






Using Systematized Nomenclature of Medicine clinical term codes to assign histological findings for prostate biopsies in the Gauteng province, South Africa: Lessons learnt

**Authors:**

Naseem Cassim^{1,2} 
Ahsan Ahmad³ 
Reubina Wadee⁴ 
Jaya A. George⁵ 
Deborah K. Glencross^{1,2} 

Affiliations:

¹National Health Laboratory Service(NHLS), National Priority Programme, Johannesburg, South Africa

²Department of Molecular Medicine and Haematology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

³Department of Urology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

⁴Department of Anatomical Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

⁵Department of Chemical Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

Corresponding author:

Naseem Cassim,
naseem.cassim@wits.ac.za

Dates:

Received: 11 Sept. 2018
Accepted: 24 June 2020
Published: 28 Sept. 2020

Read online:

Scan this QR code with your smart phone or mobile device to read online.

Background: Prostate cancer (PCa) is a leading male neoplasm in South Africa.

Objective: The aim of our study was to describe PCa using Systemized Nomenclature of Medicine (SNOMED) clinical terms codes, which have the potential to generate more timely data.

Methods: The retrospective study design was used to analyse prostate biopsy data from our laboratories using SNOMED morphology (M) and topography (T) codes where the term 'prostate' was captured in the narrative report. Using M code descriptions, the diagnosis, sub-diagnosis, sub-result and International Classification of Diseases for Oncology (ICD-O-3) codes were assigned using a lookup table. Topography code descriptions identified biopsies of prostatic origin. Lookup tables were prepared using Microsoft Excel and combined with the data extracts using Access. Contingency tables reported M and T codes, diagnosis and sub-diagnosis frequencies.

Results: An M and T code was reported for 88% ($n = 22\ 009$) of biopsies. Of these, 20 551 (93.37%) were of prostatic origin. A benign diagnosis (ICD-O-3:8000/0) was reported for 10 441 biopsies (50.81%) and 45.26% had a malignant diagnosis ($n = 9302$). An adenocarcinoma (8140/3) sub-diagnosis was reported for 88.16% of malignant biopsies ($n = 8201$). An atypia diagnosis was reported for 760 biopsies (3.7%). Inflammation (39.03%) and hyperplasia (20.82%) were the predominant benign sub-diagnoses.

Conclusion: Our study demonstrated the feasibility of generating PCa data using SNOMED codes from national laboratory data. This highlights the need for extending the results of our study to a national level to deliver timeous monitoring of PCa trends.

Keywords: prostate biopsy; Systemized Nomenclature of Medicine; SNOMED; morphology; topography; prostate cancer and adenocarcinoma.

Introduction

Prostate cancer (PCa) was a leading male cancer in South Africa in 2012, while globally it is the second most frequently diagnosed neoplasm.¹ A 2012 global cancer study reported an estimated age-specific incidence rate of 67.9 per 100 000 for South Africa, with an associated mortality rate of 26.4 per 100 000.^{1,2} Presentation with highly aggressive PCa in African men in South Africa has been described.^{3,4}

A national cancer control programme was recommended by the World Health Organization, with the aim of identifying priorities and assigning resources to sustain progress towards the reduction of cancer incidence and mortality as well as to improve the quality of life for patients with cancer.⁵ This would be attained through the equitable implementation of evidence-based strategies for prevention, early detection, treatment and palliative care, in the context of optimal use of limited healthcare resources.⁶ The 2013–2020 Global Action Plan for the Prevention and Control of Non-Communicable Diseases aims to achieve a 25% reduction in the relative risk of premature mortality from cancers and other non-communicable diseases.⁷ This initiative emphasises the

How to cite this article: Cassim N, Ahmad A, Wadee R, George JA, Glencross DK. Using Systematized Nomenclature of Medicine clinical term codes to assign histological findings for prostate biopsies in the Gauteng province, South Africa: Lessons learnt. *Afr J Lab Med*. 2020;9(1), a909. <https://doi.org/10.4102/ajlm.v9i1.909>

Copyright: © 2020. The Authors. Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

importance of both surveillance and disease registries that should be integrated into existing health information systems to improve the availability of high-quality data.⁷ Current cancer registry reporting is not integrated into any hospital information system. The purpose of a cancer registry is to establish and maintain a cancer incidence reporting system that informs planning of cancer control programs.⁸ Cancer registries should typically publish annual data within 28 months after the close of the year in which the incident case was diagnosed.⁹ A delay in cancer registry reporting is a major limitation for understanding PCa trends.⁹

The National Cancer Registry of South Africa is a passively reported registry that used the International Classification of Diseases for Oncology (ICD-O-3) – recommended international methodology – to manually code pathology reports.^{10,11,12} The ICD-O-3 describes both the anatomical site and cell type and behaviour (malignant or benign biopsy).¹³ The National Cancer Registry (NCR) had only reported data for 2014 in 2018.¹⁴ Cases are manually coded, and this results in a reporting delay. Based on the surveillance, Epidemiology, and End Results programme standard, the 2015 report should have been published by 2018.⁹ The programme itself, on which the South African reporting standard is based, is comprised of 11 registries in five states and six metropolitan areas in the United States which generate annual cancer data approximately 28 months after diagnosis.⁸

Both internationally and nationally, cancer surveillance is defined as an ongoing, timely, and systematic collection and analysis of cancer data to assess risk factors, screening, diagnosis, and cancer incidences and deaths.⁹ The aim of surveillance is to analyse and disseminate cancer data to identify challenges and opportunities in the delivery of timely cancer control programmes.⁹

The cancer registry reporting is a time-intensive process requiring trained coders to review each narrative biopsy report individually and to manually add the applicable ICD-O-3 codes to the reporting system.¹⁵ This can take hundreds of hours of manual database building to report PCa data. The coders would have to add both topography and morphology ICD-O-3 codes for each narrative biopsy report reviewed.¹⁵ This manually intensive process prolongs time between diagnosis and reporting of identified PCa cases for surveillance. PCa reporting is required at least 28 months after diagnosis (~2 years) to understand changes in incidence.⁸ Without timely data this would not be possible. Therefore, new approaches are required to reduce or automate the coding process to provide more timely cancer data.

