Application of Data Science and statistics in HIV clinical research

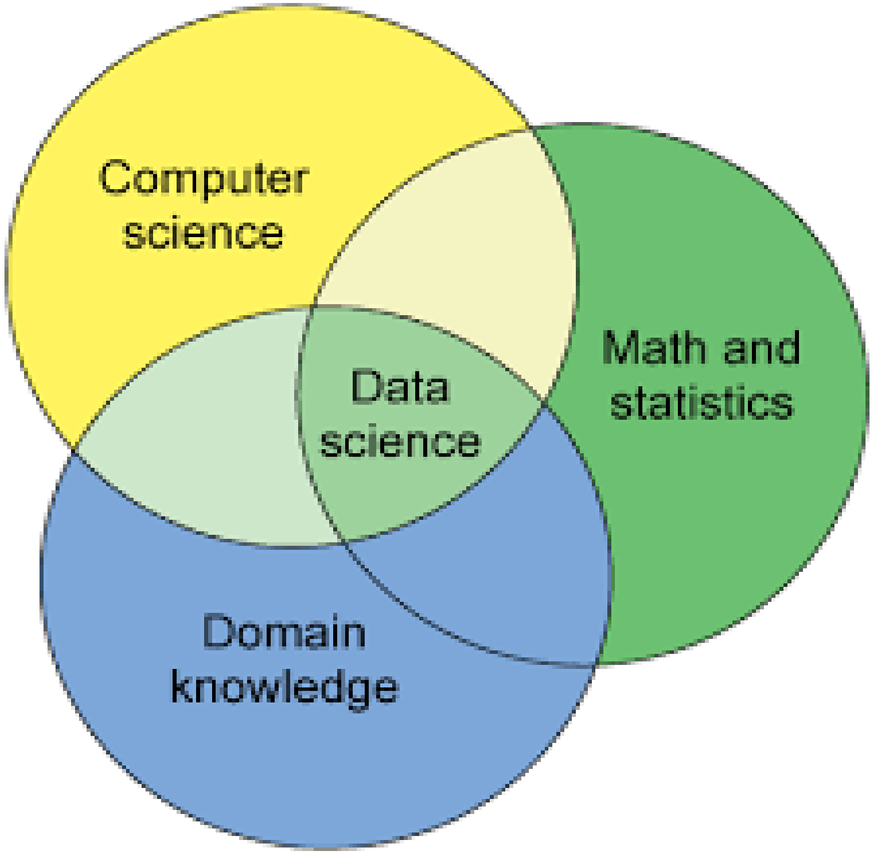
Agnes Kiragga
Head of statistics, Research Department
Infectious Diseases Institute Kampala

Definitions

Statistics (1786): A science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Computer Science (1962): A study of principles and use of computers

Data Science (2001): A multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured.



Data science methods & algorithms

Machine learning:

- A data analysis method that automates analytical model building
- Systems learn from data, identify patterns and make decisions with minimal human intervention

Two broad categories:

- **Supervised Algorithms:**
  - Regression (e.g. Linear Regression)
  - Classification (e.g. Decision trees)
- **Unsupervised Algorithms**
  - K Means Clustering
- Approximately 40 data science techniques

Availability of HIV clinical data

- Infectious Diseases Institute (IDI) clinic Mulago provides care and treatment to over 300,000 ever registered and 8000 active PLWHIV
- 10 year cohort of PLWHIV with structured laboratory and clinical monitoring
- IDI Implementing partner for CDC in four regions (Kampala, Central, Bunyoro and West Nile regions)
- Access to data from approximately nation's 300,000 PLWHIV
- Partnerships that foster access to data from MOH PrEP and EID/PMTCT dashboards

1. Can we merge HIV clinical data and data science techniques?

2. How good is your model?

- Evaluated using performance measures and “confusion matrix”

https://towardsdatascience.com/creating-intelligence-with-data-science-2fb9f697fc79

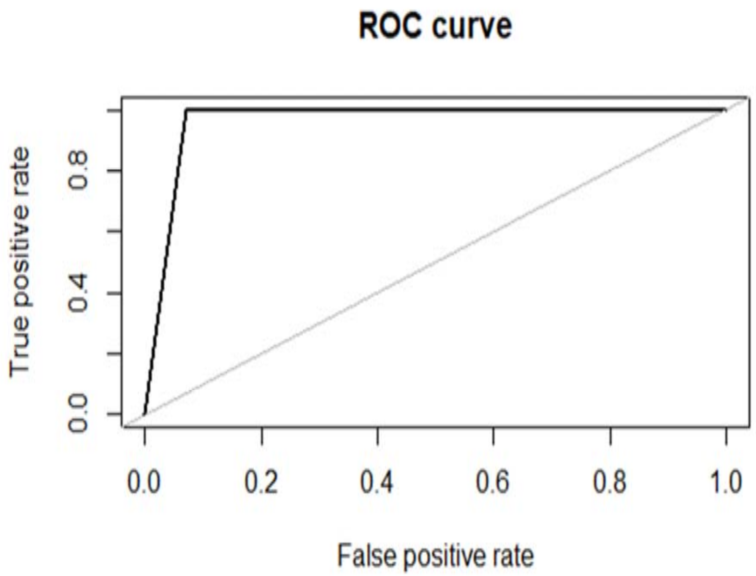
Project 1: Predicting mortality after ART initiation

