Predictive and Spatial Analysis of HIV/AIDS Outcomes in NYC

1. Research Background and Research Questions

Research Problem Description:

The ongoing public health dilemma of HIV/AIDS in New York City continues to be widespread despite the intervention efforts in place. The Department of Health and Mental Hygiene (DOHMH) HIV/AIDS Annual Report offers an in-depth examination of numerous indicators—HIV and AIDS diagnosis, access to care, viral suppression, and mortality, among others—stratified by geographic regions (boroughs and UHF areas) and demographic subgroups for adults 18 years and older. Our study project aims to apply machine learning methods and spatial analysis to predict key health outcomes, including achievement in accessing care and viral suppression, while uncovering spatial and demographic disparities. In doing so, we hope to increase the effectiveness of targeted public health interventions and streamline the distribution of resources.

Primary Research Question:

How can a unified analytical framework that integrates predictive modeling with spatial analysis effectively identify the key determinants of HIV/AIDS outcomes, including participation in care and viral suppression, among different geographic and population-based sub-divisions of New York City?

Related Literature:

HIV/AIDS is one of the most serious public health challenges in the world (World Health Organization: WHO & World Health Organization: WHO, 2024). To control its spread effectively, increase virus suppression rates, and reduce deaths, researchers have explored different prediction methods. Traditional demographic prediction factors play a role in disease monitoring, but they have limits to accurately identify high-risk groups or key factors affecting disease management (Maskew et al., 2022).

Currently, machine learning methods are becoming an important tool in HIV/AIDS prediction studies, providing stronger support for disease management decisions (Infectious Disease Advisor, 2024). However, most current studies focus on predictions for individuals, while exploration of community or regional impacts is still quite limited.

Ramachandran et al. (2020) analyzed patient retention rates at urban HIV clinics to predict which HIV-positive patients are most likely to stop medical care. This study considered geographic factors, but it still focused on individual-level medical care rather than overall trends in HIV diagnosis rates and virus suppression rates within a region or an age group.

In addition, existing studies may ignore the impact of spatial factors such as socioeconomic factors, healthcare resource distribution, and population group on HIV spread. Therefore, to fill this research gap, our study tries to combine machine learning prediction models with spatial analysis to find the influencing factors of HIV/AIDS in specific geographical areas and population segments. By combining individual characteristics with community-level data, we aim to improve the model's usefulness for predicting regional HIV spread, providing better guidance for public health decisions.

Specific Research Questions

Predictive Modeling:

- Which demographic, geographic, and clinical factors (e.g., diagnosis rates, viral suppression, borough/UHF-level indicators) are most predictive of successful linkage to HIV care within 3 months and achievement of viral suppression?
- How do traditional statistical models such as logistic regression compare with ensemble methods like random forests in predicting care linkage and suppression outcomes?

Spatial and Cluster Analysis:

- What spatial patterns or clusters emerge in HIV/AIDS-related metrics across NYC boroughs and UHF regions?
- How do these spatial clusters relate to disparities in public health outcomes such as timely linkage to care and viral suppression?

Dimensionality Reduction and Feature Interaction:

- How can Principal Component Analysis (PCA) be used to reduce dimensionality in highly correlated public health indicators and reveal key underlying dimensions of HIV-related health outcomes?
- Which combinations of indicators retain the most predictive value while minimizing redundancy across diagnosis, prevalence, and mortality variables?

2. Data Description and Variables

The dataset used in this analysis originates from the New York City Department of Health and Mental Hygiene (DOHMH) HIV/AIDS Annual Report, obtained via NYC Open Data. It contains **31,926 rows**, covering data from **calendar years 2017 through 2021**. Each row represents an observation defined by a unique combination of **year, borough, UHF region, age group, race, and gender**, along with associated public health indicators. The dataset includes borough- and UHF-level metrics related to HIV diagnoses, treatment engagement (e.g., linkage to care within 3 months), viral suppression, AIDS diagnoses, and mortality. All records pertain to individuals aged 13 and older. The most recent version of the dataset, released as of **March 31, 2022**, was used in this project.

Key Variables

The dataset includes 18 variables, categorized as follows:

Geographic and Temporal Variables

- **year** (Text): Calendar year of record.
  borough (Text): Borough of residence related to diagnosis or outcome (e.g., Bronx, Manhattan).
- **uhf** (Text): United Hospital Fund neighborhood code used for finer geographic granularity.

