An Investigation of the Presence of Cervical Schistosomiasis in Patients with Cervical Cancer and Pre-Cancer Lesions: A Text Mining Approach

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Carole Metekoua Temdemnou

Student Number: 1363322

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 - Public Health Informatics

18th June 2020

Carole Metekoua Temdemnou . *2020*
An Investigation of the Presence of Cervical Schistosomiasis in Patients with Cervical Cancer and Pre-Cancer Lesions: A Text Mining Approach .

SUPERVISORS:
Mr. Michael Mapundu, University of The Witwatersrand
Dr. Mazvita Muchengeti, National Cancer Registry

DECLARATION

I hereby declare that this research report is project work carried out by me under the guidance of Mr. Michael Mapundu 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,
18th June 2020

Carole Metekoua Temdemnou

DEDICATION

I dedicate this work to Mrs Toukam Jeanette epouse Temdemnou. You always believed in me.

ABSTRACT

Background: Cervical cancer is the second most common cancer in developing countries. Few studies have associated its incidence rate to cervical schistosomiasis which is also endemic in Sub-Saharan Africa (SSA) while others have rejected this association. This contradiction could be a result of small sample sizes which introduce bias in study findings. Therefore, there is a need to appraise cost-effective methods such as text-mining. Text-mining of histology reports in laboratory databases may provide sufficient sample size to adequately test the relationship between cervical schistosomiasis and cervical cancer and precancerous lesions and provide more robust findings.

Aim: This study aimed to use text-mining techniques on pathology reports available at National Health Laboratory Service (NHLS) to investigate the presence of cervical schistosomiasis in patients with precancerous lesions and cervical cancer.

Methods: Detailed cervical histopathology reports from Inkosi Albert Luthuli Hospital from 2011 to 2017 were obtained from NHLS. Only 50% of records had a diagnosis in terms of cervical precancerous lesions, cancer and other cervical pathologies. The reports were pre-processed and One-Versus-One (OVO) and One-versus-All (OVA) Support Vector Machine (SVM) and Random Forest (RF) classifiers were trained and optimised to assign diagnosis to records without diagnosis. The performance evaluation of all classifiers was assessed using recall, precision and F-measure. The best performing classifier was used to predict the pathology group of all records in the dataset. Furthermore, word/phrase matching was used to classify records with cervical precancerous lesions into their various grades and cancer records into various types of cervical cancer. Word/phrase was also used to extract the Human Papillomavirus (HPV) and schistosomiasis status from all selected records. Frequencies and proportions were used to describe reporting patterns of schistosomiasis and HPV. Subsequently, the prevalence of schistosomiasis was calculated using a dataset that was restricted to records with a known schistosomiasis status. The Independent t-test was used to assess if there was any relationship between schistosomiasis and age at cervical pre-cancerous lesions and cancer diagnosis. A sensitivity analysis was also done. It was done under the assumption that patients without schistosomisis related terms in their histology report were schistosomiasis negative.

Results: OVO SVM accurately classified 94.2% of records into cervical precancerous lesions, cancer and other pathologies. Schistosomiasis and HPV was reported in 5.58% (n=1,315) and 58.02% (n=13,668) of records, respectively. The

prevalence of schistosomiasis was 42.80% (n=464) in cervical pre-cancerous lesions and 23.3% (n=54) in cervical cancer in samples where schistosomiasis status was reported. This prevalence changed to 2.59% in cervical precancerous lesions and 0.96% in cervical cancer under the sensitivity analysis. The mean age of pre-cancer diagnosis in schistosomiasis positive patients was 38 ± 9.46 years while that of schistosomiasis negative patients was 40 ± 9.86 years. The mean age at cancer diagnosis in schistosomiasis positive patients was 45 ± 13.70 years while that of schistosomiasis negative patients was 50 ± 13.78 years. The mean age difference between the schistosomiasis positive patients and schistosomiasis negative patients were statistically significant in both cervical pre-cancerous lesions and invasive cancer. These findings were similar in the sensitivity analysis.

Conclusion: This study shows that text-mining techniques effectively classified records into various pathology groups. Through text-mining, this study identified the largest number of cervical precancerous lesions and cancer cases that could be used to explore the relationship between cervical schistosomiasis and cervical precancerous lesions/cancer compared to other studies. Women with schistosomiasis may develop cervical precancerous lesions and cancer 2 and 5 years earlier than patients without schistosomiasis respectively. However, findings were limited by the low schistosomiasis reporting rates in this population. There is need for pathologists to report the schistosomiasis status in cervical precancerous lesions and cancer as this type of study highly relies on routinely collected data. Moreover, further studies are required to explore the effect of HIV and schistosomiasis treatment on the control of cervical precancerous lesions and cervical cancer in SSA.

ACKNOWLEDGMENTS

I would like to acknowledge the following people and organisations, without whom I would not have been able to complete this research:

- I would like to express my sincere gratitude to my supervisors Dr Mazvita Muchengeti and Mr Michael Mapundu for their constant support, guidance and motivation.
- Thank you to NHLS for providing the dataset required for this research.
- Thank you to the School of Public Health at the University of the Witwatersrand for paying my school fees in my second year and supporting me all through my studies.
- Completing this work would have been very difficult without the technical and friendly support of Mr Victor Olago. I am indebted to him for his help.
- Thank you to my colleagues at the School of Public Health. Discussing with you always gave the strength to put in more efforts in the research
- To my family and friends, hanging out with you always gave me a better perspective in my work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CDW Corporate Data Warehouse
CIN Cervical Intraepithelial Neoplasia
DNA Deoxyribonucleic Acid
FGS Female Genital Schistosomiasis
FN False Negative
FP False Positive
HIV Human Immunodeficiency Virus
HPV Human Papillomavirus
HSIL High-grade Squamous Intraepithelial Lesions
KZN KwaZulu-Natal
LSA Latent Semantic Analysis
LSIL Low-grade Squamous Intraepithelial Lesions
ML Machine Learning
NHLS National Health Laboratory Service
NHLS-CDW National Health Laboratory Service Corporate Data Warehouse
NLTK Natural Language Toolkit
NLP Natural Language Processing
OVA One-versus-All
OVO One-Versus-One
RF Random Forest
SNOMED-CT Systemised Nomenclature of Medicine Clinical Terms
SSA Sub-Saharan Africa
SVD Singular Value Decomposition
SVM Support Vector Machine
TF-IDF Term Frequency-Inverse Document Frequency
TN True Negative
TP True Positive
WHO World Health Organisation

1

INTRODUCTION

1.1 BACKGROUND

Cervical cancer is the second most common cancer in developing countries [1]. Rural communities in sub-Saharan Africa SSA carry the highest burden of cervical cancer [1, 2]. They have an age-standardised incidence and mortality rates of 33 and 21 per 100,000 women per year respectively. [1]. In 2012, 18% of all new cervical cancer cases and 22% of all cervical cancer deaths were from SSA. Based on 2008 incidence rates, cervical cancer incidence rates are expected to increase by 90% by 2030 in most SSA sub-regions [3].

Literature provides evidence on the socio-cultural, socio-economic and biological factors that influence the distribution of the necessary cause of cervical cancer (i.e. Human Papillomavirus HPV) in SSA [2, 4]. Early and polygamous marriages, high parity, poverty, poor hygienic conditions, poor nutritional status and infections such as Human Immunodeficiency Virus (HIV) that compromise the immune system have been identified as risk factors of acquiring HPV in SSA [1]. The high incidence rates of cervical cancer in developing countries have also been associated to limited effective HPV screening [5]. However, few studies have focused on the role of infectious diseases such as schistosomiasis that are prevalent in SSA on the incidence rates of cervical cancer currently observed in the region [6–10].

Schistosomiasis is a neglected tropical disease that affects approximately 207 million people worldwide with more than 90% of cases living in SSA [11]. Schistosomiasis has been associated with cancer of the bladder [12]. It has been

reported that 0.3% of all cancers due to infectious agents were attributed to urogenital schistosomiasis in 2012 [13]. However, its role on the incidence of cervical precancerous lesions and cancer is still undefined.

Currently, literature from observational studies on the role of cervical schistosomiasis in cervical precancerous lesions and cancer is contradictory. The hypothesis of a causal relationship between cervical schistosomiasis and cervical cancer was suggested in 1970 [7]. Subsequent, cross-sectional studies conducted in Tanzania, Ghana and Togo rejected this hypothesis [6, 8, 10]. On the other hand, a cross-sectional study in rural Tanzania and a cohort study in Zimbabwe suggested that cervical schistosomiasis may provide a favourable environment for high-risk HPV infection [9, 14].

The existing contradiction in the literature on the relationship between cervical schistosomiasis and cervical cancer can be linked to the limitations of the methodology used when assessing this relationship. So far, studies have implemented methodologies that require laboratory testing and manual review of unstructured patient records (e.g. pathological reports). Therefore, they often use a small sample size which may be prone to bias in the selection of study participants and non-representativeness. Where large sample sizes are involved, errors associated with a manual review of records may bias the findings of the study. Thus, these limitations may lead to different conclusions in various studies. Moreover, the cost of implementing such methodologies may limit the rate at which literature on the relationship between cervical schistosomiasis and cervical cancer is provided. Therefore, there is a need to explore less resource-intensive and efficient methodologies such as text-mining of electronic histological reports to assess the relationship between cervical schistosomiasis and cervical cancer.

Text-mining is a discovery process that is used to automatically extract important knowledge from narratives such as admission histories, discharge summaries, reports of physical examination, radiology and pathology reports [15]. These

reports are rich sources of data that could be used to improve the quality and cost of healthcare; describe disease patterns and provide information for effective disease surveillance [16]. However, the free text in the report does not enable quantitative analysis to draw all these benefits. Therefore, key information needs to be efficiently extracted from these reports for quantitative analysis that will strengthen health systems. When human coders are performing this task, they are required to read each record to extract relevant information. This is a time-consuming process and might require many coders depending on the size of narratives. This might lead to delays in providing key evidence for policy making. The automation involved in text-mining reduces the time and number of human coders required to extract information from narratives. Thus, text-mining is a cost-effective alternative to human coders when analysing large numbers of narratives.

Studies have demonstrated that text-mining can be used to efficiently extract information from pathological and radiological reports. In cancer research, text-mining has been successfully used to extract information on clinical description of malignancies, Tumour-Nodes-Metastasis classification, International Classification of Disease in Oncology (ICD-O3) classification, cancer death classification and high-risk HPV types from clinical narratives [17]. However, the application of text-mining techniques in the extraction of information on cervical precancerous lesions, cervical cancer and schistosomiasis research has not been explored.

1.2 PROBLEM STATEMENT

Studies have demonstrated that HPV is a necessary cause for cervical cancer. Despite the global distribution of HPV, SSA currently has the highest incidence and fatality rates of cervical cancer in the world. Literature suggests this burden could be associated with the high prevalence of cervical schistosomiasis in the region. However, to date there is limited evidence on this relationship and the data are

contradictory. This could be due to the cost of implementation of current methodologies and their weaknesses. Thus, there is a need to explore efficient methodologies like text-mining in producing valid findings on the role of cervical schistosomiasis on the burden of cervical cancer in SSA.

1.3 MOTIVATION

The incidence rate of cervical cancer in SSA is expected to increase by 90% by 2030 if nothing is done [3]. Previous studies assessing the role of cervical schistosomiasis on the burden of cervical cancer in SSA have been expensive, time-consuming and produced results with limited generalizability. Thus, it is crucial to explore the use of less resource-intensive methodologies such as text-mining to assess the role of cervical schistosomiasis on the burden of cervical cancer in SSA. Text-mining can provide timely evidence to inform current efforts to address the burden of cervical cancer in SSA. Furthermore, the National Health Laboratory Service Corporate Data Warehouse (NHLS-CDW) provides a unique opportunity to access national electronic cervical histology reports in South Africa.

1.4 RESEARCH AIM

The aim of this study was to investigate the presence of cervical schistosomiasis in patients with cervical precancerous lesions and cancer using text-mining techniques on pathology reports available at NHLS.

1.5 SPECIFIC OBJECTIVES

To achieve this aim, the following objectives applied;

1. Establish the performance of text classification techniques in the study of cervical precancerous lesions, cancer and schistosomiasis.

2. Determine the yearly prevalence of cervical schistosomiasis in patients with precancerous and cervical cancer lesions from 2011 to 2017.

3. Describe the patterns, trends and distribution of cervical schistosomiasis in patients with cervical cancer and precancerous from 2011 to 2017.

1.6 DEFINITION OF KEY TERMS

Regular expressions: These are specially encoded text strings used as patterns for matching sets of strings [18].

Features: also known as attributes are individual measurable properties or characteristics of a phenomenon being observed [19].

Classes: This can be considered as the groups to which various data points in a dataset belong. They are also called labels or categories [20].

Classification: This is the process of predicting the class/category of a data point [20].

Classifiers: These are algorithms that learn the relationship between features and classes and uses this knowledge to assign a class label to a data point [21].

Confusion matrix: This is a matrix showing the predicted and actual classification [22].

1.7 OUTLINE OF RESEARCH REPORT

This report is organised as follows. Chapter 2 presents the literature that underpins the relationship between cervical schistosomiasis and cervical precancerous lesions and cancer. The literature on the implementation of

text-mining techniques to address public health problems is also presented in this chapter. Chapter 3 discusses the text-mining techniques and statistical techniques applied in this study. Chapter 4 presents the findings of this study while Chapter 5 discusses the interpretation and limitations of these findings. Chapter 6 provides conclusion and recommendations.

2

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter provides an overview of the literature on cervical precancerous lesions, cervical cancer and cervical schistosomiasis as well as their possible relationship. It also presents the literature on the implementation of text-mining to address public health issues and current gaps in the literature.

2.2 CERVICAL PRECANCEROUS LESIONS AND CANCER

Cervical cancer is the second most common cancer in developing countries [1]. It is preceded by precancerous phases characterised by a series of abnormal changes in the cervical cells. A three-tiered classification was previously used to denote the severity of these changes; Cervical Intraepithelial Neoplasia (CIN)1 was used for mild changes, CIN2 for moderate changes and CIN3 for severe changes and carcinoma in situ [23]. This classification was revised into a two-tiered classification by the Bethesda System for Reporting Cervical Cytology (TBS) for cytology specimens in 1988 and was adopted by the World Health Organisation (WHO) in 2014 [24]. In this new classification, CIN1 is classified as Low-grade Squamous Intraepithelial Lesions (LSIL) while CIN2 and CIN3 are classified as High-grade Squamous Intraepithelial Lesions (HSIL).

Cervical precancerous lesions and cancer have been linked to persistent infection with HPV. HPV is a group of sexually transmissible viruses with about 100 types. Thirteen types of HPV are known carcinogens, with two (HPV16 and HPV18) being responsible for 70% of cervical precancerous lesions and cancer cases. Risk factors

for contracting HPV include early sexual intercourse, multiple sexual partners, tobacco use and suppressed immune system due to HIV infection and prolonged use of steroid contraceptives [25, 26].

Cervical infection with HPV might remain stable, regress to undetectable levels or lead to an LSIL within months or years with few cases progressing to HSIL and cervical cancer [25]. A study among adolescent and young women between the age of 13 and 22 years conducted in San Francisco-USA showed that 40% (n=207) of women infected with HPV develop LSIL within 40 months of follow-up. However, approximately 70% of these women regressed to undetectable levels within 20 to 30 months and only 5% of women developed HSIL within two years [27]. A meta-analysis of 36 studies suggests that approximately 50% of CIN2 cases will also regress after two years with only one fifth progressing to CIN3 or cervical cancer [28]. Furthermore, a cohort study conducted in Denmark showed that the probability of progressing to CIN3 or worse in 12 years after infection with HPV16 was 26.7% and 19.1% following infection with HPV18 [29].

Despite studies showing high HPV infection clearance rates, rural communities of SSA currently have the highest incidence and fatality rates of cervical cancer [2]. These incidence and fatality rates have been associated with limited access to health education, HPV vaccination, inadequate HPV screening and care in advanced stages of cervical cancer [5, 25].

2.3 SCHISTOSOMIASIS, CERVICAL PRECANCEROUS LESIONS AND CERVICAL CANCER

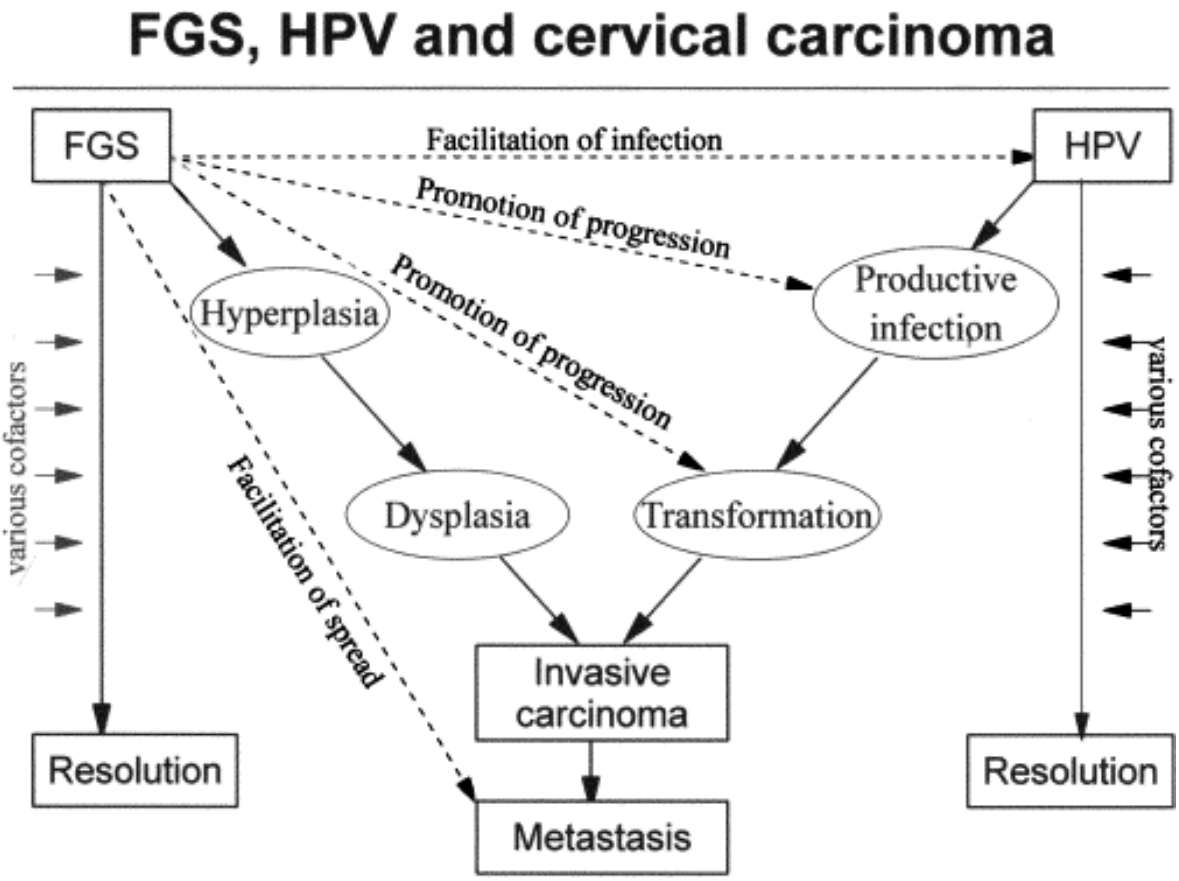
Schistosomiasis is a neglected tropical disease contracted when larvae of a Schistosoma genus released by freshwater snails penetrate the skin. There are five species of Schistosoma with three predominantly present in Africa, namely, *Schistosoma mansoni*, *Schistosoma guineensis* and *Schistosoma haematobium* [30]. Upon penetration into the body, the larvae travel through venous circulation to the

lungs, heart and liver where they develop into adult worms. The adult worms then migrate in pairs to various mesenteric venules (depending on the species). The male and female adult worms copulate at this location. The female adults then lay eggs which are transported to various organs where they get trapped and/or eliminated [31].

S. haematobium has been identified as the main Schistosoma specie involved in urogenital schistosomiasis [30]. Their adult worms migrate to the vesicular and pelvic venous plexus of the bladder and their eggs are often transported to the bladder, ureters and reproductive organs [31]. In some instances, *S. mansoni* and *S. japonica* whose ova have a predilection for large and small intestines respectively, are also deposited in the female genital organ [32, 33]. Female Genital Schistosomiasis (FGS) is characterised by the deposit of Schistosoma ova in the female genital organ. It mostly affects the cervix and vagina [34]. The clinical features of FGS were first described in an Egyptian woman in 1899 [35]. Since then, there have been growing evidence of its negative impact on the female reproductive organ [36, 37]. Schistosoma ova often causes moderate to severe inflammation, ulcerative lesions and scars in the areas surrounding the affected part that may cause infertility, abortion, ectopic pregnancy, weakened local immune system etc. [38].

Studies have shown that *S. haematobium* is the carcinogenic agent involved in most bladder cancers [12, 39]. However, its role in cervical precancerous lesions and cancer is not yet defined. Figure 2.1 illustrates the hypothesised pathways through which FGS can affect the natural history of cervical cancer.

Youssef *et al.* suggested a causal relationship between schistosomiasis and cervical precancerous lesions/cancer [7]. The cervical tissues infected with non-viable schistosoma ova usually shows a fibrous connective reaction described as scars [38]. Youssef *et al.* viewed the extensive epithelial scaring associated with cervical schistosomiasis as a predecessor to cervical precancerous lesions and



Source:[40]

Figure 2.1: Hypothesised relationships between schistosomiasis and cervical cancer

cancer [7].

Other researchers have suggested that schistosomiasis can alter the natural history of HPV and HIV [34] which can lead to the high burden of cervical cancer observed in developing countries. The ulcerative lesions induced by schistosomiasis enhance the acquisition of HPV [34]. Moreover, schistosomiasis weakens the cervical local immune systems reducing the ability of the cervix to clear HPV infections [41]. This may lead to persistent infection with HPV and increased local HPV viral load which will assist in the rapid development of HPV induced lesions [42].

A recent study suggested that infection with schistosomiasis can reduce the impact of the HPV vaccine in preventing cervical cancer [43]. Gent V. *et al.* studied

the effect of chronic infection with *S. mansoni* on the antibodies induced by the HPV vaccine was examined in animals [42]. Findings showed that animals infected with *S. mansoni* developed HPV-specific IgG antibodies at a lower rate compared to the group without the infection. This rate was relatively higher among those who received Praziquantel compared to those who did not receive any treatment. While the antibodies developed over six weeks persisted in the group without the infection and the treated infected group, it decreased in the infected group without any treatment for schistosomiasis. Thus, reducing the long term protective effect of the HPV vaccine against cervical cancer.