Some studies have used approaches such as natural language processing and text mining.^{16,17,18} For our study, we decided to use the Systematized Nomenclature of Medicine (SNOMED) clinical terms.¹⁹ The Systematized Nomenclature of Medicine is a comprehensive and precise health terminology used globally that incorporates a structured list of health terms or concepts.^{19,20} One of the benefits of using

SNOMED is that codes can be mapped to other coding systems to facilitate interoperability.¹⁹ All references to SNOMED relate to SNOMED CT.²¹ In an Anatomical Pathology setting, SNOMED CT is used to capture the histological finding in the form of morphology (M) and topography (T) codes. The M and T codes are captured directly in the laboratory information system (LIS) by the pathologist after examining prostate cores. The same histological findings are also reported as a narrative pathology report.¹⁹ The M and T code(s) are captured separately in defined test items in the LIS, which has a dictionary of all the SNOMED codes that may be reported and the anatomical pathologist selects the appropriate codes to add based on the histological findings. For each biopsy, more than one SNOMED M/T code may be captured. For example (personal communication, Vreede H, August 12, 2017, sharing the laboratory information system SNOMED CT code table extract), for a biopsy of prostatic origin with an adenocarcinoma finding, the following codes would be reported (description in brackets); (1) T-28 000 or T-92 000 (Prostatic structure), (2) M-80 003 (Malignant neoplasm, primary) and (3) M-81 403 (Adenocarcinoma, no subtype).

Studies outside South Africa have used SNOMED codes to transform laboratory and other reported data for cancer registry reporting.²² A good example is the Danish PCa registry that analysed SNOMED data for biopsies with PCa, that is histologically verified.²² This study confirmed that the SNOMED codes generated clinically useful data.²²

Our study described here is the first attempt in South Africa to investigate reporting PCa using SNOMED codes by collating this information contained in the national laboratory data repository. It is anticipated that this could, in future strengthen surveillance activities. Additionally, using the SNOMED codes to report on patients without PCa could provide important presentation information that is currently poorly understood.

The majority of local studies have manually coded biopsy reports to extract PCa information. The development of SNOMED CT lookup tables have the potential to automate this process and improve timely PCa reporting. The objective of this study was to describe the methodology used to report PCa data using SNOMED CT lookup tables.

Methods

Ethical considerations

Ethics clearance for this study was obtained from the University of the Witwatersrand (M170419). This study used national laboratory data that does not contain any patient identifiers. No patient recruitment was required.

Study design

This was a retrospective descriptive study that analysed prostate biopsy data between 2006 and 2016 for men ≥ 30 years in the Gauteng province.

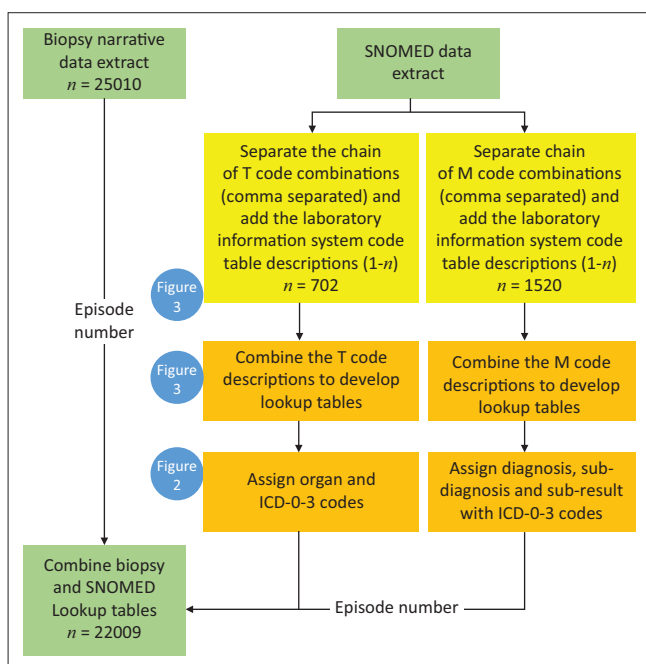
Data extraction and preparation

Retrospective prostate biopsy data for the period 2006–2016 were extracted from the national health laboratory repository of patient-related data where the term ‘prostate’ was captured in the narrative pathology report. Simple text mining approaches were used in the Netezza Aginity (Marlborough, Massachusetts, United States) query tool which employed pattern matching by fuzzy string search function.²³ Two data extracts were received: (1) prostate biopsy, and (2) chained M and T code(s) captured for each biopsy.

The prostate biopsy extract included the following variables: (1) episode number, (2) unique patient identifier (generated using a probabilistic matching algorithm²⁴), (3) age, (4) gender, (5) race (where populated), (6) facility code, (7) facility name, (8) reviewed date, and (9) biopsy results text (unstructured narrative report detailing histological findings).

The SNOMED data extract included the following variables: (1) episode number, (2) chained (comma separated) M code(s) (comma delimited), e.g. M-00 100, M-43 000, M-72 450, and (3) chained T code/s, e.g. T-92 000, T-74 000).

From the prostate biopsy data extract, the unique SNOMED CT code combinations were extracted, and a lookup table



ICD-O, International Classification of Diseases for Oncology; SNOMED, Systemized Nomenclature of Medicine.

FIGURE 1: High-level overview of the steps taken to code the unique chained Systemized Nomenclature of Medicine (SNOMED) code combinations extracted from the biopsy narrative data extract. There were two separate SNOMED data extracts from the laboratory information system: morphology (M) and topography (T). The colour coding indicates the various processes; (1) green: data extracts, (2) yellow SNOMED code manipulation in preparation for lookup table development, and (3) orange: lookup tables with coded variables. The extracted unique chained SNOMED code combinations were used to prepare the two lookup tables to generate the following new coded variables; (1) organ, (2) organ ICD-O-3, (3) diagnosis, (4) diagnosis ICD-O-3, (5) sub-diagnosis, (6) sub-diagnosis ICD-O-3, and (7) sub-result (for an inflammation sub-diagnosis). The number of biopsies is indicated for each data extract. The blue circles indicate which figures provide additional details on each step.

was developed (Figure 1). The lookup tables were developed in a two-step process: (1) code manipulation to combine descriptions, and (2) coding the lookup tables. The blue circles in the figure indicate which figures provide additional details for each step, that is: data manipulation (Figure 2) and query to combine data extracts and lookup tables in a single database query (Figure 3).

Combining Systemized Nomenclature of Medicine descriptions for the lookup table

The chained SNOMED codes were provided in a format that could not readily be analysed and had to be separated into individual columns with the applicable descriptions added from the LIS code table. Data were prepared using Microsoft Excel (Microsoft Corporation, Redmond, Washington, United States)²⁵ (Figure 2). The unique code combinations and descriptions were extracted for the development of the lookup table.