Demographic Variables

- **age** (Text): Age group of the individual (e.g., 20–29, 30–39).
  gender (Text): Gender identity (Male, Female, Transgender).
- **race** (Text): Self-reported race/ethnicity. "Other/Unknown" includes multiracial and Native American.

Diagnosis and Prevalence Metrics

- **hiv_diagnoses** (Number): Count of new HIV diagnoses (age 13+).
- **aids_diagnoses** (Number): Count of new AIDS diagnoses.
  concurrent_diagnoses (Number): Cases where HIV and AIDS diagnoses occurred within 31 days.
- **hiv_diagnosis_rate** (Number): Rate of new HIV diagnoses per 100,000 NYC population.
- **aids_diagnosis_rate** (Number): Rate of new AIDS diagnoses per 100,000 population.
  plwdhi_prevalence (Number): Prevalence of people living with diagnosed HIV infection (PLWDHI) per 100 people.

Treatment and Outcome Metrics

- **linked_to_care_within_3_months** (Number): Proportion of individuals newly diagnosed with HIV who were linked to care within 3 months. This is a key binary target variable for one of our models.
- **viral_suppression** (Number): Percentage of PLWDHI whose last viral load was ≤200 copies/mL during the calendar year. This is used as a second outcome variable, modeled as either a percentage or binary indicator.
- deaths (Number): Count of deaths among PLWDHI.
- **death_rate** (Number): Age-adjusted death rate per 1,000 PLWDHI.
  hiv_related_death_rate (Number): Rate of HIV-related deaths per 100 PLWDHI.
  non_hiv_related_death_rate (Number): Rate of non-HIV-related deaths among PLWDHI.

*__PLWDHI:__ People Living with Diagnosed HIV Infection

3. Data Preparation and Processing

Data Cleaning and Transformation

- **Cleaning:**
    - Identify and handle missing or ambiguous entries (e.g., Replace placeholder values like 99999 with NA).
    - Remove duplicates and ensure consistency across reported metrics.
    - Ensure consistency across reported metrics: Confirmed that rates (e.g., hiv_diagnosis_rate, x_viral_suppression) remain within logical bounds (e.g., 0–100 for percentages).
- **Transformation:**
    - Encode categorical variables (Borough, UHF, and binary health indicators) into numeric formats.
    - Scale continuous variables (diagnosis counts, rates) to ensure comparability.
    - Aggregate or stratify data if necessary (e.g., by year or demographic subgroup).
- **Feature Engineering:**
    - Create derived variables such as ratios (e.g., diagnosis_to_death_ratio = hiv_diagnosis_rate / death_rate) to highlight relationships between metrics.
    - Use PCA to reduce dimensionality and capture the most informative features for further analysis.

4. Modeling and Findings

4.1 Logistic Regression

Purpose:

To predict whether individuals newly diagnosed with HIV are successfully linked to care within 3 months, using interpretable linear relationships between predictors and the binary outcome.

Data Interaction:

- **Input Variables**: Diagnosis metrics (e.g., hiv_diagnosis_rate), treatment indicators (e.g., x_viral_suppression), and categorical variables such as borough and uhf.
- **Processing**: Set the target variable x_linked_to_care_within_3_months to a binary label based on a threshold of 0.8 (representing the percentage linked to care). Assign the label **"high"** if the value is greater than 0.8, otherwise label it as **"low"**. Rename this new variable as linked_to_care_binary. Then, use other variables as predictors to build the model.
- **Output**: A classification model and the importance level of each variable. Model evaluation was conducted using confusion matrix and ROC/AUC.

Why Useful:

Logistic regression offers interpretable coefficients, which are valuable for understanding which factors most significantly affect care access.

Result:
The logistic regression model was used to estimate the probability that individuals would be linked to HIV care within 3 months of diagnosis. The model achieved a moderate discriminative ability with an AUC of 0.6895. The most statistically significant predictor was the viral suppression rate (x_viral_suppression, $p < 2e\text{-}16$), indicating a strong relationship between treatment engagement and care linkage.

Additionally, several UHF regions emerged as statistically significant predictors. Residents in neighborhoods such as Greenpoint, Williamsburg–Bushwick, and Sunset Park were more likely to be linked to care early. While the model offered valuable interpretability through odds ratios, its performance in identifying low-linkage cases was limited, as indicated by lower sensitivity.

4.2 Random Forest Classifier

Purpose:
To improve prediction accuracy and uncover non-linear interactions in outcomes such as linkage to care and viral suppression.

