https://dspace.org/

Record linkage of national health laboratory service (NHLS) HIV datasets to cancer registry datasets using supervised learning techniques

Olago, Victor

2019
https://hdl.handle.net/10539/29782

Downloaded from a DSpace repository. https://dspace.org/

Record Linkage of National Health Laboratory Service (NHLS) HIV Datasets to Cancer Registry Datasets using Supervised Learning Techniques

VICTOR OLAGO

STUDENT NUMBER: 1739470

A Research Report submitted to the Faculty of Health Sciences in partial fulfilment of the requirements for the degree of *Master of Science (MSc)* in Epidemiology - Research Data Management

August, 2019

Victor Olago . *2019*
Record Linkage of National Health Laboratory Service (NHLS) HIV Datasets to Cancer Registry Datasets using Supervised Learning Techniques.

Copyright © University of the Witwatersrand, Johannesburg, South Africa.

All rights reserved. No part of this research report may be stored in a retrieval system, transmitted, or reproduced, in any form or by any means, including but not limited to photocopy, photograph, magnetic or other record, without prior agreement and written permission of the copyright holder.

SUPERVISORS:
Dr. Gideon Nimako, University of The Witwatersrand
Dr. Mazvita Sengayi, National Cancer Registry

SUPPORTED BY:
This project was supported by the National Institutes of Health (NIH) administrative supplement to Existing NIH Grants and Cooperative Agreements (Parent Admin Supplement) (The South African HIV/AIDS Match Study (SAM); U01AI069924 – 09, Principal Investigator (PI) Matthias Egger, co- PI Julia Bohlius) President's Emergency Plan for AIDS Relief (PEPFAR) supplement (PI Matthias Egger) and the Swiss National Science Foundation (SNSF) (The SAM, 320030_169967, PI Julia Bohlius). The National Cancer Registry (NCR) provided office space and technical support and supervision for the study.

DECLARATION

I hereby declare that this research report is project work carried out by me under the guidance of Dr. Gideon Nimako and Dr. Mazvita Sengayi. I have taken care in all respects to honour the intellectual property rights of others and have acknowledged the contribution of others. I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Parktown, Johannesburg, August, 2019

Victor Olago

DEDICATION

For Nicholas and Mary, its more than a decade but somehow I trust you are both happy with me.

ABSTRACT

INTRODUCTION

The National Health Laboratory Service (NHLS) is a national network of public health laboratories that serves more 80% of the South African population. All the laboratories are connected to a single data repository called the corporate data warehouse. The South African NCR is a pathology based registry housed within the NHLS. The NCR collates and analyses cancers diagnosed in pathology laboratories nationwide. These two data repositories present the opportunity to link HIV and cancer data to improve cancer surveillance among HIV positive people. Such linkage studies have made major contributions in understanding the epidemiology of HIV related cancers in developed countries. Although probabilistic methods have been used to link HIV and cancer datasets in South Africa before, it is computationally intensive and not scalable at national level. Supervised machine learning has been shown to be scalable and efficient in linking records accurately. In this work, our aim was to use Support Vector Machine (SVM) algorithms to link national HIV laboratory data to NCR data for the period 2004 to 2014.

METHODS

We used Cluster of Differentiation 4 (CD_4) counts, DeoxyriboNucleic Acid - Polymerase Chain Reaction (DNA-PCR) and Enzyme-Linked ImmunoSorbent Assay (ELISA) tests for the HIV data and laboratory confirmed cancers in the NCR. We linked the two datasets using names, surname, gender and date of birth since there was no common unique identifier. We used Python 3.6 running on Spyder terminal for the linkage. The linkage process involved data pre-processing, deterministic de-duplication, chunking, probabilistic de-duplication, blocking, pairwise comparison then records pair classification using SVM. After the linkage we performed high dimensional clustering using Gaussian Mixture Model (GMM).

RESULTS

NHLS HIV dataset had 39,249,147 HIV test records while the NCR cancer dataset had 664,869 laboratory confirmed cancers for the period 2004 to 2014. The de-duplication of the HIV dataset resulted to 15,157,685 HIV positive patients, 3,696,121 HIV negative patients and 41,147 patients had no valid HIV results. The matched dataset resulted in 309,741 linked records. A total of 231,945(74.88%) records had an HIV positive result compared to 69,648(22.49%) and 8,148(2.63%) records with HIV negative and no valid HIV result respectively. The matched dataset had 212,993(68.76%) and 96,718(31.23%) females and males respectively. The distribution of the race was 78.45%, 9.42%, 9.19% and 0.85% for Blacks, Whites, Coloured and Asians respectively. The age at the time of cancer diagnosis was 10 years younger for HIV positive compared to HIV negative cancer patients. The proportion with AIDS-defining cancers was 50.67% compared to 49.33% non-AIDS defining cancers. The precision, recall and F-measure for the linkage were 0.883,

0.997 and 0.937 respectively based on the records with national identification (ID) numbers as the ground truth.

CONCLUSION

Our study demonstrated that SVM algorithms are an effective way of linking large datasets in the absence of unique identifiers. Such techniques enable the linkage of disease registries in developing countries with accuracy. This methodology provides opportunities for enriching HIV cohort data with routinely collected laboratory and treatment data of other co-morbidities to inform public health actions.

ACKNOWLEDGMENTS

I wish to acknowledge the following people and organisations:

- I would like to express my utmost gratitude to Dr. Gideon Nimako and Dr. Mazvita Sengayi for their guidance, supervision, and for providing necessary information regarding the project as well as for their support in completing the project.
- To the South African NCR, thank you for providing space, datasets, computer and technical support during the entire research period.
- Sincere thanks to my family for their support and patience throughout the study period.
- I would like to express my gratitude to University of Witwatersrand - School of Public Health (SPH) for paying my tuition fees for the first academic year and Institute of Social Preventive Medicine (ISPM), NIH and SNSF for providing financial support through a stipend which enabled me to undergo this master's programme at University of the Witwatersrand, Johannesburg, South Africa.
- To other students at the NCR, it was nice having you around as we would always have short meetings to think through ideas related to the research work.
- My sincere gratitude to Dr. Timothy and Dr. Denielle, you both have always been part of my life.
- To Mr. Goldstine, you were the brother I never had.

DEFINITION OF TERMS

In this report the following terms will be defined as follows;

- Confidence Interval (CI). This is the probability that the population parameter lies within the interval, in our case we used 95% CI.
- Links. These are record pairs assessed as referring to the same person.*
- Matches. These are record pairs that actually refer to the same person.*
- True Positive (TP) link. These are records which refer to the same entity and have been correctly linked.*
- False Positive (FP) link. These are records that have been linked, but do not belong together (also known as false links).*
- True Negative (TN) link. These are records are which do not have a match and are correctly classified as a non-link.*
- False Negative (FN) link. These are records that have not been linked, but do belong together (also known as missed links or missed matches).*
- Accuracy rate. The proportion of all record pair comparisons that are true positive links or true negative links. The denominator for this rate is the number of all record pair comparisons, while the numerator is the number of record pairs that are correctly classified as true matches or false matches.*
- False-negative rate. The proportion of all record pairs belonging to the same individuals or entities that are incorrectly assigned as non-links.*
- False-positive rate. The proportion of all record pairs belonging to two different individuals or entities that are incorrectly assigned as links.*
- Precision (Link accuracy). The proportion of all classified links that are true links as opposed to classified links that are false links. It is calculated by dividing the number of links that are ascertained as true, by the total number of classified links.*
- Sensitivity (match rate). The proportion of all records in a file or database with a match in another file that were correctly accepted as a link.*
- Specificity or true-negative rate. The proportion of all records on one file or database that have no match in the other file that were correctly not accepted as a link.*

*Sourced from Australian Government National Statistical Service. Data linking, Information sheet five.

CONTENTS

DECLARATION iii
DEDICATION iv
ABSTRACT v
ACKNOWLEDGMENTS vii
DEFINITION OF TERMS viii
ACRONYMS xi
List of Figures xiii
List of Tables xiii
1 INTRODUCTION 1
1.1 Background . . . 1
1.2 Problem statement . . . 2
1.3 Justification . . . 3
1.4 Motivation . . . 3
1.5 Contributions . . . 3
1.6 Outline of research report . . . 4
2 BACKGROUND AND RELATED WORKS 5
3 USING SUPERVISED MACHINE LEARNING TO LINK NCR AND NHLS HIV DATASET 8
3.1 Introduction . . . 8
3.2 Study site . . . 9
3.3 Study population and data sources . . . 10
3.4 Data pre-processing . . . 11
3.5 Deterministic de-duplication . . . 14
3.6 Chunking . . . 15
3.7 Probabilistic de-duplication . . . 16
3.8 Blocking . . . 17
3.9 Record field comparison . . . 18
3.10 Support vector machine learning . . . 18
3.11 Creating training set and test set . . . 20
3.12 Evaluation of classification . . . 21
3.13 High dimensional clustering . . . 22
3.14 Ethical consideration . . . 24
4 RESULTS 25
4.1 Introduction . . . 25
4.2 Linkage results . . . 25
4.3 Results for high dimensional clustering . . . 36
4.4 Evaluation of classification results . . . 39
5 DISCUSSION 41
5.1 Introduction . . . 41
5.2 Characteristics of the linked dataset . . . 41
5.3 Evaluation measures . . . 44
5.4 Limitations . . . 44

5.5 Recommendations 44
6 CONCLUSIONS AND FUTURE DIRECTIONS 46

BIBLIOGRAPHY 48
A ETHICS CLEARANCE CERTIFICATE FOR THE STUDY 53
B ETHICS CLEARANCE CERTIFICATE FOR THE PARENT STUDY 54
C LETTER FROM THE GATEKEEPER 55
D LETTER FROM THE GATEKEEPER FOR THE PARENT STUDY 56
E ANTI-PLAGIARISM DECLARATION 57
F ORIGINALITY REPORT 58

ACRONYMS

ADC	AIDS-Defining Cancer
ART	Anti-retroviral Treatment
BCC	Basal Cell Carcinoma
CD_4	Cluster of Differentiation 4
CDW	Corporate Data Warehouse
CI	Confidence Interval
DNA-PCR	DeoxyriboNucleic Acid - Polymerase Chain Reaction
EC	Eastern Cape
ELISA	Enzyme-Linked ImmunoSorbent Assay
EM	Expected-Maximum
FN	False Negative
FP	False Positive
FS	Free State
GAU	Gauteng
GMM	Gaussian Mixture Model
HDSS	Health and Demographic Surveillance System
HIV	Human Immunodeficiency Virus
ID	identification
ISPM	Institute of Social Preventive Medicine
JW	Jaro-Winkler
KS	Kaposi Sarcoma
KZN	Kwa Zulu-Natal
LIM	Limpopo
LIS	Laboratory Information System
MAP	Maximum A Posteriori
MPU	Mpumalanga
NADC	Non-AIDS Defining Cancer

NaN Not A Number

NC Northern Cape

NCR National Cancer Registry

NHL Non-Hodgkin Lymphoma

NHLS National Health Laboratory Service

NIH National Institutes of Health

NW North West

PCR Polymerase Chain Reaction

PEPFAR President's Emergency Plan for AIDS Relief

PI Principal Investigator

PLHIV People Living with HIV

PMTCT Prevention of Mother-to-Child Transmission

PRL Probabilistic Record Linkage

RAM Random Access Memory

ROC Receiver operating characteristic

SAM South African HIV/AIDS Match Study

SCC Squamous Cell Carcinoma

SNSF Swiss National Science Foundation

SVM Support Vector Machine

SPH School of Public Health

STATS SA Statistics South Africa

TN True Negative

TP True Matches

TP True Positive

VCT Voluntary Counselling and Testing

WC Western Cape

LIST OF FIGURES

Figure 1 General process of record linkage . . . 9
Figure 2 Splitting data into sections . . . 16
Figure 3 General process flow of pairwise comparison on each block in each chunk . . . 17
Figure 4 Age at cancer diagnosis . . . 26
Figure 5 The age distribution at cancer diagnosis for HIV positive versus HIV negative patients . . . 27
Figure 6 The correlation matrix for various variables in the linked dataset . . . 36
Figure 7 The relationship between various cancer types and age at cancer diagnosis in the linked dataset . . . 37
Figure 8 The relationship between various cancer types and age at cancer diagnosis for the HIV negative population . . . 38
Figure 9 The relationship between various cancer types and age at cancer diagnosis for the HIV positive population . . . 39
Figure 10 Sensitivity and specificity of gender classification . . . 40
Figure 11 Ethics clearance certificate for the study . . . 53
Figure 12 Ethics clearance certificate for the parent Study . . . 54
Figure 13 Letter from the gatekeeper . . . 55
Figure 14 Letter from the gatekeeper for the parent study . . . 56
Figure 15 Plagiarism declaration . . . 57
Figure 16 Originality report . . . 58

LIST OF TABLES

Table 1 Assigning row identifier NHLS HIV dataset . . . 12
Table 2 Assigning row identifier NCR cancer dataset . . . 13
Table 3 Group ID assigned (not real data) . . . 14
Table 4 One instance of group ID (not real data) . . . 15
Table 5 Record field comparison . . . 18
Table 6 Number of matches per province . . . 28
Table 7 Number of matches per age group and their HIV status . . . 29
Table 8 Top ten cancer types in the matched records . . . 29
Table 9 Top ten cancer types for HIV positive population . . . 30
Table 10 Number of matches per race stratified by HIV status . . . 31
Table 11 Top five cancer types stratified by race and gender in the linked dataset . . . 32
Table 12 Top five cancer types stratified by race and HIV status in the linked dataset . . . 34
Table 13 HIV treatment province versus cancer diagnosis province . . . 35

1

INTRODUCTION

This chapter begins with the background information on HIV/AIDS cancer match studies. It also highlights the gaps presented in the literature, the problem statement, justification, motivation, research aims and then presents literature on how record linkages have been used in the public health domain in the past.