Coding the Systemized Nomenclature of Medicine M and T lookup tables

We used the prepared M and T code descriptions to start populating the lookup tables. For the M lookup table, we used each unique code description combination to populate the following new variables: (1) diagnosis, (2) diagnosis ICD-O-3 code, (3) sub-diagnosis, (4) sub-diagnosis ICD-O-3 code, and (5) sub-result with guidance from an anatomical pathologist and a urologist. The team reviewed each code combination and assigned values to be captured in the lookup table. An example of the coding is provided for four biopsies in Table 1. We captured the matching ICD-O-3 codes for predominantly malignant findings. The diagnosis reports the overall biopsy finding, whereas the sub-diagnosis was used to differentiate the diagnosis, e.g. Benign, negative for malignancy (ICD-O-3: 8000/0) and ‘Hyperplasia’, respectively (Episode A). Assigning the malignant code descriptions was fairly easy, but we struggled with benign findings given the number of findings reported and the order thereof. To clarify coding, we defined the reporting order to assign a benign sub-diagnosis as follows: (1) inflammation, (2) hypertrophy, (3) hyperplasia, (4) edema, (5) atrophy, and (6) adenosis. This made it easier to assign these finding in logical order. The sub-result was used to identify the type of inflammation reported, e.g. acute, chronic, granulomatous, etc. Similarly, the T code lookup table reported the organ (Prostate/Other) and associated ICD-O-3 code (C61.9 for Prostate).

Combining the lookup tables with the prostate biopsy data

Microsoft Access (Microsoft Corporation, Redmond, Washington, United States) was used to combine the datasets in a single query: (1) prostate data extract table, (2) SNOMED M lookup table, and (3) SNOMED T lookup table. Tables were combined using a left outer join in a relational database (Figure 2).²⁶ This join type ensures that all rows from the prostate data extract table were populated with

Step 1: Unique SNOMED M code combination in data extract (comma separated)

SNOMED M CODES
M-00100,M-72440,M-43000

Step 2: Copy and paste SNOMED M code combination into a new column (to preserve the original value)

SNOMED M CODES	SNOMED M CODES Copy
M-00100,M-72440,M-43000	M-00100,M-72440,M-43000

Step 3: Use the Microsoft Excel text to column wizard to separate SNOMED M codes. Once separated, name columns (1-n)

SNOMED M CODES	SNOMED M1 CODE	SNOMED M2 CODE	SNOMED M3 CODE
M-00100,M-72440,M-43000	M-00100	M-72440	M-43000

Step 4: Name SNOMED M code description columns (1-n)

SNOMED M CODES	SNOMED M1 CODE	SNOMED M1 CODE DESCRIPTION	SNOMED M2 CODE	SNOMED M2 CODE DESCRIPTION	SNOMED M3 CODE	SNOMED M3 CODE DESCRIPTION
M-00100,M-72440,M-43000	M-00100		M-72440		M-43000	
M-00000	M-00000	Morphologic finding (finding)				
M-00001	M-00001	Morphology unknown (finding)				
M-00002	M-00002	Morphology not applicable (finding)				
M-00003	M-00003	Morphology not assigned in SNOMED (finding)				
M-00004	M-00004	Endometrium normal (finding)				
M-00005	M-00005	Villous atrophy (finding)				
M-00006	M-00006	Partial villous atrophy (finding)				
M-00100	M-00100	Normal tissue (finding)				

Step 5: Use the Microsoft Excel VLOOKUP function to add SNOMED code descriptions 1-n

SNOMED M1 CODE	SNOMED M1 CODE DESCRIPTION	SNOMED M2 CODE	SNOMED M2 CODE DESCRIPTION	SNOMED M3 CODE	SNOMED M3 CODE DESCRIPTION
M-00100	Normal tissue (finding)	M-72440	Glandular and muscular hyperplasia (morphologic abnormality)	M-43000	Chronic inflammation (morphologic abnormality)

Step 6: Insert a new column labelled SNOMED M DESCRIPTIONS and use the Microsoft Excel CONCATENATE function to combine the SNOMED descriptions (1-n). This is required to enable the development of the lookup table. The diagnosis, sub-diagnosis and sub-result would be added to start populating these values based on the description, i.e. Inflammation and hyperplasia.

SNOMED M CODES	SNOMED M DESCRIPTIONS
M-00100,M-72440,M-43000	Normal tissue (finding),Glandular and muscular hyperplasia (morphologic abnormality),Chronic inflammation (morphologic abnormality)

SNOMED, Systemized Nomenclature of Medicine; M, morphology; T, topography.

FIGURE 2: Six-step procedure used to transform the chained comma separated Systemized Nomenclature of Medicine M codes into individual columns (leaving the original value intact) to add the laboratory information system code table descriptions (in preparation for lookup table development). The same procedure was conducted for T codes. The manipulation was achieved using standard Microsoft Excel functions (screenshots included next to each step). The steps are as follows: (1) copy unique chained codes to a new worksheet and then copy and paste to a new column for processing (leaving the original values intact), (2) use the Microsoft Excel text to column function to separate the chained codes and name new columns, e.g. M Code 1-n, (3) insert a new column next to each code column and label as a description column, e.g. M Code Descr 1-n, (4) add the alphabetically sorted laboratory information system systemized nomenclature of medicine code description in a new worksheet, (5) use the Microsoft Excel VLOOKUP function to add the code description (range lookup set at 1 for an exact match), e.g. M-00 100 code description is 'Normal tissue (finding)', and (6) Microsoft Excel CONCATENATE function used to combine code descriptions combined in a new column.

only the matching values from the other tables reported using referential integrity (Table 1).^{26,27} This query contained all the variables for the data analysis.

Systemized Nomenclature of Medicine M code descriptive analysis

The number of prostate biopsies with an M and T code populated was assessed as a contingency table using SAS Enterprise Guide 7.1 (SAS Institute Incorporated, Cary, North Carolina, United States).²⁸ Descriptive analysis was conducted for prostatic biopsies with the M code captured. Test volumes for the top 10 M code combinations with their chained descriptions were reported (in descending order). The number of biopsies of prostatic origin was also reported.

Descriptive analysis of diagnosis and sub-diagnosis

The diagnosis and sub-diagnosis volumes were reported for biopsies with M code populated of prostatic origin. For each

diagnosis, the sub-diagnoses were then reported. Where more than 10 sub-diagnoses were reported, the first 10 were reported and the remaining grouped as 'Other'.

Results

Using our methodology, 25 010 biopsy results were extracted and analysed. For the lookup tables, there were unique 1520 M and 702 T code combinations.