Data Interaction:

- **Input Variables**: Demographics (e.g., race, gender, age), geography (borough, uhf), and diagnosis counts (hiv_diagnoses, aids_diagnoses).
- **Processing**: Classify both linked to care within 3 months and viral suppression. If the value is higher than 0.85, label it as **"high"**; otherwise, label it as **"low"**. Use the classification results as new target variables named linked_class and vs_class.
- **Output**: Classification results for both care linkage and viral suppression. Evaluation included accuracy, AUC, and variable importance rankings.

Why Useful:
Random forests modeled complex interactions and performed better than logistic regression.

Result:
Two random forest models were trained:

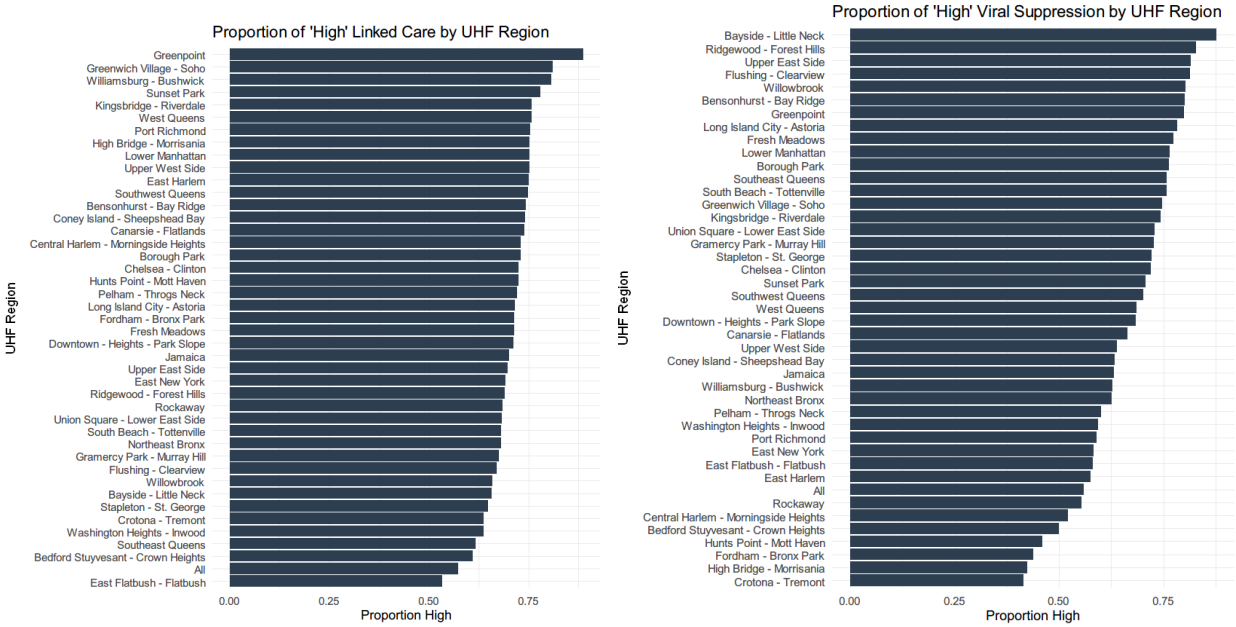
- One for predicting care linkage (linked_class)
- Another for predicting viral suppression (vs_class)

Both models achieved high accuracy and generalizability, with better classification performance than the logistic regression model.

- The care linkage model achieved:
    - Accuracy: 80.8%
    - AUC: 0.88
    - Top features: uhf, race, hiv_diagnoses, and borough
- The viral suppression model achieved:
    - Accuracy: 85.9%

- Balanced Accuracy: 83.1%
- Sensitivity: 92.3% (for identifying "High" suppression areas)

Variable importance plots revealed that geographic and demographic indicators consistently ranked highest, reaffirming the need for location-specific interventions.



4.3 Spatial Analysis and Clustering (K-Means)

Purpose:

To identify clusters of UHF regions in NYC that share similar HIV-related public health profiles.

Data Interaction:

- **Input Variables**: Region-level averages of hiv_diagnosis_rate, aids_diagnosis_rate, x_viral_suppression, and death_rate.
- **Processing**: Data were grouped by uhf, then standardized. K-means clustering was applied, and PCA was used for cluster visualization.
- **Output**: A three-cluster

Why Useful:

The clustering revealed meaningful regional segmentation, identifying neighborhoods like Central Harlem and Crotona as high-priority intervention zones. Although formal spatial autocorrelation tests (e.g., Moran's I) were not conducted, clustering alone revealed clear disparities.

