**THE DATA COLLECTION AT NACSCOP**

**PURPOSE OF THE STUDY**

The World Health Organization has set some standards on the data collected regarding HIV tests. 0ver the past few years they have changed the standards which makes it hard to pull records regarding the matter. This makes the data to be categorized as recorded using two different data tools I.e. the new data tools and the old data tools.

In the new data tools one will find that the age difference is narrower than the age difference in the old data tool. There is also the gender issues where we find that in the new data tool we don’t have gender in the 1-9 age difference.

Our main aim for the study is to make sure that the data is pulled using the most suitable data tool avoiding the bias-ness of any data tool.

**JUSTIFICATION OF THE STUDY**

The purpose of this data collection is to facilitate work for those who use it effectively and comprehend it thoroughly without complicating or overthinking it. In light of the task we were given, we determined as a group that Python is the best language to use in data science because the libraries are better suited for such algorithms. It includes models that make manipulating data very simple.

The study's design makes use of the provided datasets and streamlines them by combining the new data with the older versions, just in case someone wants to compare changes between a certain time period covered by the old data tool and the new data tool.

Data is collected from the outdated versions to the more up-to-date and accurate new data, where the files are provided by the HealthIT, according to the data collection recorded from Bungoma county.

In this study, data integration is necessary, as it a combination of both data tools during the time span of the transitioning.

**MAKING HYPOTHESIS**

Do the new data tool bring the best results ?

The new tool is quite detailed, straightforward, and easy to read, so it produces the best results, according to the datasets provided by HealthIT.

If the new data is integrated then the best results will be the acquired

The more the data is integrated ,it will bring a positive effect in the performance in the HealthIT

**DEFINITION OF TERMS**

**Data Pivoting**

Enables you in rearranging the columns and rows in report so you can view data from different perspective.

**Data Models**

The process of creating a visual representation of an information system as a whole or parts of it to communicate connections between data points and structures.

**Data cleaning**

Is dataset's incorrect, corrupted, improperly formatted, duplicate, or incomplete data being fixed or removed

**BRIEF OVERVIEW OF THE STUDY (ABSTRACT)**

NACSCOP's data collection tools are in transition. Two sets of data collection and reporting tools will be used during this transition. Selecting the metrics to use requires an objective approach that excludes human biases, including the characteristics of other relevant institutions.

**DESCRIPTION OF THE RESEARCH DESIGN DESCRIPTION OF THE SAMPLE**

The aim of this project is based on the quantitative approach which helps in dealing with the numerical data and predictions which are the main agendas in the project

+++++++ serves all the HIV affected people using the probability sampling for it allows you to make strong statistical inferences about the population of the infected ones.

As we all know the aim is to create models that advices on the indicators whether new or old by using the secondary data and quantitative methods for we want to test the effectiveness of the system and we can use this method to measure the outcomes.

Use of questionnaires or forms is the most appropriate for consistent data is collected from many people because the questionnaires offers closed questions with limited options and therefore more common in quantitative research

**DESCRIPTION OF THE INSTRUMENTS USED (Including scoring procedures )**

**Analyzing and pivoting data**

The datasets were where analyzed and pivoted for ease evaluation, reading and understanding.

**Data cleaning**

The data had some empty cells and after pivoting the cells were labeled empty. We used python pandas function to replace it to 0.

**Filtering**

We created a criteria that would help in searching through the organizations names and if there was a typo, it would not prevent there being a result

**TESTING AND IMPLEMENTATION**

we were able to give the program a file and was able to create tables that held the organizations that used the old data tool only, the new data tool only and the ones which had collected the data with both of the data tools, we created a table that merged both the new data tool and old one.

This meant that there were some columns that were being discarded, but this was to avoid being bias and also to avoid data generation.

**SUMMARY AND CONCLUSION**

Data collection in this times are very important to keep track of people’s day – day activities. When it comes to health, the same should be considered to make sure people lead a better life.

We got to realize that even though the new data tool was very data specific(which is very recommended for decision making), the old one was not as specific as the new one but the challenge of combining both is cumbersome.

The new data tool should be used to collect data for it is more precise.