1.1 BACKGROUND

The SAM study aims to link cancer data to Human Immunodeficiency Virus (HIV) laboratory data in order to determine the incidence of cancers among HIV positive people. The HIV data includes ELISA tests, rapid HIV tests, HIV DNA-PCR results (in infants) or CD_4 cell counts. Such record linkage work (HIV and cancer registry match studies) has mainly been implemented in developed countries [16, 43], with low HIV prevalence as compared to sub-Saharan Africa. In Uganda, the HIV/AIDS cancer match study was conducted but this was in the pre- Anti-retroviral Treatment (ART) era [28]. The HIV/AIDS cancer match studies have made major contributions in understanding the epidemiology of HIV related cancers such as Kaposi Sarcoma (KS), Non-Hodgkin Lymphoma (NHL) and cervical cancer [20].

The South African NCR was established in 1986 and is South Africa's main cancer statistics source [10, 41]. It collates and analyses cancer cases diagnosed in pathology laboratories (both public and private) nationwide and reports annual cancer incidence rates stratified by sex, age and population groups [10, 55].The NCR is a department within the NHLS which is the largest diagnostic pathology service in South Africa. Its main man-

date is to support the national and provincial health departments in the delivery of healthcare.The NHLS provides laboratory and public health related services to over 80% of the South Africa population through a national network of laboratories. All the laboratories are connected to a single data repository called the Corporate Data Warehouse (CDW) for which the NHLS is the curator [35]. Data repositories like the CDW provide an opportunity for observational linkage studies that use routinely collected data [21] thus enabling the study of co-morbidities.The SAM is a data match study that enables the study of HIV and cancer epidemiology by using record linkage.

Although Probabilistic Record Linkage (PRL) techniques have been used in the SAM study to link records, the number of pairwise computations grows quadratically with the size of the input data thus making the scale-up of the process to large datasets a problem [2]. In order to improve yield and accuracy on large datasets, supervised machine learning techniques provide improved performance, hence its use in this work. Supervised machine learning algorithms use training data to determine a model that can be used to classify records with unknown matching status into *true matches*, *non-matches* and *possible matches* [2, 5]. The current study is a sub-study within the SAM study.

1.2 PROBLEM STATEMENT

There are large datasets that are routinely being collected in the public health sector that do not have unique identifiers. In order to use these datasets we need to refine methods of identifying records that belong to the same person within the datasets. To improve cancer surveillance among HIV positive people, we needed to link the HIV dataset to the cancer dataset efficiently. Currently, there are minimal efforts to link nationwide cancer and HIV data in South Africa. This work aimed to fill this gap by exploiting supervised machine learning techniques.

1.3 JUSTIFICATION

In order to study the burden, spectrum and incidence of cancer in HIV positive South Africans in the era of ART, it is important to carry out the SAM study. Although probabilistic methods have been used to link these datasets, it is computationally intensive and not scalable to nationwide datasets. In this work, supervised machine learning techniques which has improved accuracy and is scalable on big datasets was applied enabling us to link HIV and cancer records.

1.4 MOTIVATION

This study was motivated by a pilot study conducted by the NCR team that linked NHLS HIV data from Northern Cape province to NCR cancer data for the period between 2004 to 2013 [50]. The work demonstrated the feasibility of integrating HIV and cancer data using probabilistic record linkage approaches based on machine learning models using the Fellegi-Sunter framework [11]. In this work, we focused on the use of SVM approach to link the two datasets with the aim of improving accuracy of the linkage and reducing the computational cost.

1.5 CONTRIBUTIONS

This research work linked the NHLS HIV dataset to NCR cancer dataset using machine learning algorithms. We then clustered the linked dataset with the aim of studying the inherent characteristics produced by the linkage. The main results being reported here include:

1. The implementation of supervised learning models for linking NCR and NHLS HIV datasets for the period of 2004 to 2014; with the aim of improving the accuracy and computational cost of the linkage.

2. The implementation of high dimensional clustering algorithms to discover and describe inherent clusters in the matched dataset; with the aim of studying other factors associated with cancer and HIV co-morbidity

3. The validation of the linked dataset to establish the accuracy and performance of the linkage.

1.6 OUTLINE OF RESEARCH REPORT

The remainder of the report is organised as follows: Chapter 2 presents the literature on related works on record linkage studies and data integration. Chapter 3 describes the techniques and models used in this study. Chapter 4 presents the results of record linkage using SVM, and the results of the high dimensional clustering. Chapter 5 discusses the findings of the study and provides an overview of how the tools and models presented in the report can be integrated in other health information systems. Finally, Chapter 6 provides conclusions and future directions.

2

BACKGROUND AND RELATED WORKS

Record linkage, otherwise known as data integration or entity resolution is the process of finding matches of records or duplicates within or across files [1, 5, 31, 32, 38, 60]. This is based on a person's attributes such as names, dates of birth, gender, initials etc. In many instances, data coming from different sources may not have the same structure or format though the attributes will be similar. For example, one database might have gender as 0, 1 while another database might have gender as *male, female* or *m, f*. In such cases, the data needs to be cleaned and standardized [40].

Record matching can employ deterministic algorithm (exact matching) [14]. One can use a unique identifier or form a unique identifier from a combination of variables such as first name, last name, gender and date of birth. However, this might not be usable in the absence of a unique identifier field or when data has typographical errors. In such cases, the use of probabilistic algorithms [3] are ideal. Probabilistic record linkage involves pairwise comparison of records [46]. Then, the similarity weights are calculated in order to determine matched (records that belongs to the same pair) and unmatched (records that does not belong to the same pair).

There has been a significant amount of work on probabilistic and deterministic record linkage algorithms. Clark et al performed a comparison study on probabilistic and deterministic record linkage [8]. This study highlighted the importance of linking disease registries. They de-duplicated 7 registries using deterministic and probabilistic methodologies [8]. The number of individual records reduced slightly in probabilistic de-duplication as compared to deterministic de-duplication. The study recommended the use of probabilistic record linkage for health services research and epidemiology [8].

In the analysis of a probabilistic record linkage technique without human review [19], the researchers reported on the performance of both deterministic and probabilistic record linkage algorithms. Making reference to their previous article [18], the performance of deterministic linkage yield sensitivities approaching 90% and specificities approaching 100%. The results of deterministic linkage were poor compared to probabilistic linkage methods [19]. The performance of probabilistic methods were more than 95% for both sensitivity and specificity.

At Agincourt Health and Demographic Surveillance System (HDSS) in South Africa, probabilistic record linkage was used to assess the uptake of health services in resource limited settings [23]. By using finger prints deterministic matches as the gold standard for comparing the accuracy of the linked records, the team demonstrated the feasibility of using record linkage to identify patients in routinely collected health data [23]. Agincourt HDSS have also used probabilistic record linkage to link their mortality data to civil death registration data. A total of 87% of the records in the two datasets linked successfully [23]. An extension of this work compared death data from 2006 to 2009 through deterministic and probabilistic techniques. This was to relate the cause of death as stated by verbal autopsy reports against routine civil registration death certification cause of death which is written by a government pathologist. The study reported a matching rate of 61% [22].

In the state of Rio de Janeiro, Brazil, 43,825 tuberculosis records from 2009 to 2011 were linked using deterministic and probabilistic algorithms [37]. This study was used to ascertain the accuracy of probabilistic and deterministic record linkage as a technique for identifying records that belong to the same person. Sensitivity accuracy ranged from 87.2% to 95.2% while the specificity was from 99.8% to 99.9%. Dennis et al studied the importance of linkage algorithm in population based AIDS and cancer registries [12]. They applied probabilistic record linkage techniques in matching HIV and cancer datasets in order to study the risk of cancers among People Living with HIV (PLHIV).

In assessing the uptake of preventive measures and cancer surveillance [36, 47], records of the general population who participated in these preventive measures were linked to records of treatment uptake. The linkage was done using probabilistic record linkage. This study aided the evaluation of the effectiveness of such measures in the early detection of either breast, or cervical cancers cases.

At Pune cancer registry in India, 32,575 records of HIV positive patients seeking treatment at a government hospital were linked to 31,754 Pune cancer registry records for the period 1996 to 2008 using probabilistic record linkage [17]. The study demonstrated the feasibility of using record linkage to study cancer/other co-morbidities among HIV positive patients seeking treatment in India. It also gave insight on population based estimates of cancer risk in PLHIV in India. A recent study in South Africa used record linkage to correct under-ascertainment of cancers in HIV cohorts at the Sinikithemba HIV clinic in Kwa Zulu-Natal (KZN) [49]. Records of HIV positive individuals aged >= 16 years who were on treatment at the clinic were linked to NCR data using probabilistic record linkage. The study showed that record linkage is feasible and important for cancer ascertainment [49]. In a similar study, Bohlius et al performed probabilistic record linkage on HIV cohort data to four paediatric oncology units in South Africa in order to study the incidence of AIDS-Defining Cancer (ADC) and Non-AIDS Defining Cancer (NADC) in HIV positive children [4].

In this work however, supervised machine learning methods were used to link cancer data to national HIV laboratory data to improve the record linkage accuracy of the SAM study. High dimensional clustering was also implemented on the linked records to discover and describe characteristics of the inherent clusters.

3

USING SUPERVISED MACHINE LEARNING TO LINK NCR AND NHLS HIV DATASET

3.1 INTRODUCTION

In this chapter, the details of the methods and tools for linking the NCR and the NHLS HIV dataset are presented. The chapter discusses the study population, data sources, ethical consideration, data pre-processing, deterministic linkage, chunking, probabilistic linkage as a way of data de-duplication, SVM classification, validation of the linked dataset as well as high dimensional clustering.

Record linkage using supervised machine learning makes use of training data to classify and match the dataset. The training data is used to discover a predictive relationship in a given dataset (in some occasions training data is termed as known truth) [61]. The record linkage process in this study involved the following steps;

1. Data pre-processing/ cleaning and standardization
2. Blocking
3. Record Comparison/ weighting
4. SVM classification
5. Evaluation of classifications
6. High dimensional clustering of the linked dataset using GMM

Figure 1 shows the flow of the record linkage process while each step has been discussed in the subsections that follow.

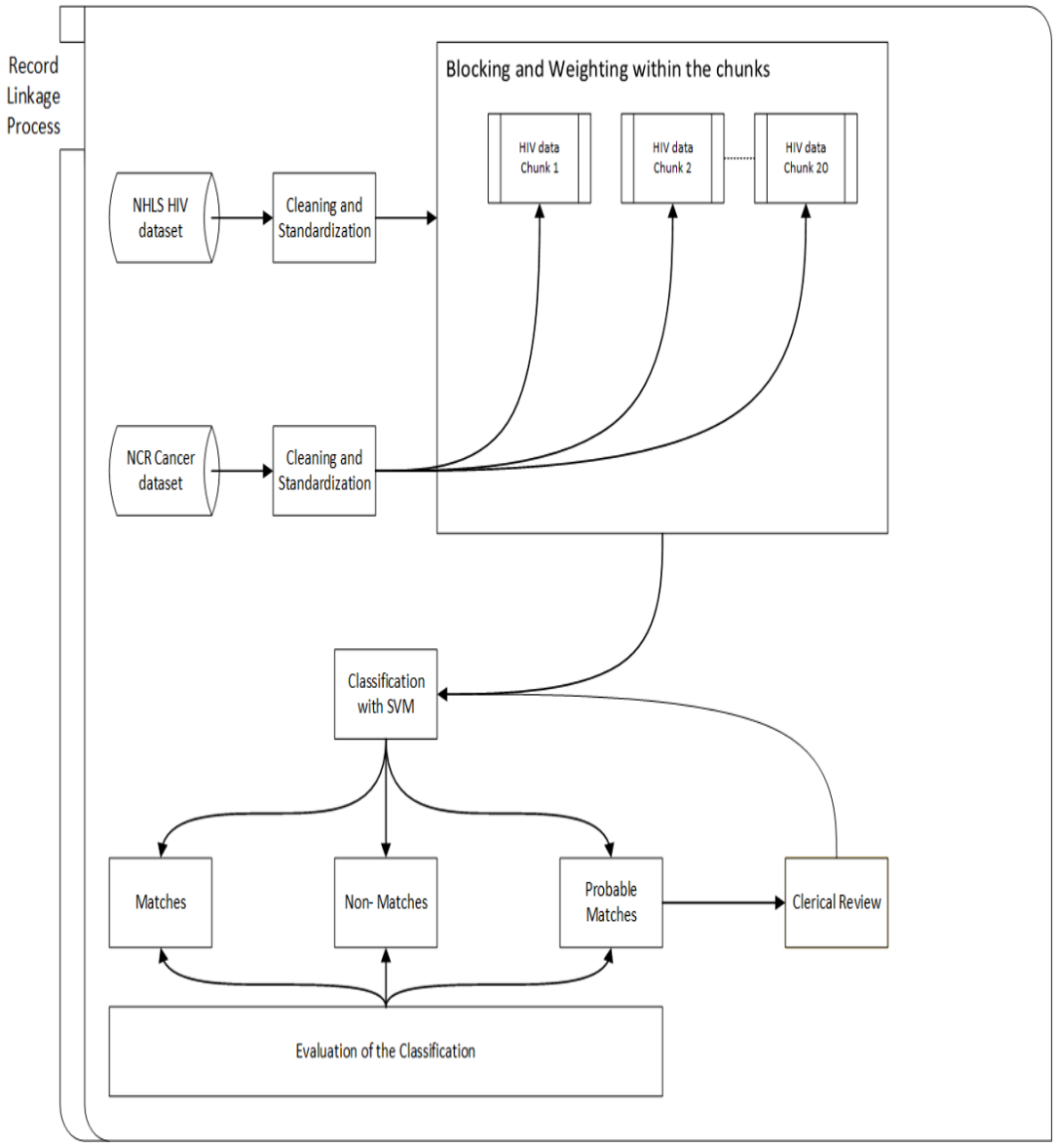


Figure 1: General process of record linkage

3.2 STUDY SITE

The work was conducted at the NCR which is a division within the NHLS. The NHLS is the largest diagnostic pathology service in South Africa [35]. It is responsible for supporting the national and provincial health departments in the delivery of healthcare by providing laboratory and related public health services [35]. It supports over 80% of the South African population through a national network of laboratories [35]. The NCR is the primary

cancer surveillance system and most extensive repository of cancer data in South Africa. It was established in 1986 as a voluntary, pathology-based cancer surveillance system, and continues to operate [10, 41, 55]. Its database contains over 1.2 million cancer records with approximately 80,000 cancer notifications yearly of which about 60,000 are incident cases. The NCR is located at the NHLS campus in Sandringham, Johannesburg South Africa.