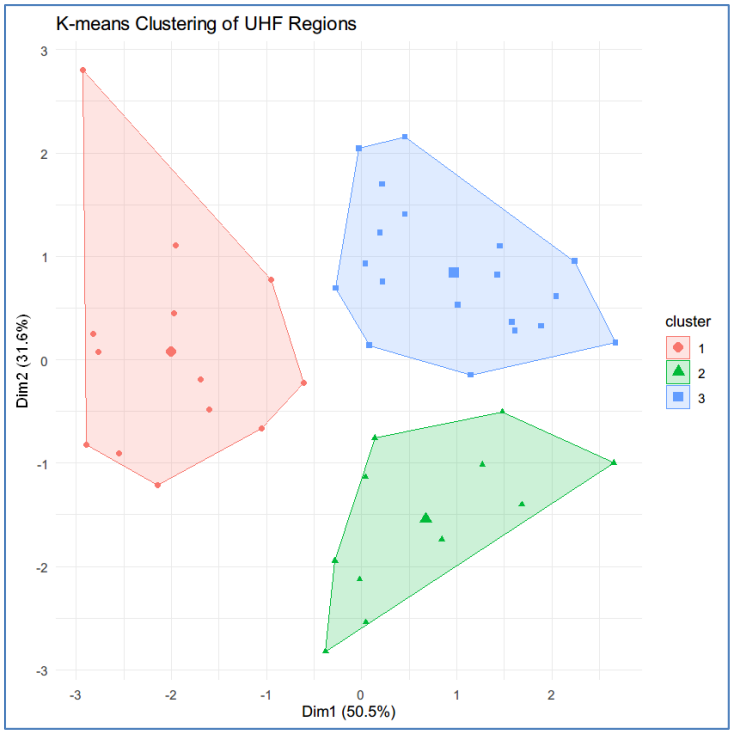
Result:

K-means clustering was performed on standardized averages of UHF-level variables: hiv_diagnosis_rate, aids_diagnosis_rate, x_viral_suppression, and death_rate. Based on the Elbow Method, we identified three spatial clusters:

- **Cluster 1** – High diagnosis / Low suppression (e.g., Central Harlem, Crotona)
- **Cluster 2** – Moderate diagnosis / High death rates
- **Cluster 3** – Low diagnosis / High suppression (e.g., Bay Ridge, Stuyvesant Town)

These clusters revealed clear geographic disparities and provide a framework for targeting public health resources, community outreach, and educational programs.

Cluster	Avg HIV Diagnosis Rate	Avg Viral Suppression Rate	Avg Death Rate	Top Regions
Cluster 1	38.7	14.8	8.4	Chelsea – Clinton Central Harlem - Morningside Heights Bedford Stuyvesant- Crown Heights
Cluster 2	13.9	16	9.9	Downtown - Heights - Park Slope Canarsie - Flatlands Port Richmond
Cluster 3	18.9	16.4	5.3	Gramercy Park - Murray Hill Long Island City - Astoria Union Square - Lower East Side



4.4 Principal Component Analysis (PCA)

Purpose:

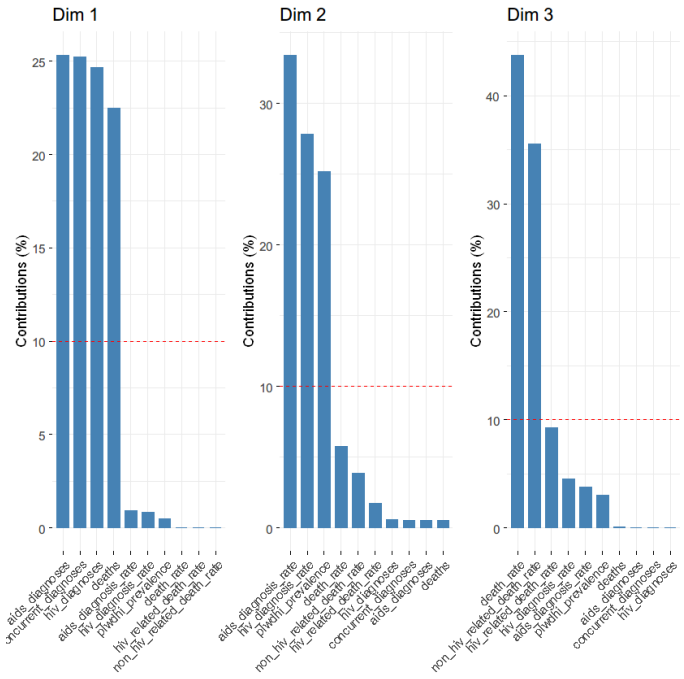
To reduce multicollinearity and simplify the modeling process, PCA was applied to 10 numeric public health indicators.