Systemized Nomenclature of Medicine M and T code descriptive analysis

M codes were provided for 22 195/25 010 biopsies (89%; Table 2). The T code was provided for 24 546/25 010 (98%) biopsies. There were 22 009/25 010 biopsies with both an M and T code populated (88%). There were 2815/25 010 (11%) biopsy reports that could not be analysed using lookup tables; they did not have both M and T codes (Table 2). Descriptive analyses were conducted for 20 551/25 010 (82%) biopsies of prostatic origin. Of the 22 195 biopsies with M codes, M-40 000

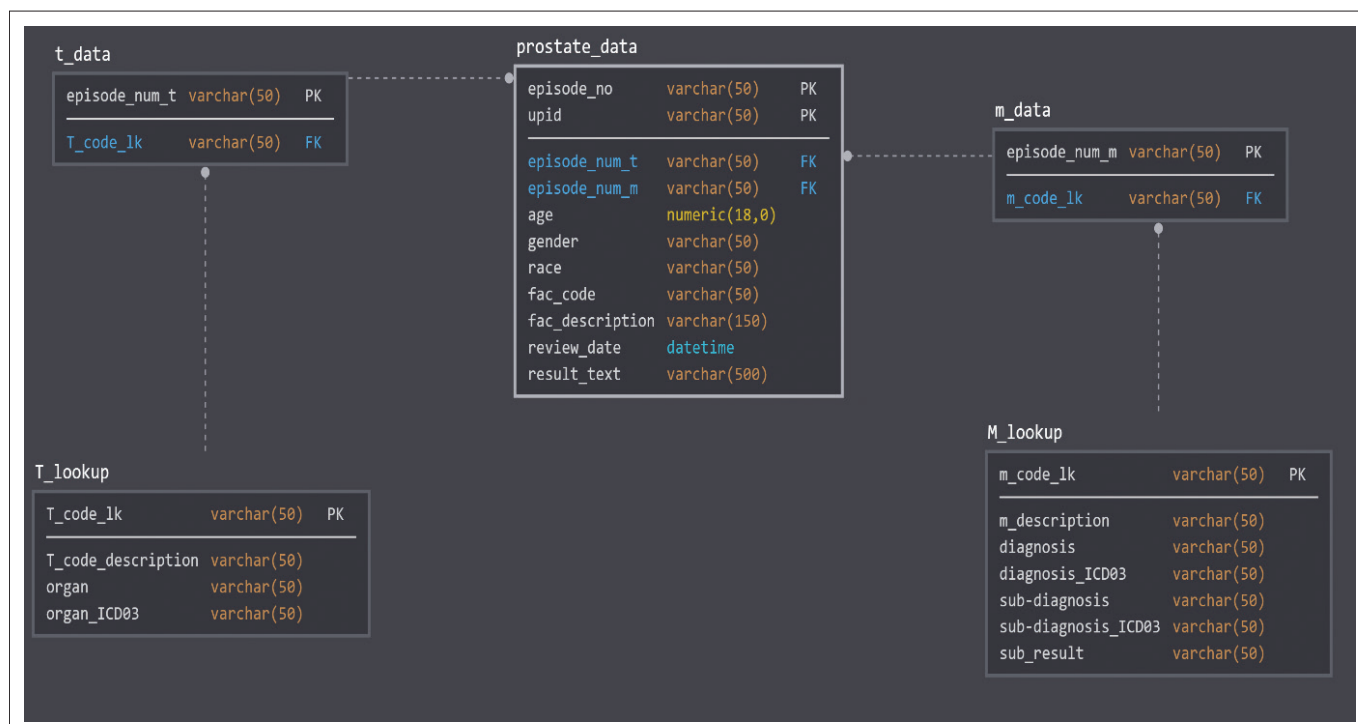


FIGURE 3: Relational database diagram describes how the various tables were joined (left outer). For each table, the primary and foreign keys are provided. The lookup tables contain only the unique Systemized Nomenclature of Medicine code combinations. The lines indicate a relationship join between tables. Once the table joins are implemented, variables from any table can be reported. Structured query language could be used to combine the required data for analysis.

TABLE 1: Example of four prostate biopsies where the systemized nomenclature of medicine code descriptions were assigned a diagnosis and sub-diagnosis including the allocation of ICD-O-3 codes.

Episode no.	SNOMED M codes (T Code)	Biopsy result text	Date of review	Gleason score	SNOMED M descriptions	Diagnosis (ICD-O-3)	Sub-diagnosis (ICD-O-3)
A	M-72 450,M-00 100 (T-92 000)	†	2009/10/19	N/A	Adenofibromyomatous hyperplasia (morphologic abnormality) Normal tissue (finding)	Benign/negative for malignancy (8000/0)	Hyperplasia
B	M-72 000 (T-92 000)	‡	2013/03/12	N/A	Hyperplasia (morphologic abnormality)	Benign/negative for malignancy (8000/0)	Hyperplasia
C	M-00 100,M-81 403 (T-92 000)	§	2014/02/16	4 + 3 = 7	Normal tissue (finding) Adenocarcinoma, no subtype (morphologic abnormality)	Neoplasm, malignant (8000/3)	Adenocarcinoma (8140/2)
D	M-71 000,M-40 000 (T-92 000)	¶	2016/10/28	N/A	Hypertrophy (morphologic abnormality) Inflammation (morphologic abnormality)	Benign/negative for malignancy (8000/0)	Inflammation and hypertrophy

Note: The biopsy result text is also provided. Episode numbers are anonymised.

SNOMED, Systemized Nomenclature of Medicine; M, morphology; T, topography; ICD-O, v International Classification of Diseases for Oncology; N/A, not applicable.

†, episode number: a clinical history: an 86-year-old male patient with a prostate specific antigen = 14.1. Microscopy: sections of the specimen confirm the presence of multiple cores of prostatic tissue with features of benign fibromuscular and adenomatous hyperplasia. Diagnosis: prostate – benign fibromuscular and adenomatous hyperplasia.

‡, episode number: b clinical history: a 57-year-old man. Re-turp. Macroscopy: 8.5 g of chips received, all processed. Microscopy: sections show fragments of prostate tissue exhibiting features of benign nodular glandular and stromal hyperplasia and mild chronic prostatitis. There is no evidence of high-grade prostatic intraepithelial neoplasia or infiltrating malignancy. Pathological diagnosis: Turp: – prostatic hyperplasia and chronic prostatitis – no evidence of malignancy.

§, episode number: c clinical history: an 82-year-old patient with prostate specific antigen = 273. Hard, multinodular prostate. Trus biopsy done. macroscopy: twelve biopsies, the longest 1.8 cm. Microscopy: histological examination shows numerous cores of prostatic tissue, most of which contain extensive infiltration by adenocarcinoma, Gleason 4, 3. focally is perineural invasion seen. Pathological diagnosis: prostate gland: extensive involvement by adenocarcinoma, Gleason 4, 3 focally perineural invasion is seen.