2.4 EPIDEMIOLOGY OF CERVICAL SCHISTOSOMIASIS AND PRECANCEROUS LESIONS AND CANCER

The role of cervical schistosomiasis on the burden of cervical precancerous lesions and cancer is contradictory. A study conducted in Egypt in 1970 reported a causal relationship between cervical cancer and precancerous lesions [7]. The researchers retrospectively assessed the laboratory reports of 121 females with cervical cancer from two hospitals in Egypt. They found out that 15 cases were infected with cervical schistosomiasis. Based on these findings, it was concluded that cervical schistosomiasis might lead to cervical cancer. A subsequent review of a larger sample of patient records in Tanzania and Togo identified a 1.7% and 1.9% prevalence of cervical schistosomiasis respectively. Thus, they concluded that there was not enough evidence to support the association between cervical schistosomiasis and cervical cancer highlighted in the earlier studies [8, 10]. A cross-sectional study and a cohort study were conducted in Tanzania and Zimbabwe respectively to further investigate this association [8, 14]. In these studies, samples were collected from females living in rural areas and tested for HIV, HPV, *S. haematobium* and precancerous lesions. The Tanzanian study hypothesised an association between persistent high-risk HPV and cervical schistosomiasis [9]. However, the study in Zimbabwe demonstrated that this association disappears after 5 years of follow up [14]. Nonetheless, both studies

concluded that damages to genital tissues and weakening of the local immune system induced by FGS might favour a persistent infection with HPV.

2.5 LIMITATIONS OF CURRENT METHODS

The extensive use of laboratory testing and manual review of records might explain the differences in current findings of the relationship between cervical schistosomiasis and cervical cancer. Laboratory testing generally involves the cost of buying reagents to carry out the required tests within a specific population. Most studies using this approach had a sample size of less than 40 study participants [6, 14, 44] with only one having a sample size of 218 participants [9]. However, they may not be representative of the study population and may lead to variation in results and validity. Moreover, where studies have a larger sample size, manual human coding is the general method used in the review of records before analysis [7, 8, 10]. This is time-consuming, requires extensive human input, and is prone to many errors [45, 46]. Therefore, text-mining can be used to overcome the above-stated limitations and improve the quality of results.

2.6 TEXT-MINING IN HEALTH

Text-mining techniques have been widely used in the field of public health to extract information from clinical narratives that are relevant to understand disease patterns. They have also been used to classify and retrieve medical records and establish surveillance systems. A recent application of text-mining in public health includes the development of the Clinical History Extractor for Syndromic Surveillance in Singapore [47]. A keyword-based algorithm was used to automatically extract 48 signs of respiratory infections, gastrointestinal infections, constitutional and symptoms associated with other infectious diseases from patients' records. It had an overall accuracy of 97.6% when compared to a human classifier. The integration of such systems in the provision of health care service

will provide a reliable surveillance system to efficiently monitor and treat respiratory and gastrointestinal infections in the region.

Furthermore, text-mining techniques have been used in pathology studies for auto-coding, case detection, critical value detection, de-identification, negation and extracting structured information over the past 25 years [17]. The classification of 5,121 pathology reports from 35 pathologists into cancer cases and non-cases in France shows an application of these techniques in cancer research [48]. In this application, text-mining achieved a specificity of 96.7% for topography and morphology in the classification of pathology reports based on the International Agency for Research (IARC) guidelines. However, it had lower performance on ICD-03 classification (topography specificity 75.1%, morphology specificity 85.4%). Nonetheless, this performance could be improved by increasing the sample size, carefully selecting the variables to be considered and techniques to be used in various text-mining processes.

2.7 GAP IN THE LITERATURE

Despite the efficiency and the cost-effectiveness of text-mining in extracting information from pathology reports, very little attention has been given to its use for determining the effect of cervical schistosomiasis on the burden of cervical cancer in SSA. Thus, this study explored the use of text-mining techniques to extract and analyse information on cervical schistosomiasis from cervical histology reports.

2.8 CONCLUSION

This chapter presented the current literature on cervical precancerous lesions/cancer and cervical schistosomiasis. it also discussed the possible pathways through which schistosomiasis can cause or favour the progression of

precancerous lesions as well as epidemiological knowledge on the association of cervical schistosomiasis and precancerous lesions/cancer. This chapter highlighted limitations in methodologies currently used to explore the impact of schistosomiasis on the burden of cervical cancer in SSA. Furthermore, it provided a review of the use of text-mining techniques to strengthen health systems. The next chapter will present detailed steps on how text-mining techniques were implemented to convert histology reports into a structured dataset and quantitatively analysed.

3

METHODOLOGY

3.1 INTRODUCTION

This chapter describes the study population, the data source and data cleaning. It also describes the text-mining methods used to convert cervical histology reports from histology reports to a structured dataset. The data management, descriptive methods and ethical considerations are also described in this chapter.

3.2 STUDY POPULATION, DATA SOURCE AND DESCRIPTION

3.2.1 *Study Population*

This study involved all cervix related histology reports uploaded to the NHLS Corporate Data Warehouse (CDW) database by Inkosi Albert Luthuli Central Hospital from 2011 to 2017. Inkosi Albert Luthuli Central Hospital is a public/private hospital situated in Durban, KwaZulu-Natal (KZN). KZN is the province situated in the south-eastern part of South Africa. The prevalence of schistosomiasis in school-going children in this province varies between 37.5% and 56% [49, 50]. Between 2013 and 2014, KZN had the highest prevalence of HSIL and the second-highest prevalence of LSIL in South Africa [51]. Moreover, cervical cancer is the most common cause of cancer death among women in this province [52].

3.2.2 *Data Source*

The datasets used in this study are extracts from the NHLS central CDW. NHLS is the largest provider of diagnostic pathology in South Africa. It supports the public health sector by providing diagnostic services to approximately 85% of South Africans; as well as evidence for effective disease surveillance. NHLS achieves this through its national network of more than 260 laboratories [53]. Each laboratory has a laboratory information system (LIS) into which the demographics, test, and results of each patient is captured. These data from all laboratories in the NHLS network are automatically fed into a central corporate data warehouse in the form of reports [54]. These reports are often semi-structured with the results sections often presented as narratives (see Appendix B for a sample report). This corporate data warehouse provides a rich source with pathology reports from which information could be extracted to investigate disease patterns, trends and associations.

3.2.3 *Dataset Description*

NHLS provided cervical histology laboratory reports from 2011 to 2017 as four separate datasets. Each dataset consisted of seventeen variables. There was a unique identifier for patients and a unique identifier for each histology sample. Patients demographic information were captured in nine variables. Histology information was captured in six variables among which one had the detailed histology report as narratives (see Appendix C for list of variables).

All the four datasets had the same demographic information and unique identifiers but different information on the SNOMED-CT. SNOMED-CTs are clinical terms that have been translated into codes to enable the exchange of healthcare information globally [55]. In these datasets, the SNOMED-CT defined the specimen site (topography), the procedure performed and the diagnosis (morphology and

diagnosis). Each dataset contained information on one SNOMED-CT only i.e. there was a dataset with topography SNOMED-CT only, one with the procedure, one with morphology and another with the diagnosis. The number of records in each dataset varied given that all records did not have all four SNOMED-CT. Only 50% of records had a diagnosis in terms of cervical precancerous lesions, cancer and other cervical pathologies.

The methodology sub-sections above presented the population, data source and datasets from which information on cervical precancerous lesions, cancer and schistosomiasis was extracted and analysed. The following sub-sections will describe the methods applied to extract and analyse information from the dataset described above. Figure 3.1 presents an overview of the methods applied in this study.

Two main techniques were applied; text-mining and data analysis

3.3 TEXT-MINING

Text-mining is a process within which, Natural Language Processing (NLP), data mining and machine learning (see 3.3.2) techniques are used to analyse large amounts of texts to extract and summarise important information [15, 17]. NLP is the application of computational techniques to analyse and synthesizes naturally occurring text to achieve human-like processing in certain tasks or applications [56]. Text-mining was applied in this study to assign diagnosis to records without any and extract information on the precancerous lesions grade, cancer type, HPV, HIV and schistosomiasis status. This was achieved through the set of iterative steps as presented in Figure 3.2.

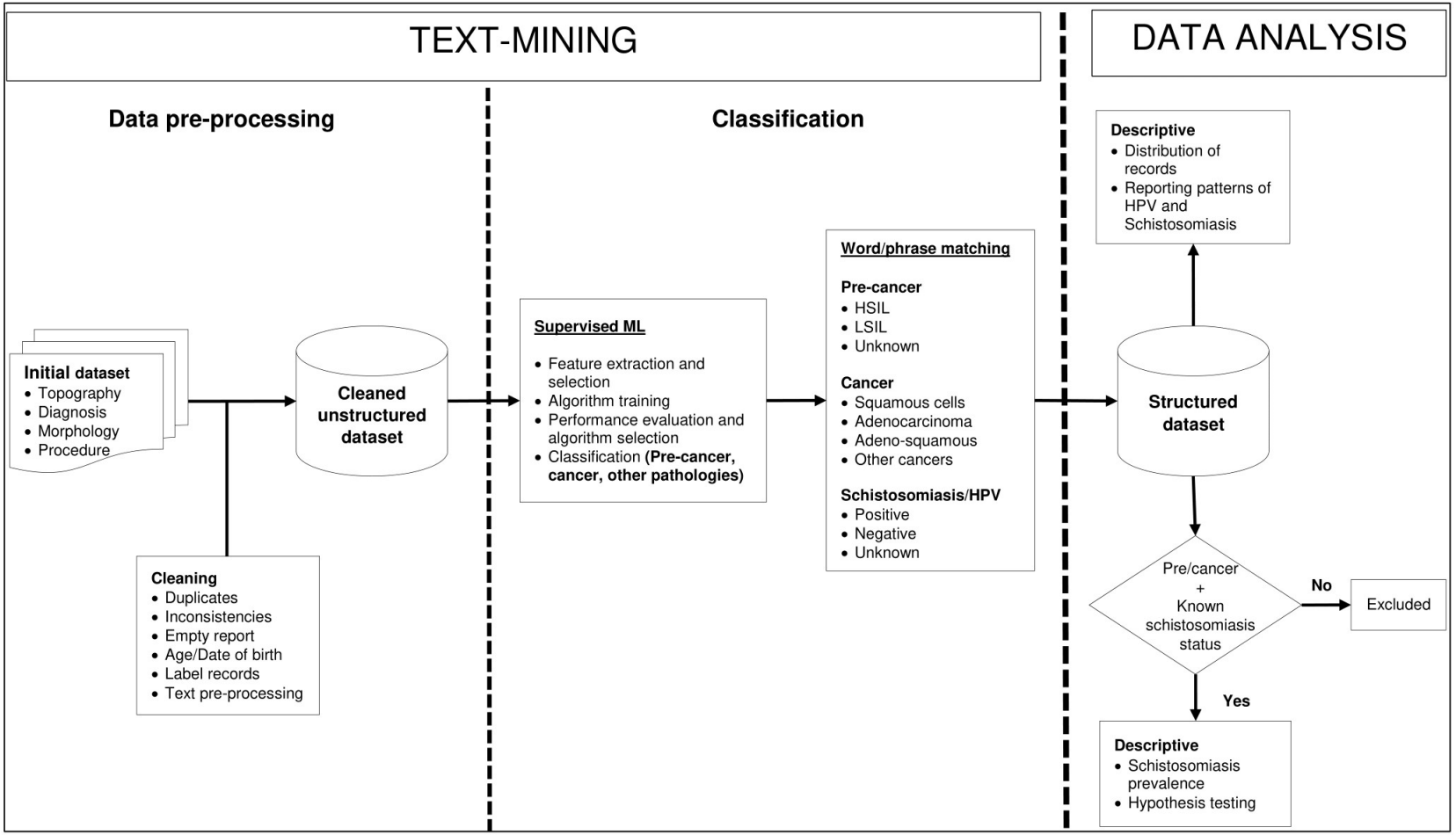


Figure 3.1: Overview of the Methodology

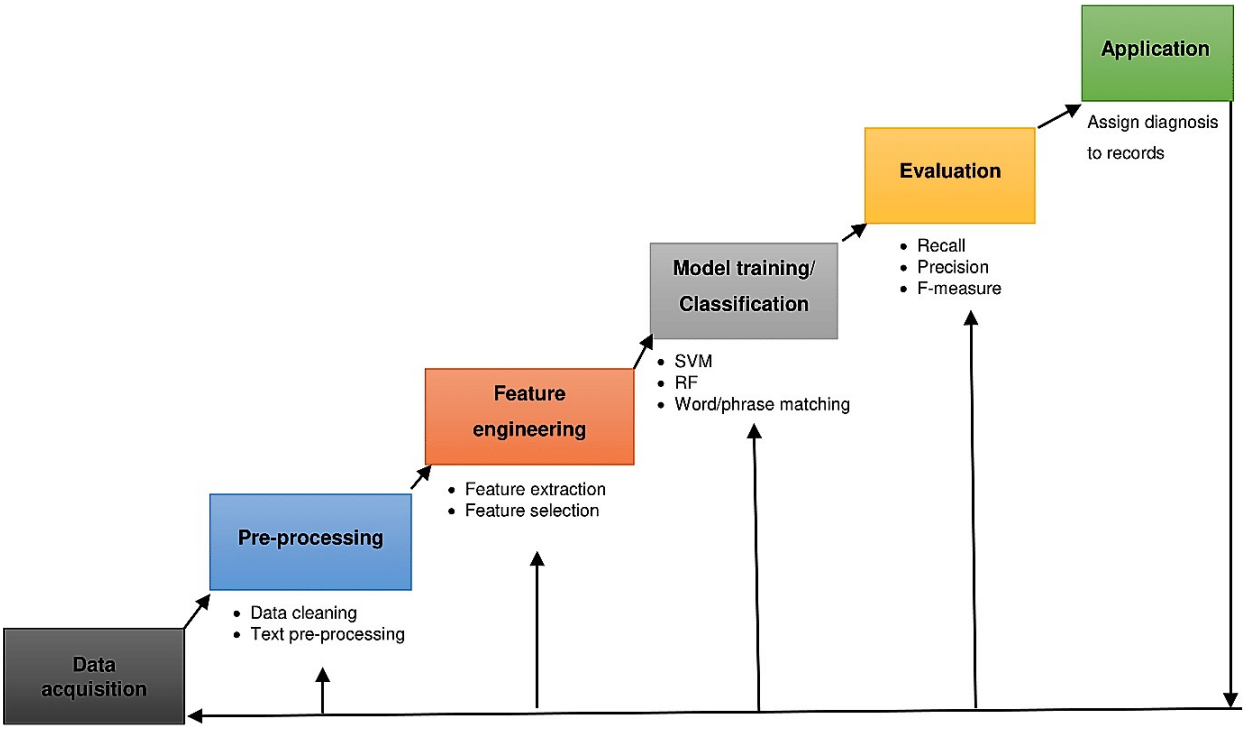


Figure 3.2: Text-mining process

3.3.1 *Pre-Processing*

Pre-processing consisted of eliminating duplicate records, dealing with implausible date of births and age, correcting inconsistent information and integration of the four datasets provided by NHLS.

Duplicates removal

There were 6 (0.02%), 12,866 (34%), 23,800 (98%) and 4,806 (46%) duplicates records in the topography, procedure, morphology and diagnosis datasets respectively. These duplicates were due to SNOMED-CTs (*"snomed_code"*) with more than one definition (*"snomed_name"*). An example of this is illustrated in Figure 3.3below.

snomed_code	snomed_name
D7-71514	CHRONIC CERVISITIS (DISORDER)
D7-71514	CHRONIC CERVISITIS

Figure 3.3: Example of SNOMED-CT definition variations

Most records in the procedure, morphology and diagnosis datasets were duplicated to accommodate all these variations in SNOMED-CT definitions. Also, there were few records with a variation the test methods names. This variation was the main source of duplicates records in the topography dataset.

To eliminate duplicates, the SNOMED-CT definitions and test method name variables were eliminated from the main dataset and duplicates eliminated based on the remaining variables. The test method was permanently eliminated as it was not an important variable of interest in the study. On the other hand, the SNOMED-CT definitions were standardised by giving only one of the corresponding names and merging them back to the main datasets.

Cleaning date of birth and age

Approximately 10% of records had their date of birth before 1900 and they were set to missing. Regular expressions were used to extract patients' age from the detailed histology reports. Regular expressions are specially encoded text strings used as patterns for matching sets of strings [18]. In other words, a sequence of symbols and characters were used to find and extract age captured by the pathologist in the histology report. In cases where the age was not mentioned in the histology report or the patients were less than ten years old, the age was calculated from the date of birth and the report date.

Inconsistent information

Under normal circumstances, the dataset on cervical pre-cancerous lesions and cancer should only have female patients. However, approximately 419 (1%) of patients were captured with a male or unknown sex. The manual review of fifty

records was used to confirm that these were reports related to the cervix. After confirmation, all records were considered to belong to female patients.

Integration of datasets

The four datasets were appended and cleaned through the procedures described above. The SNOMED-CT codes were separated from the main dataset. All morphology and diagnosis codes corresponding to each histology sample were combined in one cell before merging back to the main datasets using the histology sample unique identifier. Duplicates created by appending all datasets were excluded. This approach was motivated by the fact that certain histology samples had supplementary reports that could be found in any of the four datasets provided. Thus, appending all datasets before excluding the duplicates limited information loss which could be associated with extracting the SNOMED-CT from other datasets and merging them to one dataset.

Data labelling

Records with a diagnosis and/or morphology SNOMED-CTs were grouped into three main pathology groups; cervical precancerous lesions, cancer and other cervical pathology. The precancerous lesions group was made up of all HSIL, LSIL and carcinoma in-situ. The cancer group was made up of all cancer types and any other record that was not a precancerous lesion or cervical cancer was grouped under other cervical pathologies. This is because pathology reports in various groups had similar terms in their histology report. Moreover, further disaggregation into smaller groups (various cervical precancerous lesions and cervical cancers) could negatively impact the accuracy of the classification process [57].

Each pathology group was assigned a numerical value as shown in Table 3.1. The numerical equivalents of various groups were used as labels to train a multiclass classified.

Table 3.1: Labels and numbers

Category Name	**Numerical equivalent**
Other	0
Pre-cancer	1
Cancer	2

The distribution of each pathology group in the labelled dataset was explored to identify any need to oversample or under-sample records in certain pathology groups. It also guided the selection of the appropriate algorithm evaluation metric.

Text pre-processing

Text pre-processing is an important process before feature extraction. It enables the removal of features that are not necessary for the classification e.g. (punctuation marks, stop words, etc.) [58]. Features also known as attributes are individual measurable properties or characteristics of a phenomenon [19]. In this study, text pre-processing involved converting all text into lower case, removal of line breaks, extra space between words, punctuation marks, special characters and stop words, unnecessary information as well as the expansion of contracted words (short words made by putting two words together e.g. can't).

Regular expressions were used to remove line breaks, extra space between words, laboratory addresses and section headings. Subsequently, the histology report was broken up into individual words called tokens (tokenisation) and all special characters were removed. A JavaScript Object Notation (JSON) file with one hundred and seventeen contracted English words and their expansions were used to expand contracted words in the histology reports. Stop words were removed using a modified list of stop words provided in the Natural Language Toolkit (NLTK) package of Python 3.7. Words like *"no"* and *"not"* were excluded from the list of stop words proposed by NLTK as they are key in identifying a

negative diagnosis.

Visualisation was applied using word cloud and density plots before and after text pre-processing. They provided visuals of the key changes in the contents of the histology reports after cleaning.

Feature extraction

Classification algorithms are unable to read and interpret narratives like human beings. They rely on the representation of narratives as numerical values for interpretation and classification of reports [59]. Feature extraction also known as vectorisation or indexing is the process of transforming text data into a numerical vector space [60]. This was achieved through the N-grams and Term Frequency-Inverse Document Frequency (TF-IDF) metric. N-gram is a method where n-number of words are extracted in the order in which they occur in a text to generate features that are used to represent a text [58]. TF-IDF is a metric that measures how important these features are to a document in relation to a group of documents [61]. It is calculated as

$$TF - IDF(t, d) = TF(t, d) \times \log(\frac{N}{DF(t)}) \tag{3.1}$$

Where

- *t*: term (i.e. a word in a document)
- *d*: document
- *TF(t)*: term frequency (i.e. how many times the term *t* appears in the document *d*)
- *N*: number of documents in the corpus
- *DF(t*: number of documents in the corpus containing the term *t*

In other words, the histology reports (documents or records)were broken down into single words (uni-gram) and groups of two adjacent words (bigrams). This

method was used to consider the order of words in the report text before weighting. Each feature was allocated a weight based on their frequency of occurrence in each report compared to their occurrence in the whole dataset. Features that occurred in less than 2 reports and in more than 95% of all the reports were excluded. Higher weights were assigned to features with a low frequency of occurrence in all reports and features that were frequent in the corpus were assigned lower weights. The end product of this process was a term-document matrix with $54,629$ features. Each row represents a report and each column represent a feature. The data points contained the weight of the feature in each report.

Feature selection

Feature selection is a process where some criteria are used to choose the optimal subset of features from the original features of the dataset. Through feature extraction, texts or documents are presented as a bag of words and/or n-words. This representation generally leads to a high dimensionality feature space with redundant features and sparsity (features with mostly missing values) [62]. Training a model with such features can drastically reduce the accuracy of the model. Thus, it is important to construct a new vector space with features that will enable the model to reach its highest possible accuracy [63]. This was achieved through Latent Semantic Analysis (LSA).