Data Interaction:

- **Input Variables**: Ten quantitative features including diagnosis counts, prevalence rates, and mortality indicators.
- **Processing**: Variables were scaled; PCA was applied to extract orthogonal principal components.

Why Useful:
PCA highlighted the major dimensions of variation in the dataset and reduced redundancy across correlated features.

Result:
The top three PCs explained 76.5% of total variance.
PC1: Diagnosis/concurrent case burden
PC2: Diagnosis rates and prevalence
PC3: Mortality indicators



To assess the predictive capability of PCA-derived factors on viral suppression outcomes, a linear regression model was constructed using the three principal components (PC1, PC2, and PC3) as explanatory variables.

```
Call:
lm(formula = x_viral_suppression ~ ., data = trainComponents)

Residuals:
   Min     1Q Median     3Q    Max
-92.78 -13.77 -11.66 -11.01  88.12

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.76363    0.18589  79.421  < 2e-16 ***
Dim.1        1.87165    0.09475  19.754  < 2e-16 ***
Dim.2        2.11194    0.13310  15.868  < 2e-16 ***
Dim.3        0.66029    0.13650   4.837 1.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.72 on 25558 degrees of freedom
Multiple R-squared:  0.02537,   Adjusted R-squared:  0.02526
F-statistic: 221.8 on 3 and 25558 DF,  p-value: < 2.2e-16
```

The resulting model shows a statistically significant relationship between these latent dimensions and viral suppression (all p-values < 0.05). However, its overall explanatory power remained limited. This suggests that while each component—capturing aspects such as disease burden, prevalence, and mortality—contributes meaningfully, they do not fully account for the variation in viral suppression across UHF regions.

Several factors may explain this limited explanatory capacity, including the absence of individual-level data, the exclusion of key social determinants of health (e.g., income, healthcare accessibility), and the inherent complexity of treatment adherence and engagement in HIV care.

Despite these limitations, using PCA still plays an important role. It effectively combines highly correlated public health variables and extracts key dimensions that reflect disease burden, prevalence, and mortality risk. For example, it can support the identification of high-risk areas or hidden relationships between variables, helping to improve the overall understanding of disparities in HIV treatment.

5. Summary of Findings

This study explored HIV-related public health disparities across New York City by applying statistical modeling, machine learning, and unsupervised learning techniques to Department of Health and Mental Hygiene (DOHMH) surveillance data. The key findings are summarized as follows:

- **Logistic regression** identified viral suppression rate and specific UHF (United Hospital Fund) regions as significant predictors of care linkage within 3 months. Neighborhoods such as Greenpoint, Williamsburg–Bushwick, and Sunset Park were consistently associated with higher probabilities of early linkage to care.

- **Random forest classifiers** outperformed logistic regression in predictive accuracy for both care linkage (AUC = 0.88) and viral suppression (accuracy = 85.9%). Feature importance rankings confirmed that geography (uhf), race, and HIV diagnoses are among the strongest predictors of treatment outcomes.
- **K-means clustering** grouped NYC's UHF regions into three distinct clusters based on diagnosis rates, suppression levels, and death rates. This segmentation exposed public health hotspots—most notably Central Harlem and Crotona—which were characterized by high disease burden and low suppression.
- **Principal Component Analysis (PCA)** reduced 10 correlated public health metrics into 3 uncorrelated components explaining 76.5% of the variance. These components were statistically significant in predicting viral suppression, enabling more efficient modeling and interpretation of latent health trends across NYC.

6. Conclusion & Recommendations

This study applied a comprehensive analytical framework to understand and predict disparities in HIV/AIDS outcomes across New York City, integrating machine learning, traditional statistical modeling, clustering, and dimensionality reduction. Using publicly available surveillance data from the NYC Department of Health and Mental Hygiene, we were able to identify key geographic and demographic factors that influence early linkage to care and viral suppression, revealing significant regional inequalities.