- Antiretroviral therapy (ART) has significantly improved survival of HIV patients and changed HIV infection to a manageable chronic disease[1]
- Mortality during early ART roll-out ranged between 15 – 35% and has since decreased with changing ART initiation guidelines[2]
- Aimed to apply machine-learning (ML) techniques to predict all-cause mortality amongst patients previously in a 5 year randomized ART trial
- Data from 377 patients (153 men and 224 women)
- Data randomly split into nine-tenths to train and the rest used to test
- Used Random Forest (RF) and Support Vector Machine (SVM) techniques

1. Quinn et al, 2008, Kambugu et al CID 2009

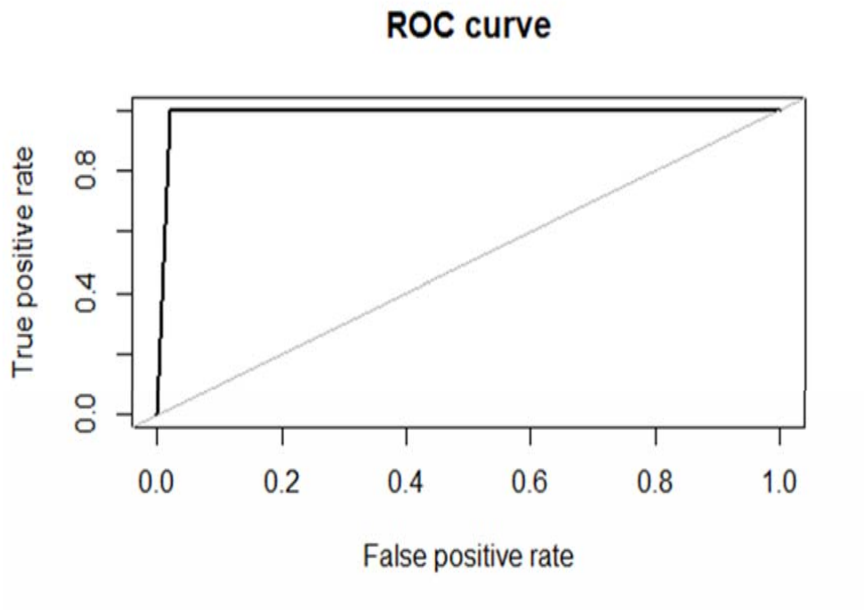
Receiver Operating Characteristic curve (ROC)

ROC for Support Vector Machine



Accuracy (95% CI) = 0.93 (0.92, 0.95)
Area under curve (AUC) = 0.97
Sensitivity = 1.00
Specificity = 0.93

ROC for Random Forests (RF)