LSA is a dimensionality reduction method similar to principal component analysis. It applies Singular Value Decomposition (SVD) to construct concepts based on word similarity [64]. SVD takes as input the dataset produced during the feature extraction process. This dataset is decomposed into three matrices. Two orthogonal matrices that measure the document-to-concept similarity and the feature-to-concept similarity and a diagonal matrix that measures the strength of each concept. The product of these matrices is subsequently computed to identify the similarity between words and built relevant concepts. The output of SVD is a document-to-concept matrix. Thus, reducing the number of features in the dataset,

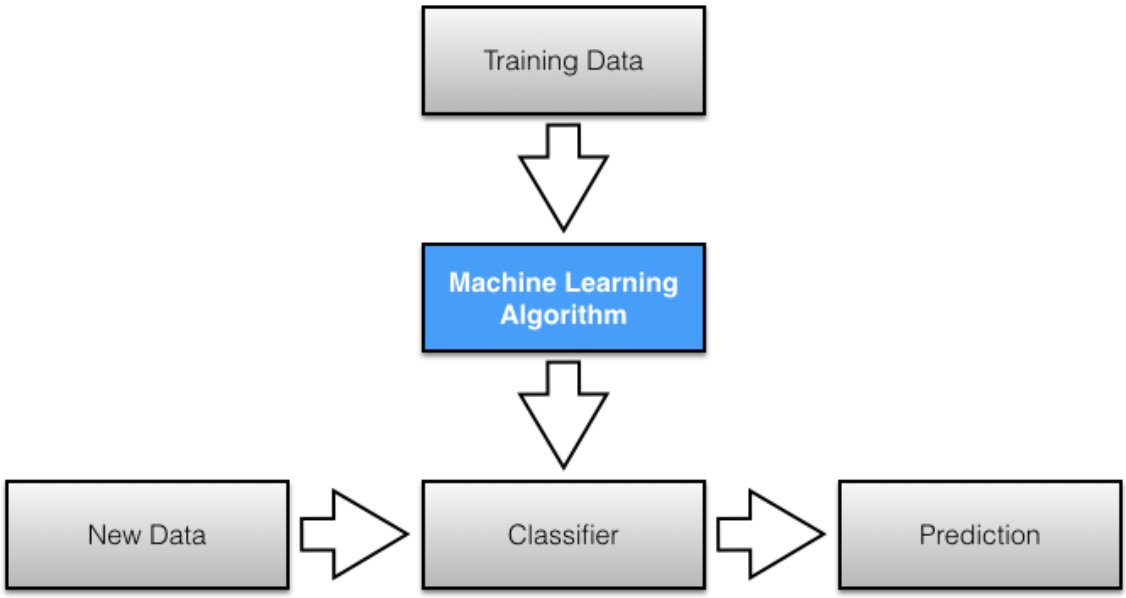
a process known as dimensionality reduction.

Although LSA considerably reduces dimensionality, it may not be low enough to produce optimal model performance. In this study, the optimum number of features to be included in the model was determined manually after LSA. First, the classifiers were trained with 20 features and their F-measure were assessed. The number of features used to train the classifier was gradually increased until the point where the performance started dropping. The number of features included just before the performance dropped was considered as the optimum number of features required to train the classifier.

3.3.2 *Classification of Cervical Precancerous Lesions, Cancer and Other Cervical Pathologies*

Classification is the process of automatically assigning documents to predefined target/outcome categories based on their features [20]. This can be achieved using a rule-based or a learning approach. In a rule-based approach, a set of rules are defined and integrated into an algorithm to classify records into specific categories. The learning approach requires the algorithm to learn patterns from the data through a training process and uses this knowledge to perform classification tasks on future data. In cancer research, the rule-based approach is the preferred approach of classification among clinicians. This involves manually defining the logical rules that are used to classify documents into various categories. However, this approach is time-consuming and requires the input of domain experts. On the other hand, Machine Learning (ML) approaches reduce these efforts [17].

ML is a part of artificial intelligence where algorithms learn from data and make predictions on new data without being explicitly programmed to do so [57]. Figure 3.4 displays the machine learning process.



Source:[65]

Figure 3.4: Machine Learning process

When dealing with narratives or free text, the algorithm is programmed to learn the characteristics of documents and use them to classify records into specific groups [17]. This can be achieved through supervised, unsupervised and semi-supervised methods. Supervised machine learning is when the algorithm learns with the help of a training dataset where the outcome or the targeted class has been defined by humans i.e. a labelled dataset. The labels enable the algorithm to learn from the important characteristics in each group and uses these characteristics to classify new records under various groups. In contrast, unsupervised machine learning algorithms are trained using an unlabelled dataset (the outcome/target is not defined). The algorithm has to learn from characteristics of the dataset and group records/observations with similar characteristics together. Semi-supervised machine learning is a hybrid approach. The training dataset is made up of both labelled and unlabelled records. This is generally used when there is a large training dataset without labels. Labels are assigned to a small portion of the dataset and both label and unlabelled records

are used to train the algorithm and use in prediction [57].

The datasets provided by NHLS were made up of records with and without a diagnosis in terms of cervical precancerous lesions, cancer and other cervical pathologies. Therefore, the purpose of the classification at this stage was to assign a diagnosis to records without a diagnosis. Unsupervised ML was inadequate as some records had labels and there was no need to discover new groups in the dataset. Semi-supervised ML was inadequate as well as the aim was not to establish a classification system for future use but identify a classifier that will best assign a diagnosis to the records lacking a diagnosis in the dataset. Thus, supervised ML approach was used to achieve our classification goals at this point.

Supervised machine learning algorithms

There are many supervised ML algorithms that can be used for text classification. The learning capacity of each algorithm varies with each dataset. Thus, it is important to identify the algorithm which will provide a better fit for each dataset. The performance of scikit-learn dummy, SVM, and RF classifiers were explored on this dataset. Scikit-learn dummy classifier was used to assess the need for using trivial classifiers like SVM and random forest. SVM is the most frequently used in pathology studies [66]. On the other hand, RF is not frequently used in pathology studies, however, it has been identified as a promising method for text classification due to its simplicity and ability to handle high dimensional data [67]. Moreover, random forest algorithms are relatively faster to train and use in prediction [68].

Scikit-learn dummy classifier

This is a baseline classifier in the sci-kit-learn library. It applies simple rules to classify data points. For instance, the algorithm may always predict the class with a higher proportion of records in the dataset. The performance of the dummy classifier provides the baseline against which other classifiers can be

compared [69]. The higher the performance of such classifiers, the less need to implement trivial classifiers like SVM and RF.

Support vector machine

SVM classifier is one of the most popular machine learning algorithms. SVM solves classification problems by creating a decision boundary (hyperplane) between two categories in a multidimensional environment enabling the classification of labelled vectors. It creates a hyperplane that is as far away as possible from the closest data points (support vectors) in both categories [70].

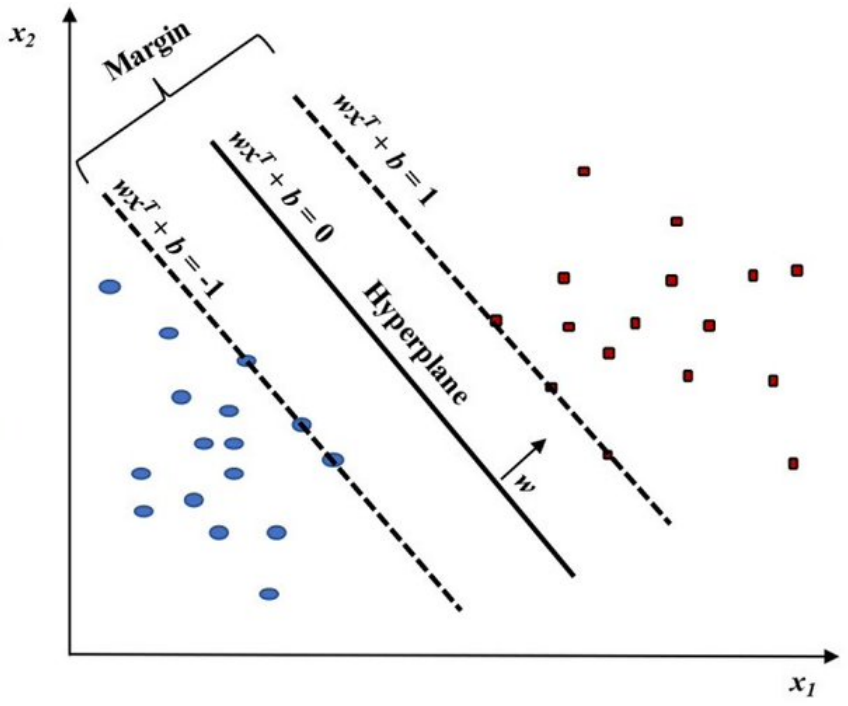


Figure 3.5: Linear SVM model classifying red and blue

For instance:

Given a labelled training dataset:

$$(x_1, y_1), \ldots, (x_n, y_n), \text{ so that } x_i \in R^d \text{ and } y_i \in (-1, +1) \tag{3.2}$$

where

x_i is a feature vector representation and

y_i the class label (negative or positive) of a training compound i.

The optimal hyperplane can then be defined as:

$$\mathbf{w}\mathbf{x}^T + b = 0 \tag{3.3}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input feature vector, and b is the bias.

The w and b would satisfy the following inequalities for all elements of the training set:

$$y_i = \begin{cases} 1, & \text{if } \mathbf{w}\mathbf{x}_i^T + b \geqslant +1 \\ -1, & \text{if } \mathbf{w}\mathbf{x}_i^T + b \leqslant -1 \end{cases} \tag{3.4}$$

Thus, we train the SVM model to find the w and b so that the hyperplane separates the data and maximizes the margin 1 divided by the Euclidean distance squared

$$1/\|w\|^2 \tag{3.5}$$

Vectors x_i for which $|y_i|$ $(\mathbf{w}{x_i}^T + b = 1)$ will be termed support vector (Figure 4.6). In cases where the categories are non-linear, an adequate kernel function is used to add more dimensions to raw data and convert it to a linear problem prior to classification using the SVM.

Random forest

RF is an ensemble classification approach i.e. classification by several basic classifiers and combining their weighted output for decision making [71]. RF uses a decision tree as the basic classifier. Decision tree algorithms process the data using a tree-like structure for classification. It consists of internal nodes that are made up of features, branches that are decisions (they depart from the nodes) and leaves that represents the class. Thus, structured queries run from the root creating internal nodes and branches based on predefined rules and ends with a

classification outcome [60]. The accuracy of the decision tree algorithm decreases with high dimensional data. But this can be improved by combining the output of decision trees built using random features. This is applicable through a random forest approach to classification where an object is independently classified by a set of independent but identical decision trees with random vectors, each decision tree cast a vote as to which class the object belongs to and this object is finally classified in the class with the maximum votes [72]. This is illustrated in Figure 4.7.

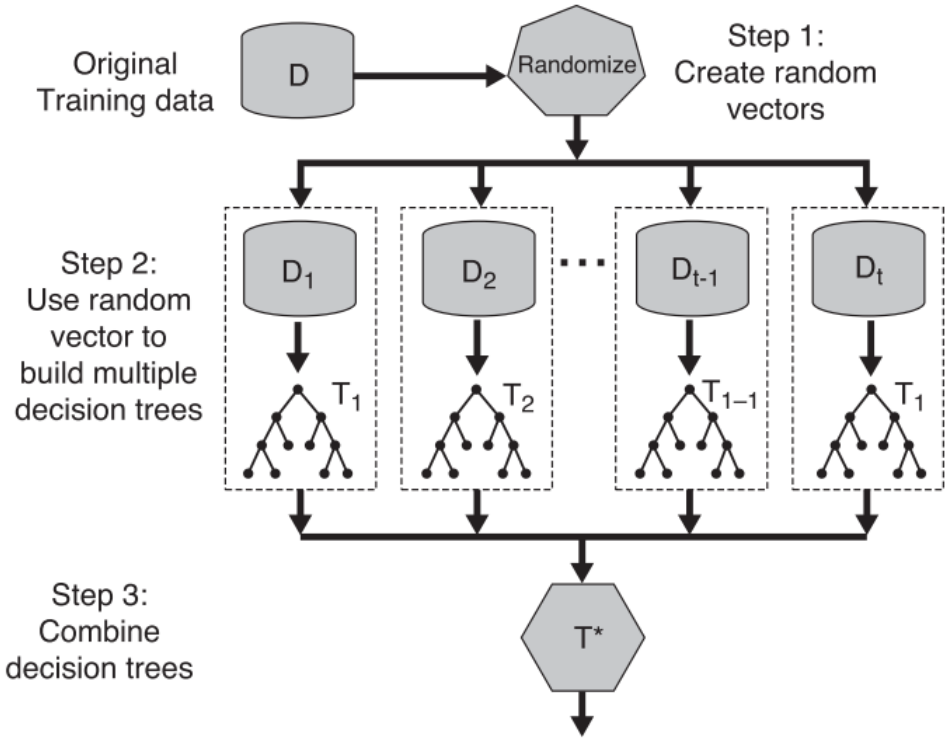


Figure 3.6: Random Forests

Source:[72]

Multiclass classifier

The records without diagnosis needed to be classified into three distinct groups where one record can only belong to one pathology group. This is known as a multiclass classification problem. Most baseline classification algorithms are developed to solve binary classification problems i.e. classify records into two classes. However, they can be extended to solve classification problems with more

than two classes by using multiple binary classifiers. Each binary classifier is trained to classify each class in the dataset independently and the class of a record is determined through a voting system of all the binary classifiers [73]. This is usually achieved by using OVO or OVA approach. In the OVO approach, binary classifiers are trained for each pair of classes such that C*(C-1)/2 classifiers are trained (C is the number of classes) and used for classification. On the other hand, the OVA trains binary classifiers for each class, where the one class is considered as the positive outcome and the remaining classes are considered as negative [73].

Hyperparameter optimisation

ML algorithms are made up of parameters and hyperparameters. Parameters are internal characteristics of the classifiers whose values are estimated from the data during the fitting process to reduce the algorithm error. In contrast to parameters, hyperparameters control how the algorithm work and cannot be determined from the data. They are considered as the settings of ML algorithms. For example, the number of estimators in a random forest. Classifiers have default values for hyperparameters [74]. Changing these default values can influence the performance of the algorithm. Thus, finding the optimal values for the hyperparameters can improve the predictive power of an ML algorithm [74].

Randomised search and Grid-search are two methods that are frequently used to find the optimal hyperparameters of an ML algorithm. Randomised search randomly selects a combination of hyperparameters in the provided grid to fit models and provide optimal hyperparameters based on sample combination. On the other hand, Grid-search identifies the optimal parameters of a model by building several models with various parameters that lie within a grid. It trains the model with a combination of all the parameters provided and assess the output of each parameter combination using a cross-validation approach and/or any other model scoring metrics. Grid-search is computationally expensive compared to randomised search and cannot be used to optimise all parameters [74].

Preparation of the training and testing set

The pre-processed data was split into training and test sets using stratified sampling. The training set included 80% (n=12,608)of the labelled dataset while the test set included 20% (n=3,152). Stratified sampling was to ensure a similar distribution of all the pathology groups in the training and the test set.

Machine learning algorithm training process

OVO and OVA algorithm with its default hyperparameters were trained and evaluated using the train and test dataset respectively. The performance of each algorithm was evaluated using recall, precision and F-measure. Random search to maximize the F-measure with 50 iterations and 5-folds cross-validation was used to identify regions where the algorithm performed well. Subsequently, values provided from the randomised search and their neighbouring values were used in a grid-search to maximize the F-measure that identifies optimal hyperparameters of the algorithm. The list of hyperparameters that were optimised are summarised in Table 3.2.

Table 3.2: Hyperparameters optimised

ML Algorithm	Hyperparameter	Description
SVM	C	Penalty parameter C of the error term
	Kernel	Specifies the kernel type to be used in the algorithm.
	gamma	Kernel coefficient.
	degree	Degree of the polynomial kernel function.
Random Forest	n_estimators	Number of trees in the forest
	max_depth	The maximum number of levels in each decision tree.
	max_features	Maximum number of features considered for splitting a node
	min_sample_leaf	Minimum number of data points allowed in a leaf node
	min_sample_split	The minimum number of data points placed in a node before the node is split.
	bootstrap	Method for sampling data points (with or without replacement)

Source: scikit-learn documentation

ML algorithms with the optimum hyperparameters were trained and evaluated.

3.3.3 *Performance Evaluation*

ML involves programming an algorithm to learn from existing data to make predictions on new data without being explicitly programmed [57]. Performance

evaluation measures are used to assess how well an ML algorithm has learned from the data. This involves passing new dataset through the algorithm after training it to predict the class to which it belongs. The predicted classes are compared with the actual classes of all test records and various measures are computed to provide the overall performance of the model. In this study, precision, recall and the F-measure were used to determine the performance of the text classifier. Based on a binary outcome where there are usually two classes for the outcome (positive or negative) these measures are defined and computed as follows:

1. Precision also is known as the positive predictive value, is the proportion of true positives given by the classifier.

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

2. Recall also referred to as the true positive rate or sensitivity, is used to measure the percentage of actual positives which are correctly identified by the classifier.

$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$

3. F-measure or specificity is the harmonic mean of precision and recall.

$$F - measure = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{3.8}$$

In a multiclass problem, the precision, recall and F-measure are calculated for each class (one class against all).

Figure 3.7 shows an example of how TP, FP, TN, FN when computing the above measures for one class in a multiclass classification. The average score for all classes is reported as the overall performance of the model. This can be a micro average or macro average depending on how the data is skewed. The micro average adjusts for the proportion of each class in the testing set while the macro average does not.

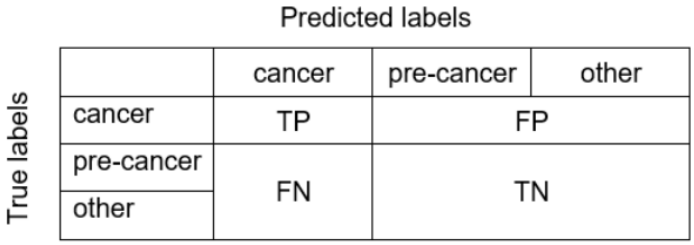


Figure 3.7: Example of TP, TN, FP, and FN when computing precision, recall and F-measure for one class (cancer) in a multiclass classification

A confusion matrix (3 by 3) was generated. The values provided by the confusion matrix were used to compute various performance measures for each class and macro averages given that the classes in the testing set were relatively balanced.

3.3.4 *Qualitative Evaluation of the Classification Error*

Some records misclassified by the best classifier were reviewed to understand the possible cause of misclassification by the algorithm.

3.3.5 *Application*

The ML algorithm with the best F-measure was used to classify all the records into cervical precancerous lesions, cervical cancer and other cervical pathologies. All records classified as other cervical pathologies were excluded from the analysis.

3.3.6 *Classification of Other Pathologies*

Word/phrase matching was used to further identify cervical precancerous lesions grades and cancer types after classification in records that were classified as cervical precancerous lesions and cervical cancer respectively. The HPV, HIV and schistosomiasis status of each record was also extracted using the same approach. Some records were explored to identify terminologies used by pathologists to report on these pathologies. Terms and patterns used to report positive and

negative cases were identified and integrated into regular expressions to extract information on all pathologies of interest.

Classification of precancerous lesions grade

A subset of the dataset made up of records classified as precancerous lesions by the ML algorithm was generated. Regular expressions were used to identify all cases of CIN3 and 2 as well as records with high grade squamous intraepithelial lesion. These records were kept as a separate dataset and removed from the main dataset with all precancerous lesions cases. Subsequently, CIN1 and Low grade squamous intraepithelial lesion were also extracted and stored in a separate dataset. Approximately 10% of the remaining records were reviewed manually. Based on the findings of the manual review process these records were classified as having an unknown precancerous lesions grade. All the datasets created were appended to recreate the initial dataset with a column to identify the precancerous lesion grade.

Classification of cancer types

A subset of the dataset made up of cancer records classified by the ML algorithm was generated. Regular expressions were used to identify the most common cancers in the cervical region (squamous cells carcinoma, adenocarcinoma and adenosquamous carcinoma). Any other record with the pattern specified in regular expressions was classified as other cancers.

Classification of HIV status

The HIV status of all records was classified using regular expressions where negative records were classified first, followed by positive records and records without any term related to HIV were classified as unknown HIV status.

Classification of HPV status

A look-up approach was used to identify the records that referred to any term related to HPV (*hpv* and *papillo*). 10% of the records with the term HPV were reviewed to identify any negative reporting on HPV. Terms and patterns used by pathologists to denote positive and negative HPV were used in a regular expression to assign HPV status filtered records. Records without any term related to HPV were given an *"unknown"* HPV status.

Classification of schistosomiasis status

Like the approach used to extract HPV status from records, regular expressions were used to find all records with terms related to schistosomiasis (*schisto, bilha, katayama, and snail*). These records were separated from the main dataset. Regular expressions were used to further classify schistosomiasis cases into positive and negative cases before appending with the records remaining from the main dataset. Any record without any term related to schistosomiasis was classified as having an "unknown" schistosomiasis status.

3.4 DATA MANAGEMENT

At each stage of the information extraction process, various datasets were produced. These datasets were stored in separate tables in a PostgreSQL 12 database. All relevant fields from various tables were merged to generate a structured dataset from which patterns could be explored using the statistical methods. At the end of the process, eight tables were created in the database and include; (appended dataset, cleaned text, classification of cancer, pre-cancer and other pathologies, schistosomiasis, HPV and HIV status, CIN level, cancer types and the structured dataset)

3.5 DATA ANALYSIS

Descriptive statistics using frequencies and proportions were used to describe the study population and the distribution pathologies of interest. The distribution of records with of HPV, Schistosomiasis and HIV related terms were also explored. A within group analysis was done to describe patterns in each group taking into account the group size.

Furthermore, records with cervical precancerous lesions or cancer with known schistosomiasis status were used to describe the prevalence of schistosomiasis in this population. The age distribution of patients with and without schistosomiasis was assessed using a density plot. The hypothesis on the effect of schistosomiasis on the age at cervical precancerous lesions and cervical cancer diagnosis was tested using an independent t-test.

A sensitivity analysis was conducted to validate the findings of the analysis described above. In this analysis, it was assumed that patients with an *unknown* schistosomiasis status were schistosomiasis negative.

3.6 ETHICAL CONSIDERATIONS

Ethical approval was obtained from the University of Witwatersrand Human Research Ethics Committee (clearance certificate number M181160) (Appendix D). Approval to access data from NHLS CDW was also obtained from the academic affairs and research department at NHLS (Appendix E). The data provided by NHLS contained identifiable patients' information. As such, the following measures will be put in place to ensure that the confidentiality of patients is maintained:

- The researcher's laptop was configured to NHLS security standards and all the security parameters checked and administered by the NHLS information technology support team.
- The computer of the primary investigator had password protection and could not be accessed by any unauthorised person.
- After the text-mining process, patients' identifiable information was removed from the structured dataset.

3.7 CONCLUSION

This section of the report described study population, the methods applied to extract and analyse information from cervical histology reports. The text-mining process included pre-processing, feature extractions and selections, algorithm training, hyperparameter optimisations, performance evaluation and word phrase matching. While the data analysis process included descriptive statistics and hypothesis testing. The next chapter will present the results from the methods discussed in this chapter.