The logistic regression model offers good interpretability, identifying viral suppression rate and UHF region as significant predictors of care linkage. However, its predictive power is moderate, and sensitivity is low, indicating limitations in identifying high-risk populations.

In contrast, the random forest model performs better in both care linkage and viral suppression prediction tasks, with higher accuracy and stronger classification ability. It highlights UHF region, race, and number of HIV diagnoses as key variables, further emphasizing the critical role of geographic and racial disparities in healthcare access and treatment outcomes.

K-means clustering provided a spatial lens to our analysis. We found that UHF regions could be segmented into three clusters with distinct public health profiles—those with high diagnosis and low suppression, those with elevated death rates, and those with relatively better outcomes. These clusters support actionable targeting: regions like Central Harlem and Crotona, which consistently fell into the highest-risk cluster, can be prioritized for additional funding, outreach, or case management services.

Principal Component Analysis (PCA) further strengthened our conclusions by reducing multicollinearity among public health indicators and revealing latent dimensions that strongly

align with outcome variables. Disease burden (PC1), prevalence patterns (PC2), and mortality indicators (PC3) all significantly predicted viral suppression in a linear model, suggesting that these unobserved structures are valuable for both descriptive and predictive modeling.

Beyond modeling performance, our findings have practical implications for public health decision-making in NYC. They suggest that current disparities in HIV care are not evenly distributed but rather concentrated in specific demographic and geographic pockets. These insights can be used by city health officials, community organizations, and policymakers to design more tailored interventions, improve resource allocation, and close care gaps in the most underserved communities.

Based on this, we suggest the following policies for NYC public health decision-makers to consider.

1. **Prioritize Targeted Interventions in Cluster 1 UHF Regions**
   Areas like Central Harlem and Crotona consistently showed high HIV burden and low suppression. These regions should receive priority for outreach, education, and support programs.
2. **Address Regional Disparities in Care Linkage**
   UHF-level effects in both logistic and random forest models indicate significant geographic inequality. Tailored care navigation services should be expanded in underperforming neighborhoods.
3. **Integrate Suppression Monitoring into Care Linkage Strategies**
   Viral suppression was the strongest predictor of care linkage. Programs focused on improving ART adherence may indirectly enhance early care access as well.
4. **Using Predictive Models for Resource Planning**
   Random Forest models can help predict where care engagement may lag. Embedding these models in DOHMH planning workflows could improve allocation of limited resources.
5. **Extend Data Infrastructure to Include Spatial Autocorrelation Tools**
   Though not used in this study, Moran's I and other spatial statistics should be included in future DOHMH evaluations to detect clustering beyond descriptive methods.

Finally, this project illustrates the value of combining data science techniques with public health policy thinking. Our methodology is scalable and replicable across other urban datasets and disease areas, supporting more equitable and data-driven health strategies. Future research could extend this work by incorporating patient-level longitudinal data, applying spatial autocorrelation measures, or building real-time decision-support tools for public health agencies.

In conclusion, our integrated approach not only predicted HIV-related outcomes with high accuracy but also offered strategic insights into where and for whom public health interventions

are most urgently needed. As cities like New York strive to end the HIV epidemic, such analytic tools will be critical in guiding the next generation of targeted, equitable, and impactful care strategies.

Reference List

1.DOHMH HIV/AIDS Annual Report | NYC Open Data. (2023, June 1). https://data.cityofnewyork.us/Health/DOHMH-HIV-AIDS-Annual-Report/fju2-rdad/about_data

2.Infectious Disease Advisor. (2024, May 13). Machine Learning AI model accurately predicts HIV incidence in patients with STIs. https://www.infectiousdiseaseadvisor.com/news/machine-learning-artificial-intelligence-model-predicts-hiv-incidence/

3.Maskew, M., Sharpey-Schafer, K., De Voux, L., Crompton, T., Bor, J., Rennick, M., Chirowodza, A., Miot, J., Molefi, S., Onaga, C., Majuba, P., Sanne, I., & Pisa, P. (2022). Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-16062-0

4.Ramachandran, A., Kumar, A., Koenig, H., De Unanue, A., Sung, C., Walsh, J., Schneider, J., Ghani, R., & Ridgway, J. P. (2020). Predictive analytics for retention in care in an urban HIV clinic. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-62729-x

5.World Health Organization: WHO & World Health Organization: WHO. (2024, July 22). HIV and AIDS. https://www.who.int/news-room/fact-sheets/detail/hiv-aids