3.3 STUDY POPULATION AND DATA SOURCES

The record linkage in this study involved HIV and cancer datasets. The cancer dataset included all individuals who received cancer diagnosis between 2004 to 2014 from either public or private hospitals reported or recorded in the NHLS's CDW. The HIV dataset included all individuals who were tested for HIV in public health laboratories in South Africa under the same time frame from all the health facilities that are supported by the NHLS around the country.

The CDW is a data warehouse that provides a central repository for data that is entered from all the NHLS laboratories. The data gets synchronised to the main server. The laboratories that feed data through the CDW use a web-based interface called *TrackCare* [51]. This was the source of the laboratory HIV data for the period of 2004 to 2014. It included HIV test results for the entire study period. The results included CD_4 count, HIV investigation, HIV P24 antigen, HIV serology, HIV-1 Polymerase Chain Reaction (PCR), HIV-1 western blot, HIV-1/2 antibody, HIV-1/2 rapid, and HIV sputum. The results of the tests were either positive, negative (especially for the screening tests), invalid or sometimes not applicable. The dataset had about 40 million records collated from all the public laboratories that used *TrackCare* Laboratory Information System (LIS) [51]. The second dataset from the NCR consisted of pathology based cancer records collated from all public and private hospitals in South Africa for the period 2004 to 2014. The dataset had 664,869 records. The variables common to both datasets used for record linkage were; first name, surname,

gender, day of birth, month of birth, year of birth and initial of the first name. Records with South African national ID numbers were used to validate our linkages. About 5% of the records had ID numbers.

3.4 DATA PRE-PROCESSING

The aim here was to prepare the data in order to maximize the matches in the records to be linked [7]. The process started by removing unwanted characters in selected field values. This involved removing punctuation marks that appeared in the names. All similar attributes in the two datasets were formatted uniformly; for example the attribute gender was coded as **M, F** for male and female respectively in both datasets.

The row identifier constituting a province identifier combined with a 9 digit auto-incremental number was created. Table 1 shows how the row identifier was created. After creating the row identifier both the row number and the province codes were dropped. The purpose of creating a row identifier as a dummy identifier was to aid the merge of the dataset after splitting in portions and tracing back merged records to the raw data.

Table 1: Assigning row identifier NHLS HIV dataset

Province Name	**Province Code**	**Row Number**	**Row identifier**
Northern Cape	111	000000365	111000000365
Eastern Cape	222	000065987	222000065987
Gauteng	333	000008574	333000008574
Kwa Zulu-Natal	444	000000905	444000000905
Western Cape	555	000321456	555000321456
Limpopo	666	000789654	666000789654
Free State	777	000654123	777000654123
Mpumalanga	888	000000963	888000000963
North West	999	000852147	999000852147

In the NHLS HIV dataset the variable *patient_name_lis* contained all the names of a patient in one column separated by comma. In order to get surname, initial and first-names, the *patient_name_lis* was split at the commas. The name fields with words like *anonymous, abandoned, baby of, mother of, father of, child to, child of, Voluntary Counselling and Testing (VCT), HIV, etc.* (which had varied spelling mistakes) were removed. Names that were just all numbers were dropped while names with partial numbers such as *124Victor,Olago673* had the digits dropped to give *Victor, Olago*. Name fields with numbers only were mostly from clinical trials that had names coded for purposes of confidentiality [30, 56, 57]. Names with special characters were transformed to normal alphabetical name; for instance *Victör* was changed to *Victor*. Names that are in short forms were expanded, in the South African context for example *VD Merwe* in the *patient_name_lis* field is a short form for *VanDerMerwe*. Names like *Van Vywk* were joined to form one name e.g. *VanVywk*. When a name has embedded blank spaces, this reduces the matching probability as the blank spaces are considered characters therefore, such spaces were truncated. Names with hyphens in the middle had the hyphens removed e.g. *Mary-Anne* would be *MaryAnne*.

System generated date of births in the NHLS HIV dataset for instance 1800:01:01 were all set to *null* or Not A Number (NaN). If the date of birth was more recent than the date of specimen collection, then the date of birth was also set to *null* or NaN. For records that had less than two years difference between date of birth and date of specimen collection and patient name contained the word *'BABY'* such records were dropped. Such records were assumed to belong to babies who had not yet been named at the time of specimen collection. Date of birth was split into; day, month and year.

The cleaning of the *patient_name_lis* was done using regular expressions [25] in Python, where specific patterns were analysed and extracted. This created a difference between the original dataset and the extracted data. The difference was used in the next extraction. This process was repeated for all the name patterns. After data preprocessing, NHLS HIV dataset had 39 million records.

The NCR cancer dataset, had its variables renamed to the format of NHLS HIV dataset to enable linkage of the two datasets. A row identifier for the NCR cancer dataset was created as well. Standard codes were created for the cancer dataset but this was not broken down into province level. Table 2 shows how the creation of row identifier in the NCR cancer dataset was computed.

Table 2: Assigning row identifier NCR cancer dataset

row	**NCR Code**	**Row Number**	**Row identifier**
row1	123	000000001	123000000001
row2	123	000000002	123000000002
row3	123	000000003	123000000003

3.5 DETERMINISTIC DE-DUPLICATION

The de-duplication step involved use of a unique record identifier (i.e. linkage key), to classify the records as either having agreement or disagreement [3, 5, 6, 34, 37, 46, 62]. A linkage key from first name, surname, gender and date of birth was created. Records that were exactly similar on the linkage key were assigned a group id as shown in the Table 3.

Table 3: Group ID assigned (Note: not real data)

First Name	**Surname**	**gender**	**date of birth**	**group id**
Victor	Olago	Male	01/01/1920	1
Victor	Olago	Male	01/01/1920	1
Peter	Nyaoke	Male	01/01/1921	2
Millicent	Achieng	Female	01/01/1924	3
Millicent	Achieng	Female	01/01/1924	3
Millicent	Achieng	Female	01/01/1924	3
Catherine	Sihoho	Female	01/01/1929	4
Catherine	Sihoho	Female	01/01/1929	4
Catherine	Sihoho	Female	01/01/1929	4
Catherine	Sihoho	Female	01/01/1929	4

The group ID variable acted as our unique identifier for records that had identical characteristics, that is first name, surname, gender and date of birth. Another table with only single instances of group id was created as shown in Table 4.

Table 4: One instance of group ID (Note: not real data)

First Name	Surname	gender	date of birth	group id
Victor	Olago	Male	01/01/1920	1
Peter	Nyaoke	Male	01/01/1921	2
Millicent	Achieng	Female	01/01/1924	3
Catherine	Sihoho	Female	01/01/1929	4

Deterministic de-duplication is very important, especially on big data as it eliminates unnecessary pairwise comparisons in the probabilistic de-duplication. This improves probabilistic de-duplication accuracy and reduces the computation time. However it is not possible to pick up typographical errors using deterministic de-duplication and that is the gap that probabilistic de-duplication addresses. Before probabilistic de-duplication we performed chunking as explained in the following section.

3.6 CHUNKING

Chunking is the process of dividing big-data into meaningful sections to enable ease of computational processing and algorithms [13]. Big data often has the challenge of memory management as it is hard to process such data without running out of Random Access Memory (RAM) [13]. The NHLS HIV dataset was divided into 20 sections while performing pairwise comparison for probabilistic de-duplication and 20 sections while performing pairwise comparison of the NCR cancer dataset with NHLS HIV dataset. This greatly improved the efficiency in terms of memory management and reduced the time involved in performing pairwise comparison. Pairwise comparison is the most resource intensive step in record linkage and data de-duplication. When a huge amount of data is called to the computer memory for processing, often times the computer runs out of space to perform computation. Only a small portion of RAM is allocated for computation while

the rest of the RAM is left to just hold the data thus slowing the data processing speed and increasing the processing time. This can halt the programme running or in some cases generate an error. Figure 2 is a diagrammatic representation of the chunking procedure.

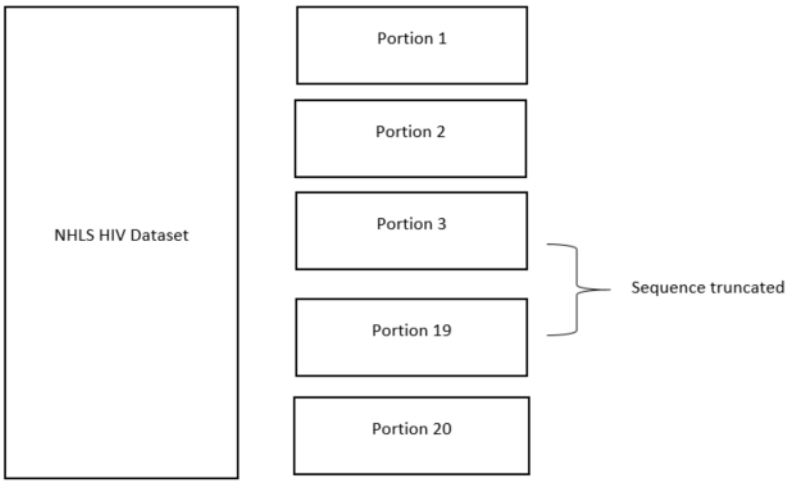


Figure 2: Splitting data into sections

3.7 PROBABILISTIC DE-DUPLICATION

One of the shortfalls of deterministic record linkage is the inability to detect typographical errors, for this reason it is not able to fully de-duplicate all records in a dataset. In order to fully de-duplicate the NHLS HIV dataset probabilistic de-duplication was used. The similarity weights were generated for surname, first name, gender, day of birth, month of birth and year of birth for the records that fell within each block [3, 5, 6, 34, 37, 46, 62]. String records comprising of first names and surnames were classified in appropriate groups using the weights generated. The similarity threshold for string comparison was set to 0.85.

3.8 BLOCKING

Blocking splits records into equally selective groups that have the probability of being linked (blocks). Linkage is done within each block [2]. Thus, records within a given block are matched; the linkage algorithm may still compare similarity across blocks. Blocking drastically reduces the number of record comparison and thus reduces the record linkage time. The two datasets were blocked on surname; records with surnames that were *phonetically* similar but not necessarily the same were put in the same block. Blocking has a weakness of creating large blocks that makes the linkage practically impossible [2, 62], thus the use of the chunking technique discussed in the previous section.

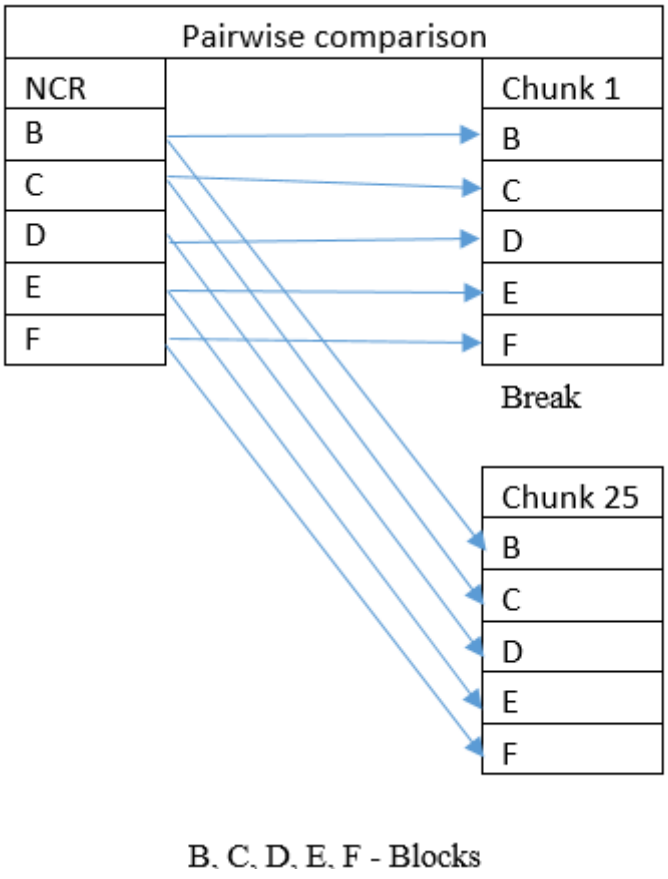


Figure 3: General process flow of pairwise comparison on each block in each chunk

3.9 RECORD FIELD COMPARISON

Records in each block in each chunk were compared in order to determine their similarity agreements as shown in Figure 3. This process is also known as weighting [11, 46]. Since different fields had different data types, different methods were used to generate linkage weights. For fields with string variables like surname and first name, the use of Jaro-Winkler (JW) string comparison metrics where the overall similarity threshold was set to 0.85 in order to cover for typographical errors was applied [59]. For gender, initial, day of birth, month of birth and year of birth the exact match comparison method was applied. In creating the datasets of record pairs, fields were considered to agree using the criteria given in Table 5.

Table 5: Record field comparison

NCR cancer Variable	**NHLS HIV Variables**	**Agreement Criteria**
Surname	Surname	JW-Score>= 0.85
Firstname	Firstname	JW-Score>= 0.85
Gender	Gender	Exact
Initial	Initial	Exact
Day of birth	Day of birth	Exact
Month of birth	Month of birth	Exact
Year of birth	Year of birth	Exact

3.10 SUPPORT VECTOR MACHINE LEARNING

This is a branch of machine learning that is used in mathematical and engineering problems like recognition of objects, face detections, voice recognition, handwriting digit recognition and target detection [33, 39]. Suppose one has points of set S $X_i \in R^N$, where

$i = 1, 2, ..., N$ and that every point X_i, fits in either two category labelled $Y_i \in \{-1, 1\}$, the aim is to put the equation of the hyperplane that divides S separating all the points of the similar category on one side. This classification is done by way of building an N-dimensional hyperplane that perfectly divides the data into two groups [33, 39].