4

RESULTS

4.1 INTRODUCTION

Chapter 4 presents the results of the text-mining process and the data analysis. It provides a description of the population in the dataset provided by NHLS as well as pre-processing and classification results. It also presents the prevalence of schistosomiasis in women with cervical precancerous lesions and cancer and tests the association between schistosomiasis and age at precancerous lesions and cancer diagnosis.

4.2 DESCRIPTION OF THE STUDY POPULATION

The final dataset used for this analysis was obtained by appending and cleaning four different datasets provided by NHLS. It consisted of 31,568 reports from histology samples analysed at Albert Luthuli Hospital from 2011 to 2017. After excluding records with little or no information on the detailed results column (n=52), the number of records reduced to 31,516. Approximately 95.08% (n=28,326) of patients had one histology sample while 4.92% (n=1,467) had two or more histology samples in the dataset.

The number of reports from histology samples increased with time, except in 2013 where there was a decline compared to the previous year (Figure 4.1).

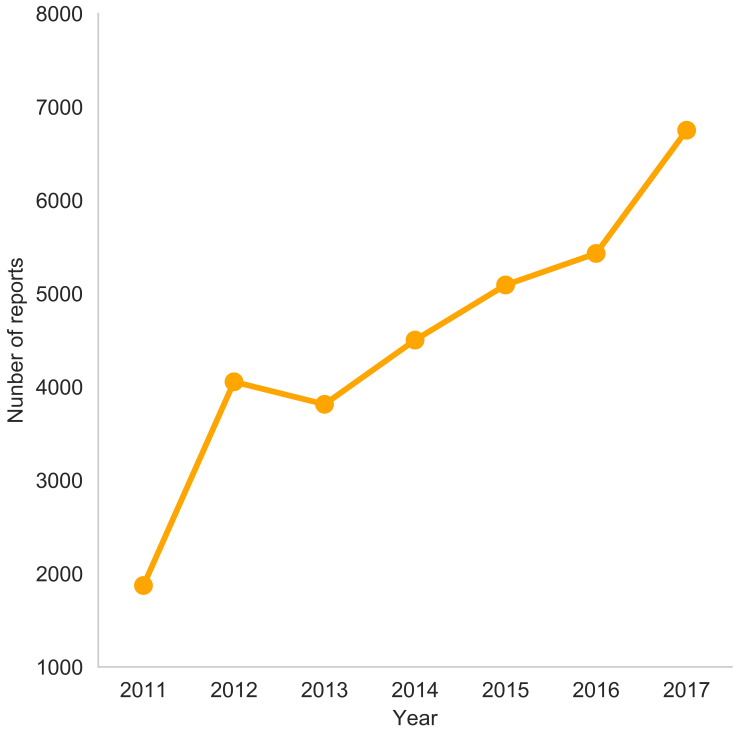


Figure 4.1: Distribution of reports by year

Black Africans were the most dominant population in the dataset constituting 76.68% (n=24,166) of all patients. They were followed by Indians 1.88% (n=598), Coloured 1.10% (n=348), Whites 0.27% (n=85), and Asians 0.17% (n=53), (Figure 4.2). Those with a missing race contributed to 19.90% (n=6,273) of the population in the dataset.

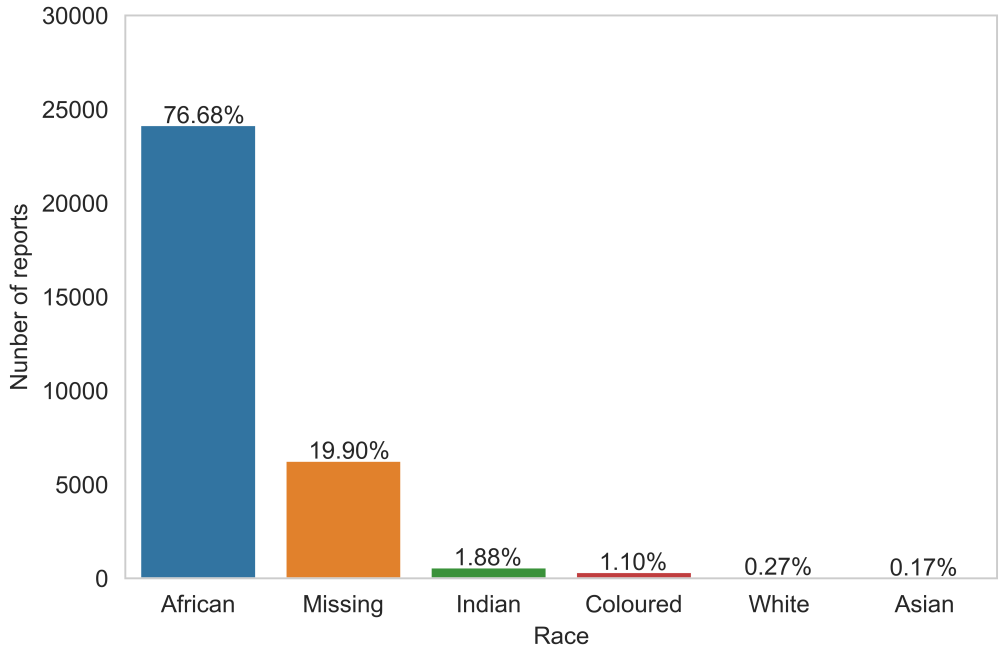


Figure 4.2: Distribution of records by race

Pathologists captured patients ages in 31.03% (n=9,779) of the detailed histology report. Besides the age captured by pathologists, 89.20% (n=28,115) of the date of birth variable had dates of birth which were used to calculate patients age. After combining both sources of patients age, 91.84% (n=28,945) of records had an age. The age for the remaining 8.16% (n=2,571) could not be determined. Patients were aged between 11 to 95 years old with a median age of 41 years (IQR=34;51).

Of the 31,516 records, only 50% (15,760) had a diagnosis in terms of pre-cancer, cancer and/other pathologies. Figure 4.3 shows the distribution of records by the pathology group. A higher proportion of labelled records belong to patients with pre-cancerous lesions (n=6,272), followed by cancer patients (n=5402) and other pathologies (n=4086).

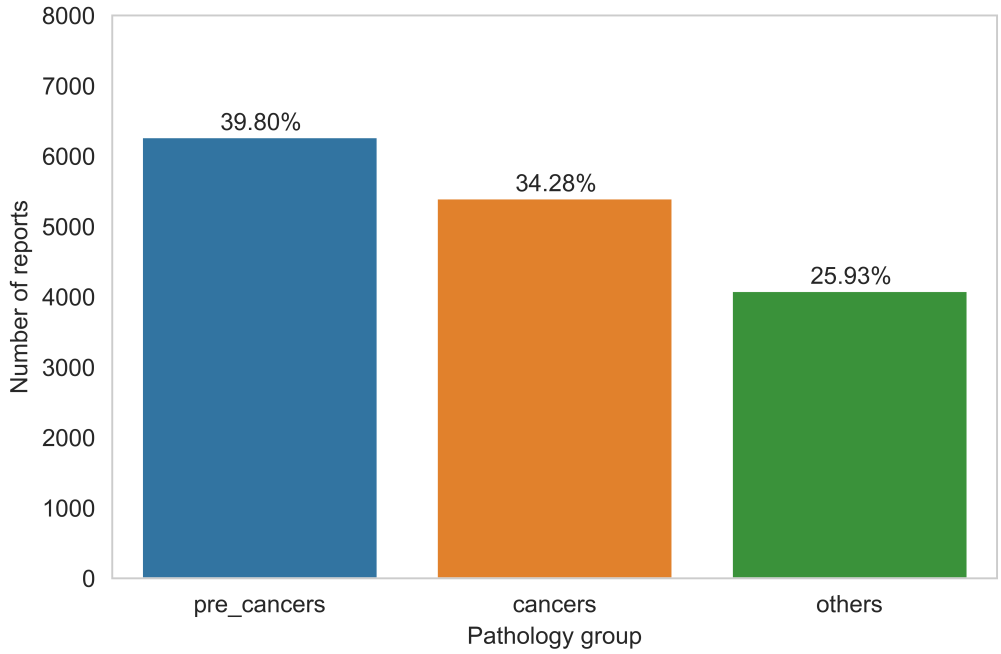
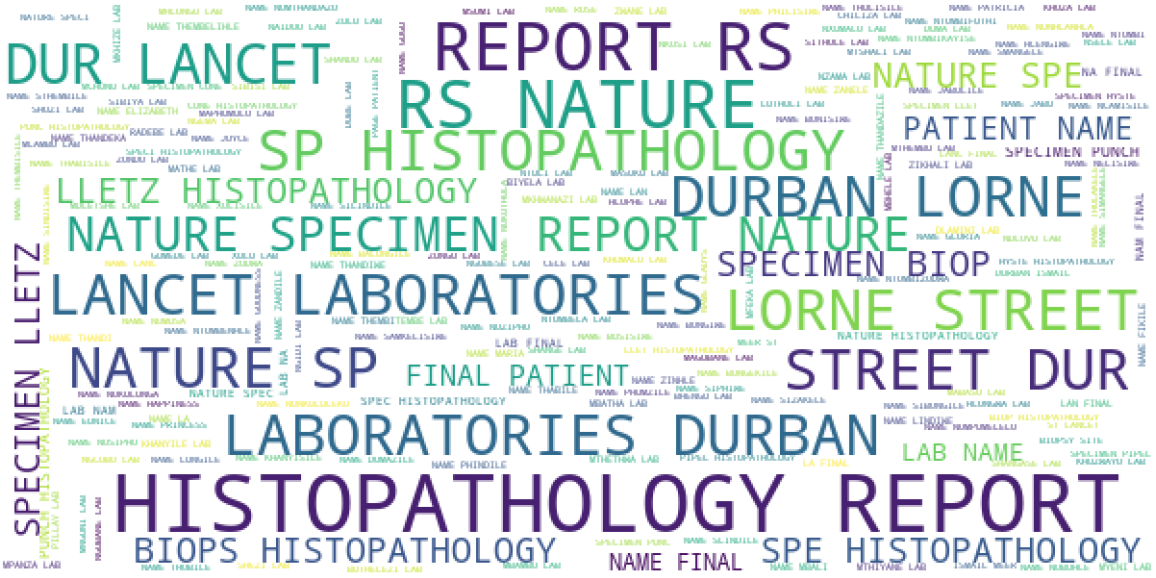


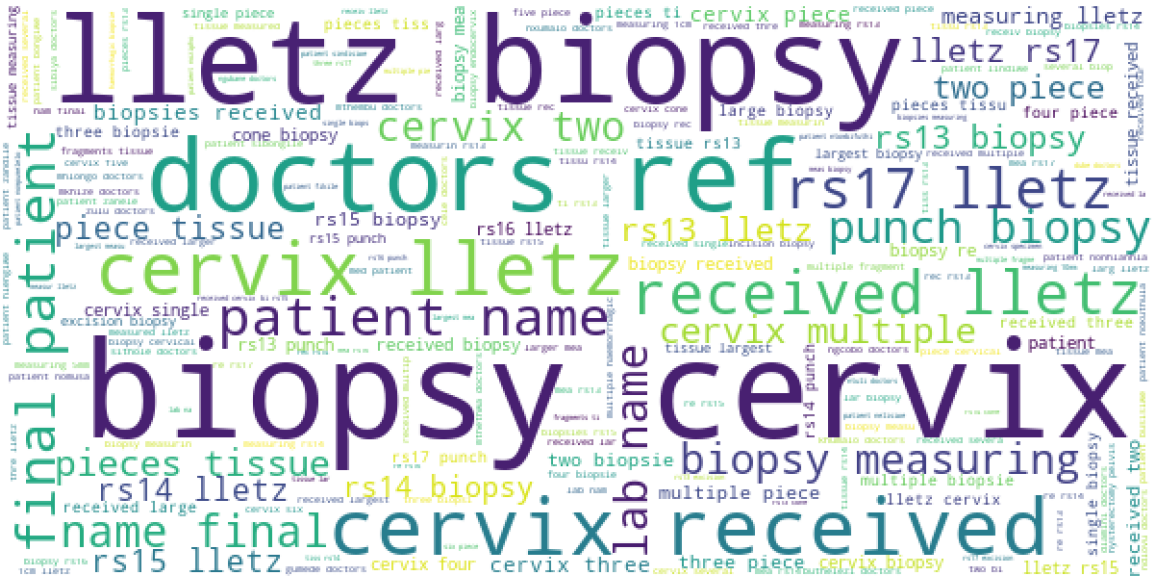
Figure 4.3: Distribution of labelled records by pathology group

4.3 TEXT PRE-PROCESSING RESULTS

Figure 4.4a illustrates a partial view of the content of all histology reports before pre-processing. The font size of words increased with an increase in the frequency of occurrence in the corpus. For instance, the phrase *"HISTOPATHOLOGY REPORT"* is the most dominant phrase before text pre-processing. This is because most reports start with that heading. Laboratory names and addresses were also dominant in the histology report before cleaning. The length of these reports ranged from 31 to 607 words. Among records with labels, the median length of reports was 100, 102 and 138 for other pathologies, cancer and precancerous lesions respectively. All three pathology groups had a similar distribution of report length with most records having less than 900 characters.



a



b

Figure 4.4: Overview of histology reports before and after pre-processing

Following text pre-processing, words in relation to the description of the histology samples such as *"biopsy"*, *"cervix"*, *"lletz"* became more dominant (Figure 4.4b). There was a considerable reduction in the length of the report after pre-processing. The pre-processed reports length ranged from 15 to 378 characters, with a median of 65, 64 and 68 for other pathologies, cancer and precancerous lesions respectively. The word count of most reports was reduced by approximately 35%. This reduction rate was consistent across all pathology groups (Figure 4.5).

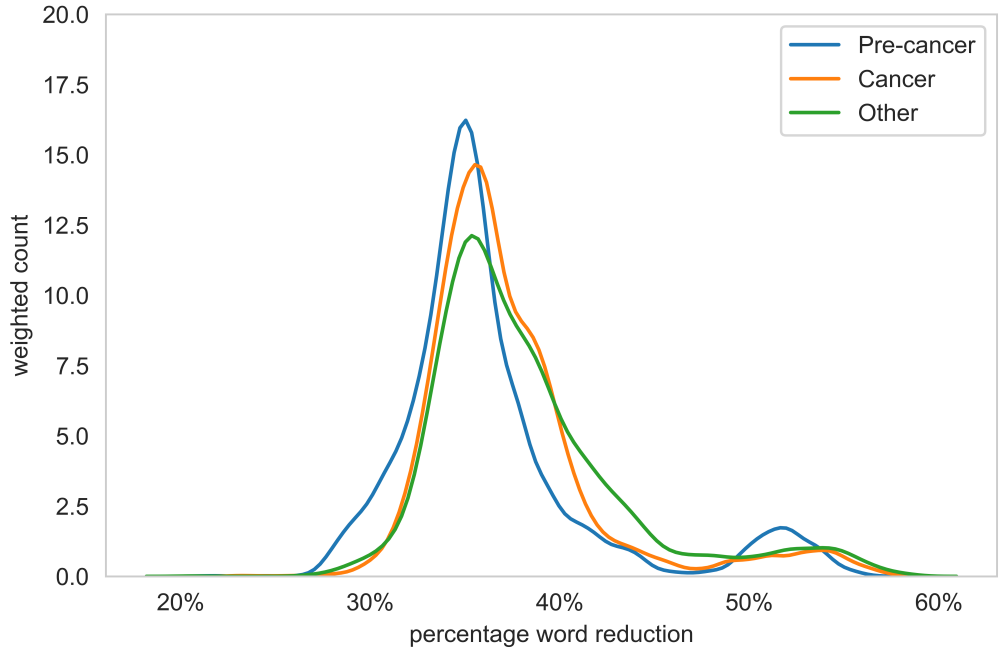


Figure 4.5: Word percentage reduction

4.4 FEATURE EXTRACTION AND SELECTION RESULTS

A total of 50,672 features made up of uni-gram and bigrams were generated from the cleaned histology reports. However, the model only required 60 features to perform at an optimum level.

4.5 EVALUATION OF MACHINE LEARNING ALGORITHMS RESULTS

4.5.1 *Dummy Classifier Compared to Selected OVO Algorithms*

The performance of the dummy classifier, SVM OVO and Random forest OVO is displayed in Table 4.1 below. With an F-measure of 0.339, the dummy classifier performed poorly compared to the other two classifiers. SVM OVO was the best performing classifier with F-measure=0.921.

Table 4.1: Performance of Dummy classifier and OVO classifiers

ML Algorithm	Precision	Recall	F-measure
SVM OVO	0.926	0.918	0.921
RF OVO	0.917	0.917	0.917
Dummy	0.339	0.339	0.339

4.5.2 *OVO Compared to OVA Algorithms*

Table 4.2 compares the performance of OVO and OVA versions of evaluated machine learning algorithms. RF OVA had the best precision, recall and F-measure in the classification of pathology reports into their three main groups (precancerous lesions, cancer and other cervical pathology). On the other hand, RF OVO was the algorithm with the least performance measures.

Table 4.2: Performance of the model OVO and OVA ML algorithms

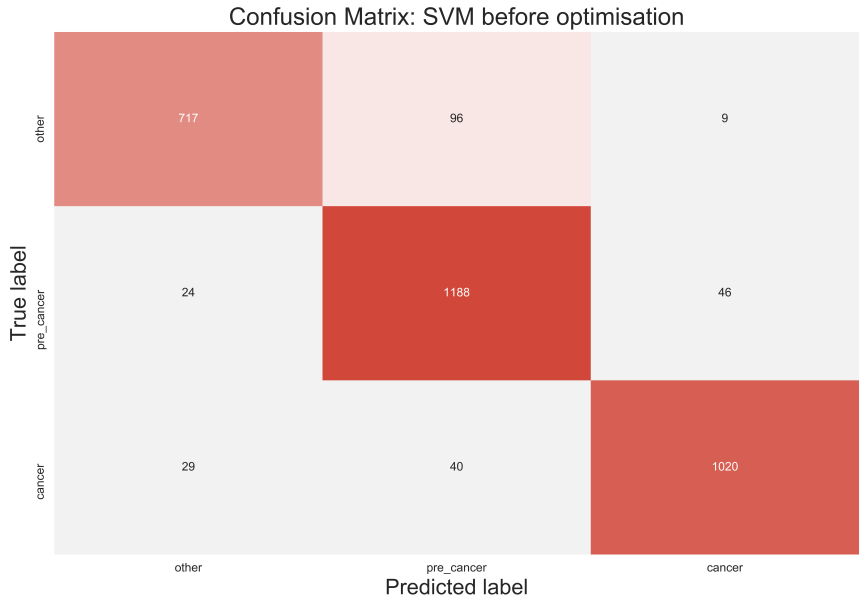
	OVO			OVA		
ML Algorithm	*Precision*	*Recall*	*F-measure*	*Precision*	*Recall*	*F-measure*
RF	0.917	0.917	0.917	0.925	0.922	0.924
SVM	0.926	0.918	0.921	0.923	0.919	0.921

The precision of the SVM OVO (0.926) was slightly better than that of SVM OVA (0.923), while the recall of SVM OVA (0.919) was slightly better than that of SVM OVO (0.918). However, there was no difference in their F-measures (0.921)

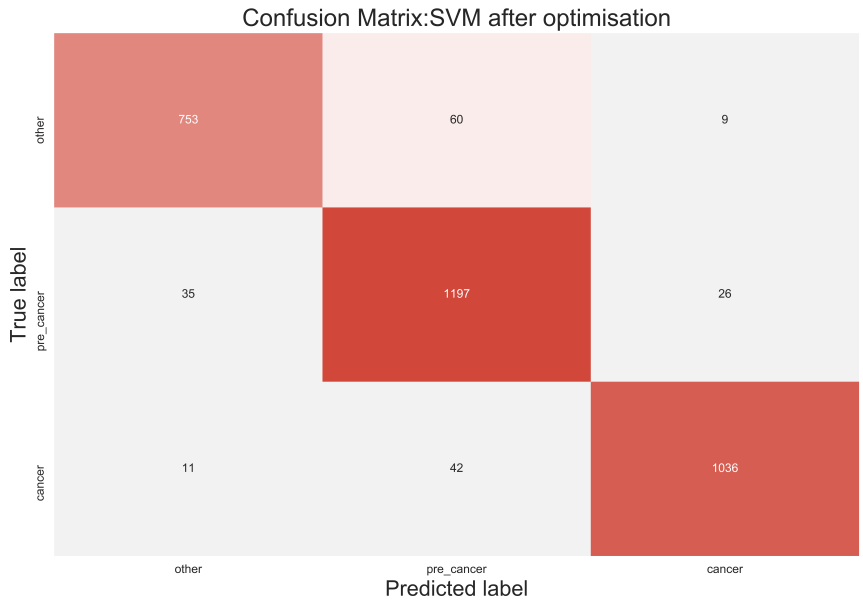
4.5.3 *Hyperparameter Optimisation*

Support Vector Machine

Figure 4.6a and Figure 4.6b show the confusion matrix of the SVM OVO before and after optimisation respectively. The number of records accurately classified increases with hyperparameter optimisation. For instance, the number of records accurately classified as other cervical pathologies in the test set increased from 717 before optimisation to 753 after optimisation.



a



b

Figure 4.6: SVM confusion matrix before and after optimisation

This increase lead to an improved overall F-measure and pathology group specific F-measure (Table 4.3).