¶, episode number: d clinical details: the patient is a 61-year-old male. The patient has a prostate specific antigen level of 10.4. Macroscopy: 13 cores, the longest of which measured 11 mm and the shortest 2 mm. pathological diagnosis: prostate: (1) sections show representation predominantly of seminal vesicles. (2) There is focal representation of rectal mucosa. (3) Foci of chronic inflammation are identified. (4) Features of benign prostatic enlargement (BPE) are present. There is no evidence of prostatic intraepithelial neoplasia (pin) or invasive adenocarcinoma in the sections examined.

TABLE 2: Contingency table to assess the percentage of Systemized Nomenclature of Medicine M and T codes populated using the mapping table for prostate biopsies between 2006 and 2016 in the Gauteng province, South Africa.

Status	T Code(s) captured		T Code(s) not captured		Total	
	n	%	n	%	n	%
M Code/s captured	22 009	88	186	1	22 195	89
M Code/s not captured	2537	10	278	1	2815	11
Total	24 546	98	464	2	25 010	100

M, morphology; T, topography.

(‘Inflammation’) was the most commonly reported sub-diagnosis ($n = 3400$; 15.3%). This was followed by two adenocarcinoma combinations: M-81 403 ($n = 2677$; 12.1%) and M-81 403, M-80 003 ($n = 2166$; 9.8%) (Table 3).

Descriptive analysis of diagnosis and sub-diagnosis

Using this approach, we noted 10 441 (50.81%) biopsies with a benign diagnosis (ICD-O-3:8000/0) and 9302 (45.26%) biopsies with a malignant diagnosis (8000/3). Atypia was reported for 760 biopsies (3.7%). An uncertain (whether benign or malignant) diagnosis (8000/1) was reported for 48 (0.23%) biopsies (Table 4).

Inflammation was reported as a sub-diagnosis for 4075 benign biopsies (39.03%). This was followed by no pathologic diagnosis and hyperplasia at 26.90% ($n = 2809$) and 20.82% ($n = 2174$) respectively. Inflammation and hyperplasia were

TABLE 3: Top 10 most commonly requested Systemized Nomenclature of Medicine M code combinations from the prostate biopsy data between 2006 and 2016 in the Gauteng province, South Africa.

M codes	M code descriptions	Total	
		n	%
M-40 000	Inflammation (morphologic abnormality)	3440	15.5
M-81 403	Adenocarcinoma, no subtype (morphologic abnormality)	2677	12.1
M-81 403, M-80 003	Adenocarcinoma, no subtype (morphologic abnormality) Malignant neoplasm, primary (morphologic abnormality)	2166	9.8
M-00 100	Normal tissue (finding)	1668	7.5
M-80 003, M-81 403	Malignant neoplasm, primary (morphologic abnormality) Adenocarcinoma, no subtype (morphologic abnormality)	1596	7.2
M-00 100, M-81 403	Malignant neoplasm, primary (morphologic abnormality) Adenocarcinoma, no subtype (morphologic abnormality)	1299	5.9
M-00 100, M-72 440	Normal tissue (finding) Glandular and muscular Hyperplasia (morphologic abnormality)	1081	4.9
M-80 003	Malignant neoplasm, primary (morphologic abnormality)	706	3.2
M-09 350	Morphologic description only (finding)	394	1.8
M-72 440	Glandular and muscular hyperplasia (morphologic abnormality)	370	1.7

M, morphology.

TABLE 4: Descriptive analysis of diagnosis and sub-diagnosis where both a Systemized Nomenclature of Medicine M and T code are populated of prostatic origin between 2006 and 2016 in the Gauteng province, South Africa.

Lookup table value	n	%
Diagnosis		
Benign/negative for malignancy (8000/0)	10 441	50.81
Neoplasm, malignant (8000/3)	9302	45.26
Atypia/dysplasia (8000/0)	760	3.70
Neoplasm, uncertain whether benign or malignant (8000/1)	48	0.23
Total	20 551	-
Sub-diagnosis		
Benign/negative for malignancy		
Inflammation	4075	39.03
No pathologic diagnosis	2809	26.90
Hyperplasia	2174	20.82
Inflammation and hyperplasia	788	7.55
Inflammation and hypertrophy	123	1.18
Atrophy	120	1.15
Hypertrophy	82	0.879
Inflammation and atrophy	63	0.60
Hyperplasia and atrophy	55	0.53
Inflammation, hypertrophy and hyperplasia	31	0.30
Other	121	1.16
Total	10 441	-
Neoplasm, malignant sub-diagnosis		
Adenocarcinoma (8140/3)	8201	88.16
Carcinoma (8010/3)	408	4.39
Malignant neoplasm (8000/3)	693	7.45
Total	9302	-
Atypia/dysplasia sub-diagnosis		
Atypia	616	81.05
Dysplasia	87	11.45
Atypia/dysplasia	44	5.79
High grade intraepithelial lesion (8148/2)	13	1.71
Total	760	-
Neoplasm, uncertain whether benign or malignant sub-diagnosis		
Neoplasm, uncertain whether benign or malignant (8000/1)	48	0.23

Note: Where appropriate, the ICD-O-3 codes are provided in brackets.

M, morphology; T, topography; ICD-O, International Classification of Diseases for Oncology.

reported in two combinations contributing 7.85% of the top 10 benign sub-diagnoses. Cumulatively, the top 10 sub-diagnoses reported represented 98.8% ($n = 10\ 320$) (Table 4) of all benign diagnoses. The majority of samples with a malignant diagnosis reported an adenocarcinoma (8201, 88.16%)

sub-diagnosis. There were 408 biopsies with a carcinoma (4.39%) sub-diagnosis. The malignant neoplasm sub-diagnosis was reported for 693 (7.45%) biopsies where it was not possible to differentiate the tissue type, reporting predominantly the M-80 003 code (Malignant neoplasm, primary) (personal communication, Vreede H, August 12, 2017). The majority of biopsies with an atypia/dysplasia diagnosis reported an atypia sub-diagnosis ($n = 616$; 81.05%) followed by dysplasia ($n = 87$; 11.45%). Only 1.71% of biopsies with an atypia/dysplasia diagnosis reported a high grade intraepithelial lesion ($n = 13$). Only 48 (0.23%) biopsies reported an uncertain sub-diagnosis (8000/1).