Within the probable hyperplanes, choose the ones where the distance of the hyperplane from the closest data points is as large as possible, this distance is called the *margin*. The criteria are based on good training data, where the entire potential test vector is in a given radius r of the training vector. The training data will accurately separate all test data when the chosen hyperplane is at least r from any training vector. The further the hyperplane from any data, the larger r becomes [33, 39].

The desired hyperplane (i.e. maximum margin) is also the separator of the line between the nearest points on the convex hulls. Learning in this respect is getting the maximum margin separating the hyperplane between two points. Within each training point there are related coefficients [33, 39]. The final function for any given test points is given by the strength with which the points are incorporated into the final decision function for any given test points using kernel function K;

$$x_1, x_2 \in XK(x1, x2) = \langle \phi(x1) \cdot \phi(x2) \rangle \tag{1}$$

The dot product is used to calculate the match of two data points in the feature space. Since the kernel function defines the feature space in which the training set is defined, it is good to choose an appropriate kernel [33, 39]. It is the work of a kernel function to explain the feature space through which the training set examples are categorised by expressing the prior knowledge of the data being modelled. The classification of points in a feature space can be located by a support vector machine without characterizing the space. This is achieved simply by defining a kernel function in the feature space that plays

a role of the dot product [33, 39]. The prediction of the classification of X will be done using the formula;

$$f(X) = sign(\langle w \cdot \phi(X) \rangle - b) \tag{2}$$

The function returns either 1 or -1 depending on which class X belongs to, while $\langle w \cdot \phi(X) \rangle$ forms the dot product of vector w that forms the origin of the point $\phi(X)$ and b which is the shift of hyperplane from the original to system's co-ordinates [33, 39].

SVM performed classification on the weights generated for surname, first name, gender, initial, day of birth, month of birth, and year of birth. A randomised test sample of classification on the matches and non-matches generated by comparing data vectors on each variable was treated as the training data. The training data was then used to perform classification of the entire datasets [33, 39].

3.11 CREATING TRAINING SET AND TEST SET

Having completed the weighting process, a training data was created from a sample of the dataset that had been classified as match, non-match and probable match. The training was then used as labels for the classification of the weighted dataset.

The dataset had seven column vectors that is surname, first name, initial, gender, day of birth, month of birth and year of birth. A record was considered a match if the row sum was above six and both the first name and surname had a positive outcome score of 1. That means even if the row sum was 6 but the first name column had a 0, such records were dropped for record review. The record review compared if there were other names or there was a change in surname before the classification. The dataset was divided into two parts, 20% for the training set, then the training set used to classify the whole dataset. One of the advantages of the SVM classifier is that it only requires a small proportion of training data to achieve good results in-terms of precision, recall and F-measure [61].

During the process of developing the algorithm, we used clerical review to determine an appropriate threshold for the matching record pairs. The records with a column sum of 5 were clerically reviewed while those records that had a column sum of less than 5 were considered a non-match pairs.

3.12 EVALUATION OF CLASSIFICATION

Evaluation of classifications plays a very important role in record linkage. In order to perform evaluation, it is necessary to express the classification quality in terms of precision, recall and F-score [5, 24]. These are defined as follows;

1. Precision: the proportion of true matches from the record pairs given by a decision model

$$Precision = \frac{TM}{(TM + FM)} \tag{3}$$

2. Recall: the proportion of true matches that are correctly classified by a decision model

$$Recall = \frac{TM}{(TM + FN)} \tag{4}$$

3. F-measure: the harmonic mean of Precision and Recall, calculated as

$$F - measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{5}$$

The NHLS HIV dataset and NCR cancer dataset had a column for government issued national ID numbers. However, it is not a requirement to have a government issued ID number to access care and treatment in the Republic of South Africa. As a result only 5% of the records had ID numbers. The validation of the linkage quality was done with records that had government issued ID numbers. This aided in calculating precision, recall

and F-measure of the linked dataset. The merge of records with government issued ID numbers acted as the ground truth.

3.13 HIGH DIMENSIONAL CLUSTERING

Clustering is the act of dividing data into meaningful groups in order to summarise and understand the relationship within the data better [29, 42]. For example, in clustering records it is possible to discover trends and ascertain the natural layout of data [29, 42]. In this project clustering of the linked dataset to study the inherent characteristics of cancer and HIV co-morbidity in the matched dataset was done using high dimensional clustering through GMM. GMM otherwise known as kernel density estimation based clustering, is a weighted sum of Gaussian component densities with a parametric probability density function [29, 42, 52]. The clusters are formed based on the Gaussian distribution of the centres.

A well trained prior model is used to estimate parameters using recursive Expected-Maximum (EM) algorithm or Maximum A Posteriori (MAP) [42].

These are explained by the equations that follow;

$$p(X|\lambda) = \sum_{i=1}^{M} W_i g(X|\mu_i, \sum_i) \tag{6}$$

Where x is a D-dimensional continuous-valued data vector (features)

$$w_i, i = 1, \ldots M, \tag{7}$$

are the mixture weights, and

$$g(X|\mu_i, \sum_i), i = 1, \ldots M \tag{8}$$

are components Gaussian densities. With each component density is a D-variate Gaussian function;

$$g(X|\mu_i, \sum_i) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(X-\mu)\prime \sum_i{}^{-1}(X-\mu)\} \tag{9}$$

With mean vector

$$\mu_i \tag{10}$$

and covariance matrix

$$\sum_i \tag{11}$$

The mixture weights meet the constraint

$$\sum_{i=1}^{M} W_i = 1 \tag{12}$$

The complete model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. Collectively represented by the notation,

$$\lambda = \{w_i, \mu, \sum_i\} i = 1, \ldots, M \tag{13}$$

GMM's have the capabilities of representing large classes of sample distributions and forming smooth approximations to arbitrary shaped densities. This is intuitive in modelling some underlying set of hidden classes [29, 42, 52].

Binary variables as well as continuous variables were selected to study the high level distribution and relationships in the data. The variables correlation tested before performing the high dimensional clustering by computing the correlation matrix and then plotting a heatmap [58]. The first cluster was for the whole linked records then for records with HIV positive results and lastly records with HIV negative results. This presented the high level characteristics of the linked dataset which is the main function of high dimensional clustering.

3.14 ETHICAL CONSIDERATION

Although this work was a sub study within the SAM study, which has ethical approval from the University of Witwatersrand Human Research Ethics Committee (clearance certificate number M140602), ethical clearance was sought before commencing the research work. This clearance was granted by the University of Witwatersrand Human Research Ethics Committee (clearance certificate number M171176). The work involved the use of patient identifiable information, as such the dataset was at all times stored on the NHLS servers. The researcher was issued with an NHLS computer that is connected to the NHLS domain, and was also given login details which conform with the NHLS Information Technology security standards. The server could only be accessed via the NHLS accredited computers connected to the network of NHLS. The NHLS has firewalls to prevent unauthorised external connections to the server. After the linkage, all patient identifiable information were removed and a unique record identifier was assigned to all records. Only de-identified records were used for cluster analysis.

4

RESULTS

4.1 INTRODUCTION

This chapter presents the results of the machine learning linkage on NHLS HIV dataset to NCR cancer dataset as well as the de-duplication of the NHLS HIV dataset for the period 2004 to 2014. Descriptive statistics on data cleaning, deterministic de-duplication and probabilistic de-duplication will be presented first. This will be followed by the results of the linked NHLS HIV dataset and NCR cancer dataset as tables of results and graphical presentations.

4.2 LINKAGE RESULTS

The NCR cancer dataset had 664,869 distinct cancer cases collected from South African public and private laboratories for the period of 2004 to 2014. The NHLS HIV dataset had 41,035,009 records for the period of 2004 to 2014. After cleaning and standardization, the records were reduced to 39,249,147. This further reduced to 18,894,953 after probabilistic de-duplication.

The linkage of the de-duplicated NHLS HIV dataset to NCR cancer dataset resulted in 157,430 linked records. The linked dataset was merged dataset with HIV follow up visits using group id. This was to obtain all instances of the records that needed elimination to avoid unnecessary pairwise comparison and improve efficiency of computational intensity and time involved in the linkage. The merge produced 419,675 successful links. Post

processing on the linked and merged dataset was performed by removing records that merged but had more than two years gap in the date of birth. This resulted in 309,741 linked records.

Of the 309,741 records, 212,993 (68.76%) were females, 96,718 (31.23%) were males while 30 (0.01%) had unknown gender. The mean age at the time of cancer diagnosis was 43.07 years for the entire population with females having a mean age of 43.99 years and men 42.65 years. The proportions for HIV status in the linked records were 231,945 (74.88%), 69,648 (22.49%) and 8,148 (2.63%) for positive, negative and unknown respectively. The mean age of HIV negative cancer patients was 51.08 (CI 50.95,51.21) while that of HIV positive cancer patients was 40.67 (CI 40.62,40.71). It is important to note that there were 152,792 NADC's, with a mean age of 45.32 (CI 45.25,45.40) and 156,949 ADC's, with a mean age of 40.88 (CI 40.82,40.94).

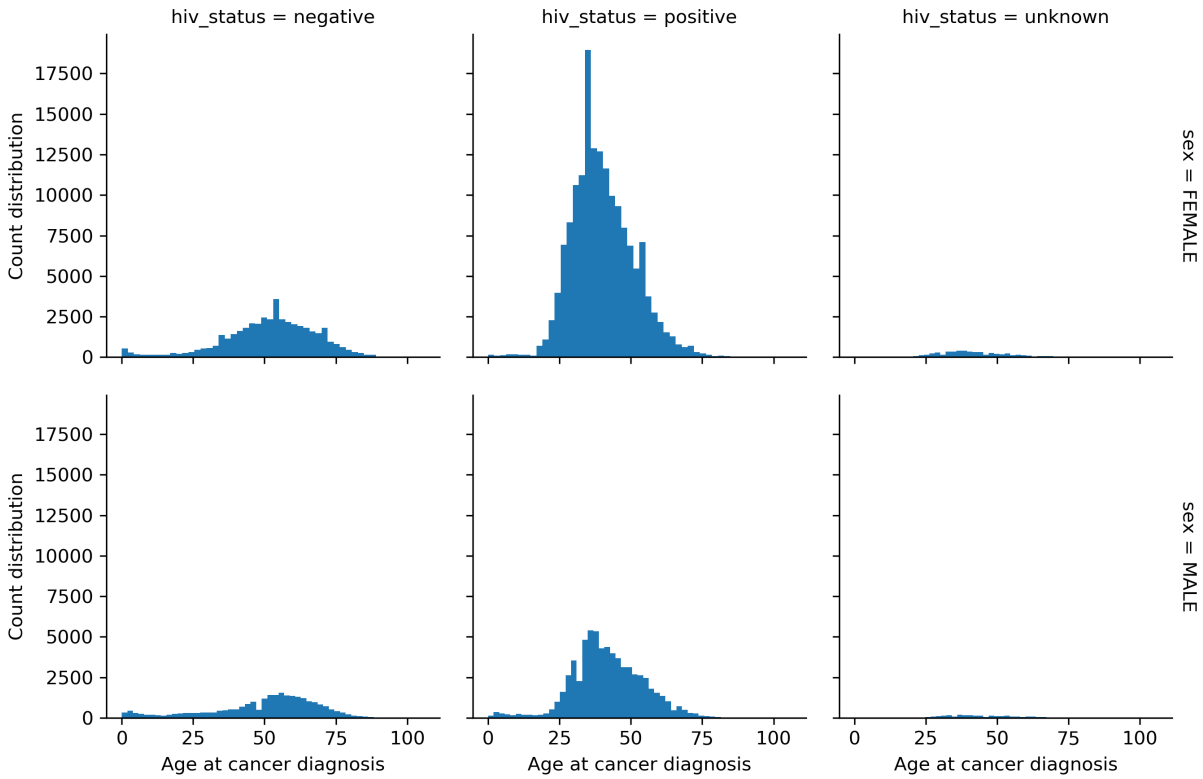


Figure 4: Age at cancer diagnosis

Figure 4 shows a histogram of age at cancer diagnosis stratified by gender and HIV status. From the histogram there is generally higher proportion of HIV positive cancer patients than HIV negative ones. Figure 5 provides the high level indication of the relationship between the HIV status and age at cancer diagnosis. There was a sharp increase for the number of HIV positive cancer patients at the age of 18 to 25 years then the number drops sharply. From the age of 0 to 5 years, a higher proportion of cancers were noted amongst HIV negative children compared to those who were HIV positive. Between the ages 20 to 45 years, more cancers were noted amongst HIV positive individuals compared to HIV negative ones.

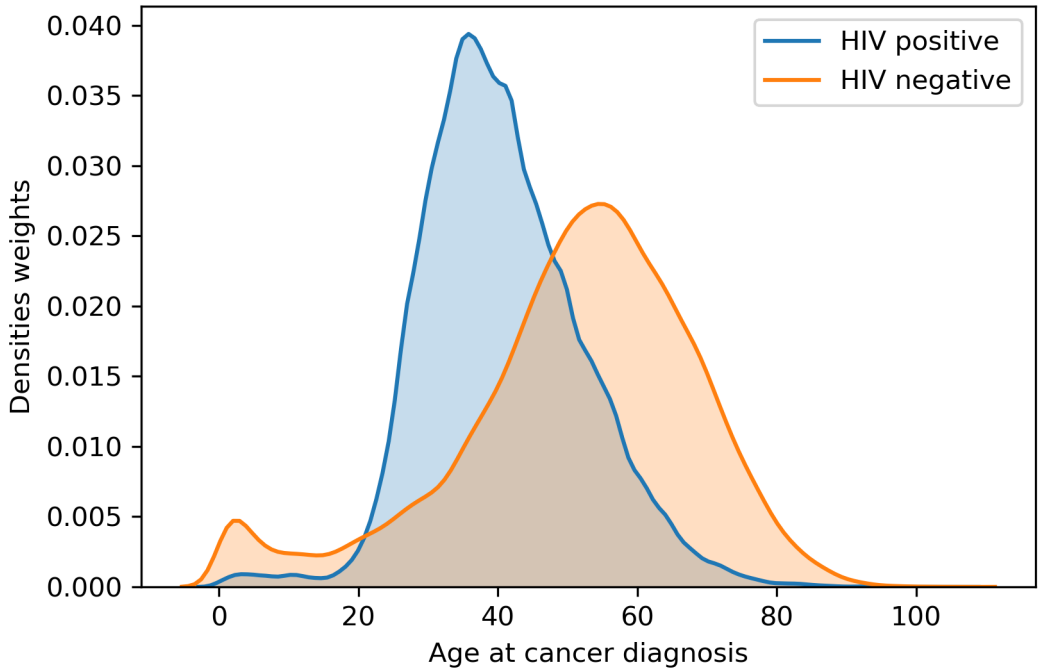


Figure 5: The age distribution at cancer diagnosis for HIV positive versus HIV negative patients

Table 6 presents the absolute numbers for the linkage for each province stratified by gender, listed in alphabetical order. Gauteng province had the highest linked records followed by Western Cape then Mpumalanga.