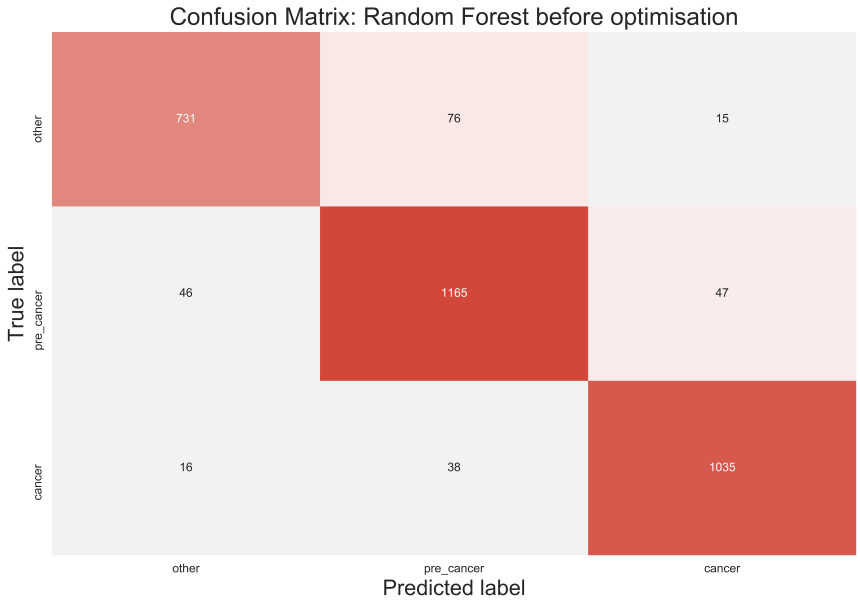
Table 4.3: F-Measure of SVM for each category before and after optimisation

	F-measure		
	Before optimization	***After optimisation***	***Percentage increase***
other	0.901	0.929	2.09%
pre-cancer	0.920	0.936	2.52%
cancer	0.943	0.959	1.91%
Overall	0.921	0.942	2.28%

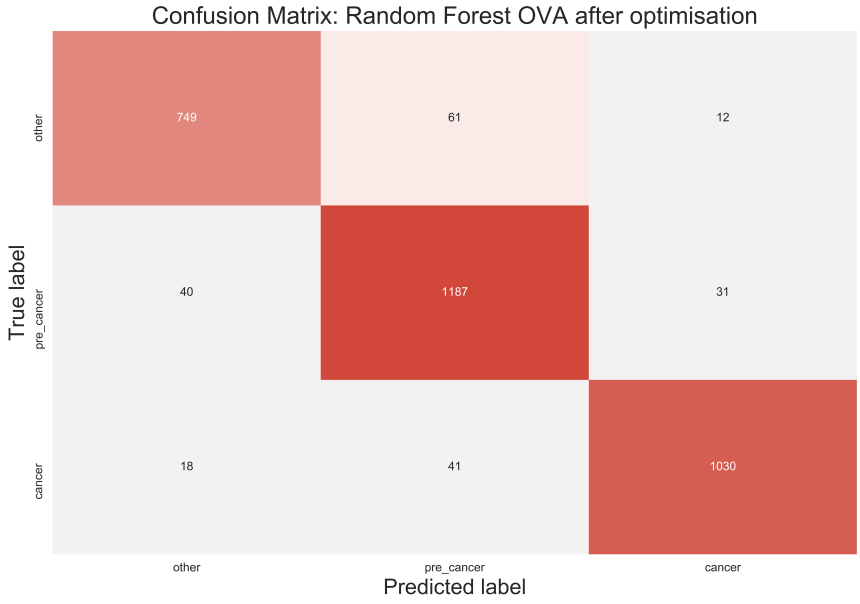
There was an overall 2.28% increase in performance of OVO SVM algorithm after optimisation. The percentage increase varied from 1.91% to 2.52% in various pathology groups. The highest improvement was in the classification of precancerous lesions while the lowest was among cancer records.

Random forest OVA

Figure 4.7a and Figure 4.7b show the confusion matrix of the Random Forest OVA before and after optimisation respectively.



a



b

Figure 4.7: RF confusion matrix before and after optimisation

Like in SVM OVO, the number of records correctly classified increased with hyperparameter optimisation which lead to an improved overall F-measure and pathology group-specific F-measure (Table 4.4).

Table 4.4: F-Measure of RF for each category before and after optimisation

	F-measure		
	Before optimization	***After optimisation***	***Percentage increase***
other	0.905	0.920	1.66%
pre-cancer	0.918	0.932	1.52%
cancer	0.947	0.953	0.6%
Overall	0.924	0.935	1.19%

Compared to SVM OVO, the performance improvements observed after hyperparameter optimisation RF OVA were relatively lower. There was an overall 1.19% increase in the performance of RF OVA after optimisation. The percentage increase varied from 0.6% to 1.66% in various pathology groups. The highest improvement was in the classification of other cervical pathologies records while the lowest was among cancer records.

4.5.4 *Comparison of Optimised ML algorithms*

Table 4.5 shows the precision, recall and F-measure of optimised algorithms. The performance of the SVM OVO was 0.5% higher than that of RF OVA.

Table 4.5: Comparison of optimised algorithms

ML algorithm	**Precision**	**Recall**	**F-Measure**
SVM	0.944	0.940	0.942
RF	0.936	0.934	0.935

4.5.5 *Qualitative Evaluation of the Classification Error*

A review of some records misclassified by SVM OVO showed that misclassification could be as a result of the two major issues.

Firstly, some histology reports were describing more than one pathology. In certain reports, the pathologist may have discussed the presence of cervical precancerous lesions and cervical cancer and assigned the SNOMED-CT corresponding the cervical cancer. However, the classifier may classify this as a record with a precancerous lesion.

Secondly, the inconsistent use of SNOMED-CTs across pathologies could have also led to misclassification of records. Certain SNOMED-CTs were used to describe both cervical precancerous lesions and certain cancer. Some SNOMED-CT used to indicate certain cancers were also used to indicate carcinoma in-situ which is a precancerous lesion.

4.6 DISTRIBUTION OF PATHOLOGIES AFTER CLASSIFICATION AND WORD/PHRASE MATCHING

Figure 4.8 shows the distribution of pathologies in the records after classification and word/phrase matching. After classification, the number of pre-cancer records increased by 11,635. The number of cancers increased by 250 and other pathologies increased by 3,871.

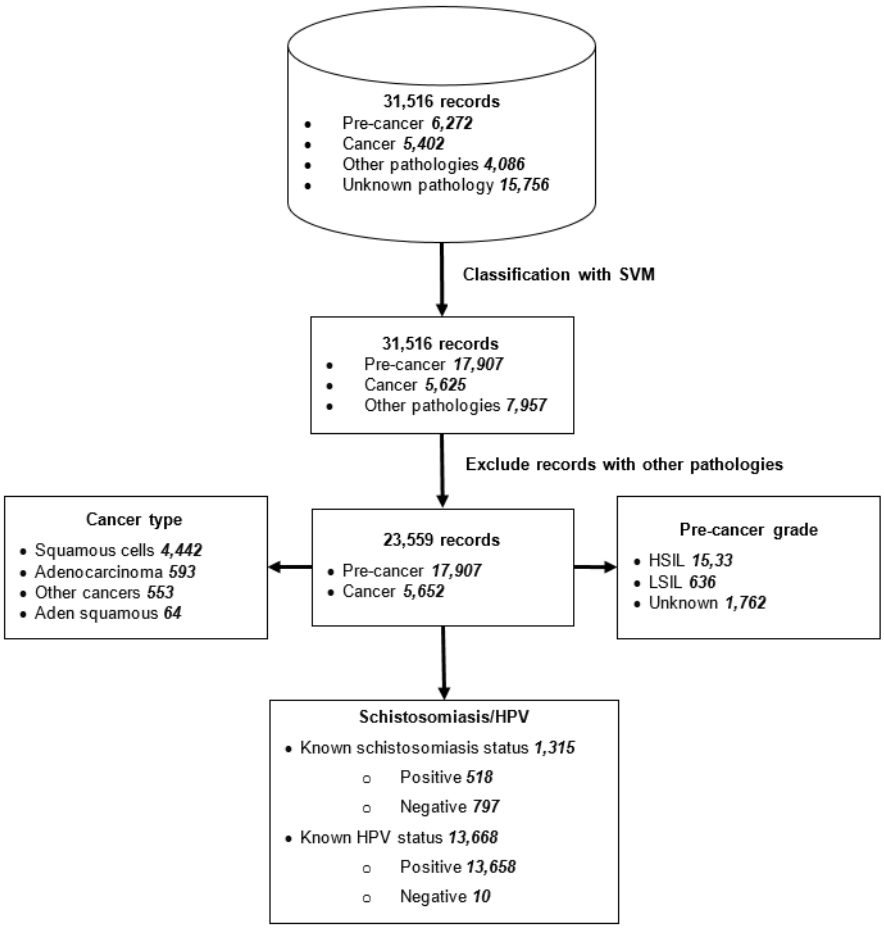


Figure 4.8: Distribution of all pathologies after classification and word/phrase matching

Patients with pre-cancers contributed to 56.82% (n=17,907) of the dataset while cancer patients and patients with other pathologies contributed to 17.93% (n=5,652) and 25.25% (n=7,957) respectively (Figure 4.9).

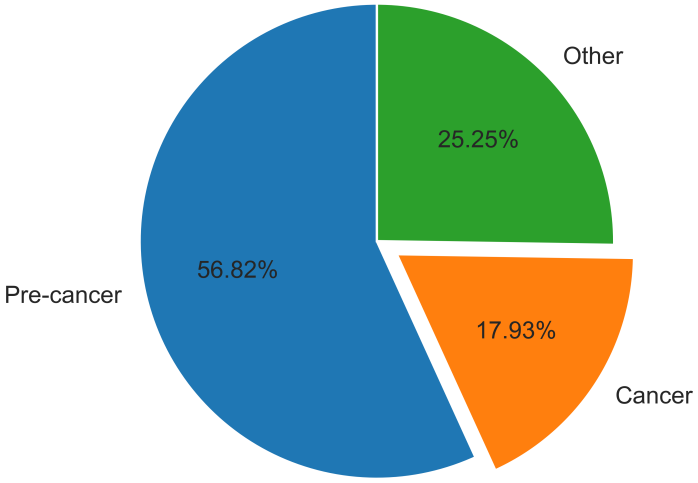


Figure 4.9: Distribution of pre-cancer, cancer and other pathologies

HSIL was the most common type of pre-cancer in the dataset (86.61%, n=15,133) while LSIL was least common (17.93%, n=636). An additional 1,762 (9.84%) records had pre-cancer however, they were not graded. This is referred to as the "*UNKNOWN*" category in Figure 4.10.

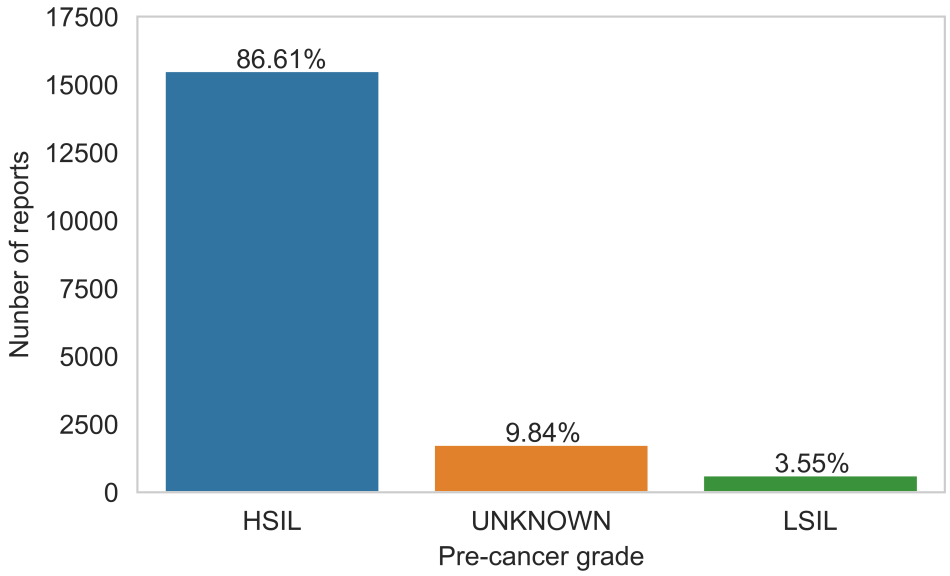


Figure 4.10: Distribution of precancerous lesions grade

Figure 4.11 shows the distribution of cancer types. The most common type of cancer was squamous cell carcinoma (78.59%, n=4,442), followed by adenocarcinoma (10.49%, n=593), other cancers (9.78%, n=553) and adenosquamous carcinoma (1.13%, n=64).

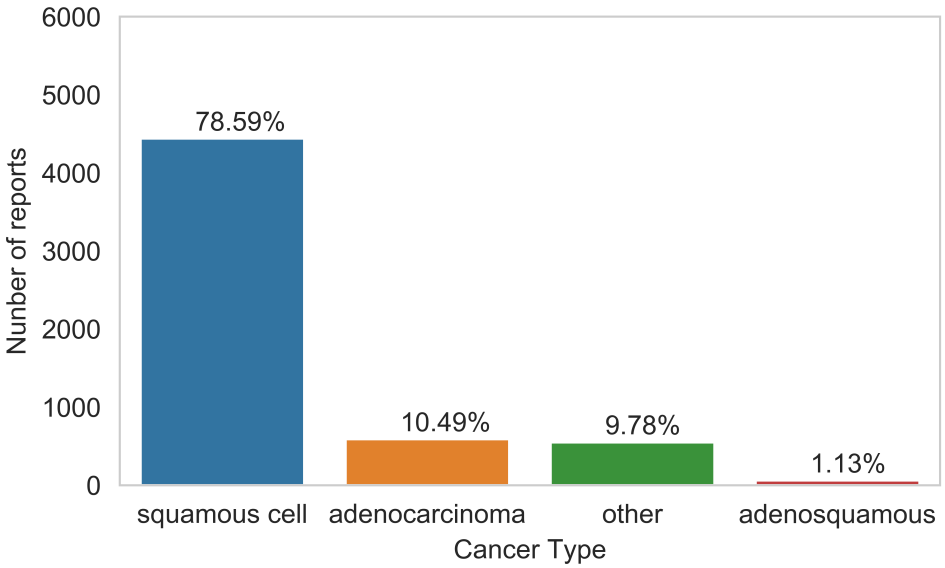


Figure 4.11: Distribution of cancer type

4.7 WORD/PHRASE MATCHING HPV RESULTS

HPV related terms were found in 58.02% (n=13,668) of the 23,559 records with cervical precancerous lesions or cancer. There was an increase in the proportion of records with HPV terms over the years (Figure 4.12). The proportion of records with HPV terms increased from 44.03% (n=557) in 2011 to 66.16% (n=5,290) in 2017.

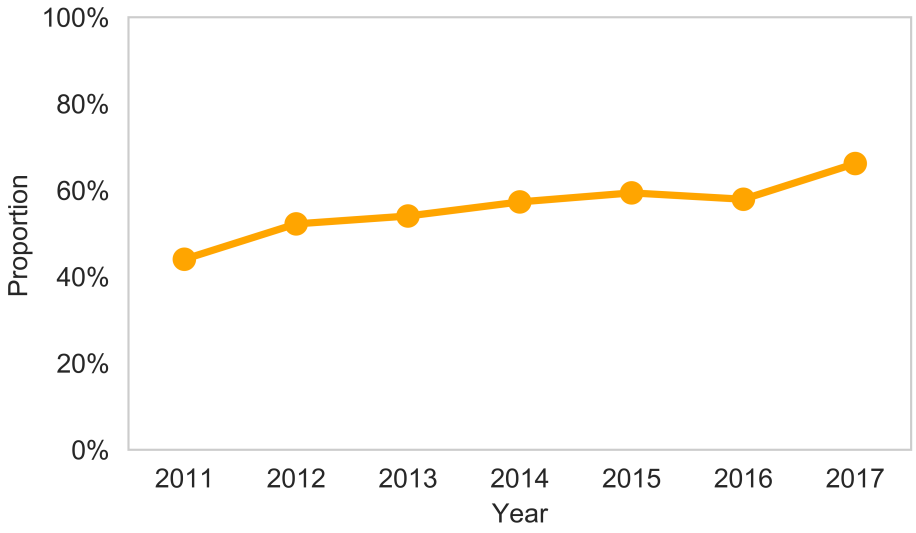


Figure 4.12: Proportion of records with HPV related terms by year

Table 4.6 shows the distribution of records with HPV terms by age group, race, and pathology group. HPV terms were mostly found among patients between the ages of 15 and 44 years old (65.34% to 76.52%). They were less present in records belonging to patients who were 55 years or older.

HPV related terms were mostly found in records belonging to black patients (58.49%). They were followed by patients with a missing race (57.66%) and coloureds (57.03%). Thirty-six percent (36.36%) of records belonging to whites patients had HPV related terms.
HPV terms were mostly found in pre-cancer records (74.81%) and least found in cancer records (4.81%). HPV related terms were found in 90.41% of LSIL records..

Table 4.6: Distribution of HPV related terms

	Records with HPV terms		
	No.	***%***	***Total***
Age group			
<15	3	60.00%	5
15-24	352	76.52%	460
25-34	3,960	73.33%	5,400
35-44	5,200	65.34%	7,958
45-54	2,173	51.25%	4,24
55+	1,002	27.36%	3,662
Race			
African	10,743	58.49%	18,367
Asian	13	46.43%	28
Coloured	150	57.03%	263
Indian	114	39.58%	288
White	20	36.36%	55

Table 4.6: Distribution of HPV related terms

	Records with HPV terms		
	No.	*%*	*Total*
Missing	2,628	57.66%	4,558
Pre-cancer	13,396	74.81%	17,907
HSIL	11,991	77.32%	15,509
LSIL	575	90.41%	636
Unknown	830	47.11%	1,762
Cancer	272	4.81%	5,652
Squamous cell	240	5.40%	4,442
Adenocarcinoma	16	2.70%	593
Adenosquamous	1	1.56%	64
Other	15	2.71%	553

4.8 WORD/PHRASE MATCHING SCHISTOSOMIASIS RESULTS

Schistosomiasis related terms were found in only 5.58% (n=1,315) of all precancerous lesion and cancer records. Figure 4.13 shows the pattern of schistosomiasis terms in pre-cancer and cancer records over time. The proportion of records with schistosomiasis terms drops from 3.48% to 2.25% between 2011 and 2013. It gradually increases from 2014 and reached its peak at 8.96% in 2016, and drops to 5.46% in 2017.

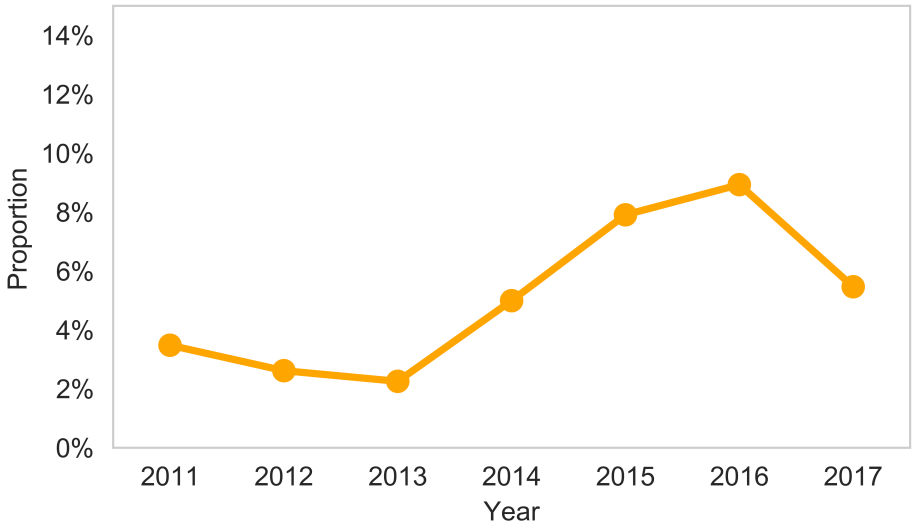


Figure 4.13: Proportion of records with schistosomiasis related terms by year

Table 4.7 shows the distribution of records with schistosomiasis related terms by age group, race, and pathology groups.

Table 4.7: Distribution of schistosomiasis related terms

	Records with schistosomiasis terms		
	No.	***%***	***Total***
Age group			
<15	0	0.00%	5
15-24	22	4.78%	460
25-34	356	6.59%	5,400
35-44	507	6.37%	7,958
45-54	230	5.42%	4,240
55+	154	4.21%	3,662
Race			
African	1,075	5.85%	18,367
Asian	1	3.57%	28
Coloured	20	7.60%	263

Table 4.7: Distribution of schistosomiasis related terms

	Records with schistosomiasis terms		
	No.	*%*	*Total*
Indian	6	2.08%	288
White	1	1.82%	55
Missing	212	4.65%	4,558
Pre-cancer	1,084	6.05%	17,907
HSIL	907	5.85%	15,509
LSIL	114	17.92%	636
Unknown	63	3.58%	1,762
Cancer	231	4.09%	5,652
Squamous cell	190	4.28%	4,442
Adenocarcinoma	21	3.54%	593
Adenosquamous	2	3.13%	64
Other	18	3.25%	553
HPV status			
Negative	7	70.00%	10
Positive	1,031	7.55%	13,658
Unknown	277	2.80%	9,891

The proportion of records with schistosomiasis terms varies from 0% to 6.59% in various age groups. The lowest proportion was observed among patients who were less than 15 years old, while the highest was among patients aged between 25 and 34 years.

Schistosomiasis related words were mostly found among records belonging to coloured patients (7.6%), followed by records belonging to black patients (5.85%).

Few records from white patients had schistosomiasis terms (1.82%).

Approximately 6% of Precancerous lesion records had schistosomiasis related terms. These terms were found in 4.09% of cancer records. Among cervical precancerous lesions records, schistosomiasis terms were mostly found in LSIL (17.92%).

The reporting patterns of positive and negative cases among patients with schistosomiasis status changed over the years. These changes are illustrated in figure 4.14. The number of records with schistosomiasis positive terms increases from 44 in 2011 to 74 in 2012. There is little variation in this number between 2013 and 2016 (62 to 79). This number of records with schistosomiasis positive terms increases to 107 in 2017. On the other hand, the number of records with a negative schistosomiasis status varies between 0 and 2 between 2011 and 2013. This number increases to 86 by 2014 and continuously increase until it reaches its peak of 302 in 2016 and drops to 182 in 2017.