Discussion

We showed that it was possible to automate prostate biopsy reporting using a commonly available relational database (Microsoft Access) and SNOMED lookup tables in the Gauteng province. The use of ICD-O-3 codes for malignant findings facilitate PCa reporting similar to cancer registries.¹³

To routinely automate the registration and surveillance of PCa in South Africa, the lookup tables developed for this study would need to be introduced to the corporate data warehouse. Lookup tables are routinely used by the Corporate Data Warehouse (CDW) to transform laboratory data for reporting, for example the HIV serology results reported as 'NEG', 'N' or 'NEGATIVE' are transformed to a single value ('NEG') for uniform reporting.²⁸ The benefit of this mechanism for PCa reporting is that as biopsies are reported in the LIS, the data replicated to the CDW will be conformed to report the biopsy diagnosis and sub-diagnosis within three months of diagnosis.

Lookup tables would facilitate a constant feed of analysed PCa data to the South African NCR to ensure timeous reporting. Over time, any new SNOMED code combinations identified would have to be added to the lookup table. By providing this data at shorter intervals, it would be possible to triangulate against other local data sources (NCR and other). Triangulation is the process used in public health to review and interpret data from multiple sources that answer the same question for decision making.³⁰ Unpublished data from this study revealed that between 2012 and 2016, PCa incidence has increased from 44.92 to 57.31 per 100 000 compared to 46.53 reported by the NCR in 2012.³¹

Another advantage of this approach is that PCa data from other African countries using a LIS could also be analysed using the developed lookup tables to dramatically improve PCa reporting across Africa. Antoni et al. assessed the methods used for reporting the 2018 Global Cancer Statistics estimates.³² For 14/51 African countries, PCa incidence estimates were based on simple average rates from neighbouring countries (27%).³² For South Africa, projections of national incidence were sourced from the NCR.³² The SNOMED lookup tables have the potential to improve both national and regional PCa incidence reporting across the African continent providing more accurate data. With better data, cancer control initiatives could be better mobilised.

The principles applied in our study could also be implemented for other cancers. The SNOMED codes are captured for all cancers of public health importance routinely. Similar lookup tables could be developed to report on lung, breast and cervical cancers with incidence rates of 17.3, 49.0 and 13.5 per 100, 1000 respectively in 2018 in South Africa.³³

The approach described in our study is not unique. Similar approaches have been employed in the Danish cancer registry, where data reported for 161 525 biopsies for the period 1995–2011 were undertaken using SNOMED codes.²² The Danish cancer registry predominantly reported data for PCa, whereas our study reported data for negative biopsies as well. The combination of the lookup tables reported in our study with text mining to extract the Gleason score reported in an unpublished study could be used to report data similar to the Danish cancer registry, for example diagnosis of 'Neoplasm, malignant', sub-diagnosis of 'Adenocarcinoma' and a Gleason score of $3 + 3 = 6$ would be coded as 'bGS3+3'.²²

It is important to provide up to date PCa data at both the national, provincial, district and health facilities levels to identify hotspots where programmatic interventions are required. SNOMED lookup tables could be used to focus programmatic interventions for geographic areas with a higher PCa burden. Similar initiatives using laboratory data have been undertaken locally for HIV and tuberculosis services.^{24,34,35} An example is the World Bank report that described spatial clustering analysis of HIV viral load suppression at the national, provincial, district, sub-district and health facility levels.²⁴ This study indicated that national laboratory data stored in the CDW has the potential to provide important strategic information on the quality and reach of the antiretroviral therapy programme by highlighting the geographical variation in the proportion of patients virally suppressed.²⁴ This information can be accessed by healthcare workers using the epidemiological dashboard developed to identify health facilities that are performing poorly.^{24,34} Similar work has also been published using cluster of differentiation 4 data to highlight areas where HIV-positive patients presenting for care have a higher burden of advanced disease.³⁴ The assimilation of health data into workable and user-friendly dashboards has had a big impact on how health data is used locally.^{36,37,38} In the medium term, it would be possible to develop a PCa epidemiological dashboard similar to the HIV example mentioned to report PCa trends by age, race group and geographical boundaries routinely. The PCa dashboard could facilitate the reporting of the number of incident cases. The dashboard could also provide insights into how and where health care services are being accessed to inform both guideline changes and programmatic improvements to facilitate equitable access to care across the country. It could also provide loss to follow up and waiting time data for patients who had presented with an elevated prostate specific antigen and who were confirmed with PCa histologically using the CDW probabilistic matching algorithm.²⁹

Data reported in our study additionally demonstrated functionality by providing data for benign histological

findings, potentially useful to identify trends for patients diagnosed with chronic inflammation who eventually progress to PCa. While the ICD-O-3 codes are particularly important for PCa cases, the importance of additional benign findings is especially important for inflammation and hypertrophy that have been shown to be linked with many other cancers.^{39,40} Several studies have indicated that chronic inflammation has a potential role in prostatic carcinogenesis and tumour progression.^{39,41,42} Nelson et al. reported that chronic or recurrent inflammation may play a role in the development of PCa.⁴³ Using the data generated using the lookup tables, patients with chronic inflammation could be followed up to identify whether they progress to PCa in an African setting.

Finally, to improve SNOMED reporting, it is recommended that the M and T codes be defined as mandatory fields. This will ensure that 100% of biopsies include these codes. This will address the 12% of biopsies without this information. Future research includes extending the lessons learnt with PCa in the use of lookup tables to other common cancers, as patient-level data is already available in the CDW and could be easily unlocked for national cancer reporting.

Limitations

One of the limitations is that the data for our study was limited to biopsies sent to the NHLS and did not include data from private sector laboratories and thus limits the generalisability of our study. As private sector laboratories also use the SNOMED code, discussions will be initiated with them to share PCa data for national reporting.

An additional limitation of our data is the 10% of biopsies excluded a SNOMED M or T code. Unfortunately, these fields are not mandatory and can be uncaptured. By amending rules and making these fields mandatory, all biopsies would prospectively be reviewed with at least one T and M code captured. The excluded data would also affect the generalisability of our study. We are not able to determine the findings for these biopsies. To address this gap, text mining and machine learning approaches are being investigated. A sample of the biopsy data will be used to train the machine learning models i.e. malignancy (1) and benign (0). The big data tools will be validated against well populated SNOMED data entailing grid search, *k*-fold cross validation, precision, recall and *F*-score. The combination of SNOMED, text mining and machine learning will hopefully address the missing data. With minor changes to the laboratory information system, this has the potential to report PCa histological findings for all biopsies.

Key messages

Existing national laboratory data has been used for the first time in South Africa to report PCa diagnosis across a province using SNOMED lookup tables. This could be implemented across South Africa to provide timely PCa trends.

Conclusion

Our study has demonstrated that it is possible to automate PCa reporting using SNOMED codes for 88% of biopsies. The value of national laboratory data as shown in our study can easily be extended to deliver the timely monitoring of PCa trends across South Africa and other African countries. This could also be applied to report data for other cancers of public health interest.