Table 6: Number of matches per province

Province Name	**Females**	**Males**	**Unknown**	**Total**
Eastern Cape	16,716	8,318	3	25,037
Free State	16,431	7,191	4	23,626
Gauteng	64,561	32,614	12	97,187
Kwa Zulu-Natal	17,072	5,456	0	22,528
Limpopo	11,995	2,291	1	14,287
Mpumalanga	27,141	8,175	0	35,316
North West	15,757	6,829	10	22,596
Northern Cape	5,442	3,279	0	8,721
Western Cape	37,878	22,565	0	60,443

Table 7 shows the absolute numbers of the matched records for various age groups stratified by HIV status. The age group 30 to 54 years had the highest number of linked records followed by age group 55 to 64 years then 15 to 29 years. This was slightly different for the HIV positive in which the age group 15 to 29 years was the second after the age group 30 to 54 years.

Table 7: Number of matches per age group and their HIV status

Age Group	**HIV Negative**	**HIV Positive**	**Unknown HIV Status**	**Total**
14 and below	3,493	2,654	94	6,241
15 to 29	4,200	29,681	1,145	35,026
30 to 54	29,624	170,997	6,456	207,077
55 to 64	17,313	21,107	1,061	39,481
65 and above	15,018	6,304	594	21,916

Table 8 shows the top ten cancer types for the linked dataset irrespective of the HIV status. Cervical cancer was the leading cancer type in the linked dataset followed by Kaposi Sarcoma then Breast.

Table 8: Top ten cancer types in the matched records

Cancer Type	**Total**
Cervix	80,482
Kaposi Sarcoma	56,702
Breast	29,624
Non Hodgkin lymphoma	19,765
Eye	11,816
SCC of skin	7,218
Lung	6,967
Hodgkin lymphoma	6,060
Leukaemia	5,790
Colorectal	5,688

Table 9 shows the top ten cancer types for the linked dataset for the HIV positive population. Cervical cancer is the leading cancer type in the linked dataset for the HIV positive population followed by Kaposi Sarcoma then Breast. Contrary to the spectrum in the overall linked dataset Vulva and Burkitt lymphoma are present in the top ten cancers for the records with HIV positive results.

Table 9: Top ten cancer types for HIV positive population

Cancer Type	**Total**
Cervix	64,304
Kaposi Sarcoma	54,028
Breast	18,133
Non Hodgkin lymphoma	16,509
Eye	11,037
SCC of skin	5,412
Vulva	4,334
Lung	3,255
Burkitt lymphoma	3,123
Colorectal	2,935

Table 10 shows the absolute numbers of cancers per race stratified by HIV status for the linked records, listed in alphabetical order. Black Africans had the highest numbers in the linked dataset followed by Whites, Coloured then Asians. The trend in the HIV negative strata is slightly different from the HIV positive strata as Coloured is second after Back Africans.

Table 10: Number of matches per race stratified by HIV status

Race	**HIV Negative**	**HIV Positive**	**Unknown HIV Status**	**Total**
Asian	1,218	1,347	74	2,639
Black	37,566	197,456	7,972	242,994
Coloured	13,603	14,343	508	28,454
Whites	15,338	13,242	607	29,187
Unknown	1,923	4,355	189	6,467

Table 11 shows the top five cancer types stratified by race and gender in the linked dataset. Among the females, cervical cancer was the leading cancer type. Breast cancer was the second leading cancer type among females except for black African females where Kaposi sarcoma was the second leading cancer type and breast cancer was third. Kaposi sarcoma was the third leading cancer type among females of Asian, Coloured and Whites origin. Kaposi sarcoma was the leading cancer type for Asian, black Africans and Coloured males. Basal Cell Carcinoma (BCC) was the leading cancer type among White males. NHL was the second leading cancer type for males of black African and Asian origin while Lung and SCC were the second for Coloured and White males respectively.

Table 11: Top five cancer types stratified by race and gender in the linked dataset

Asians			Coloured		
Gender	Cancer Type	Number	Gender	Cancer Type	Number
Females	Cervix	480(28.54%)	Females	Cervix	5,084(28.36%)
	Breast	382(22.71%)		Breast	3,929(21.92%)
	KS	170(10.11%)		KS	1,702(9.49%)
	Uterus	66(3.92%)		NHL	897(5.00%)
	SCC	57(3.39%)		Uterus	562(3.13%)
Males	KS	210(21.94%)	Males	KS	1,578(14.99%)
	NHL	106(11.08%)		Lung	1,072(10.18%)
	Lung	81(8.46%)		NHL	1,004(9.54%)
	Leukaemia	58(6.06%)		Prostrate	530(5.04%)
	Colorectal	54(5.64%)		Eye	431(4.09%)
Blacks			**Whites**		
Females	Cervix	69,493(40.09%)	Females	Cervix	4,302(26.35%)
	KS	25,156(14.51%)		Breast	3,624(22.20%)
	Breast	20,371(11.75%)		KS	1,208(7.40%)
	NHL	8,610(4.97%)		NHL	733(4.49%)
	Eye	7,320(4.22%)		BCC	667(4.09%)
Males	KS	24,169(34.72%)	Males	BCC	1,544(12.01%)
	NHL	6,978(10.02%)		SCC	1,258(9.78%)
	Eye	3,487(5.01%)		KS	1,052(8.18%)
	Prostrate	2,931(4.21%)		Lung	911(7.09%)
	Lung	2,782(4.00%)		Prostrate	906(7.05%)

Table 12 shows the top five cancer types stratified by race and HIV status in the linked dataset. Breast and cervical cancers were the first and the second common cancer types among HIV negative Asians and Whites in the linked dataset. Cervical and breast cancers were the first and the second most common cancer types among black African and Coloured populations in the linked dataset. Leukaemia was the third most common cancer type among the HIV negative population of all races except in the white population where BCC was the third most common cancer type. Kaposi sarcoma and cervical cancer were the first and second leading types among HIV positive Asians and Coloureds in the linked dataset. Cervical cancer and Kaposi sarcoma were the first and second leading cancer types among black Africans and Whites. Breast cancer was the third most common cancer type among HIV positive population.

Table 12: Top five cancer types stratified by race and HIV status in the linked dataset

Asians		
Gender	Cancer Type	Number
Negative	Breast	290(23.81%)
	Cervix	128(10.51%)
	Leukaemia	76(6.24%)
	NHL	69(5.67%)
	Lung	67(5.50%)
Positive	KS	363(26.81%)
	Cervix	338(24.96%)
	Breast	87(6.43%)
	NHL	87(6.43%)
	SCC	71(5.24%)

Coloured		
Gender	Cancer Type	Number
Negative	Cervix	302(15.70%)
	Breast	223(11.60%)
	Leukaemia	106(5.51%)
	Lung	99(5.15%)
	Colorectal	89(4.63%)
Positive	KS	3,179(22.04%)
	Cervix	3,020(20.94%)
	Breast	1,324(9.18%)
	NHL	1,191(8.26%)
	Lung	569(3.94%)

Blacks		
Negative	Cervix	9,353(24.90%)
	Breast	4,579(12.19%)
	Leukaemia	1,733(4.61%)
	Uterus	1,594(4.24%)
	Oesophagus	1,392(3.71%)
Positive	Cervix	57,950(29.20%)
	KS	47,197(23.78%)
	Breast	15,635(7.88%)
	NHL	14,027(7.07%)
	Eye	10,181(5.13%)

Whites		
Negative	Breast	2,642(17.23%)
	Cervix	1,672(10.87%)
	BCC	1,226(7.99%)
	Leukaemia	732(4.77%)
	NHL	716(4.67%)
Positive	Cervix	2,558(19.22%)
	KS	2,184(16.41%)
	Breast	990(7.44%)
	BCC	950(7.14%)
	NHL	888(6.67%)

Table 13: HIV treatment province versus cancer diagnosis province

Province	EC	FS	GAU	KZN	LIM	MPU	NC	NW	WC
EC	21,093	111	843	112	8	70	86	75	2,616
FS	70	20,166	1,731	53	61	120	594	217	553
GAU	372	1,281	87,434	1,151	1,203	3,625	162	945	869
KZN	351	243	10,029	10,239	52	1,296	23	75	207
LIM	22	129	2,137	28	11,601	258	3	58	38
MPU	102	389	11,566	679	719	21,427	29	120	167
NW	85	590	5,089	34	215	111	198	15,770	448
NC	98	265	362	6	12	29	6,899	106	888
WC	834	105	820	100	15	35	246	42	58,173

Table 13 shows the representation in absolute numbers for the province of HIV treatment versus the province of cancer diagnosis in the linked records. The row values were the HIV treatment values while column values were the cancer diagnosis values. Close to half (44.52%) of cancer patients from KZN province had their cancers diagnosed in the Gauteng province. A third (32.75%) of cancer patients from Mpumalanga province came to Gauteng province for cancer diagnosis, 22.52% patients from North West province came to Gauteng province for cancer diagnosis while 14.96 of patients from Limpopo province also came to Gauteng province. Another notable movement of patients was also seen in Eastern Cape province as 10.45% of the patients went to Western Cape province for cancer diagnosis. Overall, Gauteng province diagnosed the most cancer patients at 38.75% followed by Western Cape at 20.65%.

4.3 RESULTS FOR HIGH DIMENSIONAL CLUSTERING

Figure 6 is a correlation matrix for the variables in the dataset. This showed the strength of association between variables in the linked dataset. Dark colours represent no correlation while light colours shows a high correlation. For example there's a high correlation between HIV status and ADC or NADC while there is no correlation between age and HIV status. There is a strong correlation between cancer topography and cancer morphology.

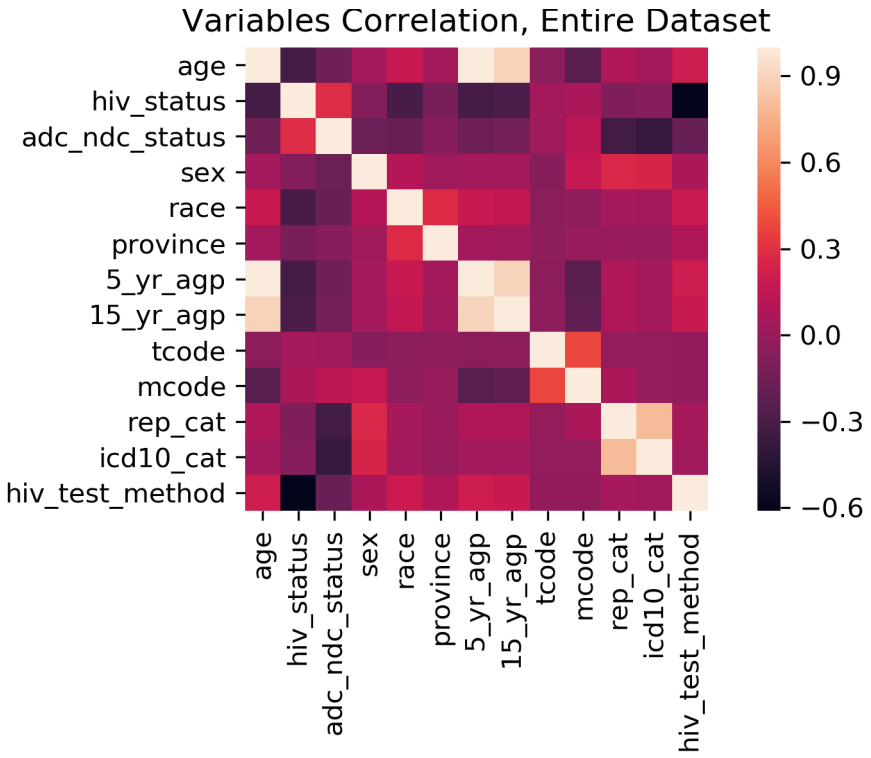


Figure 6: The correlation matrix for various variables in the linked dataset

The first cluster showed the general distributions of cancer types against the age at the time of diagnosis as shown in Figure 7. Light colours represent stronger association while dark colours show little or no association. From the figure, there is a strong association between several cancer types as people grow old, while a few cancer types do not depend on age.