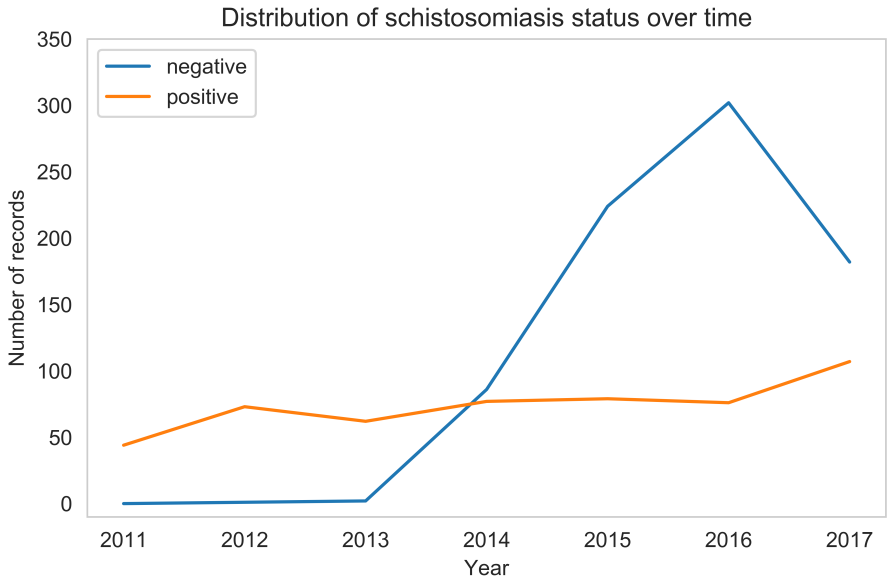


Figure 4.14: Distribution of schistosomiasis status over time

4.9 PATTERNS OF SCHISTOSOMIASIS IN WOMEN WITH PRECANCEROUS LESIONS AND CERVICAL CANCER DIAGNOSIS

The patterns of schistosomiasis in women with cervical precancerous lesions and cancer status were based on a population restricted to those with a known schistosomiasis status. Figure 4.15 shows the process used to select records for this analysis.

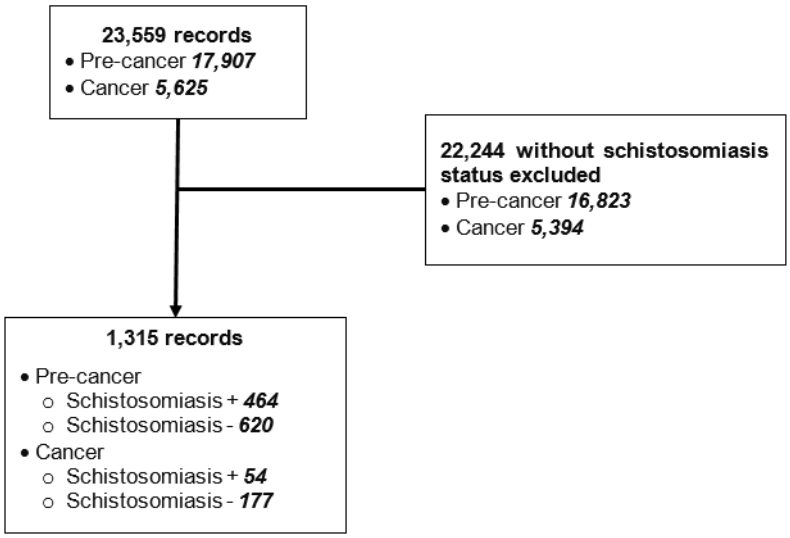


Figure 4.15: Selection of cases for schistosomiasis and age at diagnosis analysis

Out of the 23,559 records with pre-cancer and cancer, 22,244 records were excluded because they did not have a known schistosomiasis status. Thus, the analysis on the relationship between schistosomiasis and age at precancerous lesion and cancer diagnosis was based on 1,315 records (1,084 pre-cancer records and 231 cancer records). This population is made up of 81.75% (n=1,075) Africans, 16% (n=212) undefined race, 1.52% (n=20) coloured, Indians 0.46% (n=6), Asians 0.08% (n=1), and white 0.08% (n=1).

4.9.1 *Yearly Prevalence of Schistosomiasis Among Women with Cervical Precancerous Lesions/Cancer*

In total, 518 women with cervical precancerous lesions or cancer were diagnosed with schistosomiasis at Inkosi Albert Luthuli hospital from 2011 to 2017. The overall prevalence of schistosomiasis among women with cervical precancerous lesions, cancer and schistosomiasis status was 39.42%. Figure 4.16 shows the prevalence of schistosomiasis in this population over the years. There is an apparent reduction in the prevalence of schistosomiasis over the years. In 2011 the prevalence of schistosomiasis in women with cervical precancerous lesions, cancer and schistosomiasis was 100%. This reduced to 20.11% in 2016 and subsequently increased to 37.02% in 2017.

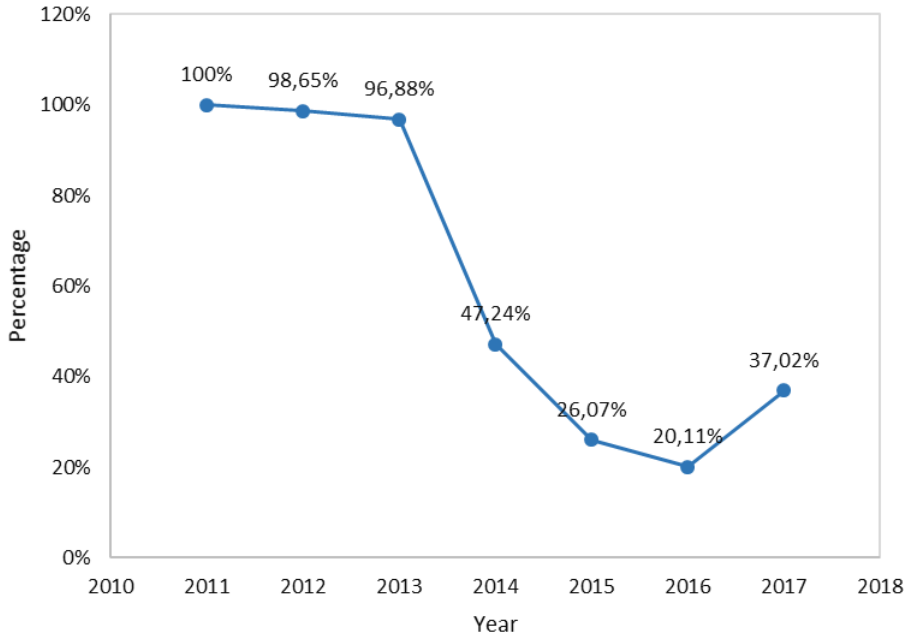


Figure 4.16: Prevalence of schistosomiasis in pre-cancer and cervical cancer patients over time

4.9.2 *Prevalence of Schistosomiasis Among Women with Cervical Precancerous Lesions*

Pre-cancer patients had a schistosomiasis prevalence of 42.87% (464). This prevalence was highest in patients with ungraded pre-cancers (58.70%; n=54) and

lowest in patients with LSIL (13.16% n=15). The prevalence of schistosomiasis was 44.99% (395) among patients with HSIL (Table 4.8).

Table 4.8: Prevalence of schistosomiasis in women with cervical precancerous lesions

Pathology	**Schistosomiasis**		
	No.	***%***	***Total***
Pre-cancer	464	42.80%	1,084
HSIL	405	44.65%	907
LSIL	15	13.16%	114
Unknown	44	69.84%	63

4.9.3 *Prevalence of Schistosomiasis Among Women with Cervical Cancer*

The prevalence of schistosomiasis in cancer patients was 23.38% (n=54). Patients with adenocarcinoma had the highest prevalence of 33.33% (n=7). They were followed by patients with squamous cell carcinoma with a prevalence of 23.16% (n=44). Other cancers had a schistosomiasis prevalence of 16.67% (n=3). Neither of the two patients with adenosquamous carcinoma tested positive for schistosomiasis (Table 4.9).

Table 4.9: Prevalence of schistosomiasis in women with cervical cancer

Pathology	**Schistosomiasis**		
	No.	***%***	***Total***
Cancer	54	23.38%	231
Squamous cell	44	23.16%	190
Adenocarcinoma	7	33.33%	21
Adenosquamous	0	0.00%	2
Other	3	16.67%	18

4.9.4 *Distribution of Age at Precancerous Lesion and Cervical Cancer Diagnosis*

Figure 4.17 shows the density distribution plot of patients' age at pre-cancer and cervical cancer diagnosis. A high proportion of precancerous lesion cases are diagnosed at around 35 years while most cancer patients are diagnosed at approximately 44 years.

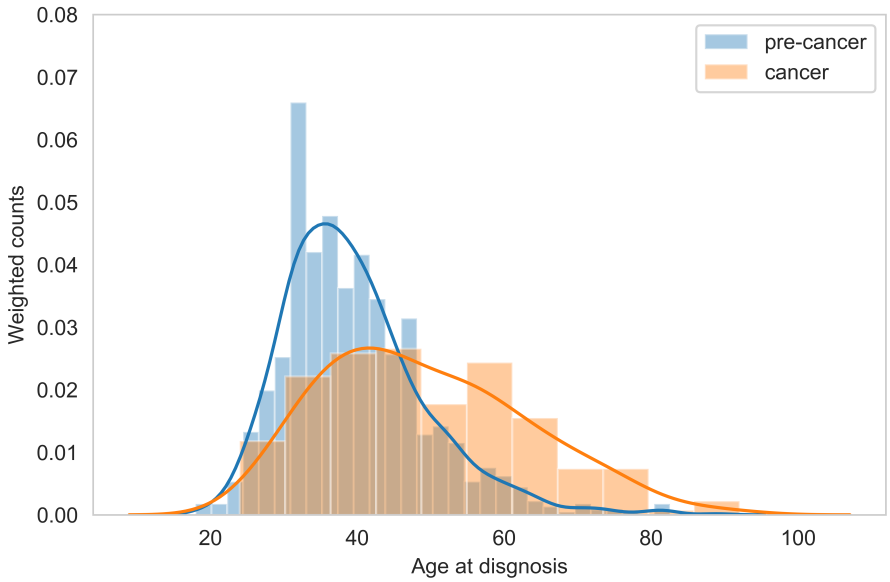


Figure 4.17: Distribution of pathologies by age at diagnosis

4.9.5 *Relationship Between Schistosomiasis and Age at Precancerous Lesions Diagnosis*

Figure 4.18 shows the distribution of age at pre-cancer by schistosomiasis status. The peak age at pre-cancer diagnosis in patients who are positive to schistosomiasis is approximately 34 years. On the other hand, the peak age at pre-cancer diagnosis in schistosomiasis negative women is 2 years later than those with schistosomiasis i.e. 36 years.

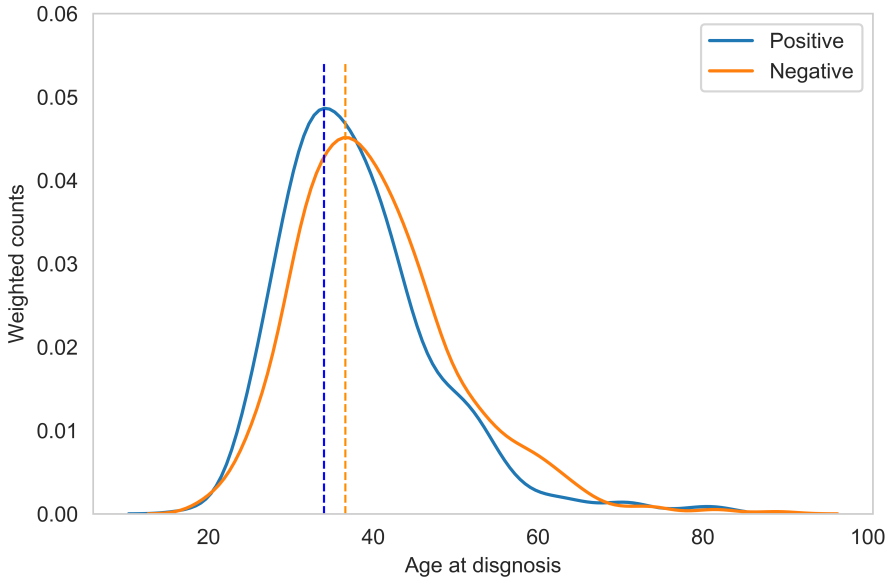


Figure 4.18: Distribution of age at pre-cancer diagnosis by schistosomiasis status

The mean age at precancerous lesion diagnosis in schistosomiasis positive patients was 38 ± 9.46 years while that of schistosomiasis negative patients was 40 ± 9.86 years. There is a two-year difference on the mean age at cervical precancerous lesion diagnosis between schistosomiasis positive and schistosomiasis negative patients. This difference is statistically significant ($p<0.001$).

4.9.6 *Relationship Between Schistosomiasis and Age at Cancer Diagnosis*

Figure 4.19 shows the distribution of age at cancer by schistosomiasis status. The peak age at cancer diagnosis in patients who are positive to schistosomiasis is approximately 39 years. The peak age at cancer diagnosis in schistosomiasis negative is approximately 46 years.

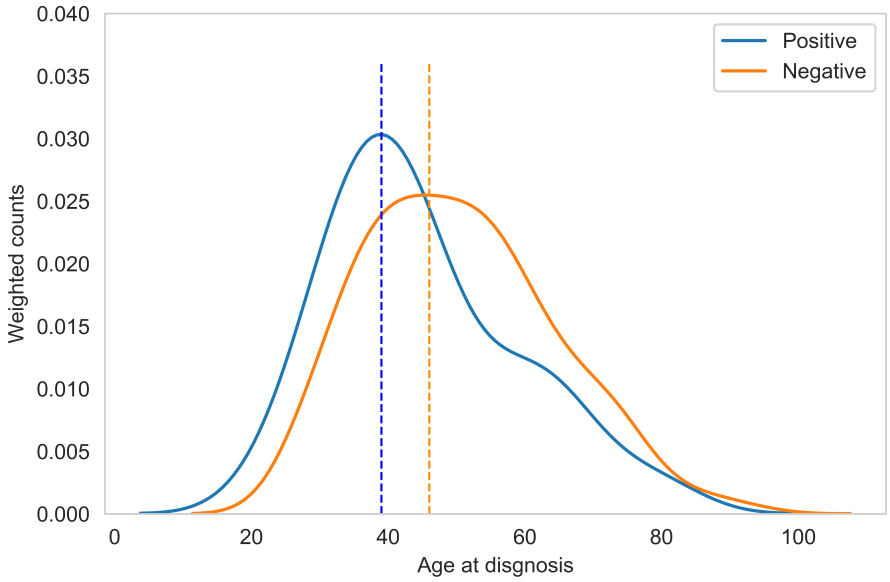


Figure 4.19: Distribution of age at cancer diagnosis by schistosomiasis status

The mean age at cancer diagnosis in schistosomiasis positive patients was 45 ± 13.70 years while that of schistosomiasis negative patients was 50 ± 13.78 years. There is a significant mean age difference of five years at cancer diagnosis between schistosomiasis positive and schistosomiasis negative patients (p=0.021).

4.9.7 *Results of the sensitivity analysis*

The sensitivity analysis was based on the assumption that women without any schistosomiasis related terms in the histology report are schistosomiasis negative. Under this assumption, there was a significant drop in the prevalence of schistosomiasis. It was 2.2% among women with cervical precancerous lesions and cancer. Women with cervical precancerous lesions had a schistosomiasis prevalence of 2.59% (n=464) while this prevalence was only 0.96% among women with cervical cancer.

There was no statistically significant difference in the age at cervical precancerous diagnosis (p=0.8941) and cervical cancer diagnosis (p=0.3027) between the negatives included for the main analysis and the patients with an *"unknown"*

schistosomiasis status added as negative cases in the sensitivity analysis. There was a slight reduction on the peak age at cervical precancerous lesion and cancer diagnosis in schistosomiasis negative patients. However, the average age at cervical precancerous lesion diagnosis did not change in this population. The difference of 2 years in age at cervical precancerous lesion diagnosis between schistosomiasis negative and schistosomiasis positive patients was statistically significant (p=0.001). On the other hand, the mean age at cancer diagnosis slightly increased from 50 years to 51 years in patients. This led to a statically significant difference (p=0.0020) of 6 years between the schistosomiasis negative and schistosomiasis positive patients. See Appendix F for detailed results on the sensitivity analysis.

4.10 CONCLUSION

This chapter presented the findings of the text-mining techniques and data analysis described in Chapter 3.

The interpretation of these results will be discussed in the next chapter.

5

DISCUSSION

5.1 INTRODUCTION

This chapter presents the discussion and limitations of the findings presented in the previous chapter. It includes the interpretation of the classification results, the word/phrase matching results of HPV and schistosomiasis, the hypothesis testing results as well as the limitations in the interpretation of such results.

The aim of this study was to investigate the presence of cervical schistosomiasis in patients with cervical cancer and precancerous lesions using text-mining techniques on pathology reports available at the NHLS. RF OVA performed better than SVM OVO in the classification of pathology reports into cervical precancerous lesions, cancer and other cervical pathologies before hyperparameter optimisation. However, SVM OVO outperformed RF after hyperparameter optimisation, making it the best classifier for this task.

Descriptive statistics showed that HPV related words were found in 52.08% of all cervical precancerous lesions and cancer patients, while schistosomiasis related terms were only present in approximately 5% of all cervical precancerous lesions and cancer patients. Assuming that records without schistosomiasis related terms were negative schistosomiasis cases, the prevalence of schistosomiasis was 2.2% and 0.96% among women with cervical precancerous lesions and cancer respectively. However, the prevalence was 42.87% in precancerous lesions and 23.38% in cervical cancer when records without a known schistosomiasis status are excluded. Moreover, the age at diagnosis of cervical precancerous lesions and cancer in the population with schistosomiasis was 38 ± 9.46 and 45 ± 13.70 years

respectively, while that in the schistosomiasis negative population was 40 ± 9.86 and 50 ± 13.78 years respectively. This suggests that patients with schistosomiasis may develop cervical pre-cancerous lesions and cancer 2 and 5 years earlier than patients without schistosomiasis respectively.

Text-mining techniques (NLP and ML) accurately classified 93.5% to 94.2% of histology records into cervical precancerous lesions, cancer and other cervical pathologies. Subsequently, word phrase matching (another text-mining technique) enabled the classification of precancerous lesion grades, cervical cancer type as well as schistosomiasis and HPV status in a shorter time frame compared to a human coder. Training algorithms to classify records into cervical precancerous lesions, cancer and other cervical pathologies increased the sample size for this study as records without diagnosis were assigned a diagnosis using the trained algorithm. Using this approach enabled this study to include a high number of observations compared to other studies that explored the relationship between cervical schistosomiasis, cervical precancerous lesions and cancer. Before, this study, the largest sample size used to explore this relationship was made up of 4,520 cervical cancer records. It was a retrospective study conducted at the cancer registry in Tanzania [8]. Pathologists reviewed records of patients with cervical cancer from 1980 to 1990 and tested histology samples when available. This method focusses only on records that were correctly labelled as having cervical cancer. Thus, excluding records which are not labelled. Moreover, the manual review of records is time-consuming and might take years to generate a large dataset for analysis. Applying text-mining techniques enabled the identification of an additional 11,635 cervical precancerous lesions cases and 250 cervical cancer cases which would have been omitted if we relied on labels provided by the pathologists. Moreover, text-mining enabled the description of cervical schistosomiasis patterns in 17,907 cervical precancerous lesions and 5,652 cervical cancer cases. However, schistosomiasis related terms were only found in 5% (n=1,315) of cervical precancerous lesions and cancer.

5.2 EVALUATION OF TEXT-MINING TECHNIQUES

The dummy classifier was only able to correctly classify 33.9% of cervical pathology report into cervical precancerous lesions, cancer and other cervical. This is because it does not use features for classification and uses a prediction process similar to guessing. Thus, showing the need to use trivial classifiers such as SVM and RF on this data.

Overall, SVM and RF classifiers performed well in the classification cervical pathology reports into cervical precancerous lesions, cancer and other cervical pathologies. They were able to classify more than 91% of records accurately. The performance of the SVM OVO and SVM OVA classifiers were the same. On the other hand, random forest OVA classifier outperformed the random forest OVO classifier in all evaluation metrics. This is similar to findings reported by Adnan MN and Islam MZ [75]. They conducted an experiment comparing the accuracy of random forest OVO classifier with random forest OVA classifier. In eight out of ten datasets, random forest OVA classifier performed better than random forest OVO yielding an average accuracy of 83.19% for random forest OVA classifier compared to 79.46% in for random forest OVO classifier.

Prior to hyperparameter optimisation, RF OVA had a better accuracy compared SVM OVO. However, optimisation improved the performance of SVM OVO so that it outperformed the optimised RF OVA. SVM OVO accurately classified 94.2% of records into cervical precancerous lesions, cancer and other cervical pathologies. Thus, SVM OVO classified records with an overall error rate of 5.8%. The performance in this study is higher than the SVM performance on the morphology classification of pathology reports in France [48]. This difference could be associated with the size of the training dataset used by both studies. The training set used in [48] was relatively lower compared to the training set used in this study.

The classification errors observed in the records classified by SVM OVO could be associated with discussing of multiple pathologies in one report and the inconsistent use of SNOMED-CTs across pathologies observed. The errors due to multiple pathologies being discussed is a single report is an indication that there is a need to develop models that can effectively classify a record in more than one class. This is known as multi-label classification.

The inconsistent use of SNOMED-CTs across pathologies observed in this study may have a negative impact on research and patients' health. Routinely collected data are often used for research, policy making, etc. If conclusions from these processes are made based on erroneous information, it might lead to health decisions that do not reflect the need of patients. Thus, negatively affecting patients care. Moreover, certain pathologists may rely on SNOMED-CTs to determine the treatment plan of the patient. As such inadequate treatment may be provided due to the use of inadequate SNOMED-CTs increasing the incidence rate of precancerous lesions and cancer.

5.3 HPV REPORTING RATES IN CERVICAL PRECANCEROUS LESIONS AND CANCER

HPV terms were only found in 58.02% of all pre-cancer and cancer records. This is an indication that HPV was only reported in 58.02% of pre-cancerous lesions and cervical cancer from 2011 to 2017. There was no clear reason in literature on screening policies that explained the steady increase in HPV reporting rates between 2011 and 2016 at Inkosi Albert Luthuli Hospital. However, the peak increase observed in 2017 (66.16%) could be associated with the introduction of the HPV Deoxyribonucleic Acid (DNA) test in public hospitals in 2017. HPV testing was not offered in the public sector before 2017 despite several proposals made by academics to include HPV DNA in cervical cancer screening [76, 77]. Vijayaraghavan A. *et al.* simulated the impact and cost of screening and treatment on the natural history of cervical cancer using a lifetime Markov Monte Carlo

simulation model [78]. They found that HPV testing was a cost-effective option for cervical cancer screening and prevention in South Africa [78]. Subsequently, the South African HPV Advisory Board updated its cervical cancer screening guidelines in 2010. In this update, they proposed three HPV based cervical precancerous lesions and cancer screening methods. However, the HPV test was only introduced in South Africa's public hospitals in 2017 [77]. This led to an increase in the proportion of people tested for and reporting on HPV.

