) as well as in methodologies and technologies allowing innovative analysis and use of such data. Linking vast numbers of records and subsequent analysis is one such 'Big Data' method that has become an*

*for governments and private sector organizations in more efficiently and e*

eries, tells the story of an inn  
which, when combined with  
on on viral suppression amo  
aphy. The analysis emphasi  
support and identifies suc  
ould inform targeted ART pro

## Analysis of Big Data for Improving Antiretroviral Treatment Programmes

### Determinants of CD4 Immune Recovery among Individuals on ART in South Africa

A National Analysis | February 2016



#### PROBLEM

the largest ART programme  
ities, but there is incomple  
el about the proportion of  
amount of HIV in their body  
w levels of HIV in ART client  
of an HIV treatment progr  
better and living longer. It a  
rough reduced HIV transmi  
smit HIV to others.

monitoring test results, including  
National Health Laborator  
mat at health facilities or th  
e faring in terms of VLS of tl



Linking databases  
from National  
Health Laboratory  
and National Health  
System through  
complex matching  
procedure and  
newly developed  
algorithm

### DETERMINANTS OF CD4 COUNT RECOVERY IN SOUTH AFRICA'S NATIONAL HIV CARE AND TREATMENT PROGRAMME: A NOVEL DATA ANALYSIS TO GUIDE ACTION

*In HIV infection, the CD4 cell count is the best known, most studied and readily available prognostic marker of disease progression. It makes sense as a marker because decline in CD4 cell numbers is an effect of HIV, and CD4 T-cell depletion causes immune deficiency. Once a person is infected with HIV, the virus begins to attack and destroy the CD4 cells of the immune system.*

In HIV infection, the CD4 cell count is the best known, most studied and readily available prognostic marker of disease progression. It makes sense as a marker because decline in CD4 cell numbers is an effect of HIV, and CD4 T-cell depletion causes immune deficiency. Once a person is infected with HIV, the virus begins to attack and destroy the CD4 cells of the immune system.

For most of the HIV treatment era, medical professionals have referred to the CD4 count to decide when to begin treatment during HIV infection. Only recently, WHO and medical guidelines have changed to recommend treatment at all CD4 counts as soon as HIV is diagnosed. Prior to launching its Universal Test & Treat policy (treatment initiation for all HIV positive individuals) in September 2016, the South Africa National Department of Health and the National Institute of Communicable Diseases wanted to take full stock of CD4 patterns across populations and geographies in South Africa. To do so, laboratory and clinic data had to be linked across data systems. In the absence of a unique identifier to match patients' records, a new linking procedure and algorithm was developed, tested and applied to almost 4 million individual records with evidence of a CD4 test (public sector laboratory results are all archived at the National Health Laboratory Services' Corporate Data Warehouse).

This policy brief, part of a World Bank series, tells the story of this innovative Big Data analysis of laboratory CD4 counts undertaken in South Africa, providing new strategic information on CD4 count recovery among individuals who initiated antiretroviral therapy in South Africa, by geography and demography. The analysis emphasizes population groups, districts and provinces that need enhanced linkage to HIV care and treatment, adherence support and continued CD4 count monitoring in order to improve ART treatment outcomes. Such strategic information should inform ART programme improvements.

## Analysis of Big Data for better targeting of ART Adherence Strategies

viral load  
ince, district,  
- March 2015)