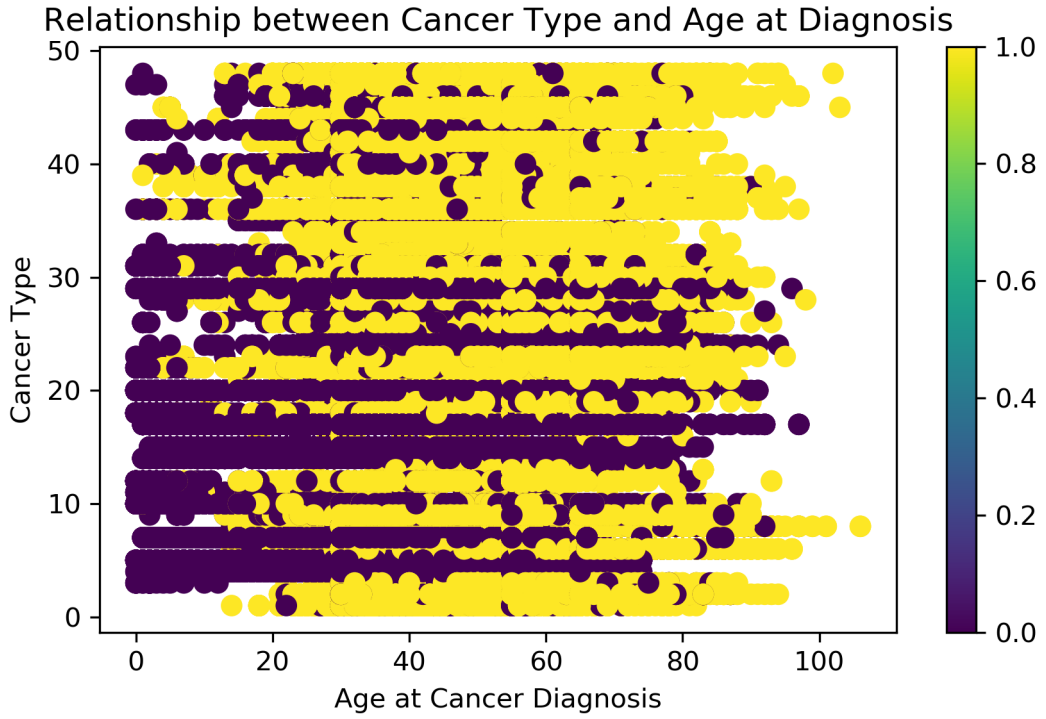


Figure 7: The relationship between various cancer types and age at cancer diagnosis in the linked dataset

Figure 8 shows the general distributions of cancer types against the age at the time of diagnosis for the HIV negative population. Light colours represent stronger association while dark colours show little or no association. For the HIV negative population in the linked dataset, there is a strong association between several cancer types for the younger population as compared to older population.

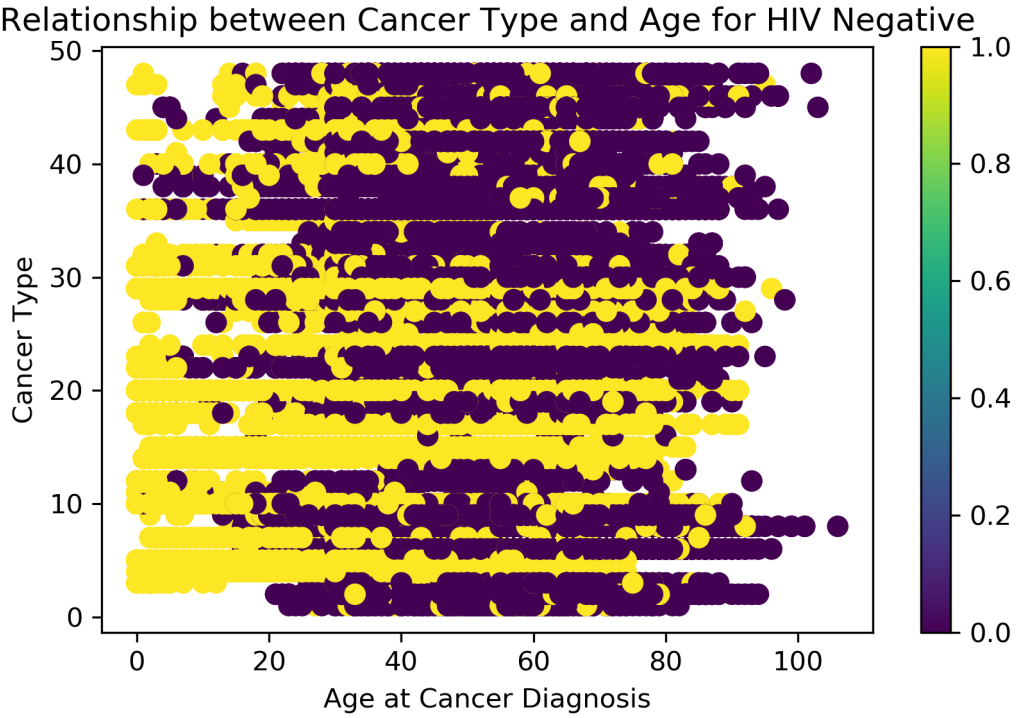


Figure 8: The relationship between various cancer types and age at cancer diagnosis for the HIV negative population

Figure 9 shows the general distribution of cancer types against the age at the time of cancer diagnosis. Light colours represent stronger association while dark colours show little or no association. For the HIV positive population in the linked dataset, there is a strong association between several cancer types for the older population as compared to the younger population.

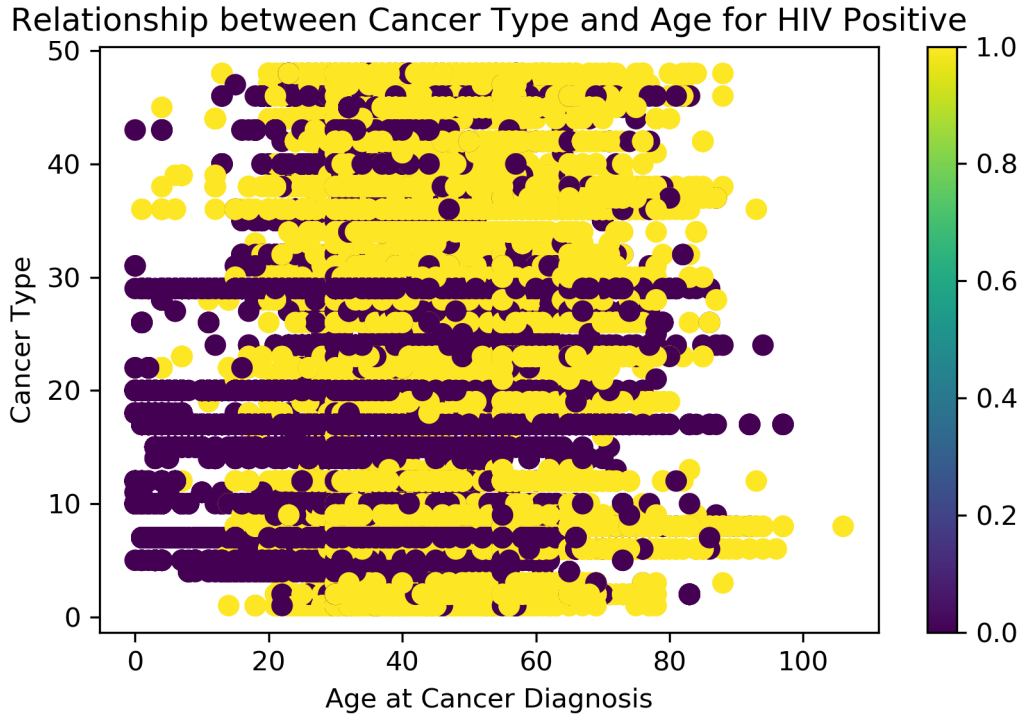


Figure 9: The relationship between various cancer types and age at cancer diagnosis for the HIV positive population

4.4 EVALUATION OF CLASSIFICATION RESULTS

The rows of records with national identification numbers aided in the calculation of precision, recall and F-measure. The TP were 2,468, while the FP were 328 and 7 were FN.

$$\text{Precision} = \frac{\text{TM}}{(\text{TM} + \text{FM})} = \frac{2468}{(2468 + 328)} = 0.883$$

$$\text{Recall} = \frac{\text{TM}}{(\text{TM} + \text{FN})} = \frac{2468}{(2468 + 7)} = 0.997$$

$$\text{F} - \text{measure} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} = 2 \times \frac{(0.883 \times 0.997)}{(0.883 + 0.997)} = 0.937$$

Figure 10 show a graphical representation of the sensitivity and specificity of the gender classification. This showed the agreement rate of the gender classification between the two datasets for the linked records. The area under Receiver operating characteristic (ROC) curve was 0.982, which is a measure of how well the algorithm was able to distinguish the gender classification between the two datasets.

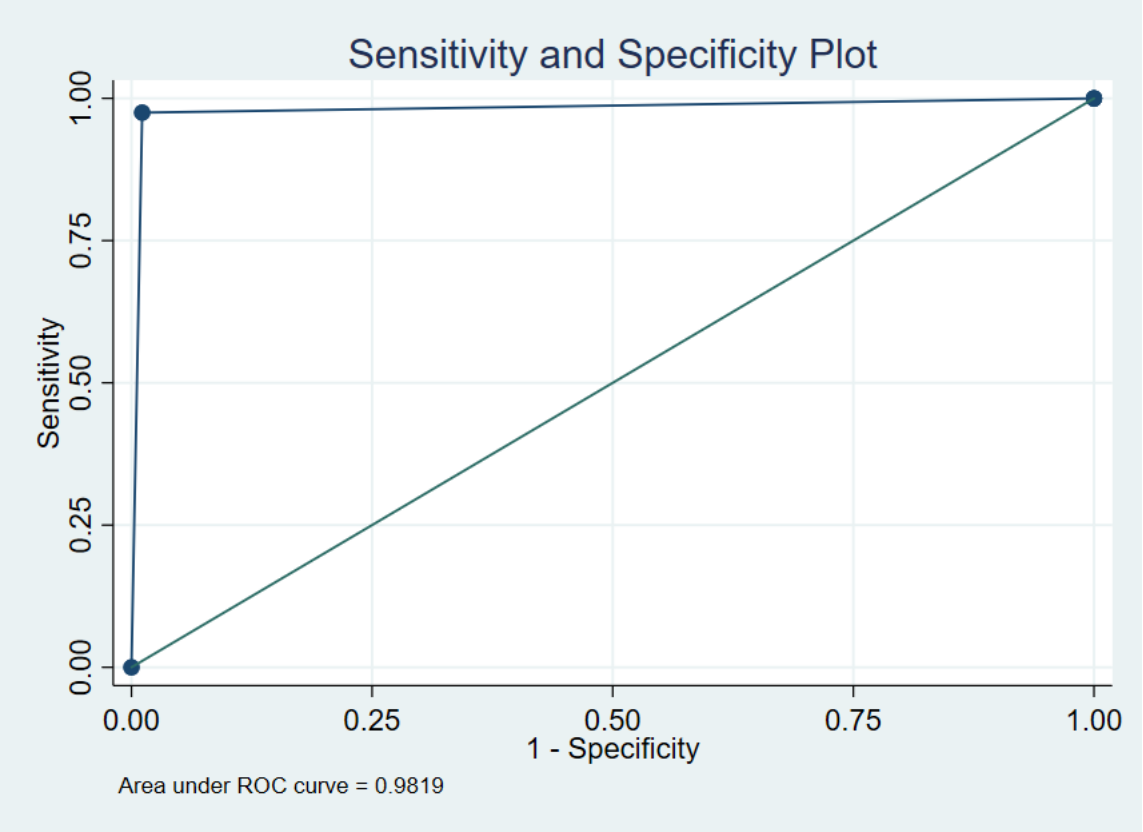


Figure 10: Sensitivity and specificity of gender classification

5

DISCUSSION

5.1 INTRODUCTION

This chapter presents the discussion of the results from the previous chapter, and how the methods and results helped achieve the study aims and objectives. It starts by explaining the general characteristics of the linked dataset and the patterns presented against the general characteristics of the population in South Africa, with elements on how the linkage quality was validated.

5.2 CHARACTERISTICS OF THE LINKED DATASET

For the study period 2004 to 2014, females had the highest burden of cancer with their proportion being 212,993(68.76%) as compared to men at 96,718(31.23%) while 30(0.01%) had unknown gender type. Since this was routine laboratory data, it is likely that doctors mostly tested patients with suspected HIV infection, hence, 231,945(76.91%) of the results were HIV positive while 69,648(23.09%) of the results were HIV negative while 8,148(2.63%) had no valid result. For the HIV positive 163,729(70.59%) were females, 68,198(29.40%) were males and 18(0.01%) had unknown gender. This also showed that females had the highest burden of HIV and cancer co-morbidity. This is likely to be attributed by the fact that women access care and treatment earlier as compared to men [53]. The distribution of the races in our linked dataset was 242,994(78.45%) Blacks, 29,187(9.42%) Whites, 28,454(9.19%) Coloured, 2,639(0.85%) Asian and 6,467(2.09%) Un-

known race. These results for race distributions mirror the mid year estimates published by Statistics South Africa (STATS SA) [45]. This reflects the use of HIV care services in the public sector by the different population groups.

The age group 30 to 54 years had the highest burden of cancer with $207,077(66.85\%)$ cases reported. This age group also reported the highest proportion of HIV positive results. This is explained by the fact that it is the reproductive age group and people in this age group are sexually active therefore are at higher risk of HIV infection [54]. For the age group 14 years and below, those that were HIV negative had a higher cancer burden as compared to HIV positive in the same age group. A possible explanation for this may be the introduction of Prevention of Mother-to-Child Transmission (PMTCT) programme [53]. In the programme all pregnant women are tested for HIV and if positive given HIV treatment during and after pregnancy [53]. This helps the mothers prevent transmission of HIV to their children. For the age group 65 years and above, $15,018(68.53\%)$ were HIV negative while $6,334(28.90\%)$ were HIV positive showing a higher cancer burden for HIV negative compared to HIV positive individuals in the same age group. The population in this age group are outside their reproductive age and are at lower risk of HIV infection [54]. On the same note, HIV positive individuals are more likely to die at a younger age compared to HIV negative individuals since our data is outside the test and treat era [54]. The results showed that black Africans had the highest burden of cancer with a proportion of $242,994(78.45\%)$ of the entire population in the linked dataset. In addition $198,477(81.68\%)$ of the black Africans were HIV positive. Although there are more blacks in South Africa as compared to other races, most of them access their HIV care and treatment through public health services and are likely to be found on the public health laboratory data [45, 54].

Cancer patients travel long distances, which might even involve moving to another province to seek treatment [15]. This was evident in our results which showed that 18.38% of the cancer patients sought cancer care and treatment in a province that was outside the area of their routine HIV care and treatment centres. Almost half of the cancer patients

from KZN province had their cancer cases diagnosed in the Gauteng province. This was likely due to declining oncology services in KZN which has been a long standing challenge. Other studies have also confirmed the shortage of oncologists in this province [26, 27]. Similar patterns were noted with cancer patients from Mpumalanga, North West, Limpopo going to Gauteng for cancer care. In the linked dataset, 27.15% of the patients did not have their HIV tests done in Gauteng province. Cancer diagnosis in other provinces are hindered by the shortage of oncology services and treatment backlog [9], thereby forcing the patients to seek treatment in other provinces with better oncology services like Gauteng and Western Cape.