Although HPV has been associated with cervical pre-cancerous lesions and cervical cancer, this study suggests that pathologists put more effort to report HPV status in cervical precancerous lesions than cervical cancer. The overall HPV reporting rates in precancerous lesions was 15 times higher than HPV reporting rates in cervical cancer. This could be explained by that fact that HPV status in precancerous lesions are expected to clear with no treatment in LSIL and adequate treatment in HSIL [79]. As such, pathologists might be more likely to report the HPV status at diagnosis as this will be an important marker needed to assess the treatment outcome of precancerous lesions over time. However, in cervical cancer, treatment outcomes are mostly assessed by the progression or regression of the tumour and little attention may be given to the changes in HPV status overtime. Thus, HPV is not reported in this population.

The reporting rates of HPV among cervical cancer patients in this study was 6 times lower than that reported in Belgium for the year 2013 to 2014. In their 2017 annual report, the Belgian government reported an HPV testing rate in cervical cancer of 29% for the 2013-2014 year [80], while this study found an HPV testing rate of only 4.81% for the 2011-2017 period. This difference could be as a result of low reporting rates of HPV status in the records used for this analysis and/or limited resources for HPV screening in developing countries compared to developed countries [5, 25]. Nonetheless, the HPV reporting rates are still low in both studies. There is increasing evidence that shows that there is a considerable proportion of non- HPV related cervical cancers. This varies between 7% and 15%

worldwide [81–84]. Cervical cancers that are not related to HPV are often more aggressive than HPV positive cervical cancers [84]. Thus, the HPV status at cancer diagnosis is essential for adequate clinical management of cervical cancer patients and further research in this area.

5.4 SCHISTOSOMIASIS REPORTING RATES IN CERVICAL PRECANCEROUS LESIONS AND CANCER

There has been a slight increase in the reporting rates of schistosomiasis among patients with cervical precancerous lesions and cervical cancer at Inkosi Albert Luthuli Hospital over the years. There was a particular increase in the reporting of negative cases. Despite the increase in the reporting of schistosomiasis in patients with cervical cancer and precancerous lesions over the years, it was still worryingly low. Inkosi Albert Luthuli hospital is in KZN where schistosomiasis is endemic [49, 50]. With extensive literature demonstrating the negative effect of female genital schistosomiasis on the female reproductive health, one would expect that it would be reported in all biopsies of the female genital organ especially in schistosomiasis endemic areas. Out of the 23,559 cervical precancerous lesions and cancer records available, only 5.15% (1,315) had a schistosomiasis status. This may negatively impact the treatment outcome of patients and research efforts to understand the effect of schistosomiasis on cervical precancerous lesions and cancer.

Prospective studies with control groups cannot be used to study such association as it is unethical to observe a disease progressing without providing any treatment. The use of cross-sectional designs to study such associations are often made up of small sample sizes which reduce the power of their results. As such, laboratory reports from routine histopathology samples should be an acceptable source of data to study such an association. The current schistosomiasis reporting rates among patients with cervical precancerous lesions and cancer at Inkosi Albert Luthuli hospital suggests that it might be challenging to study such

association using laboratory reports from routine histopathology sample if no intervention is implemented to increase schistosomiasis reporting rates.

It is difficult to ascertain the prevalence of schistosomiasis in the patients with cervical precancerous lesions and cancer using this data. There was an apparent decrease in the prevalence of schistosomiasis in patients with precancerous lesions and cancer and a schistosomiasis status over time. This decrease in prevalence is driven by the reporting patterns of schistosomiasis negative cases observed. Before 2014, pathologists were only reporting schistosomiasis positive cases. Later they increasingly reported negative cases, while the number of positive cases reported had little variation. This increased the denominator used to calculated prevalence over time and gives the impression that the prevalence of schistosomiasis in women with precancerous lesions and cancer was decreasing. Thus, there is a need for pathologists to consistently report both positive and negative schistosomiasis cases for patients with cervical precancerous lesions and cancer to determine the actual prevalence of schistosomiasis in this population.

The prevalence of schistosomiasis among patients with a schistosomiasis status was very high in cervical precancerous lesions (42.80%) and cancer (23.38%). The prevalence of schistosomiasis observed in this study was higher compared to any other prevalence reported from a retrospective record review. Moubayed P. *et al.* and Darré T. *et al.* reported a prevalence of 1.7% and 1.9% in Tanzania and Togo respectively [8, 10]. The difference in prevalence could be as a result of the difference in the population used to determine the prevalence of schistosomiasis. In fact, the prevalence reported in previous studies was similar to the prevalence calculated during the sensitivity analysis with the assumption that records without any schistosomiasis related terms were negative schistosomiasis cases (0.96%). Therefore, we assumed that previous studies considered any case where schistosomiasis was not reported as schistosomiasis negative given that the definition of negative cases were not explicitly defined.

An analysis of age at precancerous lesions and cancer diagnosis of schistosomiasis positive and schistosomiasis negative patients suggests that patients with schistosomiasis may develop cervical precancerous lesions and cancer 2 and 5 years earlier than patients without schistosomiasis respectively. This finding is comparable with the findings from other studies [7, 8]. This could be explained by a possible relationship existing between cervical schistosomiasis, HPV and HIV.

Children in communities where schistosomiasis is endemic are exposed to infested waters at an early age such that they can develop urogenital schistosomiasis as early as the age of 4 months [85]. At this age, their mothers may use infested waters to bath them. As they get older, their exposure increases as they go to infested water to bath, play and swim with friends [86]. Girls infected with schistosomiasis start showing signs and symptoms related to cervical ulcerative lesions and erosions from the age of 5 years old [87]. Many women and girls in SSA have their first sexual debut before the age of 15 years [88]. Therefore, the cervical lesions induced by schistosomiasis before their first sexual debut increases their risk of acquiring HPV at a young age. Moreover, cervical schistosomiasis facilitates the progression of HPV induced lesions [41]. This might lead to cervical precancerous lesions and cancer at a young age in females with cervical schistosomiasis.

Studies have also reported schistosomiasis as a risk factor for acquiring HIV at an early age [88]. In an environment where schistosomiasis, HIV and HPV coexist, high HIV viral loads induced by schistosomiasis might also increase HPV viral load and the rapid progression of HPV induced lesions [22, 89]. Thus, women with cervical schistosomiasis, HIV and HPV may develop cervical lesion even earlier than patients with schistosomiasis and HPV only. However, this hypothesis was not explored in this study due to the limited reporting of HIV status in the histology reports.

Moreover, the diagnosis of cervical precancerous lesions and cancer at an early age in patients with schistosomiasis can be associated to schistosomiasis symptoms that occur in early adulthood. The presence of trapped schistosomiasis eggs in the cervix often causes inflammation in the surrounding areas and scarring. This may lead to cervicitis, dysmenorrhoea, menorrhagia, leucorrhoea, lower abdominal pain, post-coital bleeding, inter-menstrual bleeding and dyspareunia [40] that may prompt patients to seek medical attention at an earlier stage compared to patients without schistosomiasis.

5.5 LIMITATIONS

The findings of this study should be considered in light of some limitations.

1. The performance of the word/phrase matching in extracting HPV and schistosomiasis status could not be assessed due to the limited number of records with SNOMED-CTs for the HPV and schistosomiasis diagnosis.

2. Findings of this study are based on a dataset where only 5% of records have a schistosomiasis status and any assumption used to analyse such data maybe biased as it might not reflect reality.

3. The HIV status of the patient is an important determinant in the progression of HPV induced lesions. However, due to limited reporting of HIV status of patients in histology narratives and constrained time for the study, this study only explored the effect of schistosomiasis on age at cervical precancerous lesions and cancer diagnosis without taking into consideration the HIV status of patients. Further studies should consider using record linkage to link this dataset to the HIV laboratory tests to study the combined effect of schistosomiasis and HIV on the burden of cervical cancer observed in SSA.

5.6 CONCLUSION

This chapter presented the discussion of the research findings, its limitation and future research directions. The next chapter, will provide a summary of the study and recommendations.

6

CONCLUSIONS AND RECOMMENDATIONS

This study aimed to apply text-mining techniques on pathology reports available at NHLS to investigate the presence of cervical schistosomiasis in patients with cervical precancerous lesions and cancer. Applying text-mining techniques to explore the relationship between cervical schistosomiasis and precancerous lesions could be a cost-effective alternative to currently applied methodologies that are resource-intensive, expensive and/or time-consuming. It might also provide findings on time to develop policies to address the burden of cervical cancer in the rural areas of developing countries.

The study shows that machine learning combined with word/phrase matching approaches are effective text-mining techniques that can be used in research focusing on the association between schistosomiasis and cervical cancer. It demonstrates that this technique can be used to convert histology reports into a structured dataset on which further statistical methods can be applied to study the relationship between schistosomiasis and cervical precancerous lesions and cancer. The pipeline proposed in this study can be implemented in other settings to study this association. It can also be used on a yearly basis to monitor the rate of HPV negative cervical cancer and inform policymakers to act accordingly.

However, the generalizability of findings using this technique is limited by the rate at which pathologists report schistosomiasis and HPV in cervical histology reports (negative and positive reporting). Therefore, it is recommended that pathologists should be educated on the importance of reporting these conditions for better patient management and research.

In addition, interventions should be implemented to ensure that pathologists report the schistosomiasis status in cervical precancerous lesions and cancer cases especially in schistosomiasis endemic areas. Reporting schistosomiasis in this population might guide clinicians on the treatment of patients with cervical precancerous lesions and cancer for schistosomiasis. This will provide data that can be used to understand the effect of schistosomiasis treatment on the progression of HPV induced lesions.

By analysing the small proportion of patients with a schistosomiasis status, this study identified a high prevalence of schistosomiasis in women with cervical precancerous lesions and cancer. However, this reduces considerably when records without schistosomiasis terms are considered as schistosomiasis negative. Also, women with schistosomiasis appeared to develop cervical precancerous lesions and cancer 2 and 5 years earlier than patients without schistosomiasis respectively. This could be attributed to the effect of cervical schistosomiasis on the natural history of HPV and HIV. Based on this conclusion, efforts to fight against schistosomiasis, HPV and HIV should be integrated for better control of these epidemics in schistosomiasis endemic areas.

This finding is based on a test of association due to the scope of this study. Further studies should consider linking this dataset to the HIV database for KZN and conduct multivariate analysis to validate these findings.

REFERENCES

[1] L Bruni, L Barrionuevo-Rosas, G Albero, B Serrano, M Mena, D Gómez, J Munoz, FX Bosch, S de Sanjose et al. *Human papillomavirus and related diseases in the world. Summary report 27 July 2017*. 2017.

[2] Atara Ntekim. 'Cervical cancer in sub Sahara Africa'. In: *Topics on cervical cancer with an advocacy for prevention*. InTech, 2012.

[3] Hugo De Vuyst, Laia Alemany, Charles Lacey, Carla J Chibwesha, Vikrant Sahasrabuddhe, Cecily Banura, Lynette Denny and Groesbeck P Parham. 'The burden of human papillomavirus infections and related diseases in sub-saharan Africa'. In: *Vaccine* 31 (2013), F32–F46.

[4] Jean Damascène Makuza, Sabin Nsanzimana, Marie Aimee Muhimpundu, Lydia Eleanor Pace, Joseph Ntaganira and David James Riedel. 'Prevalence and risk factors for cervical cancer and pre-cancerous lesions in Rwanda'. In: *Pan African Medical Journal* 22.1 (2015).

[5] Rosa Catarino, Patrick Petignat, Gabriel Dongui and Pierre Vassilakos. 'Cervical cancer screening in developing countries at a crossroad: Emerging technologies and policy choices'. In: *World journal of clinical oncology* 6.6 (2015), p. 281.

[6] E Szela, J Bachicha, D Miller, M Till and JB Wilson. 'Schistosomiasis and cervical cancer in Ghana'. In: *International Journal of Gynecology & Obstetrics* 42.2 (1993), pp. 127–130.

[7] Abdel Fattah Youssef, Mohamed M Fayad and M Alaa-El-Din Shafeek. 'Bilharziasis of the cervix uteri'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 77.9 (1970), pp. 847–851.

[8] P Moubayed, J-F Lepère, H Mwakyoma and D Neuvians. 'Carcinoma of the uterine cervix and schistosomiasis'. In: *International Journal of Gynecology & Obstetrics* 45.2 (1994), pp. 133–139.

[9] KU Petry, U Scholz, B Hollwitz, R Von Wasielewski and CJLM Meijer. 'Human papillomavirus, coinfection with Schistosoma hematobium, and cervical neoplasia in rural Tanzania'. In: *International Journal of Gynecological Cancer* 13.4 (2003), pp. 505–509.

[10] Tchin Darré, Abdoul-Samadou Aboubakari, Bingo K N'Bortche, Akila Bassowa and Gado Napo-Koura. 'Association of Schistosomiasis with Cervical Cancer in Togo: the Consequence of this Association'. In: *Pathology & Oncology Research* (2017), pp. 1–2.

[11] World Health Organization et al. 'Working to overcome the global impact of neglected tropical disease'. In: *First WHO Report on Neglected Disease). Geneva: World Health Organization* (2010).

[12] David A Mayer and Bernard Fried. 'The role of helminth infections in carcinogenesis'. In: *Advances in Parasitology* 65 (2007), pp. 239–296.

[13] Martyn Plummer, Catherine de Martel, Jerome Vignat, Jacques Ferlay, Freddie Bray and Silvia Franceschi. 'Global burden of cancers attributable to infections in 2012: a synthetic analysis'. In: *The Lancet Global Health* 4.9 (2016), e609–e616.

[14] EF Kjetland, PD Ndhlovu, T Mduluza, V Deschoolmeester, N Midzi, E Gomo, L Gwanzura, PR Mason, JB Vermorken, Henrik Friis et al. 'The effects of genital Schistosoma haematobium on human papillomavirus and the development of cervical neoplasia after five years in a Zimbabwean population-The impact of Schistosoma haematobium on persistent HPV in rural Zimbabwean women at high-risk was investigated.' In: *European journal of gynaecological oncology* 31.2 (2010), p. 169.

[15] Farhad Soleimanian Gharehchopogh and Zeinab Abbasi Khalifelu. 'Analysis and evaluation of unstructured data: text mining versus natural language processing'. In: *Application of Information and Communication Technologies (AICT), 2011 5th International Conference on*. IEEE. 2011, pp. 1–4.

[16] George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson and Paul D Clayton. 'Unlocking clinical data from narrative reports: a study of natural language processing'. In: *Annals of internal medicine* 122.9 (1995), pp. 681–688.

[17] Irena Spasić, Jacqueline Livsey, John A Keane and Goran Nenadić. 'Text mining of cancer-related information: review of current status and future directions'. In: *International journal of medical informatics* 83.9 (2014), pp. 605–623.

[18] O'Reilly. *Chapter 1. What Is a Regular Expression? - Introducing Regular Expressions*. 2020. URL: https://www.oreilly.com/library/view/introducing-regular-expressions/9781449338879/ch01.html (visited on 13/01/2020).

[19] John Maindonald. *Pattern Recognition and Machine Learning*. Vol. 17. Book Review 5. 2006. DOI: 10.18637/jss.v017.b05.

[20] Sidath Asiri. *Machine Learning Classifiers – Towards Data Science*. 2018. URL: https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623 (visited on 13/01/2020).

[21] Sebastian Raschka. *Machine Learning FAQ: What is the difference between a classifier and a model?* 2020. URL: https://sebastianraschka.com/faq/docs/difference{_}classifier{_}model.html (visited on 27/01/2020).

[22] Machine learning. 'Glossary of Terms Glossary of Terms'. In: 30.2-3 (1998), pp. 271–274.

[23] Ciaran BJ Woodman, Stuart I Collins and Lawrence S Young. 'The natural history of cervical HPV infection: unresolved issues'. In: *Nature Reviews Cancer* 7.1 (2007), pp. 11–22.

[24] Robert J Kurman, Maria Luisa Carcangiu, C Simon Herrington and Robert H Young. *WHO classification of tumours of female reproductive organs*. International Agency for Research on Cancer, 2014.

[25] WHO. *Human papillomavirus (HPV) and cervical cancer*. Accessed: 2018-06-26. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer.

[26] MJ Arends, CH Buckley and M Wells. 'Aetiology, pathogenesis, and pathology of cervical neoplasia.' In: *Journal of clinical pathology* 51.2 (1998), pp. 96–103.

[27] Anna-Barbara Moscicki, Stephen Shiboski, Jeannette Broering, Kimberly Powell, Lisa Clayton, Naomi Jay, Teresa M Darragh, Robert Brescia, Saul Kanowitz, Susanna B Miller et al. 'The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young women'. In: *The Journal of pediatrics* 132.2 (1998), pp. 277–284.

[28] Karoliina Tainio, Antonios Athanasiou, Kari AO Tikkinen, Riikka Aaltonen, Jovita Cárdenas, Sivan Glazer-Livson, Maija Jakobsson, Kirsi Joronen, Mari Kiviharju, Karolina Louvanto et al. 'Clinical course of untreated cervical intraepithelial neoplasia grade 2 under active surveillance: systematic review and meta-analysis'. In: *bmj* 360 (2018), k499.

[29] Susanne K Kjær, Kirsten Frederiksen, Christian Munk and Thomas Iftner. 'Long-term absolute risk of cervical intraepithelial neoplasia grade 3 or worse following human papillomavirus infection: role of persistence'. In: *Journal of the National Cancer Institute* 102.19 (2010), pp. 1478–1488.

[30] WHO. *Schistosomiasis*. Accessed: 2018-06-26. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/schistosomiasis.

[31] CDC. *CDC-Schistosomiasis - Biology*. 2012. URL: https://www.cdc.gov/parasites/schistosomiasis/biology.html/https://www.cdc.gov/parasites/schistosomiasis/biology.html (visited on 23/01/2020).

[32] Liu Qunhua, Zhang Jiawen, Liu Bozhao, Peng Zhilan, Zhang Huijie, Wang Shaoying, Mei Delun and Lee-Nah Hsu. *Investigation of association between female genital tract diseases and Schistosomiasis japonica infection*. Tech. rep. 2000, pp. 179–183.

[33] Victor M. Areán. 'Manson's schistosomiasis of the female genital tract'. In: *American Journal of Obstetrics and Gynecology* 72.5 (1956), pp. 1038–1053. ISSN: 00029378. DOI: 10.1016/0002-9378(56)90070-9.

[34] Gabriele Poggensee, Ireen Kiwelu, Margret Saria, Joachim Richter, Ingela Krantz and Hermann Feldmeier. 'Schistosomiasis of the lower reproductive tract without egg excretion in urine.' In: *The American journal of tropical medicine and hygiene* 59.5 (1998), pp. 782–783.

[35] Frank Cole Madden. 'A Case of Bilhazia of the Vagina'. In: *The Lancet* 153.3956 (1899), p. 1716. ISSN: 01406736. DOI: 10.1016/S0140-6736(01)51385-9.

[36] A Harouny and H Pedersen. 'Pelveo-peritoneal schistosomiasis as a cause of primary infertility'. In: *International Journal of Gynecology & Obstetrics* 27.3 (1988), pp. 467–469.

[37] OM Attia. 'A critical study of bilharziasis of the uterus'. In: *BJOG: An International Journal of Obstetrics & Gynaecology* 69.2 (1962), pp. 334–336.

[38] Gertrud Helling-Giese, Aimee Sjaastad, Gabriele Poggensee, Eyrun Floerecke Kjetland, Joachim Richter, Lester Chitsulo, Newton Kumwenda, Paul Racz, Borghild Roald, Svein Gunnar Gundersen et al. 'Female genital schistosomiasis (FGS): relationship between gynecological and histopathological findings'. In: *Acta tropica* 62.4 (1996), pp. 257–267.

[39] IARC Working Group on the Evaluation of Carcinogenic Risk to Humans. 'Schistosomes, Liver Flukes and Helicobacter pylori'. In: (1994).

[40] Gabriele Poggensee and Hermann Feldmeier. 'Female genital schistosomiasis: facts and hypotheses'. In: *Acta tropica* 79.3 (2001), pp. 193–210.

[41] H Feldmeier, GA Gastl, U Poggensee, C Kortmann, AA Daffalla and HH Peter. 'Relationship between intensity of infection and immunomodulation in human schistosomiasis. I. Lymphocyte subpopulations and specific antibody responses.' In: *Clinical and experimental immunology* 60.2 (1985), p. 225.

[42] Nathalie Ylitalo, Per Sørensen, Agnetha M Josefsson, Patrik KE Magnusson, Per Kragh Andersen, Jan Pontén, Hans-Olov Adami, Ulf B Gyllensten and Mads Melbye. 'Consistent high viral load of human papillomavirus 16 and risk of cervical carcinoma in situ: a nested case-control study'. In: *The Lancet* 355.9222 (2000), pp. 2194–2198.

[43] Vicky Gent, Rebecca Waihenya, Lucy Kamau, Ruth Nyakundi, Peris Ambala, Thomas Kariuki and Lucy Ochola. 'An investigation into the role of chronic Schistosoma mansoni infection on Human Papillomavirus (HPV) vaccine induced protective responses'. In: *PLoS neglected tropical diseases* 13.8 (2019), e0007704.

[44] P Moubayed, A Ziehe, J Peters, H Mwakyoma and D Schmidt. 'Carcinoma of the uterine cervix associated with schistosomiasis and induced by human papillomaviruses'. In: *International Journal of Gynecology & Obstetrics* 49.2 (1995), pp. 175–179.

[45] Anoop D Shah, Carlos Martinez and Harry Hemingway. 'The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records'. In: *BMC medical informatics and decision making* 12.1 (2012), p. 88.

[46] Kevin J Biese, Cory R Forbach, Richard P Medlin, Timothy F Platts-Mills, Matthew J Scholer, Brenda McCall, Frances S Shofer, Michael LaMantia, Cherri Hobgood, JS Kizer et al. 'Computer-facilitated Review of Electronic Medical Records Reliably Identifies Emergency Department Interventions in Older Adults'. In: *Academic Emergency Medicine* 20.6 (2013), pp. 621–628.

[47] Antony Hardjojo, Arunan Gunachandran, Long Pang, Mohammed Ridzwan Bin Abdullah, Win Wah, Joash Wen Chen Chong, Ee Hui Goh, Sok Huang Teo, Gilbert Lim, Mong Li Lee et al. 'Validation of a Natural Language Processing Algorithm for Detecting Infectious Disease Symptoms in Primary Care Electronic Medical Records in Singapore'. In: *JMIR medical informatics* 6.2 (2018).

