Abstract

Background: Cancer is an increasingly multidimensional global health problem that demands continuous and concerted efforts toward diagnosis, treatment and management. Developing countries, including South Africa, are expected to experience an increasing burden of cancer morbidity and mortality because of a late-stage diagnosis, poor awareness, lack of biomedical screening and limited resources for treatment. Nonetheless, there is a growing trend to improve treatment outcomes by understanding the epidemiology and dynamics of cancer. The multifaceted clinical approach toward cancer has resulted in an unprecedented production of a large amount of complex and heterogeneous clinical data. These include information describing disease symptoms, diagnostics reports, drug usage and clinical outcomes, often presented in a structured, semi-structured and unstructured data format. Physicians' intuition may not efficiently integrate these mixed data types in this situation for a realistic prognosis and treatment. Therefore, the next generation of cancer treatments may be found not by empirical science but by data scientists developing mathematical and statistical models to describe the nature, epidemiology and dynamics of cancer at various stages of the disease. Hence, cancer research is attracting the application of high-performance computing and big data analytics to enhance diagnosis, prognosis, and treatment. It has been shown that the use of data mining (DM), text mining (TM) and machine learning (ML) has efficiently and effectively uncovered trends from cancer data to support oncologists in decision-making. However, Africa is yet to embrace clinical decision-making in cancer diagnosis and prognosis using these algorithms. Current ML models for cancer detection and prognostic classifications have been chiefly designed for developed countries. Nevertheless, findings in the developed countries may not truly reflect the situation in an African population with higher diversity.

Aim: This research aimed to contribute to knowledge discovery using descriptive and predictive analytics, specifically in colorectal, breast and prostate cancers. One of the objectives of this study is to develop and evaluate methods for extracting relevant information to classify cancer pathology reports. This will be achieved by analysing a large pool of historical pathology reports and assessing performance across cancer types. Secondly, there is only limited study on biological and clinical parameters which drives cancer prognosis, especially in Africa. This study

aimed to develop a rule-based method to automatically extract important cancer prognostic parameters from free-text pathology reports, transform them into structured data, and uncover the trend of these parameters over the years. In addition, studies have shown that hospital length of stay (LOS) following surgical cancer resection varies across countries. However, cancer studies on hospital length of stay are unknown in an African population. This study aimed to present an avenue to understand the dynamics and prognostic classification of patient risk groups post-surgical recession. Finally, cancer recurrence and patient survival studies are limited in an African population. We explored the feasibility of integrating statistical and machine learning algorithms for the first time to achieve higher predictive performance and interpretability models for cancer recurrence and patient survival.

method: We conducted secondary data analysis using two data sources based on the study aims. The first data set consists of 181,000 breast, colorectal and prostate cancer patients whose diagnostic tests were carried out at the National Health Laboratory Services (NHLS) from 2008 to 2019. Each patient is described with demographic information, pathology reports, SNOMED morphology and topography codes. The data set addressed this study's first and second objectives, focusing on information extraction from pathology reports, text classification and prediction, and trend analysis. The second data set consisted of 761 colorectal cancer patients diagnosed in Johannesburg. The patients were treated in private and public hospitals from 2015 to 2019 and followed until 2020. Each patient is described with more features detailing their demographic, clinical and histological information. We used this data set to address the third and fourth objectives, which focused on prognostic classification and prediction, with machine learning and statistical algorithms.

Results: The incidence of breast, colorectal and prostate cancer diagnostic tests trends is increasing over the year. In recent years, pathology reporting on these cancers has increased in text length, with more malignancy than benign. The result also showed inconsistencies and incompleteness in reporting each year and across the year of study. However, our methodologies could standardise and accurately extract essential parameters with high performance. The second objective showed that the developed rule-based method achieved high accurate annotation for all the parameters extracted, with performance measures ranging from 83% -100%. The trend analysis result showed significant trends in the proportion of molecular subtypes and Ki67, comparable to previous studies. Further, we observed that the median hospital LOS post-surgical resection was higher than those reported in the developed world. There is no significant difference in hospital LOS when comparing private and public facilities. Factors predisposing patients to prolonged hospital LOS include preoperative, perioperative and postoperative parameters. However, surgical complications are still the primary driver of prolonged hospital LOS. Finally, younger patients experienced higher recurrence than older patients but had a comparable survival rate. There was an improved survival for patients

treated in a private hospital compared to those treated in public hospitals.

conclusion: We developed reproducible frameworks that can form the basis for future studies in South Africa, using DM, TM and ML algorithms. This study supports a nationally agreed standard in pathology reporting and the use of these algorithms for encoding, classifying, and producing high-quality information abstractions for cancer diagnosis, prognosis, incidence reporting, and research. The association established in this study may enable clinicians to understand the diagnostic and prognostic factors that influence patients' health status and implement changes in patient care pathways. Clinicians can employ this type of study as a simple and quick test to flag patients at high risk. Such a strategy would improve clinical outcomes, but it is also likely to improve efficiency, favourably impacting the cost of care for cancer patients. The findings of this study can be generalised not only to the population of South African cancer patients but in other Sub-Saharan African countries with similar trends in urbanisation and dynamics in cancer epidemiology.