The top ten cancers in the entire linked dataset were consistent with the reports published by the South African NCR for the period 2004 to 2014 [41, 48]. However, differences were noted in the spectrum of cancers in those who were HIV positive with vulva and Burkitt lymphoma also in the top ten. This reflects the effect of HIV as both cancers are associated with HIV [20].

Several correlations were observed within the linked dataset. For example, cancer type and sex, province and race, HIV status and ADC or NADC as well as HIV test and age of the person. The correlation of province of residence and race, is consistent with the geographical variation of races in South Africa as published by STATS SA [44]. For HIV status and ADC or NADC, this is attributed to the fact that there are some cancers that are common in people with HIV. With regards to HIV tests and age of the person, this relationship is clearly visible for tests that are only done in specific group such as PCR in children. The correlation between cancer type and sex, is attributed to the fact that there are gender specific cancers. These relationships portrayed in the heat map open up proxies for further research.

The high dimensional clusters shown in Figures 7, 8 and 9 showed that there were high chances of getting cancer as a person grows older. This scenario was different for the HIV negative population under 5 years, as a higher proportion were diagnosed with cancer as

compared to the HIV positive. There were more cancers in HIV negative children than the HIV positive children, which is comparable to the findings of Meredith et al [53].

5.3 EVALUATION MEASURES

The linkage quality of the NHLS HIV dataset and NCR cancer dataset was assessed by calculating precision, recall and F-measure. Despite the poor quality of the national ID number columns in both datasets, a precision of 88.3%, recall of 99.7% and F-measure of 93.7% were achieved. This shows that the linkage algorithm classified records with high level of accuracy for the true matches records and there was low misclassification. The F-measure represents the harmonic mean between precision and recall [5, 24], it shows the overall classification performance of the algorithm.

5.4 LIMITATIONS

The study only considered one model of machine learning. This is because literature supports SVM as a better classifier, compared to other machine learning models [61]. Even though we did a series of clerical reviews while testing the algorithms to make sure that a column sum of six gives a match between the records. We failed to perform a k-fold validation, that would have also improved the performance of the SVM classifier.

5.5 RECOMMENDATIONS

Routinely collected data from different sources stored in a central repository is often not clean. This is because the data collection might not be standard and so consistency is lacking during data entry. It is therefore appropriate to enforce validation checks during data entry to standardize the data being entered onto TrakCare. Constant data cleaning

is also necessary to make sure that the database has clean data which is ready for use. This would cut down time for other projects that may want to use the CDW data as data cleaning phase is reduced to minimal if not eliminated.

6

CONCLUSIONS AND FUTURE DIRECTIONS

This report focused on the application of supervised machine learning algorithms to efficiently link cancer and HIV dataset in the absence of unique identifiers. This work helps in bridging the gap by providing a platform for studying disease co-morbidities of HIV and cancer. In most cases, studying disease co-morbidities is often a challenge as there are no unique identifiers that may be used to link records belonging to the two databases. The work linked national HIV laboratory data to national cancer data, thereby allowing for the creation of national cohorts for people with HIV and cancer in the whole of South Africa. Our linkage was consistent with the available literature on cancer and HIV.

The results showed that, there are more cancer cases in HIV negative children under 5 years of age as compared to HIV positive ones in the same age bracket. In overall, the mean age for the HIV positive cancer patients was 10 years younger than the mean age for HIV negative cancer patients. Even though Gauteng province had the highest number of cancer diagnosis, about 27% of cancer diagnosis were patients who do not reside in the province. This raises concern over the oncology services in other provinces in South Africa.

Due to time limitation and scope of the project work, statistical models were not performed in the linked dataset. The results were based on the high dimensional clustering and descriptive statistics. Therefore, performing multivariate analysis in the linked dataset to get an in-depth knowledge on cancer and HIV co-morbidities is necessary. It is also necessary to perform geo-mapping for the HIV positive cancer patients who travelled to another province for cancer diagnosis. This would be important in getting the distance travelled to seek cancer diagnosis and the complications involved in travelling

based on the distance between the HIV care and treatment facility and cancer diagnosis facility. Moreover this will help the government to effect policies that improve the oncology services in other provinces.

With the learning strength of supervised machine learning algorithm, SVM can be used on python pipelines that enable the classification of continually growing datasets with minimal human input, as long as the characteristics of the datasets are known and the domain knowledge of the linkage is known. The algorithm will use the already existing labels and label characteristics in order to classify a new weighted dataset sent though the pipeline. This will save time in scripting, testing and evaluating a new algorithm, while maintaining high accuracy and efficiency. Such python pipelines can not only be applied to linking HIV to cancer datasets but to other big data linkage projects as well. These efficient pipelines can be adapted to harmonize NHLS data at large and create and enrich national cohorts of other diseases for epidemiological analyses.

BIBLIOGRAPHY

[1] Indrajit Bhattacharya and Lise Getoor. "Iterative record linkage for cleaning and integration." In: *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* (2004), pp. 11–18.

[2] Mikhail Bilenko, Beena Kamath, and Raymond J Mooney. "Adaptive blocking: Learning to scale up record linkage." In: *IEEE* (2006), pp. 87–96.

[3] Tony Blakely and Clare Salmond. "Probabilistic record linkage and a method to calculate the positive predictive value." In: *International journal of epidemiology* 31.6 (2002), pp. 1246–1252.

[4] Julia Bohlius, Nicola Maxwell, Adrian Spoerri, Rosalind Wainwright, Shobna Sawry, Janet Poole, Brian Eley, Hans Prozesky, Helena Rabie, Daniela Garone, et al. "Incidence of AIDS-defining and other cancers in HIV-positive children in South Africa: record linkage study." In: vol. 35. 6. NIH Public Access, 2016, e164.

[5] MJ Carey, S Ceri, P Bernstein, U Dayal, C Faloutsos, JC Freytag, G Gardarin, W Jonker, V Krishnamurthy, MA Neimat, et al. "Data-Centric Systems and Applications." In: *Springer* (2008).

[6] James Chipperfield, Noel Hansen, and Peter Rossiter. "Estimating Precision and Recall for Deterministic and Probabilistic Record Linkage." In: *International Statistical Review* (2018).

[7] Peter Christen. "The Data Matching Process." In: *Springer* (2012), pp. 23–35.

[8] David E Clark and David R Hahn. "Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry." In: *Proceedings of the annual symposium on computer application in medical care*. American Medical Informatics Association. 1995, p. 397.

[9] South African Human Rights Commission. "South Africa: Treatment Of Cancer In North West Is Impossible." In: *https://www.sahrc.org.za/index.php/sahrc-media/news/item/1378-south-africa-treatment-of-cancer-in-north-west-is-impossible* (2018).

[10] National Institute of Communicable Diseases. "Naitonal Cancer Registry." In: *http://www.nicd.ac.za/index.php/centres/national-cancer-registry/* (2018).

[11] Jonathan De Bruin. "Probabilistic record linkage with the Fellegi and Sunter framework: Using probabilistic record linkage to link privacy preserved police and hospital road accident records." In: *Diss. Master's thesis, Delft University of Technology* (2015).

[12] Dennis Deapen, Myles Cockburn, Rich Pinder, Sharon Lu, and Amy Rock Wohl. "Population-based linkage of AIDS and cancer registries: importance of linkage algorithm." In: vol. 33. 2. Elsevier, 2007, pp. 134–136.

[13] Marco Dozza, Jonas Bärgman, and John D Lee. "Chunking: A procedure to improve naturalistic data analysis." In: *Accident Analysis & Prevention* 58 (2013), pp. 309–317.

[14] Stacie B Dusetzina, Seth Tyree, Anne-Marie Meyer, Adrian Meyer, Laura Green, and William R Carpenter. "Linking data for health services research: a framework and instructional guide." In: (2014).

[15] Lynn Barbara Edwards and Linda Estelle Greeff. "Exploring grassroots feedback about cancer challenges in South Africa: a discussion of themes derived from content thematic analysis of 316 photo-narratives." In: *Pan African Medical Journal* 28.1 (2017).

[16] Eric A Engels, Robert J Biggar, H Irene Hall, Helene Cross, Allison Crutchfield, Jack L Finch, Rebecca Grigg, Tara Hylton, Karen S Pawlish, Timothy S McNeel, et al. "Cancer risk in people infected with human immunodeficiency virus in the United States." In: *International journal of cancer* 123.1 (2008), pp. 187–194.

[17] Sheela V Godbole, Karabi Nandy, Mansi Gauniyal, Pallavi Nalawade, Suvarna Sane, Shravani Koyande, Joy Toyama, Asha Hegde, Phil Virgo, Kishor Bhatia, et al. "HIV and cancer registry linkage identifies a substantial burden of cancers in persons with HIV in India." In: *Medicine* 95.37 (2016).

[18] Shaun J Grannis, J Marc Overhage, and Clement J McDonald. "Analysis of identifier performance using a deterministic linkage algorithm." In: *Proceedings of the AMIA symposium*. American Medical Informatics Association. 2002, p. 305.

[19] Shaun J Grannis, J Marc Overhage, Siu Hui, and Clement J McDonald. "Analysis of a probabilistic record linkage technique without human review." In: *AMIA annual symposium proceedings*. Vol. 2003. American Medical Informatics Association. 2003, p. 259.

[20] Andrew E Grulich, Marina T Van Leeuwen, Michael O Falster, and Claire M Vajdic. "Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis." In: *The Lancet* 370.9581 (2007), pp. 59–67.

[21] National Cancer Institute. "HIV/AIDS Cancer Match Study." In: *https://hivmatch.cancer.gov/* (2018).

[22] Jane Joubert, Debbie Bradshaw, Chodziwadziwa Kabudula, Chalapati Rao, Kathleen Kahn, Paul Mee, Stephen Tollman, Alan D Lopez, and Theo Vos. "Record-linkage comparison of verbal autopsy and routine civil registration death certification in rural north-east South Africa: 2006–09." In: vol. 43. 6. Oxford University Press, 2014, pp. 1945–1958.

[23] Chodziwadziwa W Kabudula, Benjamin D Clark, Francesc Xavier Gómez-Olivé, Stephen Tollman, Jane Menken, and Georges Reniers. "The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa." In: *BMC medical research methodology* 14.1 (2014), p. 71.

[24] Hanna Köpcke, Andreas Thor, and Erhard Rahm. "Evaluation of entity resolution approaches on real-world match problems." In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), pp. 484–493.

[25] Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and HV Jagadish. "Regular expression learning for information extraction." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2008, pp. 21–30.

[26] Times Live. "Cancer patients in KZN must ŕelocate to another province or dieás last specialist in Durban resigns." In: *https://www.timeslive.co.za/news/south-africa/2017-06-08-cancer-patients-in-kzn-must-relocate-to-another-province-or-die-as-last-specialist-in-durban-resigns/* (2017).

[27] Times Live. "Human Rights Commission slams Ĺack of progressín KZN oncology crisis." In: *https://www.timeslive.co.za/news/south-africa/2018-04-20-human-rights-commission-slams-lack-of-progress-in-kzn-oncology-crisis/* (2018).

[28] Sam M Mbulaiteye, Elly T Katabira, Henry Wabinga, Donald M Parkin, Phillip Virgo, Robert Ochai, Meklit Workneh, Alex Coutinho, and Eric A Engels. "Spectrum of cancers among HIV-infected persons in Africa: the Uganda AIDS-Cancer Registry Match Study." In: *International Journal of Cancer* 118.4 (2006), pp. 985–990.

[29] Andrew McCallum, Kamal Nigam, and Lyle H Ungar. "Efficient clustering of high-dimensional data sets with application to reference matching." In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2000, pp. 169–178.

[30] Deven McGraw, Sarah M Greene, Caroline S Miner, Karen L Staman, Mary Jane Welch, and Alan Rubel. "Privacy and confidentiality in pragmatic clinical trials." In: *Clinical Trials* 12.5 (2015), pp. 520–529.

[31] David Menestrina, Steven Euijong Whang, and Hector Garcia-Molina. "Evaluating entity resolution results." In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), pp. 208–219.

[32] Matthew Michelson and Craig A Knoblock. "Learning blocking schemes for record linkage." In: *AAAI* (2006), pp. 440–445.

[33] Andrew W Moore. "Support vector machines." In: *Tutorial. School of Computer Science of the Carnegie Mellon University. Available at http://www. cs. cmu. edu/awm/tutorials.[Accessed August 16, 2009]* (2001).

[34] Jared S Murray. "Probabilistic record linkage and deduplication after indexing, blocking, and filtering." In: *arXiv preprint arXiv:1603.07816* (2016).

[35] NHLS. "National Health Laboratory Service." In: *http://www.nhls.ac.za/* (2018).

[36] Dermot O'Reilly, Heather Kinnear, Michael Rosato, Adrian Mairs, and Clare Hall. "Using record linkage to monitor equity and variation in screening programmes." In: vol. 12. 1. BioMed Central, 2012, p. 59.

[37] Gisele Pinto de Oliveira, Ana Luiza de Souza Bierrenbach, Kenneth Rochel de Camargo Júnior, Cláudia Medina Coeli, and Rejane Sobrino Pinheiro. "Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis." In: *Revista de saude publica* 50 (2016), p. 49.

[38] Shirley Ong Ai Pei. "A comparative study of record matching algorithms." In: *RWTH Aachen, Thesis, Germany, University of Edinburgh, Scotland* (2008).

[39] P Jonathon Phillips. "Support vector machines applied to face recognition." In: *Advances in Neural Information Processing Systems*. 1999, pp. 803–809.

[40] Erhard Rahm and Hong Hai Do. "Data cleaning: Problems and current approaches." In: *IEEE Data Eng. Bull.* 23.4 (2000), pp. 3–13.

[41] South African National Cancer Registry. "Cancer in South Africa." In: *www.ncr.ac.za* (2014), pp. 69–90.

[42] Douglas Reynolds. "Gaussian mixture models." In: *Encyclopedia of biometrics* (2015), pp. 827–832.