Acknowledgements

The authors thank Professors Martin Hale (Anatomical Pathology) and Mohammed Haffejee (Urology) for their insights into PCa diagnosis. We would also like to thank Drs Elvira Singh and Mazvita Sengayi from the National Cancer Registry. For providing the data for the study, we would like to thank the Academic Affairs, Research and Quality Assurance department, as well as Tinyiko Ngobeni, a data analyst at the corporate data warehouse.

Competing interests

The authors declare that they have no conflicting interests.

Authors' contributions

J.A.G. and D.K.G. designed and supervised the study by providing leadership and oversight. R.W. and A.A. assisted with the SNOMED lookup table development. N.C. developed the methodology and conducted the research. All authors contributed to writing the initial draft. J.A.G. and D.K.G. were responsible for reviewing the article and overall project administration.

Sources of support

The authors declare that they have not received any funding for this study.

Data availability statement

The authors do not have permission to share the data generated for this study.

Disclaimer

The authors declare that the views expressed in the submitted article are our own and not the official position of any institution or funder.

References

- Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, cancer incidence and mortality worldwide: IARC CancerBase No. 1 [homepage on the Internet]. Lyon: International Agency for Research on Cancer (IARC); c2013 [cited 2018 Feb 28]. Available from: <http://globocan.iarc.fr>
- International Agency for Research on Cancer (IARC). GLOBOCAN 2012: Estimated cancer incidence, mortality and prevalence worldwide [homepage on the Internet]. 3rd ed. Geneva: IARC; c2013 [cited 2018 Feb 28]. Available from: http://globocan.iarc.fr/old/summary_table_site.html.asp?selection=24191&title=Prostate&sex=1&type=0&window=1&africa=1&sort=0&submit=%C2%A0Execute
- Heyns CF, Fisher M, Lecuona A, Van Der Merwe A. Prostate cancer among different racial groups in the Western Cape: Presenting features and management. *SAMJ*. 2011;101(9):267–270. <https://doi.org/10.7196/samj.4420>
- Le Roux HA, Urry RJ, Sartorius B, Aldous C. Prostate cancer at a regional hospital in South Africa: We are only seeing the tip of the iceberg. *S Afr J Surg*. 2015;53(3 and 4):57–62.
- World Health Organization (WHO). National Cancer Control Programmes (NCCP) [homepage on the Internet]. Geneva: World Health Organization (WHO); c2002 [cited 2016 May 01]. Available from: <https://apps.who.int/iris/bitstream/handle/10665/42494/9241545577.pdf;jsessionid=7B75F2254A4A5DC71445907B939A1709?sequence=1>
- World Health Organization (WHO). National Cancer Control Programme: Policies and managerial guidelines [homepage on the Internet]. Geneva: World Health Organization (WHO); c2002 [cited 2016 May 04]. Available from: <http://apps.who.int/iris/bitstream/handle/10665/42494/9241545577.pdf;sequence=1>
- World Health Organization (WHO). Global action plan for the prevention and control of NCDs 2013–2020 [homepage on the Internet]. Geneva: World Health Organization (WHO); c2013 [cited 2016 May 01]. Available from: http://apps.who.int/iris/bitstream/10665/94384/1/9789241506236_eng.pdf?ua=1
- U.S. National Institutes of Health (NIH) NCIN. SEER training modules: Cancer registry [homepage on the Internet]. Bethesda, MD: U.S. National Institutes of Health (NIH), National Cancer Institute (NCI); c2018 [cited 2018 May 09]. Available from: <https://training.seer.cancer.gov/registration/registry/>
- Centers for Disease Control and Prevention (CDC) NCCDPaHP, Division of Cancer Prevention and Control, National Program of Cancer Registries. National Program of Cancer Registries: Cancer Surveillance System Rationale and Approach [homepage on the Internet]. USA: Centers for Disease Control and Prevention (CDC), Georgia, Atlanta, USA; c1999 [cited 2018 May 09]. Available from: https://www.cdc.gov/cancer/npcr/pdf/npcr_css.pdf
- Singh E, Sengayi M, Urban M, Babb C, Kellett P, Ruff P. The South African National Cancer Registry: An update. *Lancet Oncol*. 2014;15(9):e363. [https://doi.org/10.1016/S1470-2045\(14\)70310-9](https://doi.org/10.1016/S1470-2045(14)70310-9)
- National Cancer Registry (NCR). Methodology: Data collection and data flow [homepage on the Internet]. Sandringham: National Institute of Communicable Diseases (NICD); c2006 [cited 2018 Feb 26]. Available from: <http://www.nicd.ac.za/index.php/centres/national-cancer-registry/methodology/>
- Schonfeld SJ, Erdmann F, Wiggill T, et al. Hematologic malignancies in South Africa 2000–2006: Analysis of data reported to the National Cancer Registry. *Cancer Med*. 2016;5(4):728–738. <https://doi.org/10.1002/cam4.597>
- International Agency for Research on Cancer (IARC), World Health Organization (WHO). International classification of diseases for oncology (ICD-O) [homepage on the Internet]. 3rd ed. Geneva: IARC; c2013 [cited 2018 Apr 25]. Available from: <http://codes.iarc.fr/>
- National Cancer Registry. Summary statistics of cancer diagnosed histologically in South Africa in 2014. Johannesburg: National Institute for Occupational Health (NIOH); 2014.
- National Institute for Communicable Diseases (NICD). National Cancer Registry (NCR) methodology [homepage on the Internet]. Sandringham: National Institute for Communicable Diseases (NICD); c2017 [cited 2018 Apr 25]. Available from: <http://nicd.ac.za/?page=methodology&id=251>
- Thomas AA, Zheng C, Jung H, et al. Extracting data from electronic medical records: Validation of a natural language processing program to assess prostate biopsy results. *World J Urol*. 2014;32(1):99–103. <https://doi.org/10.1007/s00345-013-1040-4>
- Banerjee I, Li K, Seneviratne M, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. *JAMIA Open*. 2019;2(1):150–159. <https://doi.org/10.1093/jamiaopen/ooy057>
- Leyh-Bannurah SR, Tian Z, Karakiewicz PI, et al. Deep learning for natural language processing in urology: State-of-the-Art automated extraction of detailed pathologic prostate cancer data from narratively written electronic health records. *JCO Clin Cancer Inform*. 2018;2(1):1–9. <https://doi.org/10.1200/CCI.18.00080>
- U.S. National Library of Medicine. Overview of SNOMED CT [homepage on the Internet]. Bethesda, MD: U.S. National Library of Medicine; c2016 [cited 2018 Feb 20]. Available from: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html
- SNOMED International. What is the difference between a classification such as ICD-10 or ICPC and a terminology like SNOMED CT? [homepage on the Internet]. London: SNOMED International; c2015 [cited 2018 Feb 20]. Available from: <https://ihtsdo.freshdesk.com/support/solutions/articles/4000052506-what-is-the-difference-between-a-classification-such-as-icd-10-or-icpc-and-a-terminology-like-snomed->
- SNOWMED International. SNOMED CT starter guide [homepage on the Internet]. London: SNOWMED International; c2017 [updated 2017 Jul 28; cited 2018 Feb 20]. Available from: https://confluence.ihtsdotools.org/display/DOCSTART?previeew=/28742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf
- Helgstrand JT, Klemann N, Roder MA, et al. Danish prostate cancer registry – Methodology and early results from a novel national database. *Clin Epidemiol*. 2016;8(1):351–360. <https://doi.org/10.2147/CLEP.S114917>
- IBM Netezza. Netezza Aginity Workbench Marlborough [homepage on the Internet]. Marlborough, MA, USA: IBM Netezza [Netezza Aginity Workbench]; c2018 [cited 2018 Feb 20]. Available from: <https://www.agnity.com/documentation/WB/NZ/Default.htm>
- World Bank. Analysis of big data for better targeting of ART adherence strategies: Spatial clustering analysis of viral load suppression by South African Province, District, Sub-District and Facility (April 2014–March 2015) [homepage on the Internet]. Washington, DC: World Bank; c2015 [cited 2018 Apr 24]. Available from: <https://openknowledge.worldbank.org/handle/10986/25399>
- Microsoft Corporation. Microsoft Office Professional Plus 2013. Microsoft Office Professional Plus 2013 ed. Redmont, Washington, DC: Microsoft Corporation; 2013. p. Microsoft Office Professional Plus 2013.