[48] Vianney Jouhet, Georges Defossez, Anita Burgun, Pierre Le Beux, P Levillain, Pierre Ingrand, Vincent Claveau et al. 'Automated classification of free-text pathology reports for registration of incident cases of cancer'. In: *Methods of information in medicine* 51.3 (2012), p. 242.

[49] Muhubiri Kabuyaya, Moses John Chimbari, Tawanda Manyangadze and Samson Mukaratirwa. 'Schistosomiasis risk factors based on the infection status among school-going children in the Ndumo area, uMkhanyakude district, South Africa'. In: *Southern African Journal of Infectious Diseases* 32.2 (2017), pp. 67–72.

[50] Nkosinathi Banhela, Myra Taylor, Siphosenkosi Gift Zulu, Linnea Sund Strabo, Eyrun Floerecke Kjetland and Svein Gunnar Gundersen. 'Environmental factors influencing the distribution and prevalence of Schistosoma haematobium in school attenders of ILembe and uThungulu Health Districts, KwaZulu-Natal Province, South Africa'. In: *Southern African Journal of Infectious Diseases* 32.4 (2017), pp. 132–137.

[51] Lynette Denny and Louise Kuhn. 'Cervical cancer prevention and early detection from a South African perspective'. In: *South African Health Review* 2017.1 (2017), pp. 189–195.

[52] The Cancer Association of South Africa. *CANSA Assisting in KZN Cervical Cancer Screening Programmes*. 2018. URL: https://www.cansa.org.za/cansa-assisting-in-kzn-cervical-cancer-screening-programmes/ (visited on 25/01/2020).

[53] NHLS. *National Health Laboratory Service Annual Report 2016/17*. Accessed: 2018-08-02. 2017. URL: http://www.nhls.ac.za/assets/files/an_report/NHLS_AR_2016_17.pdf.

[54] GG Sherman, RR Lilian, S Bhardwaj, S Candy and P Barron. 'Laboratory information system data demonstrate successful implementation of the prevention of motherto-child transmission programme in South Africa'. In: *South African medical journal* 104.3 (2014), pp. 235–238.

[55] SNOMED International. *SNOMED - Who we are*. 2019. URL: http://www.snomed.org/snomed-international/who-we-are.

[56] Elizabeth D Liddy and E D Liddy. 'Natural Language Processing Recommended Citation'. In: *Encyclopedia of Library and Information Scienc*. 2001. URL: https://surface.syr.edu/cnlp.

[57] Mariette Awad and Rahul Khanna. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress, 2015.

[58] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes and Donald Brown. 'Text classification algorithms: A survey'. In: *Information* 10.4 (2019), p. 150.

[59] Tara Rawat and Vineeta Khemchandani. 'Feature Engineering (FE) Tools and Techniques for Better Classification Performance'. In: *International Journal of Innovations in Engineering and Technology (IJIET)* 8.2 (2017).

[60] Vandana Korde and C Namrata Mahender. 'Text classification and classifiers: A survey'. In: *International Journal of Artificial Intelligence & Applications* 3.2 (2012), p. 85.

[61] TFIDF. *A Single-Page Tutorial - Information Retrieval And Text Mining*. 2014. URL: http://www.tfidf.com/.

[62] Luying Liu, Jianchu Kang, Jing Yu and Zhongliang Wang. 'A comparative study on unsupervised feature selection methods for text clustering'. In: *2005 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE. 2005, pp. 597–601.

[63] Ron Kohavi, George H John et al. 'Wrappers for feature subset selection'. In: *Artificial intelligence* 97.1-2 (1997), pp. 273–324.

[64] Piyatida Inrak and Sukree Sinthupinyo. 'Applying latent semantic analysis to classify emotions in Thai text'. In: *2010 2nd International Conference on Computer Engineering and Technology*. Vol. 6. IEEE. 2010, pp. V6–450.

[65] Sebastian Raschka. *Predictive modeling , supervised machine learning , and pattern classification — the big picture*. 2014. URL: https://sebastianraschka.com/Articles/2014{_}intro{_}supervised{_}learning.html (visited on 28/01/2020).

[66] Gerard Burger, Ameen Abu-Hanna, Nicolette de Keizer and Ronald Cornet. 'Natural language processing in pathology: a scoping review'. In: *Journal of clinical pathology* 69.11 (2016), pp. 949–955.

[67] Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng. 'An Improved Random Forest Classifier for Text Categorization.' In: *JCP* 7.12 (2012), pp. 2913–2920.

[68] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012.

[69] Scikit-learn. *sklearn.dummy.DummyClassifier — scikit-learn 0.22.1 documentation*. 2019. URL: https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html (visited on 28/01/2020).

[70] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang and Wayne Xu. 'Applications of support vector machine (SVM) learning in cancer genomics'. In: *Cancer Genomics-Proteomics* 15.1 (2018), pp. 41–51.

[71] Yangu Zhang, Saiping Chen and Yi Wan. 'An intelligent algorithm based on grid searching and cross validation and its application in population analysis'. In: *2009 International Conference on Computational Intelligence and Natural Computing*. Vol. 2. IEEE. 2009, pp. 96–99.

[72] Leo Breiman. 'Random forests'. In: *Machine learning* 45.1 (2001), pp. 5–32.

[73] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[74] Pier Paolo Ippolito. *Hyperparameters Optimization*. 2019. URL: https://towardsdatascience.com/hyperparameters-optimization-526348bb8e2d (visited on 30/01/2020).

[75] Md Nasim Adnan and Md Zahidul Islam. 'One-vs-all binarization technique in the context of random forest'. In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2015, pp. 385–390.

[76] Hennie Botha, Bruno Cooreman, Greta Dreyer, Gerhard Lindeque, Arrie Mouton, Franco Guidozzi, Tony Koller, Trudy Smith, Anwar Hoosen, Louis Marcus et al. 'Cervical cancer and human papillomavirus: South African guidelines for screening and testing'. In: *Southern African Journal of Gynaecological Oncology* 2.1 (2010), pp. 23–26.

[77] National Department of Health. *Cervical cancer prevention and control policy*. Tech. rep. 2017. arXiv: 1560-4381.

[78] Arthi Vijayaraghavan, Molly Efrusy, Gerhard Lindeque, Greta Dreyer and Christopher Santas. 'Cost effectiveness of high-risk HPV DNA testing for cervical cancer screening in South Africa'. In: *Gynecologic Oncology* 112.2 (2009), pp. 377–383.

[79] Alliance for Cervical Cancer Prevention Members. *Natural history of cervical cancer: even infrequent screening of older women saves lives.* 2003.

[80] WAA Tjalma. 'HPV negative cervical cancers and primary HPV screening'. In: *Facts, views & vision in ObGyn* 10.2 (2018), p. 107.

[81] Karl Ulrich Petry, Clemens Liebrich, Alexander Luyten, Martina Zander and Thomas Iftner. 'Surgical staging identified false HPV-negative cases in a large series of invasive cervical cancers'. In: *Papillomavirus Research* 4 (2017), pp. 85–89.

[82] GM Clifford, JS Smith, M Plummer, N Munoz and S Franceschi. 'Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis'. In: *British journal of cancer* 88.1 (2003), pp. 63–73.

[83] Eduardo González-Bosquet, Sergi Fernandez, Sally Sabra and Jose M Lailla. 'Negative HPV testing among patients with biopsy-proven cervical intraepithelial neoplasia grade 2/3 or cervical cancer'. In: *International Journal of Gynecology & Obstetrics* 136.2 (2017), pp. 229–231.

[84] Inmaculada Nicolás, Lorena Marimon, Esther Barnadas, Adela Saco, Leonardo Rodríguez-Carunchio, Pere Fusté, Cristina Martí, Adriano Rodriguez-Trujillo, Aureli Torne, Marta del Pino et al. 'HPV-negative tumors of the uterine cervix'. In: *Modern Pathology* 32.8 (2019), pp. 1189–1196.

[85] KM Bosompem, Irene A Bentum, J Otchere, WK Anyan, CA Brown, Y Osada, S Takeo, S Kojima and N Ohta. 'Infant schistosomiasis in Ghana: a survey in an irrigation community'. In: *Tropical medicine & international health* 9.8 (2004), pp. 917–922.

[86] Amadou Garba, Nouhou Barkiré, Ali Djibo, Mariama S Lamine, Boubacar Sofo, Anouk N Gouvras, Elisa Bosqué-Oliva, Joanne P Webster, J Russell Stothard, Jürg Utzinger et al. 'Schistosomiasis in infants and preschool-aged children: infection in a single Schistosoma haematobium and a mixed S. haematobium–S. mansoni foci of Niger'. In: *Acta tropica* 115.3 (2010), pp. 212–219.

[87] UF Ekpo, OM Odeyemi, SO Sam-Wobo, OB Onunkwor, HO Mogaji, AS Oluwole, HO Abdussalam and JR Stothard. 'Female genital schistosomiasis (FGS) in Ogun state, Nigeria: a pilot survey on genital symptoms and clinical findings'. In: *Parasitology Open* 3 (2017).

[88] Aoife M Doyle, Sue Napierala Mavedzenge, Mary L Plummer and David A Ross. 'The sexual behaviour of adolescents in sub-Saharan Africa: patterns and trends from national surveys'. In: *Tropical Medicine & International Health* 17.7 (2012), pp. 796–807.

[89] Stanley M F Luchters, Davy Vanden Broeck, Matthew F Chersich, Annalene Nel, Wim Delva, Kishor Mandaliya, Christophe E Depuydt, Patricia Claeys, John-Paul Bogers and Marleen Temmerman. 'Association of HIV infection with distribution and viral load of HPV types in Kenya: a survey with 820 female sex workers'. In: *BMC Infectious Diseases* 10.1 (2010), p. 18. ISSN: 1471-2334. DOI: 10.1186/1471-2334-10-18. URL: http://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-18.

Appendices

A

PLAGIARISM DECLARATION

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I **Carole Metekoua Temdemnou** (Student number: **1363322**) am a student registered for the degree of **Master of Science in Epidemiology** in the academic year **2019**.

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: **23 February 2020**

26/04/2015
1

B

SAMPLE HISTOLOGY REPORT

"HISTOPATHOLOGY REPORT

NATURE OF SPECIMEN
PUNCH BIOPSY

SITE OF SPECIMEN
CERVIX

MACROSCOPIC DESCRIPTION
MULTIPLE PIECES OF TISSUE WERE RECEIVED, THE LARGEST MEASURED 4MM.

MICROSCOPIC DESCRIPTION
SECTIONS DEMONSTRATE A VERY FRAGMENTED BIOPSY IN WHICH A CIN III WITH HPV CHANGES WAS NOTED. ON THIS BIOPSY NO SIGNS OF INFILTRATING MALIGNANCY WAS SEEN.

CONCLUSION
PUNCH BIOPSY : CERVIX: CIN III WITH HUMAN PAPILLOMA VIRUS CHANGES. SPECIMEN VERY FRAGMENTED. NO INFILTRATING MALIGNANCY IS SEEN.
"

C

LIST OF VARIABLES

1. unique_patient_id
2. episode_no
3. folder_no
4. date_of_birth
5. gender
6. race
7. province_name
8. district_name
9. sub_district_name
10. facility_name
11. ward_name
12. date_reviewed
13. specimen_type
14. result_text
15. test_method_name
16. snomed_code
17. snomed_name

D

ETHICAL CLEARANCE CERTIFICATE

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Ms CM Temdemnou

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M181160

NAME: (Principal Investigator)	Ms CM Temdemnou
DEPARTMENT:	School of Public Health Medical School University
PROJECT TITLE:	An investigation of the presence of cervical schistosomiasis in patients with cervical cancer and pre-cancer lesions: a text mining approach
DATE CONSIDERED:	30/11/2018
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Mr M Mapundu and Dr M Muchengeti
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	28/01/2019

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 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
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**. When a funder requires annual re-certification, the application date will be one year after the date of the meeting when the study was initially reviewed. In this case, the study was initially reviewed in **November** and will therefore reports and re-certification will be due early in the month of **November** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

January 30, 2019
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

E

APPROVAL LETTER TO ACCESS NHLS DATA

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

30 January 2019

Applicant: Ms Carole Metekoua Temdemnou
Institution: University of the Witwatersrand
Department: Epidemiology – Public Health Informatics
Email: 1363322@students.wits.ac.za
Cell: 060 526 6945

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

Your application to undertake a research project "**An Investigation of the Presence of Cervical Schistosomiasis in Patients with Cervical Cancer and Pre-Cancer Lesions: A Text Mining Approach**" 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 ***without patient names*** to conduct the proposed study as outlined in the submitted application.

Please note that final 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.

- Processes are discussed with the relevant NHLS departments (i.e. Information Management Unit and Operations Office) 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.
- NHLS Data cannot be used to track patients as no pre-approval/consent is obtained from Patients.

Please note that this letter constitutes approval by the NHLS Academic Affairs and Research Office. Any data related queries may be directed to NHLS Corporate Data Warehouse, contact number: 011 386 6074 email: zarina.sabat@nhls.ac.za

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

F

SENSITIVITY ANALYSIS: DETAILED RESULTS

F.1 YEARLY PREVALENCE OF SCHISTOSOMIASIS IN CERVICAL PRECANCEROUS LESIONS AND CERVICAL CANCER PATIENTS OVER TIME

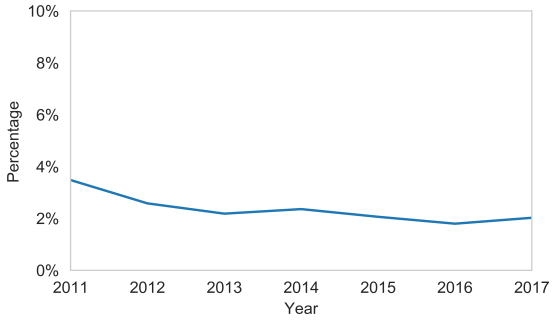


Figure F.1: Prevalence of schistosomiasis in pre-cancer and cervical cancer patients over time (sensitivity analysis)

F.2 PREVALENCE OF SCHISTOSOMIASIS AMONG WOMEN WITH CERVICAL PRECANCEROUS LESIONS

Table F.1: Prevalence of schistosomiasis in women with cervical precancerous lesions (sensitivity analysis)

Pathology	**Schistosomiasis**		
	No	%	Total
Pre-cancer	464	2.59	17,907
HSIL	405	2.61	15,509
LSIL	15	2.36	636
Unknown	44	2.50	1,762

F.3 PREVALENCE OF SCHISTOSOMIASIS AMONG WOMEN WITH CERVICAL CANCER

Table F.2: Prevalence of schistosomiasis in women with cervical cancer (sensitivity analysis)

Pathology	**Schistosomiasis**		
	No	%	Total
Cancer	54	0.96	5,652
Squamous cell	44	0.99	4,442
Adenocarcinoma	7	1.18	593
Adenosquamous	0	0.00	64
Other	3	0.54	553

F.4 AGE DIFFERENCE BETWEEN SCHISTOSOMIASIS NEGATIVE AND SCHISTOSOMIASIS UNKNOWN PATIENTS

Table F.3: Age difference between schistosomiasis negative and schistosomiasis unknown

	Mean (STD)	**Difference**	**p-value (differnce != 0)**
Pre-cancer			
Schistosomiasis Negative	40.45 (9.86)	0.59	0.8941
Schistosomiasis Unknown	40.98 (10.78)		
Cervical cancer			
Schistosomiasis Negative	50.42 (13.78)	1.11	0.3027
Schistosomiasis Unknown	51.54 (13.97)		

F.5 DISTRIBUTION OF AGE AT PRECANCEROUS LESION AND CERVICAL CANCER DIAGNOSIS

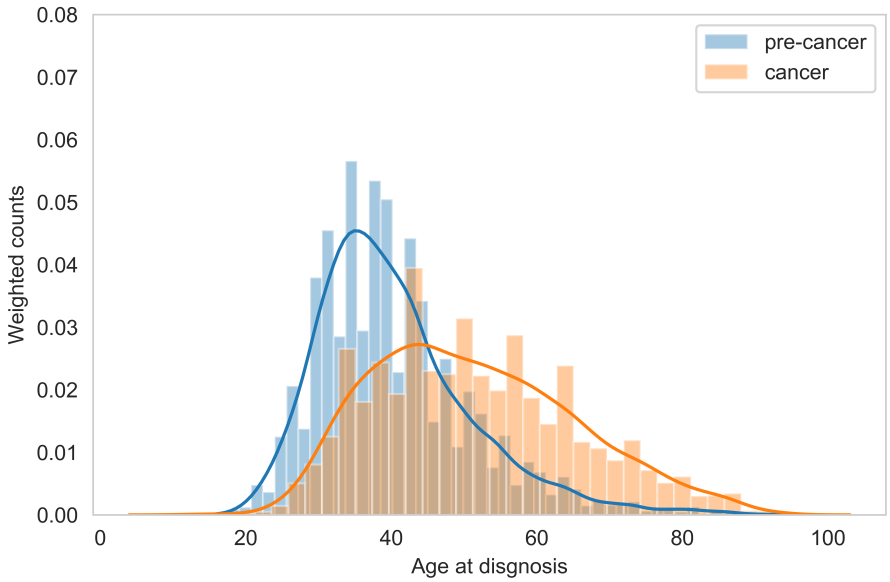


Figure F.2: Distribution of pathologies by age at diagnosis (sensitivity analysis)

F.6 RELATIONSHIP BETWEEN SCHISTOSOMIASIS AND AGE AT PRECANCEROUS LESIONS DIAGNOSIS

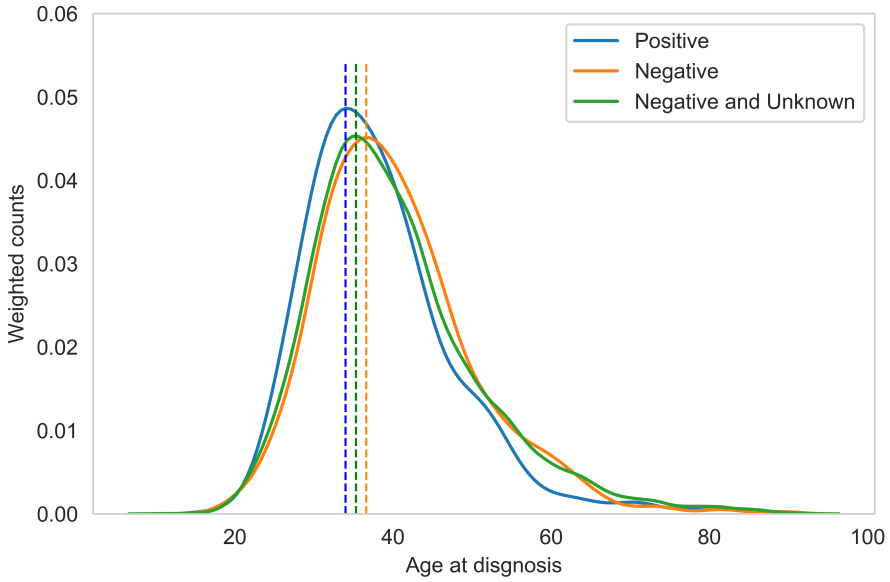


Figure F.3: Distribution of age at pre-cancer diagnosis by schistosomiasis status (sensitivity analysis)

Table F.4: Age difference between schistosomiasis status at cervical precancerous lesion diagnosis

	Mean (STD)	**Difference**	**p-value (difference >0)**
Schistosomiasis Negative & Unknown	40.39 (10.75)	2.07	<0.0001
Schistosomiasis Positive	38.32 (9.46)		

F.7 RELATIONSHIP BETWEEN SCHISTOSOMIASIS AND AGE AT CANCER DIAGNOSIS

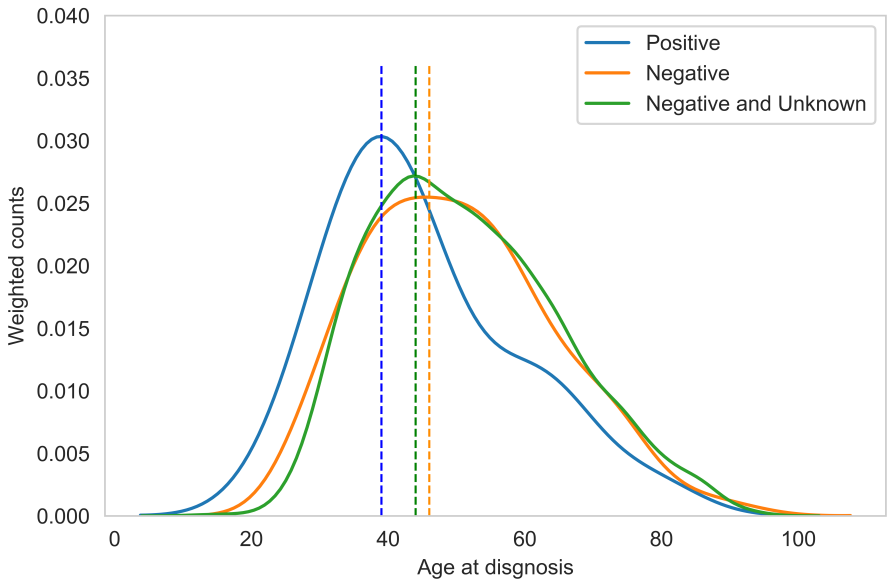


Figure F.4: Distribution of age at cancer diagnosis by schistosomiasis status (sensitivity analysis)

Table F.5: Age difference between schistosomiasis status at cervical cancer diagnosis

	Mean (STD)	**Difference**	**p-value (difference >0)**
Schistosomiasis Negative & Unknown	51.50 (13.96)	6.39	<0.0010
Schistosomiasis Positive	45.12 (13.70)		

G

ORIGINALITY REPORT

1363322:Metekoua_ResearchReport.pdf

ORIGINALITY REPORT

8%	4%	6%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	P. Belitsos, D. Papoutsis, A. Rodolakis, S. Mesogitis, A. Antsaklis. "Three-dimensional power Doppler ultrasound for the study of cervical cancer and precancerous lesions", Ultrasound in Obstetrics & Gynecology, 2012 Publication	<1%
2	Ramadhani Chambuso. "Cervical cancer and HIV/AIDS - Influence of HIV/AIDS on cervical precancerous lesions and invasive cervical cancer in Morogoro region. Tanzania", Repositório Aberto da Universidade do Porto, 2014. Publication	<1%
3	"Applications of Support Vector Machine (SVM) Learning in Cancer Genomics", Cancer Genomics & Proteomics, 2018 Publication	<1%
4	"Text and Web Mining", Studies in Computational Intelligence, 2006 Publication	<1%