[43] Hilary A Robbins, Meredith S Shiels, Ruth M Pfeiffer, and Eric A Engels. "Epidemiologic contributions to recent cancer trends among HIV-infected people in the United States." In: *AIDS (London, England)* 28.6 (2014), p. 881.

[44] STATS SA. "Census 2011 Census in brief." In: *www.statssa.gov.za* Report No. 03-01-41 (2012).

[45] STATS SA. "Mid-year population estimates, 2017." In: *www.statssa.gov.za* STATISTICAL RELEASE P0302 (2017).

[46] Adrian Sayers, Yoav Ben-Shlomo, Ashley W Blom, and Fiona Steele. "Probabilistic record linkage." In: *International journal of epidemiology* 45.3 (2015), pp. 954–964.

[47] Leo J Schouten, J Th Schlangen, JM de Rijke, and André LM Verbeek. "Evaluation of the effect of breast cancer screening by record linkage with the cancer registry, the Netherlands." In: *Journal of medical screening* 5.1 (1998), pp. 37–41.

[48] Mazvita Sengayi, Chantal Babb, Matthias Egger, and Margaret I Urban. "HIV testing and burden of HIV infection in black cancer patients in Johannesburg, South Africa: a cross-sectional study." In: *BMC cancer* 15.1 (2015), p. 144.

[49] Mazvita Sengayi, Adrian Spoerri, Matthias Egger, Danuta Kielkowski, Tamaryn Crankshaw, Christie Cloete, Janet Giddy, and Julia Bohlius. "Record linkage to correct under-ascertainment of cancers in HIV cohorts: The Sinikithemba HIV clinic linkage project." In: *International journal of cancer* 139.6 (2016), pp. 1209–1216.

[50] Mazvita Sengayi, Wenlong Chen, Adrian Spoerri, Elvira Singh, Matthias Egger, and Julia Bohlius. "SOUTH AFRICAN HIV CANCER MATCH STUDY: A PILOT STUDY TOWARDS PRECISION PUBLIC HEALTH." In: *http://www.croiconference.org/sessions/south-african-hiv-cancer-match-study-pilot-study-towards-precision-public-health* (2018).

[51] National Health Laboratory Service. "TrackCare." In: *https://trakcarelabwebview.nhls.ac.za/trakcarelab/* (2018).

[52] Noam Shental, Aharon Bar-Hillel, Tomer Hertz, and Daphna Weinshall. "Computing Gaussian mixture models with EM using equivalence constraints." In: *Advances in neural information processing systems*. 2004, pp. 465–472.

[53] Meredith S Shiels, Ruth M Pfeiffer, and Eric A Engels. "Age at cancer diagnosis among persons with AIDS in the United States." In: *Annals of internal medicine* 153.7 (2010), pp. 452–460.

[54] Olive Shisana, Thomas Rehle, Leickness C Simbayi, Khangelani Zuma, Sean Jooste, Nompumelelo Zungu, Demetre Labadarios, and Dorina Onoya. "South African national HIV prevalence, incidence and behaviour survey, 2012." In: *HSRC press* (2014).

[55] Elvira Singh, Paul Ruff, Chantal Babb, Mazvita Sengayi, Moira Beery, Lerato Khoali, Patricia Kellett, and J Michael Underwood. "Establishment of a cancer surveillance programme: the South African experience." In: *The Lancet Oncology* 16.8 (2015), e414–e421.

[56] June Smith-Tyler. "Informed consent, confidentiality, and subject rights in clinical trials." In: *Proceedings of the American Thoracic Society* 4.2 (2007), pp. 189–193.

[57] Alan F Westin and Oscar M Ruebhausen. *Privacy and freedom*. Vol. 1. Atheneum New York, 1967.

[58] Leland Wilkinson and Michael Friendly. "The history of the cluster heat map." In: *The American Statistician* 63.2 (2009), pp. 179–184.

[59] William E Winkler. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." In: *ERIC* (1990).

[60] William E Winkler. "Methods for record linkage and bayesian networks." In: *U.S. Bureau of the Census* (2002).

[61] William E Winkler et al. "Machine learning, information retrieval and record linkage." In: 2000, pp. 20–29.

[62] Ying Zhu, Yutaka Matsuyama, Yasuo Ohashi, and Soko Setoguchi. "When to conduct probabilistic linkage vs. deterministic linkage? A simulation study." In: *Journal of biomedical informatics* 56 (2015), pp. 80–86.

A

ETHICS CLEARANCE CERTIFICATE FOR THE STUDY

R14/49 Mr Vistor Olago

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M171176

NAME: (Principal Investigator)	Mr Vistor Olago
DEPARTMENT:	Public Health National Cancer Registry National Health Laboratory Services
PROJECT TITLE:	Record Linkage of National Health Laboratory Services HIV Datasets to Cancer Registry Datasets using Supervised Learning Techniques
DATE CONSIDERED:	24/11/2017
DECISION:	Approved uncoditionally
CONDITIONS:	
SUPERVISOR:	Dr Gideon Nimako and Dr Mazvita Sengayi
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	22/12/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review November and will therefore be due in the month of November each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 19/1/2018

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure 11: Ethics clearance certificate for the study

B

ETHICS CLEARANCE CERTIFICATE FOR THE PARENT STUDY

R14/49 Dr Elvira Singh et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M140602

NAME: (Principal Investigator)	Dr Elvira Singh et al
DEPARTMENT:	School of Public Health National Health Laboratory Services
PROJECT TITLE:	The South African HIV/AIDS Cancer Match Study
DATE CONSIDERED:	27/06/2014 (Initial Approval 02/06/2017)
DECISION:	Approved unconditionally
CONDITIONS:	Annual Recertification (2017)
SUPERVISOR:	
APPROVED BY:	Professor P. Cleaton-Jones Chairperson, HREC (Medical)
DATE OF APPROVAL:	17/06/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review June and will therefore be due in the month of June each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date 20/07/2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure 12: Ethics clearance certificate for the parent Study

C

LETTER FROM THE GATEKEEPER

NATIONAL HEALTH LABORATORY SERVICE
National Cancer Registry

National Cancer Registry
1 Modderfontein Road, Sandringham, Johannesburg 2131
Tel: +27 (0)11 555 0548 Fax: +27 (0)11 386 6516

02 November 2017

The University of the Witwatersrand
Human Research Ethics Committee (Medical)
Faculty of Health Sciences

RE: Authorisation to use data from the South Africa National Cancer Registry

Mr Victor Olago (student number: 1739470) is a MSc Epidemiology (Research Database Management) student at Wits School of Public Health who is currently placed at the National Cancer Registry (NCR) through a capacity building cancer epidemiology fellowship funded by the National Institutes of Health and the Swiss National Science Foundation.

This letter authorises data usage for the research topic***: Record Linkage of NHLS HIV Datasets to Cancer Registry Datasets using Supervised Learning Techniques***. The dataset will be provided to Mr Victor Olago upon receipt of ethical clearance from the University of the Witwatersrand Human Research Ethics Committee.

The nature of the research project requires patient identifiers to allow for linking HIV and cancer records routinely collected by the National Health Laboratory Service (NHLS) and the NCR, respectively using supervised machine learning. The National Cancer Registry collects identifiable cancer reports as part of its routine operations and the use of identifiers is required in quality control and identification of records belonging to the same person, which Victor will have to do for his own project. All identifiers will be removed after record linkage and data cleaning is completed and a de-identified dataset will be used in the analysis. The NCR has ethics clearance for linking HIV data to NCR data ***(The South African HIV Cancer Match Study (M140602))*** and written permission from the NHLS executive for the parent study which are attached. The current study is a sub-study of the South African HIV Cancer Match Study. He is hereby authorised to use data from the National Cancer Registry. The authorisation restricts the use of the data only towards his research project as described in his protocol.

Kind regards,

Dr Elvira Singh
Principal Investigator
The South African HIV Cancer Match Study
Head: National Cancer Registry
National Health Laboratory Service
Tel: +27(0)11 386 6407 | Mobile: +27(0)84 587 7010
elvira.singh@nioh.nhls.ac.za | www.ncr.ac.za | www.nhls.ac.za

Figure 13: Letter from the gatekeeper

D

LETTER FROM THE GATEKEEPER FOR THE PARENT STUDY

NATIONAL HEALTH LABORATORY SERVICE

Academic Affairs and Research
Modderfontein Road, Sandringham, 2031
Tel: +27 (0)11 386 6142
Fax: +27 (0)11 386 6296
Email: babatyi.kgokong@nhls.ac.za
Web: www.nhls.ac.za

09 February 2016

Applicant: Mazvita Sengayi
Institution: National Health Laboratory Service
Division: National Cancer Registry
Email: mazvita.sengayi@nhls.ac.za
Tel: 011 489 9178

Re: Approval to access National Health Laboratory Service (NHLS) Data

Your application to undertake a research project "**The South African HIV Cancer Match Study**" using data from the NHLS database has been reviewed. This letter serves to advise that the application has been approved and the required data will be made available to you to conduct the proposed study as outlined in the submitted application.

Please note that the approval is granted on your compliance with the NHLS conditions of service and that the study can only be undertaken provided that the following conditions have been met.

- Linkage will be done at the Cancer Registry, and data will be managed at the NHLS National Cancer Registry.
- Data will be anonymous and shared with the Primary Investigator at the National Cancer Registry.
- Ethics approval is obtained from a recognised SA Health Research Ethics Committee.
- Processes are discussed with the relevant NHLS departments (i.e. Information Management Unit and Operations Department) and are agreed upon.
- Confidentiality is maintained at participant and institutional level and there is no disclosure of personal information or confidential information as described by the NHLS policy.
- A final report of the research study and any published paper resulting from this study are submitted and addressed to the NHLS Academic Affairs and Research office and the NHLS has been acknowledged appropriately.

Please note that this letter constitutes approval by the NHLS Academic Affairs and Research. Any data related queries may be directed to Sue Candy, manager NHLS Corporate Data Warehouse, Tel: (011) 386 6036. Email: sue.candy@nhls.ac.za.

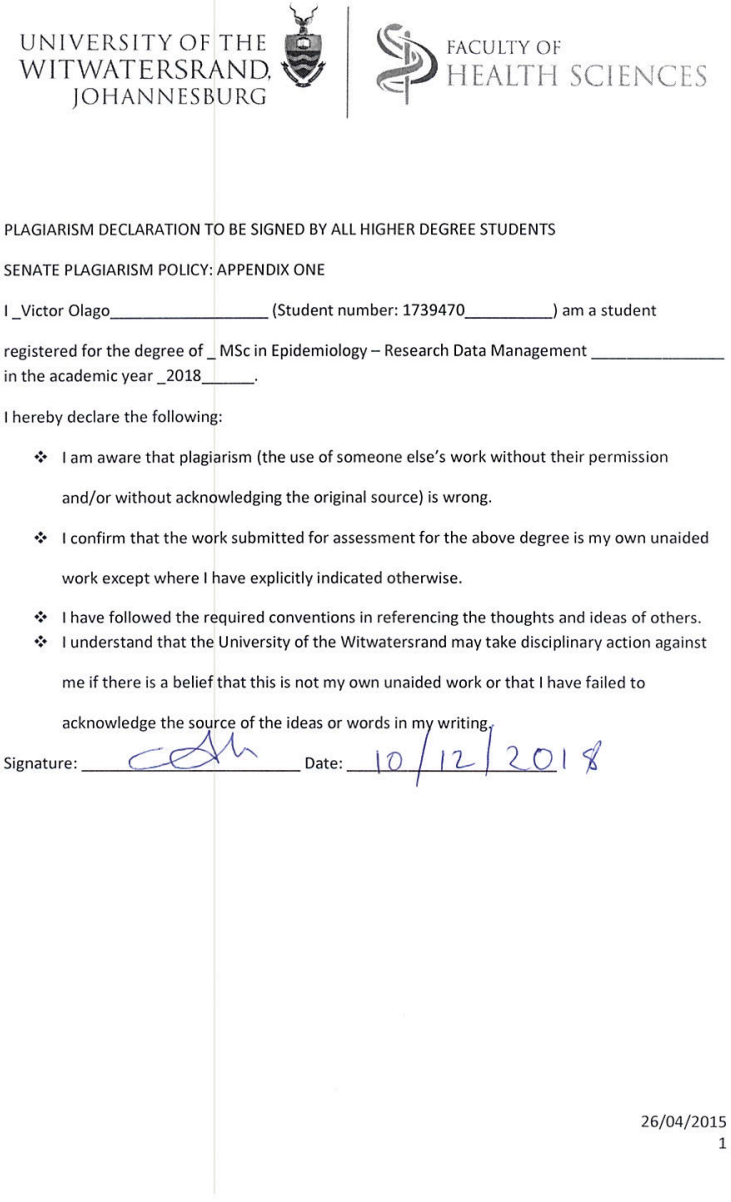
Yours sincerely,

Dr Babatyi Malope-Kgokong
National Manager: Academic Affairs and Research

Figure 14: Letter from the gatekeeper for the parent study

E

ANTI-PLAGIARISM DECLARATION



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG | FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I _Victor Olago_____ (Student number: 1739470_____) am a student registered for the degree of _ MSc in Epidemiology – Research Data Management _____ in the academic year _2018_____.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: _____ Date: 10/12/2018

26/04/2015
1

Figure 15: Plagiarism declaration

ORIGINALITY REPORT

1739470:Olago_ResearchReport.pdf

ORIGINALITY REPORT

12%	9%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	academic.oup.com Internet Source	1%
2	uir.unisa.ac.za Internet Source	<1%
3	www.nicd.ac.za Internet Source	<1%
4	www.nioh.ac.za Internet Source	<1%
5	Sengayi, Mazvita, Adrian Spoerri, Matthias Egger, Danuta Kielkowski, Tamaryn Crankshaw, Christie Cloete, Janet Giddy, and Julia Bohlius. "Record linkage to correct under-ascertainment of cancers in HIV cohorts: the Sinikithemba HIV clinic linkage project", International Journal of Cancer, 2016. Publication	<1%
6	www.slideshare.net Internet Source	<1%
7	X.-D. Sun. "Prediction of protein structural	

Figure 16: Originality report