**Research Protocol**

**Title:**

Comparative Cluster Analysis for Identifying Risk Groups for Positive VIA Outcomes in HIV-Positive Women in Eswatini

**Principal Investigator:**

[Pako Mmileng, Francis Ngema]

**Background and Rationale**

Cervical cancer remains a significant health burden among HIV-positive women, with HIV increasing susceptibility to Human Papillomavirus (HPV) infections, a known precursor to cervical cancer. Visual Inspection with Acetic Acid (VIA) is an effective, low-cost cervical cancer screening tool, particularly in low-resource settings like Eswatini. Given the high rates of HIV among women in Eswatini, identifying risk groups for positive VIA outcomes is essential to provide targeted interventions and optimise screening strategies.

This study proposes a novel approach: leveraging unsupervised machine learning techniques, such as K-means, hierarchical clustering, and DBSCAN, to identify and interpret clusters of HIV-positive women based on demographic, clinical, and behavioral characteristics. These clusters may reveal underlying risk groups, providing a basis for targeted interventions to reduce cervical cancer incidence.

**Objectives**

**Primary Objective**:

* To identify clusters of HIV-positive women in Eswatini with distinct characteristics associated with positive VIA outcomes.

**Secondary Objectives**:

1. To compare the efficacy of K-means, hierarchical clustering, and DBSCAN in identifying clinically meaningful clusters related to positive VIA outcomes.
2. To characterize each cluster to identify significant demographic, clinical, or behavioural factors associated with increased risk for positive VIA outcomes.

**Study Design**

This is a retrospective, observational, data-driven study using unsupervised learning techniques to identify and interpret clusters among HIV-positive women in Eswatini who have undergone VIA screening.

**Methods**

**Study Population**

* **Inclusion Criteria**:
  + HIV-positive women, aged 18 years and above, who have undergone VIA screening at participating health facilities in Eswatini.
  + Available demographic, clinical, and behavioural data relevant to cervical cancer risk.
* **Exclusion Criteria**:
  + Women with incomplete records or missing data for key variables related to VIA screening.

**Data Collection**

Data will be collected retrospectively from medical records at designated health facilities. The variables considered include:

* **Demographic Variables**: Age, marital status, residence type (urban/rural), education level.
* **Clinical Variables**: WHO clinical stage at ART initiation, ART regimen, duration on ART, history of opportunistic infections.
* **Behavioural Variables**: Smoking status, sexual history, contraceptive use, and any previous history of cervical abnormalities or cancer screening.

**Outcome Variable**

* **Primary Outcome**: Positive VIA outcome (binary: positive/negative).

**Data Analysis**

**Data Preprocessing**

1. **Handling Missing Data**: Missing values will be handled using imputation techniques based on the nature and distribution of the missing data.
2. **Normalization**: Numerical variables will be standardized to ensure consistent scaling across variables.
3. **Encoding**: Categorical variables will be encoded appropriately, for instance, with one-hot encoding or dummy variables as needed.

**Cluster Analysis**

Three unsupervised clustering techniques will be applied independently to identify clusters:

1. **K-means Clustering**:
   * K-means will be performed with varying numbers of clusters (k) to determine the optimal number of clusters based on the elbow method and silhouette score.
2. **Hierarchical Clustering**:
   * Hierarchical clustering will use agglomerative methods with different linkage criteria (e.g., complete, average). The number of clusters will be determined based on the dendrogram and silhouette score.
3. **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**:
   * DBSCAN will be used to identify clusters based on the density of data points, with epsilon and minimum samples tuned to optimize cluster formation and noise reduction.

**Comparative Analysis**

* **Cluster Validation**: Clusters identified by each method will be evaluated and validated using silhouette scores, Davies-Bouldin index, and other cluster validation metrics.
* **Comparative Efficacy**: The clustering techniques (K-means, hierarchical, DBSCAN) will be compared based on interpretability, stability, and ability to identify meaningful subgroups associated with VIA outcomes.

**Cluster Interpretation and Characterization**

For each clustering method:

* **Descriptive Statistics**: Analyse and report descriptive statistics for each cluster regarding demographic, clinical, and behavioural variables.
* **Correlation with Positive VIA Outcomes**: Each cluster will be evaluated for its association with positive VIA outcomes to identify specific clusters with higher relative risks.
* **Risk Group Profiles**: Develop profiles for clusters with significant associations with positive VIA outcomes, highlighting key characteristics (e.g., younger age, advanced clinical stage, rural residence) that may inform targeted screening and intervention efforts.

**Ethical Considerations**

* **Confidentiality**: Data will be de-identified to ensure patient confidentiality. All analysis will be conducted on anonymized data in compliance with data protection laws and ethical guidelines.
* **Data Access and Security**: Access to patient data will be restricted to the research team. Data will be securely stored on encrypted servers.
* **Informed Consent**: As this is a retrospective study utilizing anonymized data, a waiver of informed consent will be requested from the Institutional Review Board (IRB).

**Limitations**

* **Retrospective Design**: As the study uses retrospective data, there may be limitations in the availability or completeness of certain variables.
* **Generalizability**: Findings may be specific to the population studied and may require validation in other settings to confirm applicability.
* **Cluster Interpretation**: Unsupervised clustering may lead to clusters that are difficult to interpret clinically, requiring careful analysis to ensure findings are actionable.

**Expected Outcomes**

1. Identification of distinct clusters within the population of HIV-positive women that correlate with positive VIA outcomes.
2. Comparison of clustering methods to determine the most effective approach for identifying clinically meaningful risk groups.
3. A set of recommendations for targeted interventions based on cluster characteristics, aiming to improve cervical cancer screening and prevention efforts in HIV-positive women.