26. Microsoft Corporation: TechNet. Using outer joins Albuquerque [homepage on the Internet]. Redmont, Washington DC, USA: Microsoft Corporation; c2018 [cited 2018 Apr 24]. Available from: [https://technet.microsoft.com/en-us/library/ms187518\(v=sql.105\).aspx](https://technet.microsoft.com/en-us/library/ms187518(v=sql.105).aspx)
27. Microsoft Corporation. Guide to table relationships [homepage on the Internet]. Redmont, WA: Microsoft Corporation; c2018 [cited 2018 May 12]. Available from: <https://support.office.com/en-us/article/guide-to-table-relationships-55b8db2c-9480-4269-b1bb-f6ec09623dfd>
28. SAS Institute Inc. SAS Enterprise Guide [homepage on the Internet]. Cary, NC, USA: SAS Institute Inc; c2017 ([7.15 HF2] (7.100.5.6112) (64-bit) [cited 2019 Mar 12]. Available from: <http://support.sas.com/software/products/enterprise-guide/index.html>.
29. Ngobeni T. Corporate data warehouse lookup tables. Sandringham: National Health Laboratory Service (NHLS); 2017.
30. Rutherford GW, McFarland W, Spindler H, et al. Public health triangulation: Approach and application to synthesizing data to understand national and local HIV epidemics. *BMC Public Health*. 2010;10(1):447. <https://doi.org/10.1186/1471-2458-10-447>
31. Cassim N, Ahmad A, Wadee R, Rebbeck T, Glencross DK, George JA. Prostate cancer age-standardised incidence increase between 2006 and 2016 in the Gauteng province, South Africa: a laboratory data-based analysis. [Article in press 2020 (SAMJ)].
32. Antoni S, Soerjomataram I, Moller B, Bray F, Ferlay J. An assessment of GLOBOCAN methods for deriving national estimates of cancer incidence. *Bull World Health Organ*. 2016;94(3):174–184. <https://doi.org/10.2471/BLT.15.164384>
33. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424. <https://doi.org/10.3322/caac.21492>
34. Coetzee LM, Cassim N, Glencross DK. Analysis of HIV disease burden by calculating the percentages of patients with CD4 counts <100 cells/microL across 52 districts reveals hot spots for intensified commitment to programmatic support. *S Afr Med J*. 2017;107(6):507–513. <https://doi.org/10.7196/SAMJ.2017.v107i6.11311>
35. Nanno A, Izu A, Ismail NA, et al. Nationwide and regional incidence of microbiologically confirmed pulmonary tuberculosis in South Africa, 2004–12: A time series analysis. *Lancet Infect Dis*. 2015;15(9):1066–1076. [https://doi.org/10.1016/S1473-3099\(15\)00147-4](https://doi.org/10.1016/S1473-3099(15)00147-4)
36. Coetzee L-M, Cassim N, Glencross DK. Using laboratory data to categorise CD4 laboratory turn-around-time performance across a national programme. *Afr J Lab Med*. 2018;7(1):a665. <https://doi.org/10.4102/ajlm.v7i1.665>
37. World Bank. Analysis of big data for better targeting of ART adherence strategies: Spatial clustering analysis of viral load suppression by South African Province, District, Sub-District and Facility (April 2014–March 2015). Washington, DC: World Bank; 2015.
38. Tepper EE, Cassim N, Coetzee LM, Glencross DK. Introducing an easily assessable desktop tool for laboratory management review of global laboratory turn-around time (TAT) performance across an extended network. 56th International Congress of the Federation of South African Societies of Pathology (FSASP); 16–18 August 2018; Cape Town, South Africa; 2018.
39. Jank BJ, Kadletz L, Schnoll J, Selzer E, Perisanidis C, Heiduschka G. Prognostic value of advanced lung cancer inflammation index in head and neck squamous cell carcinoma. *Eur Arch Otorhinolaryngol*. 2019;276(5):1487–1492. <https://doi.org/10.1007/s00405-019-05381-0>
40. Kobayashi S, Karube Y, Inoue T, et al. Advanced lung cancer inflammation index predicts outcomes of patients with pathological stage IA lung adenocarcinoma following surgical resection. *Ann Thorac Cardiovasc Surg*. 2019;25(2):87–94. <https://doi.org/10.5761/atcs.0a.18-00158>
41. Sfanos KS, De Marzo AM. Prostate cancer and inflammation: The evidence. *Histopathology*. 2012;60(1):199–215. <https://doi.org/10.1111/j.1365-2559.2011.04033.x>
42. Leitzmann MF, Rohrmann S. Risk factors for the onset of prostatic cancer: Age, location, and behavioral correlates. *Clin Epidemiol*. 2012;4(1):1–11. <https://doi.org/10.2147/CLEP.S16747>
43. Nelson WG, De Marzo AM, Isaacs WB. Prostate cancer. *N Engl J Med*. 2003;349(4):366–381. <https://doi.org/10.1056/NEJMra021562>