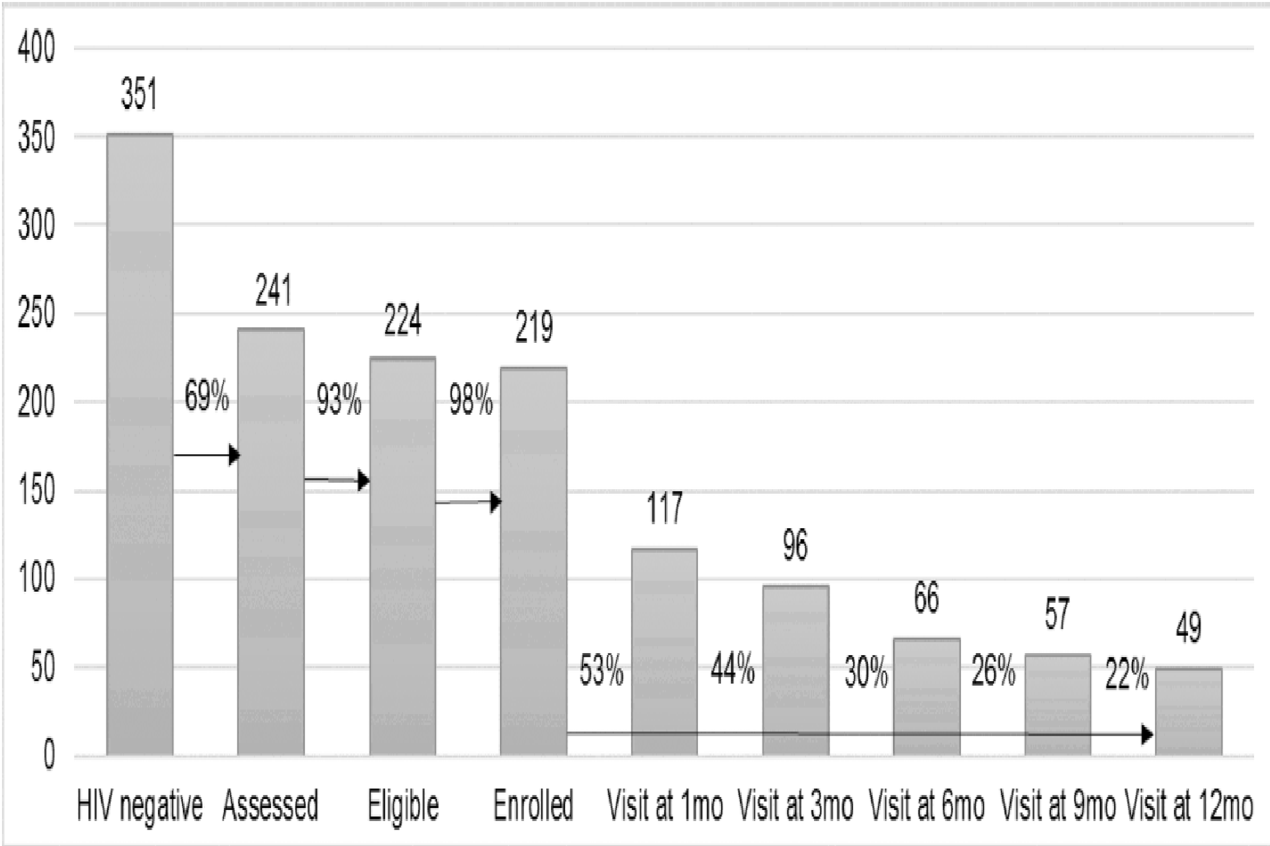


Accuracy (95% CI) = 0.98 (0.97, 0.99)
Area under curve (AUC) = 0.99
Sensitivity =1.00, Precision = 100%
Specificity = 0.98, Accuracy = 98%
F1 score = 58% Recall = 41%

Project 2: Machine learning to predict retention in PrEP programs

- WHO recommends Oral Pre-exposure prophylaxis (antiretrovirals for HIV negative persons) for HIV prevention
- Uganda MOH recommends PrEP for key and priority populations e.g. SW, TG, MSM, FF, AGYW, PWIDs, AGYW etc
- Worryingly low figures of retention in PrEP <50%
- Can we predict who will drop out and design targeted retention strategies using ML?

PrEP Cascade



Eakle R et al (2017), HIV pre-exposure prophylaxis and early antiretroviral treatment among female sex workers in South Africa

Methods

- De-identified data were extracted from an electronic web-based PrEP tracker and dashboard from at 5 implementing sites in the central (urban) and mid-western (rural) regions of the country.
- Retention was defined as having at least one follow-up visit following PrEP initiation.
- We implemented the XGBoost algorithm in Python to predict retention.
- 7800 patients initiated on PrEP (August 2018)
- Data were split into training (70%) and test datasets (30%)
- Evaluated model performance using ROC, accuracy, precision, F score

Preliminary results and next steps

- Over all retention observed among 42% of clients initiated on PrEP
- The model precision was 0.975, F score was 0.958 and a C statistics from (ROC) curve of 0.982 (95% CI: 0.965–0.995)
- FSW and persons 18-24 likely to drop out of PrEP programs
- Developing a risk score for PrEP retention/drop out

Other ongoing projects:

Project 3: Development of Risk score for predicting disengagement from PMTCT programs

Project 4: Predicting which patients are likely to be successfully found during community tracing of persons who disengage from ART programs

Conclusions

- Data science can be used to predict key outcomes such as mortality, retention in HIV clinical research
- Wide set of methods that can be combined with large clinical databases to design solutions to common problems
- Data science techniques/models can identify subset of populations for targets interventions
- Future results will guide development strategies for national HIV care programs

Acknowledgments

- Mark Okello – data scientist

- Infectious Diseases Institute staff
  - Statistics unit
  - Research
  - Outreach
  - Academy program

- Ministry of Health staff

- Monitoring Evaluation Technical Support Program